Exploring the added value of Whole Genome Sequencing in routine and pandemic viral surveillance

Laura Van Poelvoorde

Supervisors:
Prof. Dr. Xavier Saelens
Department of Biochemistry and Microbiology, Ghent University, Ghent, Belgium
VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

Dr. Nancy Roosens
Transversal activities in Applied Genomics, Sciensano, Brussels, Belgium

Onderzoek naar de toegevoegde waarde van Whole Genome Sequencing gegevens in routine en pandemische virale surveillance

---

Thesis submitted in partial fulfilment of the requirements for the degree
of Doctor of Science: Biochemistry and Biotechnology

Academic year: 2022-2023

Members of the Examination Committee:

Chair: Prof dr. Savvas Savvides
Department of Biochemistry and Microbiology, Ghent University, Ghent, Belgium
VIB-UGent Center for Inflammation Research, VIB, Ghent, Belgium

Secretary: Prof dr. Marie Joossens
Department of Biochemistry and Microbiology, Ghent University, Ghent, Belgium

Dr. Philippe Lemey
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven

Prof dr. Lieve Naesens
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven

Dr. Sebastiaan Theuns
Department of Translational Physiology, Infectiology and Public Health, Ghent University, Ghent, Belgium
PathoSense, Lier, Belgium

Prof dr. Steven Van Gucht
Viral diseases, Sciensano, Brussels, Belgium

SUMMARY

Whole genome sequencing (WGS) has opened up a lot of new possibilities for virus surveillance, however, many applications have remained unexploited in the context of routine surveillance, which is the core business of national reference centres and public health institutes such as Sciensano. The virus surveillance aims to monitor the circulating strains by collecting viral genomic data, characterising the viruses and if possible to link it to patient data. It is in particular important to evaluate the pathogenicity, vaccine and antiviral drug susceptibility of these circulating strains. WGS enables tracking viral outbreaks and estimating the virus spread in a given population, how fast the virus is mutating, as well as the impact of genetic modifications on human disease. This thesis aims to explore the added value and challenges of this genomic approach for virus surveillance and how to overcome these challenges in order to receive the highest benefit from these approaches. Within this thesis, we were mainly focused on respiratory samples in the influenza surveillance and wastewater samples in the SARS-CoV-2 surveillance.

Regarding the influenza genomic surveillance, we have first evaluated the possible added value of using WGS to identify viruses with one or more mutations that are associated with antiviral drug resistance. Due to the extensive use of neuraminidase inhibitors to treat influenza, the antiviral influenza drug susceptibility has long focussed on the neuraminidase segment. We showed that with the emergence of new antiviral drugs that target other segments, there is an increased need to obtain information about the whole influenza genome and we evaluated how WGS could be implemented to detect drug resistance mutations in clinical influenza virus isolates.

Secondly, we assessed how WGS of clinical samples can improve the routine surveillance of circulating influenza strains in humans. Indeed, the hemagglutinin segment has been the principal target region in classical influenza surveillance programmes, besides neuraminidase. Consequently, relatively little information is available about the other segments. Much needed improvement of the influenza genetic surveillance can be provided by using WGS, which can facilitate inferring potential links between genomic data from the whole genome and disease and host characteristics. In this thesis, a new way of classifying the influenza viruses based on the whole genome was proposed. This may allow an improved vaccine strain selection, but also for the future next-generation antiviral drugs and vaccines that will not solely focus on neuraminidase and hemagglutinin. Moreover, mutations across the whole genome and reassortments could be detected and linked to patient data with significant associations as a result. Furthermore, because of the high diversity within influenza subtypes, a new approach was proposed that classifies the influenza viruses based on their phylogenetic relatedness and reduces the viral genetic background before analysing the mutations in relation to the patient data. Besides the advantage of obtaining the whole genome in one reaction, WGS also offers the opportunity to sequence patient-derived virus population at sufficiently high depths to identify low-frequency variants present in a viral quasispecies. However, due to experimental errors from the PCR and NGS, it is a challenge to distinguish these low-frequency variants from the experimental errors while considering the limiting circumstances of a routine setting where it is improbable that samples will be sequenced multiple times in order to identify more easily the experimental errors. Therefore, we proposed a general approach to identify these low-frequency variants that ensures high-quality results and remains feasible using clinical samples in routine surveillance. Although the approaches were successfully developed in this thesis, the results are presented as a proof of concept because of the limited number of influenza samples that were available within the Belgian influenza dataset. The challenge of having a middle-sized collection will probably be encountered by most countries due to the cost of WGS. This highlights the need for a public worldwide database that contains patient data that is linked in a harmonised way with genomic data. This enables the analysis of results obtained at a local level, and compare them at the global level.

Not only the influenza surveillance on clinical samples can benefit from WGS, WGS can also benefit the SARS-CoV-2 surveillance. We focused on monitoring SARS-CoV-2 using wastewater surveillance. First, we looked at how the global effort to sequence SARS-CoV-2 whole genomes could be an improvement over the current surveillance based on polymerase chain reaction techniques through the design and *in silico* evaluation of primers and probes while considering a broad spectrum of variants. Therefore, to deal with this challenge we propose an approach that allows the evaluation of the *in silico* specificity of the assay based on publicly available WGS data combined with minimal experimental testing to evaluate the *in vitro* performance of the assay for respiratory and wastewater samples. Moreover, wastewater samples that contain human faeces have the advantage that it includes multiple variants and reflects the circulating strains in a given population at a specific time. Therefore, by developing an analytical sequencing strategy that identifies and measures all circulating variants in a sample, a good global picture of the epidemiological spread and evolution of circulating strains is obtained without a priori knowledge. In this thesis, we evaluated whether high-throughput sequencing of a sample that has been enriched by PCR and that targets the whole SARS-CoV-2 genome, would be able to identify and quantify low-frequency variants. To start to address this question, an *in silico* dataset has been constructed by mixing wild-type sequencing data obtained by PCR enrichment and by introducing mutations of interest in raw wild-type sequencing data. The SARS-CoV-2 B.1.1.7 lineage was used as a case study. The use of such *in silico* datasets to mimic the diversity in SARS-CoV-2 variants in wastewater, allowed the development of a workflow and the set-up of minimal quality criteria in order to take full advantage of the opportunities of NGS to try and define the population of SARS-CoV-2 and its variants present in wastewater. This will enable trying out this approach on real wastewater samples in the near future.

SAMENVATTING

Whole genome sequencing (WGS) heeft veel nieuwe mogelijkheden voor virale surveillance geopend, maar veel toepassingen zijn onbenut gebleven in de context van routinematige surveillance, de kernactiviteit van nationale referentielaboratoria en volksgezondheidsinstituten zoals Sciensano. De virale surveillance heeft als doel de circulerende influenza virussen te monitoren door virale genetische gegevens te verzamelen, de virussen te karakteriseren en indien mogelijk te linken aan patiëntgegevens. Het is namelijk belangrijk om de pathogeniteit en de gevoeligheid voor vaccins en antivirale geneesmiddelen van deze circulerende stammen te evalueren. WGS maakt het mogelijk om virale uitbraken op te volgen en het staat toe om de virale verspreiding in een bepaalde populatie, de mutatiesnelheid van het virus en de impact van genetische modificaties op menselijke ziekten in te schatten. Deze thesis heeft tot doel de toegevoegde waarde en uitdagingen van sommige van deze genomische methodes voor virale surveillance te onderzoeken en hoe deze uitdagingen kunnen worden overwonnen om het meeste voordeel uit deze methodes te halen. In deze thesis hebben we ons voornamelijk gericht op respiratoire stalen in de griepsurveillance en afvalwaterstalen in de SARS-CoV-2-surveillance.

Met betrekking tot de genomische surveillance van influenza hebben we eerst de mogelijke toegevoegde waarde geëvalueerd van het gebruik van WGS om de aanwezigheid van virussen te detecteren met een of meer mutaties die geassocieerd zijn met resistentie tegen antivirale geneesmiddelen. Als gevolg van uitgebreid gebruik van neuraminidase inhibitoren voor de behandeling van influenza, heeft de gevoeligheid voor antivirale influenzageneesmiddelen zich lang geconcentreerd op het neuraminidase segment. We laten zien dat door de opkomst van nieuwe antivirale geneesmiddelen die gericht zijn op andere segmenten, er een grotere behoefte is om informatie te verkrijgen over het hele influenzagenoom en we hebben geëvalueerd hoe WGS kan worden geïmplementeerd om resistentiemutaties tegen antivirale middelen te detecteren in klinische influenza virus stalen.

Ten tweede hebben we onderzocht hoe WGS van klinische stalen de routinematige surveillance van circulerende influenza virussen bij mensen kan verbeteren. Het hemagglutinine segment was de belangrijkste target in klassieke surveillanceprogramma's voor griep, naast neuraminidase. Er is dan ook relatief weinig informatie beschikbaar over de andere segmenten. De cruciale verbetering van de genetische surveillance van influenza kan worden tegemoetgekomen door WGS te gebruiken, wat het afleiden van potentiële verbanden tussen genomische gegevens van het hele genoom en de patiëntgegevens kan vergemakkelijken. In deze thesis werd een nieuwe manier voorgesteld om de influenza

virussen te classificeren op basis van het hele genoom. Dit maakt een verbeterde selectie van vaccinstammen mogelijk voor de huidige vaccins, maar ook voor de toekomstige antivirale geneesmiddelen en next-generation vaccins die zich niet alleen zullen richten op neuraminidase en hemagglutinine. Bovendien konden mutaties over het hele genoom en reassortanten worden gedetecteerd en gekoppeld aan patiëntgegevens met significante associaties als resultaat. Bovendien, vanwege de grote diversiteit binnen de subtypes van influenza, werd een nieuwe methode voorgesteld die de influenza virussen classificeert op basis van hun fylogenetisch verwantschap en op die manier de virale genetische achtergrond vermindert voordat de mutaties in relatie tot de patiëntgegevens worden geanalyseerd. Naast het voordeel van het verkrijgen van het hele genoom in één reactie, biedt WGS ook de mogelijkheid om van de viruspopulatie van een patiënt aan lage frequenties te sequencen om de varianten met een lage frequentie, die aanwezig zijn in een virale quasispecies, te identificeren. Vanwege de experimentele fouten van de PCR en NGS is het echter een uitdaging om deze varianten aan een lage frequentie te onderscheiden van de experimentele fouten, rekening houdend met de beperkende omstandigheden van een routinelaboratorium waar het onwaarschijnlijk is dat stalen meerdere keren worden gesequenced om experimentele fouten gemakkelijker te identificeren. Daarom stellen we een algemene aanpak voor waarbij deze varianten aan een lage frequentie worden geïdentificeerd waarbij een hoge kwaliteit van de resultaten wordt gegarandeerd en haalbaar blijft met behulp van klinische stalen in routinematige surveillance. Hoewel de methodes met succes werden ontwikkeld in deze thesis, worden de resultaten gepresenteerd als een proof of concept vanwege het beperkte aantal griepstalen binnen de Belgische griepdataset. De uitdaging van het hebben van een middelgrote collectie zal waarschijnlijk door de meeste landen worden ondervonden vanwege de kosten van WGS. Dit onderstreept de behoefte aan een openbare wereldwijde database met patiëntgegevens die op een geharmoniseerde manier is gekoppeld aan genomische gegevens. Dit zou het mogelijk maken om de resultaten die op lokaal niveau zijn verkregen, te vergelijken en op een globaal niveau.

Niet alleen de griepsurveillance op klinische stalen kan profiteren van WGS, daarom hebben we ons voor de SARS-CoV-2 surveillance gericht op het monitoren van dit virus met behulp van de afvalwater surveillance. Eerst hebben we gekeken hoe de wereldwijde inspanning om hele SARS-CoV-2 genomen te sequencen de huidige surveillance op basis van PCR zou kunnen verbeteren door het ontwerp en *in silico* evaluatie van primers en probes, terwijl we een grote waaier aan varianten in overweging nemen. Om deze uitdaging aan te gaan, stellen we daarom een methode voor die de evaluatie van de *in silico* specificiteit van de test mogelijk maakt op basis van openbaar beschikbare WGS-gegevens in combinatie met minimale experimentele tests om de *in vitro* resultaten van de test voor klinische en afvalwaterstalen te evalueren. Bovendien hebben afvalwaterstalen die menselijke faeces bevatten het voordeel dat het meerdere varianten bevat en de circulerende virussen in een bepaalde populatie op een bepaald moment weerspiegelt. Door een analytische sequentiestrategie te ontwikkelen die alle circulerende varianten in een staal identificeert en meet, wordt daarom een goed globaal beeld verkregen van de epidemiologische verspreiding en evolutie van circulerende stammen zonder a priori kennis. In deze thesis hebben we geëvalueerd of high-throughput sequencing van een staal door middel van PCR-verrijkte targeting van het hele SARS-CoV-2 genoom in staat zou zijn om varianten aan een lage frequentie te identificeren en te kwantificeren. Om deze vraag te beantwoorden, is een *in silico* dataset samengesteld door wildtype sequencing gegevens die zijn verkregen door PCR targeting te mengen en door interessante mutaties in wild-type sequencing gegevens te introduceren. De SARS-CoV-2 B.1.1.7-afstamming werd gebruikt als case studie. Het gebruik van een dergelijke *in silico* dataset om de diversiteit in SARS-CoV-2 varianten in afvalwater na te bootsen, maakte de ontwikkeling mogelijk van een workflow en het opstellen van minimale kwaliteitscriteria om ten volle te profiteren van de mogelijkheden van NGS om te proberen de populatie van SARS-CoV-2 en zijn varianten die aanwezig zijn in afvalwater te definiëren. Hierdoor zal deze aanpak in de nabije toekomst op echte afvalwaterstalen kunnen worden uitgeprobeerd.

ACKNOWLEDGMENTS

The past five years working on my thesis has been an incredible experience both on a professional and personal level. This thesis would not have been possible without many valuable collaborations and discussions and of course the feedback from collaborators, colleagues and jury members.

First and foremost, I am extremely grateful to my supervisor dr. Nancy Roosens for her invaluable supervision, support and tutelage the past five years. You were the coordinator of the .Be Ready, DIGICOVID and COVIDDIVER projects on which this thesis was based and I want to thank you for giving me the chance to participate in these projects and allowing me to become a better scientist in your lab. I would also like to thank my academic promotor, prof. dr. Xavier Saelens, for all of his help and advice with my publications and this thesis. Also an immense thanks to the members of my steering committee, dr. Cyril Barbezange, dr. Benedicte Lambrecht and dr. Sigrid De Keersmaecker for their insightful comments and suggestions and my jury members, prof. dr. Savvas Savvides, prof. dr. Marie Joossens, dr. Philippe Lemey, prof. dr. Lieve Naesens and prof. dr. Steven Van Gucht, for the feedback on the thesis and defence which has improved my understanding and context of my work.

I would also like to thank all members of the service TAG past and present who have made the service such a full and friendly environment and to everyone who has helped me with this research. I'm especially grateful to Kevin Vanneste and his bioinformatics team for their invaluable help. I am also grateful for the chance of working together with people from other services including the Viral Diseases service, the Public Health and Genome service, the Avian virology and immunology service, and the Foodborne pathogens service.

Finally, also a big thanks to my family, my boyfriend and friends for always being so supportive and helping me every step of the way.

TABLE OF CONTENTS

Summary i

Samenvatting v

Acknowledgments ix

Table of Contents 1

List of Figures xv

List of Tables xvii

Abbreviations xix

Glossary xxiii

CHAPTER 1: INTRODUCTION 1

1.1. GENERAL INTRODUCTION AND BACKGROUND 1

1.2. INFLUENZA VIRUS 1

1.2.1. Transmission and symptoms 2
1.2.2. Virus classification 3
1.2.3. Viral proteins and their functions 4
1.2.4. Life Cycle 7
1.2.5. Influenza evolution 11
1.2.6. Quasispecies 13
1.2.7. Human Influenza Surveillance 15
1.2.8. Vaccination & Antiviral drugs 16

1.3. SARS-COV-2 21

1.3.1. Transmission and symptoms 22
1.3.2. Virus classification 23
1.3.3. Viral proteins and their functions 25
1.3.4. Life Cycle 28
1.3.5. SARS-CoV-2 evolution 31
1.3.6. Quasispecies 31
1.3.1. SARS-CoV-2 Surveillance 32
1.3.2. Vaccination & Antiviral drugs 34

1.4. DIAGNOSTIC METHODS 37

1.4.1. Clinical diagnostics 39
1.4.2. In vitro diagnostics 39
1.4.3. DNA sequencing 45

1.5. IMPLEMENTATION OF NGS IN PUBLIC HEALTH 60

1.5.1. Potential advantages of the implementation of NGS 60
1.5.2. Bottlenecks for the implementation of NGS 60

CHAPTER 2: AIMS AND OBJECTIVES 63

CHAPTER 3: ADDED VALUE OF USING NGS FOR THE SURVEILLANCE OF ANTIVIRAL DRUG RESISTANCE IN INFLUENZA VIRUS 67

3.1. WHAT IS THE IMPORTANCE OF GENETIC INFLUENZA SURVEILLANCE WITH REGARD TO ANTIVIRAL RESISTANCE? 68

3.2. DISCOVERING MUTATIONS THAT CONFER ANTIVIRAL RESISTANCE 70

3.3. DETECTION OF ANTIVIRAL RESISTANCE MUTATIONS: FROM CLASSICAL SURVEILLANCE TO THE IMPLEMENTATION OF NGS 71

3.4. CASES WHEN NGS COULD HAVE ADDED VALUE FOR DETECTION OF DRUG-RESISTANCE MUTATIONS 74

3.4.1. Monitoring and Surveillance of Resistance to New Antiviral Drugs 74
3.4.2. Surveillance of Emergence of Resistance Mutations in Quasispecies 75

3.5. CONCLUDING REMARKS AND FUTURE PERSPECTIVES 76

ACKNOWLEDGMENTS 77

CHAPTER 4: NEW INFLUENZA CLASSIFICATION BASED ON THE WHOLE VIRAL GENOME 79

4.1. INTRODUCTION 81

4.2. METHODS 83

4.2.1. Sample selection, RNA isolation, PCR amplification and WGS 83
4.2.2. Generation of consensus genome sequences 85
4.2.3. Phylogenomic analysis 85
4.2.4. Reassortment detection 86
4.2.5. Inference of host characteristic associations 86

4.3. RESULTS 87

4.3.1. WGS of clinical influenza A(H3N2) samples 87
4.3.2. Classification using WHO/ECDC guidelines for the HA segment 87
4.3.3. Using whole-genome sequences and beast allows improved phylogenetic classification 90
4.3.4. Increasing the number of genomes considered with a custom-built Nextstrain instance 92
4.3.5. Detection of intra-subtype reassortments 93
4.3.6. Associations between host characteristics and phylogenetic groups 94

4.4. DISCUSSION 96

4.5. ACKNOWLEDGMENTS 99

CHAPTER 5: INTEGRATING PATIENT DATA AND MUTATIONS ACROSS THE WHOLE INFLUENZA GENOME TO IMPROVE THE ASSOCIATION ANALYSIS 101

5.1. INTRODUCTION 103

5.2. MATERIAL AND METHODS 104

5.2.1. Sample Selection, RNA Isolation, PCR Amplification, and WGS 104
5.2.2. Phylogenomic Analysis and Subsampling by Group 106
5.2.3. Inference of Associations with Patient Data 109

5.3. RESULTS 110

5.3.1. Significant Associations Between Viral Mutations and Patient Data 110
5.3.2. Significant Associations Between Mutations and Patient Data When Samples Are Stratified According to the Phylogenetic Clade 115

5.3.3. Evaluation of Significantly Associated Mutations From the Belgian Samples in an International Context 117

5.4. DISCUSSION 117

5.5. ACKNOWLEDGMENTS 121

CHAPTER 6: GENERAL APPROACH TO IDENTIFY LOW-FREQUENCY VARIANTS WITHIN ROUTINE INFLUENZA SURVEILLANCE 123

6.1. INTRODUCTION 125

6.2. METHODS 127

6.2.1. Viruses and cells 127
6.2.2. Patient samples 127
6.2.3. Creation of mixes of wild-type and mutant viruses 129
6.2.4. RNA isolation and RT-qPCR 129
6.2.5. PCR amplification and whole genome sequencing 129
6.2.6. Low-frequency variant identification 131
6.2.7. Statistical analysis 132

6.3. RESULTS 132

6.3.1. Validating an AF threshold for LFV calling using an experimental quasispecies population 132
6.3.2. Selection of patient-derived samples based on their genome copy number 135
6.3.3. Prevalence of LFV in clinical samples 137
6.3.4. Patient data associated with prevalence of LFV 138

6.4. DISCUSSION 138

6.5. ACKNOWLEDGMENTS 142

CHAPTER 7: STRATEGY TO DEVELOP AND EVALUATE A MULTIPLEX RT-DDPCR IN RESPONSE TO SARS-COV-2 GENOMIC EVOLUTION 143

7.1. INTRODUCTION 145

7.2. MATERIALS AND METHODS 147

7.2.1. Selection and evaluation of key target for PCR detection of SARS-CoV-2 using WGS data 147
7.2.2. Development of RT-ddPCR method for the detection of SARS-CoV-2 149
7.2.3. Validation of the specificity of the RT-ddPCR assay for SARS-CoV-2 149
7.2.4. Validation of sensitivity of the RT-ddPCR assay for SARS-CoV-2 150
7.2.5. Applicability assessment 150

7.3. RESULTS 151

7.3.1. In silico inclusivity evaluation for the ORF1a and RdRp_IP4 assays using SCREENED 151
7.3.2. Specificity assessment 152
7.3.3. Sensitivity assessment 153
7.3.4. Applicability assessment 154

7.4. DISCUSSION 154

CHAPTER 8: ESTABLISHING QUALITY CRITERIA TO CHARACTERISE SARS-COV-2 AND ITS VARIANTS IN NGS DATA 159

8.1. INTRODUCTION 161

8.2. METHODS 164

8.2.1. Employed Sequencing Data and Generation of Consensus Genome Sequences 164
8.2.2. Low-Frequency Variants Detection 169

8.2.3. Qualitative Evaluation of Detection of B.1.1.7 at Different Abundances....171
8.2.4. Quantitative Evaluation of Detection of B.1.1.7 at Different Abundances....172

8.3. RESULTS....173

8.3.1. Qualitative Evaluation Demonstrates That B.1.1.7 Clade-Defining Mutations Can Be Reliably Detected at Low Allelic Frequency When Sequencing Coverage Is Adequately High....173
8.3.2. Quantitative Evaluation Demonstrates That the Resulting Allelic Frequencies for B.1.1.7 Clade-Defining Mutations Are Close to Their Target Values....177

8.4. DISCUSSION....179

8.5. CONCLUSION....182

CHAPTER 9: GENERAL DISCUSSION AND FUTURE PERSPECTIVES....183

9.1. WGS: ADDED VALUE AND CHALLENGES OFFERED BY WGS FOR THE INFLUENZA SURVEILLANCE IN RESPIRATORY SAMPLES....185

9.1.1. What this thesis adds....185
9.1.2. Way forward....188

9.2. WGS: ADDED VALUE AND CHALLENGES OFFERED BY WGS FOR THE SARS-COV-2 SURVEILLANCE IN WASTEWATER SAMPLES....190

9.2.1. What this thesis adds....191
9.2.2. Way forward....192

9.3. FUTURE PERSPECTIVES....193

9.3.1. Long-read sequencing as the future of high-throughput sequencing for the genomic characterisation of a virus....193
9.3.2. Quality databases in a One Health context....194
9.3.3. Metagenomics as a surveillance tool....196

ACADEMIC CV....243

BIBLIOGRAPHY....201

LIST OF FIGURES

Figure 1.1: Structure of the influenza virus particle. 7
Figure 1.2: Influenza replication cycle. 10
Figure 1.3: Schematic representation of influenza virus quasispecies. 14
Figure 1.4: Schematic representation of genomic structure of SARS-CoV-2. 28
Figure 1.5: Life cycle of SARS-CoV-2 in human cell. 30
Figure 1.6: Overview of the available clinical, diagnostic and research strategies for the diagnosis of a COVID-19 infection. 38
Figure 1.7: Approximate timeline from SARS-CoV-2 infection. 40
Figure 1.8: Sequencing cost per megabase (August 2020). 49
Figure 1.9: Cumulative number of viral genome submissions to NCBI and GISAID (for influenza and SARS-CoV-2 for major human viral pathogens. 53
Figure 2.1: Schematic outline of the thesis. 65
Figure 3.1: Antiviral drugs against influenza virus and their target sites in the virus cycle. 69
Figure 4.1: Phylogenetic HA gene tree. 89
Figure 4.2: Phylogenetic tree based on the whole genome. 91
Figure 4.3: Time-resolved overview of influenza samples from the Belgian 2016-2017 outbreak season in the context of globally circulating influenza strains based on an in-house Nextstrain instance using only whole-genome sequences. 93
Figure 4.4: Phylogenetic tree based on the whole-genome annotated patient data for which significant associations with phylogenetic groups were detected. 95
Figure 5.1: Phylogenetic tree based on the whole H3N2 genome. 108
Figure 5.2: Comparison of the Belgian influenza samples with samples from the GISAID database for mutations that were considered significantly related to the sampling period. 114
Figure 5.3: Statically significant results using the Fisher’s exact test with FDR correction for the association between renal insufficiency and amino acid mutations in the whole genome from all of the samples and “Phylogenetic Group X” and “Group 3C2a3”. 116
Figure 6.1: ROC curve for validating an AF threshold using a A(H3N2) benchmark dataset. 134
Figure 6.2: ROC curves to validate an AF threshold using a A(H3N2) benchmark dataset at different genome copy numbers. 136
Figure 8.1: Schematic representation of the workflow. 168
Figure 8.2: Qualitative evaluation of Dataset 1 (A) and Dataset 2 (B) based on false negative proportions per condition until a targeted mutant AF of 20%. 176
Figure 8.3: Quantitative evaluation of Dataset 1 (A) and Dataset 2 (B) using the squared SD divided by the maximal squared SD per targeted AF. 178
Figure 9.1: Schematic outline of what this thesis adds and the way forward. The boxes in green represent what this thesis adds and the boxes in yellow represent what the challenges and future perspectives are and what the way forward may be. 184

LIST OF TABLES

Table 3.1: Comparison of different genotypic assays 73
Table 4.1: Samples stratified according to host characteristics. 84
Table 5.1: Sample numbers per patient data. 105
Table 5.2: Statically significant associations found between patient data and amino acid mutations in the whole genome...111
Table 6.1: Samples stratified according to patient data...128
Table 6.2: Number of TP, FP, sensitivity, and specificity at different AF thresholds for the A(H3N2) benchmark dataset. 135
Table 6.3: Number of TP, FP, sensitivity, and specificity at different AF thresholds using a A(H3N2) benchmark dataset at different genome copy numbers. 137
Table 6.4: Statistically significant associations between number of LFV in clinical samples and patient data. 138
Table 7.1. Primer and probe sets included in the multiplex RT-ddPCR assay. 147
Table 7.2. Inclusivity *in silico* evaluation of ORF1a and RdRp_IP4 assays obtained with SCREENED. 151
Table 7.3. Specificity assessment of the developed RT-ddPCR method. 152
Table 7.4. Sensitivity assessments of the developed RT-ddPCR method..........................153
Table 7.5. SARS-CoV-2 investigation in clinical samples and wastewater samples...........154
Table 8.1: Mutations linked to SARS-CoV-2 lineage B.1.1.7 [964]....................................166
Table 8.2: List of SRA accession numbers used for employed wild-type and lineage B.1.1.7 samples in this study. 167
Table 8.3: Qualitative evaluation of Dataset 1 based on false negative proportions per condition until a targeted mutant AF of 20%. 174
Table 8.4: Qualitative evaluation of Dataset 2 based on false negative proportions per condition until a targeted mutant AF of 20%. 175

ABBREVIATIONS

2019-nCoV	2019 novel-coronavirus
2'-O-ribose Mtase	2'-O-ribose methyltransferase
3CLpro	3C-like protease
A	Adenosine
ACE2	Angiotensin-converting enzyme 2
Ad26	Adenovirus serotype 26
ADCC	Antibody-dependent cell-mediated cytotoxicity
AF	Allelic Frequency
AI	Artificial intelligence
AP	Alkaline phosphate
ARDS	Acute Respiratory Distress Syndrome
BSL	Biosafety Level
C	Cytosine
Cas	CRISPR-associated
CDC	Centers for Disease Control and Prevention
CLIA	Chemiluminescent immunoassays
CLR	Continuous Long Read
COG-UK	COVID-19 Genomics UK
COVID-19	Coronavirus disease 2019
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
cRNA	Complementary RNA
CS	Cleavage site
CTL	Cytotoxic T lymphocyte
CXCL-10	C-X-C motif chemokine ligand 10
DAG	Directed Acyclic Graph
ddNTPS	Dideoxynucleotides
ddPCR	Droplet digital PCR
DMV	Double-membrane vesicles
DPP4	Dipeptidyl peptidase-4
dsDNA	Double stranded DNA
dsRNA	Double stranded RNA
E	Envelope
ECDC	European Centre for Disease Prevention and Control
ECMO	Extracorporeal Membrane Oxygenation
ELISA	Enzyme linked immunosorbent assay
endoRNase	Endoribonuclease
ER	Endoplasmic reticulum
ERGIC	Endoplasmic reticulum–Golgi intermediate compartment
ES	Effect Size
ExoN	Exonuclease
FDR	False Discovery Rate
FET	Field-effect transistor
FIA	Fluorescence immunochromatographic assay
FM	Full-text index in Minute space

FN	False Negative
FP	False Positive
G	Guanine
GIP	Global Influenza Program
GiRaF	Graph-incompatibility-based Reassortment Finder
GISAID	Global Initiative on Sharing Avian Influenza Data
GPS	Global Positioning System
gRNA	Genome RNA
HA	Hemagglutinin
HAI	Hemagglutination inhibition assay
hAPN	Human aminopeptidase N
HCoV	Human Coronavirus
HEF	Hemagglutinin-esterase fusion
HIV	Human immunodeficiency virus
HMM	Hidden Markov Model
HPAI	Highly pathogenic avian influenza
HRP	Horseradish peroxidase
IC50	Half maximal inhibitory concentration
ICT	Immunochromatographic
IFN	Interferon
IIV	Inactivated influenza vaccines
IL	Interleukins
ILI	Influenza-like-illness
INC-Seq	Intramolecular-ligated Nanopore Consensus Sequencing
Indels	Insertions and deletions
IQR	Interquartile Range
ISIRV	International Society for Influenza and other Respiratory Virus Diseases
LAIV	Live-attenuated influenza vaccines
LASL	Linker amplification shotgun libraries
LFIA	Lateral flow immunoassays
LFV	Low-Frequency Variant
LNP	Lipid nanoparticle
LOD	Limit of Detection
M	Membrane
M2e	Ectodomain of M2
MCMC	Markov Chain Monte Carlo
MCP-1	Monocyte chemoattractant protein-1
MDA	Multiple displacement amplification
MDCK	Madin Darby Canine Kidney
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
MRCA	Most recent common ancestor
mRNA	Messenger RNA
MSA	Multiple Sequence Alignment
N	Nucleocapsid
N7-MTase	Guanine-N7-methyltransferase
NA	Neuraminidase
NAAT	Nucleic acid amplification test

NAI	Neuraminidase inhibitor
NASBA	Nucleic Acid Sequence-Based Amplification
NCBI	National Center for Biotechnology Information
NEP	Nuclear export protein
NGS	Next-generation sequencing
NI	NA inhibition assays
NIAID	National Institute of Allergy and Infectious Disease
NJ	Neighbour Joining
NLS	Nuclear localisation signal
NP	Nucleoprotein
NPC	Nuclear pore complex
NRC	National Reference Centre
NS	Non-structural
NSP	Non-structural protein
ONT	Oxford Nanopore Technology
ORF	Open reading frame
PA	Polymerase acidic
PacBio	Pacific Biosciences
Pangolin	Phylogenetic Assignment of Named Global Outbreak LINeages
PB1	Polymerase basic 1
PB2	Polymerase basic 2
P-FAB	Fibre-optic absorbance biosensor
PHB	Prohibitin1
PHE	Public Health England
PLpro	Papain-like protease
POC	Point-of-care
POD	Probability of Detection
QIV	Quadrivalent vaccine
R_0	Basic reproduction number
RAxML	Randomized Axelerated Maximum Likelihood
RBD	Receptor-binding domain
RdRp	RNA-dependent RNA polymerase
RIDT	Rapid influenza diagnostic test
RSV	Respiratory Syncytial
RTC	Replication and Transcription Complex
RT-LAMP	Reverse Transcription Loop-Mediated Isothermal Amplification-Based Assay
RT-qPCR	Reverse transcription quantitative polymerase chain reaction
S	Spike
SAM	Sequence Alignment Map
SAMBA	Simple Amplification-Based Assay
SARI	Severe-acute-respiratory-infection
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SELEX	Systematic evolution of ligands by exponential enrichment
sgRNA	Subgenomic RNA
SISPA	Sequence-independent single-primer amplification
SMRT	Single molecule real time sequencing
SNP	Single Nucleotide Polymorphisms

SNV	Single nucleotide variants
SRA	Sequence Read Archive
SRH	Single radial hemolysis
TESSy	The European Surveillance System
TIV	Trivalent vaccine
TM	Transmembrane regions
TMPRSS2	Transmembrane Serine Protease 2
TN	True Negative
TNF-α	Tumour necrosis factor-α
TP	True Positive
U	Uracils
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
UTR	Untranslated region
VIF	Variance Inflation Factor
VLP	Virus-like particle
VN	Virus neutralisation assay
vRNA	Viral RNA
vRNP	Viral ribonucleoprotein
WGS	Whole genome sequencing
WHO	World Health Organization
WWTP	Wastewater Treatment Plant
ZMW	Zero-mode waveguides

GLOSSARY

- Amino acid: An amino acid is a building block of proteins, which is a combination of multiple amino acids. In some cases, a nucleotide mutation can lead to a change in amino acid, which could affect the produced protein.
- Amplicon: An amplicon is a specific DNA product that is generated by PCR.
- Antibody: An antibody is an immunoglobulin protein of the immune system that circulates in the blood. They are produced as a response to the detection of an antigen. Antibodies can be produced artificially to be used in antigen detection tests.
- Antigen: An antigen is a substance recognized by the immune system that stimulates an immune response.
- Assembly: An assembly uses sequencing reads to reconstruct larger genomic fragments
- Base pair (bp): Thymine (T) and Adenosine (A) are complementary nucleotides bound by two hydrogen bonds while Cytosine (C) and Guanine (G) are connected by three hydrogen bonds. Base pairs are used to measure the length of sequencing reads.
- Coverage: The coverage is the number of times a nucleotide at a particular position in the genome has been read. Each position in the genome is overlapped by sequenced fragments.
- De Bruijn graph: The De Bruijn graph is a directed graph that is used to represent overlapping strings in a collection of k-mers and is used for *de novo* genome assembly of short reads. K-mers are used as basic sequence elements and exact overlaps of defined lengths between each k-mer is necessary. The sequence reads are split into all possible k-mers for genome assembly. Subsequently, overlapping k-mers are linked by edges in the graph after which reads are mapped onto the graph. This greatly reduces the computational complexity of genome assembly.
- Deletion: A deletion is a mutation that has been incorporated during DNA replication or data generation. A deletion includes one or multiple nucleotides that have been deleted within the genome.
- *De novo* sequencing: *De novo* sequencing is the sequencing of a new, previously unsequenced organism or DNA segment. Moreover, this term is also used when a genome is assembled without the use of known reference sequences. The genome is then assembled by methods of sequence overlap.

- DNA: The deoxyribonucleic acids are double stranded molecules within cells that carry genetic information. The bases adenine, guanine, cytosine and thymine are the building blocks of DNA and are bonded to a sugar (deoxyribose) and a phosphate to make nucleotides.
- Epidemic: An epidemic is the higher occurrence of cases of disease compared to the normal expectancy in a defined geographic location over a particular defined period of time
- False negative: A false negative is an incorrect test result that incorrectly reads negative when it should be a positive result.
- FASTQ: The FASTQ format is a text-based format that stores biological sequences and their quality scores for NGS reads.
- HA: Hemagglutinin is a surface protein that is found on influenza viruses. This protein plays an important role in infection because it allows influenza viruses to enter a healthy cell.
- HMM: A hidden Markov model is a probabilistic model. HMM is usually used in statistical pattern recognition and classification. HMMs are frequently used in Computational Biology to model biological sequences and find homologous sequences.
- Genomic surveillance: The goal of genomic surveillance is to systematically collect and analyse pathogen genomes to understand the evolution of the pathogen and support genomic epidemiology efforts.
- IC_{50}: The half maximal inhibitory concentration is a measure of the effectiveness of a compound to inhibit the virus by 50% of its maximum value.
- ILI: ILI cases in Belgium are defined by a sudden onset of symptoms, including fever and respiratory and systemic symptoms.
- *In silico*: *In silico* refers to an experiment that was conducted or a dataset that was generated on a computer.
- Insertion: An insertion is a mutation that has been incorporated during DNA replication or data generation. An insertion includes one or multiple nucleotides that have been inserted in the genome.
- *In vitro*: *In vitro* refers to an experiment that is being performed in a controlled experimental environment instead of a natural setting or with living organisms.
- *In vivo*: *In vivo* refers to an experiment that is performed in a natural setting or with a living organism.
- Haplotype: High mutation rates in RNA viruses are caused by the lack of the proofreading ability of DNA polymerases. Consequently, copies of the viral genome of

a RNA virus will often differ from the original genome because of single nucleotide polymorphisms. These are often referred to as haplotypes.

- K-mer: A k-mer is a substring of length k in a genetic sequence. These are split during k-mer frequency-based binning approaches and de Bruijn graph-based assembly.
- MCMC: The Markov Chain Monte Carlo is a combination of a probabilistic analysis method and the Markov model. This stochastic algorithm draws samples from a posterior distribution to get an estimate of the distribution.
- Mutation: A mutation is an alteration to the genome sequence including a change of nucleotides, an insertion or deletion of genetic material. Certain mutations can cause changes in amino acid sequences and potentially their protein function.
- NA: Neuraminidase is a surface protein that is found on influenza viruses. This protein plays an important role in infection because it allows influenza viruses to exit the infected cell in order to spread the infection to other health cells.
- Next-Generation sequencing (NGS): Massively parallel or high-throughput or deep sequencing can be used to determine the nucleotide sequence of a part of a genome or whole genome in a single reaction. It is performed by a non-Sanger-based sequencing technology that can sequence multiple DNA fragments in parallel. Each position is sequenced multiple times, which results in high coverage depth to deliver insight into nucleotide sequence variation and accurate data.
- Nucleotides (dNTPs): Nucleotides are organic molecules, including guanine, adenine, cytosine and thymine, that serve as DNA building blocks. In RNA sequences, uracil will occur instead of thymine.
- Pandemic: A pandemic is a global spread of an infectious disease with a larger reach (multiple countries or continents) than an epidemic.
- Phenotype: A phenotype refers to an observable characteristic which resulted from the interaction of its genotype with the environment.
- Phylogenetic analysis: Phylogenetic analysis will analyse the characteristics and/or the evolutionary development of a pathogen to describe the relationships between different forms of the pathogen.
- Phred: The Phred quality scores assign a quality score to each base, which is equivalent to the base calling probability error. The negative log of the error probability is the phred score and has been adopted as the measure of sequence quality of NGS.
- Primer: A primer is a short string of nucleotides in a single stranded DNA format. They are designed to be the complement of a certain target within the sequence.
- Polymerase: The polymerase is an enzyme that catalyses the synthesis of a new DNA strand alongside a template DNA strand.

- PCR: Polymerase chain reaction is a widely used technique to amplify a single or multiple parts of a genome. It allows exponential amplification by using a polymerase enzyme and primers to synthesise a new complementary strand by adding new nucleotides at the end of the primer.
- Posterior probability: The posterior is an estimate that is being attempted to obtain within a MCMC analysis. It is the probability distribution over the parameter state space, given the data under the chosen evolution model.
- Read: A read is a single, unassembled nucleotide sequence fragment that is produced during a sequencing run.
- Reassortment: A reassortment occurs when a single host is infected with two or more viruses. This could result in the emergence of a novel virus.
- Reverse genetics: Reverse genetics is an experimental procedure that genetically engineers specific nucleic acid sequences that allows to elucidate the gene function by examining changes to phenotypes.
- RNA: The ribonucleic acids are single or double stranded molecules that are present in some viruses. The bases adenine, guanine, cytosine & uracil are the building blocks of RNA.
- ROC curve: A receiver operating characteristic curve is a plot of the true positive rate against the false positive rate for different threshold values.
- Quasispecies: Quasispecies are subpopulations in a biological sample that carry low frequency variants. Due to the high mutation rates in RNA viruses, a heterogeneous population within the same patient is exhibited each with specific evolutionary properties.
- SARI: A SARI case in Belgium is an acute respiratory illness with onset of fever and respiratory symptoms within the past 10 days, and requiring hospitalisation.
- Site-directed mutagenesis: Site-directed mutagenesis is a molecular technique in which a mutation is created at a defined site. This enables the investigation of certain genetic changes with reverse genetics.
- Variant of concern (VOC): The WHO working definition of a VOC is: A VOI is a variant of concern (VOC) if, through a comparative assessment, it has been demonstrated to be associated with:
    - Increase in transmissibility or detrimental change in COVID-19 epidemiology
    - Increase in virulence or change in clinical disease presentation
    - Decrease in effectiveness of public health and social measures or available diagnostics, vaccines, therapeutics

  - Assessed to be a VOC by WHO in consultation with the WHO SARS-CoV-2 Virus Evolution Working Group.
- Variant of interest (VOI): The WHO working definition of a VOI: A SARS-CoV-2 isolate is a variant of interest (VOI) if:
  - It is phenotypically changed compared to a reference isolate or has a genome with mutations that lead to amino acid changes associated with established or suspected phenotypic implications
  - Has been identified to cause community transmission/multiple COVID-19 cases/clusters, or has been detected in multiple countries
  - Is otherwise assessed to be a VOI by WHO in consultation with the WHO SARS-CoV-2 Virus Evolution Working Group

CHAPTER 1: INTRODUCTION

1.1. General introduction and background

Unfortunately, disease is a reality of our biological existence, with a staggering cost of disease as consequence. In 2016, approximately 10 million people died of infectious diseases, accounting for approximately one-fifth of all deaths worldwide [1]. Lower respiratory tract infections, enteric infections causing diarrhoea, tuberculosis, acquired immunodeficiency syndrome and malaria caused the highest mortality among infectious diseases [1]. The scientific and industrial revolutions made many advances towards control and prevention of infectious diseases. However, the crude death rate per 100 000 population, i.e. the number of deaths over a given period divided by the person-years lived by the population over that period, due to respiratory infections in Europe only slightly decreased from 32.8 in 2000 to 28.8 deaths per 100 000 population in 2019 [2]. In the region of the Americas, there is even an increase from 27.4 to 31.5 deaths per 100 000 population [2].

1.2. Influenza virus

Influenza is one of the most ubiquitous infectious diseases in modern society with a very high global disease burden, leading to considerable morbidity and mortality [3, 4]. The U.S. Centers for Disease Control and Prevention (CDC) estimated that 5 to 20% every year of the population gets infected, while 3 to 11% of the population also presents symptomatic flu illness which leads to an annual estimated economic loss of €6 billion to €14 billion in the European Union [5, 6]. Moreover, it is estimated by the World Health Organization (WHO) that 3 to 5 million infections lead to severe illness of which 290 000 to 650 000 lead to respiratory death per year or a case-fatality rate of <0.1% during a typical influenza epidemic [7, 8]. The emergence of influenza in the community can occur as seasonal influenza that causes an annual epidemic and, occasionally, as a pandemic [9]. In the northern and southern hemispheres, seasonal influenza typically occurs in the winter, while in tropical regions influenza occurs throughout the year with irregular outbreaks [8]. The drivers of the wintertime rhythm of seasonal influenza outbreaks in temperate regions are still not fully understood [10], but it has been hypothesised that fluctuations in temperature and absolute humidity may play a role [11]. Additionally, the increase in indoor crowding during the winter months and the decrease in vitamin D due to reduced sunlight exposure resulting in a weaker immune system may also contribute to an increase in circulation [10].

1.2.1. Transmission and symptoms

Influenza is caused by a respiratory virus and can be transmitted in different ways. An infected individual who coughs or sneezes causes virus-laden droplets to disperse into the air. Consequently, nearby people can be infected by inhaling these infectious droplets or even by shaking hands with infected persons [8, 12]. Additionally, influenza viruses can survive on fomites, i.e., contaminated environmental surfaces, and hands for periods consistent with the possibility of onward transmission [13, 14]. However, it has also been shown that aerosols generated by mere breathing can transmit influenza viruses [15]. On an international level, it is likely that air travel contributes to the rapid global spread of influenza viruses [16, 17]. For these reasons, influenza remains a persistent and increasingly serious global threat in both epidemic and pandemic form.

. Influenza symptoms can range from asymptomatic to severe illness leading to complications. On average an influenza infection has a two-day incubation period followed by an infectious period of approximately five days [18]. The symptoms often include a sudden onset of fever, sore throat, cough, runny nose, headache, and muscle and joint pain [8, 19]. Other symptoms can include hot and moist skin, severe malaise, flushed face and infected eyes (conjunctivitis) [20]. Generally, influenza is a self-resolving disease, however influenza can cause serious morbidity and mortality in the very young and old, persons with a compromised immune system and people with underlying chronic diseases. Severe illness can lead to complications such as pneumonia, inflammation of the heart, brain or muscle tissues, and multi-organ failure [21]. The most severe cases are also at risk of putting the body in a life-threatening state, which can lead to sepsis that is mainly caused by a secondary bacterial infection and finally death [22].

1.2.2. Virus classification

Human influenza is caused by an infection with an influenza A or B virus, which are both classified as genera in the *Orthomyxoviridae* family. This family also comprises influenza C and D viruses [23, 24]. Members of the Orthomyxoviridae are enveloped and have a segmented negative-stranded RNA genome: 8 segments for influenza A and B viruses and 7 for influenza C and D viruses. Influenza A, B and C viruses are known to infect humans, while influenza D virus has been discovered in cattle and swine [24, 25]. These 4 types are believed to share a common ancestor, but have evolved to the point that genetic material may be interchanged within a type (intra-subtype), but it is likely not shared between types (inter-subtype) [26].

Higher levels of diversity, pathogenicity and infection are associated with influenza A viruses, therefore, the majority of research tends to focus on this type [27]. Besides humans, influenza A viruses can also infect other species such as birds, pigs, horses, seals, cats, ferrets, dogs and bats [28]. Wild aquatic birds have been established as the natural reservoir for influenza A viruses [29]. The hemagglutinin (HA) and neuraminidase (NA) are encoded by two of the eight segments comprising the Influenza A and B genome. HA is responsible for receptor-binding and membrane fusion whereas NA cleaves off the receptor and is mainly responsible for the release of budding virions from the infected cell. HA and NA are the main targets of the humoral immune response. In total, there are 18 HA and 11 NA subtypes for Influenza A viruses. These subtypes can co-occur in various combinations (e.g. H1N1, H3N2, H5N8, …) and are used to further classify influenza A viruses. Currently, there are two influenza A subtypes circulating in humans, namely an A(H1N1)pdm09 and an H3N2 subtype [8]. Based on the phylogenetic similarity, HA is divided into two groups with group 1 including H1, H2, H5, H6, H8, H9, H11, H12, H13, H16, H17 and H18 and H3, H4, H7, H10, H14 and H15 belonging to group 2 [30–32]. Similarly, NA is divided into group 1 comprising N1, N4, N5 and N8 subtypes and group 2 including N2, N3, N6, N7 and N9 [33]. The current nomenclature system for influenza A viruses includes the host species of origin (except when it concerns a human isolate), geographic location of isolation, strain number and year of isolation, followed by the HA and NA subtype [34].

The host range of influenza B viruses is much narrower than influenza A viruses, infecting only humans and seals. Influenza B viruses are classified into 2 antigenically distinct lineages, namely Victoria and Yamagata, that diverged from each other in the early 1980s.

Influenza A and B viruses cause seasonal epidemics, while influenza C viruses cause mostly mild or asymptomatic infections, with many people acquiring antibodies in childhood [35]. Until recently it was accepted that only one subtype of influenza C viruses was circulating and that there was no animal reservoir. However, a novel influenza C virus has been isolated recently from dogs, pigs and cattle, showing 50% homology with influenza C virus found in humans [24].

The influenza D virus was first identified in 2011 in pigs and has also been observed in cattle. To date no human infections with influenza D virus have been observed. However, it might pose a potential threat to pig and cattle farm workers [36].

1.2.3. Viral proteins and their functions

Influenza virions are rounded with a spherical, oval, or kidney-like shape and a diameter of 80-120 nm. A lipid bilayer envelopes the influenza virus, which has a single-stranded, negative-sense, segmented RNA genome [37]. The structure of influenza A and B viruses are similar and their genomes both include eight genome segments (PB2, PB1, PA, HA, NP, NA, M and NS) coding for at least ten proteins (PB2, PB1, PA, HA, NP, NA, M1, (B)M2, NS1 and NS2) (Figure 1.1). Some influenza A and B viruses can express up to eight additional proteins, such as PB2-S1, PB1-F2, PB1-N40, PA-X, PA-N155, PA-N182, M42 and NS3 [38–40]. Each protein has a unique function in the replication cycle of the influenza virus. Embedded in the envelope are the trimeric HA, tetrameric NA and tetrameric M2, with HA (80%) being the most abundant membrane protein in the virion followed by NA (17%) [41, 42].

The HA gene is responsible for host cell attachment and membrane fusion and is synthesised as an immature, trimeric protein, HA0. Post-translational cleavage converts HA0 into a homotrimer of three protomers that comprise HA1 and HA2 [43]. HA1 forms the membrane distal globular head containing anti-parallel β sheets that contain the receptor binding domain (RBD). This domain allows virus particles to bind to the host cell receptors and fuse with the membrane during viral entry [44]. Most antigenic variation is found in the HA1 region. HA2 forms the membrane-proximal stem region composed of α-helices and contains the cleavage site and fusion domain. HA0 cleavage is essential for the fusion of the viral envelope with the cell endosome and subsequent liberation of the viral genome segments into the cytoplasm of the infected cell [45]. The host range can be limited due to the binding affinity of the RBD of the virus to the type of sialic acid receptors on cells. Avian influenza viruses will primarily recognize α(2,3)-linked sialic acid receptors that are mainly found in the gastrointestinal tract of birds and the lower respiratory tract of mammals. Human influenza viruses will primarily bind to α(2,6)-linked sialic acid receptors that are mainly found in the upper respiratory tract of mammals. Furthermore, the HA protein is associated with high pathogenicity in birds in case of a polybasic cleavage site in HA0. Moreover, the sialic acid receptor specificity and five antigenic sites of the HA protein can help evade the immune system's proteolytic cleavage site and glycosylation sites [46]. These antigenic sites are denoted Ca1, Ca2, Cb, Sa and Sb for H1 [47] and A to E for H3 [48, 49].

The main function of NA is to remove sialic acid residues that are present on newly synthesised virions and on the membrane of infected cells. This activity of NA leads to the separation of the virus particle from the plasma membrane of the host cell and prevents self-aggregation and reattachment of the virus to infected cells [50, 51].

Only 16-20 M2 molecules are present on the surface of a virion [52]. The tetrameric M2 protein acts as a passive proton channel upon acidification of the environment (e.g. in the endosome where the influenza virion resides shortly after receptor-binding) and is essential for virus growth [53]. It has also been reported that a variant of M2, named M42, may compensate for the loss of M2 that has been observed in some influenza A virus strains, thereby restoring viral growth [54]. Below the envelope, the M1 protein forms a layer and plays a role in initiating progeny virus assembly [55]. It covers the viral core that is made up of the viral ribonucleoproteins (vRNPs) that are involved in the transcription and replication of the viral genome. In the influenza B virus, the BM2 protein (BM2) also functions as a proton-selective ion channel. M2 and BM2 share little sequence similarity [56].

The three largest genome segments code for polymerase acidic (PA), polymerase basic 1 (PB1) and polymerase basic 2 (PB2). Combined, PB1, PB2, and PA form the RNA-dependent RNA polymerase (RdRp) that is responsible for complementary RNA (cRNA), messenger RNA (mRNA), and viral RNA (vRNA) synthesis. This polymerase complex is associated with the nucleoprotein (NP) via a panhandle structure formed by the 3' and 5' termini of each genome segment [57, 58] to form the vRNP. The polymerase proteins are important determinants of tissue tropism, host range and pathogenicity. PB2 recognizes and binds to the cap structure of host cell mRNA. PA is responsible for the initiation of mRNA synthesis through its cap snatching activity. PB1 is the actual RNA polymerase. PB2-S1 is translated from spliced mRNA transcribed from the PB2 segment and can inhibit *in vitro* the RIG-I dependent signalling pathway [59]. PB1-F2 is encoded by the +1 reading frame of PB1 [60] and has been described as a virulence factor [61] that affects viral dissemination, pathogenesis and transmission [62]. PB1-N40 has been associated with efficient viral replication [63]. PB1-N40 lacks a transcriptase function, but it interacts with PB2 and the polymerase complex in the cellular environment [63]. PA-X is encoded by the +1 reading frame from the PA segment and has been reported to decrease host protein expression and to contribute to pathogenicity by controlling apoptosis and inflammation [64]. Recently PA-N155 and PA-N182 have been discovered and they are translated from the eleventh and thirteenth in-frame AUG start codons in PA, respectively. These proteins do not have RNase activity, but while investigating viruses lacking the N-truncated PAs in mice, lower pathogenicity was observed compared to wild-type virus [39]. Besides the structural role of NP to bind and encapsidate vRNA, it is also responsible for the switch of the vRNA polymerase activity from mRNA synthesis to cRNA and vRNA production.

Finally, the smallest gene segment of the influenza virus codes for non-structural (NS) proteins NS1, NS2 and NS3. The NS1 protein is primarily a viral interferon (IFN) antagonist [65] that interferes with cellular pathways to counteract the transcription of the type I and III IFN genes. In addition, NS1 is a regulator of mRNA splicing and translation [66]. Furthermore, the PDZ ligand motif at the C-terminus of NS1 is thought to influence pathogenicity in mammals [67, 68]. NS2 or nuclear export protein (NEP) mediates the CRM1-mediated nuclear export pathway to export new vRNPs from the nucleus into the cytoplasm [69]. Lastly, it has been found that NS3 is associated with the adaptation of avian influenza viruses to mammalian hosts [70].

The Influenza C genome only possesses seven RNA segments that encode for nine proteins (Figure 1.1). In contrast to the other influenza viruses, it only has a single surface protein, hemagglutinin-esterase fusion (HEF). Additionally, Influenza C virus binds to 9-O-acetyl-Nacetylneuraminic acid rather than sialic acid, which is important to consider for influenza treatments.

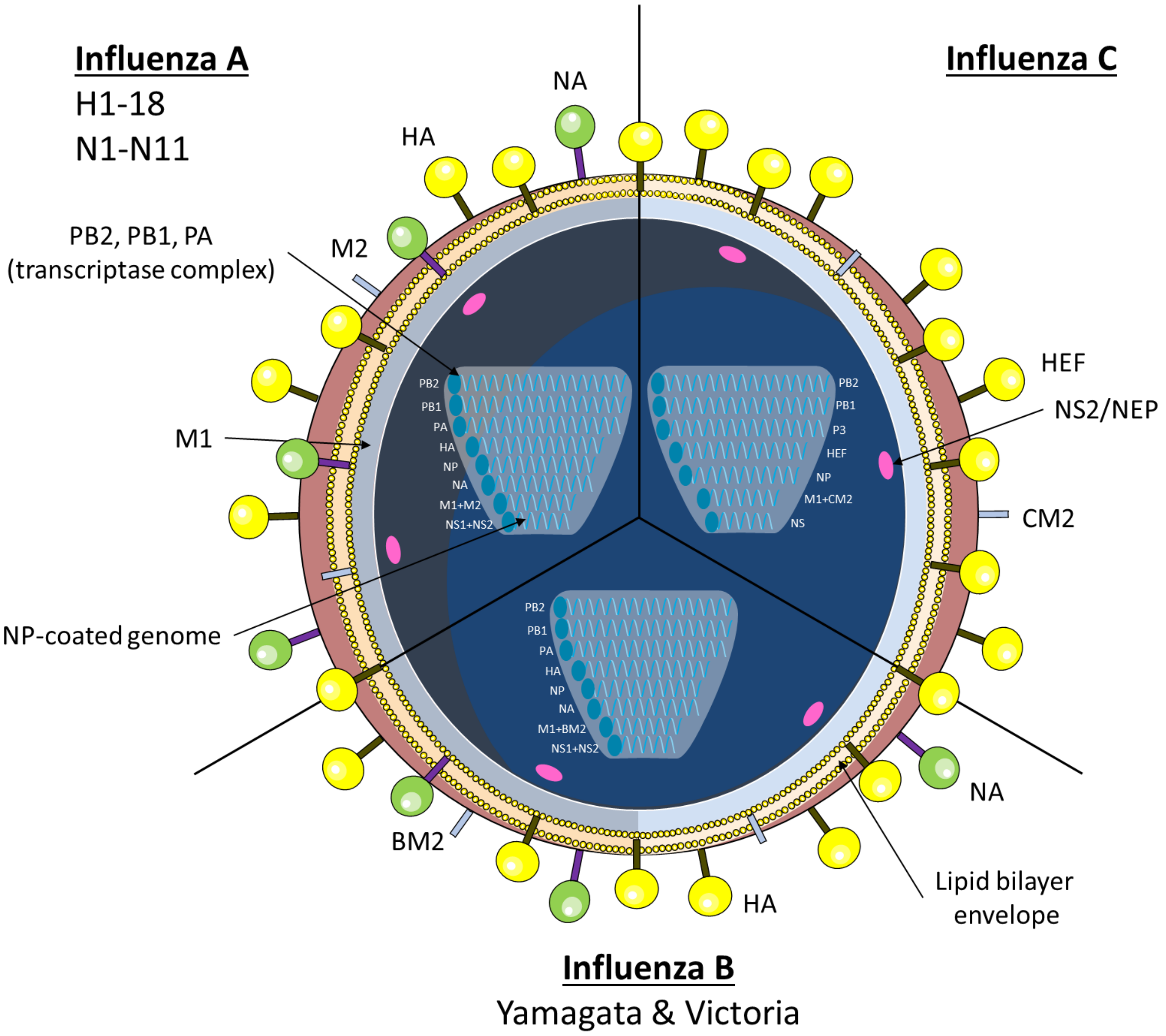


Figure 1.1: Structure of the influenza virus particle.

1.2.4. Life Cycle

Due to the significant impact of the influenza virus on human health, it is not surprising that the life cycle of the virus has been well characterised (Figure 1.2). The influenza viral genome replication is dependent on the nucleus of an active host cell for viral transcription and processing of viral mRNAs [37]. The total replication cycle of an influenza virus *in vitro* is a relatively fast process that only takes eight to ten hours in tissue cultures [29]. The virus infects respiratory epithelial cells by binding with the viral HA protein to sialic acid receptors of the host. Sialic acids are classified according to their linkage to the penultimate galactose in an N-linked glycan chain of the cellular receptor [71, 72]. Influenza A viruses have evolved to use a host specific sialic acid type. Sialic acids linked from C2 to C3 of the galactose, (α(2,3)-linked sialic acid), are primarily recognized by avian influenza A viruses. These α(2,3)-linked sialic acids are primarily found in the gastrointestinal tract of birds and in the lower respiratory tract of mammals. Sialic acids linked from C2 to C6 of galactose, (α(2,6)-linked sialic acid), are primarily bound by mammalian influenza viruses and are predominantly found in the upper respiratory tract of mammals. α(2,6)-linked sialic acids were found to be dominant on epithelial human cells in bronchi, trachea, pharynx, paranasal sinuses and nasal mucosa. However, α(2,3)-linked sialic acids are found on type II cells lining the alveolar wall and on non-ciliated cuboidal bronchiolar cells at the junction between the alveoli and respiratory bronchiole in humans [73, 74]. Consequently, influenza viruses can be present both in the upper and lower respiratory tract. Lower respiratory infections can result in pneumonia with progression to acute respiratory distress syndrome (ARDS), and ultimately death from respiratory failure [75]. Pigs are considered a likely mixing vessel for genetic reassortment as their respiratory epithelia have a mixed population of α(2,3)- and α(2,6)-linked sialic acids and can thus be infected by both avian and mammalian influenza viruses [76].

In humans and other mammals, the virus enters the cell through receptor-mediated endocytosis (Figure 1.2, step 1) [77] and the genome is released into the cytoplasm by dissociating from M1 upon fusion of the endosomal and viral membranes (Figure 1.2, step 2) [78]. For fusion to occur, the viral HA0 precursor protein must have been cleaved into the HA1 and HA2 subunits by host-specific proteases either in the preceding round of viral infection [79, 80] or within the endosome [81]. The pH decrease inside the acidifying endosome triggers a structural rearrangement of the cleaved HA protein resulting in the exposure of a hydrophobic fusion peptide in HA2. The fusion of the endosomal and viral membrane is initiated by the insertion of this fusion peptide into the endosomal membrane [82]. Simultaneously, H^+ ions pass through the viral M2 passive ion channel into the virion, acidifying the virion interior [83] and thus disrupting the weak interactions between M1 and vRNP complexes and releasing the vRNPs from the virion structure [84].

Subsequently, the vRNP complexes are actively transported to the nucleus by the importin-α/β pathway (Figure 1.2, step 3) [85]. Nuclear import of vRNP complexes is mediated by the nuclear localisation signal (NLS) of NP [86]. The NLS binds with importin-α which subsequently binds with importin-β, allowing the protein complex to dock at the nuclear pore complex (NPC) and translocate into the nucleus.

Inside the nucleus of the host cell, the transcription of negative-sense vRNA results in mRNA. Both the 5’ and 3’ ends of each vRNA contain conserved non-coding regions that operate as a promoter for the RdRp to initiate viral transcription [58, 87]. The cap-snatching mechanism is used to synthesise mRNA in which the 5’ methylated cap of a nascent host cell mRNA molecule is first bound by PB2 and subsequently cleaved off by the endonuclease activity of PA [88–91]. mRNA transcription is initiated by the capped RNA products that act as primers. Additionally at the 3’ end, the viral mRNAs are polyadenylated due to a stuttering mechanism caused by steric hindrance of bound RdRp, which leads to the addition of a poly(A) tail at the 3’ end (Figure 1.2, step 4) [58, 87, 92]. This poly(A) tail enables viral mRNA to be exported to the cytoplasm and translated to viral proteins using the host cell machinery (Figure 1.2, step 6) [93]. Viral mRNA that encodes for example NEP and M2 are spliced before the export to the cytoplasm (Figure 1.2, step 5) [94–96].

Ribosomes bound to the endoplasmic reticulum (ER) translate mRNAs for HA, NA and M2 (Figure 1.2, step 7) while the other mRNAs are translated by ribosomes in the cytoplasm (Figure 1.2, step 8) [97–99]. Afterwards M1 and NS1 proteins are transported to the nucleus where the binding of M1 together with NEP proteins induces the export of newly synthesised progeny vRNPs to the cytoplasm (Figure 1.2, step 13) [100–102]. The NP and polymerase proteins are imported to the nucleus and are involved in the synthesis of cRNA, which is primer independent (Figure 1.2, step 9) [103, 104]. This cRNA corresponds to a complete copy of the negative-sense vRNA, which is required to serve as a template for the synthesis of new negative strand vRNAs that can, in turn, serve as template for mRNA synthesis (Figure 1.2, step 10) [105]. The vRNPs are exported from the nucleus via the CRM1 dependent pathway (Figure 1.2, step 11) [106].

The HA, M2 and NA proteins are transported through the Golgi apparatus to the cell surface (Figure 1.2, step 12, 13) [107, 108] and subsequently assembled in association with M1 to become incorporated into the plasma membrane [109, 110]. The vRNPs associated with M1 and NEP proteins are transported to the cell surface and attached to regions of the plasma membrane that contain the HA, NA, M1 and M2 proteins (Figure 1.2, step 13) [111]. The eight gene segments are assembled into a newly formed virion prior to budding out of the cell due to the segmented nature of the influenza genome (Figure 1.2, step 14). Mature virions bud off from the host cell and are released through the enzymatic activity of viral NA (Figure 1.2, step 15) [112, 113]. Afterwards, the virion is liberated from the infected cell and can diffuse to attach to other cells and repeat the replication cycle [114, 115].

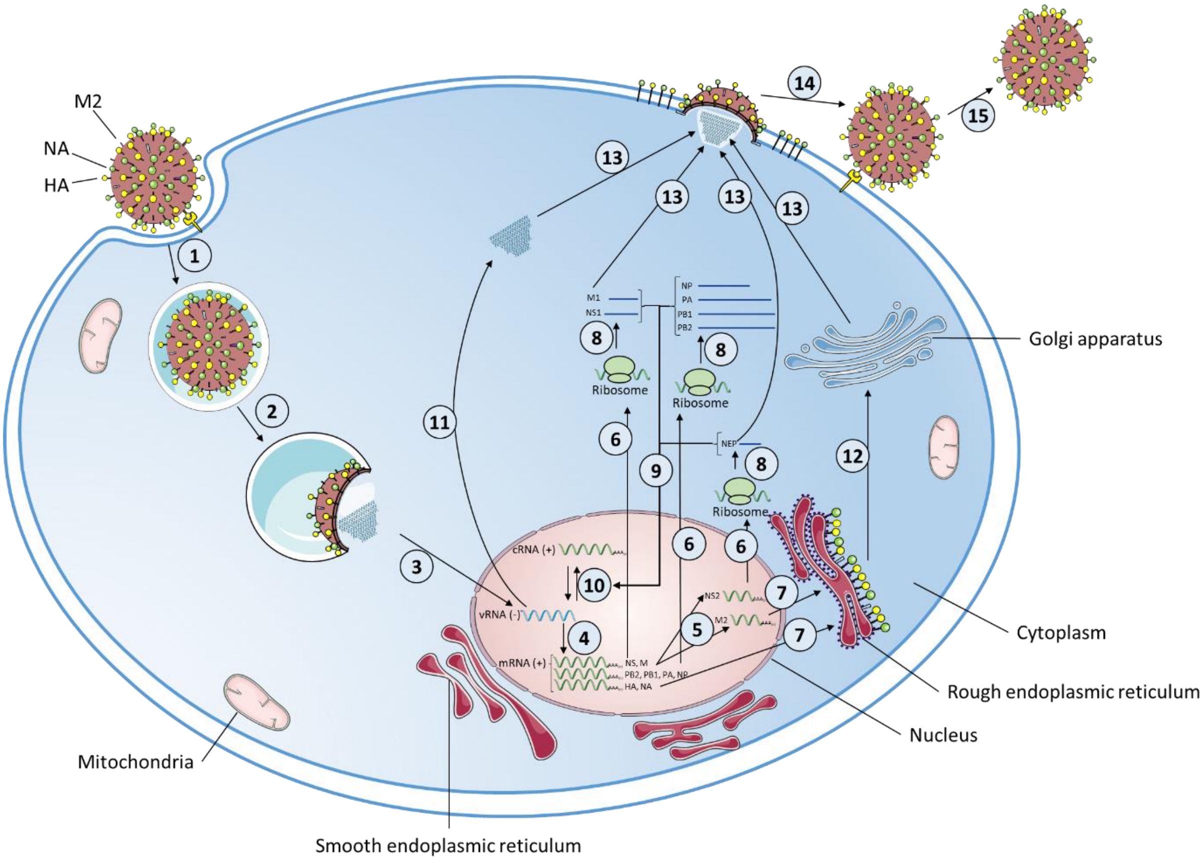


Figure 1.2: Influenza replication cycle. The different stages of the viral cycle are the binding of HA to the sialic-acid containing host receptor (1), followed by the endocytosis and the acid-induced conformational change of the HA protein (2). This acidification in the early endosomes leads to fusion of the viral and endosomal membranes, and triggers the influx of H^+ ions through the M2 channel that results in the dissociation of the vRNPs and uncoating (2). After transport of the vRNPs to the nucleus (3), viral mRNA synthesis is initiated by the viral polymerase (4, 5). The latter is also responsible for the unprimed replication of the vRNA through a cRNA intermediate (10). The viral mRNAs are exported to the cytoplasm and translated into viral proteins (6). In the ER, the membrane proteins HA, M2 and NA are processed, glycosylated and transported to the cell membrane (7, 12,13). The newly synthesised vRNPs are transported to the cytoplasm, mediated by a M1-NS2 complex that is bound to the vRNP (13). At the virion assembly and budding site, the newly produced vRNPs are incorporated into new viruses (14). Finally, the NA cleaves these sialic acid residues and virions are released from the host cell (15).

1.2.5. Influenza evolution

An influenza virus infection will trigger the innate immune response in the host. This induces an inflammatory response, with type I IFN response as a dominant outcome. Chemokines and cytokines are recruited to eliminate the pathogen at the site of infection. Influenza viruses have, in turn, evolved mechanisms to evade the innate immune system.

Antigenic drift and antigenic shift (reassortment) and recombination are the main influenza evolution mechanisms. The lack of a proofreading mechanism of the influenza RdRP leads to the incorporation of incorrect nucleotides during the replication. This results in a relatively high mutation rate of $2.3x10^{-5}$ and $1.7x10^{-6}$ substitutions per nucleotide per replicative cycle for influenza A and B viruses, respectively [116]. This gradual accumulation of point mutations can result in minor variations in the antigenic sites of the HA proteins, known as antigenic drift [117]. More generally, the accumulation of mutations throughout the whole genome is termed genetic drift. Infection with one subtype of influenza virus results in partial cross-protection against other influenza subtypes [118]. The infected person is protected for a prolonged period against reinfection by viruses of the same subtype [119]. However, the recognition by the hosts' immune system is confounded by these structural changes in key regions that are recognized by the adaptive immune system. This leads to the immune escape of the virus which renders the host susceptible to reinfection. The pre-existing neutralising antibodies are impaired and cannot recognize and neutralise the drifted virus. This allows reinfection of the host with the same virus [120]. Consequently, the immune system of the host can often no longer protect the host from this newly formed virus which results in annual epidemics [121].

When two different viral subtypes co-infect a cell, the segmented influenza genome enables the exchange of genetic material forming reassorted influenza strains containing a different HA or NA gene. This is known as antigenic shift or genetic reassortment and can lead to occasional pandemics causing unusually high morbidity and mortality. Heretofore four major pandemics have been recorded, namely in 1918 the H1N1 Spanish flu, in 1957 the H2N2 Asian flu, in 1968 the H3N2 Hong Kong flu and in 2009 the 2009 H1N1 flu pandemic [122–125]. Antigenic shift happens when two or more different subtypes of influenza virus infect the same host and undergo reassortment by swapping gene segments. This results in generating a new strain with potentially novel phenotypic properties such as an altered host range. Generally, humans are only sporadically infected by avian influenza A viruses and vice versa. However, like the H1N1 pandemic virus of 2009, that is now circulating as a seasonal strain, swine can act as a virus mixing vessel because it is both susceptible to human and avian influenza viruses. The H1N1 pandemic virus of 2009 was a reassortment of a Eurasian avian-like swine H1N1 virus and a swine triple reassortant, which, in turn, was a reassortment in the late 1990s among North American avian, classical swine H1N1, and human H3N2 viruses [126]. These alterations in influenza A viruses can be attributed to evolutionary pressure, mechanistic errors during replication of vRNA polymerase, immune pressure, a new host environment or antiviral drug pressure [127]. To date, known human influenza pandemics have been limited to H1, H2 and H3 subtypes, however, avian influenza viruses pose a zoonotic threat to public health and can result in devastating consequences. Influenza viruses of the H2, H5, H6, H7 and H9 subtypes are considered to have pandemic potential. The H2 subtype is viewed as a risk because of the emergence of the H2 subtype in the US swine population and its ability to transmit among ferrets and pigs [128]. The concern for the H6 subtype originates from the fact that they are now endemic in minor poultry in live bird markets in Asia and continue to reassort with H5N1 and H9N2 viruses [128]. The subtypes H5, H7 and H9 are considered a threat because direct transmission to humans has already occurred in the past resulting in mild to severe infections. A large outbreak of highly pathogenic avian influenza (HPAI) H5N1 virus in poultry in 1997 resulted in the first documented cases of direct transmission of HPAI H5N1 virus from poultry to humans. Six out of 18 cases resulted in a fatal outcome [129]. Since 2003, more than 800 cases of human HPAI H5N1 infections were reported of which half had a fatal outcome [130]. Currently, human-to-human transmission of HPAI H5N1 virus has not yet been detected, however it is feared that these viruses might mutate or reassort with contemporary human influenza viruses, possibly resulting in adaptation to humans. Furthermore, coinfections of human and avian influenza viruses in humans or pigs may provide new opportunities for reassortment [131–133]. Due to the large host reservoir [134], the enzootic nature of HPAI H5N1, and accumulation of mammalian adaptation mutations, HPAI H5N1 is currently considered to be the largest pandemic threat to humans. Therefore, avian influenza surveillance and research is important with the aim to predict novel potential pandemic strains and further understand virus spread to allow efficient control of human and avian infections. Furthermore, after the introduction of the influenza A(H1N1)pdm09 virus from swine in 2009, surveillance studies in swine populations were expanded as they are crucial for the preparedness.

Finally, the segmented nature of the influenza genome allows incorporating short stretches of genetic material from a different segment. Usually, recombination does not have such a destructive impact on the host. Events have been recorded, where nucleotides from other viral segments were inserted into the HA cleavage site and increased the virulence. Recombination with the NP and M genes were documented in HPAI A(H7N3) virus in Chile (2002) and Canada (2004), respectively [135, 136]. Furthermore, recombination with non-viral genes such as mitochondrial RNA and host-derived 28s ribosomal RNA also has been observed.

1.2.6. Quasispecies

The survival of a virus is dependent on the host cell machinery, the accuracy of its genome replication, proliferation and adaptation to its changing environment. As a result of the low fidelity of virus-encoded vRNA polymerases, with an estimated error rate of 1 in 10 000 copied nucleotides [137, 138], the virus can also evolve within a single host. In the first few days of an influenza virus infection, viruses are present in very large population sizes with viral loads averaging from 10^3 and 10^8 copies per microliter in collected nasopharyngeal swabs. Consequently, newly generated virus particles are expected to differ after each replication cycle with one or two mutations from their parent virus [139]. Therefore, the genetic makeup of a virus within a host can be more accurately described as a population of closely related viruses or quasispecies [140]. Quasispecies are defined as a set of related genomic variants, produced through replications with errors, which are arranged around a master sequence that usually has the highest fitness. The distribution of genomic variants is usually also called a cloud or swarm of variants (Figure 1.3). For many decades, quasispecies have been a widely accepted concept in virology for many RNA viruses, such as influenza virus, human immunodeficiency virus (HIV), poliovirus and hepatitis C virus [137, 141]. Similar to virus evolution of influenza viruses at the population level, intrahost viral generation is governed by population bottlenecks, selection and reassortment. In contrast to the population level, these processes are far less well-defined at the within-host level [142]. Each mutation could potentially affect the relative viral fitness of the newly generated viral variant, thus each new variant is subjected to a process of positive or negative selection. The frequency of each individual viral variant is dependent on the variants' ability to reproduce and survive combined with the frequency that the same variant is created from a closely related viral variant during the replication process. Additionally, if changes occur in the environment, for instance by moving between compartments within a host (e.g., from the upper to lower respiratory tract) or by infecting a new host, the elements that determine the selection pressures change. This will lead to a different steady-state spectrum of variants with the potential emergence of a new (adapted) variant. Furthermore, the seasonal influenza evolution is mainly guided by immune escape, thus it is likely that the host immune system influences the intrahost evolution as well. Antibodies that target HA and NA can be evaded by accumulating amino acid substitutions on their epitopes [143]. The antibodies, generated by previous exposure or vaccination, are affected by these substitutions resulting in annual epidemics. Therefore, antigenic changes are of great relevance in the choice of strains used for the vaccines. Hence, there is a lot of interest in understanding how such changes occur. Due to the emergence of new sequencing technologies, such as next-generation sequencing technologies (NGS), the detection of quasispecies has been applied more widely in influenza research and diagnostics. Several studies already tried to suggest possible functions of these variants or have functionally characterised particular segments in terms of their escape of the immune system [144], resistance to antivirals [145, 146] and increase in transmission and pathogenicity [147].

Moreover, recent studies indicate that viral reassortment may also be important for infections caused by a single subtype of influenza virus and do not result in antigenic shift. It is difficult to establish the precise virion haplotype due to the segmented nature of the virus. However, rearrangement events between their own segments can occur with high frequency because quasispecies are expected to have genetically similar variants [148]. Therefore, it is an advantage for these viruses to maintain a large genetic diversity of segments. They can recover or eliminate segments as they become available in quasispecies and result in an improvement in fitness [149]. Consequently, segments have different evolutionary patterns which makes analyses of diversity by segments more informative regarding the functional potential that is housed in quasispecies.

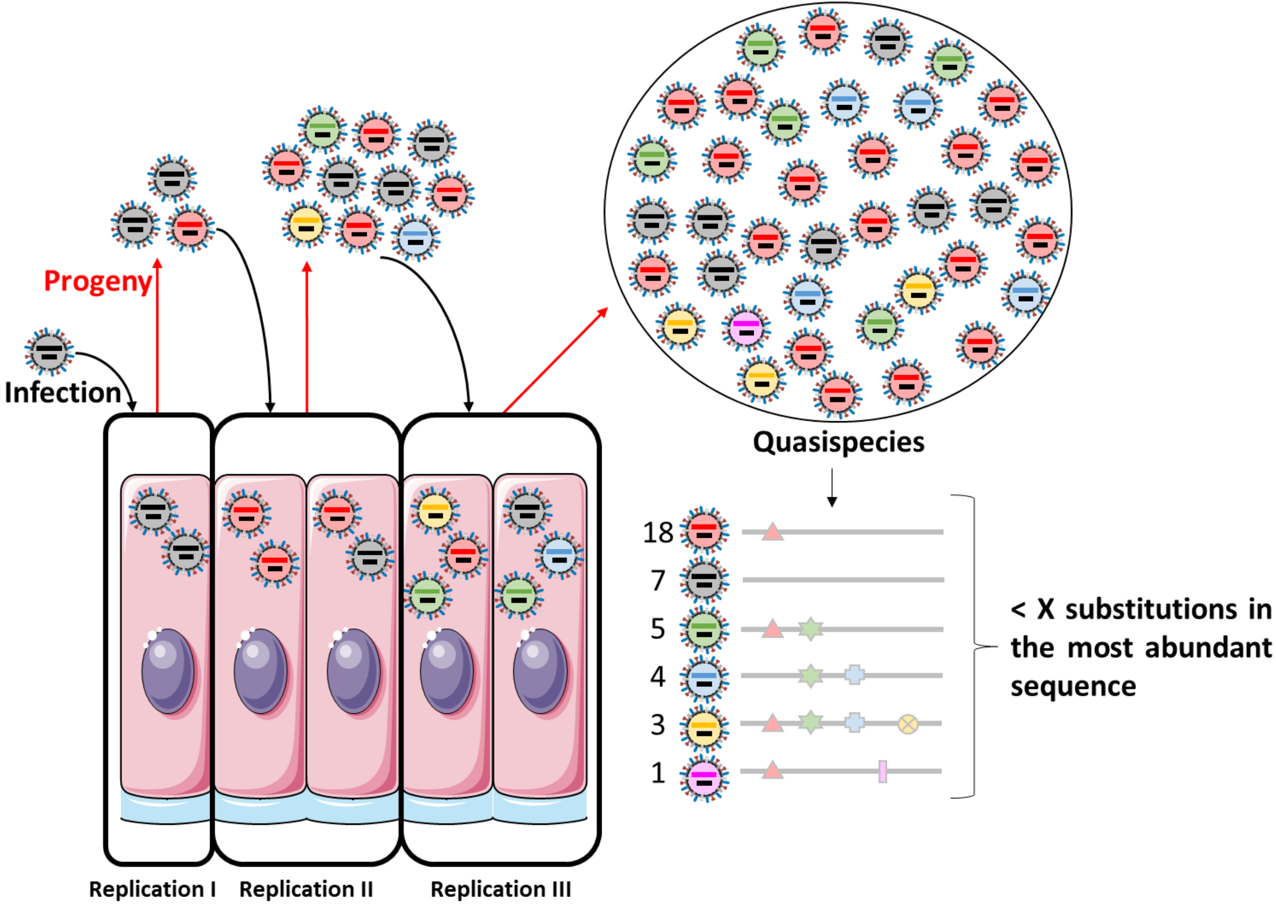


Figure 1.3: Schematic representation of influenza virus quasispecies. Several rounds of cell infections where genetic differences in comparison to the initial infective particle are highlighted by different colours in the virion. The genetic sequences differ a small number of mutations from the most frequent genotype (shown by triangles, stars…). These genetic variations emerge due to errors from the viral polymerase which limits the genetic distance between members of the same quasispecies. Due to the segmented nature of the virus, each segment must be approached individually. For simplicity, the influenza virus genome is represented by 2 segments instead of 8.

1.2.7. Human Influenza Surveillance

Vaccination against human influenza was successfully introduced after World War II. However, a drifted influenza H1N1 variant virus emerged after 2 years of implementation, resulting in a drop in vaccine effectiveness [150]. Consequently, the emerged variant virus was used for subsequent vaccine production [151]. A global institutional network was created to ensure that the vaccine would, as much as possible, always contain the most prevalent influenza variant viruses. This network has evolved into a global institution composed of WHO Global and National Influenza laboratories and several national centres that have working relationships with national and regional licensing agencies and several vaccine manufacturers. Global standards for influenza surveillance are provided by the WHO’s Global Influenza Program (GIP). Additionally, virological and epidemiological influenza surveillance data are collected globally by GIP. By regularly sharing influenza surveillance data from different countries, WHO can provide countries with information about influenza transmission in other parts of the world. This allows national policy makers to better prepare for the upcoming seasons. Additionally, critical features of influenza epidemiology are described including risk groups, impact and transmission characteristics. Also, global trends of influenza transmission are monitored and the data supports the selection of influenza strains for vaccine production.

At European level, the European Centre for Disease Prevention and Control (ECDC) coordinates the European Influenza Surveillance Network (EISN) that combines virological and epidemiological surveillance of influenza through The European Surveillance System (TESSy) database. This will provide public health experts and decision makers within the European Union the required information to better evaluate the influenza activity in Europe and take appropriate action.

In Belgium, there are two main surveillance systems, namely ‘influenza-like-illness’ (ILI) and ‘severe-acute-respiratory-infection’ (SARI). A standard survey accompanies all samples with patient information on sex, birth date, clinical features, vaccination status, administration of antiviral treatment or antibiotics, date of symptom onset and date of sample collection. ILI cases are defined by a sudden onset of symptoms, including fever and respiratory and systemic symptoms. The ILI surveillance uses a network of sentinel general practitioners and results in 500 to 1000 samples per Belgian influenza season. A SARI case is an acute respiratory illness with onset of fever and respiratory symptoms within the past 10 days, and requiring hospitalisation. The SARI surveillance uses a network of six hospitals and results in 500 to 3000 samples per Belgian influenza season. Using a multiplex RT-PCR, all Belgian samples are tested for the presence of influenza A or influenza B viruses. Positive results are further investigated to define the subtype or lineage depending on the influenza strain.

Additionally, all samples are tested for other respiratory viruses, including respiratory syncytial virus A and B, parainfluenza viruses 1, 2, 3 and 4, enterovirus D68, rhinoviruses, human metapneumoviruses, paraechoviruses, bocaviruses, adenovirus, coronaviruses SARS-CoV-2, OC43, NL63, 229E and MERS-CoV.

1.2.8. Vaccination & Antiviral drugs

1.2.8.1. Vaccination

Vaccination is recommended by the CDC to protect people against flu and prevent its spread, especially in high-risk groups [152]. In the 1940s, the first influenza vaccine was produced which was an inactivated virus vaccine containing one influenza A strain, A(H1N1) [153, 154]. When influenza B strains were discovered, a bivalent vaccine was formulated. In 1957, the vaccine was again updated due to the pandemic caused by a new H2N2 virus. Due to the emergence of a new A(H3N2) strain in 1968, a trivalent vaccine (TIV) was further formulated that contained A(H1N1), A(H3N2) and one influenza B strain. In the 1980s it became clear that influenza B viruses were diverging into two different antigenic lineages, called the Victoria and Yamagata lineages. To improve the protective potential against influenza B viruses, the FDA considered adding one more influenza B strain in 2009. Since 2012, this quadrivalent (QIV) vaccine has been recommended by the WHO for seasonal vaccination [155, 156]. The selected strains are updated by the WHO surveillance system by identifying circulating strains by monitoring antigenic drift or shift and comparing them with viruses included in the current influenza vaccine. Based on this information, they try to predict the circulating strains of the next season that should be included in the vaccine [157, 158]. Depending on the antigenic match between vaccine strains and circulating viruses, the protective efficacy of the currently licensed influenza vaccines varies each year. Additionally, the vaccine efficacy can also be affected by the immune status of the host. Unprimed young children have a reduced response to influenza vaccines, while immunocompromised individuals and elderly generally suffer from a declined immune function. In addition, elderly in general will respond poorly to the influenza vaccine. The vaccine efficacy against medically attended ILI can be as low as 10%, and generally does not exceed 60% [159–162].

Currently, there are three types of licensed vaccines available, namely inactivated influenza vaccines (IIV), live-attenuated influenza vaccines (LAIV) and recombinant HA vaccines [163]. The IIV mainly consists of split viruses or subunit influenza antigens. The production of split vaccines occurs by disrupting virus particles using detergents or chemicals. Subunit vaccines are produced by partially purifying viral NA and HA proteins after detergent of chemical splitting [164]. The LAIV are constructed from cold-adapted viruses that are not replicating well at body temperature and are administered intranasally. Local mucosal immunity is induced this way, but the LAIV is only recommended for non-pregnant individuals between 2 and 49 years [164]. While the quantity and quality of NA can vary by manufacturing process and by vaccine, the HA content of licensed vaccines is standardised to 15 micrograms per strain for IIVs. Viral components of the two influenza A strains and one or two influenza B strains are combined to generate the TIV and QIV, respectively. The recombinant HA vaccine contains recombinant HA proteins that are produced in insect cells with baculovirus vectors. Although the current licensed influenza vaccines are effective in healthy young adults, several challenges remain.

First, IIV and LAIV are still mainly produced by using embryonated chicken eggs, which is laborious and time consuming. Consequently, the decision on the vaccine strains should be taken six months before the start of the vaccinations, which allows time for antigenic drift variants to evolve that may be less matched to the chosen vaccine strain, which potentially leads to more severe epidemics [165]. Furthermore, egg-based production methods are heavily dependent on the supply of vaccine-quality eggs and the manufacturers can thus not be flexible in the number of produced doses. Especially in pandemic situations this could lead to vaccine shortages. Besides the long production time in eggs, vaccine production in eggs can also lead to adaptive viral mutations leading to lower vaccine effectiveness [166]. Consequently, several new influenza vaccines have been licensed in recent years that do not rely on the production in eggs, but use mammalian or insect cell lines. Additionally, virus-like particles (VLPs), DNA, and RNA vaccines are in clinical development and are not manufactured in eggs [167]. Furthermore, in preparation for future pandemics, vaccine seed viruses against different subtypes with pandemic potential should be stockpiled. The selection of representative viruses from each subtype should be prioritised based on epidemiological data and testing the candidate vaccines in preclinical studies and clinical trials [168, 169].

Secondly, annual revaccination is required due to the antigenic drift of influenza viruses over time and the decline in vaccine-specific antibodies. To increase the breadth of protection of influenza vaccines, several strategies are being explored. These include induction of more broadly reactive antibodies directed at the conserved HA stem [170, 171], expression of additional influenza antigens using replication-defective viruses as viral vectors that can express high, sustained levels of antigens [172–175], and computationally optimised broadly reactive antigens against the HA protein presented in a VLP vaccine that incorporates the most common amino acid at each position [176–178]. Furthermore, the breadth can also be broadened by immunisation with conserved influenza proteins that target T cell responses by primarily targeting internal highly conserved proteins such as NP and M1 and inducing cytotoxic T-lymphocyte responses that are more cross-reactive than antibody responses directed at the HA [179, 180], incorporation of an adjuvant to boost and broaden the immune response [179, 181–183], and strategies that combine different vaccine platforms in “prime-boost” formats [179, 184–187].

Thirdly, the currently available vaccines are relatively effective against seasonal influenza viruses, however, they fail to protect against antigenically new pandemic viruses. The delay in delivery of the 2009 pandemic vaccine and the inability to predict the subtype that will cause the next influenza pandemic has increased the interest in a universal vaccine. The National Institute of Allergy and Infectious Disease (NIAID) released in 2018 a strategic plan for the development of a universal vaccine. It suggests that the vaccine should (1) be suitable for all age groups, (2) be at least 75% effective against symptomatic influenza infections, (3) have durable protection that lasts at least one year, and (4) protect against group I and II influenza A viruses [188]. There are many vaccine candidates for the generation of broadly protective vaccines, but the two target candidates that have been explored intensively are the highly conserved HA stem which was mentioned earlier [170] and the M2 protein. Antibodies targeting M2 do not neutralise the virus infectivity, but the severity of infection can be reduced by clearing infected cells through antibody-dependent cell-mediated cytotoxicity (ADCC). Typically, the M2 protein is incorporated into a VLP or expressed in a recombinant vaccine by fusing the gene encoding M2 or tandem repeats of the ectodomain of M2 (M2e) to a carrier molecule or protein [179].

Fourthly, there is a need for improved immunogenicity in the elderly as they are the most vulnerable to severe complications from influenza. The effectiveness of most vaccines is poor in elderly due to immunosenescence, i.e., progressive decline in systemic immunity with increasing age [189]. The immunogenicity of IIV in elderly could be enhanced by adding an adjuvant [189] or increasing the antigen dose [163].

Finally, there is a need for an improved correlate of protection, which is currently the HAI antibody titer induced by vaccination. Approximately 50% of individuals are protected from infection at an hemagglutination inhibition (HAI) titer of ≥1:40. However, some studies suggested that T cells may be a better indicator for protection in the elderly and that a higher HAI titer may be required in children [189, 190]. Furthermore, as LAIV has been shown to be effective in the absence of a robust serum antibody response, serum antibody titer is not a reliable correlate of protection for LAIV vaccines [168]. Also, the magnitude and immune response is unknown in pandemic influenza viruses and although the immunogenicity and safety are assessed in clinical trials, the efficacy is established in animal model studies or by extrapolation from experience with human influenza virus vaccines [169]. By using the HAI antibody titer, other aspects of immune memory against the virus, including T cell responses and contribution of non-neutralising antibodies, are not taken into account.

1.2.8.2. Antiviral drugs

Although influenza vaccines are considered the most effective way to prevent seasonal influenza, they often provide suboptimal protection against influenza variants. Antiviral drugs typically remain effective against antigenic drift variants and newly emerged pandemic viruses, because they target conserved or highly stable parts of the virus. They are primarily used to treat severely ill patients and are the most effective when started in the first 48 hours after symptom onset. Currently there are two classes of influenza antivirals licensed in most countries, namely adamantanes and neuraminidase inhibiting drugs (NAIs).

Adamantanes (rimantadine and amantadine) interfere with the M2 matrix protein ion channel [191, 192]. Adamantanes block the influenza M2 ion channel, and consequently inhibit viral uncoating and release of RNA into the cell. However, amantadine resistance is currently present amongst most seasonal influenza strains [193, 194]. Furthermore, the M2 proteins of influenza A and B viruses show very little homology which results in resistance to adamantanes by influenza B. Consequently, these drugs are no longer recommended for influenza treatment or prophylaxis.

NAIs are licensed to treat uncomplicated influenza virus infections, however in most European countries they are mostly used to treat patients at risk of developing more serious complications [195]. They block the enzyme activity of NA by mimicking the binding of sialic acid in the active site of NA. Consequently, the spread of virus within the host is controlled by preventing the release of progeny virions from infected cells. NAIs include oseltamivir (Tamiflu) and zanamivir (Relenza), and peramivir. Although some NAI resistant mutations are known, NAI remains currently the most effective antiviral treatment for influenza infections [196]. Resistance variants are usually less fit, however, they can still propagate until eventually compensatory mutations are acquired that makes the variant fit. For example, mutation H275Y in influenza A(H1N1) and A(H1N1)pdm09 has shown to confer high resistance to NAIs, but it made the virus less fit. However, compensatory mutations such as T289M and N369K restored some of its fitness [197]. Currently, the prevalence of these combinations remains low according to surveillance data [198]. However, this was not always the case. During the 2007-2008 influenza season in Europe, a sporadic emergence of antiviral resistance in seasonal A(H1N1) influenza viruses started spreading rapidly in the population. During that season, 14% of the A(H1N1) virus samples were found to be resistant to oseltamivir [199]. This percentage increased to 98% during the 2008-2009 season [200]. This increase raised concerns until the extinction of this A(H1N1) subtype following the emergence of the NAI-susceptible A(H1N1)pdm09 virus in 2009. A combination of functional phenotypic NA inhibition (NI) assays and genotypic methods, such as conventional reverse transcription quantitative polymerase chain reaction (RT-qPCR) or sequencing methods [201] of the NA gene, is used to assess the influenza antiviral susceptibility to NAIs. Drug-resistant viruses with established or novel changes in NA can be detected using the NI assay and this method is typically the choice for surveillance purposes [202]. The NI assay evaluates the concentration of an NAI required to reduce the enzyme activity by 50% (IC_{50}), which provides valuable information for detecting NAI-resistant viruses. Genotypic methods can only be used to detect molecular markers that are related to resistance in the NA gene [202].

Gradually, antivirals targeting other influenza proteins are being approved and used in the population. Recently, Xofluza (baloxavir marboxil) was approved in the European Union, Japan and the US, which is a cap-dependent endonuclease inhibitor of PA [203, 204]. In China and the Russian Federation, umifenovir (arbidol) is used to treat influenza by preventing viral entry into the host cell by targeting the HA protein [205]. Favipiravir targets the viral RdRP and is available for patients infected by an influenza virus that is resistant to other antivirals in Japan. In Europe and the US, it is currently in the third phase of clinical trials [206, 207]. Other antivirals, such as nitazoxanide [208, 209] (HA maturation inhibitor) is being evaluated in Phase III clinical trials. The antiviral pimodivir [210] (PB2 inhibitor) has been arrested after Phase III clinical trials.

1.3. SARS-CoV-2

Before 2003, only four human coronaviruses, namely Human Coronavirus (HCoV-)229E, HCoV-OC43, HKU1, and (HCoV-)NL63 were known and typically these viruses cause only mild illness in humans [211–213]. However, the emergence of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) proved to the world that coronaviruses can also cause life-threatening infections [214, 215]. These two highly pathogenic coronaviruses with zoonotic origin made emerging coronaviruses a new public health concern [216]. Late in 2019, a new coronavirus, named SARS-CoV-2, emerged in the city of Wuhan, China. Most of the early infected patients were workers at the Huanan Seafood Wholesale Market [217, 218]. Therefore, it has been suggested that the market was at the origin of the virus [219, 220]. However, other studies suggested that the virus was introduced to the market by visitors, which allowed a rapid expansion of infections [221]. After its emergence in the city of Wuhan, the highly transmissible SARS-CoV-2 virus started spreading fast all over the world [222, 223] and overwhelmingly surpassed MERS-CoV and SARS-CoV both in the spatial range of epidemic areas and the number of infected people. The evolutionary history and infection source were established by phylogenetic analysis of SARS-CoV-2 genomes and other coronaviruses. This analysis indicated that the SARS-CoV-2 genomes had 96% nucleotide sequence identity to bat coronavirus RaTG13 (GenBank: MN996532.1) and 79.5% and 55% identity to SARS-CoV BJ01 (GenBank: AY278488.2) and MERS-CoV HCoV-EMC (GenBank: MH454272.1), respectively. Although further analysis is required, it is thought that bats are the hosts of origin and that SARS-CoV-2 might have been transmitted either directly from bats or through an unknown intermediate host to infect humans just like the Himalayan palm civet and Arabian camel for SARS-CoV and MERS-CoV, respectively [224–227].

Initially, this virus was labelled 2019 novel-coronavirus (2019-nCoV), however because of its genetic similarity to SARS-CoV the International Committee on Taxonomy of Viruses officially named it SARS-CoV-2 on February 11, 2020 [226]. On March 11, 2020 the WHO declared COVID-19 as a pandemic [228].

1.3.1. Transmission and symptoms

As of September 28th, 2022, SARS-CoV-2 has infected more than 610 million people, resulting in more than 6.5 million deaths worldwide. In Belgium, there are more than 4.5 million confirmed cases of COVID-19 and more than 32 600 died from the disease [229]. The illness coronavirus disease 2019 (COVID-19) is caused by the SARS-CoV-2 virus, which is highly contagious and transmits between humans through direct contact with an infected individual, touching virus contaminated surfaces or via respiratory droplets and excretions, including droplet inhalation, biological aerosols, sneeze, cough, saliva, mucus, fomites, nasal discharge, ocular fluid, and through breathing and talking [230, 231]. Other possible sources are contact with eye, nasal and oral mucous membranes, urine and faecal contamination. Aerosols, close contact and respiratory droplets exhaled during breathing, talking, coughing or sneezing have been reported as the main transmission route in the spread of COVID-19 [230, 232–237]. The incubation period, which is the time period from exposure to symptom onset, is two days, but can go up to 14 days. As soon as two days prior to the symptom onset, the infected individuals can be already contagious and transmit the virus [238]. The most infectious period is when people show symptoms, because the viral load is the highest at this time. Although asymptomatic cases have the same viral load as (pre)symptomatic cases and are able to transmit the virus [239], a shorter infectious period has been observed [240].

The pathogenesis of a SARS-CoV-2 infection in humans can range from asymptomatic, mild symptoms to severe respiratory failure. At the time of onset and throughout the disease, symptoms of COVID-19 include fever or chills, breathlessness, coughing, headache, muscle ache, fatigue, nausea or vomiting, congestion, sore throat, diarrhoea, running nose, loss of taste (ageusia) and loss of smell (anosmia) [217, 241]. SARS-CoV-2 binds to nasal epithelial cells in the upper respiratory tract and starts local replication and propagation, along with the infection of ciliated cells in the conducting airways [242]. This stage of infection can last a couple of days and the generated immune response is limited. Although infected individuals at this stage have a low viral load, they are already highly infectious [243, 244]. Next, the virus starts migrating from the nasal epithelium to the upper respiratory tract via the conducting airways. During this phase, a greater immune response is manifested with symptoms of fever, dry cough and malaise. Most patients will not progress beyond this phase as the mounted immune response, including the release of C-X-C motif chemokine ligand 10 (CXCL-10) and interferons (IFN-β and IFN-λ), is sufficient to contain the spread of infection. In about one in five of all infected patients, the disease progresses to develop severe symptoms. Type 2 alveolar epithelial cells are invaded by the virus via the host receptor ACE-2 and the virus starts to replicate to produce more nucleocapsids. These infected pneumocytes release many different cytokines, inflammatory markers, including interleukins (IL-1, IL-6, IL-8, IL-120 and IL-12), tumour necrosis factor-α (TNF-α), IFN-β and IFN-λ, CXCL-10, macrophage inflammatory protein-1α and monocyte chemoattractant protein-1 (MCP-1). This cytokine storm will act as a chemoattractant for CD8 cytotoxic T cells, CD4 helper T cells and neutrophils. These cells are responsible for fighting off the virus and consequently are responsible for subsequent inflammation and lung injury. The host cells undergo apoptosis and consequently release new virus particles that infect adjacent type 2 alveolar epithelial cells. The sequestered inflammatory cells cause persistent injury and viral replication leads to loss of type 1 and type 2 pneumocytes. These lead to diffuse alveolar damage, which will culminate in an acute respiratory syndrome [245, 246]. Although all ages of the population seem to be susceptible to SARS-CoV-2, the clinical manifestation differs with age. Most young people only have mild diseases or are asymptomatic [241, 247, 248], while a greater risk of developing ARDS and death is observed in patients with serious pre-existing diseases and patients of an older age (>60 years) [249, 250]. In some COVID-19 cases, multiple organ failure has also been reported [217, 248]. Additionally, although most infected people recover from the acute phase of the disease, some people experience long COVID, which includes a range of symptoms, such as fatigue, loss of appetite, cognitive impairment, chills,… [251].

1.3.2. Virus classification

Coronaviruses can infect a wide variety of mammalian and avian species and have a large, enveloped, positive-sense single-stranded RNA genome ranging between 26 to 32 kb in length [252]. Coronaviruses are classified under the family Coronaviridae, subfamily Coronavirinae. These are subdivided based on their genotypic and serological characteristics into four genera, including *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* and *Deltacoronavirus* [216, 253–255]. HCoV-229E and HCoV-NL63 belong to the *Alphacoronavirus*. The *Betacoronavirus* have been sorted into four lineages A, B, C and D. HCoV-OC43 and HCoV-HKU1 belong to lineage A, while SARS-CoV and SARS-CoV-2 are designated to lineage B and MERS-CoV is attributed to lineage C. Within the B lineage, SARS-CoV-2 is also grouped closely together with novel coronaviruses recently identified in pangolins and four horse-shoe bat coronavirus isolates (RaTG13, ZC45, RmYN02 and ZXC21) [243].

Currently, more than a thousand SARS-CoV-2 lineages were reported that can be grouped into larger clades. The two main SARS-CoV-2 nomenclature systems include Nextstrain, where a new clade is called with a year-letter genetic clade naming when a global frequency of 20% is reached [256], and Global Initiative on Sharing Avian Influenza Data (GISAID), where a new clade is named by the actual letters of the defining marker mutations of each cluster based on statistical distribution of genome distances in phylogenetic clusters [257]. Additionally, GISAID provides more detailed lineages assigned by the Phylogenetic Assignment of Named Global Outbreak LINeages (Pangolin) [258]. Since the emergence of SARS-CoV-2, several notable variants have emerged, that have been declared as variants of concern by the WHO [259]. Lineage B.1.1.7, termed Alpha lineage by WHO [259] and 20I/501Y.V1 by Nextstrain [260], emerged in the United Kingdom in September 2020 [261]. Lineage B.1.1.7 was found to be more transmissible [262] and may cause more severe infections [263, 264]. Lineage B.1.1.7 is defined by multiple spike protein changes, including deletion 69-70, deletion 144 and amino changes N501Y, A570D, P681H, T716I, S982A, D1118H, as well as mutations in other genomic regions [265]. Lineage B.1.351, termed Beta lineage by WHO [259] and 20H/501Y.V2 by Nextstrain [260] was first detected in South Africa at the end of 2020 and has 13 fixed mutations present in all strains, and 17 non-fixed mutations compared to the most recent common ancestor (MRCA) of the SARS-CoV-2 phylogeny [258, 266]. Lineage B.1.351 has been associated with increased transmissibility [267] and reduced effectiveness for some Covid-19 vaccines [268–270]. Lineage P.1 [258], termed Gamma lineage by WHO [259] and 20J/501Y.V3 by Nextstrain [260], has been circulating in Brazil since the end of 2020. This variant has 17 unique mutations, including ten mutations in the spike protein such as N501Y and E484K [271] and is potentially associated with vaccine escape [272]. Lineage B.1.617.2 [258], termed Delta lineage by WHO [259] and 21A/484K.V1 by Nextstrain [260], emerged in India in late 2020 [259]. In the B.1.617.2 genome, 13 non-synonymous mutations were defined. There are four signature mutations in the spike protein of particular concern, including D614G, T478K, L452R and P681R [273]. It has been reported that the B.1.617.2 lineage is 40 to 60% more transmissible than the B.1.1.7 lineage and almost twice as transmissible as the original SARS-CoV-2 Wuhan strain [274]. Some studies indicate that the B.1.617.2 lineage may cause more severe illness compared to earlier strains in unvaccinated persons [275]. Additionally, it has been reported that in case of a breakthrough infection in fully vaccinated people, they produce the same amount of virus as unvaccinated people. However, this amount goes down more quickly in fully vaccinated people, which suggests that although fully vaccinated people are likely infectious, they are likely infectious for less time than unvaccinated people [275]. Finally, lineage B.1.529 [258], termed Omicron lineage by WHO [259] and 21K and 21L by Nextstrain [260], was first reported to the WHO in South Africa at the end of 2021 [276]. Since the emergence of the original BA.1 variant, several subvariants have emerged, namely BA.2, BA.3, BA.4 and BA.5 [277]. Compared to any previous SARS-CoV-2 variant more mutations were reported for the Omicron variant. Most of these mutations are novel and not found in previous variants [278]. For the variants BA.1 and BA.2, it was observed that three doses of a COVID-19 vaccine protected against severe disease and hospitalisation [279–281]. However, it was reported that BA.4 and BA.5 was more infectious even in three-dose vaccinated individuals [277, 282, 283].

1.3.3. Viral proteins and their functions

One of the first available complete genomes of the virus was SARS-CoV-2 Wuhan-Hu-1 isolate (GenBank: MN908947.3) which comprised a 29 903-bp-long RNA genome. Similar to other coronaviruses, the SARS-CoV-2 genome has a bias against cytosine (C) and guanine (G) nucleotides [284]. This nucleotide bias arises from the mutation of guanines and cytosines to adenosines (A) and uracils (U), respectively [285]. It is likely that this bias arises to lower the energy to unbind the genome during replication and translation [285] in addition to avoiding the zinc finger antiviral protein which is produced by human cells to stop the virus spread [286].

The genome is 5'-capped and 3'-polyadenylated and includes two flanking untranslated regions (UTRs) and several open reading frames (ORFs) that encode for several structural and non-structural proteins (NSPs). The structural proteins are responsible among others for the virus assembly [287], membrane fusion [288], host infection [289], morphogenesis and release of virus particles [290], while NSPs facilitate the transcription and viral replication [291]. The genome is organised in the order of non-coding 5'-UTR, replicase genes (ORF1ab), structural proteins and accessory proteins and non-coding 3'UTR [292] (Figure 1.4). The structural proteins are situated at the 3'-terminus of SARS-CoV-2 and include the spike (S), envelope (E), membrane (M), and the nucleocapsid (N) protein [293–295].

The trimeric S protein, which protrudes from the viral envelope, is key for the viral entry into the host cell [296, 297]. Human coronaviruses recognize a variety of host receptors: MERS-CoV binds to dipeptidyl peptidase-4 (DPP4) [298], HCoV-229E recognizes human aminopeptidase N (hAPN) [299], HCoV-HKU1 and HCoV-OC43 bind to certain types of O-acetylated sialic acid [300], and SARS-CoV, SARS-CoV-2 and HCoV-NL63 recognize angiotensin-converting enzyme 2 (ACE2) [301–303]. ACE2 receptors are found in the oral mucosa, nasal epithelium, heart, lungs, intestines, vascular endothelium, testis and kidneys and the virion penetrates into human cells using endocytosis [224, 304, 305]**.** It has also been reported that SARS-CoV-2 also recognizes ACE2 receptors from ferrets, pigs, civets, rhesus monkeys, pangolins, cats, dogs and rabbits [306–309]. This broad receptor usage of SARS-CoV-2 implies a wide host range. The S protein, which is a class 1 fusion protein, has three segments: an intracellular tail, a single-pass transmembrane domain and a large ectodomain. This ectodomain includes the S1 subunit, containing the RBD that engages the host cell receptor and thus determines virus cell tropism and pathogenicity, and the membrane-fusion subunit (S2), which mediates the fusion of viral and cellular membranes upon extensive conformational rearrangements [296, 310, 311].

The E protein is a transmembrane protein that functions as an ion channel and assists in the budding and virulence [312]. Moreover, together with M and N proteins, the E protein is known to play a major role in the virus assembly [287, 290, 313].

The M glycoprotein is the most abundant structural protein in a virion and is a transmembrane protein located in the viral membrane. Its primary function is in the development of a virus-specific humoral response. The M protein stimulates the host humoral response to generate neutralising antibodies [314]. Moreover, the transmembrane domain of the M protein may contain a T cell epitope cluster, which holds dominant cellular immunogenicity [314].

The N protein is responsible for packaging the viral genome RNA (gRNA) into a helical ribonucleocapsid. Furthermore, the N protein contributes to regulating vRNA synthesis during transcription and replication [315–317].

The largest ORFs of the genome are ORF1a and ORF1b, which encode 16 NSPs (NSP1-11 and NSP12-NSP16) [318]. Additionally, because of ribosomal frameshifting ORF1a and ORF1b overlap and two polypeptides, pp1a and pp1ab, are produced. Two cysteine proteins, papain-like protease (PLpro) or NSP3 and 3C-like protease (3CLpro) or NSP5, are encoded by the viral genome. They are responsible for cleaving pp1a and pp1ab into 16 NSPs [319, 320]. These 16 NSPs provide the necessary supporting functions to accommodate the viral replication and transcription complex (RTC), such as host immune evasion, modulating intracellular membranes and providing cofactors (NSP2-11). Additionally, they contain the core enzymatic functions involved in RNA synthesis, RNA proofreading and RNA modification (NSP12-16) [321, 322]. NSP1 plays an important role in suppressing the immune response of the host cell, which allows the virus to freely infect and replicate [323–325]. NSP2 interacts with host proteins prohibitin1 and 2 (PHB1 and PHB2). PHB1 and PHB2 interact with various transcription factors modulating transcriptional activity. Consequently, the host cell environment is altered and the signal that alarms the host cell about an ongoing viral infection is disrupted [326]. NSP3 plays a role in the RTC complex formation and includes two transmembrane regions (TM1 and TM2) and eight domains. During the RTC complex formation, NSP3 interacts with other NSPs and RNA [327]. Additionally, NSP3 modifies host proteins, such as RCHY1, to support viral survival and blocks the innate immune response of the host by de-ubiquitination [328]. NSP4 is a membrane-spanning protein that is important for normal double-membrane vesicles (DMV) formation along with NSP3 and NSP6. It is thought to bind the viral RTC complexes to the modified ER [329–331]. NSP5 is the viral main protease, referred to as 3CLpro and acts as a catalyst in the maturation processing of NSP4-NSP16

[332–334]. NSP6 induces autophagosome formation via an omegasome intermediate. Host proteins involved in inhibiting replication are removed by autophagosome activation, therefore, by limiting the autophagosome expansion the viral infection is favoured because of its reduced ability to deliver viral components to lysosomes for degradation [335–337]. NSP7 and NSP8 compose a hexadecameric complex and increase the RdRp template binding and processivity [338–340]. NSP9 is a ssDNA-RNA-binding protein and engages other proteins in the RTC complex, mediating efficient transcription and replication of the virus [341, 342]. NSP10 plays important roles in the vRNA synthesis and polyprotein processing by interacting with NSP5 protease [343]. Additionally, it stimulates exonuclease (ExoN) and RNA Cap 2'-O-ribose methyltransferase (2'-O-ribose Mtase) activities by recruiting NSP14 and NSP16 for the RTC complex and is thus critical for the cap methylation of viral mRNAs [344–347]. The RdRp activity of NSP12 is the main catalyst of the replication and transcription of the large SARS-CoV-2 RNA genome [348–350]. NSP13 is a RNA helicase that unwinds double-stranded RNA (dsRNA) or DNA (dsDNA) up to several hundred base pairs using the energy of nucleotide hydrolysis. The helicase activity is stimulated by the presence of NSP12. NSP13 is involved in forming a RTC complex, improving viral replication efficiency and mediation of RNA 5'-triphosphatase activity, which suggests its involvement in capping vRNA [351–353]. NSP14 has two functions, namely the guanine-N7-methyltransferase (N7-MTase) activity in the C-terminal domain that adds 5' cap to vRNA [341] and ExoN activity in the N terminal domain that can hydrolyse ssRNA and dsRNA. It has been suggested that NSP14 is involved in proofreading, repair and recombination of the genome [345, 354]. This proofreading function is required to increase the fidelity and processivity during RNA synthesis of the RdRp because of the relatively large genome size compared to other RNA viruses [355]. NSP15 is a Mn^{2+}-dependent viral endoribonuclease (endoRNase) and is cleaved to form 2'-3'-cyclic phosphates [356]. Its endoRNase activity is important to evade activation of host immune responses [357–359]. NSP16 or 2'-O-ribose Mtase forms a complex with NSP10. This complex shields vRNA and ensures formation of a protective cap to prevent recognition by the MDA5 receptor and IFIT proteins. As a consequence of viral infection with reduced host recognition, robust viral replication in the absence of the host cell's innate immune response is possible [360–362]. The RdRp is composed of a catalytic subunit NSP12 and two accessory subunits NSP7 and NSP8 that increase RdRp template binding and processivity. The latter proposed with primase or 3'-terminal adenylyltransferase activity [321, 322, 339, 363, 364]. RdRp together with helicase (NSP13) and ExoN (NSP14) are important enzymes responsible for the transcription and vRNA replication. SARS-CoV-2 also has eight accessory proteins (3a, 3b, 6, 7a, 7b, 8b, 9b and ORF14) derived from subgenomic RNA (sgRNA) that are distributed among the structural genes [292, 318, 365].

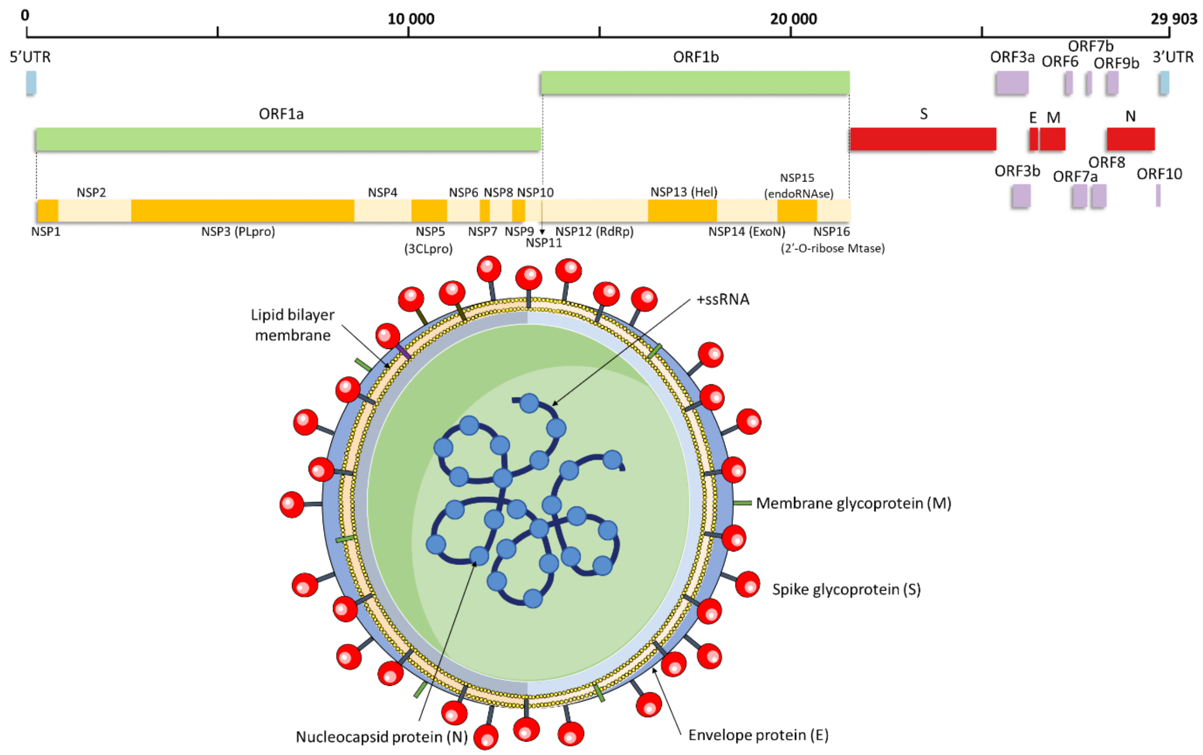


Figure 1.4: Schematic representation of genomic structure of SARS-CoV-2. Genomic distribution of open reading frames across the SARS-CoV-2 genome (29 903 bp). The UTRs (blue), non-structural proteins (green) including the 16 NSPs (yellow) that form the RTC-complex, structural proteins (red) and accessory proteins (purple) are shown. SARS-CoV-2 has spherical structure with an outer lipid envelope and covered with spike glycoproteins. ORF=Open Reading Frame; NSP=Non-structural protein; S=Spike; E=Envelope; M: Matrix; UTR=Untranslated region; N=Nucleocapsid; PLpro=papain-like protease; 3CLpro=3C-like protease; RdRp=RNA-dependent RNA polymerase; Hel=RNA Helicase; ExoN=Exonuclease; EndoRNAse=Endoribonuclease; 2'O-ribose Mtase=2'-O-ribose methyltransferase; +ssRNA=positive-sense single-stranded RNA

1.3.4. Life Cycle

The SARS-CoV-2 cell entry starts with the binding of the S protein to the human ACE2 receptor [297, 301, 366]. The entry of the virus can occur using (1) the endocytic pathway within the endosomal-lysosomal compartment including processing by lysosomal cathepsins or (2) the cell surface pathway following activation by serine proteases such as Transmembrane Serine Protease 2 (TMPRSS2) [366–368]. TMPRSS2 allows early entry into the cell, which is preferred, whereas in the absence of this protease, the virus relies on the endosomal pathway [301, 369] (Figure 1.5, step 1).

Once the RNA is released inside the cytoplasm (Figure 1.5, step 2), the genomic RNA is uncoated [370, 371] (Figure 1.5, step 3). The cell replication machinery of the host is hijacked by the vRNA and the host cell's ribosome translates ORF1a and ORF1b using ribosomal frameshifting into large overlapping polyproteins, pp1a and pp1ab (Figure 1.5, step 4). These polyproteins are further processed by viral proteases, PLpro and 3CLpro, into 16 NSPs (NSP1 to NSP16) [372]. These 16 NSPs drive the viral genome replication and transcription [373] (Figure 1.5, step 6).

The RdRp is directly involved in the replication and transcription, while the other NSPs in the RTC assist during these processes. The replication starts by the synthesis of positive-sense genomic RNA to negative-sense genomic RNA, which is mediated by the RdRp (Figure 1.5, step 7). Next, positive-sense genomic RNA is replicated from the negative-sense genomic RNA [374] (Figure 1.5, step 8). This replicated positive-sense genomic RNA will become the genome of the new viruses or generate more RTCs and NSPs. The RdRp directly mediates the synthesis of negative-sense sgRNA from the positive-sense genomic RNA. Subsequently, these negative-sense sgRNA molecules undergo transcription to the corresponding positive-sense mRNAs [374, 375]. These mRNAs are translated into accessory proteins that assist the virus and the four structural proteins that will become part of the new virus particles (Figure 1.5, step 9).

Following the replication and synthesis of sgRNA, the mRNA of the N protein undergoes translation by cytosolic ribosomes to form the viral nucleocapsid protein. The mRNAs of S, E and M proteins are translated at the cytosolic side of the ER (Figure 1.5, step 10) and undergo translation by ribosomes docked to the cytosolic phase of the ER membrane [374, 376, 377]. The proteins formed during the ER insertion and translation are transported using a secretory path that leads into the ER–Golgi intermediate compartment (ERGIC) [378, 379]. In the ERGIC compartment, viral genomes are encapsidated by nucleocapsid proteins (Figure 1.5, step 11). Subsequently, budding occurs using the ERGIC membranes containing the necessary viral structural proteins to form mature virions [380]. The E and M proteins help the formation of virus-like particles by producing viral envelopes [381], while the N protein enhances their formation in the ER and Golgi apparatus [290]. Next, the virions are released from the infected cell via budding, exocytosis or cell death. Budding will be used by undeveloped virion particles, with the N protein allowing proper orientation of the virion for budding at the plasma, endosomal, nuclear or perinuclear membranes leading to the release of virus particles (Figure 1.5, step 12). Matured virions at the ER or Golgi apparatus are released via exocytosis (Figure 1.5, step 13). Often the virus attack will destroy the host cells' machinery leading to the release of lysosomes. This disruption in cell integrity results in cell death and the consequent release of virus particles [382, 383]. The new virus particles are then ready to invade adjacent cells and provide fresh infective material for community transmission via respiratory droplets [245]. When homeostasis can no longer be maintained by the host cell, cell death or apoptosis will occur [384]. These dead cells will fill the lung airways with debris and fluid, which will cause clogging of the airways and eventually lead to pneumonia. Occasionally the immune system overreacts and healthy lung tissues are damaged, resulting in even more cells to die, further clogging the immune system and making the pneumonia worse. If the lung damage keeps increasing, the patient experiences complete respiratory failure, eventually leading to death [385].

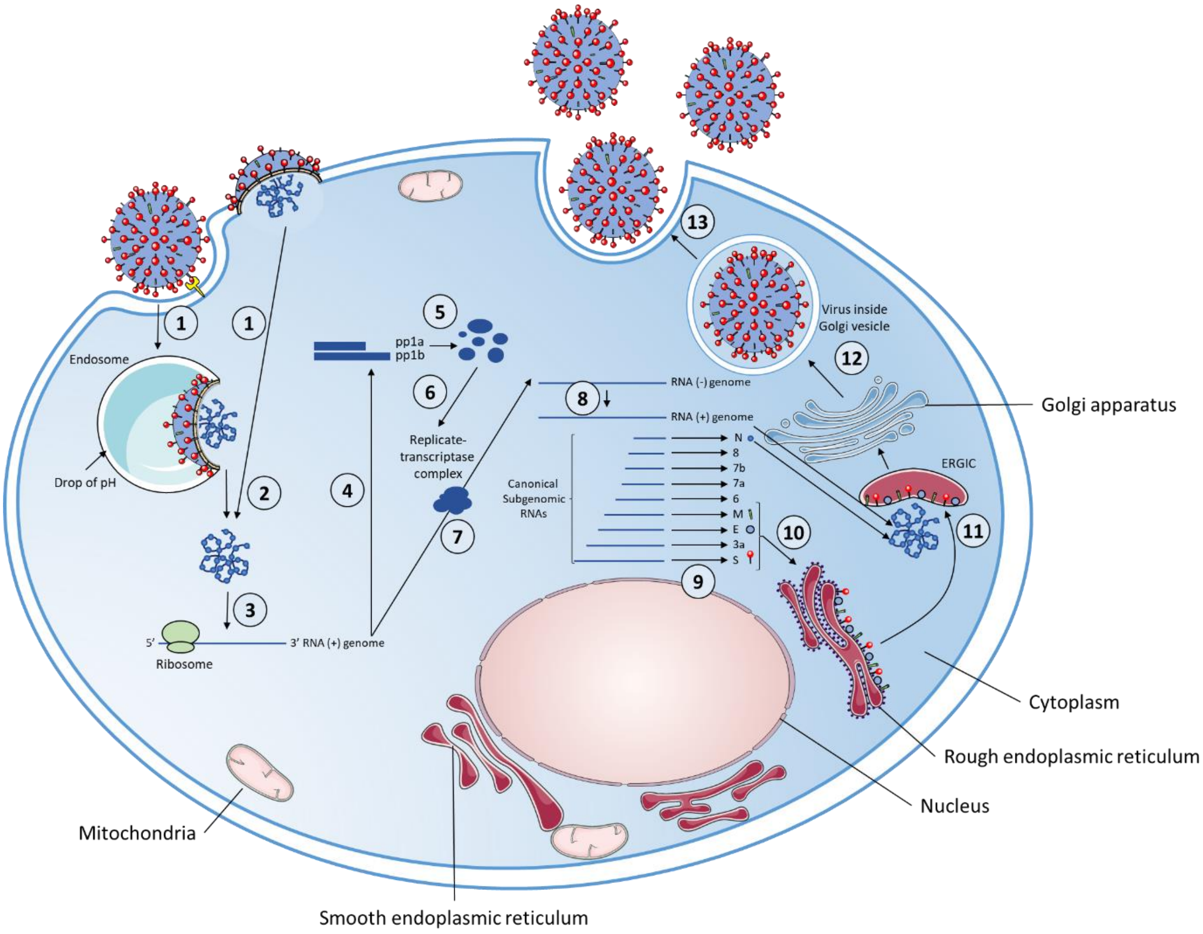


Figure 1.5: Life cycle of SARS-CoV-2 in human cell. (1) SARS-CoV-2 attaches itself to the cellular receptors and enters the cell using either the endosomal pathway or direct fusion. Following, the entry, the (2) release and (3) uncoating of the genomic RNA occur. (4) These are the subject of the translation of ORF1a and ORF1b into polyproteins pp1a and pp1ab. (5) These are processed into individual non-structural proteins (NSP) that (6) form the viral replication and transcription complex (RTC). (7) The RTC drives the production of full-length negative-sense RNA copies of the genome during replication, (8) which are used as templates for full-length positive-sense RNA genomes. (9) The characteristic nested set of subgenomic RNAs are produced during (fragmented) transcription and translated to mRNAs. (10) The translated structural proteins translocate into the endoplasmic reticulum (ER) membranes and (11) interact with N-encapsidated, newly produced genomic RNA in the endoplasmic reticulum–Golgi intermediate compartment (ERGIC). (12) Following budding into the lumen, (13) the virions are released from the infected cell using exocytosis.

1.3.5. SARS-CoV-2 evolution

The accumulation of mutations is an important feature of viral replication. These mutations can serve as viral genetic fingerprints to trace transmission routes across broad geographic regions. This genetic information can be used for contact tracing in case of superspreading events, where an individual infects a group of exposed individuals with a given viral strain [386]. The accumulation of mutations in the viral genome is due to the high error rates of the RdRp, which results in an estimated mutation rate of 6 to $8x10^{-4}$ nucleotides/genomes/year for SARS-CoV-2, with one mutation occurring every two weeks [387]. Although the mutation rate is lower compared to HIV and influenza viruses [388, 389], the frequency of replication-based mutations is sufficiently high for genetic fingerprinting analysis and its relevant applications.

Furthermore, the viral recombination process has the potential to have a severe impact on the virus evolution, host immunity evasion and transmissibility [390]. Recombination happens when host cells are co-infected with different strains of SARS-CoV-2, and the genomes are rearranged and combined during replication. These new virions could potentially possess different pathogenic properties. However, recombination detection is very challenging, because only a portion of the recombination events significantly changes the genealogy, and even then, mutations should happen on the correct branches of the genealogy to create detectable patterns [391]. Additionally, only a relatively short time period has passed since the emergence of SARS-CoV-2. Therefore, SARS-CoV-2 sequences only differ by a small number of mutations, likely resulting in undetectable recombination events. Furthermore, these recombination events can be indistinguishable from recurrent mutations [392]. However, as genetic diversity will probably increase over time, the recombination events may become more distinct.

1.3.6. Quasispecies

While on a strain-level, mutations are found at a relatively high frequency. Individual SARS-CoV-2 patients also show evidence of mutations that occur during the viral replication within the individual [393, 394]. It is believed that quasispecies are a strategy of virus evolution that allows a greater probability of changing the host range, cell tropism or overcoming internal or external selective constraints [141, 395, 396]. Deep sequencing has revealed quasispecies not only in influenza viruses, but also in SARS-CoV [397] and MERS-CoV [398] viruses. Recently, quasispecies were also observed in the case of SARS-CoV-2 [399, 400]. Based on observations in viruses and bacteria, it is well known that the genetic diversity in the quasispecies is influenced by pathogen-host interaction to adapt to different tissues and hosts. Any of these viral quasispecies or subpopulations can infect another individual. The quasispecies may be specific to an individual patient and if one of these quasispecies is transmitted to another person, this information could be used to characterise transmission patterns.

1.3.1. SARS-CoV-2 Surveillance

The initial SARS-CoV-2 outbreak in Wuhan spread rapidly and soon other parts of China and an increasing number of countries were affected by the virus, resulting in a pandemic [228]. On 31 December 2019, a media statement about a cluster of viral pneumonia cases was picked up by the WHO's Country Office in China, that subsequently notified the WHO Western Pacific Regional Office. All Member States were alerted by WHO on 5 January 2020 about the cluster that was caused by a novel coronavirus. The primary focus at global level was the rapid scaling to support the countries. WHO country offices started the support for the development of a national COVID-19 response strategy and they provided expert advice to ministries of health and its partners on priority inventions. Moreover, support for logistics, supply chain and procurement for the COVID-19 response was initiated, as well as, training relevant national and partner staff in technical areas or capacity-building [401]. Throughout 2020, WHO has taken up the considerable task of collating, validating, analysing and disseminating official daily case and death counts reported by 212 areas, territories and countries. This information was routinely published through several country- and region specific situation reports and dashboards. Besides active COVID-19 surveillance, WHO also recommends that countries use already existing syndromic respiratory disease surveillance systems such as those for the ILI and SARI surveillance. As of March 2020, WHO recommends four case definitions [402], namely a suspect case, probable case, confirmed case or contact.

Italy was the first European country that was severely hit by the epidemic at the end of February 2020, followed by the other European countries a week later [403]. For the EU level surveillance, the Early Warning and Response System (EWRS) is used by the Member States to report the number of laboratory-confirmed cases of COVID-19. Additionally, the TESSy surveillance network is also used to collect COVID-19 data. WHO Regional Office for Europe and ECDC in collaboration with their surveillance networks in the Member states coordinate the reporting of the number of performed tests, the number of cases, hospital and ICU admissions, deaths, and the number of variants [404].

The first Belgian infection was reported by the health authorities on the 4th of February 2020 in an asymptomatic Belgian citizen repatriated from Wuhan [405]. In March, the number of symptomatic cases quickly increased. Initially the outbreak was related to the return of travellers, principally from Italy, but the number of infections quickly escalated due to local-

and community-based transmission throughout the country [406]. Due to the spread of SARS-Cov-2, it became necessary to closely monitor this disease in order to inform public health decisions and actions. The Belgian institute for public health, Sciensano, collects data from different sources, including lab-confirmed COVID-19 cases, testing, hospitalised COVID-19 patients and COVID-19 deaths. Persons are diagnosed with COVID-19 using a laboratory test that was carried out by the laboratory of the National Reference Centre (NRC) or by peripheral clinical laboratories, the network of university laboratories, or by the national testing platform. These tests include antigen tests, rapid antigen tests and PCR tests. Basic demographic data, the result of these tests and the number of total tests that were performed is collected by Sciensano which is summarised in daily reports. Data about hospitalisation is collected through a survey where aggregated data on the number of hospitalised and deceased COVID-19 patients is provided by all Belgian general hospitals on a daily basis. For the COVID-19 deaths, there are several data sources, namely the daily reporting from the hospitals to Sciensano and nursing homes to the regional authorities and the mandatory declaration for general practitioners to the regional authorities [407].

Besides surveillance systems that collect data on lab-confirmed cases, governments around the world also introduced digital surveillance to help contain the spread of virus. They are mostly available in the form of smartphone apps that use Bluetooth data exchange or global positioning system (GPS) to keep track of the proximity between devices that have installed the app. If a user tests positive for the virus, other users who have been in close contact according to the proximity data are sent a message. These alerted users can then test and isolate themselves in order to reduce virus circulation in a given population [408].

In parallel to clinical surveillance, most countries, including Belgium, perform wastewater surveillance. Wastewater can also be tested for the presence or prevalence of SARS-CoV-2 in the population, because SARS-CoV-2 RNA is shed in the faeces of approximately 40% of the infected persons [409]. SARS-CoV-2 RNA loads have been reported from 550 to 1.21×10^5 copies/mL in faeces, while the viral load obtained in respiratory specimens was higher at 641 to 1.34×10^{11} copies per mL [410]. However, it was still possible to detect SARS-CoV-2 genomes in faeces several weeks after the oral swabs no longer tested positive. This would suggest that the viral excretion may last longer in faeces and it is hypothesised that the virus could be transmitted by a faecal-oral route [411, 412]. The presence of SARS-CoV-2 has been reported in raw wastewater and an association was observed between an increase of the RNA concentration in raw wastewater [413–415] and an increase in reported COVID-19 cases [414]. This renders wastewater-based epidemiology as an important tool to trace the circulating viruses in a community. Furthermore, it provides opportunities to estimate their genetic diversity, geographical distribution and prevalence [416, 417]. In addition, wastewater surveillance could provide an unbiased method not limited by the asymptomatic nature of the viral infections leading to the underdiagnosis of positive cases compared to clinical surveillance [418]. Moreover, wastewater surveillance makes it possible to evaluate the spread of infection in different areas, even where there are limited resources for clinical diagnosis or delays in test reporting [419]. However, there are several limitations to wastewater surveillance. The correlation between the level of viruses and the specific number of cases may be challenging, because the viral load in the sample is subjected to the excretion rate during the course of the infection, inconsistent capture of spatial variability due to travel and use of multiple wastewater systems in time, temporal delays, inactivation during the wastewater transport process, dilution due to precipitation and/or infrequent or absent clinical testing [420]. In addition, the virus detection and quantification can be limited due to the stability of the genome in wastewater, sampling variability, low efficiency of virus concentration methods and the lack of sensitive detection assays [413]. In Belgium, water samples are taken from the influent of 42 wastewater treatment plants (WWTP) twice per week. Three laboratories, namely Sciensano, e-Biom, and University of Antwerp, detect SARS-CoV-2 in these wastewater samples and report the results.

1.3.2. Vaccination & Antiviral drugs

1.3.2.1. Vaccination

Before the COVID-19 pandemic, a vaccine targeting an infectious disease had never been produced in less than several years. Additionally, no vaccine had been developed beyond the preclinical stage for the prevention or mitigation of disease caused by coronavirus infections in humans [421]. At least nine different technologies have been explored in an attempt to create an effective vaccine against SARS-CoV-2 [422, 423]. Most of these vaccine candidates focus on the spike protein [424]. These technologies include next-generation strategies for precise targeting of COVID-19 infection mechanisms [423–425]. Technologies that were used include RNA vaccines, adenovirus vector vaccines, inactivated vaccines, subunit vaccines, virus-like particle vaccines, lentivirus vector vaccines [426, 427], DNA vaccines [428–430, 430, 431], conjugate vaccines, and a vesicular stomatitis virus exposing the SARS-CoV-2 spike protein [432].

RNA vaccines use RNA to generate an immune response. The spike-encoded mRNA in the vaccine is translated into viral proteins and causes human cells to build the spike protein of SARS-CoV-2. This elicits an immune response and the translated viral proteins are recognized as antigens. Most of these RNA vaccines use nucleoside-modified mRNA. Furthermore, the injected mRNA is formulated in lipid nanoparticles that protect the RNA and help their absorption and delivery into the cells [433–436]. The advantages of nucleic acid vaccines are the fast design and development, the scalability, the safety as no infectious agent handling is required and the induction of humoral and cellular responses. A main disadvantage is that mRNA vaccines exhibit instability and require storage at maximum -20°C [437]. The Pfizer-BioNTech COVID-19 vaccine (Comirnaty) and Moderna COVID-19 vaccine (Spikevax) both RNA vaccines, were the first to be authorised in the United States, the United Kingdom, and the European Union [438, 439]. Pfizer's BNT162b2 vaccine encodes full-length, membrane-anchored spike proteins that is prefusion stabilised and is known as a nucleoside-modified full-length prefusion-stabilised S-2P construct. Besides the nucleoside-modified mRNA-LNP (lipid nanoparticles) vaccine that was eventually chosen, Pfizer also developed versions using self-amplifying mRNA-LNP and unmodified non-replicating mRNA LNP [440]. Moderna's mRNA-1273 also encodes the full-length prefusion stabilised spike protein and solely focused on the nucleoside-modified mRNA LNP platform [441]. Both Moderna and Pfizer/BioNTech used nucleoside-modified mRNA vaccines which had a considerably higher efficacy compared to Curevac's unmodified mRNA-LNP vaccine [440]. In June 2021, CureVac failed to prove that their COVID-19 vaccine is 50% effective [442].

Adenovirus vector vaccines are non-replicating viral vector vaccines that use an adenovirus shell containing DNA that encodes for a SARS-CoV-2 protein [443]. This vaccine does not lead to the new production of adenovirus particles, but rather evoke a systemic immune response after immunisation by letting the host cells produce antigens [441]. The advantages of these vaccines are the robust humoral and cellular responses after already one dose in addition to a good safety profile. Disadvantages are that pre-existing immunity against the viral vector can weaken the immune responses and some candidates require storage at maximum -20°C [437]. Authorised adenovirus vector vaccines in the EU are the Oxford-AstraZeneca COVID-19 vaccine, Vaxzervria, [444] and Janssen COVID-19 vaccine, Jcovden, [445, 446]. Under review at the EMA is the Sputnik V COVID-19 vaccine [447]. Vaxzevria developed by Oxford University and AstraZeneca use the modified chimpanzee adenovirus ChAdOx1 [444, 448, 449] containing the full-length codon-optimised coding sequence of the spike protein along with a tissue plasminogen activator leader sequence [449, 450]. The Janssen COVID-19 vaccine or AD26.COV2.S is a recombinant, non-replicating adenovirus serotype 26 (Ad26) vector that encodes a prefusion stabilised full-length spike protein [441]. Sputnik V employs a heterologous prime-boost vaccination strategy with recombinant adenovirus Ad26 followed by Ad5 as vectors for the expression of the spike protein [451].

Inactivated vaccines include virus particles grown in culture and then killed using methods such as formaldehyde or heat to inactivate the SARS-CoV-2 virus, while still capable of eliciting an immune response against the structural proteins of the virus [452]. Advantages of this vaccine are the safety, because the pathogen is inactivated, the transport and storage.

Disadvantages are the processing of large quantities of pathogen, the antigen immunogenicity that can be affected by the inactivation process, low immunogenicity so multiple booster doses are required and poor induction of cellular responses [437]. In the EU, the COVID-19 vaccine from Valneva is authorised for use. Other vaccines that are authorised in other parts of the world include the Indian Covaxin [453], Russian CoviVac [454], Chinese CoronaVac [455, 456], WIBP-CorV [457], BBIBP-CorV [458], Iranian COVIran Barekat [459] and Kazakhstani QazVac [460].

Recombinant subunit vaccines display one or more antigens while not introducing whole virus particles. They are often protein subunits, however they can be any molecule that is a fragment of the pathogen [461]. The advantages of recombinant subunit vaccines are the safety during production, the safe administration to immunosuppressed people and no infectious agent handling is required. Disadvantages are the small size of the antigens that diminish their uptake by antigen-presenting cells, low immunogenicity so several booster doses and adjuvants are needed, poor induction of cellular responses, a need for confirmation about the antigen integrity and finally the production is limited by antigen production scalability [437]. The Novavax COVID-19-vaccine, Nuvaxovid [462], is authorised by the EU. The peptide vaccine EpiVacCorona [463] and ZF2001 [464] are two authorised vaccines in other parts of the world, while the Sanofi GSK vaccine [465], and COVID-19 vaccine from HIPRA Human Health [466] are pending authorization by the EMA.

1.3.2.2. Antiviral drugs

Although clinical data and experimental studies suggest that the licensed vaccines are helping to prevent COVID-19, these vaccines may not be as effective against some emerging SARS-CoV-2 variants [467]. Moreover, there was an acceleration in advancing vaccine development due to the urgent need for a vaccine, resulting in licensing new vaccine technologies that lead to uncertainties regarding long-term safety issues, durability and effectiveness [467, 468]. In addition, COVID-19 vaccines are intended for prophylactic use only. Therefore, it is essential to develop antivirals for those who do contract the disease or for immune-compromised patients who respond very poorly to COVID-19 vaccines. Potential targets of antiviral therapy are inhibiting the TMPRSS2 human protease [469], preventing virus entry, targeting the RdRp that replicates SARS-CoV-2 genomes [470], preventing virus assembly by using protease inhibitors to inhibit 3CL protease that is responsible for building virus particles from polypeptides [469], and many other targets of the life cycle of SARS-CoV-2. Since August 2022, there are eight antiviral drugs or therapies authorised for use in the European Union, including Regkirona (regdanvimab) [471], Ronapreve (casirivimab/imdevimab) [472] and Veklury (remdesivir) [473], Evusheld (tixagevimab/cilgavimab) [474], Kineret (anakinra) [475], Paxlovid (PF-07321332 / ritonavir) [476], RoActemra (tocilizumab) [477], and Xevudy (sotrovimab) [478]. Regdanvimab is a monoclonal antibody and the active substance in Regkirona. It has been designed to attach to the spike protein of SARS-CoV-2 to prevent the virus from entering the cells of the host. Ronapreve, Evusheld and Xevudy include monoclonal antibodies that will attach to the spike protein of SARS-CoV-2 and prevent the virus from entering the cells. Consequently, the virus cannot multiply and is unable to cause an infection. However, although these anti-RBD mAbs can neutralise the original SARS-CoV-2 strain, many showed a reduced neutralising activity for the Omicron variant. Particularly, the mAbs casirivimab, imdevimab and regdanvimab showed a reduced or complete lack of neutralising activity in cell culture [479]. Remdesivir is the active substance in Veklury and is a viral RNA polymerase inhibitor that interferes with the production of viral RNA. This interference hampers the virus from multiplying inside cells, which helps the body to overcome the viral infection and helps patients get better more quickly. Kineret is an immunosuppressive medicine that attaches to the receptors where interleukin 1 normally attaches itself to. By blocking the activity of interleukin1, it helps to relieve the COVID-19 symptoms. Paxlovid helps to reduce the ability of SARS-CoV-2 to multiply in the body. RoActemra is a monoclonal antibody that attaches to the interleukin-6 receptor which reduces the inflammation and other symptoms. Furthermore, Olumiant (baricitinib) and Lagevrio (molnupiravir) are waiting for marketing authorization in the European Union [480].

1.4. Diagnostic methods

For both influenza and SARS-CoV-2 surveillance, mostly respiratory samples are collected. These samples should be collected as soon as possible after illness onset with a nasopharyngeal swab, nasal aspirate or wash or a combined oropharyngeal and nasopharyngeal swab. In case of intubated patients, an endotracheal, bronchoalveolar lavage or sputum specimens should be collected. The specimens should be placed into sterile viral transport media and subsequently stored at 4°C for transport to the laboratory. To prevent degradation of the RNA and viability of the pathogen, the sample should be kept at -70°C and regular freezing and thawing should be avoided. These samples are then used to perform diagnostic tests to identify the viral pathogen. These diagnostic tests are mainly divided into two broad categories, namely clinical diagnostics and *in vitro* diagnostics [481–483] (Figure 1.6).

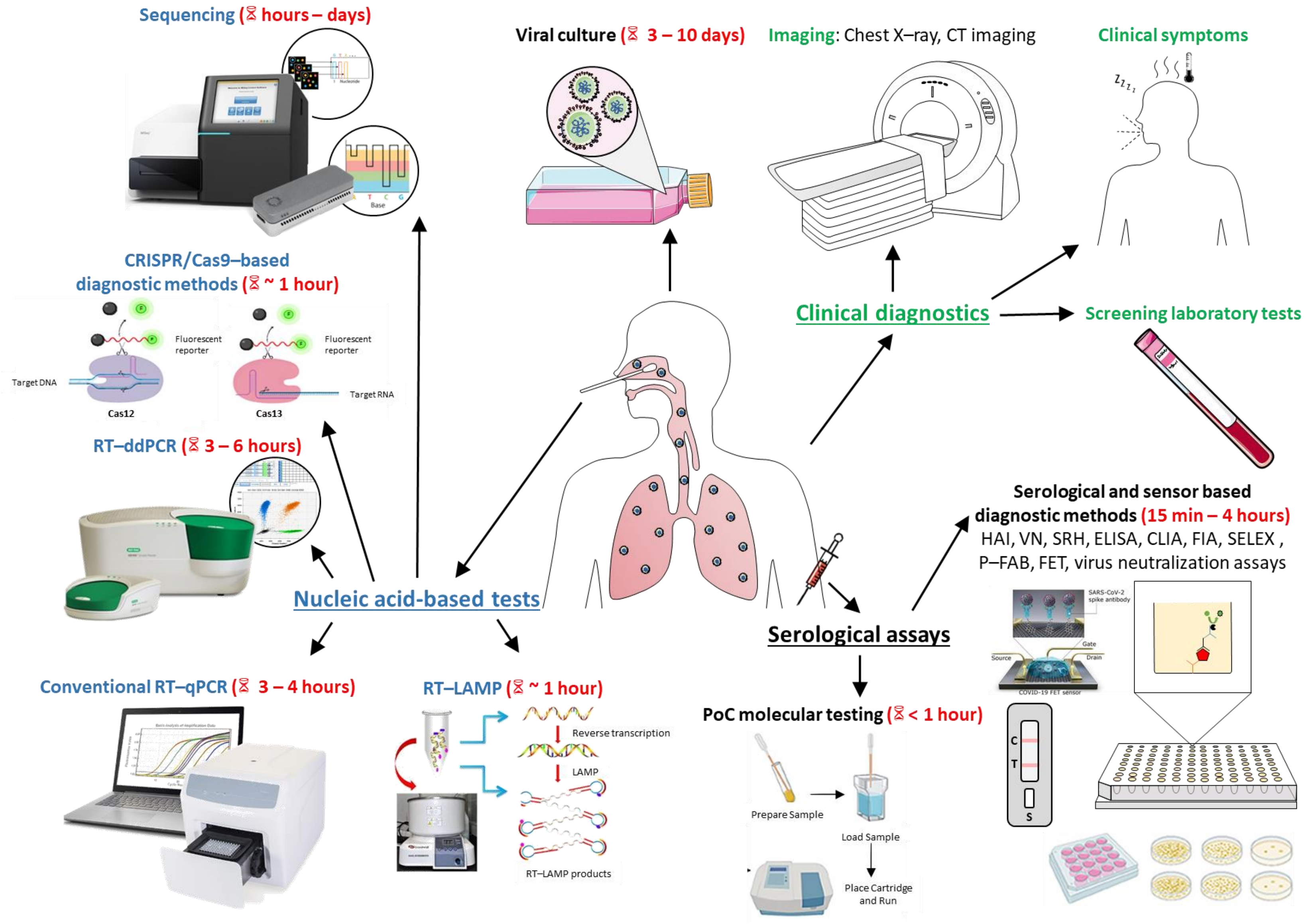

Sequencing (⌛ hours – days)
Nucleotide
Base
Viral culture (⌛ 3 – 10 days)
Imaging: Chest X–ray, CT imaging
Clinical symptoms
CRISPR/Cas9–based diagnostic methods (⌛ ~ 1 hour)
Fluorescent reporter
Fluorescent reporter
Target DNA
Cas12
Cas13
Target RNA
Clinical diagnostics
Screening laboratory tests
RT–ddPCR (⌛ 3 – 6 hours)
Serological and sensor based diagnostic methods (15 min – 4 hours)
HAI, VN, SRH, ELISA, CLIA, FIA, SELEX , P–FAB, FET, virus neutralization assays
Nucleic acid-based tests
Serological assays
SARS-CoV-2 spike antibody
Gate
Source
Drain
COVID-19 FET sensor
Conventional RT–qPCR (⌛ 3 – 4 hours)
RT–LAMP (⌛ ~ 1 hour)
Reverse transcription
LAMP
RT–LAMP products
PoC molecular testing (⌛ < 1 hour)
Prepare Sample
Load Sample
Place Cartridge and Run
C
T
S

Figure 1.6: Overview of the available clinical, diagnostic and research strategies for the diagnosis of a COVID-19 infection.

1.4.1. Clinical diagnostics

The most basic way of diagnosing infections is based on the clinical presentation of the patient. Clinical diagnostics comprise the initial assessment that may raise suspicion but does not provide definitive identification of the viral pathogen. Clinical diagnostics include symptoms, imaging, and laboratory markers not specific to the virus [482].

In the case of influenza, this method is often used in surveillance programs where they focus on the trends in illness presentation on a larger scale rather than the diagnosis of individuals. Common case definitions are ILI and SARI surveillance. Based on the symptoms, it is not possible to distinguish influenza from other respiratory pathogens. However, during seasonal epidemics it is fairly accurate and should be suspected in cases of ARDS, pneumonia, sepsis, myocarditis, encephalitis or rhabdomyolysis [484].

In the case of the COVID-19 pandemic, mainly chest CT scans have been used as a complementary approach for early SARS-CoV-2 diagnosis and evaluation of disease progression. Non-specific biomarkers, such as leukopenia, lymphopenia, elevated aminotransaminase levels, elevated lactate dehydrogenase, and elevated inflammatory markers [485, 486], can also indicate other infectious diseases. However, they were frequently relied upon at the beginning of the pandemic, when specific testing capacity was extremely limited [241, 487]. Finally, artificial intelligence (AI) shows promise for automated detection of COVID-19 using pattern recognition algorithms [488]. There have already been studies where machine learning used CT scans to distinguish COVID-19 from other pneumonia causes [489]. Smartphone-based applications using breathing and coughing sounds and results from breath-analyser tests also have been proposed to be used by machine learning algorithms for COVID-19 self-testing [490–492].

1.4.2. In vitro diagnostics

Due to the similarity between many of the viral respiratory tract illnesses, the causative agent has to be confirmed with a laboratory diagnosis. *In vitro* diagnostics include serologic antibody and antigen-based assays and nucleic acid amplification tests (NAATs) (Figure 1.6). These are broadly applicable in different settings of public health, clinical care or epidemiologic investigations and are recommended for suspected cases [493]. Though all of these tests are fundamentally different, the quality of the sample is crucial for successful detection. The collection method, type of sample (e.g. serum, blood, sputum, faecal matter, nasal swabs…) and its clinical relevance for diagnostics (e.g. for respiratory infections samples taken from the respiratory system will typically yield the highest concentration), and sample preprocessing all play an important role in the detection regardless of the chosen method.

1.4.2.1. Serological assays

Generally, serological assays are used to give information about past infections. They would not guide treatment options because antibodies are generally detected in serum samples taken at the acute and convalescent phase of the infection and the infection will be resolved by the time the second convalescent sample is collected. However, serology does provide important epidemiological data that can influence patient management, for example for the selection of the strain for influenza vaccines.

The assessment of prior viral exposure and thus potential immunity can be performed using serological measurement of specific antibodies. For most viral infections, viral RNA detection or the detection of IgM and IgG antibodies indicate an acute infection, while a significant IgG increase indicates a recent infection. For SARS-CoV-2 (Figure 1.7), neutralising antibodies are found in up to 50% of the infected, immune competent individuals by day 7 and in all infected individuals by day 14. During the first week after SARS-CoV-2 infection, the IgM levels increase, peaking after two weeks and then decrease back to near-background levels in most individuals, which makes it an indicator of early-stage infection. After one week, IgG is detectable and a high level is maintained for a long period, sometimes even more than 48 days [494]. Thus, IgG may serve as protection against reinfections. IgA can be detected between four and ten days after infection.

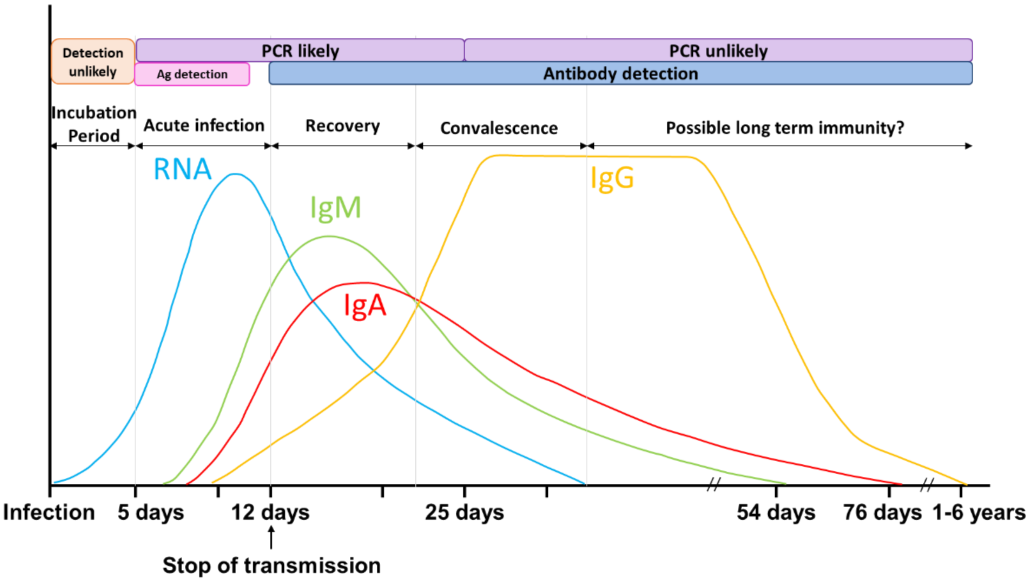


Figure 1.7: Approximate timeline from SARS-CoV-2 infection. The incubation has been reported to be 2 to 14 days. A specific IgM antibody response starts and peaks within 7 days and continues as long as the acute phase of the disease continues. Specific IgG and IgA antibodies are developed a few days later after IgM.

1.4.2.1.1. Antibody-based test

Antibody serology as a diagnostic tool can be especially interesting in case of patients with delayed clinical presentation, who may be missed by NAAT [495]. However, in most cases, antibody assays are suboptimal in a pandemic context due to delayed seroconversion and performance variability [496]. Antibody assays are of particular use for epidemiological purposes, estimation of the attack and case fatality rate [497], and to evaluate the impact of control measures. Furthermore, antibody evaluation allows assessment of vaccine immunogenicity and identification of plasma donors, especially in the elderly or otherwise immunocompromised people [495, 497, 498]. The design of antiviral drugs or vaccines can also be facilitated by the cross-reactivity with viral antibodies [499, 500]. Finally, potential zoonotic disease transmission from wild-life reservoirs can also be identified by serological surveillance [501, 502].

The most commonly used serological assays for influenza detection are HAI assay microneutralisation or virus neutralisation assay (VN) complement fixation assay, single radial hemolysis (SRH), and enzyme linked immunosorbent assay (ELISA). For SARS-CoV-2 detection, chemiluminescent immunoassays (CLIA), ELISA, and lateral flow immunoassays (LFIA) are currently the marketed platforms for serologic evaluation of antibodies.

The HAI assay is mostly used to confirm the presence of HA-specific antibodies of the influenza virus in serum due to vaccination or natural infection. This assay uses serum dilutions to find the highest dilution where complete hemagglutination is prevented by the ability of the HA-specific antibodies to hinder the attachment of the virus to erythrocytes [503]. This is a simple and inexpensive test, however, its poor sensitivity limits the usability for virus diagnosis [504].

Following a natural infection or vaccination, VN assays can measure the introduction of virus-specific antibodies. This assay is based on preventing the viral infection of cells by the ability of virus-specific antibodies to neutralise the virus. The virus neutralisation titer is the highest dilution at which virus infection is completely blocked [504]. This assay is more sensitive compared to the HAI assay, however, its application in routine is limited as infectious viruses need to be used in certified Biosafety Level (BSL) 2+ and BSL3 laboratories [503].

Complement fixation is an immunodiffusion-based approach that forms antigen-antibody complexes in case of vaccinated or infected patients which can be visualised. Due to its low sensitivity, it is not often used anymore [505].

SRH is mostly used to determine the introduction of antibodies after vaccination or natural infection. This assay measures complement-mediated hemolysis induced by antigen-antibody complex. SRH is more sensitive than HAI and no pretreatment of the serum is needed to inactivate non-specific inhibitors.

Similar principles are used by CLIA, ELISA and LFIA assays, but they differ in the method of antibody-antigen binding detection [506]. CLIAs use magnetic, protein-coated microparticles to mix patient samples and generate a light-based, luminescent readout [507, 508]. ELISAs can be quantitative or qualitative and possibly involve several manual steps which increases their time to results. In an ELISA, the antigen is immobilised either directly on the surface or by using a capture antibody that is immobilised on the surface after which a clinical sample is added. Subsequently, the antigen is complexed to a detection antibody that is labelled by a fluorophore or an enzyme (direct), such as horseradish peroxidase (HRP) or alkaline phosphate (AP) or is not labelled (indirect). LFIAs are small and portable and they are suitable for qualitative point-of-care (POC) assessment, which results in the presentation of a coloured line after the addition of specimen to the strip [509].

In addition to these conventional techniques, automated detection of antibodies are facilitated by the recent development of SARS-CoV-2 proteome-based microarrays [510]. This high-throughput format allows the generation of more systematic descriptions of antibody binding and viral antigens [511]. Furthermore, a programmable phage-display immunoprecipitation and sequencing technology platform, VirScan, has also been adapted for the detection of SARS-CoV-2. One drop of blood is required to scan over 1000 virus strains. SARS-CoV-2 exposure is predicted with 99% sensitivity and 98% specificity using a machine learning model trained on VirScan data. Although this type of approach could be interesting to understand past exposure epidemiology, it is not yet widely suitable or available for acute diagnostics [512]. Another antibody assay in development uses biosensors with polyaniline nanofibers-coated optical fibres for serological measurements. These could eventually be used in a plug-and-play format [513]. Finally, a microfluidic ELISA system is being developed to detect COVID-19 antibodies via a lab-on-chip platform. A microfluidic device separates plasma, which is used to detect antibodies using a semi-automated on-chip ELISA. Although the use is simpler than manual ELISA, this platform still needs performance evaluation [514].

1.4.2.1.2. Antigen-based test

Another type of serological assay is antigen testing, which is an attractive potential POC diagnostic. It detects protein fragments within or on the virus in specimens collected from the nasal cavity of nasopharyngeal swabs [515]. More specifically, the viral antigens will bind to the corresponding antibody which can then be detected using optical, electrochemical, magnetic, and surface plasmon resonance-based techniques [516]. These tests are more rapid compared to RT-qPCR which takes hours.

Rapid influenza diagnostic tests (RIDTs) are antigen-based tests that can be used in POC settings for rapid diagnosis of influenza virus infections. Monoclonal antibodies are used to target viral NP and make use of either immunochromatographic or enzyme immunoassay techniques. They are often used for diagnosis of influenza infections as they can be completed in less than 30 minutes, with results visually indicated based on a colour change or another optical signal [503]. However, they cannot distinguish influenza A from influenza B infections or between influenza A subtypes [503, 517].

For SARS-CoV-2, the N protein is considered an excellent target based on previous experience for a diagnostic sandwich assay using monoclonal antibodies. During replication, the N protein is abundantly secreted and the cross-reactivity with other human coronaviruses is low [518, 519]. Currently, the widely available SARS-CoV-2 antigen kits use two main approaches, namely the fluorescence immunochromatographic assay (FIA) that provides results via an automated immunofluorescence reader [520] and the immunochromatographic (ICT) assay based on colloidal gold conjugated antibodies that result in visible coloured bands to indicate positivity [521]. A SARS-CoV-2 specific antigen can also be detected by using nanotechnology in biosensor devices. A fibre-optic absorbance biosensor (P-FAB) platform and field-effect transistor (FET) based biosensing device have been developed to detect N and S proteins from SARS-CoV-2, respectively [522, 523]. However, developing effective antigen tests is challenging due to the lack of antibodies for specific proteins of SARS-CoV-2. SELEX (systematic evolution of ligands by exponential enrichment) based strategies are useful for the identification of affinity ligands such as aptamers specific to SARS-CoV-2. These aptamers are significantly cheaper and easier to produce compared to antibodies [524]. Furthermore, antigen tests have been reported to be less sensitive compared to RT-qPCR and could be less reliable in case of low viral load. Although minimal sample preprocessing is needed and the sample-to-result time of antigen assays is relatively fast [523], additional external validation is needed before its incorporation into clinical practice.

1.4.2.2. Nucleic acid-based tests

NAATs are much more sensitive in comparison to serological and antigen-based tests and can detect virus-derived nucleic acids in clinical samples much earlier in time. NAATs mainly include RT-qPCR, RT droplet digital PCR (ddPCR), Loop-Mediated Isothermal Amplification-Based Assay combined with reverse transcription (RT-LAMP), Clustered Regularly Interspaced Short Palindromic Repeats D(CRISPR) based methods, Nucleic Acid Sequence-Based Amplification (NASBA), Simple Amplification-Based Assay (SAMBA), and nucleic acid sequencing approaches that will be discussed in Section 1.4.3.

RT-qPCR is considered the gold standard for influenza diagnostics among NAATs. The extraction is followed by RT-qPCR from which the extracted RNA is reverse transcribed into cDNA using a reverse transcriptase. Reverse transcription can be conducted by using different primers, including gene-specific, random or oligo-dT, depending on the cDNA yield, type of RNA and specificity [525]. The produced cDNA can then be used for the qPCR step where the target is exponentially amplified [493, 526, 527]. Finally, the PCR products are amplified coupled with fluorescent detection of labelled PCR products [503]. RT-qPCR can easily target multiple viruses, but also different regions within a virus. A multiplex RT-qPCR enables the possibility of targeting for example both influenza A and influenza B viruses in one reaction. Targeting at least two regions of a virus helps to avoid the detection of false-negative results due to genetic modifications [528, 529]. Other false negatives can be caused by collecting the sample when the viral load is low due to the timing of sampling, poor sample collection technique resulting in reduced quantity or quality, degradation due to inappropriate transport of the unstable RNA virus, and technical limitations of the RT-qPCR test [493, 530–532]. A large disadvantage of RT-qPCR methods is that it needs to be performed in certified labs where expertise, specialised equipment and well-developed specimen management infrastructure are available. The significant burden on most labs of large-scale testing during the COVID-19 pandemic resulted in interest in a reliable POC molecular test, including rapid NAAT and rapid antigen POC tests, that produces rapid results and facilitates timely patient management decisions [533, 534]. However, the performance of these tests varies greatly. Some rapid NAATs and rapid antigen tests are considerably less sensitive compared to RT-qPCR tests. All POC tests are able to identify highly infectious cases when the viral loads are high, but sensitive tests are also important to prevent disease transmission by detecting early pre-symptomatic infections.

Another approach, RT-ddPCR can detect a target and perform absolute quantification using the principles of sample partitioning and Poisson statistics, which facilitates surveillance of intra and inter-case variability [535]. During RT-ddPCR, the sample is divided into thousands of micro-reactions of defined volume [536]. Normalisation and calibrator issues associated with RT-qPCR are overcome by RT-ddPCR, which increases the precision of the method. Furthermore, RT-ddPCR is more sensitive to detect low target copies and is relatively insensitive to potential PCR inhibitors [537].

RT-LAMP uses DNA polymerase and 4 to 6 primers that bind to distinct target regions of the genome. The analysis of the reaction products can be done using conventional DNA-intercalating dyes, UV-light illumination, agarose gel electrophoresis, real-time fluorescence, or by end-point colorimetric readouts through the detection of reaction by-products, such as protons and pyrophosphate, released during DNA polymerization [538]. Multiplex RT-LAMP assays have been designed to detect subtypes A(H1N1), A(H3N2) and influenza B viruses in a short time [539]. Due to the isothermal conditions, the greatest advantage of RT-LAMP is the low-cost field deployment because an expensive thermal cycler is not needed [540]. In addition, the short sample-to-result of approximately one hour and the adaptation to smartphones to be used as a personal POC diagnostic makes it also an interesting alternative [526, 541]. However the primer design is often complex, time-consuming and requires significant expertise [542]. Moreover, currently they are not yet sufficiently specific or sensitive. Due to the presence of multiple primer pairs, there may be an increase of non-specific by-product formation [543]. Additionally, inefficient amplification of the target sequence is caused in case of low viral loads [543].

The interest in CRISPR based methods for infectious disease applications has substantially increased the past few years [544], because of the low cost, the short sample-to-result of approximately one hour and particularly in the setting of infrastructure constraints [545]. The effector enzymes CRISPR-associated (Cas) proteins can recognize and cut CRISPR that belongs to a family of palindromic nucleic acid repeats found in bacteria [546]. These Cas proteins are very sensitive and can be programmed to identify and cut SARS-CoV-2 sequences [481] which would then generate a detectable signal [515]. However, these CRISPR-based methods still need careful validation and field testing [547].

NASBA amplifies multiple genes in a target RNA sequence using reverse transcriptase, RNaseH and RNA polymerase [548]. A forward primer with T7 promoter region binds to a target RNA sequence and is extended by reverse transcriptase after which RNaseH breaks down the original RNA target sequence. A second primer attaches to the new amplicon and using reverse transcriptase the sequence is extended. Finally, RNA is synthesised by binding the T7 RNA polymerase to the extended amplicon. These steps are repeated until the RNA is detectable [549]. SAMBA is a NASBA-based method that uses a nitrocellulose dipstick to visualise the test result. After the isothermal amplification via NASBA, a dipstick was inserted into the reaction mixture to visualise the signal [550].

Finally, genomic sequencing is essential to trace transmission patterns and for the phyloepidemiological evaluation of changes in the viral genome over time [551]. Moreover, sequencing allows the identification of the subtype or variants to which the virus belongs to, because it can take the whole genome or at least a larger part of the genome into account compared to NAATs. This method will be further discussed in the following Section 1.4.3.

1.4.3. DNA sequencing

Although the double helix structure of DNA was discovered in 1953 [552], it took another fifteen years before a segment of the DNA sequence could be determined [553]. Since the publication of the first bacterial genomes [554, 555], first shotgun-sequenced genome [556] and draft human genome [557], the cost of sequencing has been rapidly decreasing.

Consequently, methods and data from whole genome sequencing (WGS) are increasingly being used. Although DNA sequencing is currently not a first line diagnostic test in routine surveillance, it can be used to further characterise the pathogen. While WGS is too labour-intensive and costly in case limited genomic information is wished to be extracted, however, the increasing number of resistance genes and viral variants, the use of sequence data for transmission studies and other advancements in the field are driving the increased use of WGS. Moreover, as WGS characterises the whole genome, it also allows the classification of influenza viruses and SARS-CoV-2 variants.

1.4.3.1. Viral sequencing methods

After sample collection from the patient, animal or environment, the sample will contain a combination of various organisms. Typically, the viral DNA and/or RNA in respiratory samples is low for direct sequencing due to the presence of contaminating host DNA [558]. Currently, virus sequencing is achieved by ultra-deep sequencing or through viral enrichment before sequencing either directly or by concentrating virus particles. The three main viral sequencing methods are metagenomic sequencing, PCR amplicon sequencing and target enrichment sequencing.

Metagenomic sequencing has been extensively used for the characterisation of microbial diversity in clinical and environmental samples and the discovery of novel pathogens [559, 560]. The DNA and/or RNA from the host, bacteria, viruses, fungi and other pathogens is extracted from the sample and after the preparation of the library it is sequenced with RNA sequencing or shotgun sequencing. Due to the presence of contaminating nucleic acids coming from the host and commensal microorganisms, the sensitivity decreases because the proportion of viral reads is often low [558]. Consequently, additional steps are often added such as concentrating virus particles through non-specific amplification methods, sequencing at high read depth to increase the amount of virus sequences, and the depletion of host material using for example filtration, centrifugation and nuclease treatment. However, these methods add to the cost, therefore, metagenomic sequencing is typically only used on a small number of samples for research purposes [561, 562]. Moreover, appropriate bioinformatic tools and databases that need high-performance computational resources are needed for pathogen discovery or diagnosis.

PCR amplicon enrichment is the most common approach for enriching small viral genomes before sequencing by using primers complementary to a known nucleotide sequence. However, multiple overlapping sets of primers may be needed to ensure the amplification of all genotypes due to the heterogeneity of RNA viruses. PCR amplicon sequencing is favourable compared to metagenomic methods in case of low virus concentrations [563]. Although it is technically possible to use PCR-based sequencing of viruses as large as 250 kb, the technical complexity due to multiple PCR reactions makes it impractical for viruses of more than 20-50 kb with the current technologies. Moreover, highly variable pathogens may cause problems for PCR amplification, such as primer mismatches [564] and primer amplification [563, 565].

Finally, target enrichment, also known as capture, pull-down or specific enrichment methods, can be used to directly sequence viral genomes from clinical samples without the need for prior PCR or culture [566–568]. Small DNA or RNA probes that are complementary to a pathogen reference sequence or a panel of reference sequences are typically used to capture the complementary DNA sequences from the total nucleic acids in a hybridization reaction. This method allows the use of overlapping probes to cover the whole genome, unlike PCR amplicon enrichment methods. Subsequently, sequencer-specific adaptor ligation and a small number of PCR cycles are used to enrich the ligated fragments. The advantages of this method are the fewer mutations than in PCR amplified templates and more representative sequences of the original virus compared to cultured virus isolates [566, 569]. If the probes are designed against a larger panel of reference sequences they will lead to a better capture of the diversity in and between samples. However, if one probe fails, other probes can still capture overlapping and internal regions [566, 569]. Therefore, target enrichment is not appropriate for the characterisation of new viruses because of low homology to known viruses. However, because of the increase in percentage of viral reads in sequencing data and the improvement in depth and quality of sequences, more samples can be sequenced per run unlike metagenomic libraries [569].

1.4.3.2. Sequencing platforms

1.4.3.2.1. First-generation sequencing

The most commonly used first-generation DNA sequencer is based on the chain terminator or dideoxy method [570]. This method includes the purification and denaturation of DNA and subsequent amplification through cloning or PCR. The resulting DNA is divided into four tubes that each contain one of four radiolabelled dideoxynucleotides (ddATP, ddCTP, ddGTP or ddTTP) that act as chain terminators and four normal nucleotides. By chance a dideoxynucleotide (ddNTP) is incorporated instead of the normal nucleotides, resulting in fragments with varying lengths. If the ddNTPs are incorporated into a DNA strand, the extension will be terminated at this point and the last nucleotide of all strands in a tube will correspond to the added ddNTP. These double-stranded DNA fragments are subsequently denatured and used in a gel electrophoresis with one lane per ddNTP, which enables the determination of the DNA sequence with radiography, in case radiolabeled nucleotides are used. DNA reads up to 200 bases could be produced using this method. By using dye-terminators the Sanger method could be improved [571]. The four ddNTPs are tagged with a fluorescent dye, which allows one reaction for sequencing instead of four. Additionally, the genomics evolution was further aided by the development of techniques such as PCR [572, 573] and recombinant DNA technologies [574, 575]. These technologies generate high concentrations of pure DNA species required for sequencing and due to the increasing number of sequenced genomes more appropriate polymerases were found [576]. Using this method, reads up to 1000 nucleotides could be produced with an error rate of 0.001%. In 1986, this Sanger method was commercialised by Applied Biosystems [577] and the first automated sequencing machine, ABI 370A sequencer, was developed. After further improvement of the method which allowed simultaneous sequencing of hundreds of samples, this became the preferred method for large sequencing projects such as the Human Genome Project [557, 578]. Sanger sequencing also made it possible to monitor influenza evolution by characterising the prevalent genetic sequencing among many influenza viruses in a sample. Consequently, several new observations were made, including adaptive evolution of multiple co-circulating viral lineages and the prevalence of reassortments [579]. However, there are clear disadvantages to Sanger sequencing. Only minority variants at frequencies between 10% and 40% are able to be detected by Sanger sequencing. Moreover, Sanger sequencing has limited power to sequence complete genomes, with high associated costs [580].

1.4.3.2.2. Second-generation sequencing

Together with the development of large-scale Sanger sequencing efforts, second-generation DNA sequencing, also named high-throughput or next-generation technologies, emerged in 1996 [581]. NGS reduces the cost per base to a level unattainable with the traditional Sanger sequencing methods. Moreover, it has a clear advantage over Sanger sequencing due to its ability to identify individual viral genomes in complex mixtures. Additionally, it allows thorough identification of minority variants, which is only possible to a limited extent with Sanger sequencing. Until 2008, DNA sequencing costs approximately halved each two years which follows a similar pattern to Moore’s Law that predicts a doubling of computing power every two years. However, this trend was broken in January 2008 due to these high-throughput sequencers resulting in a rapid decline of sequencing costs [582] (Figure 1.8).

The most commonly used second-generation technology is based on sequencing-by-synthesis technology that uses fluorescent dyes [583]. Fluorescently labelled sequencing was developed by Solexa, which was acquired by Illumina in 2007, and was based on reversible dye-terminators technology and engineered polymerases. Multiple copies of the target are produced in clusters by immobilising DNA fragments onto a flow cell with primers where a PCR is carried out. Four types of fluorescently labelled reversible terminating bases are presented to the clusters in the presence of DNA polymerase. These four bases compete with each other for the binding sites on the template DNA and if they are incorporated, a laser will excite the labelled dyes and a camera will take images of the fluorescently labelled nucleotides on the flow cell. A next cycle can start after the chemical removal of the terminating 3' blocker dye from the DNA. All sequenced fragments will have the same length because the length is determined by the number of cycles [584]. The possibility to sequence clusters from both directions is a large advantage of this technology, because pairs of coupled reads can be provided. These Illumina sequencers can produce reads up to 300 bp with an accuracy exceeding 99.3% [585]. Illumina systems create a quality score of which Phred is the most widely used quality score. These are logarithmically related to the probability that errors occur in the base calling. For example, Phred of 10, 20 and 30 are related to miscall probabilities of 0.1 (10%), 0.01 (1%) and 0.001 (0.1%) [586]. However, they suffer from homopolymer errors and sequence specific interference errors. Moreover, because of the generation of short-read data, haplotype reconstruction is challenging. Additionally, due to problems such as mapping ambiguities it is more difficult to resolve recombination and repetitive regions. Second-generation sequencers enabled the sequencing and analysis of the intrahost diversity of influenza and SARS-CoV-2 genomes isolated from individual patients. This revealed that humans can harbour multiple variants that can have different sensitivities to antiviral drugs [587]. Moreover, due to the possibility of deep sequencing, second-generation sequencers proved to be a powerful method in combination with PCR-based and hybridization techniques for global screening of pathogens [194].

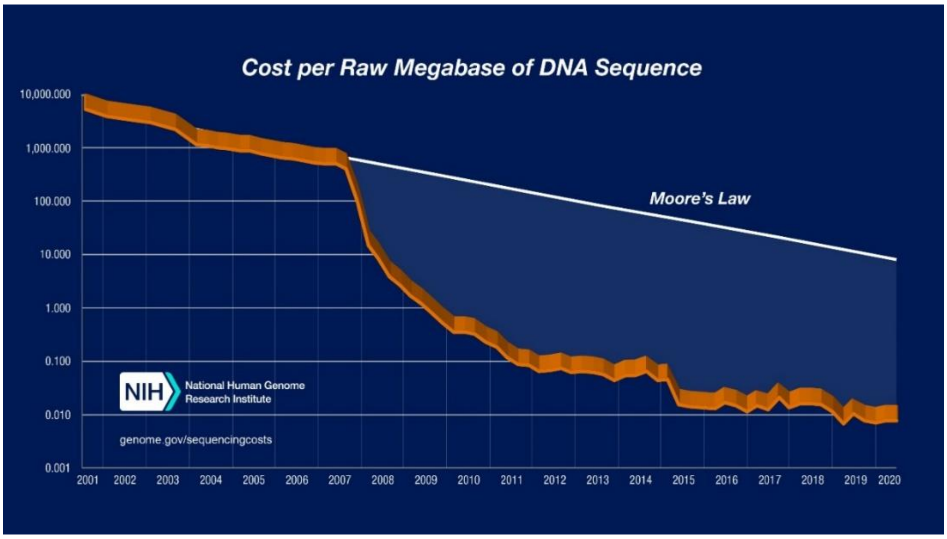


Figure 1.8: Sequencing cost per megabase (August 2020). Source: NIH https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

1.4.3.2.3. Third-generation sequencing

Third-generation sequencing or long-read sequencing is a sequencing method currently under active development [588]. Compared to second-generation sequencing, they have the ability to produce substantially longer reads that generally ranges from 10 000 to 100 000 bp [588]. Longer reads are an obvious advantage as various computational challenges such as the genomic assembly, metagenomics and transcript reconstruction will be considerably simplified [588] in addition to reliably resolving repeat sequences or large genomic rearrangements [589]. Other important advantages include the sequencing speed, portability and real-time analysis as the sequencing process is not parallelized across regions of the genome [590]. However, third-generation sequencing data is known to have a much higher error rate, but due to a great deal of research and development these error rates are continually improving [591]. The primary third-generation sequencers currently include Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT) which were commercialised in 2011 [592] and 2014 [591], respectively.

PacBio developed a parallelized single molecule real time sequencing (SMRT) which is based on the properties of nanostructures called zero-mode waveguides (ZMW) [593]. At the bottom of a ZMW, a single DNA polymerase enzyme is attached with a single DNA molecule as a template. The DNA to be sequenced exists as a single-stranded circular DNA called SMRTbell template, which is generated by ligation of hairpin adaptors (SMRTbell adaptors), to both ends of the double-stranded DNA template molecule. The ZMW in the SMRT cell is able to create an illuminated observation volume small enough to observe the incorporation of a single nucleotide of DNA by DNA polymerase. After the SMRTbell library is loaded in the SMRT cell, the DNA polymerase will bind to the adapter of the SMRTbell and the replication can begin. Four different fluorescent dyes are attached to one of the four DNA nucleotides and are used during replication. After the incorporation of a nucleotide by the DNA polymerase, a signature light pulse is produced and captured which happens during replications across all ZMWs in the SMRT cell. These light pulses are translated to nucleotide sequences and the obtained sequence from each ZMW is called a Continuous Long Read (CLR). Due to the circular DNA template, the polymerase can continue through the adaptor to replicate the second DNA strand [594]. The read depth can be increased and base call accuracy can be improved by continually amplifying the circular template multiple times. PacBio sequencing runs are limited by the finite functional life of the polymerase molecule. This results in either single reads for ultra-long templates of currently more than 135 kb reads [595] or multiple contiguous reads with read lengths of approximately 13.5 kb of both strands for shorter templates [594]. After accomplishing many improvements to the technology, the Sequel II system can produce long reads that are highly accurate with reports of up to 99.8% [596].

Additionally, as the data is collected in real-time a faster turnaround time is offered compared to second-generation technologies. Also, no PCR amplification is needed, which avoids problems like the amplification difficulties of AT- and GC-rich regions. However, significant disadvantages of the PacBio method are the comparatively large size of the instrument and the more considerable financial investment for start-up.

Sequencing platforms developed by ONT involve passing a DNA molecule through a nanoscale staphylococcal α-haemolysin protein pore structure [597] and subsequently measuring changes in the electrical field surrounding the pore [598]. Two ionic solutions are separated by the membrane in which the nanopores are present, which allows an electrical current to flow through the nanopores. An adaptor is ligated to double-stranded DNA to facilitate its capture by the protein pore and subsequently unwound. These libraries are loaded onto a flow cell that contains up to 2048 nanopores for a MinION flow cell embedded in a membrane. The ion current and a preloaded motor enzyme move the single strand through the Nanopore. A characteristic disruption in ion current is detected for each nucleotide that passes through the pore and translated to base calls [598]. The run can effectively continue until satisfactory results are achieved, because very few depletable reagents are used during the sequencing process. Long sequences can traverse uninterrupted with the limiting factor being the preparation of high molecular weight DNA [599–601], which determines ultra-long reads (> 100 kb) or standard long reads (10-100 kb). Although the accuracy currently ranges between 87 and 98%, a great deal of research and development is occurring to improve the nanopore structure and function. One strategy is using 1D^2 sequencing, that attaches a special adaptor to one end of the double-stranded DNA template molecule. This allows sequencing both the complementary and template strands contiguously, which provides higher sequence accuracy. In addition to the MinION that contains one flow cell, ONT has released PromethION and GridION X5 platforms (respectively 42 and 5 flow cell configurations with 12 000 and 2048 nanopores divided over 3000 and 512 channels per flow cell), which allows vast throughput and scaling to many whole genomes per run. Additionally, a single-use flow cell for the MinION with only 126 pores, called a Flongle, was released where lower throughput is adequate for example for smaller experiments. The Flongle allows low-cost and rapid sequencing. Nanopore sequencing has many potential advantages over other platforms, such as being highly portable, capable of sequencing when plugged into a laptop and sequencing in real-time, shortening turnaround times [598]. Furthermore, no amplification is needed which allows the preservation of nucleotide modifications such as methylation on the template and extremely long-read lengths can be produced which makes it an excellent approach for *de novo* genome assembly [598]. Also, through inversion of the voltage the pore can eject DNA molecules, which enables adaptive sampling. A potential use of adaptive sampling is rejecting host reads in clinical samples from being sequenced to target the pathogen [602].

The advent of third-generation sequencers appears to be promising for on-site real-time and large-scale influenza and SARS-CoV-2 surveillance. Moreover, within-host viral variants (haplotype), such as quasispecies in respiratory samples and various virus variants in wastewater samples, can be reliably reconstructed by using long reads. Short reads lead to poor consistency in haplotype reconstruction across the different bioinformatic tools [603, 604].

1.4.3.3. Bioinformatic analysis of viral sequencing data

WGS has proven to be invaluable and essential to the field of microbiology, because it makes systematic investigations of genes and other components within the genome possible. This leads to a better understanding of how biological systems function and to the discovery of novel species, redefining existing taxonomies and obtaining unprecedented insights into the genetic structure, diversity and evolutionary relatedness. Due to the rapid progress and innovation of NGS technologies, large volumes of sequence data have been generated and the expensive cost for WGS was reduced. After the release of the Illumina HiSeq and Illumina MiSeq sequencers in 2010, the number of sequenced genomes increased substantially. Over 35 million viral sequences are available in GenBank and GISAID (Figure 1.9), respectively, and this number is expected to increase substantially over the coming years. These advances in genome sequencing especially demonstrated its value during the COVID-19 pandemic compared to any previous outbreaks. Although genome sequencing can provide raw nucleotide sequences, further analysis is needed to understand the medical or biological meaning of these sequencing results. As massively parallel sequencing produces massive amounts of data, bioinformatics plays an important role in analysing and interpreting sequencing data. These methods to analyse sequencing data are continuously being developed and refined including advancements in sequencing technology and in computational resources and algorithms. This thesis focussed mainly on Illumina sequencing, which is also reflected here. Often traditional Illumina tools will not work for the analysis of other platforms, such as ONT sequencing.

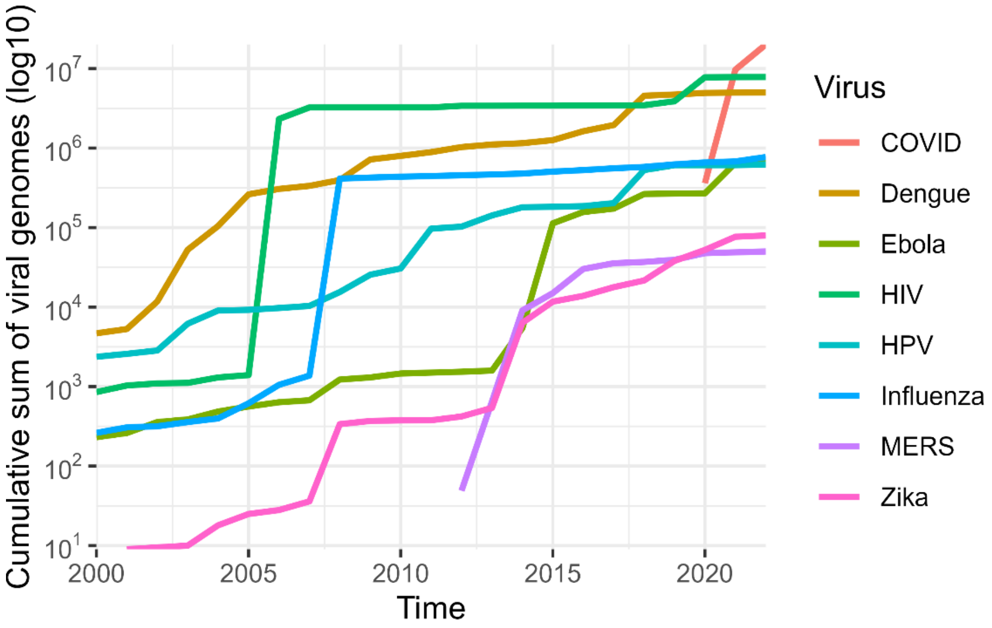


Figure 1.9: Cumulative number of viral genome submissions to NCBI and GISAID (for influenza and SARS-CoV-2 for major human viral pathogens. Y-axis shows the log10(cumulative sum of viral genomes) and the x-axis shows the date of the sample collection from 2000 until 3 November 2022.

1.4.3.3.1. Quality control, read filtering and trimming

The instrument software will usually perform the base-calling of input DNA. This will mostly result in FASTQ files, which is a combination of the base calls and their respective quality scores. The most widely used quality score code is Phred. Generally, reads with average Phred below 20 are considered as poor quality reads. These low quality base calls may be harmful for the NGS analysis as they potentially add unreliable bases [605].

Before read assembly and other downstream analysis, it is recommended to perform a quality control. Although a consensus is often lacking, these will generally include an overview of read lengths, base quality distributions, possible artefacts and contaminations [606, 607], presence of N nucleotides in reads, GC content, degree of (PCR-) duplication and sequencing bias. FastQC [608] and PICARD [609] are well-known tools available for initial quality assessment. Additionally, they can filter on low quality/complexity reads, error correction, etc. For most second-generation sequencing technologies, including Illumina, the quality will generally drop towards the end of the reads. Furthermore, sequence trimming should be performed to remove known adapter sequences originating from library preparation as they can affect the downstream analysis profoundly. Trimmomatic [610], SOAPnuke [611], and Cutadapt [612] are sequence matching-based adapter trimming tools that can be used as adapter trimmers in addition to performing maximum or window information quality filtering. Low quality bases are removed by window information quality filtering by scanning from the 5' end of the read. When the average quality of a group of bases drops below a specified threshold, the 3' end of the read is removed. Instead of a specified threshold, maximum information quality filtering will become more strict as it progresses through the read [610]. Trimming can also remove low quality portions of the sequence, while the high quality parts of NGS reads are preserved. However, over-trimming can lead to the loss of information, while retaining low-quality base reads and contaminants might decrease accuracy and performance in subsequent steps [613, 614]. Also, the quality of the throughput can also be evaluated by the number of high quality bases or reads, which is closely related to the depth of the sequencing coverage (i.e., how many times each position in the genome is covered). A threshold for the depth coverage to obtain reliable results is greatly dependent on the application. For example, the ECDC recommends a minimal depth sequencing coverage of 10X across more than 95% of the genome to obtain a complete SARS-CoV-2 genome, while a higher depth sequencing coverage of 500X across more than 95% of the SARS-CoV-2 genome should be obtained to determine low-frequency variants [615].

1.4.3.3.2. Genome assembly

Larger contiguous sequences or 'contigs' can be constructed by the assembly of preprocessed reads. Genome assembly is necessary due to the limitation of short-read sequencing that an entire genome cannot be read in one read. Here, there are two possibilities, either the genome sequence of the studied species is unknown, so the reads have to be *de novo* reassembled from scratch, or the reference genome is available and the reads can be aligned to this reference.

While *de novo* assembly is more complicated compared to reference mapping, it has the advantage that no prior information is needed, and the resulting assembly is not biased by the selection of the reference genome. The first *de novo* algorithms based themselves on pairwise evaluations of overlap between reads. However, these algorithms were very computationally intensive for large numbers of reads [616]. Therefore, most modern assembly algorithms for short reads, such as SPAdes [617], Trinity [618], Trans-ABySS [619], and Megahit [620] are based on De Bruijn graphs that convert short reads to k-mers. This type of assembler does not calculate an error-tolerant alignment between reads, but dissects the reads into overlapping k-mers that represent nodes used to build a De Bruijn graph. If the nodes share one k-mer, two nodes are linked with an edge and subsequently these shared k-mers represent overlaps between reads [613]. This is repeated until all nodes are linked or if no possible further linkage can occur. Due to the emergence of SARS-CoV-2 a specialised assembler, coronaSPAdes [621], was developed, using algorithmic assembly from rnaviralSPAdes and Hidden Markov Model (HMM)-guided algorithms of biosyntheticSPAdes. Obtaining complete genomes using *de novo* assemblers will be rarely possible due to repeats in the genome that exceed the read length of short-read sequencing data [622]. They will, however, still provide information about the putative contig order and orientation, the gaps between the contigs and a draft scaffolding of the assembled contigs. The *de novo* construction of complete genomes is expected to be more easily facilitated by further improvement and cost reduction of long-read third generation sequencing, which makes another assembly algorithm interesting to use, namely the overlap-layout-consensus approach.

Reference-guided assemblers use a reference sequence to construct a draft genome. The aim is to align each read to the reference genome, while allowing errors and structural variation because of the biological divergence or measurement errors of the sequencing machine. Variant identification or calling comprises the identification of regions where the sample and the reference genome diverge and is an important step in the alignment process. These variants include single nucleotide variants (SNV), small insertions and deletions (INDELs), larger structural changes such as inversions or translocations or copy number variations. Consequently, these are complex algorithms that are constantly optimised to improve the speed and accuracy, while preserving a low memory footprint. A major breakthrough was the “Full-text index in Minute space” (FM)-index which is an opportunistic data structure that allows the compression of input text and fast substring queries [623]. Specialised implementations of the FM-index are used in modern short-read mappers, including BWA-MEM [624] and Bowtie2 [625], for DNA sequences. These mappers perform either global (or end-to-end) alignment or a local alignment. A global alignment procures an alignment that involves all of the bases in the read, while a local alignment considers only bases in part of the read, usually omitting bases at the start or end of the read. Consequently, local alignment is often faster because the alignment process stops after a good quality unique match has been identified and is thus interesting in cases where the number of hits is of interest [626]. Generally, most mappers render the result in the Sequence Alignment Map (SAM) format, which can then easily be used for further processing such as merging, sorting and filtering with SAMtools [627] and picard [609], browsing with IGV [628] and recalibrating the quality scores and realigning the sequences to reduce artefacts with GATK [629]. The SAM files can also be converted with SAMtools to BAM files, which contains the same information in binary format.

Although reference mapping generally leads to better results and is less computationally intensive compared to *de novo* assemblers, a sufficiently similar reference genome has to be available [630]. In cases a suitable reference genome is not available or the exact taxonomic classification is unknown in advance, *de novo* assemblers are still often preferred.

As different assemblers use different heuristic approaches to tackle the problems of errors in reads and large repeats in the genome, there will be many differences in the resulting contigs. Therefore, it is important to assess the quality of the assembly. One of the most commonly used tools is QUAST, which calculates various metrics such as the total number of contigs, total number of bases in the assembly, the N50 which is the sequence length of the shortest contig at 50% of the total genome length, minimum length of contig and the L50 count which is the count of the smallest number of contigs whose length sum makes up half of the genome size.

1.4.3.3.3. Phylogenetics

The diagnostic approach of infectious diseases in routine clinical practice has been largely modified due to the introductions of advanced molecular techniques such as NGS and bioinformatics analysis. These techniques improved the ability to identify and control epidemic outbreaks and, therefore, prevent the spread of the pathogen and decrease morbidity and mortality [631]. Furthermore, they enhanced our ability to combat antimicrobial and antiviral drug resistance, develop vaccines, and detect emerging infectious diseases with an increase in accuracy, timeliness and efficiency. Also WGS enabled an improved outbreak investigation, taxonomy, source attribution, and a more detailed understanding of evolutionary history through geographic space and time [632–634]. Phylogenetic information can be extracted from sequencing data by implementing optimal criteria and methods of distance matrices, maximum parsimony, maximum likelihood and Markov Chain Monte Carlo (MCMC)-based Bayesian statistical inference. For all these methods, implicit or explicit mathematical constructs are used to model the observed evolution. These trees are either rooted or unrooted depending on the algorithm and parameters used to generate them. In case of a rooted tree, the MRCA is explicitly identified in a directed graph. A rooted tree is plotted using input sequences as leaf nodes and the genetic distances from the root are proportional to the genetic distance from the hypothesised MRCA according to the underlying model. Unrooted trees do not make assumptions regarding the descent of the input sequences and only describe the relatedness of the input sequences. Consequently, it does not start with the MRCA and does not have a root.

Distance-matrix methods start the construction of a phylogenetic tree by calculating pairwise distances between molecular sequences with a multiple sequence alignment (MSA) as input. The definition of this distance is often the fraction of mismatches at aligned positions, while gaps are either counted as mismatches or ignored [635]. Closely related sequences are thus placed under the same interior node while the branch lengths describe the observed distances between sequences. Depending on the algorithm, rooted or unrooted trees are produced. The main distance methods are based on the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Neighbour Joining (NJ) algorithms. These distance-matrix methods have the advantage of being computationally fast and are thus particularly useful for the analysis of large numbers of samples. The main disadvantage of distance-matrix methods are the several problematic assumptions they make due to overlooking a substantial amount of information in a MSA. These methods are also strongly dependent on the used model of evolution and results in only one possible tree without providing any uncertainty about the confidence of the branches.

Therefore, character-based methods, such as maximum parsimony, maximum likelihood and Bayesian methods, are often preferred as they consider each character or site in the single nucleotide polymorphisms (SNP) matrix while comparing all sequences simultaneously [636]. Methods like maximum likelihood and Bayesian inference can also incorporate various evolutionary models ranging from naïve to very complex models. However, it is important to choose a suitable model as it can have a considerable effect on the resulting phylogenetic tree [637].

Maximum parsimony methods identify the potential phylogenetic tree that needs the smallest total number of evolutionary events or SNV changes to explain the sequence data. Although maximum parsimony methods are computationally the least intensive, they may not be statistically consistent under certain circumstances [638].

Maximum likelihood methods will estimate the parameters of an assumed probability distribution for the observed sequence data. Probability distributions are assigned to particular possible trees and require a substitution model that assesses the probability of particular mutations. Bootstrapping estimates the confidence of the clades in a phylogenetic tree by resampling with replacement the characters (e.g. nucleotides) of a sequence with the same size as the original data, rebuilding the tree and testing whether the same clades are recovered. This is done through many iterations where for example a bootstrap value of 95% is obtained if the same clade is recovered in 95 of the 100 iterations. Similar to maximum-parsimony methods, a tree with more mutations at interior nodes to explain the phylogeny will be evaluated as having a lower probability. However, maximum likelihood methods will allow additional statistical flexibility by allowing varying evolution rates across lineages and sites. Therefore, it is well-suited to analyse distantly related sequences. However, this method is computationally intensive and slow and is highly dependent on the chosen evolution model. Maximum likelihood based tools that provide many evolutionary models and bootstrap settings are Randomized Axelerated Maximum Likelihood (RAxML) [639], PhyML [640] and IQ-tree [641].

Bayesian inference of phylogeny is the combination of the prior information and the information in the data likelihood to create the posterior probability of trees. The posterior probability is the likelihood that the tree is correct given the likelihood model, the priors and the input data including sequence data and optionally metadata. Bayesian inference resembles maximum likelihood methods as both seek to identify the likelihood that is proportional to observing the data conditional on a tree. However, Bayesian inference includes prior information and uses MCMC sampling algorithms to estimate the posterior probability distribution. Furthermore, these Bayesian methods can also determine virus spread patterns in time and space (phylogeography) [634]. Bayesian inference tools include MrBayes [642], BEAST1 [643], BEAST2 [644], and PhyloBayes [645].

To visualise the resulting phylogeny, FigTree [646], Dendroscope [647], and ggtree [648] can be used in addition to online services such as Evolview [649] and iTOL [650] that also can be used to annotate phylogenetic trees. Additionally, tools like Nextstrain [256] include data curation of a database of viral genomes, a bioinformatics pipeline for phylodynamics analysis using an approximate maximum likelihood approach [651] and an interactive visualisation. Nextstrain has been actively used to visualise the spread and evolution of influenza viruses, SARS-CoV-2 and other viruses.

1.4.3.3.4. Intra-host diversity and quasispecies reconstruction

An approach targeting quasispecies or minority variants is required to accurately determine the quasispecies or minority variants composition within large amounts of different viral variants. However, one of the major challenges is enhancing the specificity and sensitivity. Depending on the NGS technology and applications, error rates are still high ranging from 0.1% to 15% [652]. These errors are introduced due to a number of error-prone reverse-transcription and PCR steps during the (pre-)sequencing steps. Therefore, several bioinformatics tools and pipelines have been developed to study viral variants [653–656], calculate the complexity of a quasispecies, and measure the genetic distance between two similar quasispecies [657]. However, the error corrections made by these variant-calling tools are not optimal in case of amplicon analysis because they are based on the assumption that the error rate is randomly distributed [658]. During bioinformatics analysis, variant calling is one of the key issues. There are several popular variant callers, such as LoFreq [659], Varscan [660], MinVar [653], deepSNV [661], SAMtools, and GATK [662]. These bioinformatic and statistical methods rely on an estimation of the sequence quality to identify true SNV from false positive sequencing errors. Consequently, these approaches solely focus on sequence data and can thus be applied to a wide range of datasets regardless of experimental design. Each of these methods have their advantages and disadvantages, but the investigation of the viral diversity is very sensitive to the used variant calling method. For example, LoFreq uses the Phred of a base and a Poisson-binomial model to estimate the probability that the sequencing error alone is responsible for a putative SNV [659]. It was argued by McCrone [663] et al. that methods like LoFreq and deepSNV can very easily overestimate the viral diversity because additional sources of error on the basis of RT-PCR are not taken into consideration. There are several major experimental methods belonging to another class of variant calling approaches that apply consensus-based error correction to enhance data accuracy as they can identify errors that occur during PCR and sequencing. In the field of virus evolution, three such approaches are used, namely Cirseq, tag-based sequencing (Primer ID), sequence-independent single-primer amplification (SISPA), and intramolecular-ligated nanopore consensus sequencing (INC-Seq). However, Cirseq has poor efficiency and needs a high quantity of vRNA [664] while Primer ID can only tag small regions of a genome and is thus impractical for WGS [665]. SISPA is biased in unpredictable ways, produces uneven coverage [666], and has a much higher error rate than other library preparation methods [667]. Finally, INC-Seq requires high coverage as this approach has a high raw-read error rate of 5%-20% [652, 668].

Individual variants across the genome are detected by variant callers, but variants that are located together in individual genomes are not detected. Viral quasispecies are not a collection of individual variants, but rather a group of interactive genomes. Therefore, it can be valuable to characterise viral quasispecies and identify individual viral haplotypes. Quasispecies reconstruction is, however, computationally challenging due to the short length and error-prone nature of NGS reads. Tools like Shorah [669] and QuRe [670] use read alignments to construct overlapping windows on a genome for local haplotype reconstruction. These results are then collected from all individual windows to reconstruct global haplotypes and estimate their frequencies.

1.5. Implementation of NGS in public health

1.5.1. Potential advantages of the implementation of NGS

NGS generates vast amounts of data that could be of important use to public health. One of the major benefits of NGS is its universal applicability which enables the reuse of the same protocols for various pathogens. Additionally, complete genomic information is provided by NGS in one assay and doesn't require ensuing labour-intensive assays. Moreover, NGS makes it possible to study variations both at consensus and quasispecies level, which enables novel types of analysis. By comparing multiple whole genomes, important genetic differences can be detected [671]. The study of within-host variants has implications for understanding the evolutionary dynamics of viral populations under selection pressures such as host immune response or antiviral drugs. This sequence data could be employed to monitor the presence, spread and evolution of viral diseases. Especially in the long term, a greater understanding of the viruses that cause human disease can be acquired. Consequently, vaccine and antiviral drug development can be supported by targeting viral weaknesses. Furthermore, NGS can also be used to identify mutations and minority variants of vaccine strains and sequences of contaminant viruses in vaccine stock or seeds [672] to perform quality control of live-attenuated viral vaccines [672, 673].

In the context of pandemic management, NGS plays an important role because of the identification and monitoring of viral outbreaks. NGS is the first step in the identification of a novel viral strain [227]. Sequencing data allows the determination of possible origins [219], transmission patterns [224], and it is the first step in the design of primers and probes for RT-PCR-based assays [674]. Although currently PCR-based methods are still preferred to NGS in large-scale testing, this will likely change in the future and become a viable testing method. Further progress in the automation and integration of sample and library preparation will facilitate the widespread adoption of its capacities as it lowers the turnaround time and cost.

1.5.2. Bottlenecks for the implementation of NGS

Despite the many potential advantages of implementing NGS, several bottlenecks still hinder its successful introduction in routine. NGS is a powerful tool, but it is very complex and nuanced and needs significant experience and expertise to produce accurate and high quality data, and analyse the data. During set-up many challenges arise such as choosing the right wet lab sequencing method, the right sequencing platform, personnel with the right skills and experience, and IT and computational infrastructure [675]. The emergence of NGS sequencers and consequently a lot of NGS data led to difficulties in storage capacity and debates about what data should be stored [676].

Furthermore, all DNA is sequenced within a sample, which results in generated data containing primer residues, (host) contaminant DNA, and many low quality reads. One of the greatest challenges in viral respiratory samples, including influenza and SARS-CoV-2, are the other signals in the sample. Without viral amplification, less than 1% of the reads are non-human, and NGS is prone to contamination [677]. Also because of strict regulations regarding human DNA, it is therefore complicated to publicly share data. Additionally, in the context of molecular epidemiological studies and global WGS-based surveillance, issues regarding incomplete clinical data has often been highlighted [678].

NGS is currently not competitive with PCR-based methods because of the complex sample and library preparation, expensive platforms that require expertise to operate and analyse results, and the long turnaround time. Although the field of NGS is rapidly evolving with constant improvement leading to better outputs, and reduced error rates and cost, this makes standardisation for pathogen discovery and routine public health surveillance challenging. While some standards and guidelines have been developed, there is no consensus on the exact analysis steps and many laboratories develop in-house scripts that are often tailored to their own personal research projects [679]. Generally, there are four key steps in analysing NGS data. First, low quality reads are removed from the generated datasets. However, depending on the starting dataset the definition of low quality can change. Secondly, reads are mapped against a reference genome or *de novo* assembled. Thirdly, the mapping quality is assessed and finally a specific research question is addressed. All of these analysis steps use particular randomly chosen cut-off values that can have a significant impact on the downstream analysis.

CHAPTER 2: AIMS AND OBJECTIVES

The primary objective of this thesis is to deliver an appropriate methodology to generate and analyse NGS data generated from clinical samples of the Belgian influenza surveillance network and to explore the added value of the generated NGS sequencing data in comparison to Sanger sequencing data. A second objective has emerged from the ongoing COVID-19 pandemic. We took advantage of the knowledge and approaches developed in the context of influenza surveillance to accelerate the implementation of a NGS approach for SARS-CoV-2 genomic surveillance (Figure 2.1)**.**

Regarding the influenza genomic surveillance, we first evaluated the possible added value of using WGS **to detect the presence of viruses with one or more mutations that are associated with antiviral drug resistance, on the basis of newly generated whole genome influenza sequence data from clinical samples (Chapter 3)**.

Secondly, we assessed how the use of NGS can improve the routine surveillance of circulating influenza strains in humans using a dataset of 253 clinical samples from the NRC Influenza coming from the ILI (mild cases) and SARI (moderate and severe cases) surveillance: (a) by proposing a new way of influenza surveillance using strain classification based on the whole viral genome, which enables the detection of reassortments (Chapter 4) and (b) by tracking mutations across the entire virus genome and trying to associate allelic variations detected in the genome of influenza viruses present in a patient sample with patient data (Chapter 5). Indeed, at the moment the influenza surveillance mainly focuses on the sequence of the HA segment to have an overview of the circulating strains by classifying them according to WHO/ECDC guidelines. However, the strains could lack specific clade-defining amino acid substitutions and/or fail to cluster with reference strains. Additionally, the sequence of the other seven segments and possible reassorted strains are not taken into account. Moreover, some next-generation vaccine candidates are based on influenza virus gene products that are different from HA. Also for this reason, it is important to track the possible genetic diversity across the whole viral genome. In addition, influenza subtypes, such as the A(H3N2) subtype, show a relatively high diversity.

Therefore, the viral genetic background should be taken into account when exploring associations between a mutation and the patient’s clinical data.

Finally, our aim was to propose a **(c) methodology which can be used in routine surveillance to detect low-frequency variants in clinical samples (Chapter 6)**. Some studies indicate that low-frequency variants are interesting to analyse as the complexity of the quasispecies has an effect on the biological behaviour of the virus [684]. However, many challenges remain in order to obtain reliable sequencing results before analysing intra-host diversity can be introduced in routine surveillance. For example, during sample preparation and the WGS sequencing, experimental errors due to amplification can be introduced.

Similar to influenza, NGS can be an added value to the SARS-CoV-2 surveillance. We applied the gained knowledge and approaches that we had developed for the implementation of NGS for influenza surveillance to SARS-CoV-2. Specifically, our aims were to focus on two specific aspects of SARS-CoV-2 surveillance using NGS: (a) the use of available whole genome sequences from the SARS-CoV-2 virus to design reverse transcriptase droplet digital PCR (RT-ddPCR) targets to detect SARS-CoV-2 in patient and wastewater samples (Chapter 7) and (b) the establishment of quality criteria in order to take full advantage of introducing NGS to try to characterise the SARS-CoV-2 and its variants in wastewater (Chapter 8).

The current SARS-CoV-2 surveillance is mainly focused on monitoring the virus spread using PCR-based and whole-genome sequencing methods on clinical samples. However, as the number of new variants is rapidly increasing around the world, it is possible that PCR-based methods that were validated a few months earlier are not suitable anymore due to genetic modification appearing in the annealing site of the primers and probes of the proposed methods. Therefore, even by testing a large collection of samples, laboratories are not able to test a representative collection of samples that deals with this diversity that is continuously evolving and that needs to be seen not only locally but worldwide. Consequently, these primers and probes need to be regularly evaluated. This can be done by an *in silico* evaluation using a large set of whole genome data with sequences obtained around the world.

To estimate the prevalence of SARS-CoV-2 and assess its genetic diversity and geographical distribution, wastewater surveillance has been proposed as an independent and complementary alternative to the collection of data from individual testing for epidemiological surveillance [413–415]. Wastewater samples may include a collection of multiple strains, where the dominant strain corresponds with the most prevalent strain circulating in a community and where less prevalent strains are present as subpopulations. However, few quality criteria are in place for wastewater sequencing and they mostly only apply to constructing the consensus genome sequence. Furthermore, the establishment of quality criteria for such cases is important to take the experimental errors into account.

The supplementary material was not bundled with this dissertation because of its considerable length. It is available online (https://doi.org/10.6084/m9.figshare.21215288).

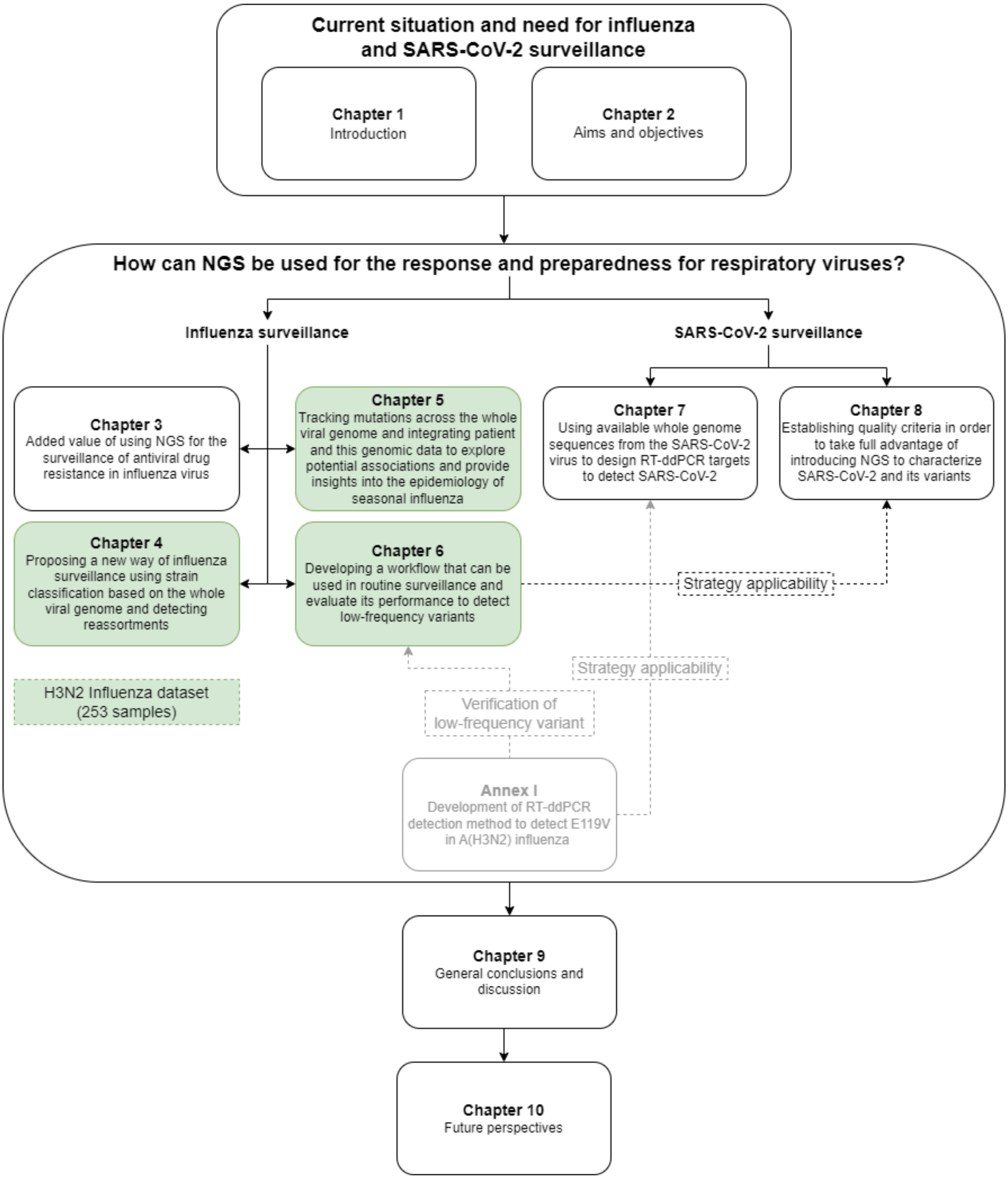

Current situation and need for influenza and SARS-CoV-2 surveillance
Chapter 1
Introduction
Chapter 2
Aims and objectives
How can NGS be used for the response and preparedness for respiratory viruses?
Influenza surveillance
SARS-CoV-2 surveillance
Chapter 3
Added value of using NGS for the surveillance of antiviral drug resistance in influenza virus
Chapter 5
Tracking mutations across the whole viral genome and integrating patient and this genomic data to explore potential associations and provide insights into the epidemiology of seasonal influenza
Chapter 7
Using available whole genome sequences from the SARS-CoV-2 virus to design RT-ddPCR targets to detect SARS-CoV-2
Chapter 8
Establishing quality criteria in order to take full advantage of introducing NGS to characterize SARS-CoV-2 and its variants
Chapter 4
Proposing a new way of influenza surveillance using strain classification based on the whole viral genome and detecting reassortments
Chapter 6
Developing a workflow that can be used in routine surveillance and evaluate its performance to detect low-frequency variants
Strategy applicability
Strategy applicability
H3N2 Influenza dataset
(253 samples)
Verification of low-frequency variant
Annex I
Development of RT-ddPCR detection method to detect E119V in A(H3N2) influenza
Chapter 9
General conclusions and discussion
Chapter 10
Future perspectives

Figure 2.1: Schematic outline of the thesis. In this diagram the different chapters and their link to either the influenza or SARS-CoV-2 surveillance is introduced in the black rectangles. The dashed lines show the links between the chapters. The strategy that was used in the draft version of “Development of RT-ddPCR detection method to detect E119V in A(H3N2) influenza has been used for also used in Chapter 7. Moreover, the method from this paper was used to verify the presence and the abundance of the low-frequency variants in Chapter 6. The workflow that was used in Chapter 6 was after some adaptations also used for the SARS-CoV-2 NGS data in Chapter 8. Finally, the same influenza dataset was used on which various analysis were applied and they were discussed in Chapter 4, 5 and 6 (boxes in green).

CHAPTER 3: ADDED VALUE OF USING NGS FOR THE SURVEILLANCE OF ANTIVIRAL DRUG RESISTANCE IN INFLUENZA VIRUS

Context of this chapter:

Monitoring the antiviral drug susceptibility of influenza virus has long focused on the NA protein, because NA inhibitors have historically been the most extensively used anti-influenza drugs. However, new antivirals targeting gene products from other segments such as baloxavir marboxil, a PA inhibitor, are becoming available for influenza treatment. This increases the need to obtain information about the whole influenza virus genome. This chapter evaluates the added value of implementing whole-genome sequencing in the context of routine surveillance of antiviral resistance.

This chapter was previously published:

Laura A. E. Van Poelvoorde, Xavier Saelens, Isabelle Thomas, Nancy H. Roosens. (2020) Next-Generation Sequencing: An Eye-Opener for the Surveillance of Antiviral Resistance in Influenza. *Trends in Biotechnology*. **38(4)**. https://doi.org/10.1016/j.tibtech.2019.09.009

Author contributions:

Conceptualization: NR, IT, XS; Project administration: NR; Investigation: LVP; Visualisation: LVP; Writing – original draft preparation: LVP; Writing – review and editing: all authors; Funding acquisition: NR; Supervision: NR, IT

Abstract:

Next-generation sequencing can enable a more effective response to a wide range of communicable disease threats, such as influenza, which is one of the leading causes of human morbidity and mortality worldwide. After vaccination, antivirals are the second line of defence against influenza. The use of currently available antivirals can lead to antiviral resistance mutations in the entire influenza genome. Therefore, the methods to detect these mutations should be developed and implemented. In this opinion, we assess how next-generation sequencing could be implemented to detect drug resistance mutations in clinical influenza virus isolates.

3.1. What Is the Importance of Genetic Influenza Surveillance with Regard to Antiviral Resistance?

One of the major technological evolutions in life sciences of the past decade is next-generation sequencing (NGS) (NGS; see Glossary). This technology could enable a more effective response to a wide range of communicable disease threats, such as influenza viruses. A major advantage of whole-genome sequencing (WGS) is that this one technique can provide broad and detailed data on the identity of pathogens that previously required multiple laboratory phenotypic and genotypic assays. Moreover, WGS provides higher resolution information than conventional genotyping tests. RT-qPCR and Sanger sequencing, for example, are often targeted to a limited fraction of the genome, meaning that crucial information can be missed. The additional genotypic information that WGS can provide, could be critical when tracking the origin of outbreaks and to forecast the spread of disease. Influenza A and B viruses are a major cause of respiratory tract infections in humans, resulting in significant morbidity and mortality [20]. The genome of these viruses consists of eight segments of single-stranded RNA of negative polarity. Influenza viruses can mutate rapidly when subjected to selection pressures, such as immunity induced by prior or vaccination [685] or antiviral drug use [686]. The hemagglutinin (HA) and neuraminidase (NA) proteins are often considered to be the most important viral antigens, because they are major targets of the immune system. HA and NA are used as targets for current anti-influenza strategies: HA is the prime target of the currently licensed influenza vaccines, and NA inhibition is the main mechanism of the most frequently used influenza antivirals. Although vaccines are considered the best way to prevent influenza, the limited use and their generally poor effectiveness in the elderly [687] (https://www.cdc.gov/flu/about/qa/vaccineeffect.htm) imply that efficient antiviral drugs are needed as a complementary or alternative line of defence.

Monitoring and detecting mutations in the influenza virus genome, especially those that confer antiviral resistance, is of paramount importance to public health surveillance (https://www.who.int/influenza/gisrs_laboratory/antiviral_susceptibility/nai_genotyping_molecular/en/). As the use of antiviral drugs continues to grow, more cases of drug-resistant viruses are expected to occur. Because of this, and the fact that a limited number of anti-influenza drugs with different mechanisms are currently available, it is important to assess whether the use of NGS could add value for the preparedness and response to the emergence of antiviral resistance. The determination of full influenza genomes, which is possible by more extensive use of NGS, will allow for a better understanding of the genetic determinants of viral resistance, and may enable the detection of minority drug resistant viral populations.

In this Opinion, we provide an overview of the potential antiviral drug resistance mutations in influenza virus genome, based on a thorough review of the literature on antiviral drugs that target different stages of the viral life cycle (Figure 3.1).

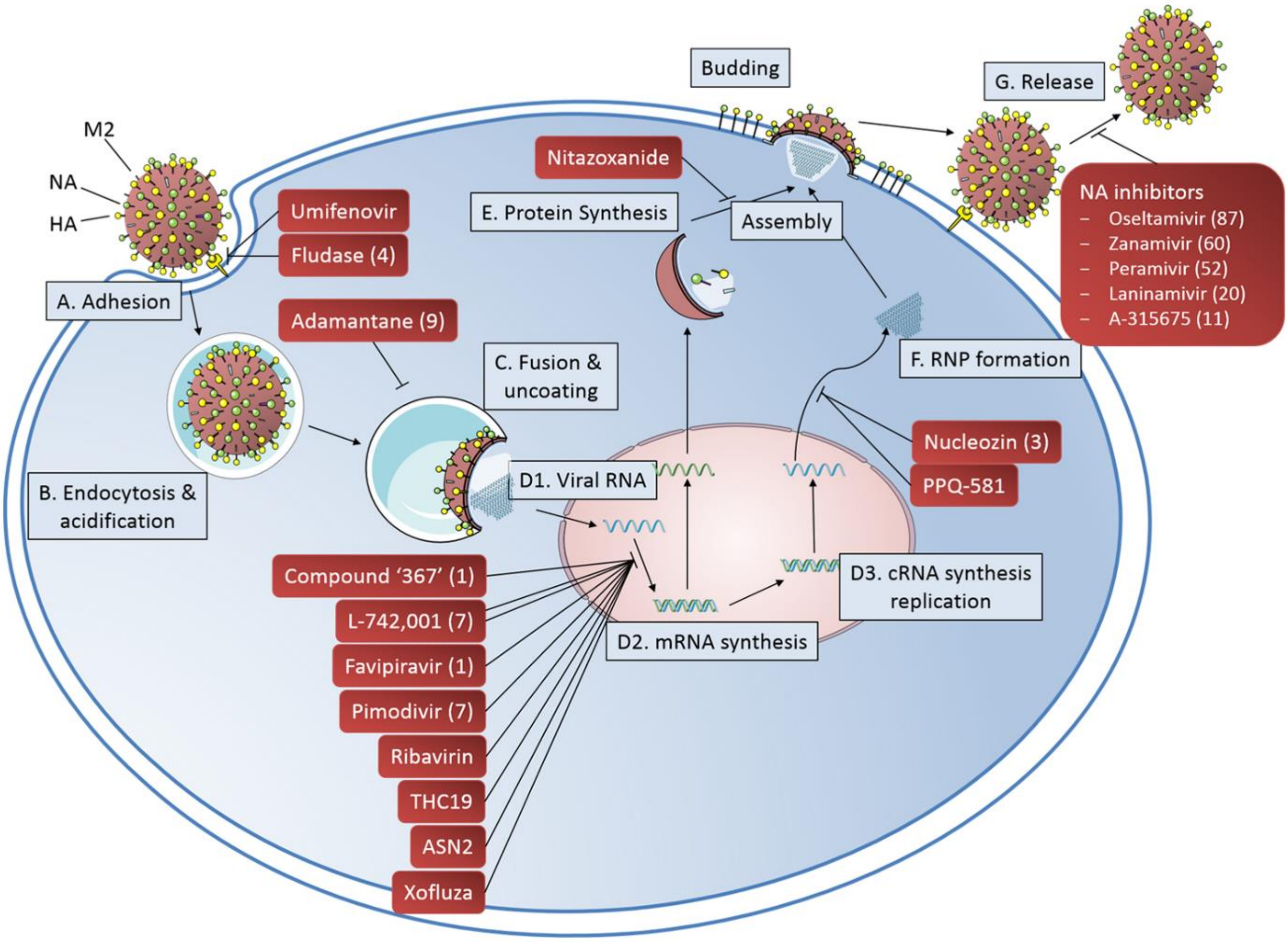


Figure 3.1: Antiviral drugs against influenza virus and their target sites in the virus cycle. The different stages of the viral cycle are the binding of HA to the sialic-acid containing host receptor (A), followed by the endocytosis and the acidification of the HA protein (B). This acidification in the early endosomes leads to fusion of the viral and endosomal membranes, and triggers the influx of H^+ ions through the M2 channel that results in the dissociation of the vRNPs and uncoating (C). After transport of the vRNPs to the nucleus, viral mRNA synthesis is initiated by the viral polymerase. The latter is also responsible for the unprimed replication of the vRNA through a cRNA

intermediate (D). The viral mRNAs are exported to the cytoplasm and translated into viral proteins. In the ER, the surface proteins HA, M2 and NA are processed, glycosylated and transported to the cell membrane (E). The newly synthesised vRNPs are transported to the cytoplasm, mediated by a M1-NS2 complex that is bound to the vRNP (F). At virion assembly and budding site, the newly produced vRNPs are incorporated into new viruses. Finally, the NA cleaves these sialic acid residues and virions are released from the host cell (G). The figure illustrates the different antiviral drugs at the position where they interfere with the viral cycle. Between brackets the number of mutations that are known to be related to antiviral resistance are indicated to the different antiviral drugs.

3.2. Discovering Mutations That Confer Antiviral Resistance

Different methods can be used to identify the appearance of antiviral drug resistance. Screening viruses in clinical samples for antiviral resistance using a specific phenotypic assay is the first method that can be used, but this implies that the antiviral drug is already commercially available and in use. Phenotypic assays can evaluate the production of virus particles in the presence of the antiviral drug in comparison to mock-treated conditions (total amount of virus using for example ELISA-based methods, or infectious particles using for example plaque reduction assays), or the activity of the enzyme/protein targeted by the antiviral (NA-inhibition assay for example).

Other approaches, such as serial passages in cell culture or in animal models [688, 689], are useful to evaluate new compounds and potential appearance of resistance. In both cases, the associated mutations can then be identified by sequencing the viral genome. Structural analyses can also be used to determine amino-acid substitutions that would likely confer resistance. Using site-directed mutagenesis and reverse genetics systems, these theoretical mutations can be tested to confirm the resistance in phenotypic assays [690].

More than 200 mutations in the influenza virus genome reportedly confer antiviral drug resistance. For each mutation (Supplementary File S3.1), the mechanism that explains the conferred resistance (if known) and the origin of that used to identify these mutations (e.g., clinical sample, reverse genetics, or serial passaging) are provided. Oseltamivir and zanamivir, two NA-activity inhibitors, are currently the most commonly used antiviral drugs; it is thus not surprising that the majority of the mutations were found in the NA segment (Supplementary File S3.2). Mutations detected in patients and reported in more than five papers that induce resistance against commercially available antivirals or antivirals in clinical trials were included in a table of antivirals in Supplementary File S3.3.

3.3. Detection of Antiviral Resistance Mutations: From Classical Surveillance to the Implementation of NGS

Once a mutation has been identified that confers resistance to an antiviral drug, it is important to be able to rapidly detect and follow its possible emergence in circulating strains. Although phenotypic tests remain the only way to confirm the resistance of the virus, genotypic assays are the most commonly used in clinical samples because they provide a rapid method to detect known mutations and eliminate the need for virus isolation and propagation in cell cultures (Table 3.1) [691–695]. In addition, if loaded into an in-house bioinformatics pipeline or another web-based application, these known mutations can be rapidly identified.

Genotyping by qRT-PCR is commonly used as a fast and relatively inexpensive method that can be performed directly on clinical specimens [691, 696]. However, this approach can only detect one targeted mutation, as it relies on limited differences in the genome. qRT-PCR can thus be difficult to develop and it has limited benefits in the context of surveillance [691, 697, 698].

Sanger sequencing is still used as a standard reference method for routine genetic surveillance of influenza viruses. However, the viral RNA genome must first be extracted, and the genomic segment of interest must be amplified by RT-PCR. Each Sanger sequencing reaction is based on one single primer pair and provides a sequence of 400–1000 bases in length [699, 700]. To obtain the sequence of an entire segment or even the whole influenza genome (approximately 14 kilobases), multiple primers must be used in parallel reactions [701]. Sanger sequencing thus becomes labour intensive and can be expensive if the whole genome sequence of a large set of samples is required [691, 698, 702]. Sanger-based technologies are also not the most suitable methodologies to detect polymorphisms or minority variants with a frequency lower than 20% [703–705].

The beginning of the 21st century has seen a gradual shift from Sanger sequencing towards newer, NGS (also known as second-generation sequencing) methods that allow a higher throughput at a relatively low cost [700, 706]. For these technologies, targeted RT-PCR might still be necessary to amplify the influenza virus genome, especially to overcome the otherwise over-represented host sequences in the sample. In 2005, Bright and colleagues were the first to use a pyrosequencing platform to monitor the emergence and spread of adamantane resistance in circulating A(H3N2) influenza virus strains over a 10-year period [707]. Pyrosequencing has since been successfully used to detect already known mutations in NA and M2, responsible for antiviral resistance. However, different sets of primers are required to detect these mutations [503].

NGS platforms remain the best choice in terms of value for money for high-throughput sequencing. NGS can be accomplished with several methodologies, namely sequencing-by-ligation (SOLiD technology), sequencing-by-hybridization (resequencing micro-array), and sequencing-by-synthesis (Illumina, Ion Torrent; now the dominant nucleic acid technology) [700, 708, 709]. NGS methods generate massive numbers of reads that are 75–700 bases in length but with a high coverage per base. Such parallel, deep sequencing allows for the reconstruction of the entire genome in a sample. NGS also provides the ability to detect quasispecies with a frequency below 20%. Van den Hoecke and colleagues arbitrarily proposed the use a threshold of 0.5% for Illumina MiSeq and Ion Torrent, below which it becomes too difficult to distinguish real mutations from background errors cumulatively introduced by the RT-PCR and the sequencing technology itself [710].

However, NGS methods come with several limitations: the short-read length requires powerful bioinformatic algorithms to assemble a consensus sequence, and the required amplification by RT-PCR may introduce biases. Third-generation sequencers may address these problems, by constructing longer reads and possibly eliminating the requirement for amplification of the virus [590]. However, the low abundance of virus in most clinical respiratory samples and the relative novelty of these approaches currently still makes it difficult to eliminate RT-PCR amplification. The two main approaches for third-generation sequencing [711] are the synthetic approach that relies on existing short-reads technologies to construct long reads [708, 712] and the single-molecule-real-time sequencing approach (SMRT) [584]. The SMRT approach is the most widely used and is represented by sequencing technologies developed by Pacific Biosciences (PacBio) and Oxford Nanopore [713–715]. PacBio is able to generate long reads, which enables the analysis of difficult regions with multiple repeats. In addition, the real-time acquisition of the signal means that there is no lag between each nucleotide addition. However, the PacBio flow cell is not as high-throughput as the Illumina platform and the error rate is still relatively high in comparison to the Illumina platform [708, 711, 716–719]. So far, only a limited number of publications have reported on the use of PacBio on clinical human influenza virus samples or on influenza viruses generated with reverse genetics [719–721]. The MinION from Oxford Nanopore is a small device that has a relatively low cost and can provide very long reads. Currently, one MinION study has used direct RNA sequencing of the influenza virus genome, without prior amplification by RT-PCR, but this was only feasible because of a very high viral load, which is rarely found in clinical samples [722–726]. The elimination of RT-PCR amplification may lead to a lower cost and a shorter execution time. MinION reads are still characterised by a lower quality with high error rates, and therefore a high depth coverage is necessary to detect antiviral resistance mutations with confidence [725].

Table 3.1: Comparison of different genotypic assays [692–695]

	Benefits	Challenges	NGS platform	Time (h)	Read length (bases)	Raw error rate (%)	Cost per Gb
qRT-PCR	• Equipment often already present in labs • High sensitivity • Quick & simple workflow	• Limited set of variants • No identification of novel variants • Low scalability • Low variant resolution	**/**	1 – 2	/	/	/
Sanger sequencing	• Cost-effective for small stretches of DNA • Quick & simple workflow	• Low sensitivity • Low variant discovery power • Low scalability	**/**	24	400 - 1000	0.001	US$ 10 000 000
NGS	• Identification of novel variants • Expanded discovery power • Higher analytical sensitivity • Great resolution • Higher sample throughput with multiplexing	• Less cost-effective for sequencing low numbers of samples • Time-consuming for sequencing low numbers of targets • Requires a dedicated data-handling workflow	**Illumina**	27 – 144	36, 75, 100, 150, 250, 300	0.1-1	US$ 7 – 2000
			Ion Torrent	2 – 7.5	200-400	1-2	US$ 80 – 2000
			PacBio	0.5 – 60	10 000 – 20 000	14 – 15	US$ 600 – 1000
			Oxford Nanopore	< 48	< 200 000	5 – 40	US$ 100 – 400

3.4. Cases When NGS Could Have Added Value for Detection of Drug-Resistance Mutations

High-throughput molecular approaches offer new possibilities for influenza virus surveillance. By determining the whole-genome of the influenza virus, higher resolution evolutionary patterns can be revealed, knowledge of reassortment events and emerging mutations across all genes can be provided and information on intrahost diversity of the virus (quasispecies) can be obtained. This information can lead to a better understanding of genetic changes in all segments for various seasons, tropism markers, antigenic characteristics, virulence, reassortment events, and of course antiviral resistance [720, 722, 727–731].

3.4.1. Monitoring and Surveillance of Resistance to New Antiviral Drugs

Currently, adamantanes and NA inhibitors, which includes oseltamivir and zanamivir, or a combination of antivirals, are the only antiviral drugs for influenza viruses licensed in Europe. Adamantanes, however, are not used anymore due to the presence of resistance mutations in almost all currently circulating influenza strains. As long as only NA inhibitors are used, whole-genome information may not be required, as Sanger sequencing of the NA segment is probably sufficient for the surveillance of antiviral emergence.

Although NA inhibitors are licensed to treat uncomplicated influenza infection, they are, in most European countries, only used to treat patients at risk of developing more serious complications, such as the elderly or people with underlying conditions (https://ecdc.europa.eu/en/seasonal-influenza/prevention-and-control/antivirals/faq).
However, the need to carefully monitor seasonal influenza virus susceptibility to NA inhibitors remains a priority for public health agencies. At present, the percentage of detected circulating NA inhibitor-resistant viruses is low, but as seen in the 2007-2008 influenza season in Europe for the seasonal A(H1N1) strain, there can be a sporadic emergence of resistance that spreads rapidly in the population; 14% of the A(H1N1) virus samples from that season were resistant to oseltamivir [199]. By the 2008-2009 season the number of resistant influenza A(H1N1) strains had increased to 98% [200]. This raised concerns until the extinction of this A(H1N1) subtype following the emergence of the susceptible A(H1N1pdm09) pandemic virus in 2009. Some mutations that confer antiviral resistance cause a decrease in fitness of the resistant viruses. However, several studies have shown that compensatory mutations may co-emerge and improve the fitness of the resistant viruses. These mutations can also arise during cell culture, which led to the CDC’s Influenza Division and WHO Collaborative Centers to shift to a sequencing first approach using NGS before performing isolation and phenotypic characterisation on a subset of samples [631].

As antivirals directed against other influenza virus proteins gradually become approved and used in the population, the need to monitor possible resistance mutations in other parts of the viral genome becomes more important (Supplementary File S3.1 and Supplementary File S3.3). For example, Xofluza (baloxavir marboxil) was recently approved in the US and Japan. This drug is a cap-dependent endonuclease inhibitor of the viral polymerase acidic (PA), for which a resistance mutation in the PA segment has already been found in clinical samples. Favipiravir, which targets the viral RNA-dependent RNA polymerase (RdRP), is available in Japan for patients infected by an influenza virus that is resistant to other available influenza drugs and it is in the third phase of clinical trials in the US and Europe. Recently, a substitution in the PB1 segment, namely K229R, was reported in an *in vitro* study to confer resistance to favipiravir. This substitution was accompanied with a PA P653L substitution, which restores the fitness of the virus [732]. In the Russian Federation and China, umifenovir (Arbidol) is used to treat influenza. This broad-spectrum antiviral can prevent virus entry into the host cell and is believed to target the HA protein. Other antivirals, such as nitazoxanide, an HA maturation inhibitor and pimodivir, a PB2 inhibitor, are being evaluated in phase III clinical trials. No specific, easy, standardised phenotypic tests have been developed yet to monitor the susceptibility of influenza viruses to these new antiviral drugs, making sequencing almost indispensable. WGS is of interest since most of these new drugs target viral proteins involved in the replication pathways where multiple viral proteins usually cooperate. In a few *in vitro* studies resistance substitutions appeared in other proteins rather than in the target protein of the antiviral drug (Supplementary File S3.2) [733].

3.4.2. Surveillance of Emergence of Resistance Mutations in Quasispecies

With traditional sequencing approaches, it is difficult to detect and quantify minority genomes present in viral quasispecies. NGS provides, for each patient, the possibility to investigate previously inaccessible aspects of viral dynamics [734]. The challenge in characterising quasispecies composition remains to define a cut-off between real mutations and false positives.

Influenza virus quasispecies analysis in the context of antiviral resistance has already been performed in clinical samples in a few studies. Trebbien and collaborators [735] followed for 6 months an immunocompromised patient treated with oseltamivir and zanamivir. They concluded that NGS was necessary to properly investigate the complex population at the sites that are considered important for antiviral resistance. Similarly, Pichon and colleagues [736] followed a child with severe combined immunodeficiency. The authors concluded that NGS allowed for the characterisation of viral variant evolution and that the quasispecies analysis could reveal a risk of decreased antiviral efficacy. These study cases clearly indicate that the characterisation of the quasispecies composition of influenza virus genome could reveal the emergence of antiviral resistance. NGS technologies provide the necessary tools to detect the appearance and emergence of resistance mutations as quasispecies, either by studying samples from a patient under treatment or by analysing a large set of circulating viruses, and thus to identify these mutations before they reach a proportion where they can affect the antiviral susceptibility phenotype or before they become dominant.

3.5. Concluding Remarks and Future Perspectives

Monitoring the antiviral drug susceptibility of influenza virus has long focused on the NA protein, because NA inhibitors have historically been the most extensively used anti-influenza drugs. However, as new antivirals that target different viral proteins become available and used to treat influenza patients, the need to obtain information about the whole genome increases. Therefore, NGS represents a more informative approach than Sanger sequencing in the context of routine surveillance.

This routine surveillance using WGS remains challenging regarding the complexity of data analysis for non-bioinformatics experts. Although, there are many tools available to analyse NGS results, many of these require substantial bioinformatic expertise because they are only available using the command line on Linux. Therefore, the spread of more web-based platforms with a user-friendly interface within the scientific community would be an advance in the use of WGS for non-bioinformatic experts [737].

The development of these NGS methods has provided an opportunity to obtain information about all the genomic segments and about the minority genomes present in viral quasispecies. Investigation of this quasispecies nature of influenza viruses thus improves preparedness by potentially forecasting the emergence of resistance substitutions.

However, despite intensive research on influenza viruses, little is still known about the role, dynamics, and spread of viral quasispecies. More work is needed to understand whether and how the quasispecies nature of influenza viruses plays a role in antiviral escape and in immune selection pressure, or whether this could be a contributing factor to disease severity. The use of WGS in routine surveillance will enable a better understanding of the association of the viral quasispecies and the host characteristics of a patient. It could also enable a quicker response when certain mutations that confer antiviral resistance are emerging within the patient. The interpretation of this genomic data is of course highly dependent on how complete and structured the epidemiological and clinical metadata is.

Use of NGS technology also comes with limitations including the cost, the requirement to amplify the genome, which introduces PCR errors, and the short-read lengths that require powerful bioinformatic tools to assemble a consensus sequence. Soon third-generation sequencers may become available and address some of these limitations by simplifying data analysis and lowering costs (see Outstanding Questions).

Acknowledgments

This research was supported by the .Be READY project financed by Sciensano. We would like to thank Cyril Barbezange for support and critically reading the manuscript. We would also like to thank Lauren Tindale for proofreading the manuscript.

CHAPTER 4: NEW INFLUENZA CLASSIFICATION BASED ON THE WHOLE VIRAL GENOME

Context of this chapter:

Currently, the hemagglutinin (HA) segment remains the principal target region for Sanger sequencing in classical influenza surveillance programmes. In this chapter and Chapter 5 and 6, we illustrate the feasibility and benefit of switching towards whole-genome-sequencing-based surveillance. This chapter reports on a proof of concept that proposes an improved method for classification of circulating human influenza A virus strains with a high resolution for genetic characterisation and reassortment detection. This also allows a better insight into virus spread and improved detection of transmission clusters that would not have been possible when solely sequencing the HA segments. Additionally, whole-genome information, i.e. from all eight segments, will improve current vaccine strain selection, and could also potentially become a requirement in the future as next-generation vaccines and antiviral drugs that do not focus solely on HA and NA, respectively, become licensed. Lastly, integration of whole-genome data with patient information also allows investigation of associations of genetic groups based on the whole genome with host characteristics.

This chapter was previously published:

Laura A. E. Van Poelvoorde, Bert Bogaerts, Qiang Fu, Sigrid C.J. De Keersmaecker, Isabelle Thomas, Nina Van Goethem, Steven Van Gucht, Raf Winand, Xavier Saelens, Nancy Roosens¶, Kevin Vanneste¶. (2021) Whole-genome-based phylogenomic analysis of the Belgian 2016–2017 influenza A(H3N2) outbreak season allows improved surveillance. *Microbial Genomics*. **7(9)**. https://doi.org/10.1099/mgen.0.000643

Authors' contributions:

Conceptualization: NR, KV, IT, XS, SDK, SVG; Project administration: NR; Data curation: IT, LVP, NVG; Formal analysis: LVP, BB, KV, RW, QF; Investigation: LVP; Visualisation: LVP; Writing – original draft preparation: LVP, KV; Writing – review and editing: all authors; Funding acquisition: NR; Supervision: NR, KV, IT

Abstract:

Seasonal influenza epidemics are associated with high mortality and morbidity in the human population. Influenza surveillance is critical for providing information to national influenza programmes and for making vaccine composition predictions. Vaccination prevents viral infections, but rapid influenza evolution results in emerging mutants that differ antigenically from vaccine strains. Current influenza surveillance relies on Sanger sequencing of the hemagglutinin (HA) gene. Its classification according to World Health Organization (WHO) and European Centre for Disease Prevention and Control (ECDC) guidelines is based on combining certain genotypic amino acid mutations and phylogenetic analysis. Next-generation sequencing technologies enable a shift to whole-genome sequencing (WGS) for influenza surveillance, but this requires laboratory workflow adaptations and advanced bioinformatics workflows. In this study, 253 influenza A(H3N2) positive clinical specimens from the 2016-2017 Belgian season underwent WGS using the Illumina MiSeq system. HA-based classification according to WHO/ECDC guidelines did not allow classification of all samples. A new approach, considering the whole genome, was investigated based on using powerful phylogenomic tools including beast and Nextstrain, which substantially improved phylogenetic classification. Moreover, Bayesian inference via beast facilitated reassortment detection by both manual inspection and computational methods, detecting intra-subtype reassortants at an estimated rate of 15 %. Real-time analysis (i.e., as an outbreak is ongoing) via Nextstrain allowed positioning the Belgian isolates into the globally circulating context. Finally, integration of patient data with phylogenetic groups and reassortment status allowed detection of several associations that would have been missed when solely considering HA, such as hospitalised patients being more likely to be infected with A(H3N2) reassortants, and the possibility to link several phylogenetic groups to disease severity indicators could be relevant for epidemiological monitoring. Our study demonstrates that WGS offers multiple advantages for influenza monitoring in (inter)national influenza surveillance, and proposes an improved methodology. This allows leveraging all information contained in influenza genomes, and allows for more accurate genetic characterisation and reassortment detection.

4.1. Introduction

Every year, 5-20% of the human population becomes infected with influenza. Worldwide, 3 to 5 million infections yearly progress into severe cases [8]. Case-fatality rates are <0.1 % during a typical influenza pandemic [7]. Severe cases predominate in certain risk groups, including the very young and very old, and patients with comorbidities such as chronic cardiac, respiratory and metabolic diseases, obese patients, immunocompromised patients and pregnant women [738]. Influenza A and B viruses are a major cause of respiratory tract infections in humans. The influenza A genome consists of eight segments, including two segments encoding hemagglutinin (HA) and neuraminidase (NA) proteins. HA and NA are considered the most important viral components because they represent key antigens due to their location on the viral envelope, rendering them the main immune response targets [739–741]. Influenza A viruses are further classified into subtypes based on the combination of their HA and NA segments. Currently, influenza A(H1N1)pdm09 and A(H3N2) are the two main influenza A subtypes circulating in humans [8].

Influenza surveillance is important to determine the vaccine composition based on circulating influenza virus strains, and to provide information to national influenza prevention and control programmes regarding the timing, impact and severity of seasonal epidemics. Additionally, surveillance allows the detection of emerging zoonotic and potentially pandemic influenza viruses [742]. According to the guidelines of the World Health Organization (WHO) and the European Centre for Disease Prevention and Control (ECDC), influenza surveillance requires classification of new samples into different clades and subclades within each influenza subtype. This surveillance is based on the combination of certain predefined genotypic amino acid variants present in the HA segment, and phylogenetic analysis for which the HA segment of samples should cluster within clades represented by indicated vaccine or reference virus strains. Additionally, the HA gene should exhibit neither many (although an exact threshold is not defined) nor critical (i.e., those that significantly affect antigenicity) amino acid differences compared to the indicated vaccine or reference strain with which they associate [743]. Based on circulating strains identified in surveillance programmes, respective HA-based clade classification, and availability of vaccine viruses, the vaccine composition for the following season is determined.

Consequently, the main focus of genetic surveillance is the HA gene, with proportionally limited data available for the other seven segments [744]. Sanger sequencing currently remains the principal approach for genetic influenza surveillance. However, next-generation sequencing (NGS) technologies are increasingly used in many countries [728, 745–747] and constitute a promising alternative, offering the possibility to simultaneously obtain the sequence of all eight segments. Whole-genome sequences can provide much additional information for influenza surveillance compared to solely sequencing the HA gene. These whole-genome sequences can assist in inferring potential links between genomic data and host characteristics, including epidemiological effect exploration of inter- and intra-seasonal evolutionary dynamics, inter- and intra-subtype reassortment detection, identification of mutations located anywhere in the genome [731, 748, 749], and genetic group strain classification based on whole-genome information. Additionally, whole-genome sequences can improve vaccine strain selection and enhance vaccine efficacy [750]. Lastly, multiple next-generation vaccines target other segments than the HA segment [751], requiring new approaches for influenza monitoring based on the whole genome. Most of these sequences are deposited in the database maintained by the Global Initiative on Sharing Avian Influenza Data (GISAID), which contains genome sequences of all influenza types and includes outbreaks and surveillance studies [752].

Since the 1968 influenza A(H3N2) pandemic, the A(H3N2) subtype has led to numerous seasonal epidemics and is considered to evolve faster than other subtypes [753]. A(H3N2) has shown extensive genetic diversity and increased morbidity and mortality in recent years, especially in the elderly [754]. Identifying and predicting current epidemiological threats from A(H3N2) is challenging due to the strain’s rapid evolution and the current limitations of existing methods for analysing the antigenic characteristics of influenza A(H3N2) [755]. This rapid evolution is caused by mutations but also reassortments. Reassortments can occur due to the segmented genome when cells are infected with different influenza viruses, when new virus particles are assembled with a mix of segments from these different viruses [756]. This rapid evolution is reflected by the continuous updates of WHO recommendations regarding vaccine strains for influenza A(H3N2) [757]. Sporadically, inter-subtype reassortment occurs, which may give rise to viruses with pandemic potential. Two recent examples of inter-subtype reassortment include a case in the Netherlands in March 2018 and a case in Sweden in January 2019, both of which resulted in Influenza subtype A(H1N2) [758]. Nevertheless, inter-subtype reassortments are rare due to potential segment incompatibility between heterologous viral components resulting in RNA or protein mismatches that decrease viral fitness and limit dispersion in the human population [759]. In contrast, intra-subtype reassortments between various lineages of the same subtype happen more frequently because of the higher genetic relatedness and functional compatibility of their segments. Intra-subtype reassortments can increase the adaptive potential and genetic diversity of circulating viruses [748]. For the A(H3N2) subtype, it was suggested that the rate of adaptive amino acid replacements within reassorted strains is temporarily increased. Although intra-subtype reassortments have not been systematically evaluated, they had a major impact on virus evolution [748].

As the role of intra-subtype reassortment is becoming increasingly clear, identifying and monitoring reassortments with whole-genome-based surveillance becomes necessary [760, 761]. However, implementing this whole-genome-based surveillance in routine surveillance requires multiple adaptations in the laboratory workflow. Additionally, in-depth bioinformatics expertise is required to process sequencing results. Hence, the availability of user-friendly tools and pipelines is paramount for the incorporation of whole-genome sequencing (WGS) into routine surveillance [762]. In this study, the feasibility of WGS is evaluated for routine influenza surveillance based on the WGS of 253 A(H3N2) samples from the 2016-2017 Belgian influenza season [763]. In particular, the suitability of current methods for defining phylogenetic groups based on the HA segment versus their whole genome is evaluated to assess the added value of incorporating whole-genome information for interpretation of strain clusters and their reassortment status. Several high-end computational methods are explored to improve classification and detection of reassortments, and this study proposes a new methodology based on WGS for genetic influenza surveillance.

4.2. Methods

4.2.1. Sample selection, RNA isolation, PCR amplification and WGS

Sample selection. Two main surveillance systems exist in Belgium, 'influenza-like-illness' (ILI) and 'severe-acute-respiratory-infection' (SARI). A sudden onset of symptoms, including fever and respiratory and systemic symptoms, define ILI cases. A SARI case is an acute respiratory illness requiring hospitalisation with fever and respiratory symptoms onset within the previous 10 days. A standard questionnaire accompanied all samples with patient information on sex, birth date, clinical features, vaccination status, administration of antiviral treatment or antibiotics, date of symptom onset and date of sample collection (Supplementary File S4.1).

From these two surveillance systems, 253 samples were selected with a Cq <32 as detected with quantitative reverse transcription PCR (RT-qPCR). Samples were mainly selected by stratifying based on the severity, patient age and sampling date. Samples from outpatients (ILI) were all categorized as mild cases (n = 93), whereas samples from hospitalised patients (SARI) were categorized as either moderate (n = 122) or severe (n = 38) cases. Patient age and sampling dates were stratified into three groups: patients <15 years, patients between 15 and 59 years, and patients ≥60 years. Samples were categorized based on their sampling history: before, during (week 4 to 6 2017), and after the epidemic peak. An overview of available host characteristics for all 253 samples is provided in Table 4.1.

Table 4.1: Samples stratified according to host characteristics.

Age (years):	**<15**	**15 – 59**	**≥60**
Beginning of epidemic (<week 4)	12	17	35
Peak of epidemic (week 4 - 6)	16	26	86
End of epidemic (>week 6)	11	16	34

ILI	93	**SARI**	160
Male*	122	**Female***	122
Vaccinated*	52	**Not vaccinated***	130
Antibiotics administered*	100	**No antibiotics administered***	126
Respiratory disease*	50	**No respiratory disease***	199
Cardiac disease*	54	**No cardiac disease***	195
Obesity	20	**No obesity**	233
Renal insufficiency	35	**No renal insufficiency**	218
Hepatic insufficiency	6	**No hepatic insufficiency**	247
Diabetes	27	**No Diabetes**	226
Immuno-deficiency	23	**No immuno-deficiency**	230
Neuromuscular disease	21	**No neuromuscular disease**	232
Stay in ICU	22	**No stay in ICU**	231
Resulting in death	19	**Not resulting in death**	234

**Samples for which certain host characteristics were unknown, were excluded for analysing that particular host characteristic.*

RNA isolation, PCR amplification and WGS. Nucleic acids of samples were extracted directly from the clinical specimens [764] using a viral RNA/DNA isolation kit (Macherey Nagel). RNA extraction was performed according to the manufacturer’s instructions, except that beads were not washed in buffer MV5 but instead dried for at least 10 min, or longer, until the pellet did not appear shiny anymore, before continuing.

Sequencing amplicons were generated in a one-step reverse transcription PCR (RT-PCR), in a 50 µl reaction volume with three primers allowing reverse transcription and amplification of each segment. This protocol is based on that of Van den Hoecke et al. [710] with optimised volumes and RT-PCR conditions (Supplementary File S4.1).

Amplified products were purified using a NucleoSpin Gel and PCR Clean-up kit (Macherey Nagel), according to the manufacturer’s instructions. Purified products were examined with the Agilent 4200 TapeStation (Agilent Technologies) using the Agilent D5000 ScreenTape system. The concentration of each purified product was quantified with a Qubit 4 fluorometer (Invitrogen) using the Qubit broad-range assay.

The purified RT-PCR products were used to prepare sequencing libraries with a Nextera XT DNA sample preparation kit (Illumina), according to the manufacturer’s instructions. All libraries were sequenced on a MiSeq (Illumina) using the MiSeq v3 chemistry according to the manufacturer's protocol, producing 2x250 bp paired-end reads. All generated WGS data have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) [765] under BioProject accession number PRJNA615341.

4.2.2. Generation of consensus genome sequences

Raw (paired-end) reads were trimmed using Trimmomatic v0.32 [610] with the following settings: 'ILLUMINACLIP:NexteraPE-PE.fa:2:30:10', 'LEADING:10', 'TRAILING:10' 'SLIDINGWINDOW:4:20', and 'MINLEN:40' retaining only paired-end reads. For each sample, a suitable reference genome for read mapping was selected from the NCBI viral genomes resource [766] (Supplementary File S4.1). Consensus sequences for all samples were obtained following the GATK 'best practices' protocol (Supplementary File S4.1). All 253 generated consensus genomes were deposited in the GISAID database (i.e., samples EPI_ISL_415199 to EPI_ISL_415452) [752]. Sequencing coverage was extracted for each position from each sample using SAMtools depth 1.3.1 [767] and positions were normalized against the vaccine strain length. The percent identity matrix was calculated using the online muscle program hosted by the EBI (European Bioinformatics Institute) [768] for each segment.

4.2.3. Phylogenomic analysis

The WHO/ECDC provide a list of references representing each HA clade [769], but provisional clustering tests indicated that these did not allow the defining of phylogenetic groups for all samples. Additional reference sequences were selected from phylogenetic trees in ECDC reports and Nextstrain (www.nextstrain.org) [256] (Supplementary Table S4.1, available in the online version of this article). Alignments of sequenced samples and reference sequences were generated for all segments employing mega 7.0.18 [770] using default parameters for ClustalW [771] alignment. Only protein-encoding sequences of each segment were retained. beast v1.10.4 [772] was used to create phylogenetic trees for every segment individually, and the whole-genome, with underlying evolutionary model and other settings as listed in detail in Supplementary File S4.1. Maximum clade credibility trees were generated afterwards using the TreeAnnotator program of beast with default settings. Generated trees were visualised in iTOL (https://itol.embl.de/login.cgi) [650].

A local Nextstrain instance [256], allowing lightweight phylogenomics comparison with much more genomes than possible with beast, was built using sequenced samples complemented with GISAID sequences. Only those GISAID whole genomes were retained that included patient sex and age information, directly sequenced without passaging in cells or eggs, resulting in 14 157 genomes. All sequences were aligned with CLC Genomics Workbench 20.0.2 with default parameters, and UTRs (untranslated regions) were stripped. Aligned segments were concatenated into a single sequence for all samples retaining only sequences with <3 gaps and/or 'N' characters. Genomes were clustered based on sequence identity with cd-hit 4.6.8 using different cut-offs to retrieve ~3000 genomes, which was reached at 99.83 % sequence identity. Sequenced Belgian samples were retained irrespective of their sequence similarity (Supplementary Table S4.2). The local Nextstrain instance was then constructed as detailed in Supplementary File S4.1. The generated tree was visualised in R using the packages 'ggtree' and 'phytools'.

4.2.4. Reassortment detection

For manual reassortment detection, individual segment trees (Figure 4.1 and Supplementary Figure S2-S8) obtained with beast were compared visually to the whole-genome tree (Figure 4.2) and reconciliation of topologies was sought between the eight influenza segment tree. When a topological inconsistency in sample positioning in one of the segment trees was present, indicative of belonging to another phylogenetic group, the posterior probability value of the ancestral nodes in the segment tree between the group of the sample defined by the whole-genome tree and the group in which the reassorted genome was present, was checked. Only if this posterior probability value was ≥0.95, the genome was retained as an intra-subtype reassortant. Reassortment detection was also computationally independently performed using Graph-incompatibility-based Reassortment Finder (GiRaF) software v1.02 [773]. Tree files from two randomly selected replicates of the whole-genome analysis produced by beast were downsampled to 1,000 trees and analysed with GiRaF using default settings. Only reassortments with a confidence level ≥0.95 were accepted. Subsequently, a consensus approach was applied by retaining only reassortments detected by both methods. To estimate the reassortment frequency, the number of reassortant genomes was divided by the total number of genomes.

4.2.5. Inference of host characteristic associations

All statistical analyses were performed using R-software (RStudio 1.0.153; R3.6.1). The two-sided Fisher's exact test was used to assess associations between host characteristics and phylogenetic groups for both the whole genome and each segment individually. The same analysis was performed to assess associations between severity outcome and reassortment presence/absence. Host characteristics included infection severity (classified into mild, moderate and severe), patient age (categorized into <15, 15-59 and ≥60 years), sampling date (before, during and after the epidemic peak), vaccination status, presence of comorbidities and severity indicators. Multiple testing correction was conducted by applying the Benjamini-Hochberg method [774] and controlling the false discovery rate (FDR) at 5 %. For statistically significant associations identified during the univariate analysis, multiple linear regression was used to identify potential confounding variables and effect modifications.

Permutation analyses were performed to investigate the effect of different sample sizes for ILI (n = 93) and SARI (n = 160) cases. SARI and ILI samples were randomly selected 10 000 times with replacement for x = 1, 2, ..., 93 cases of both the total ILI and SARI population. In addition, at x=93, a two-sided permutation test at a significance level of 0.05 was performed (i.e., attempting to answer the question, does the observed value for 93 samples lay within the 95% confidence interval when 10 000 times 93 samples are randomly selected with replacement from the respective ILI/SARI case population?). All analysis scripts are provided in Supplementary File S4.1.

4.3. Results

4.3.1. WGS of clinical influenza A(H3N2) samples

Whole-genome A(H3N2) sequences were successfully obtained for all 253 samples used in this study. Supplementary Figure S4.1 provides an overview of the sequencing coverage for each position from each sample. Obtained sequences had a median depth (i.e., number of times each base shows up in individual reads) and breadth (i.e., total recovered genome sequence length) of coverage >9077x and >99 %, respectively. Sequencing efficiency varied slightly inversely with segment size, with smaller fragments such as the M segment generally displaying a slightly greater depth. For two samples only, a part of the genes encoding for PB2, PB1 or NP had a coverage <100x. Consequently, all samples were retained for further analysis.

4.3.2. Classification using WHO/ECDC guidelines for the HA segment

Samples were first classified according to WHO/ECDC guidelines by creating a phylogenetic tree with beast software [772] for the HA segment. The terms ‘clade’ and ‘phylogenetic’ group specifically refer to grouped samples using either the current or our newly proposed classification method, respectively, whereas the term ‘genetic group’ refers to samples that cluster together without specifically considering the classification method. WHO/ECDC provide a list of influenza references representing each HA clade [769], but provisional clustering tests indicated that these did not allow proper definition of the Belgian samples within clades (data not shown). Therefore, more reference sequences were selected manually from phylogenetic trees in ECDC reports and nextstrain.org [256] by building provisional trees to evaluate whether the Belgian samples could be classified. Figure 4.1 presents the resulting phylogenetic tree based on the HA segment obtained with beast for all sequenced samples and employed additional reference/vaccine sequences. Classification following the WHO/ECDC guidelines based on clustering with reference and vaccine strains (references provided by WHO/ECDC are indicated specifically with red arrows in Figure 4.1), and identification of specific amino acid substitutions linked to clades, resulted in the identification of four clades. In total, 26, 18, 7 and 49 samples, respectively, could be attributed to the HA groups ‘3C2a1a’, ‘3C2a1b’, ‘3C2a2’ and ‘3C2a3’. Four samples clustered with a WHO/ECDC reference for group ‘3C2a’ but were phylogenetically too distant to motivate their inclusion in this group. The employed WHO/ECDC reference for Group ‘3C2a1’ did not cluster with samples that should belong to this group, but rather with samples of Group ‘3C2a1a’. Belgian samples belonging to group “3C2a1”, however, could be classified because they clustered with some of the additional reference strains that were selected from ECDC reports and nextstrain.org. In total, the approach using solely WHO/ECDC information resulted in 153 unassigned samples, of which 59 lacked at least one clade-defining amino acid substitution (highlighted by a grey strip in Figure 4.1) and 144 did not cluster with any reference strain provided by WHO/ECDC guidelines.

Considering the support of nodes by posterior probability values in relation to the additionally employed references (indicated in Figure 4.1 with the colour of their corresponding phylogenetic group but without a red arrow), and specific additionally identified substitutions, resulted in classification of 11 phylogenetic groups. Group ‘3C2a’ could now be defined because samples clustered properly with one of the additionally selected references that was defined as clade ‘3C2a’. Group ‘3C2a1’ and group ‘3C2a1(2)’, and group ‘3C2a1a’ and group ‘3C2a1a(2)’, are phylogenetically closely related to each other, but were both split into two subparts denoted with the suffix ‘(2)’ because they were clearly delineated in the phylogeny, albeit supported by nodes with low posterior probability values. Additionally, most samples of Group “3C2a1” possessed an additional amino acid substitution R142K compared to group ‘3C2a1(2)’, and group ‘3C2a1a(2)’ lacked the clade-specific substitution T135K of group ‘3C2a1a’. Group ‘X’ consists of samples that similarly clustered together in most segment trees, except the PB1 and M trees. Groups ‘HAX’ and ‘HAY’ include samples difficult to classify into the other phylogenetic groups. The classification uncertainty of these samples was often reflected in other segment trees. Additionally, these samples possess different mutations in the HA segment and in the whole genome compared to samples in other phylogenetic groups, which supports their designation to separate phylogenetic groups.

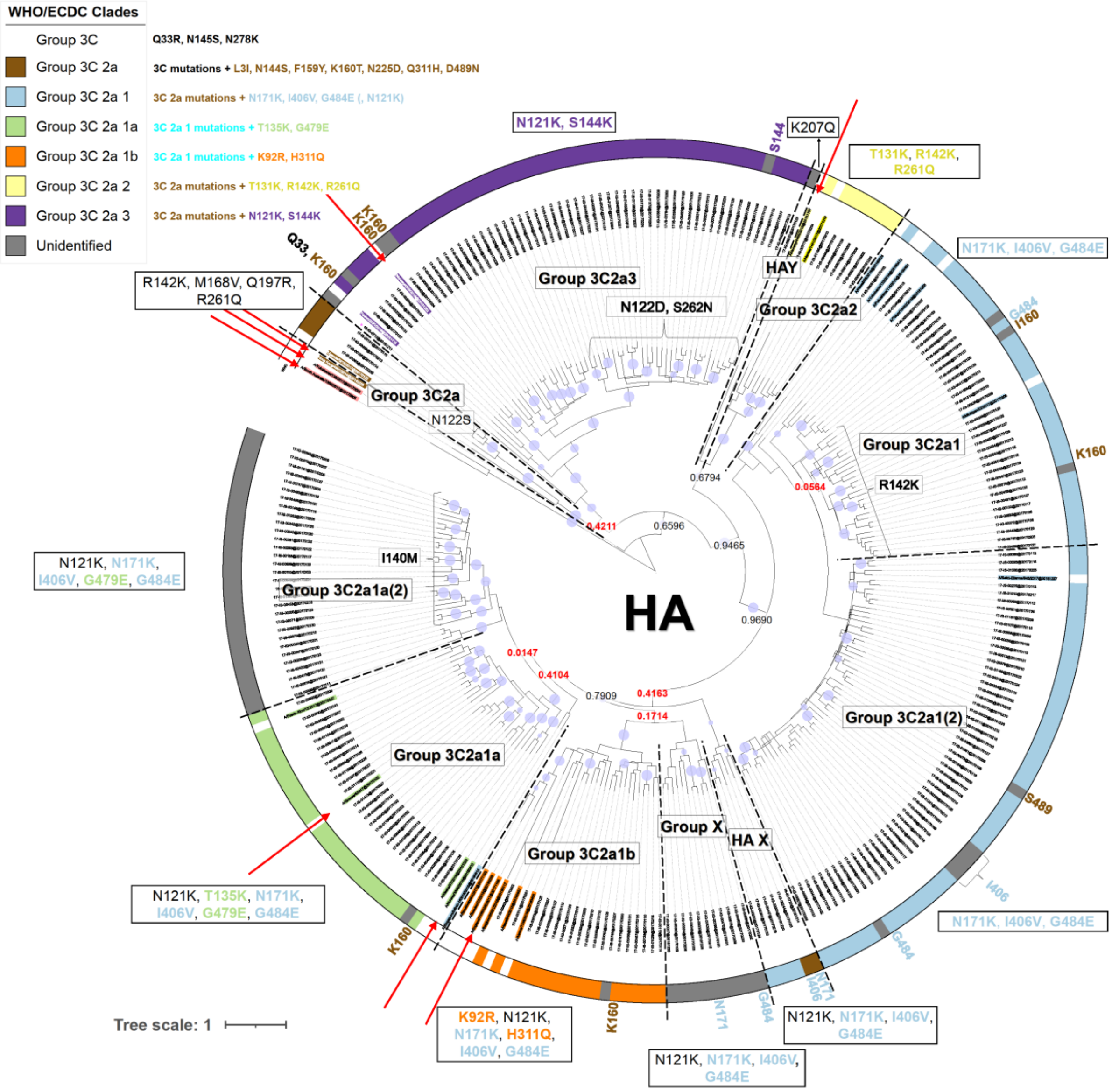


Figure 4.1: Phylogenetic HA gene tree. Red arrows indicate (coloured) reference names assigned to the various groups as defined by the WHO/ECDC, and coloured names without red arrows indicate additional references selected from ECDC reports and nextstrain.org. Specific amino acid substitutions designated to groups according to WHO/ECDC guidelines are indicated on the figure, and the circular coloured outer strip around the tree represents the assigned groups based on amino acid substitutions defined in the WHO/ECDC guidelines, according to the colour legend. Within the tree, the group labels represent the 11 phylogenetic groups that were assigned to their respective samples according to their classification based on references (coloured names) and the support of nodes by posterior probability values. Group 'X' clustered together in a separate cluster from the other groups. Groups 'HAX' and 'HAY' contain samples that could not be classified. Posterior probability values are indicated on key nodes that separate groups, and are coloured red if below 0.5. The size of the blue discs on nodes represents the posterior probability scaled between 0.5 and 1. The scale bar represents the mean number of substitutions per site.

4.3.3. Using whole-genome sequences and beast allows improved phylogenetic classification

WGS enables construction of phylogenies for the other seven segments separately, as well as the combined whole genome. A phylogenetic tree was constructed with beast using the whole-genome sequences of the Belgian isolates rather than solely the HA segment, after which classification was performed similarly by considering the support of nodes by posterior probability values in relation to employed reference/vaccine genomes, and specific additionally identified substitutions. The resulting classification is presented in Figure 4.2. Additionally, the same approach for the other seven segments was individually applied, and is presented in (Figs S2-S8). Comparison of the HA (Figure 4.1) and whole-genome (Figure 4.2) trees indicates that approximately 83.4 % of samples were classified in the same phylogenetic groups. Additionally, the whole-genome tree overall has higher posterior probability values [posterior(nodes) ≥ 0.5 : 203 (whole genome) vs 94 (HA)], increasing confidence in the overall topology, because the posterior probability of a tree is the probability that the tree is correct given the data if the underlying model is correct [775, 776].

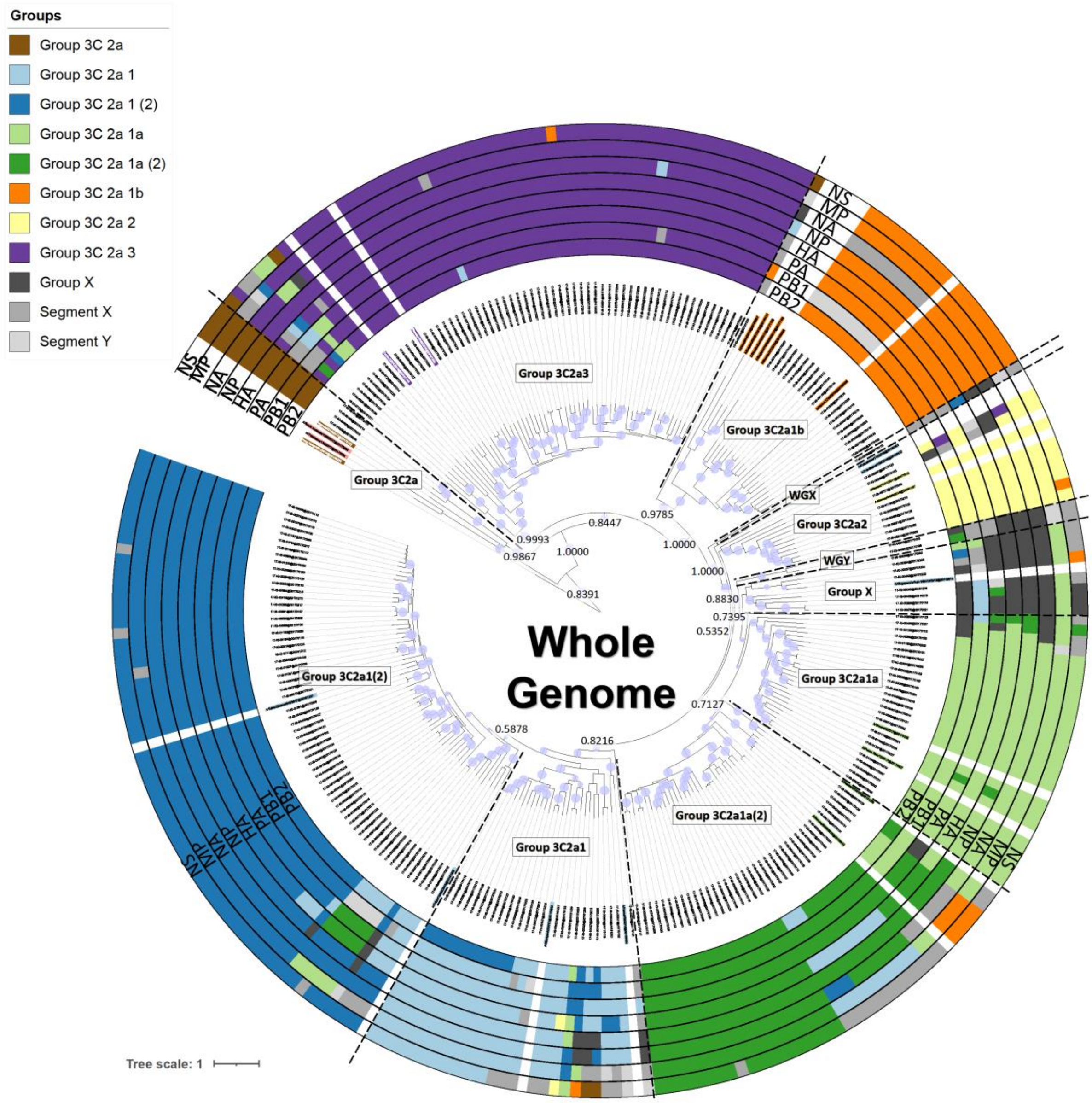


Figure 4.2: Phylogenetic tree based on the whole genome. Coloured names indicate additional references selected from ECDC reports and nextstrain.org. Coloured rings around the tree represent classification results for the eight segments separately (Figure 4.1 for the HA gene, and Figs S2-S8 for the seven other genes). Group ‘X’ clustered together in a separate cluster from the other phylogenetic groups. Groups ‘WGX’ and ‘WGY’ contain samples that could not be classified. Groups labelled with the segment name and a single letter (e.g., PB1X) similarly represent any remaining samples that could not be confidently assigned into phylogenetic groups according to their segment trees (S2-S8 Figs). Within the tree, the group labels represent the phylogenetic groups that were assigned to their respective samples according to their classification based on references (coloured names) and the support of nodes by posterior probability values. Posterior probability values are indicated on key nodes that separate phylogenetic groups. The size of the blue discs on nodes represents the posterior probability scaled between 0.5 and 1. The scale bar represents the mean number of substitutions per site.

4.3.4. Increasing the number of genomes considered with a custom-built Nextstrain instance

Proper reference selection was extremely arduous, requiring multiple explorative analysis iterations to select suitable reference genomes allowing classification to arrive at the phylogenies and resulting classification presented for HA (Figure 4.1) and the whole genome (Figure 4.2). An alternative approach was explored by using Nextstrain to reconstruct an in-house instance of the Belgian samples supplemented with more than 2500 publicly available A(H3N2) genomes available in the GISAID database, because Nextstrain is a framework meant for real-time analysis (i.e., as an outbreak is ongoing) of several hundreds to even thousands of genomes. This allowed positioning of the Belgian samples within the globally circulating context. When comparing the custom-built Nextstrain tree (Figure 4.3) with the whole-genome tree of the Belgian samples (Figure 4.2), 217 Belgian samples clustered congruently in both trees. Six samples from group ‘X’ clustered together in a separate cluster. The remaining 30 samples were placed differently between both trees. The high congruence between both trees enabled to more easily identify suitable references. Figure 4.3 illustrates considerable diversity amongst the Belgian samples, belonging to five major phylogenetic groups: (i) samples belonging to group ‘3C2a3’; (ii) group ‘3C2a1’ and group ‘3C2a1(2)’; (iii) group ‘3C2a1a’ and group ‘3C2a1a(2)’; (iv) group ‘3C2a1b’; and (v) group ‘3C2a2’. Group ‘3C2a1’ and group ‘3C2a1(2)’, and group ‘3C2a1a’ and group ‘3C2a1a(2)’, formed two distinct clusters each time, supporting their separation into different phylogenetic groups as previously suggested by both the HA (Figure 4.1) and whole-genome (Figure 4.2) trees. Lastly, four samples belonged to an additional phylogenetic group corresponding to group ‘3C2a’. The root for the samples from the Belgian 2016-2017 outbreak season goes back to 2003-2004, and phylogenetic groups containing Belgian samples are interspersed with many samples isolated in other countries, indicating that phylogenetic groups are not limited by country boundaries and that these groups have already co-circulated in the Belgian population for several years.

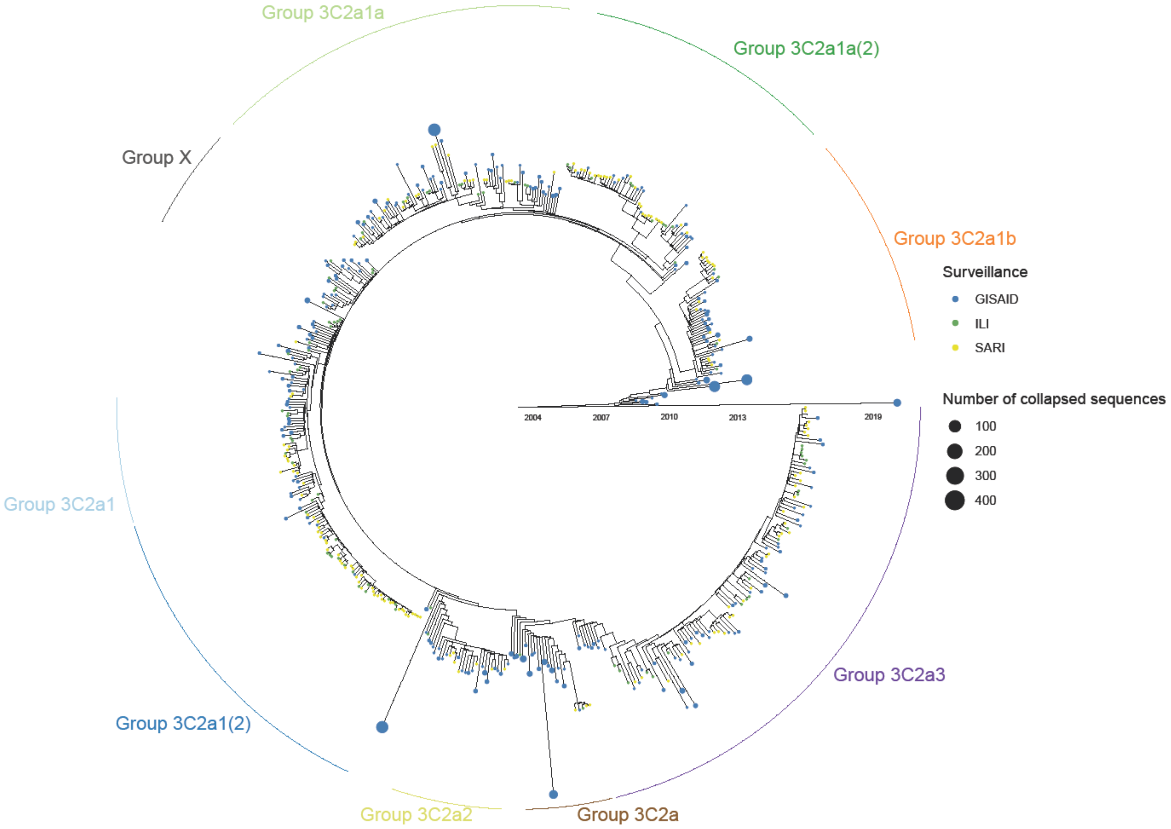


Figure 4.3: Time-resolved overview of influenza samples from the Belgian 2016-2017 outbreak season in the context of globally circulating influenza strains based on an in-house Nextstrain instance using only whole-genome sequences. Green- and yellow- coloured dots represent ILI and SARI samples sequenced in this study, respectively. Blue coloured dots represent GISAID samples. If an ancestor only included GISAID sequences, these nodes were collapsed for better visualisation with the size of such nodes proportional to the number of included samples. Phylogenetic groups based on the whole genome (Figure 4.2) are indicated around the tree. The branch lengths correspond to the sampling date of the sample. In case of grouped GISAID samples, the sampling date of the latest sample is used.

4.3.5. Detection of intra-subtype reassortments

WGS enabled identification of (intra-subtype) reassortments within the Belgian A(H3N2) samples, allowing for investigation of the influence of reassortments on virus evolution through both manual inspection by visually comparing individual segment trees obtained with beast to the whole-genome tree, as well as through computational analysis by using the GiRaF software. For both methods, the genome was retained as an intra-subtype reassortment if the posterior probability value was ≥0.95. Results of the combination of the manual and computational analysis are visualised in Figure 4.4, and a detailed list of reassortments detected by both manual inspection and computational methods is provided in Supplementary Table S4.3. Using the combination of the manual and computational approach, 38 strains were characterised as intra-subtype reassortants resulting in an intra-subtype reassortment rate of

15.02 %. With the manual detection method, 57 reassorted genomes were detected (22.53 %), whereas with GiRaF 39 reassorted genomes were identified (15.42 %).

No statistically significant result was obtained using the Fisher's exact test between the surveillance system and reassortment status. Permutation analyses were performed to investigate the association between disease severity and reassortment status while correcting for the different sample sizes of ILI and SARI samples. For all sample sizes ranging from 1 to 93 (the maximum number of ILI samples), the number of reassorted genomes was consistently higher for SARI samples compared to ILI samples (Supplementary File S4.1). Moreover, at a sample size of 93, a two-sided permutation test at a significance level of 0.05 indicated that significantly more reassorted genomes were present in the SARI population (P value=0.0208; Supplementary File S4.1).

4.3.6. Associations between host characteristics and phylogenetic groups

Associations between host characteristics and phylogenetic groups for both the whole genome and each segment individually were inferred using the Fisher's exact test with FDR correction. Statistically significant results for associations between phylogenetic groups and host characteristics are presented in Figure 4.4 for the whole genome. Results for the individual segments are presented in Supplementary Table S4.4. Significantly more male than female patients were infected with strains belonging to group '3C2a3' not only for the whole genome, but also for the HA segment with the same confounding effects (i.e., variables that have an influence on the correlation between the group and the host characteristic) (Figure 4.4, Supplementary Table S4.4). Group 'X' mainly consists of samples from ILI cases not only for the whole genome but also for the HA segment with the same confounding effects (Figure 4.4). A significant association was detected between group '3C2a1a(2)' and sampling period for the whole genome, with samples being present mainly during the seasonal beginning and peak, but decreasing in presence towards the end. Although this association was not detected for the HA segment, it was detected when samples were classified according to the PB2 and PB1 segments (Supplementary Table S4.4). A significant association between group '3C2a3' and sampling period for the whole genome was also detected, with an increasing number of samples towards the end. This significant association was also detected when samples were classified according to the PB1 segment (Supplementary Table S4.4). Although not significant, group '3C2a1a' was emerging during the course of the influenza season (Supplementary Figure S4.9), while group '3C2a2' only appeared during the seasonal peak. Several additional significant associations were detected between host characteristics and the individual segments (Supplementary Table S4.4). All genomes were assessed regarding antiviral resistance by comparing to a previously composed database of antiviral resistant mutations [749], but no known antiviral resistance mutations were observed within the genomes (results not shown).

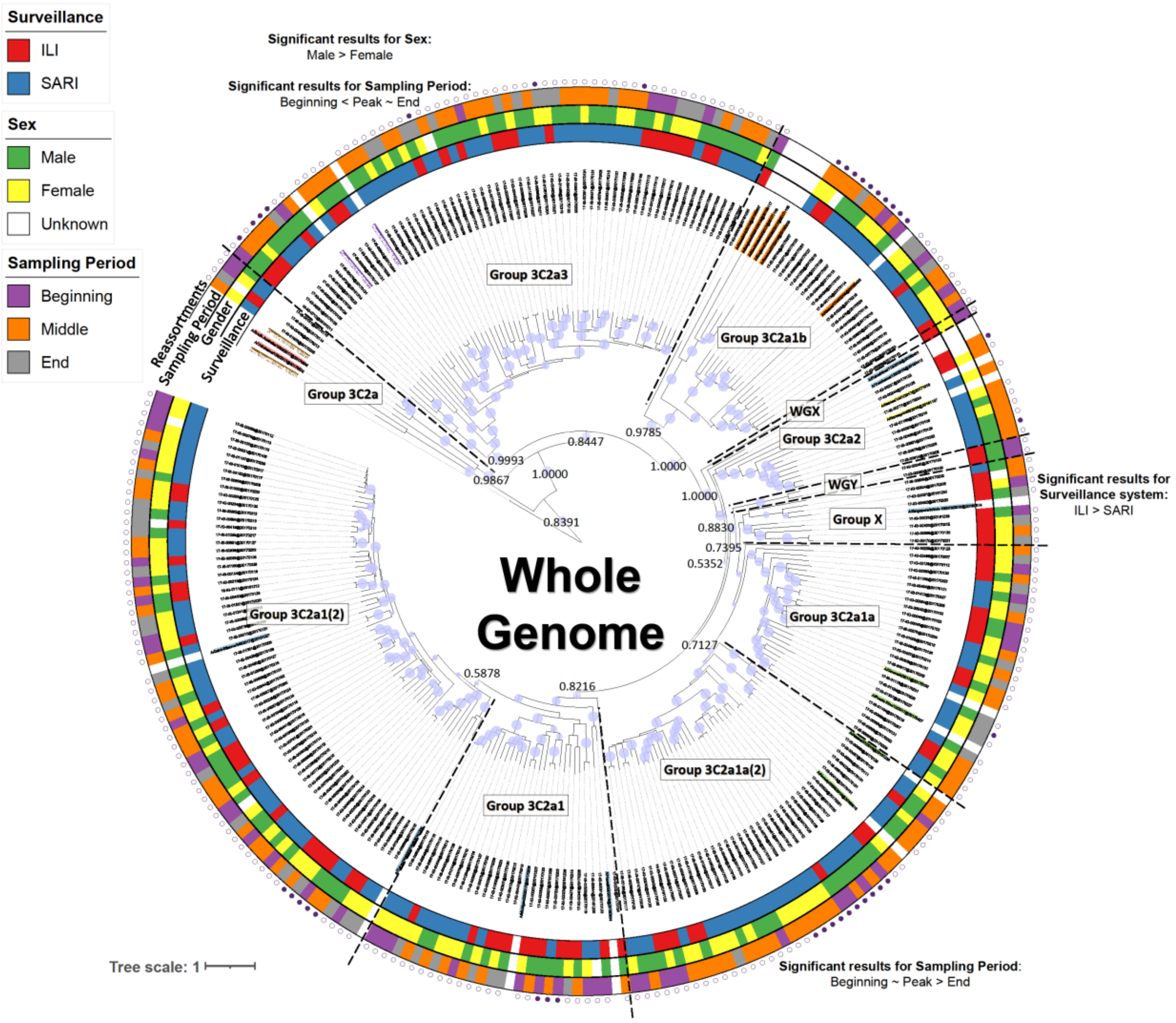


Figure 4.4: Phylogenetic tree based on the whole-genome annotated patient data for which significant associations with phylogenetic groups were detected. Coloured taxon labels indicate additional references selected from ECDC reports and nextstrain.org. Coloured rings around the tree represent patient data, including the surveillance system, sex and sampling period, for samples where this information was available. Around the outside of these strips, the presence (filled circle) or absence (empty circle) of reassorted genomes is indicated based on the consensus of both manual inspection and computational analysis with GiRaF. Statistically significant results using the Fisher's exact test with FDR correction for associations between host characteristics and sample parameters with the newly defined phylogenetic groups for the whole genome (see Figure 4.1) are presented on the figure near their respective phylogenetic group. Results for individual segments and more detailed information, including the effect size and confidence interval, are presented in Supplementary Table S4.4. The host characteristics and sample parameters for the reference genomes were excluded. Posterior probability values are indicated on key nodes that separate phylogenetic groups. The size of the blue discs on nodes represents the posterior probability scaled between 0.5 and 1. The scale bar represents the mean number of substitutions per site.

4.4. Discussion

The field of microbiology is transforming due to the decreasing turnaround times and costs, and increasing availability of whole-genome information and user-friendly data analysis tools. The added value of genomic surveillance has been showcased during the SARS-CoV-2 pandemic by tracking the virus, detecting emerging variants of concern, and applying the genomic data to a wide variety of associated biological questions [675]. As the COVID-19 restrictions also limited the spread of influenza, genomic surveillance of influenza is likely to become even more important when COVID-19 restrictions ease [777]. Although the HA gene remains the principal region for Sanger sequencing in classical influenza surveillance programmes, this study illustrates the feasibility and benefit of switching towards WGS-based surveillance. Sanger sequencing the HA gene has a lower cost compared to WGS, but when several samples are multiplexed in the same WGS run, the overall cost of the latter is lower compared to Sanger sequencing every gene separately [19]. This study proposed an improved methodology for classification of circulating strains with higher resolution for genetic characterisation and reassortment detection. Moreover, integration of WGS data with patient information allowed investigation of the associations of phylogenetic groups based on the whole genome with host characteristics.

Influenza classification according to WHO/ECDC guidelines is currently based on phylogenetic analysis of the HA segment through clustering with vaccine/reference strains and detection of predefined amino acid substitutions [743]. However, this approach enabled however only classifying a small subset of the Belgian influenza 2016-2017 outbreak samples because the majority of samples did not cluster with reference strains and/or lacked specific clade-defining HA amino acid substitutions. Our study evaluated the feasibility of shifting classification for influenza surveillance to a whole genome-based approach rather than solely the HA gene, also incorporating much more references selected from ECDC reports and nextstrain.org, to improve clade definition. We have demonstrated that employing whole genome information offers improved classification performance because more samples could be characterised into well-supported phylogenetic groups compared to only considering the HA gene.

Current surveillance programmes typically employ relatively simple phylogenetic tree reconstruction methods based on distance estimation and maximum parsimony, such as mega neighbour joining [778, 779]. More advanced methods such as RAxML maximum likelihood are used only rarely [743]. In contrast, phylogenetic tree construction through Bayesian inference has emerged as a standard in recent years for fundamental viral genomics studies. This powerful but resource-intensive approach allows robust phylogenomic investigation and facilitates deep exploration of the circulating genetic diversity [780, 781]. Here, beast software

[772] was used, which takes phylogenetic uncertainty into account and incorporates complex evolutionary models [782]. beast was for instance used to find the origin and map avian influenza A(H7N9) diversity causing human infection [783], simulating real-time evolutionary rate estimates and dating the emergence and intrinsic growth rate of the A(H1N1)pdm09 pandemic [784]. Our case study therefore demonstrates the feasibility of switching to more robust phylogenetic tree reconstruction based on Bayesian inference for genetic influenza surveillance. Ideally, more rigid model selection is performed for validating the underlying model assumptions, but this still requires computational resources beyond the capacity of the Belgian, and most other National Reference Centers (NRCs).

Being time-consuming and computationally intensive, Bayesian inference is, however, infeasible for larger datasets containing several hundreds of samples or if a quick response is required. Selection of suitable genome references for phylogenetic classification poses another challenge. Therefore, an alternative approach was explored using a custom-built Nextstrain instance, which offers several advantages. Nextstrain can analyse several hundreds to thousands of genomes much faster compared to beast and presents a real-time view. Although the public influenza instance hosted at nextstrain.org is based solely on the HA and NA genes, we demonstrated the feasibility of using whole-genome information. The feasibility of using Nextstrain with whole genomes also allows provisionally assigning samples to already existing or newly emerging groups to select suitable references to perform more powerful phylogenomics investigation with methods such as beast.

Consistent with previous studies, we found co-circulation of different A(H3N2) phylogenetic groups during influenza epidemics [780, 785, 786]. A total of 211 of 253 samples (83.4 %) classified based on the HA segment belonged to the same phylogenetic group based on the whole genome, suggesting that the other seven segments contribute important additional genetic information. Intra-subtype reassortments were observed at a rate of ~15 % in the Belgian 2016-2017 season, likely facilitated by the co-circulation of several phylogenetic groups. Although intra-subtype reassortment detection remains challenging, posterior probability values derived from beast aided reassortment detection with both manual inspection and computational approaches to avoid selecting uncertain reassortments. Intra-subtype reassortments are currently typically studied using manual inspection of different segment phylogenetic trees by checking for positional inconsistencies. This is highly time-consuming, error-prone and aggravated by unclear phylogenetic trends in segments that exhibit limited genetic diversity. Recent subtle reassortments are more challenging to detect because sequences from the same subtype are more similar [773, 787]. Through our strict requirements for reassortment detection by both a computational and manual method requiring high support values, the resulting reassortment rate of 15 % is likely underestimated. Goldstein et al. [780] studied A(H3N2) in Scotland for the 2014-2015 season, and observed less intra-subtype reassortments (5.3 %) using a similar methodology. However, Berry et al. [748] studied reassortments in 2 091 A(H3N2) globally circulating strains collected between 2009 and 2014, and observed a higher rate of intra-subtype reassortments (39.1 %) by using a computational approach. However, it should be noted that, similar to recombination for bacterial pathogens, reassortment can disturb the true phylogenetic signal in whole-genome-based phylogenetic investigations because it disturbs the underlying assumptions of the tree model, resulting in biased topologies. One strategy to circumvent this is by removing all reassorted samples, but this would typically remove large proportions of the input dataset for influenza due to its relatively high rate of reassortment. Therefore, a need exists for development of new tools and models that can take this effect into account, such as the Bacter package in beast2 developed for bacterial pathogens [788].

SARI samples were found to be more likely to comprise reassortments in agreement with Goldstein et al. [780]. Nelson et al. similarly found that reassortments could potentially trigger emergence of unusually severe seasonal A(H1N1) epidemics [789]. Several statistically significant associations were detected between patient data and the whole genomes that could not be identified using solely HA data: more males were associated with group ‘3C2a3’ and more samples occurred near the seasonal beginning in group ‘3C2a1a(2)’, which was observed for neither the HA nor NA segments, but rather the PB2 and PB1 segments, suggesting that the potential value of other segments is undervalued. Other examples include more ILI than SARI cases for group ‘3C2a1a’ for the M segment, and more intensive care unit (ICU) than non-ICU cases for group ‘3C2a1a’ for the NA segment. Several other host characteristic associations with particular segments were observed (Supplementary Table S4.4). Additional information on associations of certain groups with host characteristics can be potentially useful, for instance, to detect groups more prone to result in severe disease allowing implementation of preventive measures. To the best of our knowledge, no other studies have explored the relationship between phylogenetic groups and host data for influenza. However, it should be noted that this dataset includes a limited number of samples as sequencing all samples would still be too expensive. Although all patient and sample information was used in the analyses, the samples were selected based on the severity, patient age and sampling date. Other patient and sample information was occasionally unknown in addition to an unequal distribution for some parameters. Additionally, some phylogenetic groups contained a limited number of samples. Consequently, the sample selection was underpowered for some parameters, which could be mitigated by a larger sample dataset.

Our study illustrates that genetic surveillance should gradually shift to WGS for seasonal influenza surveillance. WGS enables employment of powerful phylogenomics methods that substantially improve phylogenetic classification, thereby providing more information to national influenza prevention and control programmes regarding the timing, impact and severity of seasonal epidemics. WGS can also improve vaccine strain selection, which will be especially relevant for next-generation vaccines that do not focus solely on the HA segment. Bayesian inference using beast, facilitates reassortment detection with both manual inspection and computational methods, enabling investigation of intra-subtype reassortment effects on public health. Tools optimised for real-time analysis, such as Nextstrain, facilitate contrasting seasonal local outbreaks to the globally circulating context and can provide a quick response. Lastly, incorporating whole-genome information allows association of phylogenetic groups with host characteristics with particular epidemiological value, such as disease severity. Future research should consider investigating whether the high diversity within the A(H3N2) influenza subtype should be considered by using the phylogenomic groups to study mutations found in the genomes.

4.5. Acknowledgments

We thank the technicians of the service Transversal activities in Applied Genomics at Sciensano, Belgium to perform the Next Generation Sequencing run.

CHAPTER 5: INTEGRATING PATIENT DATA AND MUTATIONS ACROSS THE WHOLE INFLUENZA GENOME TO IMPROVE THE ASSOCIATION ANALYSIS

Context of this chapter:

This chapter illustrates the advantages of using WGS in routine influenza surveillance. Using an A(H3N2) influenza dataset of clinical samples, mutations across the whole genome were detected. These mutations were linked to the available patient data, which resulted in significant associations between some of the identified mutation and disease outcome, suggesting a potential relevance for vaccine improvement and influenza patient management. Additionally, because of the relatively high diversity within the A(H3N2) subtype, a new approach was proposed to classify influenza viruses based on phylogenetic classification to stratify the samples and reduce the viral genetic background.

This chapter was previously published:

Laura A. E. Van Poelvoorde, Kevin Vanneste, Sigrid C. J. De Keersmaecker, Isabelle Thomas, Nina Van Goethem, Steven Van Gucht, Xavier Saelens and Nancy Roosens (2022) Whole-Genome Sequence Approach and Phylogenomic Stratification Improve the Association Analysis of Mutations with Patient Data in Influenza Surveillance. Frontiers in Microbiology. **13**. https://doi.org/10.3389/fmicb.2022.809887

Authors' contributions:

Conceptualization: NR, KV, IT, XS, SDK, and SVG; Project Administration and funding acquisition: NR; Data Curation: LVP, IT, NVG; Resources: IT, SVG, SDK, KV; Formal Analysis: LVP, KV; Investigation: LVP ; Visualisation: LVP; Writing – Original Draft Preparation: LVP, NR; Writing – Review and Editing: all authors; Supervision: NR, KV, IT

Abstract:

Each year, seasonal influenza results in high mortality and morbidity. The current classification of circulating influenza viruses is mainly focused on the hemagglutinin gene. Whole-genome sequencing (WGS) enables tracking mutations across all influenza segments allowing a better understanding of the epidemiological effects of intra- and inter-seasonal evolutionary dynamics, and exploring potential associations between mutations across the viral genome and patient's clinical data. In this study, mutations were identified in 253 Influenza A (H3N2) clinical isolates from the 2016-2017 influenza season in Belgium. As a proof of concept, available patient data were integrated with this genomic data, resulting in statistically significant associations that could be relevant to improve the vaccine and clinical management of infected patients. Several mutations were significantly associated with the sampling period. A new approach was proposed for exploring mutational effects in highly diverse Influenza A (H3N2) strains through considering the viral genetic background by using phylogenetic classification to stratify the samples. This resulted in several mutations that were significantly associated with patients suffering from renal insufficiency. This study demonstrates the usefulness of using WGS data for tracking mutations across the complete genome and linking these to patient data, and illustrates the importance of accounting for the viral genetic background in association studies. A limitation of this association study, especially when analyzing stratified groups, relates to the number of samples, especially in the context of national surveillance of small countries. Therefore, we investigated if international databases like GISAID may help to verify whether observed associations in the Belgium A (H3N2) samples, could be extrapolated to a global level. This work highlights the need to construct international databases with both information of viral genome sequences and patient data.

5.1. Introduction

Influenza A virus displays the highest diversity of all influenza viruses and remains a major public health threat in developed as well as in developing countries [790]. Although influenza infections are mostly mild [757], some population strata are at high risk for developing complications [21]. There are currently mainly two influenza A subtypes circulating in humans, namely A (H1N1) pdm09 and A (H3N2) [8]. In particular, subtype A (H3N2) has led to numerous seasonal epidemics and is considered to evolve faster than other subtypes [753]. In recent years, A (H3N2) has shown extensive clade diversity and increased morbidity and mortality, especially in the elderly [754]. This rapid evolution is mainly caused by constantly occurring mutations and intra-subtype reassortment, resulting in low vaccine effectiveness through mismatches between the vaccine strain and circulating influenza strains.

Currently, the “World Health Organization” (WHO) and “European Centre for Disease Prevention and Control” (ECDC) still focus on the genetic surveillance of the HA segment [744]. In the context of influenza surveillance and vaccine strain selection different clades and subclades within each influenza subtype are defined based on its phylogenetic analysis and amino acid differences [743]. However, as next-generation sequencing (NGS) has become more widely accessible, whole-genome sequencing (WGS) and obtaining sequences from all eight influenza segments simultaneously becomes cost-efficient [791]. WGS data can be used for several purposes including to improve the influenza surveillance, when appropriate approaches and analysis are applied on a dataset. To illustrate these approaches, we have previously sequenced influenza samples collected in the context of the surveillance of the 2016-2017 influenza season in Belgium. In a first study, this dataset was used to demonstrate that using powerful phylogenomic tools such as BEAST and Nextstrain, allows substantially improved phylogenetic classification when considering the whole genome rather than solely the HA segment [792]. Furthermore, Bayesian inference via BEAST allowed reassortment detection by both computational methods and manual inspection. These combined methods resulted in an estimated rate of 15% intra-subtype reassortment for A (H3N2) samples from the Belgian 2016-2017 outbreak season. Additionally, A (H3N2) reassortants were found to be more likely to infect hospitalised patients compared to patients with mild symptoms, which would not have been possible without considering the whole genome [792]. In another study, we have used genomic data from the hospitalised patients in this dataset in a predictive model and assessed the added value of viral genomic data in addition to clinical information [793]. The aim of the current study is to evaluate whether whole genome information may also enable exploring mutations located on all eight segments. Moreover, by integrating with patient data, associations of viral mutations and patients’ characteristics can be detected, including the disease severity. In contrast to bacterial infections, clinical studies exploring the link between mutations and the disease severity remain scarce [794]. The pathogenesis of viruses is dependent on complex and unpredictable mechanisms, including interactions formed within and between the influenza proteins. Consequently, certain mutations cannot be considered individually, but should be considered together with mutations present in the entire genome, i.e., the genetic background [795] that evolves fast due to the highly error-prone influenza replication [795, 796]. It is therefore more appropriate to include virological genetic information of all 8 segments and metadata of the host to investigate influenza [797, 798]. Most current studies focus on linking mutations to broadly defined patient outcomes related to vaccine efficacy or disease severity. However, assessing associations with the sampling period and additional patient information such as vaccination status, patient age, existing patient comorbidities, sex, and specific severity indicators has, to the best of our knowledge, not yet been performed.

In this study, 253 Influenza A (H3N2) whole genome sequences from the Belgian surveillance, for which phylogenetic classification has been previously reported [792], were used to explore potential associations between mutations positioned across the whole genome and patient characteristics and other metadata. In this analysis, the effect of sampling stratification according to the phylogenetic clade was also evaluated to consider potential effects related to the highly diverse genetic background of A (H3N2) strains. Additionally, we evaluated whether the observed associations with a restricted number of samples at the Belgian level correspond to trends observed at an international level, and highlight the necessity of constructing a large database containing both viral genome sequences and information on patient data.

5.2. Material and Methods

5.2.1. Sample Selection, RNA Isolation, PCR Amplification, and WGS

Two sentinel surveillance systems are in place in Belgium to monitor “influenza-like-illness” (ILI) in the general practices and “severe-acute-respiratory-infections” (SARI) in the hospitals. ILI cases are defined by a sudden onset of symptoms, including fever and respiratory and systemic symptoms. A SARI case is defined as an acute respiratory illness with onset within the previous 10 days of fever, respiratory symptoms, and the requirement for hospitalisation. These surveillance systems are essential for following trends of virus spread and changes in circulating influenza viruses. The present study uses 253 samples collected during the 2016-2017 influenza season in Belgium from the two surveillance systems, as previously described [792]. These include 160 hospitalised SARI patients (mean age = 70 years) and 93 ILI outpatients (mean age = 39 years). The absence of other respiratory viruses in the sample was confirmed by RT-qPCR-based testing for respiratory syncytial virus A and B, parainfluenza viruses, enterovirus D68, rhinoviruses, human metapneumovirus, paraechoviruses, bocaviruses, adenovirus, coronaviruses OC43, NL63, 229, and MERS-CoV [799, 800]. Samples of ILI outpatients were categorized as mild cases (n = 93). Samples from hospitalised SARI patients were categorized as moderate (n = 122) or severe cases (n = 38). As the requirement for hospitalisation is part of the SARI case definition, all SARI cases are consequently hospitalised patients. However, hospitalisation by itself was not considered as a disease severity indicator because patients could have been hospitalised for isolation purposes or due to other medical conditions. A severe case was therefore defined within the SARI population by the presence of at least one of the following severity indicators: death, stay in an intensive care unit (ICU), need for invasive respiratory support or extracorporeal membrane oxygenation (ECMO) or acute respiratory distress syndrome (ARDS).

Table 5.1: Sample numbers per patient data. These statistics were based on a national collection containing 93 ILI (mild) samples and 160 SARI (moderate=122; severe=38) samples.

Age (years):	**<15**	**15 – 59**	**≥60**
Beginning of epidemic (<week 4)	12	17	35
Peak of epidemic (week 4 - 6)	16	26	86
End of epidemic (>week 6)	11	16	34

ILI	93	**SARI**	160
Male*	122	**Female***	122
Vaccinated*	52	**Not vaccinated***	130
Antibiotics administered*	100	**No antibiotics administered***	126
Respiratory disease*	50	**No respiratory disease***	199
Cardiac disease*	54	**No cardiac disease***	195
Obesity	20	**No obesity**	233
Renal insufficiency	35	**No renal insufficiency**	218
Hepatic insufficiency	6	**No hepatic insufficiency**	247
Diabetes	27	**No Diabetes**	226
Immunodeficiency	23	**No immunodeficiency**	230
Neuromuscular disease	21	**No neuromuscular disease**	232
Stay in ICU	22	**No stay in ICU**	231
Fatal	19	**Not fatal**	234

*Samples for which certain patient data was unknown, were excluded for analysing that particular aspect.

Available patient data are listed in Table 5.1 in conjunction with the number of patients. The nucleic acid content of the samples was extracted directly from the clinical specimens and subjected to WGS as previously described (Van Poelvoorde et al., 2021). Generated WGS data has been deposited in the NCBI Sequence Read Archive (SRA) [33] under accession number PRJNA615341. A central ethical committee and the local ethical committees of each participating hospital approved the SARI surveillance protocol (reference AK/12-02-11/4111; in 2011: Centre Hospitalier Universitaire St-Pierre, Brussels, Belgium; from 2014 onward: Universitair Ziekenhuis Brussel, Brussels, Belgium). Informed verbal consent was obtained from all participants or parents/guardians.

The genome consensus sequences were obtained as previously described [792], and are available in the GISAID database as isolates ID EPI_ISL_415199 to EPI_ISL_415452 [39]. Identification of genome mutations requires a closely-related reference genome. In this study, the whole genome of 2016-2017 A(H3N2) vaccine strain A/HongKong/4801/2014 (GISAID: EPI_ISL_198222) was used as a reference. This strain was used as the reference because it should be genetically close to the patient samples for that season. The obtained genome consensus sequences were aligned using ClustalW in Mega 7.0.18 with default settings. The H3 numbering, excluding the signal peptide of 16 amino acids, was used to enumerate positions (both amino acid residues and the corresponding nucleotides) in the HA protein compared to this reference strain. Samtools depth 1.3.1 [767] was used to extract the coverage at each position for each sample from the BAM files. Regions with a sequencing depth lower than 100X were discarded. For two samples (A/Belgium/S0978/2017 and A/Belgium/S0182/2017), a part of the PB2, PB1, or NP fragment had a coverage lower than 100X and mutations found in these regions of these samples were consequently not considered. Additionally, mutations that occurred in less than 5% or more than 95% of all samples were also discarded as these will not contribute to the detection of associations between the mutation and the patient data.

5.2.2. Phylogenomic Analysis and Subsampling by Group

A Bayesian phylogenetic tree was created as previously described [792]. The protein-coding sequences of the sequenced samples and references were aligned using MEGA 7.0.18 [770] using default parameters for ClustalW [771] alignment. Phylogenetic trees for the whole-genome were created using BEAST v1.10.4 [772]. Classification was performed by considering the support of nodes by posterior probability values in relation to specific additionally identified substitutions and the reference genomes. Based on the whole-genome tree, eleven phylogenetic groups were identified [792]: “Group 3C2a” (n = 4), “Group 3C2a1” (n = 26), “Group 3C2a1(2)” (n = 62), “Group 3C2a1a” (n = 25), “Group 3C2a1a (2)” (n = 37), “Group 3C2a1b” (n = 20), “Group 3C2a2” (n = 9), “Group 3C2a3” (n = 59), “Group X” (n = 8), “WGX” (n = 1) and “WGY” (n = 2) (Figure 5.1). Based on this whole-genome phylogenetic tree, the viral genetic background was taken into account by grouping samples in phylogenetic groups since less diversity exists within these groups. The phylogenetic classification groups together samples with the same characteristic mutations for that particular phylogenetic group. These characteristic mutations make up the viral genetic background. To retain statistical power for the number of available samples (because too few samples were available for some phylogenetic groups to perform sound statistical inference), individual phylogenetic groups were combined based on an objective criterion by considering their sequence identity. The terms “clade” and “group” refer specifically to grouped samples using either the WHO/ECDC recommendations or our classification method, respectively. The exact sequence identity threshold was calculated from the WHO/ECDC clades, which include “Clade 3C2a1” (n = 170), “Clade 3C2a2” (n = 9) and “Clade 3C2a3” (n = 59). However, as only nine Belgian samples belonged to “Clade 3C2a 2”, these were grouped with “Clade 3C2a1” because the sequence identity showed that these ten samples are most similar to samples from “Clade 3C2a1” with a minimal sequence identity of 98.81% versus a sequence identity of 98.67% compared to “Clade 3C2a3”. The sequence identity between the concatenated genome sequences of all samples was calculated using the “Ident and Sim” tool [801]. A percent identity cut-off of 98.81% was selected for combining phylogenetic groups. This resulted in classifying the 253 samples into three groups (Figure 5.1). “Phylogenetic Group X” consisted of 190 samples from the following individual phylogenetic groups: “Group 3C2a1”, “Group 3C2a1(2)”, “Group 3C2a1a”, “Group 3C2a1a (2)”, “Group 3C2a1b”, “Group 3C2a2”, “Group X”, “WGX”, and “WGY”. The second group of 59 samples all belonged to the phylogenetic group “Group 3C2a3”. The third group of 4 samples all belonged to the phylogenetic group “Group 3C2a,” but were not retained for further analysis due to the limited number of samples. A list of the amino acid substitutions that were found in each sample of each respective group is provided in Supplementary File S5.1: AA_MUT_Phylo.

In order to compare results from the Belgian strains with the international context, a local Nextstrain instance [256], allowing light-weight phylogenomics, was built using the in-house sequenced samples complemented with GISAID sequences. Only samples that included the whole genome, patient sex, age information, and that were directly sequenced (i.e., no passaging in cells or eggs) were used, resulting in 14,157 samples (Supplementary File S5.2). All sequences were aligned with CLC Genomics Workbench 20.0.2 with default parameters and untranslated regions were stripped on both sides retaining only the protein-coding parts. Aligned segments were concatenated into a single sequence for all samples. Only sequences with less than three gaps and/or ‘N’ characters were retained (Supplementary Table S5.1). To create the local instance, the same steps were taken as described previously [792]. The Belgian samples were previously designated to their phylogenetic groups [792]. Finally, GISAID samples that clustered with these Belgian samples were assigned to the same phylogenetic group.

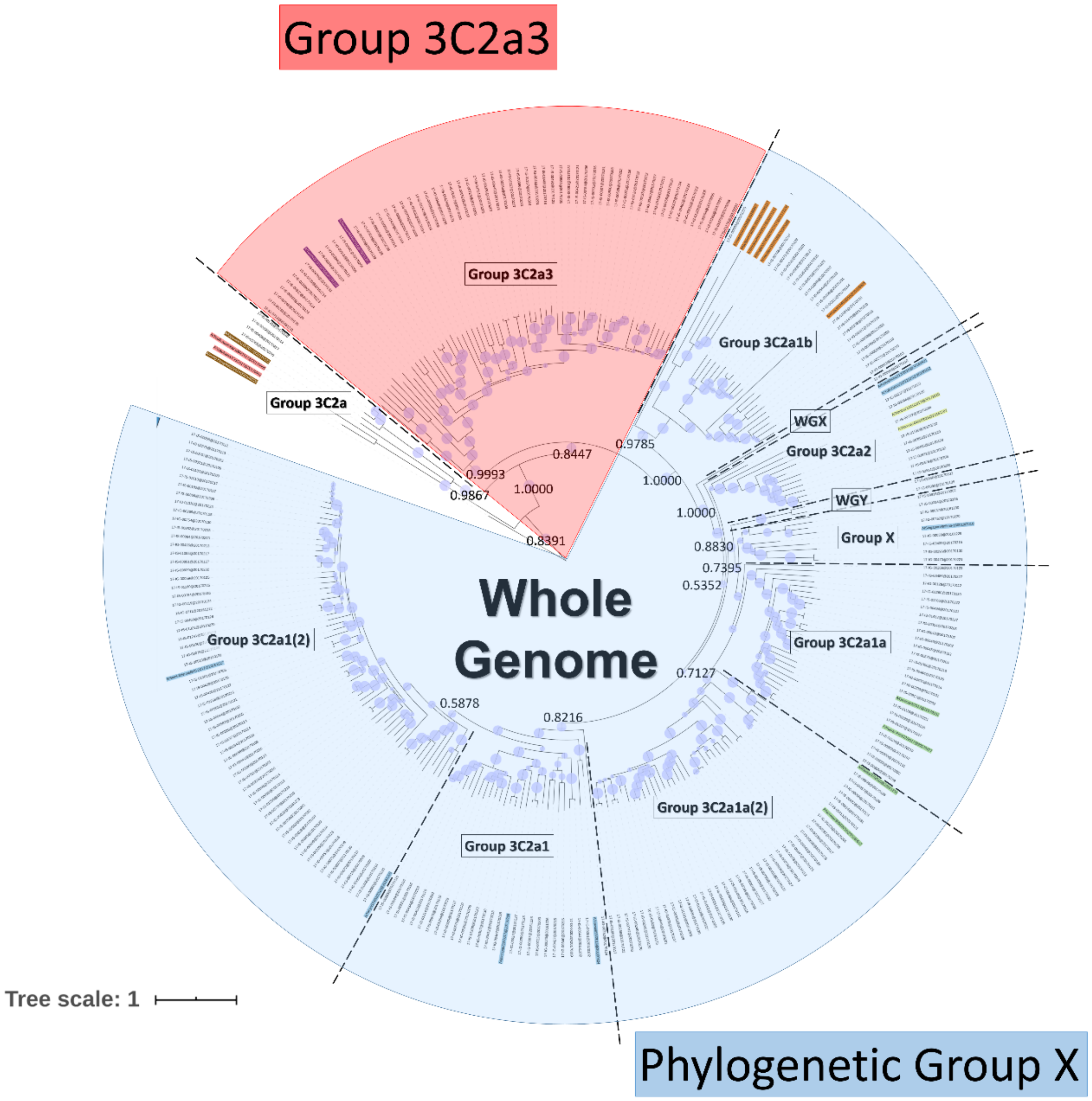


Figure 5.1: Phylogenetic tree based on the whole H3N2 genome. Within the tree, the group labels represent the phylogenetic groups that were assigned to their respective samples according to their classification based on references (colored names) and the support of nodes by posterior probability values. Posterior probability values are indicated on key nodes that separate phylogenetic groups. The size of blue disks on nodes represents the posterior probability scaled between 0.5 and 1. The scale bar represents the average number of substitutions per site. Samples belonging to ‘Phylogenetic Group X’ are indicated in blue and those belonging to Group 3C2a3 in red.

5.2.3. Inference of Associations with Patient Data

Statistical data analyses were performed using R-software (RStudio Version 1.0.153; R Version 3.6.1). For the “general approach”, the two-sided Fisher’s exact test was used to assess the association between the variables obtained from the clinical patient files and amino acid mutations from all samples that were identified in comparison to A/HongKong/4801/2014. These variables that were used in the two-sided Fisher’s exact test were obtained from the clinical patient files and include patient age (categorized into < 15, 15-59, and ≥ 60), sampling date, sex, vaccination status, the use of antibiotics, presence of comorbidities, disease severity (classified into mild, moderate, and severe). Disease severity is based on the one hand on the surveillance system: ILI (mild) and SARI (moderate + severe); and on the other hand on the absence (moderate) or presence (severe) of severity indicators. The distinction between moderate and severe among SARI patients is made based on the severity indicators. Disease severity indicators (death, stay in the ICU and advanced respiratory support) were also considered separately. Multiple testing correction was applied by employing the Benjamini-Hochberg method [774] and controlling the False Discovery Rate (FDR) at 5%. The variance inflation factor (VIF) was used to measure the amount of collinearity between mutations when inserted as a set of multiple regression variables.

Because of the overall high genetic diversity within the sequenced samples [792], the importance of the viral genetic background was explored. For this “approach considering the viral genetic background”, the same statistical analysis was conducted to detect mutations linked to the patient data within the two groups, i.e., the previously described “Phylogenetic Group X” and “Group 3C2a3.” For statistically significant associations identified during the univariate analysis, generalized linear regression with a binomial family distribution was used to identify confounding mutations and evaluate the effect modification. Confounding factors were identified by adding potential risk factors (other patient data) to the model. The effect modification was evaluated by adding interaction terms to the model. Additionally, the effect size was defined as an odds ratio as estimated by using a logistic regression analysis for the association in question.

The GISAID database contains information on patient sex and age, and sampling date. Strains from the WHO-defined clade “3C3a” were excluded from this analysis based on the custom-built Nextstrain instance of 10,583 samples because these strains were genetically too distant from the Belgian sequenced samples (Supplementary Figure S5.1 and Supplementary File S5.1: GISAID_Samples). These samples were used for the general approach as well as for the approach considering the viral background, which resulted in 8,796 samples belonging to “Phylogenetic Group X” and 831 samples belonging to “Group 3C2a3”. Samples collected between April 2016 and September 2017 were attributed to three “period groups” and these period groups each included samples from both the Northern and Southern hemispheres. The first period comprised the end of the 2015-2016 Northern hemisphere influenza season in April 2016 until week 45 of 2016 (453 samples). The second period comprised samples until week 16 of 2017, when less than 10% of lab tests were positive for influenza in Europe according to ECDC [802] (2,517 samples). The third period comprised samples until the end of September 2017 (723 samples). The first and third periods were predominated by samples from the Southern hemisphere, while the second period was predominated by samples from the Northern hemisphere, corresponding with their respective flu seasons. Because the number of genome sequences per period group varied, permutation analyses were performed to correct for sample size. In total, 1,000 subsets of 430 genome sequences were randomly selected for every period group by sampling with replacement (Figure 5.1 and Supplementary Figure S5.3) using a sample size of 95% of the smallest group. It was then assessed if the same significantly associated trends could be distinguished compared to the trends observed within the Belgian samples by performing a two-sided permutation test at a significance level of 0.05.

Finally, it was evaluated whether the mortality was significantly higher within the group of patients suffering from renal insufficiency using a Chi-square Goodness of Fit Test. All analysis scripts and results are provided in Supplementary File S5.3. Input files and detailed results are provided in Supplementary File S5.1.

5.3. Results

5.3.1. Significant Associations Between Viral Mutations and Patient Data

The aim was to identify potential associations between specific amino acid mutations in the influenza genome and available patient data in a cohort of 253 influenza patients (Table 5.1). Table 5.2 lists an overview of the identified statistically significant associations and previously described effects of these mutations reported in literature. Significant associations were detected between specific mutations and the sampling period, and the patient sex (Supplementary File S5.1: AAMut Fisher + FDR). Nine mutations were linked to their sampling period and their presence in the circulating strains significantly varied (two-sided Fisher’s exact test with FDR correction of 5%) during the season with an effect size ranging from 1.1 to 11.1 (Figure 5.2). The mutations PB2-V255I (adjusted P = 0.05), HA-S144K (adjusted P = 0.05), NA-G93D (adjusted P = 0.05), NA-P468L (adjusted P = 0.05), NS1-S99T (adjusted P = 0.05), and NS1-L146S (adjusted P = 0.05) emerged over time, whereas the mutations PB1-G216S (adjusted P = 0.05), PB1-I517V (adjusted P = 0.05), and NA-P468H (adjusted P = 0.05) decreased. At position 468 in the NA segment, the mutations NA-P468L and NA-P468H emerged and decreased, respectively, throughout the season. The VIF-analysis demonstrated that viruses containing the PB1-G216S mutation, often co-occurred with the PB1-I517V mutation. These mutations were most often observed in "Group 3C2a1a(2)" (Supplementary Figure S5.4). Additionally, samples containing the PB2-V255I mutation often possessed the other emerging mutations (HA-S144K, NA-G93D, NA-P468L, NS1-S99T, and NS1-L146S). These mutations were most often observed in "Group 3C2a3" (Supplementary Figure S5.3). For these associations, particular confounding factors existed, i.e., other variables influencing the correlation between the mutation and the patient data, that could not be excluded. These factors included vaccination status, antibiotics use, surveillance system, and/or stay in the ICU (Supplementary File S5.1: Effect Size). Ten mutations were observed to be significantly more present in either male or female patients (Supplementary Figure S5.2) (detailed results in Supplementary File S5.1: AAMut Fisher + FDR).

Table 5.2: Statically significant associations found between patient data and amino acid mutations in the whole genome. Functional sites and properties of amino acid changes are also presented. Volume categories for size are divided in "very small" [60-90 A^3], "small" [108-117 A^3], "medium" [138-154 A^3], "large" [162-174 A^3] and "very large" [189-228 A^3]. Finally, the description of the mutation was included if available in the literature. All five mutations related to renal insufficiency when considering the viral genetic background were found within "Phylogenetic Group X". "Phylogenetic Group X" includes "Group 3C2a1", "Group 3C2a1(2)", "Group 3C2a1a", "Group 3C2a1a (2)", "Group 3C2a1b", "Group 3C2a2", "Group X", "WGX" and "WGY".

AA substitution	Functional site	Amino acid properties	Previous descriptions	Citations
General approach				
Sampling period				
PB2-V255I	NP binding site [803]	Size Medium → Large	- Association between this mutation and patients that were not vaccinated.	[754]
			- No association with a significant change in pathogenicity in A(H1N1) and A(H3N2).	[804, 805]
			- Increase in pathogenicity due to this mutation in combination with seven other residues (H15R, N23S, T27I, K53R, L58S, R75H, H75L) in A(H1N1)	[806].
HA-S144K	Receptor-Binding domain [807]	Charge Neutral → Basic	- Association between this mutation and patients that were not vaccinated.	[754]
	Epitope region A [808]	Size Very small → Large	- Link with low vaccine effectiveness.	[809]

NA-G93D	Head: Enzyme active site and calcium binding domain, which stabilises the enzyme structure at low pH values [810–815]	Polarity Non-polar → Polar Charge Neutral → Acidic Size Very small → Small	- Association between this mutation and patients that were not vaccinated.	[754]
NA-P468L	Head: Enzyme active site and calcium binding domain, which stabilises the enzyme structure at low pH values [810–815]	Size Small → Large	- No studies were found.	
NS1-S99T	Effector domain [67]	Size Very small → Small	- Association between this mutation and patients that were not vaccinated.	[754]
NS1-L146S	Nuclear export signal [67]	Polarity Non-polar → Polar Hydropathy Hydrophobic → Hydrophilic Size Large → Very small	- Association between this mutation and patients that were not vaccinated.	[754]
PB1-G216S	Nuclear Localization Signal [816]	Polarity Non-polar → Polar	- A(H1N1) viruses with PB1-216G have an increased adaptability and enhancement of viral epidemiological fitness, probably due to a low-fidelity replicase. PB1-216S viruses showed a higher pathogenicity in mice in comparison to PB1-216G viruses and PB1-216S viruses had a lower mutation potential.	
PB1-I517V	Not described	Size Large → Medium	- This position in the H3N8 virus was identified as undergoing changes due to selective pressure during host shifts from birds to humans. - This mutation in a A(H1N1)pdm09 viral background was discovered in a highly complementary region between PB1 and HA and leads to an enhancement of the complementarity and consequently better binding. - In the mammalian host due to a more restricted conformation, this apparent neutral mutation is located near conserved motifs that are responsible for protein folding and this effect suggests that the mutation leads to a better compatibility with H1 in the human host [853](Nilsson, 2017).	[817, 818]

NA-P468H	Head: Enzyme active site and calcium binding domain, which stabilises the enzyme structure at low pH values [810–815]	Polarity Non-polar → Polar Charge Neutral → Positive Hydropathy Hydrophobic → Hydrophilic Size Small → Medium	- Association between this mutation and patients that were vaccinated.	[754]
			- It was demonstrated that P468H has become fixed in A(H3N2) viruses circulating since 2016. This mutation contributed to NA antigenic drift in relation to the vaccine strain Hong Kong/4801/2014. There is further research needed to understand the role of the mutation, because residue 468 is not essential for binding antibodies.	[819]
Approach considering THE VIRAL BACKGROUND				
Renal Insufficiency				
PB2-R299K	Not described	Not applicable	- It was demonstrated in A(H1N1)pdm09-infected mice that K299 is conserved, which raises the possibility that it plays some role in the adaptation to the mammalian host and might also link to the heterogeneity in A(H1N1)pdm09.	[820]
			- It has been observed that eleven amino acid mutations, including PB2-R299K, in A(H3N2) occurred between the influenza virus strains in the 2016-2017 winter season and 2017 summer season. These mutations were correlated to temperature sensitivity and viral replication, because the 2016-2017 winter season viruses were significantly restricted at 39°C. Although this mutation was identified, it had little influence on the polymerase activity at different temperatures.	[821]
PB2-K340R	Cap binding [822, 823]	Conservative	- PB2-K340R was introduced in a PR8-derived recombinant virus A(H1N1) and there was no significant increase in polymerase activity.	[824]
			- It has been observed that eleven amino acid mutations, including PB2-K340R, in A(H3N2) occurred between the influenza virus strains in the 2016-2017 winter season and 2017 summer season. These mutations were correlated to temperature sensitivity and viral replication, because the 2016-2017 winter season viruses were significantly restricted at 39°C. Although this mutation was identified, it had little influence on the polymerase activity at different temperatures.	[821]
HA-K92R	Epitope Region E [825]	Conservative	- This mutation was confirmed in this study as specific for the HA cluster 3C2a1b.	[769]
HA-H311Q	Epitope region C [826]	Charge Positive → Neutral	- This mutation was confirmed in this study as specific for the HA cluster 3C2a1b.	[769]
NP-V197I	Cytotoxic T lymphocyte (CTL) epitopes [827]	Size Medium → Large	- This mutation in a A(H3N2) virus is located in known virus CTL epitopes and they may confer a higher efficiency of escape from CTL-mediated immune responses.	[827]

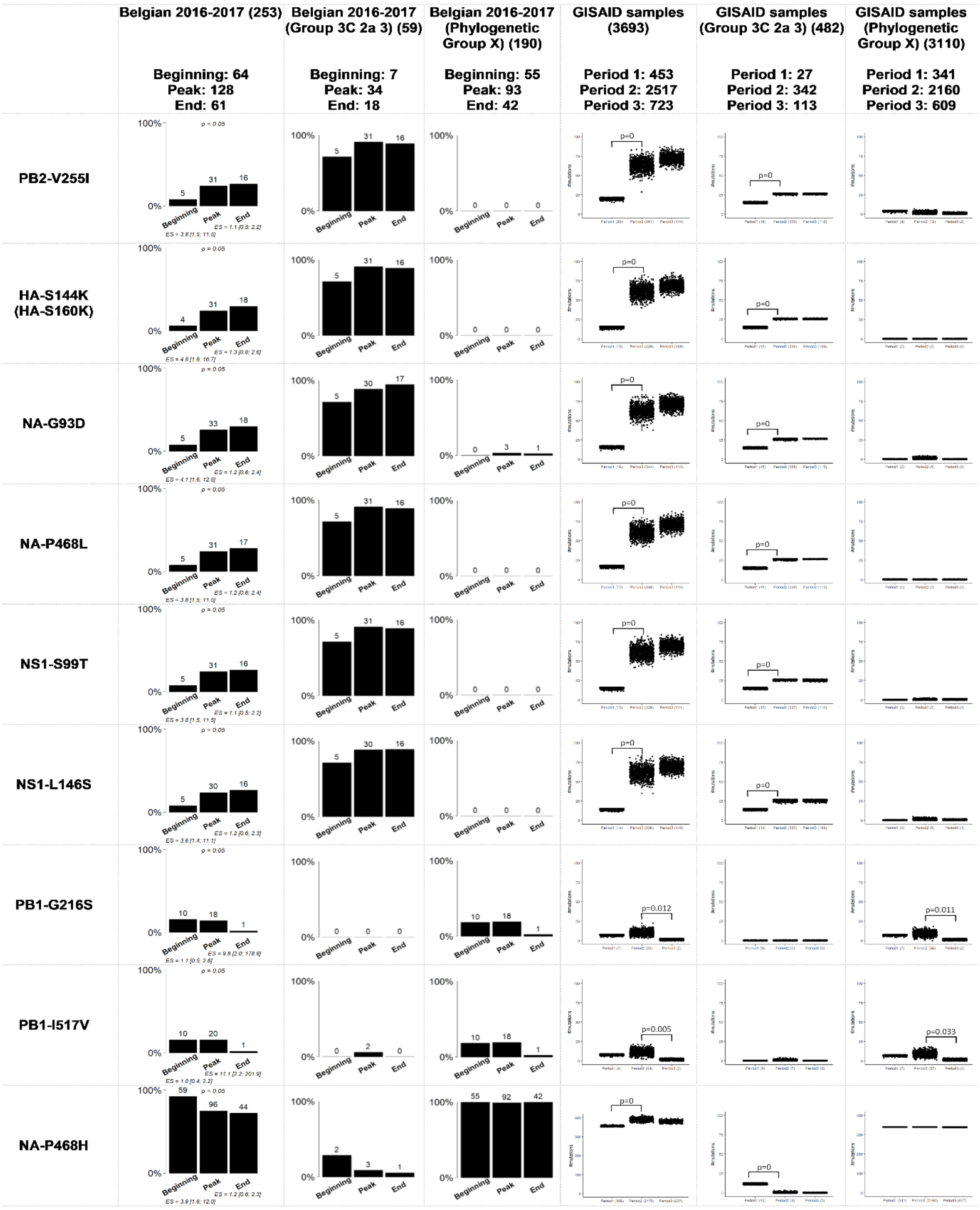


Figure 5.2: Comparison of the Belgian influenza samples with samples from the GISAID database for mutations that were considered significantly related to the sampling period. The distribution of samples in the groups “Group 3C 2a 3” and “Phylogenetic Group X” is provided for the significant results after running the Fisher’s exact test with FDR correction when the viral genetic background is not taken into account. “Phylogenetic Group X” includes “Group 3C2a1”, “Group 3C2a1(2)”, “Group 3C2a1a”, “Group 3C2a1a (2)”, “Group 3C2a1b”, “Group 3C2a2”, “Group X”, “WGX” and “WGY”. In the graphs representing the situation in Belgium, above the bars the number of samples that had this mutation are indicated. In the graphs representing the samples from GISAID, the number of samples that possessed this mutation are indicated below the chart. The magnitude of the significant association is defined by the effect size (ES) and its confidence interval. The resulting p-value of the Fisher’s exact test with FDR correction for the samples from the Belgian dataset is indicated above bar charts for which significant associations were found. The p-values of the permutation tests performed for the GISAID samples are indicated above the boxplots.

5.3.2. Significant Associations Between Mutations and Patient Data When Samples Are Stratified According to the Phylogenetic Clade

It was previously shown that similar mutations can affect viral genes in different and sometimes even contradictory ways [828, 829]. These variations can possibly be attributed to potential effects related to the highly diverse genetic background of the A (H3N2) subtype [796]. The viral background was taken into account while using phylogenetic classification based on a whole-genome tree [792]. The 253 samples were classified into three groups (Figure 5.1). "Phylogenetic Group X" (n = 190), "Group 3C2a3" (n = 59) and "Group 3C2a" (n = 4), but the latter was not retained for further analysis due to the limited number of samples. We compared the previously found associations using the general approach in this study (5.3.1) within "Phylogenetic Group X" and "Group 3C2a3" separately, by taking the viral genetic background in account based on the phylogenetic groups.

Associations between the previously identified mutations and the sampling period (Figure 5.2), that were significant using the general approach (5.3.1), presented similar trends but were no longer statistically significant. For the mutations found to be significantly more present in male or female patients, the same trends were not observed within the groups when the viral genetic background was considered (Supplementary Figure S5.2). Importantly, unequal distribution between male and female patients was observed [792] as each of these mutations was observed almost exclusively in either "Phylogenetic Group X" (Female = 102; Male = 81) or "Group 3C2a3" (Female = 18; Male = 40). Additionally, five mutations within "Phylogenetic Group X" were significantly associated with renal insufficiency (Supplementary File S5.1: AAMut Fisher + FDR (PHYLOX)) (two-sided Fisher's exact test with FDR correction of 5%). Table 5.2 presents an overview of these mutations and their previously described effects in the literature. PB2-R299K (adjusted P = 0.03), was significantly more present in samples from patients without renal insufficiency. PB2-K340R (adjusted P = 0.03), HA-K92R (adjusted P = 0.03), HA-H311Q (adjusted P = 0.03), and NP-V197I (adjusted P = 0.03), were significantly more detected in patients suffering from renal insufficiency (Figure 5.3). The VIF-analysis demonstrated samples containing the PB2-K340R mutation, often co-occurred with the HA-K92R, HA-H311Q, and NP-V197I mutations. Most of these mutations are observed within "Group 3C2a1b" (Supplementary Figure S5.6 and Supplementary File S5.1: Mutations per group). To demonstrate the limited effect of reassortment on the associations with renal insufficiency, the same analysis, namely a two-sided Fisher's exact test with FDR correction (5%), was performed for each segment tree. In most cases, the associations related to the renal insufficiency remained significant, suggesting that these associations were not related to reassortment (Supplementary File S5.1: Segment Renal).

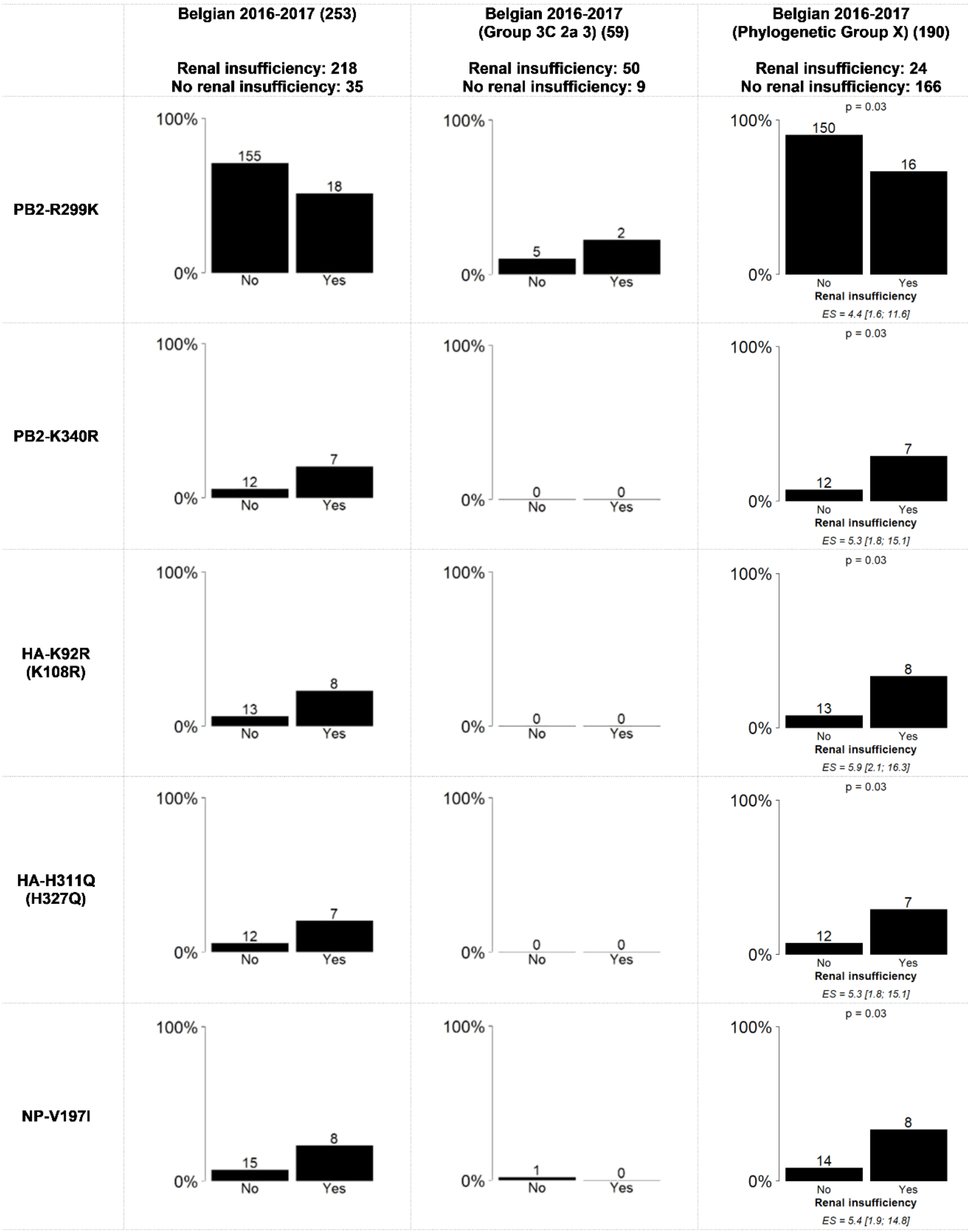


Figure 5.3: Statically significant results using the Fisher's exact test with FDR correction for the association between renal insufficiency and amino acid mutations in the whole genome from all of the samples and "Phylogenetic Group X" and "Group 3C2a3". "Phylogenetic Group X" includes "Group 3C2a1", "Group 3C2a1(2)", "Group 3C2a1a", "Group 3C2a1a (2)", "Group 3C2a1b", "Group 3C2a2", "Group X", "WGX" and "WGY". The bar graphs represent the percentage of samples per variable of the patient data that have the mutation. On top of the bars the number of samples that had this mutation are indicated. The magnitude of the significant association is defined by the effect size (ES) and its confidence interval. The resulting p-value of the Fisher's exact test with FDR correction is indicated above bar charts for which significant associations were found.

Although the exact stage of chronic renal insufficiency was not specified in our dataset and most patients suffered from other chronic diseases and/or were elderly, seven out of 35 patients with this condition did not survive, which is significantly more than in the total dataset (Chi-square Goodness of Fit Test, $p = 0.005$). For more detailed results, see the Supplementary File S5.1.

5.3.3. Evaluation of Significantly Associated Mutations From the Belgian Samples in an International Context

To evaluate significant associations observed in the Belgian dataset in a more global context, the Belgian samples were supplemented with samples from the same subtype for which patient information was available in the GISAID database (patient age, sex, and sampling date). The significant associations observed in the Belgian study were compared to the results in the GISAID database for both the general approach and the one considering the viral genetic background. Although some bias may be introduced by (i) a different selection criterion to choose the isolates to sequence by the different laboratories; (ii) different sampling population (patient) sizes, our observations are the following:

Regarding the sampling date (Figure 5.2), all of the mutations related to the sampling period, except for NA-P468H, showed the same significantly associated trend over time as observed in the Belgian study. Additionally, the observed trends when considering the viral genetic background were significant in contrast to the Belgian samples, probably due to the increase in sample size. It is therefore possible to partially extrapolate the results of the Belgian study to a global level.

Regarding patient sex, the samples from the GISAID database did not follow the same trends as the Belgian influenza samples for both the general approach and the one considering the viral genetic background (Supplementary Figure S5.2). It should be noted that in contrast to the Belgian influenza samples, the number of male and female patients extracted from the GISAID database, was more equally distributed across the groups, namely “Phylogenetic Group X” (Male = 4156; Female = 4639) or “Group 3C2a3” (Male = 431; Female = 397).

5.4. Discussion

Influenza surveillance is the basis for determining the seasonal influenza vaccine composition. Current conventional influenza vaccines are still largely based on technology from the 1940s relying on the replication of influenza in embryonated eggs and focuses on the HA segment [830]. However, next-generation vaccines also focus on other parts of the genome, consequently to track mutations across the whole genome becomes important for such vaccine candidates. Moreover, WGS enables the detection of mutations across all eight segments of the influenza genome, allowing the evaluation of associations between the patient data and mutations located on the whole viral genome instead of solely the HA segment. It is important for the influenza surveillance to provide information to national influenza prevention and control programs about the severity, impact, and timing of seasonal epidemics.

In this study, mutations were identified using WGS data of influenza A (H3N2) samples collected in the context of the influenza surveillance in Belgium. They were used in order to explore potential associations between mutations positioned across the whole genome and patient characteristics as well as other metadata. Due to the limited number of samples, that is often the case for national surveillance, it was verified, when it was possible, whether the observations at the Belgian level correspond to trends at an international level using the GISAID database. For example, significant increase or decrease over the sampling periods was observed for nine mutations located across the A(H3N2) genome during the Belgian 2016-2017 influenza season. Comparison with the GISAID database showed the same significantly associated trends worldwide for these mutations, except for NA-P468H. These mutations can probably be attributed to the fast evolutionary dynamics of influenza A (H3N2) [757]. Throughout the outbreak season, it is relevant to follow trends of emerging and disappearing mutations over the whole genome with respect to the vaccine strain, as these mutations may lead to antigenic drift from the vaccine strain. Currently, only the HA and NA segments are updated in the vaccine, the HA and NA mutations and their evolution over time should therefore be considered for the vaccine composition for the next influenza season. The vaccine strain of the 2017-2018 influenza season, which is also A/Hong Kong/4801/2014 [831], and subsequent years did not take these into account, which can be a partial explanation for the observed low vaccine efficacy in that season. The importance of following the emergence or decrease of mutations with respect to selecting the appropriate vaccine strain can be illustrated by the HA-S144K mutation, which significantly increased during the Belgian 2016-2017 influenza season. HA-S144K together with HA-N121K and HA-T135K were previously associated with outbreaks in the Northern hemisphere and suboptimal vaccine effectiveness [809, 832–836]. Noteworthy, although not significant (potentially due to the limited number of samples), HA-N121K and HA-T135K also increased during the influenza season in the Belgian surveillance (results not shown).

The substantial diversity observed within the patient derived A(H3N2) isolates during the Belgian 2016-2017 season [792] offered the opportunity to explore whether considering the viral genetic background by stratifying the samples according the phylogeny has an effect on the detection of new associations. Importantly, when stratifying the sample according the phylogeny, we discovered associations between the occurrence of certain mutations in the A

(H3N2) viruses and patient data. In our Belgian study, in particular associations related to renal insufficiency were detected after genetic stratification. In addition to severity indicators, it could be relevant for patient management to explore associations with comorbidities, including renal insufficiency. Influenza contributes to higher mortality in patients suffering from End-Stage Renal Disease, which is the last stage of chronic renal insufficiency or chronic kidney disease [837]. Patients with chronic renal insufficiency also often suffer from other diagnosed or undiagnosed risk factors possibly resulting in a poor outcome when infected with the influenza virus [838–840]. Four and one mutation(s) were detected to be significantly more likely to be present and absent in patients suffering from renal insufficiency, respectively, both within “Phylogenetic Group X”. It could be speculated that these mutations associated with renal insufficiency could be the result of a weakened immune system. To support this finding, it should be emphasized that the nasopharyngeal swabs that were obtained from the 35 patients with renal insufficiency were taken in hospitals across Belgium, and were not restricted to one of the sampling periods that were defined in this study (beginning of the influenza season, peak of the influenza season and end of the influenza season), and no evidence of epidemiological linkage could be found (Supplementary File S5.1: Metadata).

However, we cannot exclude other confounding factors that lead to these associations. Except for the vaccination status, if available, the immune status of the patient was not included in the analysis as this information was not available. These associations related to renal insufficiency could unfortunately not be confirmed with a larger number of samples from the GISAID database because this database does not contain this type of patient information. Regarding the results of this proof of concept study, it is important that the occurrence of these mutations be examined in the following years from the surveillance system in Belgium and other countries to learn if these associations can be confirmed.

A collection containing a small number of samples like in this study has limitations. Stratification according to the phylogeny has the inconvenience to further reduce this number. Indeed, on the one hand the significance of some associations obtained within the larger group may disappear due to a lack of power for the statistical analysis. On the other hand, the small number of samples and the multivariable analysis may introduce bias leading to a “false” association. Therefore, it is advised to have a confirmation with a larger dataset if possible. In this study, this was illustrated by the fact that the significant associations related to the sampling period, observed without stratification, was following the same trends but did not result in significant associations anymore when considering the viral background. This is probably due to the limited number of samples, which was confirmed using a larger dataset of GISAID while considering the viral genetic background. Using the GISAID dataset, the associations regarding the sampling period became significant for the same groups and trends.

Another limitation was related to the ten mutations identified to be significantly related to the sex of the patient with the general approach. However, this trend was not confirmed when using the GISAID database. In fact, conflicting results (opposite trends) were observed in comparison to the general approach for the mutations related to sex when the viral genetic background was considered. These results considering the viral genetic background could also not be confirmed using the sequences from the GISAID database for both the general approach and the approach taking into consideration the viral genetic background. A probable cause of this inconsistency between the general approach, the approach considering the viral genetic background and the GISAID database could be the unequal distribution of male and female patients in the Belgian dataset over the different phylogenetic groups causing a gender sampling bias. In Supplementary Figure S5.2, the unequal distribution is explained more in detail. In this study, the number of samples was limited to 253 samples due to the current infrastructure and cost of sequencing.

This proof of concept study highlights the power that WGS sequencing of influenza may offer especially when using a stratification taking into account the viral genetic background. It shows also the limitation of analysing a small number of samples. However, such size is a reality for several countries as it is already challenging to acquire the necessary funds to simply switch from Sanger sequencing the HA and NA segments to WGS in routine surveillance. Therefore, it would be of great benefit to perform such type of analysis at a European or international level using more samples to reduce the effect of sampling bias and to have more statistical power to find other associations.

The analysis of the GISAID database in this study has demonstrated that using a larger dataset could help to confirm the trends observed with a relatively limited number of samples within countries like Belgium. Also, when using genomic data, large sample sizes are needed because many mutations were included in the analysis, leading to a reduction of the statistical power due to multiple testing correction. This may be particularly crucial when using the approach taking into account the viral genetic background and working with smaller groups. In this context, a large database available to the scientific community containing genomic data with a larger set of patient data is important to be constructed. This is currently only in place to a limited extent. The WHO maintains a list of mutations linked to resistance of neuraminidase inhibitors [841]. FluSurver is an application utilizing an in-house database of curated literature annotations for mutational effects associated with antibody escape, antigenic drift, host receptor specificity, and drug resistance. Broadening the scope of such resources could allow exploring associations between particular mutations and (other) phenotypic effects and patient data [842]. GISAID maintains sequence data worldwide and could be useful to investigate if the effects of particular mutations have also been observed in other genome sequences sampled at other geographical locations and during other influenza seasons. However, available patient information in GISAID is currently mostly limited to the age and sex of the patient and the sampling date [752, 843]. In addition, the GISAID licenses should be less restrictive, allowing their data to be used more easily. The construction and the use of a database with a large dataset coming from samples selected and sequenced by different laboratories and different countries is also challenging. Indeed, this implies the need for having a common, standardised approach to collect and manage data within different laboratories or at least to provide a detailed description of the methodology used to collect the sample and the patient data in order to avoid potential bias which could result in erroneous conclusions. For example, although SARS-CoV-2 was the most sequenced virus, due to the lack of harmonisation between countries it remains difficult to draw conclusions whether certain SARS-CoV-2 mutations are related to disease severity, vaccination or other patient data mutations related to the season. Calling for a new approach to data management could enable faster solutions and improve the worldwide response of the scientific community resulting in a better surveillance.

In conclusion, the results of this study, identifying associations between the patient data and viral mutations that were not only present in the HA segment, highlight the importance of tracking mutations across the entire influenza genome. Furthermore, this study is used as a proof of concept to demonstrate how to work with real-world data coming from National Reference Centers when WGS is implemented in routine surveillance. In addition to disease severity and vaccination status, other patient data was included in this study such as age, severity indicators (stay in the ICU, death, need for invasive respiratory support, ARDS and ECMO), comorbidities (renal insufficiency, cardiac, neuromuscular and respiratory diseases, hepatic insufficiency, diabetes, and, immunodeficiency) and sampling date. This study detected associations between particular mutations and the sampling period that can be important to take into account for vaccine strain selection and clinical management of infected patients. Moreover, this study investigated the possible effect of the viral genetic background on the association between mutations and patient data and proposed a new approach based on stratification using phylogenetic groups. Using this approach, five additional mutations significantly associated with renal insufficiency were detected, indicating the potential or even necessity to take the viral genetic background of the virus into account by considering its phylogeny. Therefore, the viral genetic background could play an important role in inferring associations between genomic and patient data.

5.5. Acknowledgments

NGS was performed at the service Transversal and Applied Genomics at Sciensano.

CHAPTER 6: GENERAL APPROACH TO IDENTIFY LOW-FREQUENCY VARIANTS WITHIN ROUTINE INFLUENZA SURVEILLANCE

Context of this chapter:

An advantage of using WGS in routine surveillance is the opportunity to sequence a patient-derived virus population at sufficient depths to identify low-frequency variants (LFV) present in a quasispecies. The current focus of routine surveillance on mutations in the consensus genome may not always provide sufficient information to investigate transmission, pathogenicity, virus evolution, drug and vaccine resistant strains, and could benefit from approaches that also consider intra-host genetic diversity. However, many challenges remain for the reliable detection of low-frequency variants, mainly due to experimental errors introduced during sample preparation and sequencing. In this chapter, we propose a generally applicable approach to identify low-frequency variants that remains feasible in routine surveillance while ensuring high-quality results by limiting false positive observations. This approach was applied on the same (H3N2) influenza dataset of clinical samples as used in Chapter 4 and 5 to explore the intra-host genetic diversity. Finally, as a proof of concept, the potential clinical relevance of considering low-frequency variants in routine influenza monitoring was evaluated by assessing associations between patient data and intra-host influenza A virus sequence diversity.

This chapter was previously published:

Laura A. E. Van Poelvoorde, Thomas Delcourt, Marnik Vuylsteke, Sigrid C. J. De Keersmaecker, Isabelle Thomas, Steven Van Gucht, Xavier Saelens, Nancy Roosens, Kevin Vanneste (2022) A General Approach to Identify Low-Frequency Variants within Influenza Samples collected during Routine Surveillance. Microbial Genomics. **8**. https://doi.org/10.1099/mgen.0.000867

Authors' contributions:

Conceptualization: NR, SDK, KV, XS, SVG; Project Administration: NR; Data Curation: LVP, TD, KV, IT; Methodology: LVP, TD, NR, KV; Software: TD, LVP; Formal Analysis: LVP, TD, MV; Validation: TD, LVP, KV; Investigation: LVP, TD; Visualisation: LVP, TD; Writing – Original Draft Preparation: LVP, TD, NR, KV; Writing – Review & Editing: all authors; Funding Acquisition: NR; Resources: SDK, SVG, IT; Supervision: NR, KV.

Abstract:

Influenza viruses exhibit considerable diversity between hosts. Additionally, different quasispecies can be found within the same host. High-throughput sequencing technologies can be used to sequence a patient-derived virus population at sufficient depths to identify low-frequency variants (LFV) present in a quasispecies, but many challenges remain for reliable LFV detection because of experimental errors introduced during sample preparation and sequencing. High genomic copy numbers and extensive sequencing depths are required to differentiate false positive from real LFV, especially at low allelic frequencies (AFs). This study proposes a general approach for identifying LFV in patient-derived samples obtained during routine surveillance. Firstly, validated thresholds were determined for LFV detection, whilst balancing both the cost and feasibility of reliable LFV detection in clinical samples. Using a genetically well-defined population of influenza A viruses, thresholds of at least 10^4 genomes per microliter and AF of ≥5 % were established as detection limits. Secondly, a subset of 59 retained influenza A(H3N2) samples from the 2016-2017 Belgian influenza season was composed. Thirdly, as a proof of concept for the added value of LFV for routine influenza monitoring, potential associations between patient data and whole genome sequencing data were investigated. A significant association was found between a high prevalence of LFV and disease severity. This study provides a general methodology for influenza LFV detection, which can also be adopted by other national influenza reference centres and for other viruses such as SARS-CoV-2. Additionally, this

study suggests that the current relevance of LFV for routine influenza surveillance programmes might be undervalued.

6.1. Introduction

Influenza is a very contagious respiratory tract infection in humans, mainly caused by the Influenza A and B virus. Both the Influenza A and B genomes consist of eight segments, including the hemagglutinin (HA) and neuraminidase (NA) segments. Due to their location on the viral envelope, the proteins encoded by the HA and NA segments represent key viral antigens and are the principal targets of the humoral immune response of the host [739–741]. A(H1N1) and A(H3N2) are the two principal Influenza A subtypes that circulate in humans [8].

Influenza viruses have a low-fidelity RNA polymerase that lacks proof-reading functionality. This results in a relatively high mutation rate during viral replication [116]. Replicating influenza within a host does therefore not give rise to genetically identical progeny viruses but rather to ‘quasispecies’, i.e. closely-related viruses that differ by at least one nucleotide from each other. Viral quasispecies are defined as a population of closely-related, non-identical viral genomes in a dynamic host environment that is continuously subjected to competition and selection [137, 844, 845]. Although considerable risk exists for producing defective progeny viruses due to the low-fidelity RNA polymerase, this also provides a major opportunity for the virus to rapidly evolve and escape from neutralising antibodies [846], antiviral drugs [847] and cytotoxic T-cells [848].

The availability and cost-effectiveness of high-throughput sequencing (HTS) technologies have led to their increased use in routine influenza surveillance [749]. HTS allows to determine the sequences of all eight influenza virus segments simultaneously, which offers the opportunity to better understand between- and within-host genetic diversity [791]. Genetic surveillance of influenza virus in biological samples is currently focused on monitoring mutations that are linked to antiviral resistance [849, 850], and antigenic mutations that are relevant for selecting vaccine strains [851]. Studies examining influenza pathogenesis should consequently consider virological and immunological parameters associated to severity as a whole [754]. When investigating virus evolution, transmission, drug and vaccine resistant strains, and pathogenicity, it may not always be sufficient to only examine the consensus genome sequence. Therefore, the current focus is shifting to also include quasispecies while studying genetic diversity [852, 853]. During infection, a particular variant within a quasispecies can by chance obtain a competitive advantage over other variants [854]. This can result in positive selection, and thus an increased frequency of such a variant over time within the patient [855]. However, the spread to other hosts is limited to a small fraction of the quasispecies population and even fewer become fixed in the global viral population [845, 856]. Positive selection of specific quasispecies in hosts has thus far only been observed during long-term infection of immunocompromised patients [857] and in extreme cases of drug resistance [735, 858, 859] for the HA and NA genes.

Several recent studies have successfully identified genetic variation in viral quasispecies during clinical influenza infections using deep sequencing with HTS [663, 857, 860–862]. Deep sequencing allows higher genome coverages, and consequently more reliable estimation of the diversity within the quasispecies population present at very low abundances [710]. Apart from the increased experimental costs associated with the use of HTS, many challenges remain to detect low-frequency variants (LFV, i.e. defined as nucleotides differing from the consensus sequence at low allelic frequency at a specific genomic position), including high-quality sequencing reads to ensure that insertions and deletions (indels), and single nucleotide variants (SNVs), can be called confidently. Current variant-calling algorithms for identifying LFV are based on read quality, mapping quality, strand bias, base quality and sequence context [663]. Variants are typically accepted only when their allelic frequency (AF) exceeds the expected sequencing error rate. Several variant-calling methods have been used in multiple HTS-based studies of viral diversity [754, 858, 863]. However, these methods have not always been benchmarked against predefined viral populations, rendering their accuracy for detecting LFV largely unknown. Moreover, not only the bioinformatics approach but also the laboratory process can influence LFV detection. Experimental errors can be introduced during sample preparation, including reverse transcription and PCR amplification, and during sequencing itself [864]. The genome copy number and viral load of samples in particular affect the specificity and sensitivity of variant detection substantially, resulting in more false positive (FP) variant detections for samples with a low concentration due to propagating PCR-amplification errors [663].

In this study, we first established an approach for the quantification of low-frequency variants within influenza samples by using a genetically well-defined population of Influenza A viruses. Thresholds for LFV detection based on HTS with the Illumina technology were validated whilst ensuring that this approach remains powerful enough but also economically feasible in routine surveillance. Secondly, this approach was used to evaluate the prevalence of LFV of influenza A(H3N2) viruses recovered from the Belgian national influenza surveillance network during the 2016-2017 season, demonstrating that several LFV were identified in clinical samples. Finally, potential associations between within-host diversity and patient data were investigated as a proof of concept for the potential relevance of LFV in routine influenza monitoring.

6.2. Methods

6.2.1. Viruses and cells

A reverse genetics system of Influenza A/Bretagne/7608/2009 (A(H1N1)pdm09) and Influenza A/Centre/1003/2012 (A(H3N2)) in a bi-directional pRF483 plasmid were provided by Institute Pasteur Paris, France. Influenza viruses with a point mutation in the NA segments were obtained by reverse genetics using the QuikChange II Site-Directed Mutagenesis Kit (Agilent Technologies) and GeneJET Plasmid Miniprep Kit (Thermo Fischer) according to the manufacturer's instructions. For A/Bretagne/7608/2009, the NA-H275Y mutation (CAC → TAT) was introduced (consisting out of two nucleotide mutations). For A/Centre/1003/2012, NA-E119V (GAA → GTA) was introduced (consisting of one nucleotide mutation). The NA-plasmids were verified using Sanger sequencing on an Applied Biosystems Genetic Analyzer 3500 using the Big Dye Terminator Kit v3.1 following the manufacturer's instructions using primers described in Supplementary Table S6.1.

A co-culture of Madin-Darby canine kidney (MDCK) cells and 293T cells was maintained in Dulbecco's modified Eagle medium (DMEM) (Gibco) and 1 % Penicillin Streptomycin (Gibco). The cells were transfected using FuGene HD Transfection Reagent (Promega) and Opti-MEM (Gibco). The viruses were rescued from transfected cells using an 8-plasmid reverse genetic system containing each a genomic segment. Afterwards these viruses were amplified by two cell passages.

6.2.2. Patient samples

Patient-derived samples were collected from the two main surveillance systems in Belgium, 'influenza-like-illness' (ILI) and 'severe-acute-respiratory-infection' (SARI). ILI cases are defined by a sudden onset of symptoms, including respiratory and systemic symptoms and fever. A SARI case is defined as an acute respiratory illness with onset within the last 10 days of respiratory symptoms, fever, and requiring hospitalisation for at least 24 hours. These surveillance systems are in place to follow trends of virus spread and changes in circulating influenza viruses. From these two surveillance systems, initially 253 samples were selected [792, 865]. Only samples with a genome copy number above 10^4 genomes per microliter were retained for the LFV validation (see Results), resulting in 59 retained samples, comprising 44 samples from hospitalised SARI patients and 15 from ILI outpatients, spread over the influenza season (beginning, peak and end of epidemic). The genome copy number of 10^4 genomes per microliter is based on the Cq values from the routine diagnostic surveillance with qPCR [866] and corresponds with a Cq of 19.53. The samples tested negative using reverse transcription polymerase chain reaction (RT-qPCR) for other respiratory viruses, including respiratory syncytial virus A and B, parainfluenza viruses 1, 2, 3 and 4, enterovirus D68, rhinoviruses, human metapneumoviruses, paraechoviruses, bocaviruses, adenovirus, coronaviruses OC43, NL63, 229 and MERS-CoV [799, 800]. Samples from ILI outpatients were categorized as mild cases (n = 15). Samples from hospitalised SARI patients were categorized as moderate (n = 34) or severe cases (n = 10). Hospital admission (i.e., the SARI case definition) is not a disease severity indicator itself because patients could have been admitted to hospital care for isolation purposes or other medical conditions. A severe case was defined by the presence of at least one severity indicator: death, stay in an intensive care unit, need for invasive respiratory support or extracorporeal membrane oxygenation (ECMO), or the patient having acute respiratory distress syndrome (ARDS). Available patient data are listed in Table 6.1 with the number of patients exhibiting these characteristics.

Table 6.1: Samples stratified according to patient data.

Age (years):	**<15**	**15 – 59**	**≥60**
Beginning of epidemic (<week 4)	**4**	**2**	**12**
Peak of epidemic (week 4 - 6)	**2**	**3**	**20**
End of epidemic (>week 6)	**4**	**1**	**11**

ILI	15	**SARI**	44
Male*	25	**Female***	32
Vaccinated*	11	**Not vaccinated***	26
Antibiotics administered*	23	**No antibiotics administered***	29
Respiratory diseases	9	**No respiratory disease**	50
Cardiac disease	18	**No cardiac disease**	41
Obesity	6	**No obesity**	53
Renal insufficiency	9	**No renal insufficiency**	50
Diabetes	6	**No Diabetes**	53
Immuno-deficiency	5	**No immuno-deficiency**	54
Neuromuscular disease	7	**No neuromuscular disease**	52
Stay in ICU	5	**No stay in ICU**	54
Resulting in death*	7	**Not resulting in death***	46

*Samples for which certain patient data were unknown, were excluded for analyzing that particular characteristic.

Additionally, the median, first quartile and third quartile copy numbers of genomes per microliter of 1273 A(H3N2) positive influenza samples from the influenza seasons 2015-2019 in Belgium were calculated and plotted with an in-house script (python 3.6) and the matplotlib 3.3.4 library [867] hiding the outliers. The boxplot including the in outliers is shown in Supplementary Figure S6.1.

6.2.3. Creation of mixes of wild-type and mutant viruses

To assess the minimal percentage (i.e., AF) for a LFV to be considered truly present and not constitute a FP observation, mixes were made from the wild-type (WT) and mutant virus, created as described above, for both Influenza A/Bretagne/7608/2009 (A(H1N1)pdm09) and Influenza A/Centre/1003/2012 (A(H3N2)) with eight ratios (0, 0.1, 0.5,1, 5, 10, 20 and 100% mutant virus) (Supplementary Table S6.3). Mixes were made in triplicate based on the plaque forming units (PFU/mL; concentration of virus) of the infectious virus of the WT and mutant. Constructed mixes were situated mainly in the 0-5 % range (Supplementary Table S6.4), since previous studies [663, 857, 860–862] have reported most FP being present in this range. RT-ddPCR was used to determine the genome copy numbers of the introduced mutations in the respective mixes (Supplementary File S6.1).

6.2.4. RNA isolation and RT-qPCR

RNA of the A/Bretagne/7608/2009 (A(H1N1)pdm09) and A/Centre/1003/2012 (A(H3N2)) influenza virus mixes was extracted from culture supernatants using the Easy Mag platform (BioMérieux, #280130-#280134 and #280146) according to the manufacturer's instructions. Extraction of nucleic acids of clinical specimens was performed using the Viral RNA/DNA isolation kit (Macherey Nagel, Germany, cat No: MN 740691.4). The RNA extraction was done according to manufacturer's instructions except that the beads were not washed in buffer MV5 but instead left to dry for 10 minutes until the pellet did not appear shiny anymore.

Using 5 µL RNA for each sample, a RT-qPCR was performed using the SuperScript™III Platinum® One-Step Quantitative Kit (Invitrogen) with primers InfA_Forward, InfA_Reverse and InfA_probe. These bind to an influenza M gene section [868]. Each reaction contained 0.5 µL primer/probe, 1 µL SuperScript III RT/Platinum Taq mix, 5 µL nuclease-free water, 12.5 µL PCR Master Mix and 5 µL RNA.

6.2.5. PCR amplification and whole genome sequencing

To amplify RNA extracts, primers designed to target the 3' and 5' conserved ends of all eight segments were used as described previously [792]. Concisely, RT-PCR was used to generate sequencing amplicons in a reaction volume of 50 µL. The used protocol is based on Van den Hoecke et al. (2015) [710] with optimised volumes and RT-PCR conditions. Primers included CommonA-Uni12G (GCCGGAGCTCTGCAGATATCAGCGAAAGCAGG), CommonA-Uni12 (GCCAGAGCTCTGCAGATATCAGCAAAAGCAGG) and CommonA-Uni13G (GCCGGAGCTCTGCAGATATCAGTAGAAACAAGG) [710]. The reaction volumes included 25 µL RT-PCR buffer, 1 µL SuperScript III One-Step RT-PCR Platinum® Taq HiFi DNA Polymerase (Invitrogen, USA), 17.375 µL dH2O, 0.375 µL of each primer (20 µM), 0.5 µL RnaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen, USA) and 5 µL of RNA extract. An error rate (number of misincorporated nucleotides per total number of nucleotides polymerized) of lower than 1x10-3 by Invitrogen was estimated for the SuperScript III One-Step RT-PCR Platinum® Taq HiFi DNA Polymerase [869]. The following PCR conditions were used: one cycle at 42°C for 15 minutes, one cycle at 55°C for 15 minutes, one cycle at 60°C for 5 minutes, one cycle at 94°C for 2 minutes (ramp rate: 2.5 °C/s); 5 cycles at 94°C for 30 seconds, 45°C for 30 seconds (ramp rate: 2.5 °C/s) and 68°C for 5 minutes (ramp rate: 0.5 °C/s); 37 cycles at 94°C for 30 seconds, 55°C for 30 seconds and 68°C for 5 minutes; and one cycle at 68°C for 5 minutes (ramp rate: 2.5 °C/s). After purifying the generated amplicons with the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel, Germany) according to the manufacturers' instructions, the concentration of each purification product was quantified with the Qubit 4 Fluorometer (Invitrogen, USA) using the Qubit broad-range assay. Purified products were examined with the Agilent TapeStation (Agilent Technologies, USA) using the Agilent D5000 ScreenTape system.

Sequencing libraries using the Nextera XT DNA Sample Preparation Kit (Illumina, USA) were prepared with the purified RT-PCR products according to the manufacturer's instructions. All libraries were sequenced on an Illumina MiSeq (Illumina, USA) platform using the MiSeq V3 chemistry, as described by the manufacturer's protocol, to produce 2 x 250 bp paired-end reads. Generated WGS data are available in the NCBI Sequence Read Archive (SRA) [765] under accession number PRJNA692424 for the reverse genetics samples (Supplementary Table S6.3) and PRJNA615341 for the patient-derived samples (Supplementary Table S6.5).

Consensus genome sequences were obtained as described previously [792]. Concisely, using Trimmomatic v0.32 [610], the raw (paired-end) reads were trimmed with the following settings: 'ILLUMINACLIP:NexteraPE-PE.fa:2:30:10', 'LEADING:10', 'TRAILING:10', 'SLIDINGWINDOW:4:20', and 'MINLEN:40' retaining only paired-end reads. An appropriate reference genome for read mapping was selected from the NCBI viral genomes resource [766] for each sample. Following the GATK 'best practices' protocol [662] using Picard v2.8.3 (https://broadinstitute.github.io/picard/) and GATK v3.7, the consensus sequences for all samples were obtained. First, following best practices in the field [870–873], duplicated reads were marked with PICARD MarkDuplicates in order to remove reads originating from PCR duplicates of the same original DNA molecule which could artificially inflate AF of identified variants. This was followed by indel realignment with GATK and variant calling using GATK UnifiedGenotyper with the following options: '-ploidy 1', '--stand_call_conf 30', and '--genotype_likelihoods_model BOTH'. Subsequently, only high-quality variants with a read depth ≥200 were retained using GATK VariantFilter. Next, GATK FastaAlternateReference-Maker was used to obtain the consensus sequence based on the called variants and selected reference sequence.

6.2.6. Low-frequency variant identification

Only samples with a viral load ≥10^4 genomes/µL (see above), and a genome median coverage higher than 1000x calculated as described previously [792], were retained. For LFV calling, the consensus genome fasta files were first indexed using Samtools faidx 1.3.1. Bowtie2-build 2.3.0 [625] was then used to generate indexes. Reads were aligned to the consensus sequence using Bowtie2 align 2.3.0 in end-to-end mode for each sample, producing SAM files that were converted into BAM with Samtools view 1.3.1. Reads were then sorted using Picard SortSam 2.8.3 (http://broadinstitute.github.io/picard/) with the option ‘SORT ORDER=coordinate’. A dictionary of the reference fasta files was created using Picard CreateSequenceDictionary 2.8.3. Reads originating from PCR duplicates which could bias the observed AF of LFV were removed from read alignments using Picard MarkDuplicates 2.8.3 with the option ‘REMOVE_DUPLICATES=true’. The “LB”, “PL”, “PU” and “SM” flags are required for downstream analysis by GATK and were set to the placeholder value “test” using Picard AddOrReplaceReadGroups 2.8.3. The resulting BAM files were indexed by Samtools index 1.3.1 and used as input for GATK RealignerTargetCreator 3.7 [662] followed by GATK IndelRealigner 3.7 for indel realignment. The generated BAM files were then indexed using Samtools index 1.3.1 and LoFreq 2.1.3.1 [659] was used to detect LFV in ‘call mode’. LoFreq separates true LFV from erroneous variant calls by using Phred-scores as probability error in a Poisson-binomial distribution. The consensus sequence of each sample was used as its own reference to call LFV, in order to avoid calling high-frequency non-reference bases due to an inadequate choice of a single reference sequence for all samples used by LoFreq to call variants, i.e. nucleotides at low allelic frequency differing from the consensus at a specific genomic position [659]. Average read position values were added to called variants using an in-house script (python 3.6) [874] (Supplementary File S6.2) based on the one provided by McCrone et al. [663]. Only variants with a mean reads location within the central 50% positions (i.e., between bases 62 and 188) were retained for further analysis as advised by McCrone et al. [663]. Variants were not further filtered based on Phred-score or mapping quality as was explored in other work, because these metrics are already internally considered by LoFreq for variant calling [659, 663]. An archive containing the code used to call variants and instructions to run it is available as part of the Supplementary File S6.2.

To determine an AF threshold, the workflow described above was used to call variants in the mixes of WT and mutated A(H1N1)pdm09 and A(H3N2) strains. Receiver operating characteristic (ROC) curves for both subtypes were created using an in-house script (python 3.6) and the matplotlib 2.2.2 library [867]. Briefly, called variants were first sorted by decreasing observed AF and then numbers of true and FP variants were calculated at each called AF and plotted as a ROC curve.

6.2.7. *Statistical analysis*

All statistical analyses were performed using R-software (RStudio 1.0.153; R3.6.1) (Supplementary File S6.8). Sequencing depth and virus concentration were not introduced as covariates, because we assume that the number of amplification and sequencing errors will be limited due to the validated thresholds set up beforehand (virus concentration = 10^4 copies/µL; allelic frequency = 5% - see Results). Furthermore, any remaining amplification and sequencing errors are expected to be distributed randomly over the genome, and these should consequently not have an influence on the statistical analysis. A glm (link function = quasipoisson) was used to assess the association between number of detected LFV and individual patient data parameters, which included disease severity (classified into mild, moderate and severe), patient age, sampling date, sex, vaccination status, presence of comorbidities and disease severity indicators. Patient data were only evaluated if at least 5% of the retained patient samples met the condition. For example, asthma was not retained because only two out of 59 patients suffered from this condition (3.4%), whereas vaccination status was retained since 11 out of 59 patients were vaccinated (18.6%). Afterwards, all identified significant associations ($p<0.05$) were fitted simultaneously in a glm with the same link function and only significant associations were retained. In addition to the median, the interquartile range (IQR) and the effect size were calculated.

6.3. Results

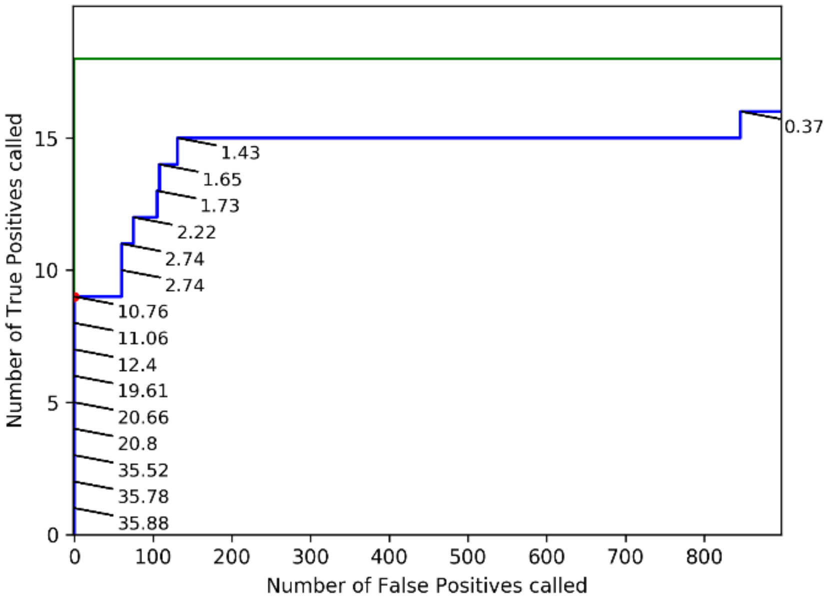
6.3.1. *Validating an AF threshold for LFV calling using an experimental quasispecies population*

Sequencing errors affect the frequencies at which variants can reliably be called. At decreasing frequencies, even for high-coverage datasets, the amount of reads containing a certain variant becomes too limited to discriminate real LFV from sequencing errors. Decreasing AF thresholds for accepting LFV will consequently increase sensitivity by identifying more true positive (TP) variants, but also decrease specificity by incorporating more

FP variants. It is therefore necessary to establish a validated threshold for the observed AF for accepting LFV. A mutated version of Influenza A/Centre/1003/2012 (A(H3N2)) with high genomic copy number (WT=98,475 genomes/µL; MUT=312,625 genomes/µL) was used to create a validation dataset in triplicate, for which the ground truth was known, to determine an AF threshold for accepting called LFV. The mutant included a specific mutation in the NA segment present at 100%, i.e. the well-known A(H3N2) oseltamivir resistance mutation NA-E119V [849], which served as a marker when mixing the WT and mutant virus in different ratios (Supplementary Table S6.4). The resulting mixes of the eight ratios (theoretically: 0%, 0.1%, 0.5%,1%, 5%, 10%, 20% and 100% mutant virus), and their triplicates, were then subjected to WGS. High sequencing coverages were obtained for all samples and segments (Supplementary Figure S6.2), after which LFV were called with LoFreq. Consequently, 18 TP were expected (i.e. one mutation times 6 ratios (0.1%, 0.5%,1%, 5%, 10%, 20%) times 3 replicates). Levels of read deduplication were relatively limited (min=21%, max=61%, average=36%; Supplementary Table S6.6), and an additional investigation of variants called with and without read deduplication confirmed that read deduplication did not cause any major bias in the numbers of called variants (Supplementary File S6.3). Noteworthy, seven additional variants were detected where the mean of the called frequencies over the triplicates corresponded to expected frequencies based on the TP dilution values, as observed at least in one dilution mix with an AF >5% (Supplementary File S6.4). This indicates that during the propagation in cells of both the WT and mutant, other variants emerged even in the absence of external selection pressure. These seven variants were therefore removed from the variant sets used for AF threshold determination as these unexpected variants were not part of the 'ground truth', but showed sufficient evidence for being true variants instead of FP (Supplementary File S6.4). Afterwards, TP variants (i.e. the introduced NA mutation in the different mixes) and FP variants (i.e. any variant called in the different mixes that did not correspond with the WT, excluding the seven aforementioned variants) observed at varying observed AFs were expressed in a ROC curve (Figure 6.1), considering triplicate values as independent values. The AFs used in the ROC curve are the observed percentages of the NA mutation as determined with Lofreq. A ROC curve expresses the relationship between sensitivity and specificity for a benchmarked experiment where the ground truth is known by varying a discrimination threshold (here the AF) and plotting the false positive rate (i.e., 1-specificity) and sensitivity on the x- and y-axis, respectively. A perfect assay where all FP are separated from TP is characterised by a ROC curve with a right angle that follows the upper left boundary of the plot (Figure 6.1).

For A(H3N2), no FP and 50.00% of TP (n=9/18) were called at an observed AF of 4.82% or higher. This seemingly low sensitivity is explained by the construction of the dataset which aimed at providing a high resolution at low AF to determine the limit of detection and therefore contained half of the variants at an AF lower than 5%. Decreasing the AF threshold increased the sensitivity but impaired a high cost in specificity (Table 6.2). At an observed AF of 1%, 83.33% of TP (n=15/18) were recovered at a cost of 289 FP. The highest sensitivity was obtained at an observed AF of 0.37%, where 88.89% (n=16/18) of variants were called at a cost of 847 FP. An AF cut-off of 5% was therefore selected as a conservative AF threshold to explicitly minimize the amount of called FP variants to be used for exploring potential associations with host characteristics (see below). Evaluation of the benchmark dataset created for A(H1N1)pdm09 exhibited the same trends, and confirmed 5% to be an adequate threshold to avoid the inclusion of FP observations (Supplementary File S6.5).

Figure 6.1: ROC curve for validating an AF threshold using a A(H3N2) benchmark dataset. The green line represents a theoretical scenario where a perfect variant caller identifies all 18 TP before any FP are called (i.e. perfect sensitivity and specificity). The blue line represents the numbers of observed TP and FP in the benchmark dataset for A(H3N2) at decreasing thresholds for the observed AF of called variants. Observed AF of TP are indicated on the graph. AF thresholds used to create the ROC curve are the numbers plotted in the figure (as percentages). The numbers of FP and TP at the threshold of 5% AF employed for the analysis of patient-derived datasets is depicted by a red dot (no additional TP or FP were observed between 5% and 10.76%). More detailed values are available in Table 6.2. AF=Allelic Frequency; ROC=Receiver operating characteristic; FP=False Positive; TP=True Positive



Observed AF (%)	Number of TP	Number of FP	Sensitivity (%)*	Specificity (%)
10.0	8	0	44.44	100.00
5.0	9	0	50.00	100.00
2.0	12	86	66.67	99.97
1.0	15	289	83.33	99.90
0.5	15	678	83.33	99.75

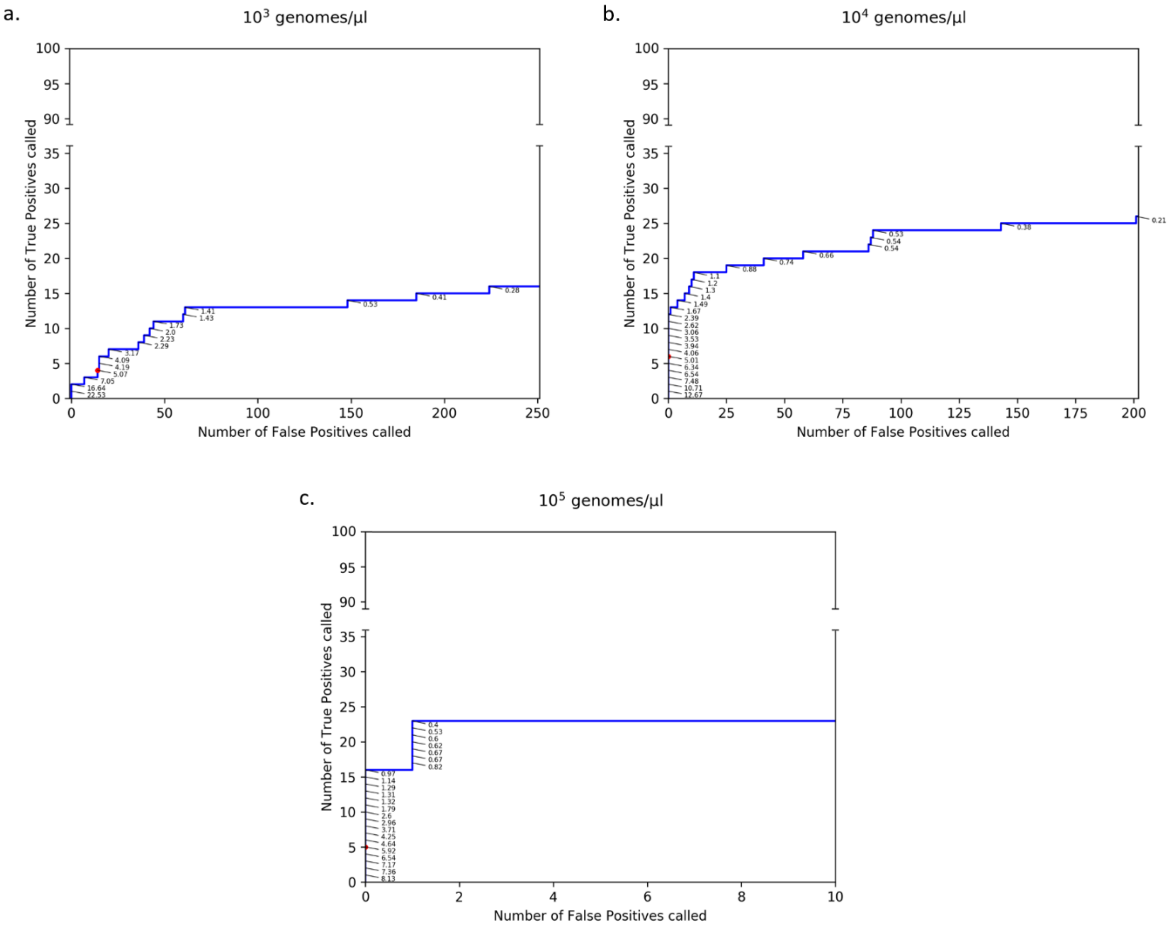
Table 6.2: Number of TP, FP, sensitivity, and specificity at different AF thresholds for the A(H3N2) benchmark dataset. Although the specificity remains high due to the size of the negative class (all positions in the genome that are not positives), the number of FP increases dramatically at lower AF, rapidly exceeding more than ten-fold the number of TP. AF=Allelic Frequency; FP=False Positive; TP=True Positive. *****: Sensitivity is considered over the full dataset, and not only variants expected at specific AF; see results for further details.

6.3.2. Selection of patient-derived samples based on their genome copy number

For the described validation of an AF threshold of 5% based on the experimentally constructed benchmark dataset, all mixes always contained very high genome copy numbers (≥10^5 genomes/µL, see above). It has been previously established that the genome copy number and titer of samples can also impact LFV calling. Prior research by McCrone et al. indicated that samples with a copy number of ≥10^5 genomes/µL are acceptable, while samples with a copy number ranging between 10^3-10^5 genomes/µL should be sequenced in duplicate to reduce FP [663]. In routine surveillance, only a limited number of samples however have a copy number of ≥10^5 genomes/µL. Only 12 out of 253 sequenced samples of the Belgian influenza season 2016-2017 had a genomic copy number ≥10^5 genomes/µL (Supplementary Table S6.5). This was not due to sample selection bias, since the median of 1273 A(H3N2) positive influenza samples from the influenza seasons 2015-2019 in Belgium was 1168.85 genomes/µL (IQR: 88.70-8907.89 genomes/µL) (Supplementary Figure S6.1), with a median associated Cq value of 22.52 (IQR: 19.48-26.68), which corresponds to other observations from the literature [875–877].

To evaluate the impact of adopting a more relaxed genome copy number threshold, we investigated the sensitivity and specificity of the LFV calling workflow on a benchmark dataset containing lower genome copy numbers, for which reference samples of mixes of specific variants at varying targeted AFs and varying initial genomics copy numbers produced and sequenced by McCrone et al. [663] were analysed with the same method as described previously. Samples used for this analysis were produced by McCrone et al. as an experimental within-host population by inserting 20 mutations in a WSN33 virus genetic background and then diluted to generate five targeted allelic frequencies (5%, 2%, 1%, 0.5% and 0.2%) and three genomic titers (10^3, 10^4 and 10^5 genomes/µl) [663]. Titers, targeted allelic frequencies and SRA accession numbers of the samples used can be found in Supplementary Table S6.2. For samples with 10^3 genomes/µl, no FP and 2% of TP (n=2/100) were called at an observed AF of ≥16.64%. These particularly low sensitivities are again the result of the dataset encompassing a majority of low allelic frequency variants. The highest sensitivities, 23%, 26% and 16% for genomic titers of respectively 10^5, 10^4 and 10^3, were obtained at an observed AF of 0.40%, 0.21% and 0.28% at a cost of 1, 201 and 224 called FP, respectively (Figure 6.2, Table 6.3).

Figure 6.2: ROC curves to validate an AF threshold using a A(H3N2) benchmark dataset at different genome copy numbers. Observed TP (out of the 100 expected) and FP counts in the benchmark datasets provided by McCrone et al. **[663]** at variable genome copy numbers. The blue line represents observed TP and FP counts in the benchmark dataset for A(H3N2) at variable thresholds for the AF. Observed AF of called TP are plotted in the figure as percentages. The numbers of observed FP and TP at the threshold of 5% AF employed for the analysis of patient-derived datasets is depicted by a red dot. More detailed values are available in Table 6.3. Abbreviations: AF=Allelic Frequency; FP=False Positive; TP=True Positive



Viral load (genomes/µl)	Observed AF (%)	Number of TP	Number of FP	Sensitivity (%)*	Specificity (%)
10^5	10.0	0	0	0.00	100.00
	5.0	5	0	5.00	100.00
	2.0	10	0	10.00	100.00
	1.0	15	0	15.00	100.00
	0.5	22	1	22.00	99.99
10^4	10.0	2	0	2.00	100.00
	5.0	6	0	6.00	100.00
	2.0	12	0	12.00	100.00
	1.0	18	17	18.00	99.97
	0.5	24	67	24.00	99.90
10^3	10.0	2	1	2.00	99.99
	5.0	4	14	4.00	99.98
	2.0	9	41	9.00	99.87
	1.0	13	83	13.00	99.36
	0.5	14	154	14.00	99.76

Table 6.3: Number of TP, FP, sensitivity, and specificity at different AF thresholds using a A(H3N2) benchmark dataset at different genome copy numbers. Although the specificity remains high due to the size of the negative class (all positions in the genome that are not positives), the number of FP increases dramatically at lower observed AF, an effect which is more pronounced at lower genome copy numbers. *: Sensitivity is considered over the full dataset, and not only variants expected at specific AF; see results for further details.

Comparison of results for a viral load of ≥10^5 genomes/µL of Table 6.2 and Table 6.3, indicates similar trends with increasing AF increasing specificity whilst penalizing sensitivity. The sensitivities of the two benchmark datasets in Table 6.2 and Table 6.3 are however not directly comparable because the truth set of mutations is present at different AF, resulting in lower sensitivity values for the McCrone dataset because more real variants were present in the observed AF range of 1%-5%. The previously selected AF threshold of 5% was therefore shown to be a conservative value for filtering out FP variants in datasets obtained from samples with low initial genomic copy numbers because despite removing many TP variants, it also effectively safeguards against including FPs for genome copy numbers at 10^4-10^5, but not at 10^3, genomes/µL. A minimal genome copy number of 10^4 genomes/µL was therefore enforced for the clinical dataset.

6.3.3. Prevalence of LFV in clinical samples

LFV calling was performed on the 59 retained samples with a genome copy number of ≥10^4 genomes/µL from the Belgian influenza 2016-2017 A(H3N2) season. When the selected threshold of 5% AF was used, at least 20 LFV were detected in seven samples, while for 30 samples between 0 and 20 LFV were detected. Finally, 22 samples did not reveal any LFV

(Supplementary File S6.7). Across all samples, LFV at 56 genomic positions were detected in two or more patients, including eight located in PB2, six in PB1, 14 in PA, 12 in HA, six in NP, three in NA, one in MP and six in NS. The majority of these variants were detected at a low observed AF of 5-20%.

6.3.4. Patient data associated with prevalence of LFV

To investigate the potential relevance of LFV for routine influenza monitoring, a proof of concept investigation based on associations of LFV with patient data was performed. The association of patient data with the number of detected LFV was investigated. After an initial glm analysis where all patient data were evaluated individually, disease severity, antibiotics use and age resulted in a significant association. In a second step, a glm was fitted including the three significant patient data simultaneously, which only resulted into a significant result for disease severity. The number of detected LFV was observed to be significantly higher in ILI cases (i.e. mild cases) compared to SARI cases (i.e. moderate and severe cases) (Table 6.4; Supplementary File S6.7).

Table 6.4: Statistically significant associations between number of LFV in clinical samples and patient data. Results include the median, first quartile and third quartile of the number of detected LFV across the 59 retained samples, and also p-value and effect size. The interpretation of the odds ratio values commonly published in the literature are: <1.68 (small effect), 1.68 - 3.47 (moderate effect) and >= 6.71 (large effect) [878]. ILI cases comprise the mild cases, while the SARI cases include moderate and severe cases. CI=Confidence interval

Patient data	Median	P-value	Effect size [CI]
Disease Severity	**Mild**: 19 [3.5-60] **Moderate/Severe**: 1 [0-3]	2.67E-08	26.40 [10.89-83.88]

Additionally, associations between patient data and the proportion of nucleotides at their specific genomic positions, including both LFV and high-frequency variants, were evaluated. Although several associations were identified, these were all below acceptable statistical thresholds. These results are therefore provided in the Supplementary File S6.6 for informative purposes only and not further considered below.

6.4. Discussion

Since the dynamics of quasispecies can afford influenza a considerable advantage on genetic fitness during within-host evolution, quasispecies information might be relevant for future clinical interventions and epidemiological investigation. HTS renders it nowadays feasible to explore viral quasispecies in patient-derived samples by detecting LFV. However, many challenges remain to obtain reliable results in order to introduce LFV in routine surveillance, in which sampling and funding are often limited. Although HTS enables deep sequencing, it becomes difficult to distinguish sequencing errors from real LFV at low AF. The first goal of this study was to establish an AF threshold for retaining LFV using mixes of a WT and NA-E119V-mutant influenza A(H3N2) virus with different proportions to create a benchmark population that was sequenced followed by LFV calling with LoFreq. While multiple other low-frequency variant callers exist [655, 661, 879–881], LoFreq has been shown to perform particularly well on short read sequencing of virus samples, especially when considering specificity [882, 883]. Other variant callers could alternatively be used as part of the validation approach presented in the current study by other scientists using other software packages. An AF cut-off of 5% was selected as the minimal AF at which no FP variants were called in the experimentally constructed benchmark A(H3N2) population. An additional exploratory analysis with mixes from the A(H1N1) subtype, which included two nucleotide mutations resulting in the NA-H275Y amino acid mutation, confirmed this as being a robust threshold also applicable to other subtypes (Supplementary File S6.5). Since the A(H3N2) and A(H1N1) benchmark populations only contained a single and two nucleotide mutations, respectively, publically available data containing more mutations were also considered. The dataset from McCrone et al. includes 20 point mutations and also an extra data point at a theoretical AF of 2%, in contrast to our sequenced A(H3N2) population containing a theoretical AF gap between 1% and 5%. Analysis of this dataset with our workflow similarly confirmed 5% to be a robust AF threshold (Figure 6.2). This threshold prioritises specificity over sensitivity, but is context-dependent for three reasons. Firstly, although the established sensitivity of 50% at 5% observed AF (Table 6.2) may appear low, the benchmark dataset was purposefully constructed to assess the limit of detection of our workflow, and therefore contained half of the inserted variants at frequencies lower than 5%. Conversely, as a result of the choice of thresholds, all variants present at ≥5% in the benchmark dataset were correctly called. Secondly, since our aim was to evaluate associations of LFV with patient data as a proof of concept, we prioritised specificity to minimize potential FP LFV included within the statistical analysis. Depending on the application scope, this AF threshold can be decreased to increase sensitivity if the cost in specificity is deemed acceptable (e.g. approaches that prioritise finding as many LFV as possible). Thirdly, AF thresholds are coverage-dependent once coverage drops below a certain turnkey point [884], with decreasing coverages typically requiring increased AF thresholds. As both the validation dataset and clinical samples consisted of high-coverage data, our established value of 5% should only be applied to high-coverage influenza datasets. Through our emphasis on specificity, the selected AF threshold of 5% is high compared to other AF thresholds reported in other studies in the literature. Gelbart et al. [885] investigated the genetic diversity of different viruses, and used a minimum AF threshold of 1%

for highly concentrated samples including human immunodeficiency virus, respiratory syncytial virus, and cytomegalovirus. Orton et al. [886] focussed on modelling sequencing errors and distinguishing them from real viral variants using foot-and-mouth disease virus as case study. They established a minimum AF threshold of 0.5%, although this was only tested on control samples that were very highly concentrated (10^6 plasmid/µl). King et al. [887] evaluated laboratory and bioinformatic pipelines to accurately identify LFV in viral populations using foot-and-mouth disease as a case study. King et al used an AF threshold of 0.2% for highly concentrated samples (10^7 copies), but observed more errors when a reduced RNA input (10^5 copies) was used and even found consensus-level errors at (very) low RNA inputs (10^3 copies).

Previous research has indicated that besides correcting for sequencing errors, the viral load and genome copy number of samples also affect LFV calling, independently of sequencing considerations. In this study, the SuperScript III One-Step RT-PCR Platinum® Taq HiFi DNA Polymerase with an estimated error rate of less than $1\text{x}10^{-3}$ misincorporated nucleotides per total number of nucleotides polymerized was used to amplify the virus. This error rate will have a larger impact on samples with low viral loads, because they are more likely to propagate PCR-amplification errors that can result in increased FP variant detections [663]. A genome copy number of 10^5 genomes/µL was recommended by McCrone et al. and a copy number of 10^3-10^5 genomes/µL was considered acceptable if sequenced in duplicate. However, the application of these recommendations to routine surveillance may prove too restrictive as 10^5 genomes/µL is an extremely high copy number for samples encountered in routine influenza surveillance (Supplementary Figure S1), where it is already a considerable challenge to acquire the necessary funds to simply switch from Sanger sequencing the HA and NA segments to WGS. As the genome copy number of our experimental dataset was very high (>10^5 genomes/µL), we employed the experimental within-host population produced by McCrone et al. [663] at a genomic input of 10^3, 10^4 and 10^5 genomes/µL with our workflow to evaluate FP counts at lower genome copy numbers when enforcing the same 5% AF threshold. We found that also at 10^4 genomes/µl, no FP were detected, but FP were found at 10^3 genomes/µl (Table 5). Similar to our experimentally constructed A(H3N2) benchmark dataset, sensitivities were (very) low because the large majority of LFV were present at AF below 5%. Notwithstanding, a direct comparison of our results with those reported by McCrone et al. is not possible for several reasons. Firstly, McCrone et al. used p-values as a threshold with either deepSNV or LoFreq to determine effects on sensitivity and specificity in samples of varying targeted AF, whereas we used the observed AF as a threshold with LoFreq with default settings (i.e., p-value dynamically adapted as part of a Bonferroni multiple test correction) to determine an AF threshold favouring optimal specificity. Secondly, high specificity at low AF could be obtained by McCrone et al. by using deepSNV on both mutated samples and control samples containing the same genetic background. This was initially done with LoFreq on our benchmark datasets using the WT samples as controls and resulted in overall higher specificity and lower sensitivity at very low AF (unpublished results), but does not reflect routine influenza monitoring where no control samples are available for clinical samples to begin with. Thirdly, the samples used by McCrone et al. were biased toward very low AF for the TP, which had a large effect on the sensitivity.

The second goal of this study was to evaluate the prevalence of LFV in actual clinical samples collected during routine influenza monitoring, using 59 influenza A(H3N2) samples from the 2016-2017 Belgian Influenza season with a genome copy number $\geq 10^4$ genomes/µL and retaining only LFV detected at ≥5% AF. It was observed that seven of the 59 samples had at least more than 20 LFV, 30 of the 59 samples had between 0 and 20 LFV, and 22 of the 59 samples did not contain any LFV.

The third goal of this study was to explore potential associations between patient data and the presence and frequency of LFV as a proof of concept for the relevance of LFV analysis in routine influenza surveillance. Statistically significant associations were found between high numbers of LFV and mild cases. It has been suggested in the literature for other viruses that within-host diversity can be driven by host selection pressure [888, 889]. In contrast to our results where more LFV were observed in mild cases, Simon et al. observed higher diversity within the PA, HA and NA segments in severe cases compared to mild cases [754]. Additionally, we evaluated potential associations between patient data and the proportion of nucleotides at specific genomic positions. Several associations were found, however, these were below acceptable statistical thresholds (Supplementary File S6.6). We are aware, however, of the low statistical power of the association study due to the small sample size of 59 patients and unequal representation of LFV among the patient data groups. More reliable associations will therefore require larger sample sizes in future studies. However, these results show the potential added value to understand virus evolution in relation to the host, but more research is needed.

In conclusion, HTS of clinical influenza samples allows to examine LFV during human infections. Our work provides a general approach for LFV detection by delineating thresholds that balance the number of FP against the feasibility of quasispecies investigation in actual samples collected in the context of routine surveillance programmes. As a proof of concept, several relevant associations with patient data were found while considering LFV, which suggests that the relevance of LFV for influenza monitoring is currently under-valued and could contribute to a better understanding of disease. Although additional validation will be necessary, it could be of great benefit to apply the proposed approach on samples collected during routine influenza monitoring.

6.5. Acknowledgments

We thank the technicians of the service Transversal activities in Applied Genomics at Sciensano, Belgium for performing the Sanger and Next Generation Sequencing runs.

CHAPTER 7: STRATEGY TO DEVELOP AND EVALUATE A MULTIPLEX RT-DDPCR IN RESPONSE TO SARS-COV-2 GENOMIC EVOLUTION

Context of this chapter:

This chapter illustrates the added value of NGS for the SARS-CoV-2 surveillance. Due to the worldwide emergence and spread of SARS-CoV-2, rapid and reliable diagnostic testing has become important to try to prevent and control the viral transmission. The monitoring of the virus spread can be performed based on individual diagnostics in clinical samples and global detection of SARS-CoV-2 in wastewater samples. As SARS-CoV-2 research is rapidly evolving together with the number of new variants appearing around the world, it is possible that PCR based methods that were validated only a few months ago are not suitable anymore due to genetic modification appearing in the annealing site of the primers and probes of the proposed methods. In this chapter we performed an *in silico* evaluation of commonly used primers and probes against available WGS data. Based on this, we propose that a minimal set-up should be designed for experimental testing, including a sensitivity, specificity and applicability test. In this chapter, the RT-ddPCR platform was chosen as it offers several advantages over RT-qPCR. The duplex RT-ddPCR method that we propose was evaluated with regard to internationally recognized performance parameters including specificity, sensitivity and applicability.

This chapter was previously published:

Laura A. E. Van Poelvoorde, Mathieu Gand, Marie-Alice Fraiture, Sigrid C. J. De Keersmaecker, Bavo Verhaegen, Koenraad Van Hoorde, Ann Brigitte Cay, Nadège Balmelle, Philippe Herman and Nancy H.C. Roosens (2021) Strategy to Develop and Evaluate a Multiplex RT-ddPCR in Response to SARS-CoV-2 Genomic Evolution. Current issues in Molecular Biology. **43**. https://doi.org/10.3390/cimb43030134

Authors' contributions:

Conceptualization: NR, SDK; Project Administration: NR; Data Curation: LVP, MG, BV, KVH., ABC, NB; Methodology: LVP, NR; Formal Analysis: LVP, MG; Investigation: LVP, MG; Visualisation: LVP; Validation: LVP; Writing – Original Draft Preparation: LVP, MG, MAF, SDK, NR; Writing – Review and Editing: all authors; Funding Acquisition: NR, PH; Supervision: NR

Abstract:

The worldwide emergence and spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) since 2019 has highlighted the importance of rapid and reliable diagnostic testing to prevent and control the viral transmission. However, inaccurate results may occur due to false negatives (FN) caused by polymorphisms or point mutations related to the virus evolution and compromise the accuracy of the diagnostic tests. Therefore, PCR-based SARS-CoV-2 diagnostics should be evaluated and evolve together with the rapidly increasing number of new variants appearing around the world. However, even by using a large collection of samples, laboratories are not able to test a representative collection of samples that deals with the same level of diversity that is continuously evolving worldwide. In the present study, we proposed a methodology based on an *in silico* and *in vitro* analysis. First, we used all information offered by available whole-genome sequencing data for SARS-CoV-2 for the selection of the two PCR assays targeting two different regions in the genome, and to monitor the possible impact of virus evolution on the specificity of the primers and probes of the PCR assays during and after the development of the assays. Besides this first essential *in silico* evaluation, a minimal set of testing was proposed to generate experimental evidence on the method performance, such as specificity, sensitivity and applicability. Therefore, a duplex reverse-transcription droplet digital PCR (RT-ddPCR) method was evaluated *in silico* by using 154 489 whole-genome sequences of SARS-CoV-2 strains that were representative for the circulating strains around the world. The RT-ddPCR platform was selected as it presented several advantages to detect and quantify SARS-CoV-2 RNA in clinical samples and wastewater. Next, the assays were successfully experimentally evaluated for their sensitivity

and specificity. A preliminary evaluation of the applicability of the developed method was performed using both clinical and wastewater samples.

7.1. Introduction

The ongoing coronavirus disease 2019 (COVID-19) pandemic is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a positive-sense single-stranded RNA virus. The symptoms of COVID-19 include cough, respiratory problems, fever, aches and pains, fatigue, diarrhoea and taste and smell disorders [890]. SARS-CoV-2 can also cause severe complications, including death, mostly in the elderly or in people suffering from comorbidities [891, 892]. To monitor the spread of the COVID-19 pandemic and to reduce transmission, many governments have implemented intensive contact tracing, testing and isolation [531, 893–895].

The gold standard for the detection of SARS-CoV-2 is reverse-transcription quantitative polymerase chain reaction (RT-qPCR) on extracted RNA from nasopharyngeal swabs for individual diagnostics. In order to rationalize the monitoring of the virus spread at the level of a country or region regarding the number of samples, the monitoring of wastewater was also proposed for the surveillance of SARS-CoV-2 [413–415]. However, there are several limitations associated with wastewater surveillance; generally, a low virus concentration is observed in such samples, which makes detection challenging. Furthermore, the virus detection and quantification can be limited due to the instability of the genome in wastewater, the low efficiency of virus concentration methods and the lack of sensitive detection assays [413].

Although the estimation of the mutation rate of SARS-CoV-2 is lower compared to other RNA viruses [896], the virus is continuously evolving, leading to the emergence of new variants carrying multiple mutations. Current and potential future variants have the potential to be more transmissible, causing more infections and/or leading to vaccine escape [272, 627, 897, 898]. Therefore, it is important to monitor these variants in order to control the epidemic. Furthermore, the emergence of variants can potentially lead to false negative results. The impact of false negative results due to viral mutations in the target region can be reduced by using multiple targets for the detection of the virus genome as well as a constant monitoring of the effect of mutations on the performance of the PCR method [899]. In the case of a false negative result, the sample should be sequenced to pinpoint what mutation is causing it and the primers and probe of the PCR assay need to be adapted. The importance of using an *in silico* analysis using publicly available sequences to identify potential false negative results has already been stated previously [899]. Of course, as mentioned by Gand et al., an *in silico* study should be backed up by an *in vitro* study that validates the design using actual samples

[899]. Although RT-qPCR methods are the standard for clinical diagnostics and consequently are often used in wastewater samples due to the availability of these methods, many drawbacks were reported related to the use of this technology. First, the tests are expressed in cycle quantification (Cq). The Cq represents the PCR cycle at which the sample produced a fluorescent signal above the background. These Cq values are laboratory- and instrument-specific and a calibration to a quantitative standard is necessary to determine the absolute virus concentration. Furthermore, Cq values are not directly comparable across assays or technology platforms due to differences in nucleic acid extraction methods, viral targets and other parameters [900], thereby affecting inter-laboratory harmonisation in the interpretation of the test results. Finally, RT-qPCR is not adapted for wastewater samples, which often contain inhibitors that might influence the Cq values. This could affect the accuracy of viral quantification [901], which was shown for multiple sample matrices by Whale et al. [902].

Reverse-transcriptase droplet digital PCR (RT-ddPCR), may offer an interesting alternative for the detection and quantification of SARS-CoV-2 RNA [903, 904]. Comparable with the RT-qPCR method, a target-specific fluorescent probe coupled with primers is used, which makes the adaptation of the existing RT-qPCR assays straightforward. In a ddPCR, a reaction is emulsified into thousands of nanodroplets, of which a proportion does not contain the template molecule [905]. The nanodroplets are used as unique and small bioreactors to amplify the template [906–909]. At the end-point, the number of positive droplets are digitally counted relative to the total number of droplets. Furthermore, their known volume while flowing through microfluidic devices allows absolute target quantification using Poisson statistics [910, 911], which enables an easier comparison between different laboratories and tests compared to RT-qPCR. To the best of our knowledge, eight RT-ddPCR methods designed to detect SARS-CoV-2 were published, of which two are commercial kits designed by Bio-Rad [904, 912–918]. The performance of these methods was tested using reference standards, and four of the methods were tested on clinical samples of infected patient’s throat and nasopharyngeal samples. Three of these methods were tested on wastewater samples [915–917]. Moreover, four of these RT-ddPCR methods were tested on respiratory samples [904, 913, 914, 918], and in some cases were found to be positive compared to the negative RT-qPCR results [904, 913]. Additionally, the sensitivity of the RT-ddPCR methods for the detection of SARS-CoV-2 has been described previously as comparable or even higher compared to RT-qPCR methods [904, 912, 913, 918]. Therefore, in the case of a low virus concentration, this technology can be interesting to use. Furthermore, inhibition can be encountered in some matrices, such as wastewater. RT-ddPCR separates DNA, inhibitors and reagents in droplets and is an end-point measurement, only measuring after the PCR amplification. Consequently, a reduction in the biases linked to the inhibitors are often observed in RT-ddPCR [919], which makes RT-ddPCR an interesting method for wastewater surveillance. In this study, we propose a methodology using an *in silico* and *in vitro* analysis. First, available whole-genome sequencing data for SARS-CoV-2 was used to select primers and probes for PCR assays, as well as to evaluate and monitor the possible impact of virus evolution on the developed PCR assays. Second, a minimal set of *in vitro* testing was proposed to validate in-house a new duplex RT-ddPCR method specific for the detection of SARS-CoV-2, including specificity and sensitivity assessments. Additionally, the applicability of the proposed RT-ddPCR method was investigated using clinical and wastewater samples. The duplex RT-ddPCR method was developed based on the RT-qPCR methods previously developed by Institute Pasteur [920] and Lu et al. [921].

7.2. Materials and Methods

7.2.1. Selection and evaluation of key target for PCR detection of SARS-CoV-2 using WGS data

For the development of the RT-ddPCR method, two sets of primers and probe were selected from publicly available RT-qPCR assays, namely RdRp_IP4 assay from Institut Pasteur (Paris) [920], and the ORF1a assay from Lu et al., 2020 [921], that target two separate locations specific to the SARS-CoV-2 genome (Table 7.1). These assays were evaluated *in silico* [899] for their inclusivity and exclusivity in a previous study in May 2020, which determined the RdRp_IP4 assay [920], S assay from Chan et al., 2020 [922] and ORF1a assay [921] as the most specific and stable assays over time. However, due to the emergence of the B.1.351 lineage in South Africa, a mismatch located in the probe sequence of the S assay was identified, which could lead to a lower sensitivity [923]. Therefore, from the three previously described, only the ORF1a and RdRp_IP4 assays were retained in this study.

Table 7.1. Primer and probe sets included in the multiplex RT-ddPCR assay.

Name	5’ → 3’ Sequence	Target	Nucleotide Position	Concentration	Ref.
ORF1a-F	AGAAGATTGGTTAGATGATGATAGT	ORF1a	3193 – 3217	0.9 µM	[920]
ORF1a-R	TTCCATCTCTAATTGAGGTTGAACC		3286 – 3310	0.9 µM	
ORF1a-P	5’6-FAM/TCCTCACTG-ZEN-CCGTCTTGTTGACCA-3’IABkFQ		3229 – 3252	0.25 µM	
RdRp_IP4-F	GGTAACTGGTATGATTTCG	RdRp	14080 – 14098	0.9 µM	[921]
RdRp_IP4-R	CTGGTCAAGGTTAATATAGG		14167 – 14186	0.9 µM	
RdRp_IP4-P	5’HEX-TCATACAAA-ZEN-CCACGCCAGG-3’IABkFQ		14105 – 14123	0.25 µM	

A second, internal ZEN-quencher was added to the probes to obtain greater overall dye quenching in addition to the Iowa Black FQ (IABkFQ) quencher. The indicated positions refer to the reference sequence NC_045512.

The *in silico* inclusivity of ORF1a and RdRp_IP4 assays was evaluated using the bioinformatics tool SCREENED v1.0 [924], previously used for *in silico* SARS-CoV-2 assay assessment [899, 923], and recent whole-genome SARS-CoV-2 sequences. A total of 296 187 SARS-CoV-2 genomes, obtained from samples collected between 1 November, 2020 and 28 February, 2021 were obtained from the GISAID database [752] on 7 March, 2021. Only complete genomes with high coverage for which the collection date was available were selected, and genomes with low coverage were excluded. Additionally, genomes containing undetermined nucleotides “N” and degenerate nucleotides were excluded from the dataset to retain only high-quality genomes (154 489 genomes) (Supplementary File S7.1 and S7.2). These genomes were divided per month according to their collection date (November: 13 678 genomes; December: 41 128 genomes; January: 58 484 genomes; February: 41 199 genomes). From these datasets, SCREENED performed a two-step BLAST approach to find in each genome the complete amplicon sequence targeted by the ORF1a and RdRp_IP4 primers and probe sets, and subsequently produced mismatch statistics from the hybridization between the nucleotides of the primers and probes and their corresponding annealing sites in the amplicon. Based on these mismatch scores, SCREENED considered that a theoretical positive RT-ddPCR signal was produced if no mismatch in the first five nucleotides of the 3’ end of the primers was reported, if the total number of reported mismatches did not exceed 10% of the oligonucleotide length and if at least 90% of the oligonucleotide sequence aligned correctly to their targets. For the primers and probes evaluated here, this resulted in no more than one or two mismatches being tolerated. These criteria were selected because it has been previously reported that two or more mismatches can lead to potential total test failure, especially if located at the 3’ end [925, 926]. Two mismatches or less can result in potential loss of sensitivity but is less likely to lead to total test failure. For each analysed SARS-CoV-2 genome, a negative SCREENED detection signal was considered as a theoretical FN result, which was used for the *in silico* inclusivity evaluation (Equation (1)):

$$\text{Inclusivity (\%)} = (1 - (\text{Number of FN} / \text{Total Number of high quality SARS-CoV-2 genomes})) \times 100 \qquad (1)$$

FASTA files for November, December, January and February containing 13 678, 41 128, 58 484 and 41 199 SARS-CoV-2 genomes, respectively (Accession ID: Supplementary File S7.1), and a tab-delimited text file (Supplementary File S7.3), containing the primer and probe sequences and their corresponding amplicon sequence to be mined in the genomes, were used as input for SCREENED.

7.2.2. Development of RT-ddPCR method for the detection of SARS-CoV-2

The RT-ddPCR assay was evaluated using purified RNA from the SARS-CoV-2 virus (Vircell, Granada, Spain – MBC137-R). The RT-ddPCR was performed using the One-Step RT-ddPCR Advanced Kit for Probes (Bio-Rad, Hercules, CA, USA). All the components from the kit were thawed on ice for 30 min and thoroughly mixed by vortexing each tube at maximum speed for 30 seconds. The reagents were made into larger master mixes and then aliquoted into individual reactions. Each reaction had a total volume of 22 µL that was set up on ice, including 0.99 µL of each primer with an initial concentration of 20 µM and 0.55 µL of each probe with an initial concentration of 10 µM, 1.1 µL of 300 mM DTT, 0.14 µL of dH2O, 2.2 µL Reverse Transcriptase, 5.5 µL One-Step Supermix and 8 µL of sample. The primers were obtained from Eurogentec (Seraing, Belgium), while the ZEN-probes were supplied by Integrated DNA Technologies (Coralville, IA, USA). According to the manufacturer’s instructions, 20 µL of the reaction mix and 70 µL of Droplet Generation Oil for Probes were loaded into a QX200TM droplet generator (Bio-Rad) and to increase the number of droplets, the cartridge was kept for two min at room temperature. After the droplet generation, 40 µL of droplets were recovered per reaction. The amplification was performed in a T100TM Thermal Cycler (Bio-Rad) with the following conditions: one cycle at 25 °C for 3 min, one cycle at 50 °C for 60 min (RT), one cycle at 95 °C for 10 min (Taq polymerase activation); 40 cycles at 95 °C for 30 seconds (denaturation), 55 °C for 60 seconds (annealing); one cycle at 98 °C for 10 min (enzyme inactivation) and finally one cycle at 4 °C for 30 min (stabilization). Next, the plate was transferred to the QX200 reader (Bio-Rad) and the results were acquired using the HEX and FAM channel, according to the manufacturer’s instructions. The QuantaSoft software v1.7.4.0917 (Bio-Rad) was used for the interpretation of the results and the threshold was set manually.

7.2.3. Validation of the specificity of the RT-ddPCR assay for SARS-CoV-2

The specificity of the method was experimentally established using a set of DNA and RNA controls from Bacillus subtilis Si0005 (Sciensano collection, Brussels, Belgium), Escherichia coli LMG 2092T (BCCM collection, Brussels, Belgium), Aspergillus acidus IHEM 26285 (BCCM collection), Candida cylindracea MUCL 041387 (BCCM collection) and Zea mays (ERM-BF413ak). These were extracted as described in Fraiture et al., 2020 [927]. Additionally, Homo sapiens (Promega, G3041) and viruses including SARS-CoV (Vircell, MBC136-R), MERS-CoV (Vircell, MBC132), influenza H1N1 (Vircell, MBC082), influenza H3 (Vircell, MBC029), influenza B (Vircell, MBC030), adenovirus (Vircell, MBC001), enterovirus D68 (Vircell, MBC125), norovirus (Vircell, MBC111), respiratory syncytial virus A (RSV A) (Vircell, MBC041), rhinovirus (Vircell, MBC091), rotavirus (Vircell, MBC026), coronavirus OC43 (Vircell, MBC135-R) and coronavirus 229E (Vircell, MBC090) were used. The SARS-CoV-2 RNA (Vircell, MBC137-R) was used as a positive control. Each material was tested in duplicate and included 200 copies/µL for the viruses, while the bacterial, fungal, plant and human DNA contained 2 ng/µL.

7.2.4. Validation of sensitivity of the RT-ddPCR assay for SARS-CoV-2

The evaluation of the sensitivity was carried out using serial dilutions of purified RNA from the SARS-CoV-2 virus. Seven serial dilutions were prepared, ranging from 0.5 to 200 copies/µL, and each dilution was tested in 12 replicates. The limit of detection (LOD95%) was calculated using the web application Quodata with the number of copies of the target that is required to ensure a probability of detection (POD) of 95% [928].

7.2.5. Applicability assessment

To assess the applicability of this RT-ddPCR assay on non-artificial samples, five samples collected from patients showing clinical signs of COVID-19 were collected. From these five samples, three samples (clinical samples 1, 2, 3) previously tested positive for SARS-CoV-2 with RT-qPCR, with a high, moderate and low Cq, while two tested negative for SARS-CoV-2 (clinical samples 4, 5) (Supplementary File S7.4). The clinical samples were obtained from a biobank (allowed by the Biobank compendium of the Federaal Agentschap voor Geneesmiddelen en Gezondheidsproducten [929]). All experiments were performed in accordance with relevant guidelines and regulations. In addition, three wastewater samples (wastewater sample 1, 2, 3) were included that also previously tested positive for the SARS-CoV-2 virus with RT-qPCR, with a high, moderate and low Cq (see Supplementary File S7.4). Due to the high concentration of clinical sample 3, the sample was diluted 80 times. Consequently 0.1 µL of sample and 7.9 µL of dH2O were used in the reaction (dilution: 80X).

7.3. Results

7.3.1. In silico inclusivity evaluation for the ORF1a and RdRp_IP4 assays using SCREENED

The ORF1a and RdRp_IP4 assays were evaluated for their inclusivity with four datasets corresponding to the months November 2020, December 2020, January 2021 and February 2021 (Table 7.2) using 13 678, 41 128, 58 484 and 41 199 SARS-CoV-2 genomes, respectively. Both for the ORF1a and RdRp_IP4 assays, excellent inclusivity was obtained for the four datasets, because all assays had an inclusivity of more than 99.5%. The little variation observed between the months can mainly be attributed to random and rare mutation events that did not spread in the viral population.

Table 7.2. Inclusivity *in silico* evaluation of ORF1a and RdRp_IP4 assays obtained with SCREENED.

Month	Number of genomes	Assay	FN	Inclusivity
November	13 678	RdRp_IP4	20	99.85%
		ORF1a	17	99.88%
December	41 128	RdRp_IP4	21	99.95%
		ORF1a	95	99.77%
January	58 484	RdRp_IP4	52	99.91%
		ORF1a	67	99.89%
February	41 199	RdRp_IP4	31	99.92%
		ORF1a	28	99.93%

The number of genomes that were used in SCREENED are indicated per month. Additionally, the number of False Negative results and the inclusivity are included per assay per month. FN = False Negative

In addition, it was verified that when an FN result was obtained for a given genome, this was limited to either only the forward or reverse primer or the probe. Moreover, if an FN result was obtained for a genome for one of the assays, a positive signal was obtained for the other assay. Consequently, the inclusivity of the multiplex method using the combination of the ORF1a assay and RdRp_IP4 assay is 100%.

Finally, the dataset included 33 611 and 293 genomes belonging to the B.1.1.7 and B.1.351 lineage, respectively. The number of FN that were attributed to these lineages was limited to 0, 1, 6 and 7 for ORF1a assay and 0, 1, 22 and 21 for the RdRP_IP4 assay on a total of 18, 103, 7291 and 26 199 genomes belonging to the B.1.1.7 lineage in the months November 2020, December 2020, January 2021 and February 2021, respectively. For the B.1.351 lineage, the number of FN was limited to 0 for the ORF1a assay and 0, 1 and 0 for the RdRP_IP4 assay on a total of 21, 138, and 134 genomes belonging to B.1.1.7 in the months

December 2020, January 2021 and February 2021. No genomes belonging to the B.1.351 lineage were included from the month November 2020.

7.3.2. Specificity assessment

The specificity of the RT-ddPCR method was experimentally tested for each positive and negative material (Table 7.3). SARS-CoV-2 RNA was used as a positive control, while four closely related coronaviruses, 10 other viruses, human DNA, plant (Zea mays), two bacteria and two fungi were used as negative controls. Excellent exclusivity was observed because no amplification was observed for all negative controls, while the positive control presented an amplification (Table 7.3).

Table 7.3. Specificity assessment of the developed RT-ddPCR method.

Kingdom	Genus	Species	Strain number	RT-ddPCR
Animalia	Homo	sapiens	/	-
Plantae	Zea	mays	/	-
Bacteria	Bacillus	subtilis	SI0005	-
	Escherichia	coli	MB1068	-
Fungi	Aspergillus	acidus	26285	-
	Candida	cylindracea	041387	-

	Family	Species	RT-ddPCR
Viruses	Picornaviridae	Rhinovirus B	-
	Reoviridae	Rotavirus	-
	Orthomyxoviridae	Influenza A (H1N1)	-
	Orthomyxoviridae	Influenza A (H3)	-
	Orthomyxoviridae	Influenza B	-
	Adenoviridae	Adenovirus	-
	Picornaviridae	Enterovirus D68	-
	Caliciviridae	Norovirus	-
	Pneumoviridae	RSV A	-
	Coronaviridae	SARS-CoV	-
	Coronaviridae	MERS-CoV	-
	Coronaviridae	Corona OC43	-
	Coronaviridae	Coronavirus control	-
	Coronaviridae	SARS-CoV-2	+

The absence and presence of amplification is symbolized by a - or +, respectively. The RT-ddPCR method was performed in duplicate on each sample. As positive control SARS-CoV-2 RNA was included.

7.3.3. Sensitivity assessment

The sensitivity of the designed RT-ddPCR method was tested using SARS-CoV-2 RNA with different estimated target copy numbers, namely 200, 50, 25, 10, 5, 1, 0.5 and 0 copies/µL. An amplification for all 12 replicates was observed until five estimated target copies/µL (Table 7.4). The $LOD_{95\%}$ for the ORF1a assay was determined at 4.57 [2.74,7.61] estimated target copies/µL, while the RdRp_IP4 assay proved to be more sensitive with a $LOD_{95\%}$ of 1.59 [0.95,2.67] estimated target copies/µL. Notably, in 4/12 and 9/12 replicates for the ORF1a assay and RdRp_IP4 assay, respectively, it also tested positive for samples with an estimation of 0.5 and 1 copies/µL (Table 7.4, Supplementary Files S7.5 and S7.6).

Table 7.4. Sensitivity assessments of the developed RT-ddPCR method

Estimated target copy number	Sensitivity assessment (ORF1a)	Sensitivity assessment (RdRp_IP4)
200 copies/µL	+ (12/12) 117.59 ± 7.68 copies/µL	+ (12/12) 138.46 ± 8.44 copies/µL
50 copies/µL	+ (12/12) 25.53 ± 8.02 copies/µL	+ (12/12) 27.98 ± 7.82 copies/µL
25 copies/µL	+ (12/12) 10.95 ± 2.37 copies/µL	+ (12/12) 12.54 ± 1.95 copies/µL
10 copies/µL	+ (12/12) 4.45 ± 0.82 copies/µL	+ (12/12) 4.70 ± 1.06 copies/µL
5 copies/µL	+ (12/12) 1.82 ± 0.66 copies/µL	+ (12/12) 2.20 ± 0.90 copies/µL
1 copies/µL	+ (4/12) 0.11 ± 0.16 copies/µL	+ (9/12) 0.37 ± 0.29 copies/µL
0.5 copies/µL	+ (4/12) 0.19 ± 0.31 copies/µL	+ (9/12) 0.48 ± 0.44 copies/µL
0 copies/µL	- (0/12)	- (0/12)

The absence and presence of amplification are indicated by - or +, respectively. For each estimated target copy number, 12 replicates were tested and the number of positive replicates is indicated between brackets at the middle line of each box. In addition, the average of the observed copies/µL (± the standard deviation, as obtained with the RT-ddPCR measurement, is indicated between brackets at the lower line.

7.3.4. Applicability assessment

The presence and quantity of SARS-CoV-2 was investigated in five clinical (nasopharyngeal swabs) and three wastewater samples. Among the five clinical samples, three samples tested positive for both the ORF1a and RdRp_IP4 assay (Table 7.5). The three wastewater samples also tested positive for SARS-CoV-2 (Table 7.5). These detection results corresponded to their previous results obtained with RT-qPCR, where wastewater sample 1 and clinical sample 1 had the lowest concentration, while wastewater sample 3 and clinical sample 3 had the highest concentration. The detailed results of the RT-ddPCR method on the clinical and wastewater samples are presented in Table 7.5 and Supplementary File S7.7.

Table 7.5. SARS-CoV-2 investigation in clinical samples and wastewater samples.

Sample	SARS-CoV-2 (ORF1a)	SARS-CoV-2 (RdRp_IP4)	RT-qPCR
Wastewater sample 1	+ 2.48 copies/µL	+ 1.93 copies/µL	+
Wastewater sample 2	+ 6.33 copies/µL	+ 2.20 copies/µL	+
Wastewater sample 3	+ 29.43 copies/µL	+ 36.29 copies/µL	+
Clinical sample 1	+ 2.75 copies/µL	+ 2.75 copies/µL	+
Clinical sample 2	+ 26.13 copies/µL	+ 32.18 copies/µL	+
Clinical sample 3	+ 88440 copies/µL	+ 91080 copies/µL	+
Clinical sample 4	-	-	-
Clinical sample 5	-	-	-

The sample name and the kind of sample are given in addition to the results of the detection of SARS-CoV-2 using the ORF1a assay and the RdRp_IP4 assay. The presence or absence of PCR amplification is symbolized by + or - respectively. For each RT-ddPCR, the observed copies/µL is given between brackets. Detailed results from the RT-qPCR can be found in Supplementary File S**7.**4.

7.4. Discussion

Using a total of 154 489 SARS-CoV-2 high-quality genomes, two simplex RT-qPCR assays that were designed previously to target the conserved regions of ORF1a and RdRp genes were selected for the development of a novel RT-ddPCR multiplex assay for the detection and quantification of SARS-CoV-2. The main advantage of targeting two regions is to anticipate FN results that could occur due to mutations that lead to possible mispriming of the primers and/or probes, and consequently to a lack of viral detection. Indeed, FN results have been reported previously in clinical samples due to the genetic evolution of the virus [400, 923, 930]. The use of multiple targets for the detection of the viral genome [931–933] can reduce the impact of FN results related to viral mutations in the region of the annealing of the primers and/or probe. The failure of one region can be compensated for by the detection of the other, as was shown in this study for the *in silico* evaluation. Evidently, in the case of a false negative result for one of the targets, further investigation is necessary to identify the mutation causing the false negative result by sequencing the sample. Furthermore, the primers and probe should then be adapted to minimize the impact on the test.

During the development of any new method for pathogen detection, it is of utmost importance to carefully assess its specificity, i.e., inclusivity and exclusivity. For inclusivity, a large number of various strains belonging to the targeted organism should ideally be tested. However, in the case of SARS-CoV-2, it is difficult to obtain a representative collection of all the circulating strains, and to test it experimentally. To overcome this issue, the specificity evaluation can be carried out *in silico* using bioinformatics and the large number of SARS-CoV-2 high-quality sequences publicly available, as previously performed for ORF1a and RdRp_IP4 assays [920, 921]. Moreover, after development, the detection assays need to be under constant monitoring over time, because the virus evolves and a mutation could be introduced within these targets. Currently, several new SARS-CoV-2 variants have emerged, carrying an unusually high number of mutations, and assessing all assays for FN is important. Therefore, in the present study, the latest WGS published data of SARS-CoV-2 (154 489 high-quality whole-genome sequences) were used to perform an *in silico* analysis of ORF1a and RdRp_IP4 assays, which both showed excellent results, i.e., an inclusivity of more than 99.5% from the beginning of November 2020 to the end of February 2021. Hence, no new mutations impacted the inclusivity, including the mutations linked to the variants of concern that emerged at the end of 2020. Most of the primers and probe sets used in other multi-target RT-ddPCR assays developed for SARS-CoV-2 detection [904, 912, 914, 916, 918] have also been previously analysed for their inclusivity using the same *in silico* approach [899, 923]. Most of these sets showed excellent inclusivity results (>99%), except for the primers and probe set targeting the gene N (June-December 2020: 63.89% inclusivity) used in Kinloch et al. and Suo et al., and initially designed by the China CDC [899, 923]. Therefore, the N target used in these assays should preferably not be chosen for developing SARS-CoV-2 detection methods. Concerning the exclusivity, this one has also been previously evaluated *in silico* for ORF1a and RdRp_IP4 assays successfully, with thousands of non-SARS-CoV-2 genomes [899]. Additionally, following the earlier *in silico* specificity assessment, a minimal experimental set-up was designed to evaluate the performance of the developed method. First, using a set of DNA and RNA references, the exclusivity of ORF1a and RdRp_IP4 assays was successfully confirmed, with no false positives detected for other viral, bacterial, plant and human RNA and

DNA, including closely related viruses such as SARS-CoV, MERS-CoV and coronavirus OC43. This result was expected based on the *in silico* analysis using the primer and probes selected in the present study performed by Gand et al. [899], where a 100% exclusivity was observed for these two assays, including closely related viruses. In contrast, the specificity of most other RT-ddPCR methods currently published were not experimentally evaluated using non-target DNA, such as that from bacteria [904, 912, 914, 916, 918].

Secondly, the sensitivity of our method was estimated at 4.6 and 1.6 estimated target copies/µL (LOD95%) for the ORF1a and RdRp_IP4 assays, respectively. This means that false negative results can possibly occur in the case of samples with a lower viral load than the LOD95%; however, positive results are still possible, as observed in both the sensitivity and applicability assessment. Although other targets were used by most other previously published RT-ddPCR methods, similar LODs were observed [934]. When comparing the LOD to RT-qPCR methods, the RdRp_IP4 assay using RT-ddPCR was found to be more sensitive compared to using RT-qPCR for the same target, with LOD95% of 7.9 estimated copies/µL [934]. Information on the LOD of RT-qPCR could not be found in the literature for the ORF1a assay. In Suo et al. [904], it was demonstrated that negative RT-qPCR results could be identified as positive when repeating the analysis with the optimised RT-ddPCR targeting the ORF1ab and N gene. In Alteri et al. [913], Deiana et al. [914], de Kock et al. [912] and Kinloch et al. [918], targeting the RdRP gene, ORF gene, E gene and N gene, the RT-ddPCR assay was found to be more sensitive than the RT-qPCR assay. Therefore, we expect that this RT-ddPCR assay would be at least as sensitive or even more sensitive [904, 912, 918] compared to RT-qPCR. In this study, no comparison could be made between the RT-qPCR methods used to characterise the clinical and wastewater samples (Supplementary File S7.4) and the developed ddPCR method, because different primers and probes were used.

In addition, a preliminary assessment of the applicability of the method was performed on RNA extracted from nasopharyngeal swabs and wastewater samples. The samples were selected on the basis of their different target concentrations, according to Cq values (low, medium and high) previously obtained by RT-qPCR (reflecting, respectively, high, medium and low contamination levels), and their different origins. The main goal of this experimental design was to evaluate a potential matrices effect on the PCR results using a minimum number of samples. The positive results obtained in low Cq samples using our newly developed RT-ddPCR method suggest a sensitivity of at least as high as the RT-qPCR assays used for these samples. Although the price of the RT-ddPCR method was calculated at approximately EUR 6.5 per sample, which is indeed more expensive compared to most RT-qPCR methods, RT-ddPCR reduces the work in the case of absolute quantification. One of the advantages of using RT-ddPCR instead of RT-qPCR is also the absolute quantification of the viral RNA without calibration, which enables comparison between different assays and laboratories without the necessity of a standard curve. Additionally, the accuracy of the RT-ddPCR methods should be less influenced by the inhibitors that are often present in wastewater samples. However, there are some drawbacks to RT-ddPCR, such as the longer turnaround time of the RT-ddPCR compared to RT-qPCR. Moreover, clinical samples may contain a high virus concentration that would need to be diluted in the RT-ddPCR method. The possible repetition of the detection of the samples that need to be diluted takes more time and makes the RT-ddPCR method a less appropriate method for routine surveillance. However, the virus concentration in wastewater samples is often low, making dilutions often unnecessary. Moreover, the lower impact of inhibition on the RT-ddPCR method makes it an appropriate method for wastewater surveillance. Due to its absolute quantification, the RT-ddPCR method can also be used to evaluate the performances in different laboratories for the inter-laboratory reproducibility and cross-validation of the methods. Because of its potential higher sensitivity, it could also complement the current RT-qPCR diagnostics to improve the rapid identification of SARS-CoV-2 infections, by detecting the virus before the virus concentration peak is reached and antibodies appear in a diagnostic sample.

In addition to the successful development and validation of the proposed multiplex RT-ddPCR method, a methodology to systematically evaluate and monitor PCR-based methods targeting evolving viruses such as SARS-CoV-2 is provided in this manuscript. This methodology includes a method performance assessment in terms of specificity (*in silico* and experimentally tested), sensitivity and applicability. The main added value of this methodology is related to the first *in silico* inclusivity assessment step, using a large set of SARS-CoV-2 strains with a high level of diversity, which is not experimentally achievable by collecting samples and testing them. Indeed, even by testing a large collection of samples, laboratories are not able to test a representative collection of samples that deals with this diversity that is continuously evolving and that needs to be seen not only locally but worldwide. Therefore, we believe that at the present time, this first *in silico* inclusivity assessment step is essential for the development and validation of PCR-based methods targeting the virus, as well as for its continuous evaluation using the newest available WGS data, which are generated over time. Moreover, an additional added value of this methodology is related to the essential experimental testing. Indeed, for the sake of efficiency and simplicity, it should be designed to use a minimal number of critical samples (as proposed in the present study) to assess the performance of the methods (specificity, sensitivity, applicability).

CHAPTER 8: ESTABLISHING QUALITY CRITERIA TO CHARACTERISE SARS-COV-2 AND ITS VARIANTS IN NGS DATA

Context of this chapter:

This chapter illustrates the added value of using NGS for the surveillance of SARS-CoV-2 based on wastewater sampling. Although RT-qPCR and RT-ddPCR methods are fast, simple and relatively inexpensive methods, these methods are ill suited to detect newly emerging variants. RT-qPCR and RT-ddPCR methods have been developed to detect a selection of the mutations assigned to specific variants of concern. However, a variant typically contains multiple mutations that sometimes overlap with other variants. Therefore, an additional step of whole genome sequencing is required to fully characterise the variant's sequence. This chapter is a first step to explore SARS-CoV-2-targeted nucleotide sequencing of wastewater as an epidemiological surveillance method to estimate the prevalence, the genetic diversity and geographical distribution of SARS-CoV-2. Few quality criteria are available when sequencing wastewater samples, and these are generally only applicable for consensus sequence construction. Therefore, the previously developed low-frequency variant detection workflow for influenza from Chapter 6 was adapted for the detection of several SARS-CoV-2 variants in wastewater samples. By using *in silico* modified sequencing data, thresholds for the allelic frequency and coverage were established. The work developed in this chapter, is to our knowledge, the first study paving the way for the detection and quantification of several variants including minority variants in wastewater.

This chapter was previously published:

Laura A. E. Van Poelvoorde, Thomas Delcourt, Wim Coucke, Phillippe Herman, Sigrid C.J. De Keersmaecker, Xavier Saelens, Nancy H.C. Roosens and Kevin Vanneste (2021) Strategy and Performance Evaluation of Low-Frequency Variant Calling for SARS-CoV-2 Using Targeted Deep Illumina Sequencing. Frontiers in Microbiology. **12**. https://doi.org/10.3389/fmicb.2021.747458

Authors' contributions:

Conceptualization: NR, KV, and XS; Project administration: NR; Data curation, investigation, and visualisation: LVP; Methodology: LVP, TD, WC, SDK, NR, and KV; Software and formal analysis: LVP, TD, and WC; Validation: LVP and TD; Writing – original draft preparation: LVP, TD, NR, and KV; Funding acquisition: NR and PH Supervision: NR and KV; Writing – review and editing: All authors

Abstract:

The ongoing COVID-19 pandemic, caused by SARS-CoV-2, constitutes a tremendous global health issue. Continuous monitoring of the virus has become a cornerstone to make rational decisions on implementing societal and sanitary measures to curtail the virus spread. Additionally, emerging SARS-CoV-2 variants have increased the need for genomic surveillance to detect particular strains because of their potentially increased transmissibility, pathogenicity and immune escape. Targeted SARS-CoV-2 sequencing of diagnostic and wastewater samples has been explored as an epidemiological surveillance method for the competent authorities. Currently, only the consensus genome sequence of the most abundant strain is taken into consideration for analysis, but multiple variant strains are now circulating in the population. Consequently, in diagnostic samples, potential coinfection(s) by several different variants can occur or quasispecies can develop during an infection in an individual. In wastewater samples, multiple variant strains will often be simultaneously present. Presently, quality criteria are mainly available for constructing the consensus genome sequence, and some guidelines exist for the detection of coinfections and quasispecies in diagnostic samples. The performance of detection and quantification of low-frequency variants using whole genome sequencing (WGS) of SARS-CoV-2 remains largely unknown. Here, we evaluated the detection and quantification of mutations present at low abundances using the mutations defining the SARS-CoV-2 lineage B.1.1.7 (alpha variant) as a case study. Real sequencing data were *in silico* modified by introducing mutations of interest into raw wild-type sequencing data, or by mixing wild-type and mutant raw sequencing data, to construct mixed samples subjected to WGS using a tiling amplicon-based targeted metagenomics approach and Illumina sequencing. As anticipated, higher variation and lower sensitivity were observed at lower coverages and allelic frequencies. We found that detection of all low-frequency variants at an abundance of 10%, 5%, 3% and 1%, requires at least a sequencing coverage of 250X, 500X, 1500X and 10,000X, respectively. Although increasing variability of estimated allelic frequencies at decreasing coverages and lower allelic frequencies was observed, its impact on reliable quantification was limited. This study provides a highly sensitive low-frequency variant detection approach, which is publicly available at https://galaxy.sciensano.be, and specific recommendations for minimum sequencing coverages to detect clade-defining mutations at certain allelic frequencies. This approach will be useful to detect and quantify low-frequency variants in both diagnostic (e.g., coinfections and quasispecies) and wastewater (e.g., multiple VOCs) samples.

8.1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of the ongoing COVID-19 pandemic [935]. To limit the spread of disease, governments were forced to take drastic measures due to the high potential for human-to-human transmission and the lack of immunity in the population [936]. SARS-CoV-2 spreads very easily during close person-to-person contact [937]. Consequently, the individual diagnostic testing for SARS-CoV-2 on respiratory samples using reverse transcription quantitative polymerase chain reaction (RT-qPCR) is essential for the diagnosis of patients presenting COVID-19 symptoms for appropriate clinical treatment and isolation, as well as for tracing potential contact transmissions, including asymptomatic individuals. Systematic individual SARS-CoV-2 diagnostics are also used to test certain population cohorts, such as primary caregivers, to avoid transmission of the virus to vulnerable people, such as the elderly [938].

Data from individual diagnostics are also collected and analysed for surveillance by National Reference Centres to assist governments to monitor the epidemiological situation. The efficiency of this strategy for epidemiological monitoring depends greatly on the extent of testing the complete population. Additionally, it may be biased by the willingness of individuals, covering all population ages, to get tested, whether individuals are aware of being infected, and visitors to a certain country not always being included in the testing strategy. Moreover, despite having a relatively low per-sample cost, the high volume of required tests incurs substantial costs for public health systems for which testing capacities can be exceeded during periods of intense circulation of the virus [939]. The detection of newly emerging SARS-CoV-2 strains may be delayed by the lack of testing during such periods. As SARS-CoV-2 virus particles and mRNA have been isolated from faeces of COVID-19 patients [940, 941], monitoring of wastewater for SARS-CoV-2 has been explored as a complementary and independent alternative for epidemiological surveillance for the competent authorities [942]. Various studies have observed an association between an increase in reported COVID-19 cases and an increase of SARS-CoV-2 RNA concentrations in wastewater [414, 415]. Wastewater-based monitoring could therefore be a cost-effective, non-invasive, easy to collect, and unbiased approach to track circulating virus strains in a community [943]. Compared to clinical surveillance, wastewater surveillance could also provide opportunities to estimate the prevalence of the virus and assess its geographical distribution and genetic diversity [416, 417], and can be used as a non-invasive early-warning system for alerting public health authorities to the potential (re-)emergence of COVID-19 infections [944]. Alternatively, the absence of the virus in wastewater surveillance could indicate that an area can be considered at low risk for SARS-CoV-2 infections [942].

Although the mutation rate of SARS-CoV-2 is estimated as being low compared to other RNA viruses [896], several new variants carrying multiple mutations have already emerged. Some of these variants are characterised by a potential enhanced transmissibility, and can cause more severe infections and/or potential vaccine escape [262–264, 272, 898]. Consequently, monitoring current and potential future variants is crucial to control the epidemic by taking timely measures because these variants can affect epidemiological dynamics, vaccine effectiveness and disease burden.

To monitor SARS-CoV-2 variants, RT-qPCR methods were designed to detect a selection of the mutations that define specific variants of concern (VOCs). VOCs, are, however defined by a combination of multiple mutations and only few mutations can be targeted by RT-qPCR assays. This approach is not sustainable because it is likely that the ongoing vaccination and increased herd immunity will result in the selection of new mutations and emergence of new VOCs [945], as has been observed with other viruses [121, 946]. Since only a few mutations can be targeted by a RT-qPCR assay, an additional step of whole genome sequencing (WGS) is required to fully confirm the variant's sequence [947].

Whole genome sequencing has been used to understand the virus evolution, epidemiology and impact of SARS-CoV-2 resulting in, as of July 2021, more than 2,000,000 publicly available SARS-CoV-2 genome sequences, mainly derived from respiratory samples that are frequently submitted to the Global Initiative on Sharing Avian Influenza Data (GISAID) database [752]. Most of these sequences were obtained using amplicon sequencing in combination with the Illumina or Nanopore technology, with Illumina still being the most commonly used method [752, 948]. This large amount of genomes allows reliable detection of variants based on the consensus genome sequence in patient samples [387, 949–951]. The

European Centre for Disease Prevention and Control (ECDC) has defined several quality criteria for diagnostic samples depending on the application. For most genomic surveillance objectives, a consensus sequence of the (near-)complete genome is sufficient and a minimal read length of 100 bp and minimal coverage of 10× across more than 95% of the genome is recommended. To reliably trace direct transmission and/or reinfection, a higher sequencing coverage of 500× across more than 95% of the genome is recommended for determining low-frequency variants (LFV) that can significantly contribute to the evidence for reinfection or direct transmission. In-depth genome analysis, including recombination, rearrangement, haplotype reconstruction and large insertions and deletions (indel) detection, should be investigated using long-read sequencing technologies with a recommended read length of minimally 1000 bp and a sequencing coverage of 500× across more than 95% of the genome [615]. A few studies evaluated quasispecies in diagnostic samples by only evaluating positions with a minimum depth of 100× [952], by employing a minimum AF of 2% and a minimum depth of 500× [953] or by using LoFreq with a false discovery rate cut-off of 1%, minimum coverage of 10×, dynamic Bonferroni correction for variant quality and strand bias filtering [954]. Due to the high cost of sequencing large quantities of samples from individual patients, samples that tested positive for a selection of mutations related to VOCs using RT-qPCR and have a sufficiently high viral load are typically sequenced. Consequently, only a subset of all circulating variants is detected during routine clinical surveillance. Since wastewater samples contain both SARS-CoV-2 RNA from symptomatic and asymptomatic individuals, sequencing wastewater samples can provide a more comprehensive picture of the genomic diversity of SARS-CoV-2 circulating in the population compared to individual diagnostic testing and sequencing. Wastewater surveillance of SARS-CoV-2 may therefore be of considerable added value for SARS-CoV-2 genomic surveillance by providing a cost-effective, rapid, and reliable source of information on the spread of SARS-CoV-2 variants in the population.

Sequencing of wastewater samples is, however, currently mainly used to reconstruct the consensus genome sequence of the most prevalent SARS-CoV-2 strain in the sample and LFV are often not investigated [414, 955–957]. This consensus sequence can be useful to demonstrate that the detected strain in wastewater corresponds to the dominant strain that circulates in individuals within the same community [956]. However, similarly to diagnostic samples, only limited quality criteria are in place when sequencing wastewater samples and those available often only apply for consensus sequence construction. The EU recommends the generation of one million reads per sample and a read length of more than 100 bp [942]. A few studies evaluated LFV in wastewater samples, by using local haplotype reconstruction with ShoRAH [958] or iVar and setting up a minimum coverage of 50×, Phred of ≥30 and a minimal allelic frequency (AF) of 10% [409] or a minimum base quality filter of 20 with a minimum coverage of 100× [959]. However, none of these studies evaluated their approach on well-defined populations nor determined detection thresholds for retaining LFV. Since multiple VOCs may co-circulate in a given population, their relative abundance is expected to vary and potentially be very low in wastewater samples. While genome consensus variant calling workflows can only identify mutations present at high AFs, LFV calling methods have been specifically designed to call mutations at lower-than-consensus AFs, and are required to detect VOCs in wastewater samples that are present at an AF below 50%. Appropriate tools and statistical approaches should be provided to ensure reliable and comparable collection and analysis of data, because the detection of LFV is challenging due to the drop in confidence of called mutations at low AFs and sequencing coverages [659, 960, 961]. High-quality sequencing reads are required to ensure that single nucleotide variants (SNVs) and indels can be reliably called and quantified. Most LFV calling algorithms therefore consider multiple sequencing characteristics such as strand bias, base quality, mapping quality, sequence context, and AF [663] to delineate true variants from sequencing errors. Although the viral diversity in multiple WGS-based studies has been explored using several variant calling methods [754, 858, 863], they are often not benchmarked against defined viral populations, rendering the feasibility of using these methods for detecting SARS-CoV-2 VOCs in mixed samples for wastewater surveillance largely unknown.

In this study, we evaluate the performance of LFV detection and quantification based on targeted SARS-CoV-2 sequencing for mutations present at low abundances via the Illumina technology. We used mutations that define the B.1.1.7 lineage as a proof of concept. Using two real sequencing datasets that were *in silico* modified by either introducing mutations of interest into raw wild-type sequencing datasets or mixing wild-type and mutant raw sequencing data, we provide guidelines for minimum sequencing coverages to detect clade-defining mutations at specific AFs. This approach can be used to detect and quantify LFV in diagnostic samples (e.g., to detect co-infections and quasispecies) and wastewater samples (e.g., to detect multiple strains circulating in the population).

8.2. Methods

8.2.1. Employed Sequencing Data and Generation of Consensus Genome Sequences

SARS-CoV-2 raw sequencing data from 316 samples was downloaded from the Sequence Read Archive (SRA) [765]. A random selection of samples was done on the 27th of January 2021 from the COVID-19 Genomics UK (COG-UK) consortium (PRJEB37886) including only samples with a submission date in January 2021, sequenced with Illumina Novaseq 6000 and using an amplicon-based enrichment strategy (Supplementary File S8.1).

To ensure correct pairing of fastq files, all samples were re-paired using BBMap v38.89 repair.sh with default settings [962] (Figure 8.1: Step 1). The consensus genome sequences were generated for all these samples (Figure 8.1: Step 2). The workflow was built using the Snakemake workflow management system using python 3.6.9 [963]. Next, the re-paired paired-end reads were trimmed using Trimmomatic v0.38 [610] setting the following options: “LEADING:10”, “TRAILING:10” “SLIDINGWINDOW:4:20”, and “MINLEN:40”. As reference genome for read mapping of the SRA samples, the sequence with GISAID accession number EPI_ISL_837246 was used for the wild-type samples, while EPI_ISL_747518 was used for the mutant samples. Both references were chosen based on the fact that they should have a complete genome according to GISAID. Additionally, these were chosen to be as close to the SRA data as possible based on their location of sampling (i.e., United Kingdom), sampling date that was in the same period as the data obtained from SRA (i.e., December 2020-January 2021), and whether or not it was classified as belonging to the B.1.1.7 lineage. These reference genomes were indexed using Bowtie2-build v2.3.4.3 [625]. Trimmed reads were aligned to their respective reference genomes using Bowtie2 v2.3.4.3 using default parameters. The resulting SAM files were converted to BAM files using SAMtools view v1.9 [627] and sorted and indexed using the default settings of respectively SAMtools sort and SAMtools index v1.9. Using the sorted BAM file, a pileup file was generated with SAMtools mpileup v1.9 using the options “--count-orphans” and “--VCF”. Next, the variants were called with bcftools call v1.9 using the options “-O z”, “--consensus-caller”, “--variants-only” and “ploidy 1”, and converted and indexed to uncompressed VCF files with respectively bcftools view v1.9 using the options “--output-type v” and bcftools index v1.9 using the option “--force”. Lastly, a temporary consensus sequence was generated using bcftools consensus v1.9 with default settings, providing the reference genome and produced VCF file as inputs. Afterward, the previous steps were repeated once with the same options using the generated temporary consensus sequence as fasta reference to generate the final consensus sequence. These sequences were used to confirm either the presence or absence of the clade-defining mutations of the B.1.1.7 mutant for both the mutant and wild-type samples respectively (Table 8.1). To extract the sequencing coverage for each position and subsequently calculate the median coverage for each sample, SAMtools depth v1.9 was used on the BAM files. Additionally, bamreadcount v0.8.01 (https://github.com/genome/bam-readcount) was run on all samples using the BAM files to determine the coverage at each position.

Table 8.1: Mutations linked to SARS-CoV-2 lineage B.1.1.7 [964]. The first, second, and third columns present respectively the gene name, cDNA-level mutation and protein-level mutation. The last column describes whether the position is covered by one or two amplicons from the enrichment panel (Supplementary Table S8.1). (*) One adaptation was observed for position 26 801. In the wild-type strains a G was observed in contrast to Rambaut et al. where a T was observed. (**) Due to the tiled amplicon approach used to amplify the samples prior to sequencing, the regions where amplicons overlapped resulted in a double coverage. Mutation C27972T was positioned in such an overlap in the wild-type, but not in the mutant. (WT = wild-type).

Gene	Nucleotide-level mutation	Amino Acid-level mutation	Number of amplicons covering the position?
ORF1ab	C913T	Synonymous	1
	C3267T	T1001I	1
	C5388A	A1708D	1
	C5986T	Synonymous	1
	T6954C	I2230T	1
	11288-11296 deletion	SGF 3675-3677 deletion	1
	C14676T	Synonymous	1
	C15279T	Synonymous	1
	C16176T	Synonymous	2
S	21765-21770 deletion	HV 69-70 deletion	1
	21991-21993 deletion	Y144 deletion	2
	A23063T	N501Y	1
	C23271A	A570D	1
	C23604A	P681H	1
	C23709T	T716I	1
	T24506G	S982A	1
	G24914C	D1118H	2
M	G26801C*	Synonymous	1
Orf8	C27972T	Q27stop	WT: 2; B.1.1.7: 1**
	G28048T	R52I	1
	A28111G	Y73C	2
N	G28280C A28281T T28282A	D3L	2
	C28977T	S235F	1

From the initial 316 samples, ten mutant samples were selected that presented similar coverage depths at the positions of interest after normalisation (see below). These samples contained the mutations assigned to the B.1.1.7 variant. Ten wild-type samples were also chosen that did not contain any of these mutations (Table 8.1, Table 8.2) and also presented similar coverage depth at the positions of interest after normalisation. Lineage B.1.1.7, termed Variant of Concern (VOC) 202012/01 by Public Health England (PHE) [965], 20I/501Y.V1 by Nextstrain [260] and alpha variant by the World Health Organization [259], was first reported in the United Kingdom but became the dominant strain in many European countries until the emergence of the delta variant since mid-April 2021 [966]. The B.1.1.7 variant was found to be more transmissible [262] and may cause more severe infections [263, 264]. Lineage B.1.1.7 is defined by multiple spike protein changes, including deletion 69-70 and deletion 144 in the N-terminal domain, amino changes N501Y in the receptor-binding domain, and amino acid changes A570D, P681H, T716I, S982A, D1118H, as well as mutations in other genomic regions [265]. More recently PHE has reported B.1.1.7 cases with an additional mutation, E484K [965]. Median coverages of the selected samples were consistently high (minimum 13,848×; maximum 36,255×) and median read lengths were always 221 and 201 for the forward and reverse reads respectively (Table 8.2). Additionally, as suggested by ECDC, more than 95% of the genome was covered by reads with a minimal coverage of 500× [615].

Table 8.2: List of SRA accession numbers used for employed wild-type and lineage B.1.1.7 samples in this study. Sample IDs, categorized as WT or mutant and the median coverage calculated using Samtools depth v1.9 **[627]**. (WT = wild-type)

Sample	WT/lineage B.1.1.7	Median coverage
ERR5058968	lineage B.1.1.7	13,848
ERR5059033	lineage B.1.1.7	21,874
ERR5059072	lineage B.1.1.7	14,628
ERR5059092	lineage B.1.1.7	16,106
ERR5059123	lineage B.1.1.7	17,349
ERR5059204	lineage B.1.1.7	18,149
ERR5059226	lineage B.1.1.7	22,194
ERR5059238	lineage B.1.1.7	27,681
ERR5059260	lineage B.1.1.7	23,975
ERR5059282	lineage B.1.1.7	27,349
ERR5039162	WT	20,071
ERR5040499	WT	24,440
ERR5059083	WT	18,220
ERR5059114	WT	14,580
ERR5059133	WT	19,866
ERR5059154	WT	28,295
ERR5059253	WT	23,798
ERR5059257	WT	25,894
ERR5059283	WT	36,255
ERR5059286	WT	29,847

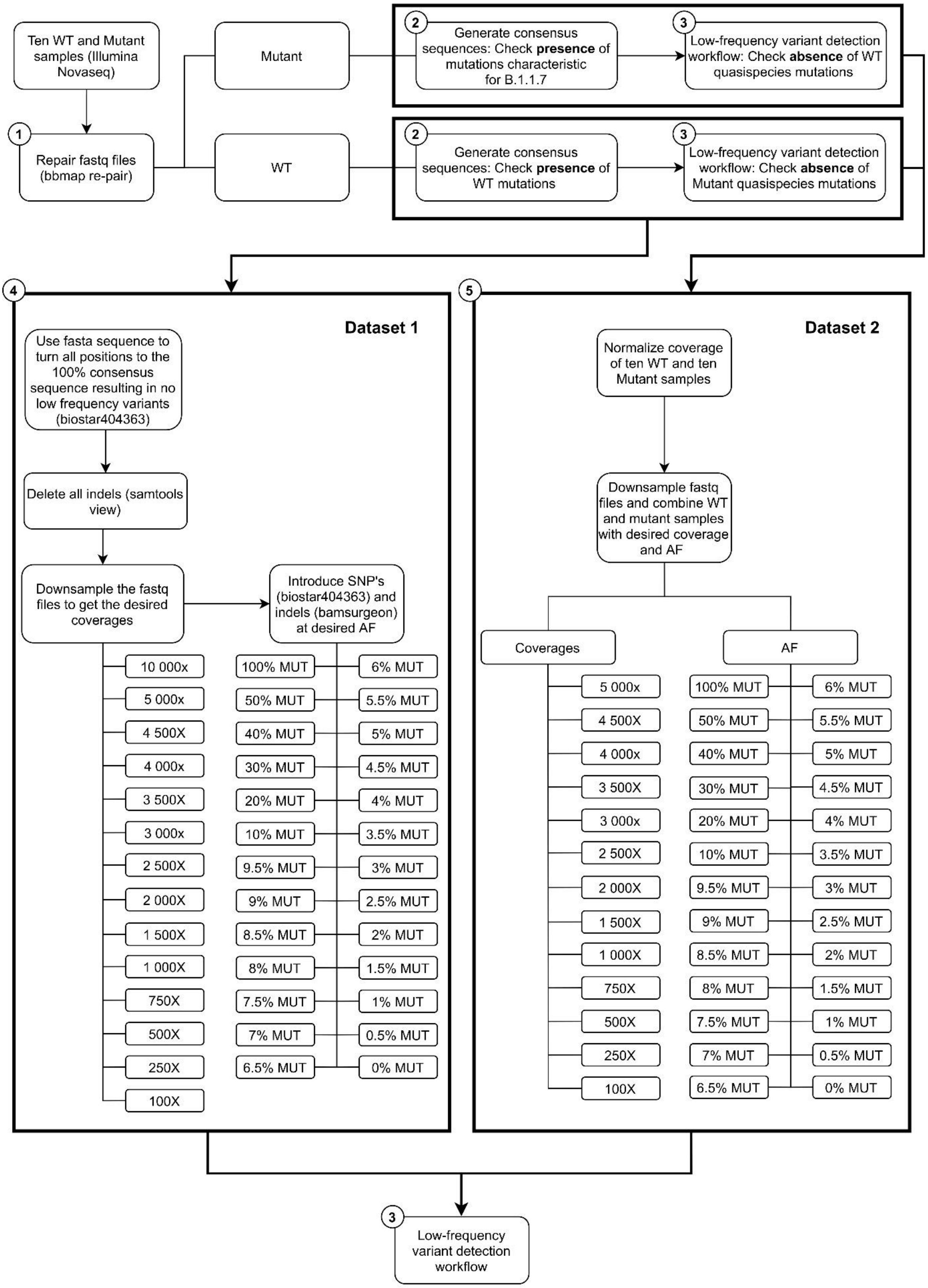

Ten WT and Mutant samples (Illumina Novaseq)
1
Repair fastq files (bbmap re-pair)
Mutant
WT
2
Generate consensus sequences: Check **presence** of mutations characteristic for B.1.1.7
3
Low-frequency variant detection workflow: Check **absence** of WT quasispecies mutations
2
Generate consensus sequences: Check **presence** of WT mutations
3
Low-frequency variant detection workflow: Check **absence** of Mutant quasispecies mutations
4
Dataset 1
Use fasta sequence to turn all positions to the 100% consensus sequence resulting in no low frequency variants (biostar404363)
Delete all indels (samtools view)
Downsample the fastq files to get the desired coverages
Introduce SNP's (biostar404363) and indels (bamsurgeon) at desired AF
10 000x
5 000x
4 500X
4 000x
3 500X
3 000x
2 500X
2 000X
1 500X
1 000X
750X
500X
250X
100X
100% MUT
50% MUT
40% MUT
30% MUT
20% MUT
10% MUT
9.5% MUT
9% MUT
8.5% MUT
8% MUT
7.5% MUT
7% MUT
6.5% MUT
6% MUT
5.5% MUT
5% MUT
4.5% MUT
4% MUT
3.5% MUT
3% MUT
2.5% MUT
2% MUT
1.5% MUT
1% MUT
0.5% MUT
0% MUT
5
Dataset 2
Normalize coverage of ten WT and ten Mutant samples
Downsample fastq files and combine WT and mutant samples with desired coverage and AF
Coverages
AF
5 000x
4 500X
4 000x
3 500X
3 000x
2 500X
2 000X
1 500X
1 000X
750X
500X
250X
100X
100% MUT
50% MUT
40% MUT
30% MUT
20% MUT
10% MUT
9.5% MUT
9% MUT
8.5% MUT
8% MUT
7.5% MUT
7% MUT
6.5% MUT
6% MUT
5.5% MUT
5% MUT
4.5% MUT
4% MUT
3.5% MUT
3% MUT
2.5% MUT
2% MUT
1.5% MUT
1% MUT
0.5% MUT
0% MUT
3
Low-frequency variant detection workflow

Figure 8.1: Schematic representation of the workflow.

8.2.2. Low-Frequency Variants Detection

The absence of pre-existing wild-type and mutant LFV at the positions defining lineage B.1.1.7 (Table 8.1) was verified in both the mutant and wild-type samples (Figure 8.1: Step 3), respectively, by calling all LFV in these samples and subsequently checking the positions of interest. Python 3.6.9 was used with the packages pysam 0.16.0.1 [767] and numpy 1.19.5 [967]. Each generated (final) consensus FASTA file for each sample coming from SRA was used as reference for its respective sample and indexed using SAMtools faidx v1.9 and Bowtie2-build v2.3.4.3. Bowtie2 v2.3.4.3 was then used to align the reads of each sample to its reference sequence, producing a SAM file that was converted into BAM using SAMtools view v1.9. Next, reads were sorted using Picard SortSam v2.18.14 [609] with the option “SORT_ORDER=coordinate” and Picard CreateSequenceDictionary v2.18.14 [609] was used to generate a dictionary of the reference FASTA file. Picard AddOrReplaceReadGroups v2.18.14 [609] was afterward run on the reads with the flags “LB”, “PL”, “PU” and “SM” set to the arbitrary placeholder value “test”. The resulting BAM files were indexed using SAMtools index v1.9 and used as input for GATK RealignerTargetCreator 3.7 [662], which was followed by indel realignment using GATK IndelRealigner v3.7 [662]. Next, generated BAM files were indexed using SAMtools index v1.9. The call function of the LoFreq v2.1.3.1 package [659] was used to call LFV in the BAM files and generate a VCF file using the options “--call-indels” and “--no-default-filter” and using the consensus sequence as reference to call LFV. Next, the unfiltered VCF file was filtered using the filter function of the LoFreq v2.1.3.1 package, setting the strand bias threshold for reporting a variant to the maximum allowed value by using the option “--sb-thresh 2147483647” to allow highly strand-biased variants to be retained, to account for the non-random distribution of reads due to the design of the amplification panel. All employed scripts are available in Supplementary File S8.2. Additionally, the workflow is also available at the public Galaxy instance of our institute at https://galaxy.sciensano.be as a free resource for academic and non-profit usage. The presence of the nucleotides assigned to the B.1.1.7 lineage or the wild-type (Table 8.1) was verified for the mutant and wild-type samples, respectively. Additionally, it was checked that there were no LFV at these positions, so that the wild-type nucleotide or mutant nucleotide was always present at 100% for the retained 10 WT and 10 mutant samples.

8.2.2.1. Dataset 1: *In silico* Insertion of Mutations of Interest into Raw Sequencing Datasets

For the first dataset (Figure 8.1: Step 4), all low-frequency single nucleotide polymorphisms (SNPs) were removed from the raw sequencing data of all samples. SNPs were removed using Jvarkit employing biostar404363 [968] by converting all nucleotides to the consensus fasta sequence. Next, all ten WT samples were down-sampled using "seqtk sample" with argument "-s100" (https://github.com/lh3/seqtk) to 14 different (median) coverages (100, 250, 500, 750, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, and 10,000×). The 22 SNP mutations characteristic for the B.1.1.7 lineage (Table 8.1) were introduced at 26 different AF (mutant: 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 20, 30, 40, 50, and 100%) at the various coverages mentioned above employing biostar404363. This resulted in 10 samples at 364 conditions (i.e., combination of coverage and AF). Next, all reads containing indels were removed from these samples using SAMtools view v1.9. Finally, the three deletions associated with the B.1.1.7 lineage were introduced at the 26 AF mentioned above using BAMSurgeon 1.2 [969], which was adapted to decrease runtime, with the options "-p 10", "--force", "-d 0", "--ignorepileup", "--mindepth 1", "--minmutreads 1", "--maxdepth 1000000", "--aligner mem", and "--tagreads". A minority of reads that were lacking a mate in the targeted regions were removed by using an in-house script making use of Python 3.6.9 and the package pysam 0.16.0.1. Samples in BAM format were then converted back to FASTQ format using bedtools bamtofastq v2.27.1 [970]. Finally the LFV detection workflow (Figure 8.1: Step 3) described in section "Low-Frequency Variants Detection" was used on these 10 samples for all 364 conditions using the FASTA file that was generated for the wild-type samples from SRA as reference with LoFreq.

8.2.2.2. Dataset 2: Introduction of Mutations of Interest by Mixing Wild-Type and Mutant Raw Sequencing Read Datasets

For the second dataset (Figure 8.1: Step 5), the coverage of all 20 samples (Table 8.2) was normalized to 5000× using BBMap v38.89 bbnorm.sh [962] with the options "target=5000", "mindepth=5", "fixspikes=f", "passes=3" and "uselowerdepth=t". However, due to the tiled amplicon approach used to amplify these samples prior to sequencing, regions where amplicons overlapped subsequently had double coverage resulting in two coverages, i.e., 5000 and 10,000×, after normalisation (Supplementary Table S8.1). *In silico* datasets were then generated by mixing the appropriate number of reads for every combination of the ten wild-type and ten mutant samples, resulting in a total of 100 mixed samples, which were down-sampled using "seqtk sample" (with option "–s100") to the appropriate fractions for the required combination of 13 final coverages (100, 250, 500, 750, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500 and 5000×) and 26 AF (mutant: 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 20, 30, 40, 50, and 100%). This resulted in 100 mixed samples at 338 conditions (i.e., combination of coverage and AF). Finally, the LFV detection workflow (Figure 8.1: Step 3) described in section “Low-Frequency Variants Detection“ was used on these samples for all conditions using the FASTA file that was generated for the wild-type samples from SRA as reference, except for samples with 100% AF for the mutant positions where the FASTA file of the mutant sample was used.

Although the second dataset was normalized for total coverage at every genomic position, the tiled amplicon approach resulted in some genomic positions being covered by two overlapping amplicons. Two groups of mutations were therefore obtained for every coverage (Table 8.2), i.e., for a targeted coverage of 5000×, 17 mutations were present at ~5000× (C913T, C3267T, C5388A, C5986T, T6954C, 11288-11296 deletion, C14676T, C15279T, 21765-21770 deletion, A23063T, C23271A, C23604A, C23709T, T24056G, G26801C, G28048T, and C28977T) and 7 mutations were present at ~10,000× (T16176C, 21991-21993 deletion, G24914C, A28111G, G28280C, A28281T, and T28282A). Mutation C27972T was excluded from further analysis, because this position in the wild-type samples was located in a region where amplicons overlapped resulting in a coverage of approximately 10,000×, while in mutant samples it was in a region with no overlap and where a coverage of 5000× was therefore observed (Supplementary Table S8.1). For further analysis, the results were pooled together per theoretical coverage resulting in 24 mutations per coverage but only 17 and 7 mutations at the lowest (i.e., 100×) and highest (i.e., 10,000×) coverage, respectively (Supplementary Table S8.2). The actual median coverage was calculated per theoretical targeted coverage using the output of bamreadcount v0.8.0 of each sample. Using this output, the coverage of each position of interest was extracted (Supplementary Table S8.2).

8.2.3. *Qualitative Evaluation of Detection of B.1.1.7 at Different Abundances*

Since samples of Dataset 1 were normalized for the total median coverage, different individual positions of interest could exhibit deviating coverages. For the qualitative evaluation of LFV detection (i.e., can mutant positions of interest be correctly detected?), the number of false negatives was counted per condition (i.e., combination of AF and coverage) and divided by the total number of observations [i.e., the number of samples (n = 10) and number of mutations considered for that condition (n = 25)]. A mutant position of interest was considered as correctly detected as soon as it was detected by LoFreq, irrespective of its estimated AF.

Dataset 2 was subjected to the same qualitative evaluation as described for Dataset 1. The number of false negatives per condition was divided by the number of observations (i.e., the number of samples (n=100) and number of mutations considered for that condition [either n = 7, n = 17 or n = 24)].

The visualisation of the qualitative evaluation was performed using a contour plot from the R package plotly (RStudio 1.0.153; R3.6.1) [971]. The false negative (FN) proportion in the qualitative evaluation plots ranged from 0 to 1 with a step size of 0.1.

8.2.4. Quantitative Evaluation of Detection of B.1.1.7 at Different Abundances

For the quantitative evaluation of LFV detection (i.e., is the estimated AF of correctly detected mutant positions of interest close to the true AF?) of both datasets, FN values were considered as ‘below the quantification limit’ with the quantification limit equal to the lowest recorded value for that condition (i.e., combination of AF and coverage). Outliers were identified for each condition using the Grubbs test that was sequentially applied by first searching for two outliers at the same side, followed by a search for exactly one outlier. If the p-value of the Grubbs test was below 0.05, outliers were excluded. The standard deviation (SD) and mean value of AF for every condition were estimated by a maximum likelihood model based on the normal distribution that took the FN into account as censor data. Data were modelled according to a normal distribution. If the percentage of FN results was above 75%, the condition was, however, excluded from quantitative evaluation. Finally, a performance metric describing closeness to the true AF was calculated for each targeted AF individually by dividing each pooled squared SD by the maximal pooled squared SD. This metric will range between 0, relatively the closest to the targeted AF, and 1, relatively the furthest from the targeted AF.

As described for the qualitative evaluation, contour plots from the R package plotly (RStudio 1.0.153; R3.6.1) were used for the visualisation of the quantitative evaluation. The performance metric in the quantitative evaluation plots ranged from 0 to 1 with a step size of 0.1.

8.3. Results

8.3.1. *Qualitative Evaluation Demonstrates That B.1.1.7 Clade-Defining Mutations Can Be Reliably Detected at Low Allelic Frequency When Sequencing Coverage Is Adequately High*

To construct samples using targeted SARS-CoV-2 sequencing with a VOC present at low abundances in the viral population, B.1.1.7 clade-defining mutations were first *in silico* introduced at well-defined AFs and coverages in real sequencing data (“Dataset 1”) of ten wild-type samples, without, however, using any coverage normalisation so that individual mutations could be present at higher or lower coverages compared to the total median genomic coverage due to unevenness of coverage. To assess whether introduced mutations were correctly detected, or alternatively missed as FN, samples of this dataset were analysed using a LFV calling workflow based on LoFreq.

Figure 8.2A depicts the proportion of FN observations, and corresponding values are presented in Supplementary Figure S8.1 and Supplementary Table S8.3, for all evaluated coverages and targeted AFs until 20%. Results for all targeted AFs (including higher values) are presented in Supplementary Figure S8.1 and Supplementary Table S8.3. All LFV could be detected at an AF of 1% at a median coverage of 10,000×. As the coverage decreased, the AF threshold at which no single FN occurred (i.e., perfect sensitivity) increased to 1.5% at 5000×, 3% at 1000×, 5% at 500×, 9.5% at 250×, and 20% at 100×. When allowing a maximum of 10% FN (i.e., sensitivity of 90%), the AF thresholds decreased substantially to 1% at 5000×, 1.5% at 1000×, 2.5% at 500×, 4% at 250×, and 7.5% at 100×. No false positive mutations related to the mutant and wild-type were observed at, respectively, 0 and 100% AF.

A second approach was also considered for constructing samples using targeted SARS-CoV-2 virus sequencing with a VOC present at low abundances, by *in silico* mixing real raw sequencing reads from ten B.1.1.7 samples into ten wild-type samples (“Dataset 2”) for a total of 100 mixes at well-defined AFs and coverages, while applying coverage normalisation so that individual mutations were present at approximately similar coverages for all B.1.1.7 clade-defining positions.

Figure 8.2B depicts the proportion of FN observations, and actual values are presented in Supplementary Figure S8.2 and Supplementary Table S8.4, for all evaluated coverages and targeted AF until 20%. Results for higher targeted AF are presented in Supplementary Figure S8.2 and Supplementary Table S8.4. All LFV could be detected at an AF of 1% at a median coverage of 9792×. As the coverage decreased, the AF thresholds at which no single FN occurred (i.e., perfect sensitivity) increased to 1.5% at 4851×, 3.5% at 969×, 4% at 482×, 7% at 237×, and 20% at 97×. However, when allowing a maximum of 10% FN (i.e., reducing the sensitivity to 90%), the AF thresholds decreased substantially to 1% at 4851×, 2% at 969×, 3% at 482×, 4% at 237×, and 7% at 97×. No false positive mutations related to the mutant and wild-type were observed at 0 and 100%, respectively. Overall, the results for Dataset 1, using the median coverages, and Dataset 2, using the coverages at the positions of interest, were qualitatively similar.

Table 8.3: Qualitative evaluation of Dataset 1 based on false negative proportions per condition until a targeted mutant AF of 20%. The percentage of FN is coloured ranging from 0 (dark) to 1 (light) according to the gradient depicted in Figure 8.2A.

Coverage → **AF ↓**	100	250	500	750	1000	1500	2000	2500	3000	3500	4000	4500	5000	10,000
20.00%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
10.00%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9.50%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9.00%	7%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
8.50%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
8.00%	9%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
7.50%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
7.00%	10%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
6.50%	15%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
6.00%	15%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
5.50%	19%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
5.00%	22%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
4.50%	26%	4%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
4.00%	31%	6%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
3.50%	45%	12%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
3.00%	47%	18%	4%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2.50%	62%	21%	7%	2%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2.00%	70%	32%	14%	7%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%
1.50%	84%	52%	24%	16%	9%	5%	1%	2%	0%	0%	0%	0%	0%	0%
1.00%	96%	77%	54%	35%	28%	15%	8%	6%	6%	3%	2%	2%	2%	0%
0.50%	98%	95%	85%	77%	70%	57%	46%	41%	33%	29%	22%	22%	16%	7%
0.00%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 8.4: Qualitative evaluation of Dataset 2 based on false negative proportions per condition until a targeted mutant AF of 20%. The percentage of FN is coloured ranging from 0 (dark) to 1 (light) according to the gradient depicted in Figure 8.2B.

Coverage → **AF ↓**	**97**	**201**	**237**	**482**	**728**	**969**	**1454**	**1937**	**2413**	**2904**	**3383**	**3872**	**4358**	**4851**	**5855**	**6834**	**7801**	**8790**	**9792**
20.00%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
10.00%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9.50%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9.00%	5%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
8.50%	6%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
8.00%	8%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
7.50%	8%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
7.00%	9%	34%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
6.50%	18%	35%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
6.00%	28%	38%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
5.50%	31%	47%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
5.00%	35%	56%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
4.50%	43%	57%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
4.00%	51%	59%	6%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
3.50%	58%	63%	18%	4%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
3.00%	68%	73%	23%	8%	2%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2.50%	77%	82%	40%	21%	4%	3%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2.00%	81%	84%	55%	33%	11%	6%	4%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
1.50%	89%	86%	69%	53%	24%	21%	12%	8%	4%	2%	1%	0%	0%	0%	0%	0%	0%	0%	0%
1.00%	92%	86%	91%	80%	57%	52%	34%	22%	8%	15%	6%	7%	6%	4%	0%	0%	0%	0%	0%
0.50%	100%	98%	98%	92%	92%	89%	80%	70%	55%	62%	34%	41%	24%	35%	62%	55%	46%	35%	28%
0.00%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

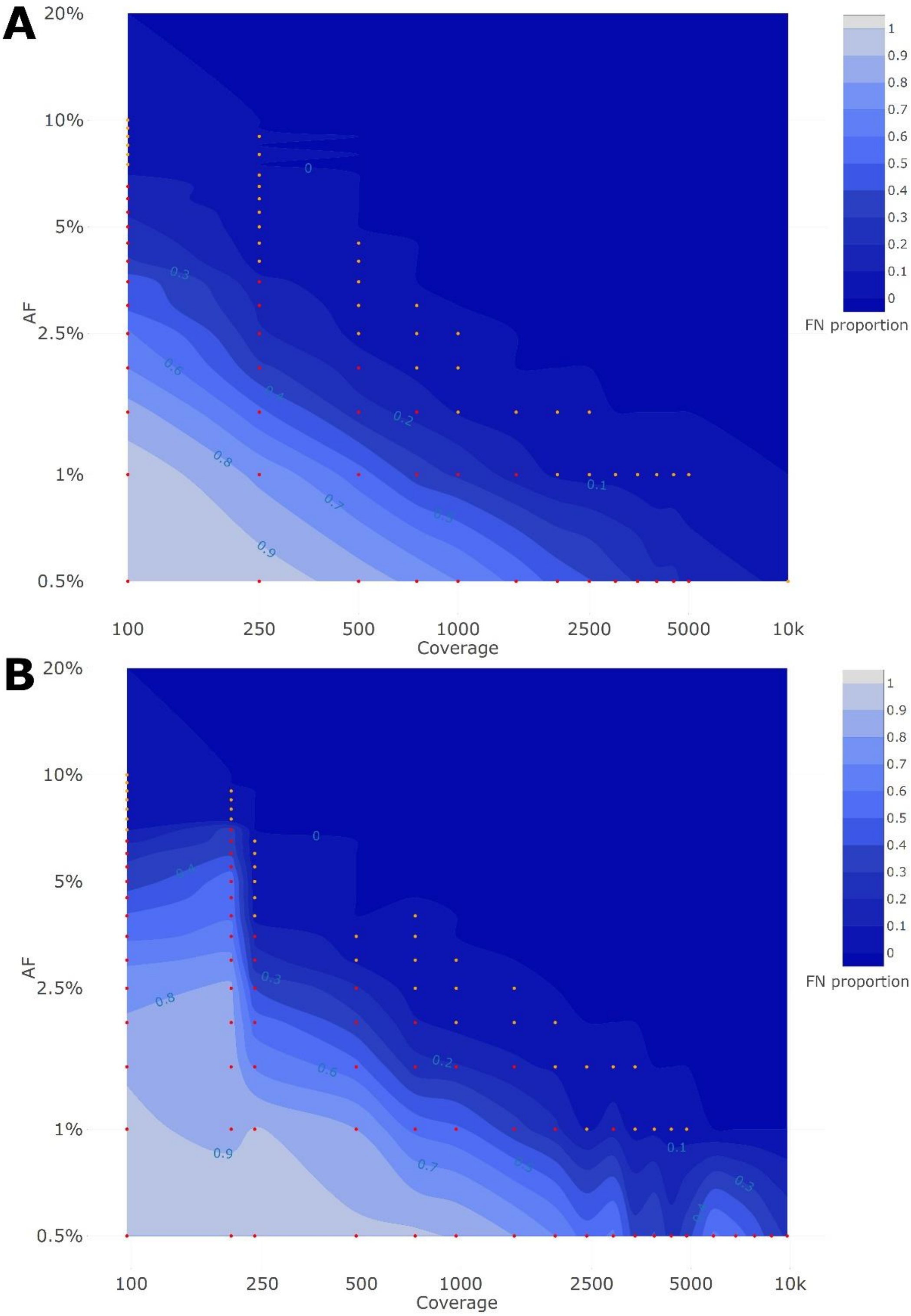


Figure 8.2: Qualitative evaluation of Dataset 1 (A) and Dataset 2 (B) based on false negative proportions per condition until a targeted mutant AF of 20%. Orange and red dots represent conditions with a FN proportion between 0 and 0.1, and between 0.1 and 1, respectively. The percentage of FN is coloured ranging from 0 (dark) to 1 (light) in intervals of 0.1 as extrapolated using a contour plot in the R package plotly **[971]** (actual FN proportions are presented in Table 8.3 for Dataset 1 and Table 8.4 for Dataset 2. Results for targeted mutant AF values >20% are presented in Supplementary Figure S8.1 for Dataset 1 and Supplementary Figure S8.2 for Dataset 2. Both the x- and y-axis follow a logarithmic scale.

8.3.2. Quantitative Evaluation Demonstrates That the Resulting Allelic Frequencies for B.1.1.7 Clade-Defining Mutations Are Close to Their Target Values

To evaluate the possibility of quantifying LFV in both datasets, the SDs of available observations were first evaluated for each condition (i.e., combination of AF and coverage). This provisional analysis indicated that for both Dataset 1 (Supplementary Figure S8.3 and Supplementary Table S8.5) and Dataset 2 (Supplementary Figure S8.4 and Supplementary Table S8.6), the SD systematically decreased per target AF as coverage increased. This provisional analysis also indicated that for both datasets, irrespective of coverage, the SD generally increased between a targeted AF of 1 to 10%, after which it plateaued for targeted AFs above 20%. We therefore employed the squared SD per AF divided by the maximal squared SD per target AF to describe closeness of observed AF to the true AF, for which results are presented in Figure 8.3A for Dataset 1. As expected, the variation in AF estimates fluctuates in function of the median coverage and targeted AF, with variation decreasing per target AF as coverage increased, but also variation being generally more pronounced at low AFs irrespective of coverage. Notwithstanding, even for regions in Figure 8.3A exhibiting high variation, the variability overall remained small (Supplementary Figure S8.3 and Supplementary Table S8.5). The interquartile range (IQR) (Supplementary Figure S8.3D) of the observed AF was still limited at the various targeted AF ranging from 0.62–6.26% at an AF of 50%, 0.36–3.49% at an AF of 10% and 0.27–2.07% at an AF of 5% with the highest IQR observed at lower coverages.

Results for the quantitative evaluation of Dataset 2 are presented in Figure 8.3B, and are in accordance with the trends observed for Dataset 1 with the variation decreasing per target AF as coverage increased, and lower target AFs exhibiting increasing variation irrespective of coverage. Notwithstanding, similarly to Dataset 1, the observed total variation remained small (Supplementary Figure S8.4 and Supplementary Table S8.6). The IQR (Supplementary Figure S8.4D) of the observed AF was limited at the various targeted AF ranging from 0.73–3.93% at an AF of 50%, 0.41–3.93% at an AF of 10% and 0.29–2.27% at an AF of 5% with the highest IQR observed at lower coverages.

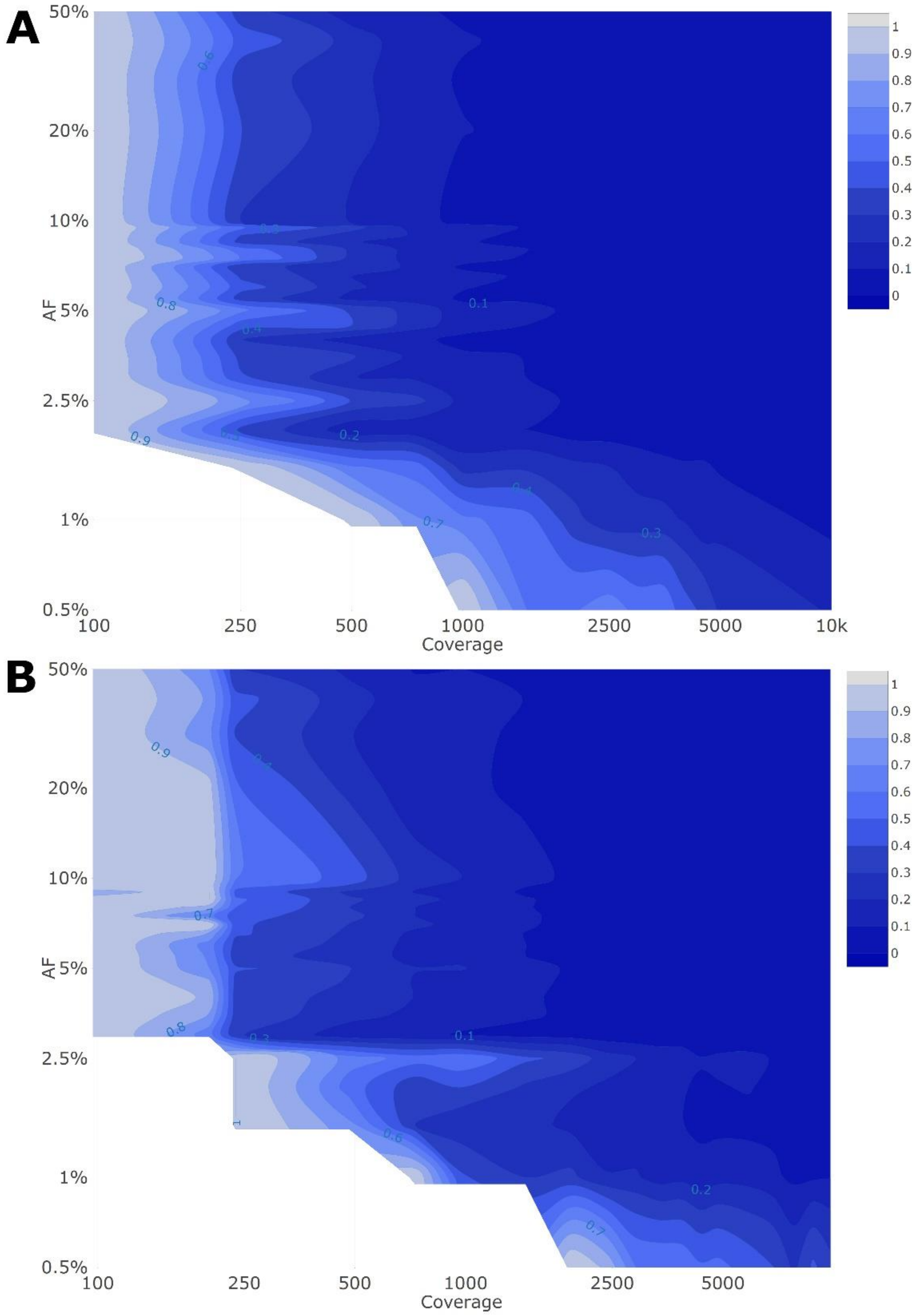


Figure 8.3: Quantitative evaluation of Dataset 1 (A) and Dataset 2 (B) using the squared SD divided by the maximal squared SD per targeted AF. The figure is coloured ranging from 0 (dark) to 1 (light) in intervals of 0.1 as extrapolated using a contour plot in the R package plotly **[971]** (actual values are presented in Supplementary Figure S8.3 and Supplementary Table S8.5 for Dataset 1 and Supplementary Figure S8.4 and Supplementary Table S8.6 for Dataset 2). Both the x- and y-axis follow a logarithmic scale. Conditions with a FN proportion higher than 75% were excluded and correspond to the white plane in the lower left corner.

8.4. Discussion

Whole genome sequencing is a more powerful approach than RT-qPCR to track both existing and newly emerging SARS-CoV-2 variants. WGS is currently, however, mainly used to construct the consensus genome sequence and determine the most prevalent strain in communities, but interest exists in its potential for detecting LFV both within diagnostic samples to detect co-infections and quasispecies, and wastewater samples to determine all circulating variants in a population [942]. To evaluate the potential of targeted amplicon-based SARS-CoV-2 WGS to detect and quantify LFVs at low abundances, we assessed the performance of a workflow designed for LFV detection in WGS data. Mutations defining lineage B.1.1.7 were employed as a proof of concept using an approach based on *in silico* modifying real sequencing data to construct two datasets with the Illumina technology. These two datasets comprise in total 35,100 different samples, which results in a thorough *in silico* analysis requiring a considerable amount of computational calculation hours to validate this approach. For the first dataset, lineage B.1.1.7-defining mutations were introduced *in silico* into raw wild-type sequencing datasets. For the second dataset, the same mutations were introduced by mixing wild-type and B.1.1.7 raw sequencing datasets. In Dataset 1, the coverage profiles of samples corresponded to a typical real dataset including large fluctuations in sequencing coverage at certain positions. In Dataset 2, sequencing coverages were normalized, which allowed evaluating with high precision how reliable AF detection is at specific coverages. Afterward, the ability to both detect and quantify LFV was evaluated. Results demonstrated that WGS enabled detecting LFV with very high performance. As expected, lower coverages and AFs resulted in lower sensitivity and higher variability of estimated AFs. We found, employing the most conservative thresholds from either Datasets 1 or 2, that a sequencing coverage of 250, 500, 1500, and 10,000× is required to detect all LFV at an AF of 10, 5, 3 and 1%, respectively (Table 8.3 and Table 8.4). For quantification of variants, the variability remained overall small for all conditions respecting the thresholds above, resulting in reliable abundance estimations, despite the variability of estimated AF increasing at lower coverages and AF. Of note, it was observed that the profile of the genome coverage differed at some positions between wild-type and mutant samples indicating that the amplicon-based enrichment approach could possibly introduce a bias. Consequently, this should be considered when examining and quantifying the proportion of mutants in samples. Our results can serve as a reference for the scientific community to select appropriate thresholds for the AF and coverage. These could also be context-specific as a smaller or larger degree of false negatives might be warranted for specific applications, and can also be used as a baseline for determining the number of samples that can be multiplexed per run to optimise cost-efficiency of WGS.

With respect to diagnostic samples, this study illustrates it is feasible to use targeted amplicon-based metagenomic approaches to detect co-infections and quasispecies in diagnostic samples. There are currently only limited guidelines available regarding the coverage and AF for such samples and those criteria were not assessed using predefined populations. ECDC has provided limited quality criteria regarding the sequencing coverage, namely 500× across 95% of the genome to detect LFV, but has not indicated the corresponding AF thresholds this corresponds to for reliable LFV detection [615]. Based on the results obtained in this study, a coverage of 500× allowed to detect LFV until an AF of 5% with perfect sensitivity and would therefore be less suited to detect LFV at lower AFs. Lythgoe et al. (2021) recommended a depth of at least 100 reads with an AF of at least 3% to detect the LFV in diagnostic samples with high viral loads (50,000 uniquely mapped reads) [952], while Siqueira et al. (2021) used an AF threshold of 2% and a minimal depth coverage of 500 reads [953] and Karim et al. (2021) adopted an AF of 1% and a minimal depth coverage of 10× [954]. Based on the results in this study, these recommendations appear not sufficiently strict, since we observed that an AF of 1, 2 and 3%, requires at least a sequencing coverage of 10,000, 2500, and 1500× to detect all LFV or 3500, 1000, and 500× to detect 90% of LFV, respectively. However, our study is limited to *in silico* modified data from real diagnostic samples, so these results will need to be validated using real samples with well-established existing LFV in future research.

With respect to wastewater samples, our findings also corroborate the feasibility of using targeted amplicon-based metagenomics approaches for wastewater surveillance, as such samples comprise a collection of different strains, among which the dominant strain will define the consensus sequence of the sample and the detected LFV will represent the circulating strains present at lower frequencies. Only very limited recommendations regarding wastewater sequencing are available by the competent authorities. The EU has recommended the generation of one million reads per sample with a read length of minimum 100 bp which corresponds to a minimum coverage of 3333× using the Lander/Waterman equation [942]. Based on the results obtained in this study, a coverage of 3000 and 3500× allowed to detect LFV until an AF of 2 and 1.5% respectively with perfect sensitivity. Other studies that investigated LFV in wastewater have provided limited quality criteria regarding the coverage and AF. Furthermore, the quality criteria in those studies were not evaluated using a defined population [409, 958]. Izquierdo-Lara used a minimum depth coverage of 50× and minimum AF of 10% [409], while Rios et al. (2021) adopted a minimum depth coverage of 100× without indicating an AF threshold [959]. Based on the results in this study, these recommendations appear not sufficiently strict as a sequencing coverage of 100× and 250× at an AF of 20 and 10% respectively was required to observe all LFV. Obtaining high-quality sequencing reads for wastewater samples may, however, be challenging under real-world conditions. In contrast to diagnostic samples in which viral loads are typically high, ranging from 10^4 to 10^7 copies/mL [410], viral RNA loads in wastewater samples are often low, ranging from 10^{-1} to $10^{3.5}$ copies/mL [972]. This renders it more challenging to sequence samples with a low viral load in addition to the RNA degradation that occurs in wastewater samples. Additionally, variants circulating at low frequencies in a community are expected to be present at a low AF in wastewater samples. Nevertheless, employing the most conservative thresholds from either Datasets 1 or 2, 90% of LFV present at an AF of 10, 5, 3, and 1% were still detected at a sequencing coverage of 100, 250, 500, and 2500×, respectively (Table 8.3 and Table 8.4).

This study focused on the sensitivity of LFV detection and did not explore the false positive rates (i.e., specificity). Although our recommendations for AFs and coverages ensure high sensitivity, often an inverse relationship exists between sensitivity and specificity and we can therefore not exclude that false positives occur for AF and coverage combinations considered as providing qualitative results in this study. A false positive detection is, however, typically less problematic compared to a false negative result as the former can still be discovered in follow-up investigation in contrast to the latter. Additionally, false positive observations typically occur randomly over the genome [663] and it is unlikely that all VOC-defining mutations would be simultaneously falsely detected, even at low AFs and coverages. The issue of low viral loads, low expected AF and potential false positives could be mitigated by sequencing samples in duplicate when necessary. Possible false positive results could be investigated using RT-qPCR or RT-ddPCR assays that target those specific positions.

In this study, the B.1.1.7 variant and a WT (i.e., non-VOC) background of the same time period and location were used as a proof of concept, but can be considered to also apply to other combinations (e.g., two VOCs), since additional VOCs in the sample material will translate into more VOC-defining mutations in the background genomic material that will be independently identified by the variant calling engine. In the presence of multiple VOCs, the VOCs can be identified by composing all possibly existing combinations of LFV as a conservative strategy, although multiple VOCs in one sample will also make the estimation of the relative abundance of each VOC more complicated. If multiple VOCs with partially overlapping defining mutations would be present in a wastewater sample, some mutations of interest would consequently be present at different AFs. Haplotype reconstruction methods could be used in such situations to delineate VOCs. However, most haplotype reconstruction programs perform poorly under higher levels of diversity, and haplotype populations with rare haplotypes are often not recovered [973]. Although haplotype reconstruction has been described for short reads, Nanopore sequencing might offer a substantial advantage for such cases due to its longer reads, despite their higher error rate, to perform haplotype estimation to delineate actual VOCs.

8.5. Conclusion

There exists a pressing need for recommendations for detecting LFV for both diagnostic samples and wastewater surveillance. Further investigation will be required to investigate the specificity and possibility to detect VOCs instead of just mutations, including for other existing and employed methodologies such as probe-based capture, other amplicon-based methods, and Nanopore sequencing. Nevertheless, using *in silico* modified data derived from WGS of real diagnostic samples, this study demonstrates the feasibility of a targeted metagenomics approach for highly sensitive LFV detection with acceptable relative abundance estimations using a tiled-amplicon enrichment based on the Illumina technology. This approach enables the detection of mutations associated with specific VOCs. Our approach could be used to evaluate the potential occurrence of co-infections with other SARS-CoV-2 variants with different strains in diagnostic samples. It can also be employed to detect multiple strains for wastewater surveillance, although several additional challenges exist for wastewater samples such as low viral load and potential RNA degradation. Since in this study, high-quality data from diagnostic samples was used and modified *in silico* to construct datasets to provide guidelines for sequencing wastewater and diagnostic samples with co-infections, future work will need to consider data coming from samples that are closer to real data from actual diagnostic and wastewater surveillance. In light of the pandemic urgency, and the multiple SARS-CoV-2 wastewater surveillance initiatives that are being established and also being integrated into overarching coordination and preparedness initiatives such as the recently announced European Health Emergency Preparedness and Response Authority [942], we hope that our results will help establishing guidance and recommendations for wastewater surveillance and other relevant applications.

CHAPTER 9: GENERAL DISCUSSION AND FUTURE PERSPECTIVES

WGS has the potential to significantly transform virus surveillance. However, many promising applications remain unexploited in the context of routine surveillance which is the core business of NRCs and public health institutes such as Sciensano. Therefore, the aim of this Ph.D. research was to demonstrate how some of these applications can be of added value for virus surveillance but also what may be the limitations that have to be overcome in order to deliver the highest benefit of these approaches. In this thesis, the analysis focused in particular on the influenza genomic surveillance using respiratory samples and SARS-CoV-2 genomic surveillance of wastewater. As a first objective, we assessed the possible added value of WGS to detect the presence of antiviral resistant influenza viruses (Chapter 3). Subsequently, a dataset of 253 influenza A(H3N2) samples was used as a proof of concept to provide a new way of classifying the viruses (Chapter 4), to track mutations across the whole virus genome and link this genomic data with the patient data (Chapter 5), and to propose a methodology which can be used in routine surveillance to detect low-frequency variants in clinical samples (Chapter 6). The advantages of using WGS are not limited to the influenza surveillance, but can also be used for other surveillance systems, such as the SARS-CoV-2 surveillance. By using whole genome sequences, targets for RT-ddPCR assays to detect SARS-CoV-2 in patient and wastewater samples can be designed and evaluated (Chapter 7). Moreover, quality criteria were established based on *in silico* datasets in order to take full advantage of using NGS to try and characterise SARS-CoV-2 and its variants in wastewater (Chapter 8). This dissertation presented several applications of using WGS in routine and pandemic surveillance, demonstrating how virus surveillance can exploit WGS to unlock its full potential. As summarised in Figure 9.1, we will describe in the general discussion what this thesis added and what the way forward is to tackle the remaining challenges. Publications in peer-review journals and presentations at scientific events were used to share this work with the scientific community of which an overview is provided in the Academic CV.

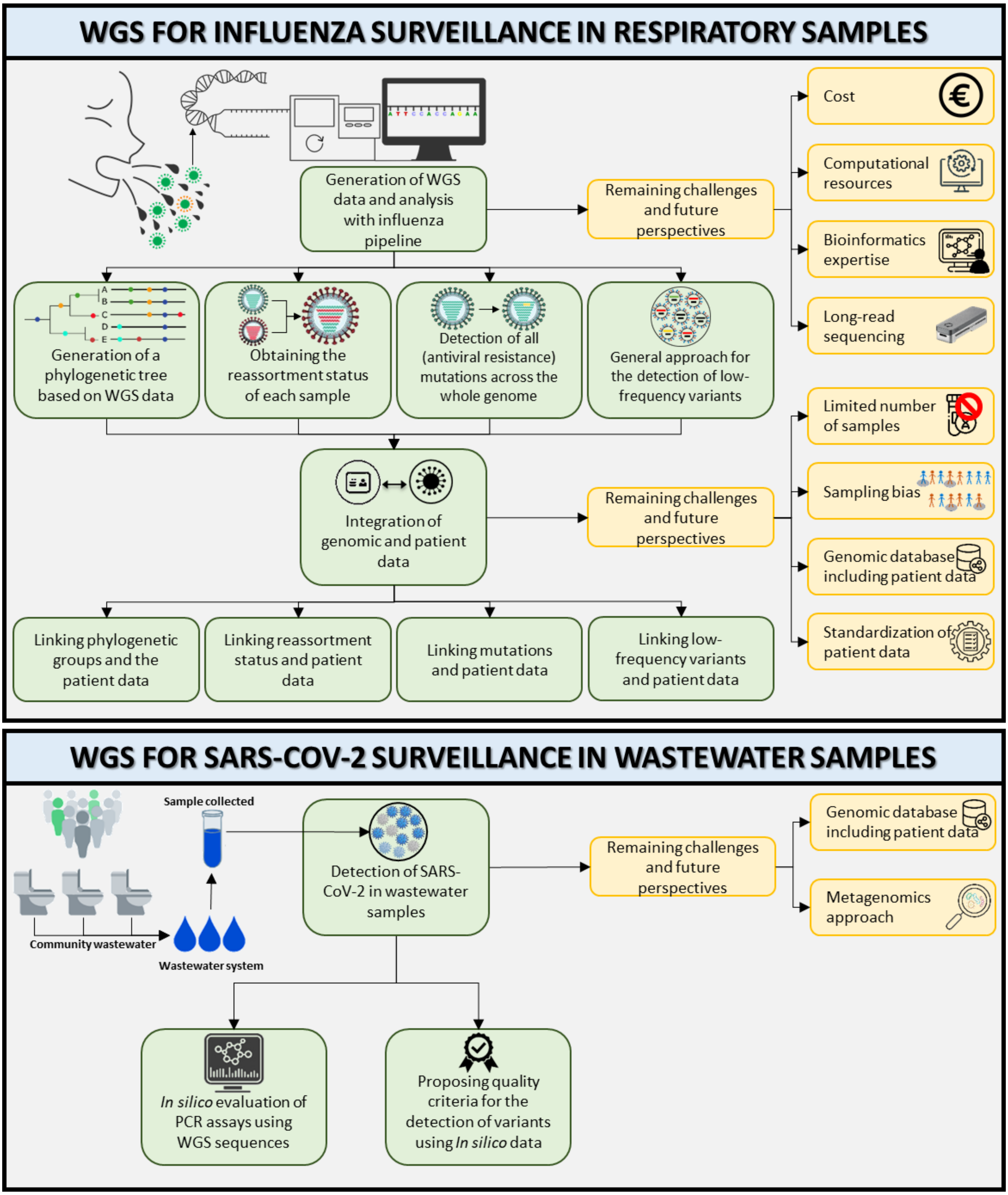

WGS FOR INFLUENZA SURVEILLANCE IN RESPIRATORY SAMPLES
Generation of WGS data and analysis with influenza pipeline
Remaining challenges and future perspectives
Cost
Computational resources
Bioinformatics expertise
Long-read sequencing
A
B
C
D
E
Generation of a phylogenetic tree based on WGS data
Obtaining the reassortment status of each sample
Detection of all (antiviral resistance) mutations across the whole genome
General approach for the detection of low-frequency variants
Integration of genomic and patient data
Remaining challenges and future perspectives
Limited number of samples
Sampling bias
Genomic database including patient data
Standardization of patient data
Linking phylogenetic groups and the patient data
Linking reassortment status and patient data
Linking mutations and patient data
Linking low-frequency variants and patient data
WGS FOR SARS-COV-2 SURVEILLANCE IN WASTEWATER SAMPLES
Sample collected
Community wastewater
Wastewater system
Detection of SARS-CoV-2 in wastewater samples
Remaining challenges and future perspectives
Genomic database including patient data
Metagenomics approach
In silico evaluation of PCR assays using WGS sequences
Proposing quality criteria for the detection of variants using In silico data

Figure 9.1: Schematic outline of what this thesis adds and the way forward. The boxes in green represent what this thesis adds and the boxes in yellow represent what the challenges and future perspectives are and what the way forward may be.

9.1. WGS: Added value and challenges offered by WGS for the influenza surveillance in respiratory samples

The NRC Influenza, which is part of Sciensano, coordinates a sentinel network of hospitals, laboratories and general practitioners that ensure the permanent surveillance of influenza activity, including the impact on the population and its severity and intensity during epidemics. At the start of this thesis, the diagnosis of an influenza sample was mainly done using RT-qPCR and in some cases Sanger sequencing of the HA segment. This thesis is one of the initiatives to implement WGS into public health practices. During this thesis, a protocol was adapted and implemented allowing to obtain genomic information from the whole genome of 253 influenza samples selected from a historic collection of the NRC. WGS allows the genetic characterisation of influenza virus at the highest possible resolution. By obtaining sequence information on the whole genome, the identification of the pathogen is more accurate and it enables numerous applications that can be used for surveillance purposes such as (1) detecting the emergence of antiviral resistance in all segments, (2) classifying the influenza strains based on the whole genome and linking it to patient data, (3) tracking mutations across the whole genome and integrating it with the patient data, and (4) detecting low-frequency mutations that are present in the influenza samples and linking it to patient data (Figure 9.1). Due to the decrease in costs and turnaround times, high-throughput genomic sequencing technologies are becoming a possibility for clinical and public health laboratories. Data in this thesis thus may contribute to the transition from Sanger sequencing to WGS and subsequently integrate this genomic information with patient data.

9.1.1. What this thesis adds

9.1.1.1. The generation and analysis of WGS data

Historically, genomic influenza surveillance has long focused on Sanger sequencing of the HA and the NA segments mainly because of the vaccines and most common antivirals, which target HA and NA, respectively, that are currently used for the management of human influenza. Moreover, the type and subtype of influenza positive samples are determined by a RT-qPCR protocol. In this thesis we adapted and implemented a protocol that enables WGS in a routine surveillance setting and applied this approach to a historical collection of samples from the Belgian NRC, consisting of a dataset of 253 influenza A(H3N2) samples from the 2016-2017 Belgian influenza season. We opted for a method where the virus was enriched using a targeted PCR method, because metagenomics or target enrichment would currently be too costly, laborious and time-consuming for routine surveillance. Moreover, an influenza genome analysis pipeline was made available on the Sciensano Galaxy internal instance (also available for external use) (https://galaxy.sciensano.be/policy/disclaimer.html) which allows obtaining the consensus sequence from influenza samples using Illumina. This facilitates the use of WGS by non-expert bioinformaticians. The availability of the whole genome of influenza has opened several doors that have been explored in this thesis (Figure 9.1).

First, the added value of WGS was illustrated in case of genotypic antiviral resistance surveillance (Chapter 3). Although NA inhibitors are still the most frequently used anti-influenza drugs, new antivirals that target the gene products of other segments are emerging, which increases the need to obtain information about the whole genome. Additionally, WGS provides the opportunity to obtain information about the low-frequency variants in the sample in all eight segments. This presents the opportunity to forecast the emergence of antiviral resistance mutations within a sample from a patient under treatment or a large set of circulating viruses. Consequently, the clinical management of the patient can be adjusted and on a larger scale it can improve the preparedness for a potential outbreak of resistant strains [749].

Furthermore, using the historical collection of the Belgian NRC, the added value of WGS was also demonstrated. At the start of this thesis, Sanger sequencing of the HA segment was the standard for genomic surveillance in Belgium. In Chapter 4**,** we have shown the benefits of WGS and how WGS can improve the current phylogenetic surveillance and reassortment detection of influenza. Indeed, at present, influenza classification is based on the HA segment and follows the guidelines of WHO/ECDC. These guidelines comprise the phylogenetic analysis of the HA segment and the detection of HA amino acid substitutions that are linked to specific clades. However, we have demonstrated that a considerable number of samples could not be classified within these clades either because they did not cluster with reference strains and/or lacked clade-defining amino acid substitutions. Furthermore, the studies within the thesis have shown that the genetic information of the whole genome and the use of more advanced phylogenetic methods has several advantages. It could better inform national influenza prevention and control programmes regarding the timing, impact and severity of seasonal epidemics. First, current surveillance programmes use mostly relatively simple phylogenetic tree reconstruction methods. However, a more robust phylogenomic investigation and deeper exploration of the circulating genetic diversity was obtained using phylogenetic tree construction through Bayesian inference. Also, by using tools, such as Nextstrain, an improved reference selection was obtained. A custom-built Nextstrain instance including the whole genome allows the analysis of several hundreds to thousands of influenza genomes, which allows a temporary classification of the samples and allows the selection of suitable references. Second, more genetic information by including the whole genome allows the classification of more samples into well-supported phylogenetic groups in comparison to only including genetic information from the HA gene, which enabled an improved classification performance. Third, reassortments can be detected by analysing the whole genome. In this study, strict requirements were used by applying both a manual and computational method requiring high support values. This resulted in an observed reassortment rate of 15% which is likely underestimated. Finally, whole genome information cannot only improve the current vaccine strain selection, but it may also even be a requirement in the future as next-generation vaccines and antiviral drugs do not solely target the HA and NA segments of the influenza genome.

Influenza viruses are known for their considerable diversity between hosts. Using the consensus sequences that were obtained using the influenza pipeline, we were able to detect mutations across the whole genome compared to the reference sequence (Chapter 5). In the context of current and future vaccines and antivirals, it may be important to track mutations across the whole genome as they may influence protein functions and interactions, the host environment or the disease progression within the host.

Besides the diversity between hosts, several low-frequency variants can be found within the same host. One of several advantages of deep sequencing with NGS is the higher genome coverage and consequently more reliable estimation of the diversity within the quasispecies population present at very low abundances. Information from low-frequency variants can provide opportunities to broaden our knowledge about its impact on virus evolution, transmission, drug and vaccine resistant strains and pathogenicity. However, experimental errors can be introduced during the PCR and NGS amplification steps. Therefore, there is a need to set up thresholds for the minimal viral load and the minimal allelic frequency to reduce false positive variant detections as much as possible. In this thesis (Chapter 6), a well-defined population was used to set up detection limits of at least 10^4 genomes/µL for the viral load and an allelic frequency of ≥5%. Subsequently, these thresholds were used to analyse the A(H3N2) dataset that was previously used in Chapters 4 and 5, which resulted in a subset of 59 retained influenza A(H3N2) samples that can be used for more in-depth analysis regarding the low-frequency variants.

9.1.1.2. The integration of patient and genomic data

Currently, the patient data are primarily linked to either the presence or absence of influenza in a sample, as determined by RT-qPCR. In this thesis, as one of the first studies, patient data was linked to the genomic influenza data for the Belgian influenza surveillance. We looked at several possibilities using the collection of the NRC as a proof of concept and proposed statistical approaches to analyse this data (Figure 9.1).

In Chapter 4, the patient data was integrated with the phylogenetic groups and reassortment status. This allowed the detection of several associations that would be otherwise missed if only the HA segment would have been considered, such as the observation that more hospitalised than non-hospitalised patients were infected with A(H3N2) reassortants. Moreover, phylogenetic groups can also be linked to disease severity indicators, which could be relevant for epidemiological monitoring.

The collection of influenza samples was also used as a proof of concept to explore potential associations between mutations positioned across the whole influenza genome and patient data in Chapter 5. Using this limited dataset, associations were detected between particular mutations and the sampling period at the Belgian level. The GISAID database is well-known for collecting genomic data worldwide from influenza viruses. Using the GISAID database, it was possible to confirm the associations regarding the sampling period seen at the Belgian level. These mutations may possibly be important for the vaccine strain selection and clinical management of infected patients. Furthermore, the highly diverse genetic background of A(H3N2) strains was considered in Chapter 5 by using a new approach based on sample stratification according to their phylogenetic groups. This approach resulted in the identification of five additional mutations that are significantly associated with renal insufficiency. This result illustrates the potentially important role of the viral genetic background in inferring associations between genomic and patient data.

Besides the mutations that were observed in the consensus sequences, there are also low-frequency variants detected within the influenza samples. Using the approach developed to detect low-frequency variants and considering the determined thresholds, the genomic information, including low-frequency variants, of 59 samples were linked to patient data as a proof of concept. Significant associations between the detected low-frequency variants and patient data were found, which indicates the potential relevance of low-frequency variant detection in routine influenza surveillance programmes.

9.1.2. Way forward

This thesis provides a PCR enrichment-based protocol to obtain the whole genome of influenza, the assembly and the consensus genome sequence of the virus in a given biological sample. An easy-to-use pipeline was built to analyse the Illumina MiSeq data in order that non-bioinformaticians can analyse WGS data. Moreover, it provides several statistical approaches to analyse and link the genomic data with the patient data. However, some challenges remain to be tackled. The higher throughput that is generated by NGS compared to classic Sanger sequencing has created a need for computational resources and data storage solutions, which demands substantial financial resources and technical expertise. By now, simple analysis of NGS data can be performed on high-end computers, however, servers with more RAM, processing power and storage will be needed when throughput increases or if results need to be obtained quickly. Although in our institute sufficient computational power is available for running the influenza pipeline, for some analysis it is still the limiting factor. For example, in Chapter 4 the phylogenetic tree was constructed using Bayesian inference. Although Bayesian inference will probably be the preferred statistical method in the future, the high computational cost and long execution time are often still the limiting factor.

Before NGS can be widely used in routine settings, easy-to-use bioinformatic tools need to be available. A user-friendly influenza pipeline that allows to generate an accurate consensus sequence was developed. However, user-friendly tools for constructing phylogenies, the detection of mutations and low-frequency variants and their association to the patient data have not yet been established. Due to the complexity and the computational resources associated with Bayesian inference, it is currently not trivial to provide an easy-to-use tool. However, a Nextstrain instance based on the whole genome should be possible to implement and would be a more thorough method compared to the current method which only considers a limited number of references. The current approaches used for the extraction of the mutations and low-frequency variants within a sample should also be integrated into the pipeline. However, still some challenges remain regarding the link between the genomic and patient data and the detection of low-frequency variants. The integration of genomic data and associated patient data often represents an additional challenge and current data integration strategies and statistical analysis approaches need to be revised. Rapid and accurate interpretation of the data can be facilitated by interactive platforms and flexible bioinformatic workflows. Moreover, as shown in Chapter 6, it is important to provide quality criteria before analysis can be implemented in routine laboratories. By standardising these criteria over time and over countries, it will be possible to compare samples coming from different laboratories and over time. Additionally, this dissertation has been focussed on short-read sequencing which comes with several limitations. Some of these limitations, such as the short reads, and amplification and sequencing biases can be resolved by using long-read sequencing which is discussed in 9.3.1.

By integrating patient and genomic data, a powerful synergy can be provided for public health. This integration can contribute to describe nearly every aspect of transmission dynamics and could possibly also improve treatment for entire patient populations. For example, if a large amount of samples is sequenced, a particular mutation that influences the severity of the virus infection in patients that suffer from obesity could potentially be found. When these mutations are detected in an infected obese patient, he can be immediately treated with antivirals as a precaution to prevent severe infection. However, the resulting metadata often lacks consistency in reporting or is often incomplete as presented in Chapter 5. Chapter 5 highlights the need to construct or enhance international databases to include both genomic and patient data that are easily accessible to the public health authorities taking into account privacy considerations to protect sensitive patient data. This aspect will be discussed more elaborately in Chapter 9.3.2.

Besides the technical challenges, Chapters 4 to 6 underlined the importance of including a sufficient number of samples in addition to a representative and unbiased collection. Non-representative sampling strategies may lead to selection bias. Additionally, the size of the dataset can lead to statistical models that are underpowered which leads to unreliable results. Causal Directed Acyclic Graphs (DAGs) can be used to recognize the problem of selection bias [974] because they provide a transparent and simple way to explicitly state the qualitative underlying assumptions about the data-generating process [975]. Causal DAGs enable scientists with less advanced mathematical training to understand and recognize how and when selection bias may hamper causal inference [974].

9.2. WGS: Added value and challenges offered by WGS for the SARS-CoV-2 surveillance in wastewater samples

Of course, not only influenza surveillance benefits from using NGS in routine surveillance, but NGS can also be of added value to other surveillance systems, including the SARS-CoV-2 surveillance. The current SARS-CoV-2 pandemic has proven that NGS is indispensable from the pathogen discovery over variant characterisation to novel vaccine development. Due to the SARS-CoV-2 pandemic and the specific role of Sciensano in pandemic surveillance, a strategy based was developed on the use of WGS data for a better SARS-CoV-2 surveillance with a special focus on wastewater surveillance. Indeed, wastewater surveillance is a complementary approach to surveillance based on clinical samples and provides an unbiased method that is not limited by asymptomatic cases. Moreover, with limited resources it is possible to evaluate the spread of the infection in different areas and trace the circulating variants in a community. In Belgium, there are 42 wastewater treatment plants that are sampled twice per week. Three laboratories, including Sciensano, University of Antwerp and e-Biom, are responsible for the detection of SARS-CoV-2 in these samples and report the results weekly. During the SARS-CoV-2 pandemic, we explored RT-ddPCR methods to detect SARS-CoV-2 in wastewater samples while also exploring the opportunities of performing NGS on wastewater samples. In the context of the pandemic, protocols need to be implemented quickly, therefore, RT-ddPCR assays were used because they can be easily set up and can produce results quickly.

Therefore, in this thesis an approach using WGS data of SARS-CoV-2 is proposed for the evaluation of the primers and probes designed for the RT-ddPCR methods, which is now even more critical because of the emergence of SARS-CoV-2 variants. Furthermore, this thesis takes a first step in providing a strategy and quality criteria for sequencing SARS-CoV-2, including its variants at lower frequencies, in wastewater samples.

9.2.1. What this thesis adds

Although COVID-19 vaccines are available, new emerging variants may have increased infectivity, transmissibility, and immune evasion properties which could threaten global health again. These new variants can possibly pose problems for the current detection methods because of mutations. This thesis shows a way to develop new methods and evaluate existing methods to detect SARS-CoV-2. As mentioned in Chapter 5, international WGS databases, such as GISAID, can be a valuable asset to routine surveillance. Not only is it possible to use this database to investigate associations between mutations and patient data at an international level as demonstrated in Chapter 5, it can also evaluate and improve the design of RT-ddPCR targets to detect SARS-CoV-2 in patient and wastewater samples. There is a rapid worldwide increase of SARS-CoV-2 variants. Therefore, PCR-based methods need to be regularly evaluated because a possible false negative result may occur due to polymorphisms or point mutations related to the virus evolution which could impact the accuracy of the diagnostics tests. In Chapter 7, a methodology is provided to systematically evaluate and monitor PCR-based methods targeting rapidly evolving viruses such as SARS-CoV-2. First, an *in silico* evaluation using WGS sequences from around the world was performed to assess the inclusivity of the primers and probes. This ensured the inclusion of almost all circulating variants worldwide, which is not feasible using real samples. This *in silico* evaluation should be performed continuously using each time the newest available WGS data. Subsequently, the assay should be evaluated for experimental testing using a minimal set-up. However, RT-qPCR and RT-ddPCR methods are not the ideal methods for identifying SARS-CoV-2 variants because they consist of a combination of mutations while RT-qPCR and RT-ddPCR methods are often limited to a small part of the genome.

Therefore, the possibility of deep sequencing with NGS provides here an opportunity to detect multiple variants in a wastewater sample. Wastewater samples contain a collection of multiple strains, where the most abundant strain in the sample corresponds to the most prevalent strain circulating in a community and subpopulations correspond to less prevalent strains. Wastewater surveillance could provide a complementary alternative to individual testing for the epidemiological surveillance. However, as the virus concentrations are often low, PCR amplification is needed. Nevertheless, the PCR and NGS amplification steps contribute to experimental errors. Currently, quality criteria are mostly lacking regarding sequencing wastewater samples and analysing low-frequency variants. This thesis is a first step towards exploring SARS-CoV-2-targeted nucleotide sequencing of wastewater as an epidemiological surveillance method to estimate the prevalence, the genetic diversity and geographical distribution of SARS-CoV-2. WGS data derived from real diagnostic samples was *in silico* modified to construct mixed samples sequenced with Illumina and subjected to WGS using tiling amplicon-based targeted metagenomics approach. As expected, lower sensitivity and higher variation were observed at lower allelic frequencies and coverages. Chapter 8 provides specific recommendations for minimum sequencing coverages to detect clade-defining mutations at certain allelic frequencies. These recommendations can be a first step in establishing guidelines for the detection of low-frequency variants in both clinical and wastewater samples. Of course, further investigation needs to consider real data from actual diagnostic and wastewater surveillance.

9.2.2. Way forward

As mentioned in 9.1.2, there are several technical challenges that need to be tackled such as the computational resources and the development of easy-to-use bioinformatic tools before routine implementation of NGS for virus surveillance can be implemented. Additionally, it is also important that the obtained sequences are of high quality and shared with the scientific community, which will be described in more detail in 9.3.2. Sharing genomic data allows the characterisation and evaluation of new and already designed PCR assays that use primers and probes that target the SARS-CoV-2 genome and its variants. Additionally, by sharing genomic data the emergence of possible new variants can be tracked across the world.

However, even by integrating the genomic data for optimal PCR development, a qPCR strategy for the surveillance of wastewater has its limitations. Indeed, the collected genomic information relies on the WGS data from the surveillance of respiratory samples after which PCR assays are designed to collect data from clinical samples. The design of a qPCR assay can be done only when a certain number of WGS of a new variant has been reached. Then the newly designed assay targeting the new variant should be tested and validated [976], which takes time and manpower. Such a drawback hampers the early detection of new variants in wastewater. In the future, it is therefore important to further develop the sequencing strategy to sequence all variants in wastewater. Our study performed based on *in silico* data should be applied on wastewater samples in order to establish quality criteria for sequencing wastewater samples and improve the variant surveillance, especially the earlier detection of a new emerging variant.

Wastewater can be an interesting alternative to follow the emergence of the SARS-CoV-2 variants and their prevalence in a community. However, wastewater samples can also be interesting to include in the surveillance system of other viruses or bacteria that can be found in wastewater. Already numerous publications show the possibly long list of microbial targets that have already been found in wastewater in the past [414, 944, 977–998]. It would thus be very time-consuming and inefficient to develop and validate a qPCR, ddPCR or NGS method targeting each target individually with endless possibilities. Therefore, an untargeted approach, metagenomics, should be considered to monitor the microbial diversity within a wastewater sample (further discussed in more detail in 9.3.3).

9.3. Future Perspectives

9.3.1. Long-read sequencing as the future of high-throughput sequencing for the genomic characterisation of a virus

The global emergence of viruses has caused and will continue to cause a considerable impact on human and animal health and welfare. Both local outbreaks and pandemics have a major economic and social impact. Therefore, rapid identification and characterisation of the pathogen causing the infectious disease outbreak is crucial to implement measures that limit further spread and overall impact. The work presented in this dissertation focused on how virus surveillance can be improved by using NGS, more specifically within Sciensano, which is representative of the needs of scientific institutes of Public health. At the time of this thesis, the Illumina MiSeq was implemented transversally within Sciensano for research and routine surveillance. Therefore, this dissertation mainly focuses on sequence data obtained by Illumina technology. However, tens to hundreds of samples are needed to make Illumina MiSeq run economically viable to use WGS for routine analysis of influenza [999]. In the context of the surveillance, the number of samples is not limiting, however, if only a few samples need to be analysed in case of an emergency, crisis or for research, the number of samples will lead to unacceptable delays or more expensive sequencing costs. Furthermore, when samples need to be sequenced due to an outbreak, the data needs to be available as soon as possible. Second-generation sequencers do not allow real-time generation and analysis in contrast to ONT technology, which provides rapid *in situ* amplicon-based or metagenomic sequencing analysis [700]. However, in contrast to the Illumina sequencing technology which is stable and standardised, improvements of the nanopore technology are crucial to further increase the reproducibility, data throughput and the shelf life of the reagents, reduce error rates, and allow real-time analysis [999]. Moreover, at the moment nanopore sequencing is a new and fast growing technology with continuous upgrades and improvements. Therefore, challenges such as consumable changes and software updates need to be continuously tackled, which is not feasible in a routine setting. Most tools for ONT technology were developed by research groups and are often not stable, not maintained, not user-friendly and difficult to implement without bioinformatics expertise. Additionally, nanopore signal data are very large with computationally expensive base-calling and downstream analysis steps leading to computational bottlenecks [1000]. The increasing demand for accurate and fast analysis of clinical and environmental samples promotes the advancement of nanopore technology. There have already been several publications where ONT technology has been used for the detection of influenza [850, 1001, 1002] and SARS-CoV-2 [959, 1003, 1004]. However, in all these publications the relatively high error rates of the ONT sequencing technology remain a challenge that needs to be tackled before long-read sequencing can be implemented in routine surveillance. The reduction of the error rate is especially crucial for downstream analysis because it can have a significant impact on the performance of assays based on variant calling. Besides real-time generation and analysis, ONT also has the advantage of producing long reads that are able to define more accurately genetic haplotype compositions in viral quasispecies. The reconstruction of haplotypes using short-read sequencing is limited by conserved regions longer than the read length. Consequently, it is possible that there is more than one way to connect these relatively short reads. There are some algorithms that use the linkage information provided by paired-end reads [1005] or use the relative frequencies [670, 1006, 1007] to resolve this. However, these strategies are subjected to severe limitations, because the linkage information from the paired-end reads relies on the location of at least one of the pairs in a heterogeneous region. Moreover, amplification and sequencing biases can lead to deviations from the true underlying frequencies of the viral strains. Additionally, long reads can improve *de novo* assembly because it can span repetitive regions. Therefore, long-read sequencing technologies may offer opportunities as they offer read lengths that are comparable to the size of many RNA viral genomes. This haplotype reconstruction using long-read sequencing offers opportunities and improvement in the strategies developed in Chapter 6 and 8 especially in order to detect the quasispecies in clinical samples and variants in wastewater samples.

9.3.2. Quality databases in a One Health context

The need for genomic databases including genomic and patient data was already briefly discussed in Chapter 5. Many institutions and countries will often be limited in the number of samples that can be sequenced due to the sometimes low concentration of the virus and the availability of the samples but also because of a lack of resources to sequence every sample. Consequently, as demonstrated in Chapter 5, due to the limited number of samples, there will be a reduction of the statistical power and the sampling bias will be more pronounced. The epidemiological and clinical metadata can have a strong influence on the interpretation of the genomic data. Successful integration can be hindered by unstructured or incomplete metadata. For example in Chapter 5, the vaccination status of 28% patients was missing of the included ILI and SARI patients. This could affect the validity of the results and its confounding factors. For example, a recent review supports the hypothesis that influenza vaccination may attenuate the course of disease among individuals with breakthrough influenza virus infection [1008]. Incomplete data and increasing the quality of the data can be improved by establishing a system that links patient-level data from different independent healthcare registers within a secured environment. Additionally, as demonstrated in Chapter 4, viruses such as influenza and SARS-CoV-2 are often not limited to country borders. Consequently, virus strains are frequently reintroduced into the population leading to the co-circulation of two or more viral strains in the population. Therefore, a database or linked databases that includes patient and genomic data across borders could improve the understanding of the virus evolution. Sharing data enables physicians to quickly learn how to detect the symptoms of the disease, track the disease spread and give the hospitals the ability to share the best antiviral treatments. Moreover, the information about the patient's immune responses and the virus can also be used by researchers and pharmaceutical companies to develop new antiviral drugs targeting the host immune system.

Since the beginning of the SARS-CoV-2 pandemic, an almost exponential growth of the number of deposited sequences is observed within large publicly available databases. This pandemic is the first time that NGS technologies have been used to sequence a massive amount of viral sequences. However, there is a bias towards a limited number of countries with a high sequencing capacity [1009] which also translates into differences in the quality and structure of metadata. Several institutions provide resources and databases to deposit viral sequences. Some databases, such as NCBI's Genbank [1010] and Sequence Read Archive (SRA) [1011], already existed before the COVID-19 pandemic started including thousands of viral species that are also a threat to humanity including other potentially dangerous viruses such as Ebola, SARS, and Dengue. Additionally, other virus-specific databases, such as GISAID [752], produced a new data collection specifically dedicated to hosting SARS-CoV-2 sequences. GISAID includes the most complete collection of genetic sequence data, and related clinical and epidemiological data of influenza viruses, but is now also the predominant data source for SARS-CoV-2. However, the availability of the clinical and epidemiological data is often limited, if already available, to the specimen source, collection date, patient sex, patient age, and country and it is often not standardised. Furthermore, ideally these databases are easily accessible to upload and download data, to stimulate scientists to share their data and also compare their obtained results at a local level with data from across the globe. However, the submission guidelines should be sufficiently strict to ensure that fewer contaminated genomes are uploaded to the repositories. At the European level, the HERA Incubator project was launched in February 2021 to enhance the response to the COVID-19 pandemic and to improve preparedness for future epidemics and pandemics [1012]. A national infrastructure for data exchange will be implemented by the HERA-BE-Incubator project. This project aims to enhance the infrastructures for data exchange and genomic-epidemiological analyses and the national infrastructure for WGS-analyses. However, there are still some challenges, because most health data contains personal and sensitive details about the patients. Data should be shared in a secure way, but it should also be standardised across borders to be able to rapidly collect and widely share the data. Moreover, data misuse and uncertainties about permission and data ownership should be addressed to convince researchers to share their data.

The importance of the One Health concept has become more apparent over the last few years. One Health recognizes that the health of humans, wild and domestic animals, plants and the wider environment are closely linked to each other. It can help contribute to global health security by addressing all aspects of disease control including prevention, detection, preparedness, response and management. The rise in antiviral resistance and the emergence of infectious diseases can benefit from a One Health approach where authorities involved in various disciplines interact with one another. Communication and collaboration would be promoted among the human-animal-eco sectors by creating a common health data space through the One Health dimension, which should cover international, cross-sectoral and multi-stakeholder collaboration which could benefit the entire world and could be very helpful to spot future potential pandemics earlier. These sectors should be harmonised to complement and build upon one another to define what gaps exist, to reduce the duplication of efforts and to reduce public health risk.

9.3.3. Metagenomics as a surveillance tool

The work that was done in the context of this dissertation focused on the analysis of data that was generated after virus enrichment of the sample using PCR technology targeting solely influenza or SARS-CoV-2. Besides NGS approaches based on enrichment methods like PCR amplification and probe hybridisation [1013, 1014], there are other common approaches that are more universal and open such as shotgun metagenomics [1015]. Each approach has its strength to better understand the virus genomes and achieves different purposes. Both the strength and weakness of PCR amplicon-based methods is that they generate mostly reads specific to the target pathogen at relatively high coverage which was previously illustrated and discussed in this dissertation. Currently, the PCR based enrichment method remains probably the most efficient approach when a single pathogen needs to be studied. However, due to multiple environmental and societal factors, there is an increased need for tools allowing more global surveillance in an open way in order to monitor pathogens including re-emerging ones or to discover novel ones. For this purpose, the PCR target approach is not appropriate. Hybrid capture-based target enrichment is able to sequence a viral genome or several viral genomes together without the need for prior culture or PCR amplification, which results in the introduction of less bias [1016]. Probes need to be designed for a panel of reference sequences and could help to better capture the diversity of the target virus genomes and analyse viral populations. Although hybrid sequencing allows the detection of multiple targets using thousands of probes [1017], there is still prior knowledge needed about the target, thus entirely new virus discovery is unlikely by using hybrid sequencing. In contrast to the enrichment methods, a metagenomics approach is carried out directly from the extraction of the DNA/RNA from a sample in connection with a diagnosis without prior culturing and without specific amplification or probe capture. Therefore, metagenomics has the potential of becoming a universal pathogen detection tool regardless of the type of microbe and can even be applied for novel organism discovery [1018, 1019]. Furthermore, it can also be an interesting tool for epidemiological surveillance and environmental surveillance, for example using wastewater samples. In recent years, metagenomics of complex samples has become increasingly common. It has been used, for example, to identify the novel coronavirus SARS-CoV-2 in a bronchoalveolar lavage fluid sample obtained from a suspected case using metagenomic RNA sequencing on 5 January 2020 [521]. Furthermore, metagenomics next-generation sequencing approaches have also already been used to explore the human virome in human stools [1020, 1021], cerebrospinal fluid [1022, 1023], blood [1024, 1025], respiratory tract samples [727, 1026–1029], and human tissues [1023, 1030]. It also allows studying the pathogen community present in environmental samples, like wastewater. The few existing publications have demonstrated that viral metagenomics of wastewater is a versatile tool to monitor, identify and discover the viral diversity among human and livestock populations [1031–1036]. As viruses are the most abundant biological entities on the planet, considerable genetic complexity and diversity within viral populations can be uncovered by using viral metagenomics. These new sequences will not share homology with sequences that are found in the reference database and by using viral metagenomics the number of reported viruses can be drastically increased [1037–1039].

To unlock the full potential of metagenomics, several technical improvements and designed strategies need to be elaborated in the future. Generally, the viral metagenomics process includes four major steps: (1) the collection of a clinical and/or environmental sample, (2) sample preparation, (3) library preparation and sequencing, and (4) sequence analysis by comparing the sequences to a reference database.

(1) For the first step, a well-thought out sampling strategy is needed. Novel diseases will most likely already circulate for some time in animals and humans before clinical cases are detected. Therefore, it is critical that the surveillance system is sensitive enough to detect emerging infectious agents by collecting samples at potential hot spots where there are enhanced chances of the emergence of novel human pathogenic viruses leading to potential outbreaks. These hot spots include places where animal habitats interact with humans [1040], but can also include samples coming from animal waste or the environment. The monitoring of animal metagenomes is also important because more than 60% of the reported novel viruses are of zoonotic origin [1041, 1042]. However, other potential sampling sites can also be urban metro-stations, long distance planes or urban sewage [414, 1043–1045]. To avoid generating massive amounts of data without meaningful surveillance information, the sampling strategies need to be guided by ecological and epidemiological knowledge. Additionally, standardised samples that represent both human and animal microbiome should be collected in a comparable way between countries and over time. Besides the detection of known and unknown microorganisms, it will also allow the detection of different variants of a specific virus such as SARS-CoV-2. To the best of our knowledge, such sampling strategy is not yet currently used in Belgium except in some research projects. So, an effort for the global harmonisation strategy of sampling in the context of the surveillance of virus discovery should be considered.

(2 & 3) The sample preparation step often includes procedures to remove as much of the cellular genomes as possible, including a combination of filtration, centrifugation, and nuclease treatment. Additionally, a retro-transcription step for RNA viruses is necessary before starting with the library preparation and in case of limited viral genetic material, a random amplification approach is often used. There are three widely employed random amplification protocols: multiple displacement amplification (MDA) [1046, 1047], linker amplification shotgun libraries (LASL) [1048] and SISPA [1049, 1050]. However, all three protocols are associated with a bias that results in uneven coverage across sequenced genomes or alters the relative abundance of the viruses. Compared to the targeted enrichment protocols that were used in this thesis, these methods will often also take more time and will have a higher cost. Several papers described random amplification in a research project on clinical samples such as nasopharyngeal swabs [1051, 1052]. Such research protocols should be tested for their feasibility when several samples are sequenced in the context of routine surveillance. This

step will also need to be evaluated and if necessary enhanced to be able to reach enough sensitivity in the context of the detection of minority variants of SARS-CoV-2 in wastewater. Wastewater samples often have a low virus concentration. Especially to have an overview of all circulating variants within a certain population, it is important that low abundance variants are detected so that the emergence of new variants can be traced. It is probably also an essential step in the context of the detection of emerging viral pathogens in wastewater. Already a few studies have tried this approach on this type of sample [1033, 1053, 1054], however in these studies bias towards certain viral genome regions were observed. Therefore, further optimization will be necessary.

(4) NGS technologies are mainly responsible for the metagenomics boom of the last few years. The species present in a sample based on the NGS primary reads can be identified by comparing against database(s), for example using kmer-based methods. These tools have to reconstruct the different genomes present, which is much more difficult or even impossible for a complex sample using short-read second-generation sequencing technology. Moreover, it is also difficult to link antibiotic resistance or virulence genes to specific bacteria or viruses present in the sample. Third-generation sequencing, on the other hand, allows a reading of several Mb which facilitates these types of analyses. Moreover, third generation sequencing methods offer significant advantages in terms of speed because they read the native DNA strand and makes real-time data analysis feasible [1055, 1056]. The longer read length makes it easier to reconstruct the genomes. However, all disadvantages related to long-read sequencing remain for the moment (discussed in 9.3.1). As mentioned, Nanopore sequencing is evolving and improving all the time and only recently Oxford Nanopore published a new protocol for rapid metagenomic characterisation of RNA and DNA viruses that has been applied on the monkeypox virus [1057]. However, this is only the start, the popularity of metagenomics received a boost due to the emergence of SARS-CoV-2 and the monkeypox virus. However, more research is needed to achieve large-scale application of nanopore sequencing. Moreover, these are often very complex samples that will generate a large amount of data resulting in storage and computational bottlenecks. Additionally, because of the complexity of these samples there are often no user-friendly tools available so there is also a need for bioinformatics expertise [1058]. At the moment, no generally validated bioinformatics pipeline exists that can perform a sensitive, rapid and specific analysis of the metagenomics data on a benchtop computer. Also, physicians will need to be trained and guided to deal with the obtained breadth of data if metagenomics is to be implemented in a clinical setting. Furthermore, the wide implementation of metagenomics approaches is limited due to the lack of standardisation, the duration and cost of sequencing which is especially important when analysing samples for diagnostic or surveillance purposes. Some studies already tried to

determine the specificity and sensitivity of metagenomic assays compared to conventional methods [1022, 1059]. The specificity and sensitivity of metagenomic assays depend on the optimization of the protocol like other methods such as PCR and sequencing. Due to the often complicated metagenomic assays that include a wide range of matrices and a multi-step custom laboratory and analytical workflow, there are many sources that can cause variability. By optimising, standardising and validating the protocols, intra-laboratory reproducibility should be achieved [1022, 1059]. Efforts towards standardising protocols are already gaining momentum [1059–1061] and this should be further stimulated by a reduction in sequencing cost, multiplexing, automation and benchmark bioinformatic tools [1062]. Also, adaptive sequencing can be done by using nanopore sequencing that can eject DNA templates so that it provides the ability to select sequence DNA in real time. By further developing adaptive sequencing, resources requirements and turnaround time should be reduced. Some research consortia such as METASTAVA (2018-2019; H2020-SFS-2017-1) and METAMORPHOSE (2021-2025; SRP-2020), to which Sciensano contributes as a partner, have already put some efforts to standardise metagenomics approaches. These kinds of efforts should be pursued further to develop protocols for a large range of samples and targets and strive for standardisation of these protocols.

In conclusion metagenomics has the potential to mature into a reliable and affordable technology for pathogen detection and surveillance in the future, especially since the onset of second- and third-generation sequencing technology [1063]. Nanopore sequencing as a metagenomic diagnostic tool has already been used by veterinary practitioners for example to detect veterinary diseases [1064–1066]. Nevertheless, several challenges remain to be tackled in order to consistently produce accurate results, especially in the case of viruses that often only represent a small fraction of the sequenced sample. Presently, pandemic surveillance plans usually target a specific virus and depend on the targeted detection of specific viral threats. Consequently, novel and unanticipated viruses are often not detected until it is too late. The current SARS-CoV-2 pandemic has shown that unbiased identification of potential pathogens and data exchange is necessary. The early detection of the pathogen is crucial to initiate infection control measures. By using metagenomics, all nucleic acids in a sample are sequenced without any assumptions. Therefore, as soon as the approach will become more common, metagenomics will probably become one of the most interesting surveillance tools that allows simultaneous characterisation of complete genome sequences, resistance and epidemiological markers and virulence factors. We believe that metagenomics will be an unavoidable surveillance tool that can unify microbial surveillance and transform public health efforts to proactively screen for threats in a One Health context in order to be better prepared for the next epidemic and pandemic.

BIBLIOGRAPHY

1. **GBD 2016 DALYs and HALE Collaborators**. Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 2017;390:1260–1344.
2. **World Health Organization**. Global health estimates: Leading causes of death. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death (2019, accessed 26 January 2022).
3. **Kondrich J, Rosenthal M**. Influenza in children. *Current opinion in pediatrics* 2017;29:297–302.
4. **Gordon A, Reingold A**. The Burden of Influenza: a Complex Problem. *Current epidemiology reports* 2018;5:1–9.
5. **Centers for Disease Control and Prevention**. Key Facts About Influenza (Flu). https://www.cdc.gov/flu/about/keyfacts.htm (2022, accessed 29 September 2022).
6. **de Courville C, Cadarette SM, Wissinger E, Alvarez FP**. The economic burden of influenza among adults aged 18 to 64: A systematic literature review. *Influenza Other Respir Viruses* 2022;16:376–385.
7. **Taubenberger JK, Morens DM**. 1918 Influenza: the Mother of All Pandemics. *Emerging Infectious Diseases* 2006;12:15–22.
8. **World Health Organization**. Influenza (Seasonal). https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal) (2022, accessed 29 September 2022).
9. **Moghadami M**. A Narrative Review of Influenza: A Seasonal and Pandemic Disease. *Iranian journal of medical sciences* 2017;42:2–13.
10. **Lipsitch M, Viboud C**. Influenza seasonality: Lifting the fog. *PNAS* 2009;106:3645–3646.
11. **Deyle ER, Maher MC, Hernandez RD, Basu S, Sugihara G**. Global environmental drivers of influenza. *Proc Natl Acad Sci U S A* 2016;113:13081–13086.
12. **Brankston G, Gitterman L, Hirji Z, Lemieux C, Gardam M**. Transmission of influenza A in human beings. *The Lancet Infectious diseases* 2007;7:257–65.
13. **Lagacé-Wiens PRS, Rubinstein E, Gumel A**. Influenza epidemiology--past, present, and future. *Crit Care Med* 2010;38:e1-9.
14. **Killingley B, Nguyen-Van-Tam J**. Routes of influenza transmission. *Influenza and Other Respiratory Viruses* 2013;7:42–51.
15. **Yan J, Grantham M, Pantelic J, Bueno de Mesquita PJ, Albert B, *et al.*** Infectious virus in exhaled breath of symptomatic seasonal influenza cases from a college community. *Proceedings of the National Academy of Sciences* 2018;115:1081–1086.
16. **Cooper BS, Pitman RJ, Edmunds WJ, Gay NJ**. Delaying the international spread of pandemic influenza. *PLoS Med* 2006;3:e212.
17. **Grais RF, Ellis JH, Glass GE**. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *Eur J Epidemiol* 2003;18:1065–1072.
18. **Carrat F, Vergu E, Ferguson NM, Lemaitre M, Cauchemez S, *et al.*** Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *American journal of epidemiology* 2008;167:775–85.
19. **Paules C, Subbarao K**. Influenza. *Lancet (London, England)* 2017;390:697–708.
20. **Taubenberger JK, Morens DM**. The pathology of influenza virus infections. *Annual review of pathology* 2008;3:499–522.
21. **Centers for Disease Control and Prevention**. Are You at High Risk for Serious Illness from Flu?
22. **Centers for Disease Control and Prevention**. Flu Symptoms & Complications. *Seasonal Influenza*.
23. **Su S, Fu X, Li G, Kerlin F, Veit M**. Novel Influenza D virus: Epidemiology, pathology, evolution and biological characteristics. *Virulence* 2017;8:1580–1591.
24. **Hause BM, Collin EA, Liu R, Huang B, Sheng Z, *et al.*** Characterization of a Novel Influenza Virus in Cattle and Swine: Proposal for a New Genus in the Orthomyxoviridae Family. *mBio*;5. Epub ahead of print May 2014. DOI: 10.1128/mBio.00031-14.
25. **Chiapponi C, Faccini S, De Mattia A, Baioni L, Barbieri I, *et al.*** Detection of Influenza D Virus among Swine and Cattle, Italy. *Emerging Infectious Diseases* 2016;22:352–354.
26. **Petrova VN, Russell CA**. The evolution of seasonal influenza viruses. *Nat Rev Microbiol* 2018;16:47–60.

27. **Davlin SL, Blanton L, Kniss K, Mustaquim D, Smith S, *et al.*** Influenza Activity — United States, 2015–16 Season and Composition of the 2016–17 Influenza Vaccine. *MMWR Morbidity and Mortality Weekly Report* 2016;65:567–575.
28. **Capua I, Alexander DJ (eds)**. *Avian Influenza and Newcastle Disease*. Milano: Springer Milan; 2009. Epub ahead of print 2009. DOI: 10.1007/978-88-470-0826-7.
29. **Wright PF, Neumann G, Kawaoka Y**. *Fields' Virology*. 2007.
30. **Air GM**. Sequence relationships among the hemagglutinin genes of 12 subtypes of influenza A virus. *Proc Natl Acad Sci U S A* 1981;78:7639–7643.
31. **Chen R, Holmes EC**. Avian influenza virus exhibits rapid evolutionary dynamics. *Mol Biol Evol* 2006;23:2336–2341.
32. **Sutton TC, Chakraborty S, Mallajosyula VVA, Lamirande EW, Ganti K, *et al.*** Protective efficacy of influenza group 2 hemagglutinin stem-fragment immunogen vaccines. *npj Vaccines* 2017;2:1–11.
33. **Fouchier RAM, Munster V, Wallensten A, Bestebroer TM, Herfst S, *et al.*** Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J Virol* 2005;79:2814–2822.
34. **Alexander DJ, Brown IH**. History of highly pathogenic avian influenza. *Revue scientifique et technique (International Office of Epizootics)* 2009;28:19–38.
35. **Cox NJ, Subbarao K**. Influenza. *The Lancet* 1999;354:1277–1282.
36. **White SK, Ma W, McDaniel CJ, Gray GC, Lednicky JA**. Serologic evidence of exposure to influenza D virus among persons with occupational contact with cattle. *J Clin Virol* 2016;81:31–33.
37. **Lamb RA, Choppin PW**. The Gene Structure and Replication of Influenza Virus. *Annual Review of Biochemistry* 1983;52:467–506.
38. **Wu YY, Wu YY, Tefsen B, Shi Y, Gao GF**. Bat-derived influenza-like viruses H17N10 and H18N11. *Trends in microbiology* 2014;22:183–91.
39. **Muramoto Y, Noda T, Kawakami E, Akkina R, Kawaoka Y**. Identification of novel influenza A virus proteins translated from PA mRNA. *Journal of virology* 2013;87:2455–62.
40. **Vasin A V, Temkina OA, Egorov V V, Klotchenko SA, Plotnikova MA, *et al.*** Molecular mechanisms enhancing the proteome of influenza A viruses: an overview of recently discovered proteins. *Virus research* 2014;185:53–63.
41. **Harris A, Cardone G, Winkler DC, Heymann JB, Brecher M, *et al.*** Influenza virus pleiomorphy characterized by cryoelectron tomography. *Proc Natl Acad Sci U S A* 2006;103:19123–19127.
42. **Shaw M, Palese P**. Orthomyxoviridae: The viruses and their replication. *Fields Virology* 2007;1647–1689.
43. **Bertram S, Glowacka I, Steffen I, Kühl A, Pöhlmann S**. Novel insights into proteolytic cleavage of influenza virus hemagglutinin. *Rev Med Virol* 2010;20:298–310.
44. **Skehel JJ, Wiley DC**. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annual review of biochemistry* 2000;69:531–69.
45. **Steinhauer D a**. Role of hemagglutinin cleavage for the pathogenicity of influenza virus. *Virology* 1999;258:1–20.
46. **Perdue ML, Suarez DL**. Structural features of the avian influenza virus hemagglutinin that influence virulence. *Veterinary microbiology* 2000;74:77–86.
47. **Gerhard W, Yewdell J, Frankel ME, Webster R**. Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies. *Nature* 1981;290:713–717.
48. **Webster RG, Laver WG**. Determination of the number of nonoverlapping antigenic areas on Hong Kong (H3N2) influenza virus hemagglutinin with monoclonal antibodies and the selection of variants with potential epidemiological significance. *Virology* 1980;104:139–148.
49. **Wiley DC, Wilson IA, Skehel JJ**. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* 1981;289:373–378.
50. **Luo G, Chung J, Palese P**. Alterations of the stalk of the influenza virus neuraminidase: deletions and insertions. *Virus research* 1993;29:141–53.
51. **Palese P, Tobita K, Ueda M, Compans RW**. Characterization of temperature sensitive influenza virus mutants defective in neuraminidase. *Virology* 1974;61:397–410.
52. **Nayak DP, Balogun RA, Yamada H, Zhou ZH, Barman S**. Influenza virus morphogenesis and budding. *Virus Research* 2009;143:147–161.
53. **Watanabe T, Watanabe S, Ito H, Kida H, Kawaoka Y**. Influenza A virus can undergo multiple cycles of replication without M2 ion channel activity. *Journal of virology* 2001;75:5656–62.
54. **Wise HM, Hutchinson EC, Jagger BW, Stuart AD, Kang ZH, *et al.*** Identification of a novel splice variant form of the influenza A virus M2 ion channel with an antigenically distinct ectodomain. *PLoS pathogens* 2012;8:e1002998.

55. **Reid AH, Fanning TG, Janczewski TA, McCall S, Taubenberger JK**. Characterization of the 1918 “Spanish” Influenza Virus Matrix Gene Segment. *Journal of Virology* 2002;76:10717–10723.
56. **Ma C, Wang J**. Functional studies reveal the similarities and differences between AM2 and BM2 proton channels from influenza viruses. *Biochim Biophys Acta* 2018;1860:272–280.
57. **Hsu MT, Parvin JD, Gupta S, Krystal M, Palese P**. Genomic RNAs of influenza viruses are held in a circular conformation in virions and in infected cells by a terminal panhandle. *Proceedings of the National Academy of Sciences of the United States of America* 1987;84:8140–4.
58. **Fodor E, Pritlove DC, Brownlee GG**. The influenza virus panhandle is involved in the initiation of transcription. *Journal of virology* 1994;68:4092–6.
59. **Yamayoshi S, Fukuyama S, Yamada S, Zhao D, Murakami S, *et al.*** Amino acids substitutions in the PB2 protein of H7N9 influenza A viruses are important for virulence in mammalian hosts. *Sci Rep* 2015;5:8039.
60. **Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, *et al.*** A novel influenza A virus mitochondrial protein that induces cell death. *Nature medicine* 2001;7:1306–12.
61. **Ozawa M, Basnet S, Burley LM, Neumann G, Hatta M, *et al.*** Impact of amino acid mutations in PB2, PB1-F2, and NS1 on the replication and pathogenicity of pandemic (H1N1) 2009 influenza viruses. *Journal of virology* 2011;85:4596–601.
62. **James J, Howard W, Iqbal M, Nair VK, Barclay WS, *et al.*** Influenza A virus PB1-F2 protein prolongs viral shedding in chickens lengthening the transmission window. *The Journal of general virology* 2016;97:2516–2527.
63. **Wise HM, Foeglein A, Sun J, Dalton RM, Patel S, *et al.*** A Complicated Message: Identification of a Novel PB1-Related Protein Translated from Influenza A Virus Segment 2 mRNA. *Journal of Virology* 2009;83:8021–8031.
64. **Jagger BW, Wise HM, Kash JC, Walters K-A, Wills NM, *et al.*** An overlapping protein-coding region in influenza A virus segment 3 modulates the host response. *Science (New York, NY)* 2012;337:199–204.
65. **García-Sastre A**. Inhibition of interferon-mediated antiviral responses by influenza A viruses and other negative-strand RNA viruses. *Virology* 2001;279:375–84.
66. **Steinhauer DA, Skehel JJ**. Genetics of influenza viruses. *Annual review of genetics* 2002;36:305–32.
67. **Hale BG, Randall RE, Ortín J, Jackson D, Ortin J, *et al.*** The multifunctional NS1 protein of influenza A viruses. *The Journal of general virology* 2008;89:2359–2376.
68. **Jackson D, Hossain MdJ, Hickman D, Perez DR, Lamb RA**. A new influenza virus virulence determinant: The NS1 protein four C-terminal residues modulate pathogenicity. *Proceedings of the National Academy of Sciences* 2008;105:4381–4386.
69. **Boulo S, Akarsu H, Ruigrok RWH, Baudin F**. Nuclear traffic of influenza virus proteins and ribonucleoprotein complexes. *Virus research* 2007;124:12–21.
70. **Selman M, Dankar SK, Forbes NE, Jia J-J, Brown EG**. Adaptive mutation in influenza A virus non-structural gene is linked to host switching and induces a novel protein by alternative splicing. *Emerging Microbes & Infections* 2012;1:1–10.
71. **Connor RJ, Kawaoka Y, Webster RG, Paulson JC**. Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. *Virology* 1994;205:17–23.
72. **Couceiro JN, Paulson JC, Baum LG**. Influenza virus strains selectively recognize sialyloligosaccharides on human respiratory epithelium; the role of the host cell in selection of hemagglutinin receptor specificity. *Virus Res* 1993;29:155–165.
73. **Long JS, Mistry B, Haslam SM, Barclay WS**. Host and viral determinants of influenza A virus species specificity. *Nat Rev Microbiol* 2019;17:67–81.
74. **Shinya K, Ebina M, Yamada S, Ono M, Kasai N, *et al.*** Avian flu: influenza virus receptors in the human airway. *Nature* 2006;440:435–436.
75. **Peteranderl C, Herold S, Schmoldt C**. Human Influenza Virus Infections. *Semin Respir Crit Care Med* 2016;37:487–500.
76. **Ito T, Couceiro JN, Kelm S, Baum LG, Krauss S, *et al.*** Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. *J Virol* 1998;72:7367–7373.
77. **Edinger TO, Pohl MO, Stertz S**. Entry of influenza A virus: host factors and antiviral targets. *J Gen Virol* 2014;95:263–277.
78. **Hamilton BS, Whittaker GR, Daniel S**. Influenza virus-mediated membrane fusion: determinants of hemagglutinin fusogenic activity and experimental approaches for assessing virus fusion. *Viruses* 2012;4:1144–1168.
79. **Boulay F, Doms RW, Wilson I, Helenius A**. The influenza hemagglutinin precursor as an acid-sensitive probe of the biosynthetic pathway. *EMBO J* 1987;6:2643–2650.

80. **Chen J, Lee KH, Steinhauer DA, Stevens DJ, Skehel JJ, *et al.*** Structure of the hemagglutinin precursor cleavage site, a determinant of influenza pathogenicity and the origin of the labile conformation. *Cell* 1998;95:409–417.
81. **Zhirnov OP, Ikizler MR, Wright PF**. Cleavage of influenza a virus hemagglutinin in human respiratory epithelium is cell associated and sensitive to exogenous antiproteases. *J Virol* 2002;76:8682–8689.
82. **Bullough PA, Hughson FM, Skehel JJ, Wiley DC**. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature* 1994;371:37–43.
83. **Pinto LH, Holsinger LJ, Lamb RA**. Influenza virus M2 protein has ion channel activity. *Cell* 1992;69:517–528.
84. **Bui M, Whittaker G, Helenius A**. Effect of M1 protein and low pH on nuclear transport of influenza virus ribonucleoproteins. *J Virol* 1996;70:8391–8401.
85. **Hutchinson EC, Fodor E**. Transport of the influenza virus genome from nucleus to nucleus. *Viruses* 2013;5:2424–2446.
86. **Cros JF, García-Sastre A, Palese P**. An unconventional NLS is critical for the nuclear import of the influenza A virus nucleoprotein and ribonucleoprotein. *Traffic* 2005;6:205–213.
87. **Tiley LS, Hagen M, Matthews JT, Krystal M**. Sequence-specific binding of the influenza virus RNA polymerase to sequences located at the 5’ ends of the viral RNAs. *J Virol* 1994;68:5108–5116.
88. **Beaton AR, Krug RM**. Selected host cell capped RNA fragments prime influenza viral RNA transcription *in vivo*. *Nucleic Acids Res* 1981;9:4423–4436.
89. **Dias A, Bouvier D, Crépin T, McCarthy AA, Hart DJ, *et al.*** The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* 2009;458:914–918.
90. **Plotch SJ, Bouloy M, Ulmanen I, Krug RM**. A unique cap(m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription. *Cell* 1981;23:847–858.
91. **Yuan P, Bartlam M, Lou Z, Chen S, Zhou J, *et al.*** Crystal structure of an avian influenza polymerase PAN reveals an endonuclease active site. *Nature* 2009;458:909–913.
92. **Pritlove DC, Poon LLM, Devenish LJ, Leahy MB, Brownlee GG**. A Hairpin Loop at the 5′ End of Influenza A Virus Virion RNA Is Required for Synthesis of Poly(A)+ mRNA *In vitro*. *J Virol* 1999;73:2109–2114.
93. **Chen Z**. Influenza A virus NS1 protein targets poly(A)-binding protein II of the cellular 3’-end processing machinery. *The EMBO Journal* 1999;18:2273–2283.
94. **Lamb RA, Lai CJ, Choppin PW**. Sequences of mRNAs derived from genome RNA segment 7 of influenza virus: colinear and interrupted mRNAs code for overlapping proteins. *Proc Natl Acad Sci U S A* 1981;78:4170–4174.
95. **Lamb RA, Choppin PW**. Identification of a second protein (M2) encoded by RNA segment 7 of influenza virus. *Virology* 1981;112:729–737.
96. **Lamb RA, Lai CJ**. Sequence of interrupted and uninterrupted mRNAs and cloned DNA coding for the two overlapping nonstructural proteins of influenza virus. *Cell* 1980;21:475–485.
97. **Hershey JWB, Sonenberg N, Mathews MB**. Principles of translational control: an overview. *Cold Spring Harb Perspect Biol* 2012;4:a011528.
98. **Roux PP, Topisirovic I**. Regulation of mRNA translation by signaling pathways. *Cold Spring Harb Perspect Biol* 2012;4:a012252.
99. **Sonenberg N, Hinnebusch AG**. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell* 2009;136:731–745.
100. **Bui M, Wills EG, Helenius A, Whittaker GR**. Role of the influenza virus M1 protein in nuclear export of viral ribonucleoproteins. *J Virol* 2000;74:1781–1786.
101. **Huang X, Liu T, Muller J, Levandowski RA, Ye Z**. Effect of influenza virus matrix protein and viral RNA on ribonucleoprotein formation and nuclear export. *Virology* 2001;287:405–416.
102. **Neumann G, Hughes MT, Kawaoka Y**. Influenza A virus NS2 protein mediates vRNP nuclear export through NES-independent interaction with hCRM1. *EMBO J* 2000;19:6751–6758.
103. **Vreede FT, Brownlee GG**. Influenza virion-derived viral ribonucleoproteins synthesize both mRNA and cRNA *in vitro*. *J Virol* 2007;81:2196–2204.
104. **York A, Hengrung N, Vreede FT, Huiskonen JT, Fodor E**. Isolation and characterization of the positive-sense replicative intermediate of a negative-strand RNA virus. *Proc Natl Acad Sci U S A* 2013;110:E4238-4245.
105. **Deng T, Vreede FT, Brownlee GG**. Different de novo initiation strategies are used by influenza virus RNA polymerase on its cRNA and viral RNA promoters during viral RNA replication. *J Virol* 2006;80:2337–2348.
106. **Huang S, Chen J, Chen Q, Wang H, Yao Y, *et al.*** A second CRM1-dependent nuclear export signal in the influenza A virus NS2 protein contributes to the nuclear export of viral ribonucleoproteins. *J Virol* 2013;87:767–778.

107. **Chen BJ, Leser GP, Morita E, Lamb RA**. Influenza virus hemagglutinin and neuraminidase, but not the matrix protein, are required for assembly and budding of plasmid-derived virus-like particles. *J Virol* 2007;81:7111–7123.

108. **Leser GP, Lamb RA**. Influenza virus assembly and budding in raft-derived microdomains: a quantitative analysis of the surface distribution of HA, NA and M2 proteins. *Virology* 2005;342:215–227.

109. **Ali A, Avalos RT, Ponimaskin E, Nayak DP**. Influenza virus assembly: effect of influenza virus glycoproteins on the membrane association of M1 protein. *J Virol* 2000;74:8709–8719.

110. **Noton SL, Medcalf E, Fisher D, Mullin AE, Elton D, *et al.*** Identification of the domains of the influenza A virus M1 matrix protein required for NP binding, oligomerization and incorporation into virions. *J Gen Virol* 2007;88:2280–2290.

111. **Noda T, Sagara H, Yen A, Takada A, Kida H, *et al.*** Architecture of ribonucleoprotein complexes in influenza A virus particles. *Nature* 2006;439:490–492.

112. **Rossman JS, Jing X, Leser GP, Balannik V, Pinto LH, *et al.*** Influenza Virus M2 Ion Channel Protein Is Necessary for Filamentous Virion Formation. *J Virol* 2010;84:5078–5088.

113. **Rossman JS, Lamb RA**. Influenza virus assembly and budding. *Virology* 2011;411:229–236.

114. **Air GM, Laver WG**. The neuraminidase of influenza virus. *Proteins* 1989;6:341–356.

115. **Huang I-C, Li W, Sui J, Marasco W, Choe H, *et al.*** Influenza A Virus Neuraminidase Limits Viral Superinfection. *J Virol* 2008;82:4834–4843.

116. **Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R**. Viral Mutation Rates. *Journal of Virology* 2010;84:9733–9748.

117. **Carrat F, Flahault A**. Influenza vaccine: the challenge of antigenic drift. *Vaccine* 2007;25:6852–62.

118. **Sonoguchi T, Naito H, Hara M, Takeuchi Y, Fukumi H**. Cross-subtype protection in humans during sequential, overlapping, and/or concurrent epidemics caused by H3N2 and H1N1 influenza viruses. *The Journal of infectious diseases* 1985;151:81–8.

119. **Gill PW, Murphy AM**. Naturally acquired immunity to influenza type A: a further prospective study. *The Medical journal of Australia* 1977;2:761–5.

120. **de Jong JC, Rimmelzwaan GF, Fouchier RAM, Osterhaus ADME**. Influenza Virus: a Master of Metamorphosis. *Journal of Infection* 2000;40:218–228.

121. **Shao W, Li X, Goraya MU, Wang S, Chen J-LL**. Evolution of Influenza A Virus by Mutation and Re-Assortment. *International Journal of Molecular Sciences* 2017;18:1650.

122. **Morens DM, Fauci AS**. The 1918 influenza pandemic: insights for the 21st century. *The Journal of infectious diseases* 2007;195:1018–28.

123. **Viboud C, Simonsen L, Fuentes R, Flores J, Miller MA, *et al.*** Global Mortality Impact of the 1957-1959 Influenza Pandemic. *The Journal of infectious diseases* 2016;213:738–45.

124. **Viboud C, Grais RF, Lafont BAP, Miller MA, Simonsen L, *et al.*** Multinational impact of the 1968 Hong Kong influenza pandemic: evidence for a smoldering pandemic. *The Journal of infectious diseases* 2005;192:233–48.

125. **Shrestha SS, Swerdlow DL, Borse RH, Prabhu VS, Finelli L, *et al.*** Estimating the burden of 2009 pandemic influenza A (H1N1) in the United States (April 2009-April 2010). *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2011;52 Suppl 1:S75-82.

126. **Tscherne DDM, García-Sastre A**. Virulence determinants of pandemic influenza viruses. *The Journal of clinical …* 2011;121:6–13.

127. **Landolt GA, Olsen CW**. Up to new tricks - a review of cross-species transmission of influenza A viruses. *Animal health research reviews* 2007;8:1–21.

128. **Yen H-L, Webster RG**. Pandemic Influenza as a Current Threat. In: Compans RW, Orenstein WA (editors). *Vaccines for Pandemic Influenza*. Berlin, Heidelberg: Springer. pp. 3–24.

129. **de Jong JC, Claas EC, Osterhaus AD, Webster RG, Lim WL**. A pandemic warning? *Nature* 1997;389:554.

130. **World Health Organization**. *Avian Influenza Weekly Update Number 797*. https://www.who.int/docs/default-source/wpro---documents/emergency/surveillance/avian-influenza/ai-20210618-rev.pdf?sfvrsn=30d65594_147 (18 June 2021, accessed 7 June 2021).

131. **Cong Y, Wang G, Guan Z, Chang S, Zhang Q, *et al.*** Reassortant between Human-Like H3N2 and Avian H5 Subtype Influenza A Viruses in Pigs: A Potential Public Health Risk. *PloS one* 2010;5:e12591.

132. **Nidom CA, Takano R, Yamada S, Sakai-Tagawa Y, Daulay S, *et al.*** Influenza A (H5N1) viruses from pigs, Indonesia. *Emerg Infect Dis* 2010;16:1515–1523.

133. **Pasma T, Joseph T**. Pandemic (H1N1) 2009 Infection in Swine Herds, Manitoba, Canada. *Emerg Infect Dis* 2010;16:706–708.

134. **Reperant LA, Rimmelzwaan GF, Kuiken T**. Avian influenza viruses in mammals. *Rev Sci Tech* 2009;28:137–159.

135. **Pasick J, Handel K, Robinson J, Copps J, Ridd D, *et al.*** Intersegmental recombination between the haemagglutinin and matrix genes was responsible for the emergence of a highly pathogenic H7N3 avian influenza virus in British Columbia. *The Journal of general virology* 2005;86:727–731.
136. **Suarez DL, Senne DA, Banks J, Brown IH, Essen SC, *et al.*** Recombination Resulting in Virulence Shift in Avian Influenza Outbreak, Chile. *Emerging Infectious Diseases* 2004;10:693–699.
137. **Lauring AS, Andino R**. Quasispecies Theory and the Behavior of RNA Viruses. *PLoS Pathogens* 2010;6:e1001005.
138. **Peck KM, Lauring AS**. Complexities of Viral Mutation Rates. *J Virol* 2018;92:e01031-17.
139. **Tsang TK, Cowling BJ, Fang VJ, Chan K-H, Ip DKM, *et al.*** Influenza A Virus Shedding and Infectivity in Households. *J Infect Dis* 2015;212:1420–1428.
140. **Eigen M**. Viral quasispecies. *Scientific American* 1993;269:42–9.
141. **Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R**. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 2006;439:344–8.
142. **Grenfell BT**. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science* 2004;303:327–332.
143. **Knossow M, Skehel JJ**. Variation and infectivity neutralization in influenza. *Immunology* 2006;119:1–7.
144. **Seidel N, Sauerbrei A, Wutzler P, Schmidtke M**. Hemagglutinin 222D/G polymorphism facilitates fast intra-host evolution of pandemic (H1N1) 2009 influenza A viruses. *PLoS ONE* 2014;9:1–11.
145. **Eshaghi A, Shalhoub S, Rosenfeld P, Li A, Higgins RR, *et al.*** Multiple influenza a (H3N2) mutations conferring resistance to neuraminidase inhibitors in a bone marrow transplant recipient. *Antimicrobial Agents and Chemotherapy* 2014;58:7188–7197.
146. **Ghedin E, Holmes EC, DePasse JV, Pinilla LT, Fitch A, *et al.*** Presence of oseltamivir-resistant pandemic A/H1N1 minor variants before drug therapy with subsequent selection and transmission. *J Infect Dis* 2012;206:1504–1511.
147. **Wei K, Sun H, Sun Z, Sun Y, Kong W, *et al.*** Influenza A virus acquires enhanced pathogenicity and transmissibility after serial passages in swine. *J Virol* 2014;88:11981–11994.
148. **Marshall N, Priyamvada L, Ende Z, Steel J, Lowen AC**. Influenza Virus Reassortment Occurs with High Frequency in the Absence of Segment Mismatch. *PLOS Pathogens* 2013;9:e1003421.
149. **McDonald SM, Nelson MI, Turner PE, Patton JT**. Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nat Rev Microbiol* 2016;14:448–460.
150. **Stöhr K, Bucher D, Colgate T, Wood J**. Influenza virus surveillance, vaccine strain selection, and manufacture. *Methods Mol Biol* 2012;865:147–162.
151. **Sartwell PE, Long AP**. The Army experience with influenza, 1946-1947; epidemiological aspects. *Am J Hyg* 1948;47:135–141.
152. **Centers for Disease Control and Prevention**. Key Facts About Seasonal Flu Vaccine. *Centers for Disease Control and Prevention*. https://www.cdc.gov/flu/prevent/keyfacts.htm (2021, accessed 15 September 2021).
153. **Parodi V, de Florentiis D, Martini M, Ansaldi F**. Inactivated influenza vaccines: recent progress and implications for the elderly. *Drugs & aging* 2011;28:93–106.
154. **Barberis I, Myles P, Ault SK, Bragazzi NL, Martini M**. History and evolution of influenza control through vaccination: from the first monovalent vaccine to universal vaccines. *Journal of preventive medicine and hygiene* 2016;57:E115–E120.
155. **Ambrose CS, Levin MJ**. The rationale for quadrivalent influenza vaccines. *Human Vaccines & Immunotherapeutics* 2012;8:81–88.
156. **Tisa V, Barberis I, Faccio V, Paganino C, Trucchi C, *et al.*** Quadrivalent influenza vaccine: a new opportunity to reduce the influenza burden. *Journal of preventive medicine and hygiene* 2016;57:E28-33.
157. **Gerdil C**. The annual production cycle for influenza vaccine. *Vaccine* 2003;21:1776–9.
158. **Sridhar S, Brokstad KA, Cox RJ**. Influenza Vaccination Strategies: Comparing Inactivated and Live Attenuated Influenza Vaccines. *Vaccines* 2015;3:373–89.
159. **Belongia EA, Kieke BA, Donahue JG, Coleman LA, Irving SA, *et al.*** Influenza vaccine effectiveness in Wisconsin during the 2007-08 season: comparison of interim and final results. *Vaccine* 2011;29:6558–6563.
160. **Belongia EA, Kieke BA, Donahue JG, Greenlee RT, Balish A, *et al.*** Effectiveness of Inactivated Influenza Vaccines Varied Substantially with Antigenic Match from the 2004–2005 Season to the 2006–2007 Season. *The Journal of Infectious Diseases* 2009;199:159–167.
161. **Rolfes MA, Flannery B, Chung JR, O'Halloran A, Garg S, *et al.*** Effects of Influenza Vaccination in the United States During the 2017–2018 Influenza Season. *Clinical Infectious Diseases* 2019;69:1845–1853.
162. **Treanor JJ, Talbot HK, Ohmit SE, Coleman LA, Thompson MG, *et al.*** Effectiveness of Seasonal Influenza Vaccines in the United States During a Season With Circulation of All Three Vaccine Strains. *Clinical Infectious Diseases* 2012;55:951–959.

163. **Grohskopf LA, Olsen SJ, Sokolow LZ, Bresee JS, Cox NJ, *et al.*** Prevention and control of seasonal influenza with vaccines: recommendations of the Advisory Committee on Immunization Practices (ACIP) -- United States, 2014-15 influenza season. *MMWR Morb Mortal Wkly Rep* 2014;63:691–697.

164. **Soema PC, Kompier R, Amorij J-P, Kersten GFA**. Current and next generation influenza vaccines: Formulation and production strategies. *European journal of pharmaceutics and biopharmaceutics : official journal of Arbeitsgemeinschaft fur Pharmazeutische Verfahrenstechnik eV* 2015;94:251–63.

165. **Kim H, Webster RG, Webby RJ**. Influenza Virus: Dealing with a Drifting and Shifting Pathogen. *Viral Immunol* 2018;31:174–183.

166. **Zost SJ, Parkhouse K, Gumina ME, Kim K, Perez SD, *et al.*** Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. *PNAS* 2017;114:12578–12583.

167. **Houser K, Subbarao K**. Influenza Vaccines: Challenges and Solutions. *Cell Host Microbe* 2015;17:295–300.

168. **Coelingh KL, Luke CJ, Jin H, Talaat KR**. Development of live attenuated influenza vaccines against pandemic influenza strains. *Expert Rev Vaccines* 2014;13:855–871.

169. **Subbarao K, Joseph T**. Scientific barriers to developing vaccines against avian influenza viruses. *Nat Rev Immunol* 2007;7:267–278.

170. **Krammer F, Palese P**. Influenza virus hemagglutinin stalk-based antibodies and vaccines. *Current Opinion in Virology* 2013;3:521–530.

171. **Corti D, Lanzavecchia A**. Broadly neutralizing antiviral antibodies. *Annu Rev Immunol* 2013;31:705–742.

172. **Tripp RA, Tompkins SM**. Virus-vectored influenza virus vaccines. *Viruses* 2014;6:3055–3079.

173. **Antrobus RD, Berthoud TK, Mullarkey CE, Hoschler K, Coughlan L, *et al.*** Coadministration of seasonal influenza vaccine and MVA-NP+M1 simultaneously achieves potent humoral and cell-mediated responses. *Mol Ther* 2014;22:233–238.

174. **Kreijtz JHCM, Goeijenbier M, Moesker FM, van den Dries L, Goeijenbier S, *et al.*** Safety and immunogenicity of a modified-vaccinia-virus-Ankara-based influenza A H5N1 vaccine: a randomised, double-blind phase 1/2a clinical trial. *Lancet Infect Dis* 2014;14:1196–1207.

175. **Price GE, Soboleski MR, Lo C-Y, Misplon JA, Quirion MR, *et al.*** Single-dose mucosal immunization with a candidate universal influenza vaccine provides rapid protection from virulent H5N1, H3N2 and H1N1 viruses. *PLoS One* 2010;5:e13162.

176. **Giles BM, Crevar CJ, Carter DM, Bissel SJ, Schultz-Cherry S, *et al.*** A computationally optimized hemagglutinin virus-like particle vaccine elicits broadly reactive antibodies that protect nonhuman primates from H5N1 infection. *J Infect Dis* 2012;205:1562–1570.

177. **Giles BM, Bissel SJ, Dealmeida DR, Wiley CA, Ross TM**. Antibody breadth and protective efficacy are increased by vaccination with computationally optimized hemagglutinin but not with polyvalent hemagglutinin-based H5N1 virus-like particle vaccines. *Clin Vaccine Immunol* 2012;19:128–139.

178. **Giles BM, Ross TM**. A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets. *Vaccine* 2011;29:3043–3054.

179. **Lee Y-T, Kim K-H, Ko E-J, Lee Y-N, Kim M-C, *et al.*** New vaccines against influenza virus. *Clin Exp Vaccine Res* 2014;3:12–28.

180. **Sridhar S, Begom S, Bermingham A, Hoschler K, Adamson W, *et al.*** Cellular immune correlates of protection against symptomatic pandemic influenza. *Nat Med* 2013;19:1305–1312.

181. **Even-Or O, Samira S, Ellis R, Kedar E, Barenholz Y**. Adjuvanted influenza vaccines. *Expert Rev Vaccines* 2013;12:1095–1108.

182. **Khurana S, Chearwae W, Castellino F, Manischewitz J, King LR, *et al.*** Vaccines with MF59 adjuvant expand the antibody repertoire to target protective sites of pandemic avian H5N1 influenza virus. *Sci Transl Med* 2010;2:15ra5.

183. **Khurana S, Verma N, Yewdell JW, Hilbert AK, Castellino F, *et al.*** MF59 adjuvant enhances diversity and affinity of antibody-mediated immune response to pandemic influenza vaccines. *Sci Transl Med* 2011;3:85ra48.

184. **Talaat KR, Luke CJ, Khurana S, Manischewitz J, King LR, *et al.*** A live attenuated influenza A(H5N1) vaccine induces long-term immunity in the absence of a primary antibody response. *J Infect Dis* 2014;209:1860–1869.

185. **Gurwith M, Lock M, Taylor EM, Ishioka G, Alexander J, *et al.*** Safety and immunogenicity of an oral, replicating adenovirus serotype 4 vector vaccine for H5N1 influenza: a randomised, double-blind, placebo-controlled, phase 1 study. *Lancet Infect Dis* 2013;13:238–250.

186. **Babu TM, Levine M, Fitzgerald T, Luke C, Sangster MY, *et al.*** Live attenuated H7N7 influenza vaccine primes for a vigorous antibody response to inactivated H7N7 influenza vaccine. *Vaccine* 2014;32:6798–6804.

187. **Price GE, Soboleski MR, Lo C-Y, Misplon JA, Pappas C, *et al.*** Vaccination focusing immunity on conserved antigens protects mice and ferrets against virulent H1N1 and H5N1 influenza A viruses. *Vaccine* 2009;27:6512–6521.

188. **Erbelding EJ, Post DJ, Stemmy EJ, Roberts PC, Augustine AD, *et al.*** A Universal Influenza Vaccine: The Strategic Plan for the National Institute of Allergy and Infectious Diseases. *J Infect Dis* 2018;218:347–354.

189. **Haq K, McElhaney JE**. Immunosenescence: Influenza vaccination and the elderly. *Curr Opin Immunol* 2014;29:38–42.

190. **Black S, Nicolay U, Vesikari T, Knuf M, Del Giudice G, *et al.*** Hemagglutination inhibition antibody titers as a correlate of protection for inactivated influenza vaccines in children. *Pediatr Infect Dis J* 2011;30:1081–1085.

191. **Pinto LH, Lamb RA**. Understanding the mechanism of action of the anti-influenza virus drug amantadine. *Trends Microbiol* 1995;3:271.

192. **Wang C, Takeuchi K, Pinto LH, Lamb RA**. Ion channel activity of influenza A virus M2 protein: characterization of the amantadine block. *Journal of virology* 1993;67:5585–94.

193. **Deyde VM, Xu X, Bright RA, Shaw M, Smith CB, *et al.*** Surveillance of Resistance to Adamantanes among Influenza A(H3N2) and A(H1N1) Viruses Isolated Worldwide. *The Journal of Infectious Diseases* 2007;196:249–257.

194. **Dawood FS, Jain S, Finelli L, Shaw MW, Lindstrom S, *et al.*** Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N Engl J Med* 2009;360:2605–2615.

195. **European Centre for Disease Prevention and Control**. Antiviral treatment of influenza. *European Centre for Disease Prevention and Control*. https://www.ecdc.europa.eu/en/seasonal-influenza/prevention-and-control/antivirals (2021, accessed 13 July 2021).

196. **Moscona A**. Global Transmission of Oseltamivir-Resistant Influenza. *New England Journal of Medicine* 2009;360:953–956.

197. **Abed Y, Pizzorno A, Bouhy X, Rheaume C, Boivin G**. Impact of Potential Permissive Neuraminidase Mutations on Viral Fitness of the H275Y Oseltamivir-Resistant Influenza A(H1N1)pdm09 Virus *In vitro*, in Mice and in Ferrets. *Journal of Virology* 2014;88:1652–1658.

198. **Hussain M, Galvin HD, Haw TY, Nutsford AN, Husain M**. Drug resistance in influenza a virus: The epidemiology and management. *Infection and Drug Resistance* 2017;10:121–134.

199. **Lackenby A, Hungnes O, Dudman SG, Meijer A, Paget WJ, *et al.*** Emergence of resistance to oseltamivir among influenza A(H1N1) viruses in Europe. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*;13. http://www.ncbi.nlm.nih.gov/pubmed/18445375 (2008).

200. **World Health Organization**. *Influenza A(H1N1) virus resistance to oseltamivir - 2008/2009 influenza season, northern hemisphere - 18 March 2019*. https://www.who.int/influenza/resources/documents/H1N1webupdate20090318_ed_ns.pdf (2009).

201. **Sheu TG, Deyde VM, Okomo-Adhiambo M, Garten RJ, Xu X, *et al.*** Surveillance for neuraminidase inhibitor resistance among human influenza A and B viruses circulating worldwide from 2004 to 2008. *Antimicrobial Agents and Chemotherapy* 2008;52:3284–3292.

202. **Okomo-adhiambo M, Sheu TG, Gubareva LV**. Assays for monitoring susceptibility of influenza viruses to neuraminidase inhibitors. *Influenza and other Respiratory Viruses* 2013;7:44–49.

203. **Hayden FG, Sugaya N, Hirotsu N, Lee N, de Jong MD, *et al.*** Baloxavir Marboxil for Uncomplicated Influenza in Adults and Adolescents. *New England Journal of Medicine* 2018;379:913–923.

204. **Roche**. *Roche announces FDA approval of Xofluza (baloxavir marboxil) for influenza*. https://www.roche.com/media/releases/med-cor-2018-10-24.htm (2018).

205. **Leneva IA, Russell RJ, Boriskin YS, Hay AJ**. Characteristics of arbidol-resistant mutants of influenza virus: Implications for the mechanism of anti-influenza action of arbidol. *Antiviral Research* 2009;81:132–140.

206. **Baranovich T, Wong S-S, Armstrong J, Marjuki H, Webby RJ, *et al.*** T-705 (Favipiravir) Induces Lethal Mutagenesis in Influenza A H1N1 Viruses *In vitro*. *Journal of Virology* 2013;87:3741–3751.

207. **Furuta Y, Komeno T, Nakamura T**. Favipiravir (T-705), a broad spectrum inhibitor of viralRNApolymerase. *Proceedings of the Japan Academy*;93.

208. **Haffizulla J, Hartman A, Hoppers M, Resnick H, Samudrala S, *et al.*** Effect of nitazoxanide in adults and adolescents with acute uncomplicated influenza: A double-blind, randomised, placebo-controlled, phase 2b/3 trial. *The Lancet Infectious Diseases* 2014;14:609–618.

209. **Rossignol JF**. Nitazoxanide: A first-in-class broad-spectrum antiviral agent. *Antiviral Research* 2014;110:94–103.

210. **Johnson & Johnson**. Pimodivir Alone or in Combination with Oseltamivir Demonstrated a Significant Reduction in Viral Load in Adults with Influenza A. https://www.jnj.com/media-center/press-releases/pimodivir-alone-or-in-combination-with-oseltamivir-demonstrated-a-significant-reduction-in-viral-load-in-adults-with-influenza-a (2017).

211. **Hamre D, Procknow JJ**. A new virus isolated from the human respiratory tract. *Proc Soc Exp Biol Med* 1966;121:190–193.

212. **Kahn JS, McIntosh K**. History and recent advances in coronavirus discovery. *Pediatr Infect Dis J* 2005;24:S223-227, discussion S226.

213. **Tyrrell DA, Bynoe ML**. Cultivation of viruses from a high proportion of patients with colds. *Lancet* 1966;1:76–77.

214. **Holmes KV**. SARS-associated coronavirus. *N Engl J Med* 2003;348:1948–1951.

215. **Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM**. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *N Engl J Med* 2012;367:1814–1820.

216. **Cui J, Li F, Shi Z-L**. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* 2019;17:181–192.

217. **Chen N, Zhou M, Dong X, Qu J, Gong F, *et al.*** Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet* 2020;395:507–513.

218. **Huang C, Wang Y, Li X, Ren L, Zhao J, *et al.*** Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* 2020;395:497–506.

219. **Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF**. The proximal origin of SARS-CoV-2. *Nat Med* 2020;26:450–452.

220. **Cyranoski D**. Mystery deepens over animal source of coronavirus. *Nature* 2020;579:18–19.

221. **Yu W-B, Tang G-D, Zhang L, Corlett RT**. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. *Zool Res* 2020;41:247–257.

222. **Hui DS, Azhar EI, Madani TA, Ntoumi F, Kock R, *et al.*** The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases* 2020;91:264–266.

223. **Wu JT, Leung K, Leung GM**. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* 2020;395:689–697.

224. **Guo Y-R, Cao Q-D, Hong Z-S, Tan Y-Y, Chen S-D, *et al.*** The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. *Military Medical Research* 2020;7:11.

225. **Ou X, Liu Y, Lei X, Li P, Mi D, *et al.*** Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun* 2020;11:1620.

226. **World Health Organization**. Naming the coronavirus disease (COVID-19) and the virus that causes it. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it (2020, accessed 6 August 2021).

227. **Zhu N, Zhang D, Wang W, Li X, Yang B, *et al.*** A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* 2020;382:727–733.

228. **Cucinotta D, Vanelli M**. WHO Declares COVID-19 a Pandemic. *Acta Biomed* 2020;91:157–160.

229. **World Health Organization**. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int (2021, accessed 31 January 2022).

230. **Jf C, S Y, Kh K, Kk T, H C, *et al.*** A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* 2020;395:514–523.

231. **Liu J, Liao X, Qian S, Yuan J, Wang F, *et al.*** Community Transmission of Severe Acute Respiratory Syndrome Coronavirus 2, Shenzhen, China, 2020. *Emerg Infect Dis* 2020;26:1320–1323.

232. **Meselson M**. Droplets and Aerosols in the Transmission of SARS-CoV-2. *N Engl J Med* 2020;382:2063–2063.

233. **Morawska L, Cao J**. Airborne transmission of SARS-CoV-2: The world should face the reality. *Environment International* 2020;139:105730.

234. **Sommerstein R, Fux CA, Vuichard-Gysin D, Abbas M, Marschall J, *et al.*** Risk of SARS-CoV-2 transmission by aerosols, the rational use of masks, and protection of healthcare workers from COVID-19. *Antimicrob Resist Infect Control* 2020;9:100.

235. **Tang X, Wu C, Li X, Song Y, Yao X, *et al.*** On the origin and continuing evolution of SARS-CoV-2. *National Science Review* 2020;7:1012–1023.

236. **van Doremalen N, Bushmaker T, Morris DH, Holbrook MG, Gamble A, *et al.*** Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *N Engl J Med* 2020;382:1564–1567.

237. **Zhang R, Li Y, Zhang AL, Wang Y, Molina MJ**. Identifying airborne transmission as the dominant route for the spread of COVID-19. *Proc Natl Acad Sci USA* 2020;117:14857–14863.

238. **Parasher A**. COVID-19: Current understanding of its Pathophysiology, Clinical presentation and Treatment. *Postgrad Med J* 2021;97:312–320.

239. **Meyerowitz EA, Richterman A, Gandhi RT, Sax PE**. Transmission of SARS-CoV-2: A Review of Viral, Host, and Environmental Factors. *Ann Intern Med* 2020;M20-5008.

240. **Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, *et al.*** SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *Lancet Microbe* 2021;2:e13–e22.
241. **Guan W, Ni Z, Hu Y, Liang W, Ou C, *et al.*** Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* 2020;382:1708–1720.
242. **Sims AC, Baric RS, Yount B, Burkett SE, Collins PL, *et al.*** Severe acute respiratory syndrome coronavirus infection of human ciliated airway epithelia: role of ciliated cells in viral spread in the conducting airways of the lungs. *J Virol* 2005;79:15511–15524.
243. **Hu B, Guo H, Zhou P, Shi Z-L**. Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol* 2021;19:141–154.
244. **Li W**. Delving deep into the structural aspects of a furin cleavage site inserted into the spike protein of SARS-CoV-2: A structural biophysical perspective. *Biophysical Chemistry* 2020;264:106420.
245. **Cascella M, Rajnik M, Aleem A, Dulebohn SC, Di Napoli R**. Features, Evaluation, and Treatment of Coronavirus (COVID-19). In: *StatPearls*. Treasure Island (FL): StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK554776/ (2021, accessed 13 August 2021).
246. **Xu Z, Shi L, Wang Y, Zhang J, Huang L, *et al.*** Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir Med* 2020;8:420–422.
247. **Lu X, Zhang L, Du H, Zhang J, Li YY, *et al.*** SARS-CoV-2 Infection in Children. *N Engl J Med* 2020;382:1663–1665.
248. **Wu Z, McGoogan JM**. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA* 2020;323:1239–1242.
249. **Liu Y, Mao B, Liang S, Yang J-W, Lu H-W, *et al.*** Association between age and clinical characteristics and outcomes of COVID-19. *Eur Respir J* 2020;55:2001112.
250. **Wu C, Chen X, Cai Y, Xia J, Zhou X, *et al.*** Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Intern Med* 2020;180:934.
251. **Callard F, Perego E**. How and why patients made Long Covid. *Social Science & Medicine* 2021;268:113426.
252. **Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ**. Nidovirales: evolving the largest RNA virus genome. *Virus Res* 2006;117:17–37.
253. **Coronaviridae Study Group of the International Committee on Taxonomy of Viruses**. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 2020;5:536–544.
254. **Fung TS, Liu DX**. Human Coronavirus: Host-Pathogen Interaction. *Annu Rev Microbiol* 2019;73:529–557.
255. **Woo PCY, Lau SKP, Lam CSF, Lau CCY, Tsang AKL, *et al.*** Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J Virol* 2012;86:3995–4008.
256. **Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, *et al.*** Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018;34:4121–4123.
257. **Elbe S, Buckland-Merrett G**. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* 2017;1:33–46.
258. **Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, *et al.*** A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology* 2020;5:1403–1407.
259. **World Health Organization**. Tracking SARS-CoV-2 variants. https://www.who.int/emergencies/emergency-health-kits/trauma-emergency-surgery-kit-who-tesk-2019/tracking-SARS-CoV-2-variants (2021, accessed 19 August 2021).
260. **Centers for Disease Control and Prevention**. SARS-CoV-2 Variants. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html (2021).
261. **Reichmuth M, Hodcroft E, Riou J, Althaus CL, Schibler M, *et al.*** Transmission of SARS-CoV-2 variants in Switzerland. https://ispmbern.github.io/covid-19/variants/ (2021, accessed 4 March 2021).
262. **Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, *et al.*** Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 2021;372:eabg3055.
263. **SAGE-EMG, SPI-B, Tranmission Group**. Mitigations to Reduce Transmission of the new variant SARS-CoV-2 virus. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/948607/s0995-mitigations-to-reduce-transmission-of-the-new-variant.pdf (2020, accessed 4 March 2021).
264. **GOV.UK - Scientific Advisory Group for Emergencies**. NERVTAG: Update note on B.1.1.7 severity. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/961042/S1095_NERVTAG_update_note_on_B.1.1.7_severity_20210211.pdf (2021, accessed 4 March 2021).

265. **Rambaut A, Loman N, Pybus O, Barcly W, Barrett J, *et al.*** Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (2021, accessed 4 March 2021).
266. **Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, *et al.*** Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*;10.
267. **Pearson CAB, Russel TW, Davies NG, Kucharski AJ, Group CC-19 working, *et al.*** Estimates of severity and transmissibility of novel South Africa SARS-CoV-2 variant 501Y.V2. https://cmmid.github.io/topics/covid19/reports/sa-novel-variant/2021_01_11_Transmissibility_and_severity_of_501Y_V2_in_SA.pdf (2021, accessed 4 March 2021).
268. **Johnson & Johnson**. Johnson & Johnson COVID-19 Vaccine Authorized by U.S. FDA For Emergency Use - First Single-Shot Vaccine in Fight Against Global Pandemic. 27 February 2021.
269. **Novavax**. Novavax COVID-19 Vaccine Demonstrates 89.3% Efficacy in UK Phase 3 Trial. *Press release,* 28 January 2021.
270. **Rawlinson K, Sample I**. Oxford Covid vaccine has 10% efficacy against South African variant, study suggests. *The Guardian*, 8 February 2021.
271. **Sabino EC, Buss LF, Carvalho MPS, Prete CA, Crispim MAE, *et al.*** Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *The Lancet* 2021;397:452–455.
272. **Hoffmann M, Arora P, Groß R, Seidel A, Hörnich BF, *et al.*** SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. *Cell*. Epub ahead of print March 2021. DOI: 10.1016/j.cell.2021.03.036.
273. **Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, *et al.*** SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. *Microorganisms* 2021;9:1542.
274. **Hagen, Ashley**. How Dangerous Is the Delta Variant (B.1.617.2)? *American Society for Microbiology*. https://asm.org/Articles/2021/July/How-Dangerous-is-the-Delta-Variant-B-1-617-2 (2021, accessed 19 August 2021).
275. **Centers for Disease Control and Prevention**. Delta Variant. *Coronavirus Disease 2019 (COVID-19)*. https://www.cdc.gov/coronavirus/2019-ncov/variants/delta-variant.html (2021, accessed 19 August 2021).
276. **Gowrisankar A, Priyanka TMC, Banerjee S**. Omicron: a mysterious variant of concern. *Eur Phys J Plus* 2022;137:100.
277. **Yao L, Zhu K-L, Jiang X-L, Wang X-J, Zhan B-D, *et al.*** Omicron subvariants escape antibodies elicited by vaccination and BA.2.2 infection. *Lancet Infect Dis* 2022;22:1116–1117.
278. **Khandia R, Singhal S, Alqahtani T, Kamal MA, El-Shall NA, *et al.*** Emergence of SARS-CoV-2 Omicron (B.1.1.529) variant, salient features, high global health concerns and strategies to counter it amid ongoing COVID-19 pandemic. *Environ Res* 2022;209:112816.
279. **Ahmed SF, Quadeer AA, McKay MR**. SARS-CoV-2 T Cell Responses Elicited by COVID-19 Vaccines or Infection Are Expected to Remain Robust against Omicron. *Viruses* 2022;14:79.
280. **Altarawneh HN, Chemaitelly H, Ayoub HH, Tang P, Hasan MR, *et al.*** Effects of Previous Infection and Vaccination on Symptomatic Omicron Infections. *N Engl J Med* 2022;387:21–34.
281. **Nemet I, Kliker L, Lustig Y, Zuckerman N, Erster O, *et al.*** Third BNT162b2 Vaccination Neutralization of SARS-CoV-2 Omicron Infection. *N Engl J Med* 2022;386:492–494.
282. **Cao Y, Yisimayi A, Jian F, Song W, Xiao T, *et al.*** BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature* 2022;1–10.
283. **Hachmann NP, Miller J, Collier AY, Ventura JD, Yu J, *et al.*** Neutralization Escape by SARS-CoV-2 Omicron Subvariants BA.2.12.1, BA.4, and BA.5. *N Engl J Med* 2022;387:86–88.
284. **Kandeel M, Ibrahim A, Fayez M, Al-Nazawi M**. From SARS and MERS CoVs to SARS-CoV-2: Moving toward more biased codon usage in viral structural and nonstructural genes. *J Med Virol* 2020;92:660–666.
285. **Wang Y, Mao J-M, Wang G-D, Luo Z-P, Yang L, *et al.*** Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames. *Sci Rep* 2020;10:12331.
286. **Rice AM, Castillo Morales A, Ho AT, Mordstein C, Mühlhausen S, *et al.*** Evidence for Strong Mutation Bias toward, and Selection against, U Content in SARS-CoV-2: Implications for Vaccine Design. *Mol Biol Evol* 2021;38:67–83.
287. **Vennema H, Godeke GJ, Rossen JW, Voorhout WF, Horzinek MC, *et al.*** Nucleocapsid-independent assembly of coronavirus-like particles by co-expression of viral envelope protein genes. *EMBO J* 1996;15:2020–2028.
288. **Hofmann H, Pöhlmann S**. Cellular entry of the SARS coronavirus. *Trends Microbiol* 2004;12:466–472.
289. **Lan J, Ge J, Yu J, Shan S, Zhou H, *et al.*** Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 2020;581:215–220.

290. **Siu YL, Teoh KT, Lo J, Chan CM, Kien F, *et al.*** The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. *J Virol* 2008;82:11318–11330.
291. **Gao Y, Yan L, Huang Y, Liu F, Zhao Y, *et al.*** Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* 2020;368:779–782.
292. **Wang C, Liu Z, Chen Z, Huang X, Xu M, *et al.*** The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J Med Virol* 2020;92:667–674.
293. **Masters PS**. The molecular biology of coronaviruses. *Adv Virus Res* 2006;66:193–292.
294. **Mortola E, Roy P**. Efficient assembly and release of SARS coronavirus-like particles by a heterologous expression system. *FEBS Lett* 2004;576:174–178.
295. **Wang C, Zheng X, Gai W, Zhao Y, Wang H, *et al.*** MERS-CoV virus-like particles produced in insect cells induce specific humoural and cellular imminity in rhesus macaques. *Oncotarget* 2017;8:12686–12694.
296. **Li F**. Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu Rev Virol* 2016;3:237–261.
297. **Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, *et al.*** Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020;367:1260–1263.
298. **Raj VS, Mou H, Smits SL, Dekkers DHW, Müller MA, *et al.*** Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. *Nature* 2013;495:251–254.
299. **Wentworth DE, Holmes KV**. Molecular determinants of species specificity in the coronavirus receptor aminopeptidase N (CD13): influence of N-linked glycosylation. *J Virol* 2001;75:9741–9752.
300. **Hulswit RJG, Lang Y, Bakkers MJG, Li W, Li Z, *et al.*** Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. *Proc Natl Acad Sci U S A* 2019;116:2681–2690.
301. **Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, *et al.*** SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 2020;181:271-280.e8.
302. **Hofmann H, Pyrc K, van der Hoek L, Geier M, Berkhout B, *et al.*** Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. *Proc Natl Acad Sci U S A* 2005;102:7988–7993.
303. **Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, *et al.*** Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 2003;426:450–454.
304. **Hikmet F, Méar L, Edvinsson Å, Micke P, Uhlén M, *et al.*** The protein expression profile of ACE2 in human tissues. *Mol Syst Biol* 2020;16:e9610.
305. **Venkatakrishnan AJ, Puranik A, Anand A, Zemmour D, Yao X, *et al.*** Knowledge synthesis of 100 million biomedical documents augments the deep expression profiling of coronavirus receptors. *Elife*;9. Epub ahead of print 1 May 2020. DOI: 10.7554/elife.58040.
306. **Chandrashekar A, Liu J, Martinot AJ, McMahan K, Mercado NB, *et al.*** SARS-CoV-2 infection protects against rechallenge in rhesus macaques. *Science* 2020;369:812–817.
307. **Shi J, Wen Z, Zhong G, Yang H, Wang C, *et al.*** Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. *Science* 2020;368:1016–1020.
308. **Zhao X, Chen D, Szabla R, Zheng M, Li G, *et al.*** Broad and Differential Animal Angiotensin-Converting Enzyme 2 Receptor Usage by SARS-CoV-2. *J Virol* 2020;94:e00940-20.
309. **Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, *et al.*** A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;579:270–273.
310. **Letko M, Marzi A, Munster V**. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol* 2020;5:562–569.
311. **Tortorici MA, Veesler D**. Structural insights into coronavirus entry. *Adv Virus Res* 2019;105:93–116.
312. **Schoeman D, Fielding BC**. Coronavirus envelope protein: current knowledge. *Virol J* 2019;16:69.
313. **Voss D, Kern A, Traggiai E, Eickmann M, Stadler K, *et al.*** Characterization of severe acute respiratory syndrome coronavirus membrane protein. *FEBS Lett* 2006;580:968–973.
314. **Liu J, Sun Y, Qi J, Chu F, Wu H, *et al.*** The membrane protein of severe acute respiratory syndrome coronavirus acts as a dominant immunogen revealed by a clustering region of novel functionally and structurally defined cytotoxic T-lymphocyte epitopes. *J Infect Dis* 2010;202:1171–1180.
315. **Chang C, Hou M-H, Chang C-F, Hsiao C-D, Huang T**. The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral Res* 2014;103:39–50.
316. **Mu J, Xu J, Zhang L, Shu T, Wu D, *et al.*** SARS-CoV-2-encoded nucleocapsid protein acts as a viral suppressor of RNA interference in cells. *Sci China Life Sci* 2020;1–4.
317. **Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, *et al.*** Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 2020;181:281-292.e6.

318. **Wu A, Peng Y, Huang B, Ding X, Wang X, *et al.*** Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* 2020;27:325–328.

319. **Chan JF-W, Kok K-H, Zhu Z, Chu H, To KK-W, *et al.*** Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect* 2020;9:221–236.

320. **Harcourt BH, Jukneliene D, Kanjanahaluethai A, Bechill J, Severson KM, *et al.*** Identification of severe acute respiratory syndrome coronavirus replicase products and characterization of papain-like protease activity. *J Virol* 2004;78:13600–13612.

321. **Perlman S, Netland J**. Coronaviruses post-SARS: update on replication and pathogenesis. *Nat Rev Microbiol* 2009;7:439–450.

322. **Snijder EJ, Decroly E, Ziebuhr J**. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. *Adv Virus Res* 2016;96:59–126.

323. **Narayanan K, Huang C, Lokugamage K, Kamitani W, Ikegami T, *et al.*** Severe acute respiratory syndrome coronavirus nsp1 suppresses host gene expression, including that of type I interferon, in infected cells. *J Virol* 2008;82:4471–4479.

324. **Pfefferle S, Schöpf J, Kögl M, Friedel CC, Müller MA, *et al.*** The SARS-Coronavirus-Host Interactome: Identification of Cyclophilins as Target for Pan-Coronavirus Inhibitors. *PLOS Pathogens* 2011;7:e1002331.

325. **Züst R, Cervantes-Barragán L, Kuri T, Blakqori G, Weber F, *et al.*** Coronavirus Non-Structural Protein 1 Is a Major Pathogenicity Factor: Implications for the Rational Design of Coronavirus Vaccines. *PLOS Pathogens* 2007;3:e109.

326. **Cornillez-Ty CT, Liao L, Yates JR, Kuhn P, Buchmeier MJ**. Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. *J Virol* 2009;83:10314–10318.

327. **Ziebuhr J, Thiel V, Gorbalenya AE**. The autocatalytic release of a putative RNA virus transcription factor from its polyprotein precursor involves two paralogous papain-like proteases that cleave the same peptide bond. *J Biol Chem* 2001;276:33220–33232.

328. **Serrano P, Johnson MA, Chatterjee A, Neuman BW, Joseph JS, *et al.*** Nuclear Magnetic Resonance Structure of the Nucleic Acid-Binding Domain of Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein 3. *J Virol* 2009;83:12998–13008.

329. **Clementz MA, Kanjanahaluethai A, O'Brien TE, Baker SC**. Mutation in murine coronavirus replication protein nsp4 alters assembly of double membrane vesicles. *Virology* 2008;375:118–129.

330. **Gadlage MJ, Sparks JS, Beachboard DC, Cox RG, Doyle JD, *et al.*** Murine hepatitis virus nonstructural protein 4 regulates virus-induced membrane modifications and replication complex function. *J Virol* 2010;84:280–290.

331. **Manolaridis I, Wojdyla JA, Panjikar S, Snijder EJ, Gorbalenya AE, *et al.*** Structure of the C-terminal domain of nsp4 from feline coronavirus. *Acta Crystallogr D Biol Crystallogr* 2009;65:839–846.

332. **Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R**. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* 2003;300:1763–1767.

333. **Lu Y, Lu X, Denison MR**. Identification and characterization of a serine-like proteinase of the murine coronavirus MHV-A59. *Journal of Virology* 1995;69:3554–3559.

334. **Yang H, Yang M, Ding Y, Liu Y, Lou Z, *et al.*** The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *PNAS* 2003;100:13190–13195.

335. **Cottam EM, Whelband MC, Wileman T**. Coronavirus NSP6 restricts autophagosome expansion. *Autophagy* 2014;10:1426–1441.

336. **Cottam EM, Maier HJ, Manifava M, Vaux LC, Chandra-Schoenfelder P, *et al.*** Coronavirus nsp6 proteins generate autophagosomes from the endoplasmic reticulum via an omegasome intermediate. *Autophagy* 2011;7:1335–1347.

337. **Oostra M, Hagemeijer MC, van Gent M, Bekker CPJ, te Lintelo EG, *et al.*** Topology and membrane anchoring of the coronavirus replication complex: not all hydrophobic domains of nsp3 and nsp6 are membrane spanning. *J Virol* 2008;82:12392–12405.

338. **Imbert I, Guillemot J-C, Bourhis J-M, Bussetta C, Coutard B, *et al.*** A second, non-canonical RNA-dependent RNA polymerase in SARS Coronavirus. *EMBO J* 2006;25:4933–4942.

339. **Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, *et al.*** One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proc Natl Acad Sci U S A* 2014;111:E3900-3909.

340. **Zhai Y, Sun F, Li X, Pang H, Xu X, *et al.*** Insights into SARS-CoV transcription and replication from the structure of the nsp7–nsp8 hexadecamer. *Nat Struct Mol Biol* 2005;12:980–986.

341. **Chen Y, Cai H, Pan J, Xiang N, Tien P, *et al.*** Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase. *PNAS* 2009;106:3484–3489.

342. **Miknis ZJ, Donaldson EF, Umland TC, Rimmer RA, Baric RS, *et al.*** Severe acute respiratory syndrome coronavirus nsp9 dimerization is essential for efficient viral growth. *J Virol* 2009;83:3007–3018.
343. **Donaldson EF, Graham RL, Sims AC, Denison MR, Baric RS**. Analysis of murine hepatitis virus strain A59 temperature-sensitive mutant TS-LA6 suggests that nsp10 plays a critical role in polyprotein processing. *J Virol* 2007;81:7086–7098.
344. **Bouvet M, Lugari A, Posthuma CC, Zevenhoven JC, Bernard S, *et al.*** Coronavirus Nsp10, a critical co-factor for activation of multiple replicative enzymes. *J Biol Chem* 2014;289:25783–25796.
345. **Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, *et al.*** RNA 3’-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. *PNAS* 2012;109:9372–9377.
346. **Chen Y, Su C, Ke M, Jin X, Xu L, *et al.*** Biochemical and Structural Insights into the Mechanisms of SARS Coronavirus RNA Ribose 2′-O-Methylation by nsp16/nsp10 Protein Complex. *PLoS Pathog* 2011;7:e1002294.
347. **Decroly E, Debarnot C, Ferron F, Bouvet M, Coutard B, *et al.*** Crystal structure and functional analysis of the SARS-coronavirus RNA cap 2’-O-methyltransferase nsp10/nsp16 complex. *PLoS Pathog* 2011;7:e1002059.
348. **te Velthuis AJW, Arnold JJ, Cameron CE, van den Worm SHE, Snijder EJ**. The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. *Nucleic Acids Res* 2010;38:203–214.
349. **te Velthuis AJW, van den Worm SHE, Sims AC, Baric RS, Snijder EJ, *et al.*** Zn(2+) inhibits coronavirus and arterivirus RNA polymerase activity *in vitro* and zinc ionophores block the replication of these viruses in cell culture. *PLoS Pathog* 2010;6:e1001176.
350. **Xu X, Liu Y, Weiss S, Arnold E, Sarafianos SG, *et al.*** Molecular model of SARS coronavirus polymerase: implications for biochemical functions and drug design. *Nucleic Acids Res* 2003;31:7117–7130.
351. **Adedeji AO, Marchand B, Velthuis AJW te, Snijder EJ, Weiss S, *et al.*** Mechanism of Nucleic Acid Unwinding by SARS-CoV Helicase. *PLOS ONE* 2012;7:e36521.
352. **Ivanov KA, Thiel V, Dobbe JC, van der Meer Y, Snijder EJ, *et al.*** Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase. *J Virol* 2004;78:5619–5632.
353. **Ivanov KA, Ziebuhr J**. Human coronavirus 229E nonstructural protein 13: characterization of duplex-unwinding, nucleoside triphosphatase, and RNA 5’-triphosphatase activities. *J Virol* 2004;78:7833–7838.
354. **Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR**. High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *J Virol* 2007;81:12135–12144.
355. **Sexton NR, Smith EC, Blanc H, Vignuzzi M, Peersen OB, *et al.*** Homology-Based Identification of a Mutation in the Coronavirus RNA-Dependent RNA Polymerase That Confers Resistance to Multiple Mutagens. *J Virol* 2016;90:7415–7428.
356. **Ricagno S, Egloff M-P, Ulferts R, Coutard B, Nurizzo D, *et al.*** Crystal structure and mechanistic determinants of SARS coronavirus nonstructural protein 15 define an endoribonuclease family. *Proc Natl Acad Sci U S A* 2006;103:11892–11897.
357. **Deng X, Hackbart M, Mettelman RC, O’Brien A, Mielech AM, *et al.*** Coronavirus nonstructural protein 15 mediates evasion of dsRNA sensors and limits apoptosis in macrophages. *PNAS* 2017;114:E4251–E4260.
358. **Hackbart M, Deng X, Baker SC**. Coronavirus endoribonuclease targets viral polyuridine sequences to evade activating host sensors. *PNAS* 2020;117:8094–8103.
359. **Kindler E, Gil-Cruz C, Spanier J, Li Y, Wilhelm J, *et al.*** Early endonuclease-mediated evasion of RNA sensing ensures efficient coronavirus replication. *PLOS Pathogens* 2017;13:e1006195.
360. **Decroly E, Imbert I, Coutard B, Bouvet M, Selisko B, *et al.*** Coronavirus nonstructural protein 16 is a cap-0 binding enzyme possessing (nucleoside-2’O)-methyltransferase activity. *J Virol* 2008;82:8071–8084.
361. **Menachery VD, Debbink K, Baric RS**. Coronavirus non-structural protein 16: evasion, attenuation, and possible treatments. *Virus Res* 2014;194:191–199.
362. **Züst R, Cervantes-Barragan L, Habjan M, Maier R, Neuman BW, *et al.*** Ribose 2′-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. *Nat Immunol* 2011;12:137–143.
363. **Millet JK, Whittaker GR**. Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein. *Proc Natl Acad Sci U S A* 2014;111:15214–15219.
364. **Tvarogová J, Madhugiri R, Bylapudi G, Ferguson LJ, Karl N, *et al.*** Identification and Characterization of a Human Coronavirus 229E Nonstructural Protein 8-Associated RNA 3’-Terminal Adenylyltransferase Activity. *J Virol* 2019;93:e00291-19.
365. **Zumla A, Chan JFW, Azhar EI, Hui DSC, Yuen K-Y**. Coronaviruses - drug discovery and therapeutic options. *Nat Rev Drug Discov* 2016;15:327–347.
366. **Yang N, Shen H-M**. Targeting the Endocytic Pathway and Autophagy Process as a Novel Therapeutic Strategy in COVID-19. *Int J Biol Sci* 2020;16:1724–1731.

367. **Millet JK, Whittaker GR**. Physiological and molecular triggers for SARS-CoV membrane fusion and entry into host cells. *Virology* 2018;517:3–8.

368. **Tang T, Bidon M, Jaimes JA, Whittaker GR, Daniel S**. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Res* 2020;178:104792.

369. **Kawase M, Shirato K, van der Hoek L, Taguchi F, Matsuyama S**. Simultaneous treatment of human bronchial epithelial cells with serine and cysteine protease inhibitors prevents severe acute respiratory syndrome coronavirus entry. *J Virol* 2012;86:6537–6545.

370. **Greber UF, Singh I, Helenius A**. Mechanisms of virus uncoating. *Trends Microbiol* 1994;2:52–56.

371. **Tok TT, Tatar G**. Structures and Functions of Coronavirus Proteins: Molecular Modeling of Viral Nucleoprotein. *International Journal of Virology* 2017;2:7.

372. **Nakagawa K, Lokugamage KG, Makino S**. Viral and Cellular mRNA Translation in Coronavirus-Infected Cells. *Adv Virus Res* 2016;96:165–192.

373. **King AMQ, Adams MJ**. *Virus Taxonomy Ninth Report of the International Committee on Taxonomy of Viruses*. San Diego, CA, USA: Elsevier Science & Technology Books. http://international.scholarvox.com/book/88812192 (2011, accessed 9 August 2021).

374. **Fehr AR, Perlman S**. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Coronaviruses: Methods and Protocols* 2015;1–23.

375. **Sola I, Mateos-Gomez PA, Almazan F, Zuñiga S, Enjuanes L**. RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. *RNA Biol* 2011;8:237–248.

376. **Fung TS, Liu DX**. Coronavirus infection, ER stress, apoptosis and innate immunity. *Front Microbiol*;0. Epub ahead of print 2014. DOI: 10.3389/fmicb.2014.00296.

377. **Ma H-C, Fang C-P, Hsieh Y-C, Chen S-C, Li H-C, *et al.*** Expression and membrane integration of SARS-CoV M protein. *J Biomed Sci* 2008;15:301–310.

378. **Krijnse-Locker J, Ericsson M, Rottier PJ, Griffiths G**. Characterization of the budding compartment of mouse hepatitis virus: evidence that transport from the RER to the Golgi complex requires only one vesicular transport step. *J Cell Biol* 1994;124:55–70.

379. **Tooze J, Tooze S, Warren G**. Replication of coronavirus MHV-A59 in sac- cells: determination of the first site of budding of progeny virions. *Eur J Cell Biol* 1984;33:281–293.

380. **de Haan CAM, Rottier PJM**. Molecular interactions in the assembly of coronaviruses. *Adv Virus Res* 2005;64:165–230.

381. **Bos ECW, Luytjes W, Meulen HVD, Koerten HK, Spaan WJM**. The production of recombinant infectious DI-particles of a murine coronavirus in the absence of helper virus. *Virology*;218. Epub ahead of print 1 April 1996. DOI: 10.1006/viro.1996.0165.

382. **Buratta S, Tancini B, Sagini K, Delo F, Chiaradia E, *et al.*** Lysosomal Exocytosis, Exosome Release and Secretory Autophagy: The Autophagic- and Endo-Lysosomal Systems Go Extracellular. *Int J Mol Sci* 2020;21:E2576.

383. **Malik YA**. Properties of Coronavirus and SARS-CoV-2. *Malays J Pathol* 2020;42:3–11.

384. **Mehrbod P, Ande SR, Alizadeh J, Rahimizadeh S, Shariati A, *et al.*** The roles of apoptosis, autophagy and unfolded protein response in arbovirus, influenza virus, and HIV infections. *Virulence* 2019;10:376–413.

385. **Moldoveanu B, Otmishi P, Jani P, Walker J, Sarmiento X, *et al.*** Inflammatory mechanisms in the lung. *J Inflamm Res* 2009;2:1–11.

386. **Rockett RJ, Arnott A, Lam C, Sadsad R, Timms V, *et al.*** Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nat Med* 2020;26:1398–1404.

387. **van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, *et al.*** Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution* 2020;83:104351–104351.

388. **Cushing A, Kamali A, Winters M, Hopmans ES, Bell JM, *et al.*** Emergence of Hemagglutinin Mutations During the Course of Influenza Infection. *Sci Rep* 2015;5:16178.

389. **Flaherty P, Natsoulis G, Muralidharan O, Winters M, Buenrostro J, *et al.*** Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Research* 2012;40:e2–e2.

390. **Simon-Loriere E, Holmes EC**. Why do RNA viruses recombine? *Nat Rev Microbiol* 2011;9:617–626.

391. **Hein J, Schierup M, Wiuf C**. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory*. Oxford University Press, USA; 2004.

392. **McVean G, Awadalla P, Fearnhead P**. A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. *Genetics* 2002;160:1231–1241.

393. **Capobianchi MR, Rueca M, Messina F, Giombini E, Carletti F, *et al.*** Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. *Clinical Microbiology and Infection* 2020;26:954–956.

394. **Jary A, Leducq V, Malet I, Marot S, Klement-Frutos E, *et al.*** Evolution of viral quasispecies during SARS-CoV-2 infection. *Clin Microbiol Infect* 2020;26:1560.e1-1560.e4.

395. **Domingo E, Sheldon J, Perales C**. Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews* 2012;76:159–216.

396. **Sun F, Wang X, Tan S, Dan Y, Lu Y, *et al.*** SARS-CoV-2 Quasispecies Provides an Advantage Mutation Pool for the Epidemic Variants. *Microbiol Spectr*. Epub ahead of print 4 August 2021. DOI: 10.1128/Spectrum.00261-21.

397. **Xu D, Zhang Z, Wang F-S**. SARS-Associated Coronavirus Quasispecies in Individual Patients. *New England Journal of Medicine* 2004;350:1366–1367.

398. **Park D, Huh HJ, Kim YJ, Son D-S, Jeon H-J, *et al.*** Analysis of intrapatient heterogeneity uncovers the microevolution of Middle East respiratory syndrome coronavirus. *Cold Spring Harb Mol Case Stud* 2016;2:a001214.

399. **Armero A, Berthet N, Avarre J-C**. Intra-Host Diversity of SARS-Cov-2 Should Not Be Neglected: Case of the State of Victoria, Australia. *Viruses* 2021;13:133–133.

400. **Shen Z, Xiao Y, Kang L, Ma W, Shi L, *et al.*** Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. *Clinical Infectious Diseases* 2020;71:713–720.

401. **World Health Organization**. *Looking back at a year that changed the world: WHO's response to COVID-19*. https://www.who.int/publications/m/item/looking-back-at-a-year-that-changed-the-world-who-s-response-to-covid-19 (22 January 2021, accessed 2 August 2022).

402. **World Health Organization**. *Global surveillance for COVID-19 caused by human infection with COVID-19 virus: interim guidance, 20 March 2020*. WHO/2019-nCoV/SurveillanceGuidance/2020.6; World Health Organization. https://apps.who.int/iris/handle/10665/331506 (2020, accessed 24 August 2021).

403. **Spiteri G, Fielding J, Diercke M, Campese C, Enouf V, *et al.*** First cases of coronavirus disease 2019 (COVID-19) in the WHO European Region, 24 January to 21 February 2020. *Eurosurveillance*;25. Epub ahead of print 5 March 2020. DOI: 10.2807/1560-7917.ES.2020.25.9.2000178.

404. **European Centre for Disease Prevention and Control**. EU level surveillance of COVID-19. *European Centre for Disease Prevention and Control*. https://www.ecdc.europa.eu/en/covid-19/surveillance (accessed 2 August 2022).

405. **Federal Public Service (FPS) Health, Food Chain Safety and Environment**. One repatriated Belgian has tested positive for the novel coronavirus. https://www.info-coronavirus.be/en/news/one-repatriated-belgian-has-tested-positive-for-the-novel-coronavirus/ (2020, accessed 24 August 2021).

406. **Sciensano**. COVID-19 epidemiological update. https://covid-19.sciensano.be/fr/covid-19-situation-epidemiologique (2020, accessed 24 August 2021).

407. **Sciensano**. *COVID-19 Surveillance: Frequently Asked Questions*. https://covid-19.sciensano.be/sites/default/files/Covid19/COVID-19_FAQ_ENG_final.pdf (2022, accessed 2 August 2022).

408. **Blasimme A, Ferretti A, Vayena E**. Digital Contact Tracing Against COVID-19 in Europe: Current Features and Ongoing Developments. *Frontiers in Digital Health* 2021;3:61.

409. **Izquierdo-Lara R, Elsinga G, Heijnen L, Munnink BBO, Schapendonk CME, *et al.*** Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater Sequencing, the Netherlands and Belgium. *Emerg Infect Dis* 2021;27:1405–1415.

410. **Pan Y, Zhang D, Yang P, Poon LLM, Wang Q**. Viral load of SARS-CoV-2 in clinical samples. *The Lancet Infectious Diseases* 2020;20:411–412.

411. **Chen C, Gao G, Xu Y, Pu L, Wang Q, *et al.*** SARS-CoV-2–Positive Sputum and Feces After Conversion of Pharyngeal Samples in Patients With COVID-19. *Annals of Internal Medicine* 2020;172:832–834.

412. **Yeo C, Kaushal S, Yeo D**. Enteric involvement of coronaviruses: is faecal-oral transmission of SARS-CoV-2 possible? *The lancet Gastroenterology & hepatology* 2020;5:335–337.

413. **Ahmed W, Angel N, Edson J, Bibby K, Bivins A, *et al.*** First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. *Science of The Total Environment* 2020;728:138764.

414. **Medema G, Heijnen L, Elsinga G, Italiaander R, Brouwer A**. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. *Environmental Science & Technology Letters* 2020;7:511–516.

415. **Wu F, Zhang J, Xiao A, Gu X, Lee WL, *et al.*** SARS-CoV-2 Titers in Wastewater Are Higher than Expected from Clinically Confirmed Cases. *mSystems*;5. Epub ahead of print 21 July 2020. DOI: 10.1128/mSystems.00614-20.

416. **Sinclair RG, Choi CY, Riley MR, Gerba CP**. Pathogen Surveillance Through Monitoring of Sewer Systems. 2008. pp. 249–269.

417. **Xagoraraki I, O'Brien E**. Wastewater-Based Epidemiology for Early Detection of Viral Outbreaks. In: O'Bannon DJ (editor). Cham: Springer International Publishing; 2020. pp. 75–97.

418. **Qi R, Huang Y, Liu J, Sun Y, Sun X, *et al.*** Global Prevalence of Asymptomatic Norovirus Infection: A Meta-analysis. *EClinicalMedicine* 2018;2–3:50–58.

419. **Nemudryi A, Nemudraia A, Wiegand T, Surya K, Buyukyoruk M, *et al.*** Temporal Detection and Phylogenetic Assessment of SARS-CoV-2 in Municipal Wastewater. *Cell Reports Medicine* 2020;1:100098.
420. **La Rosa G, Muscillo M**. Molecular detection of viruses in water and sewage. In: *Viruses in Food and Water*. Elsevier; 2013. pp. 97–125.
421. **Lurie N, Saville M, Hatchett R, Halton J**. Developing Covid-19 Vaccines at Pandemic Speed. *N Engl J Med* 2020;382:1969–1973.
422. **Shrotri M, Swinnen T, Kampmann B, Parker EPK**. An interactive website tracking COVID-19 vaccine development. *Lancet Glob Health* 2021;9:e590–e592.
423. **Thanh Le T, Andreadakis Z, Kumar A, Gómez Román R, Tollefsen S, *et al.*** The COVID-19 vaccine development landscape. *Nat Rev Drug Discov* 2020;19:305–306.
424. **Le TT, Cramer JP, Chen R, Mayhew S**. Evolution of the COVID-19 vaccine development landscape. *Nature Reviews Drug Discovery* 2020;19:667–668.
425. **Diamond MS, Pierson TC**. The Challenges of Vaccine Development against a New Virus during a Pandemic. *Cell Host Microbe* 2020;27:699–703.
426. **Chang L-J**. *Safety and Immunity Evaluation of A Covid-19 Coronavirus Artificial Antigen Presenting Cell Vaccine*. Clinical Trial Registration NCT04299724; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04299724 (6 March 2020, accessed 25 August 2021).
427. **Shenzhen Geno-Immune Medical Institute**. *Phase I/II Multicenter Trial of Lentiviral Minigene Vaccine (LV-SMENP) of Covid-19 Coronavirus*. Clinical Trial Registration NCT04276896; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04276896 (17 March 2020, accessed 25 August 2021).
428. **AnGes, Inc.** *A Non-randomized, Open-label, Non-controlled Phase I/II Study to Assess Safety and Immunogenicity of Two Doses of Intramuscular AG0301-COVID19 (1mg/2mg) in Healthy Adults*. Clinical Trial Registration NCT04463472; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04463472 (17 August 2021, accessed 25 August 2021).
429. **Genexine, Inc.** *A Phase 1/2a, Multi-center, Randomized, Double-blind, Placebo-controlled Study to Investigate the Safety, Tolerability, and Immunogenicity of GX-19, a COVID-19 Preventive DNA Vaccine in Healthy Subjects*. Clinical Trial Registration NCT04445389; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04445389 (24 July 2020, accessed 25 August 2021).
430. **Inovio Pharmaceuticals**. *Phase 1 Open-label Study to Evaluate the Safety, Tolerability and Immunogenicity of INO-4800, a Prophylactic Vaccine Against SARS-CoV-2, Administered Intradermally Followed by Electroporation in Healthy Volunteers*. Clinical Trial Registration NCT04336410; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04336410 (24 August 2021, accessed 25 August 2021).
431. **Reuters**. S.Korea's Genexine begins human trial of coronavirus vaccine. 19 June 2020. https://www.reuters.com/article/health-coronavirus-genexine-vaccine-idUSL4N2DW1T3 (19 June 2020, accessed 26 August 2021).
432. **Israel Institute for Biological Research (IIBR)**. *A Phase I/II Randomized, Multi-Center, Placebo-Controlled, Dose-Escalation Study to Evaluate the Safety, Immunogenicity and Potential Efficacy of an rVSV-SARS-CoV-2-S Vaccine (IIBR-100) in Adults*. Clinical Trial Registration NCT04608305; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04608305 (25 April 2021, accessed 25 August 2021).
433. **Kowalski PS, Rudra A, Miao L, Anderson DG**. Delivering the Messenger: Advances in Technologies for Therapeutic mRNA Delivery. *Mol Ther* 2019;27:710–728.
434. **Krammer F**. SARS-CoV-2 vaccines in development. *Nature* 2020;586:516–527.
435. **Park KS, Sun X, Aikins ME, Moon JJ**. Non-viral COVID-19 vaccine delivery systems. *Adv Drug Deliv Rev* 2021;169:137–151.
436. **Verbeke R, Lentacker I, De Smedt S, Dewitte H**. Three decades of messenger RNA vaccine development. *NANO TODAY*;28. Epub ahead of print 2019. DOI: 10.1016/j.nantod.2019.100766.
437. **Kyriakidis NC, López-Cortés A, González EV, Grimaldos AB, Prado EO**. SARS-CoV-2 vaccines strategies: a comprehensive review of phase 3 candidates. *npj Vaccines* 2021;6:1–17.
438. **Centers for Disease Control and Prevention**. Vaccine Recommendations and Guidelines of the ACIP. https://www.cdc.gov/vaccines/hcp/acip-recs/vacc-specific/covid-19.html (2021, accessed 26 August 2021).
439. **European Commission**. Safe COVID-19 vaccines for Europeans. *European Commission - European Commission*. https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans_en (2021, accessed 26 August 2021).
440. **Hogan MJ, Pardi N**. mRNA Vaccines in the COVID-19 Pandemic and Beyond. *Annual Review of Medicine* 2022;73:17–39.
441. **Ling Y, Zhong J, Luo J**. Safety and effectiveness of SARS-CoV-2 vaccines: A systematic review and meta-analysis. *J Med Virol* 2021;jmv.27203.
442. **Dolgin E**. CureVac COVID vaccine let-down spotlights mRNA design challenges. *Nature* 2021;594:483–483.

443. **Machhi J, Shahjin F, Das S, Patel M, Abdelmoaty MM, *et al.*** Nanocarrier vaccines for SARS-CoV-2. *Advanced Drug Delivery Reviews* 2021;171:215–239.

444. **University of Oxford**. *A Phase 2/3 Study to Determine the Efficacy, Safety and Immunogenicity of the Candidate Coronavirus Disease (COVID-19) Vaccine ChAdOx1 nCoV-19*. Clinical Trial Registration NCT04400838; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04400838 (18 June 2021, accessed 25 August 2021).

445. **Janssen Vaccines & Prevention B.V.** *A Randomized, Double-blind, Placebo-controlled Phase 1/2a Study to Evaluate the Safety, Reactogenicity, and Immunogenicity of Ad26COVS1 in Adults Aged 18 to 55 Years Inclusive and Adults Aged 65 Years and Older*. Clinical Trial Registration NCT04436276; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04436276 (7 July 2021, accessed 25 August 2021).

446. **Janssen Vaccines & Prevention B.V.** *A Randomized, Double-blind, Placebo-controlled Phase 3 Study to Assess the Efficacy and Safety of Ad26.COV2.S for the Prevention of SARS-CoV-2-mediated COVID-19 in Adults Aged 18 Years and Older*. Clinical Trial Registration NCT04505722; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04505722 (3 August 2021, accessed 25 August 2021).

447. **Nogrady B**. Mounting evidence suggests Sputnik COVID vaccine is safe and effective. *Nature* 2021;595:339–340.

448. **O'Reilly P**. A phase III study to investigate a vaccine against COVID-19. Epub ahead of print 2021. DOI: 10.1186/ISRCTN89951424.

449. **Watanabe Y, Mendonça L, Allen ER, Howe A, Lee M, *et al.*** Native-like SARS-CoV-2 spike glycoprotein expressed by ChAdOx1 nCoV-19/AZD1222 vaccine. *bioRxiv* 2021;2021.01.15.426463.

450. **Arashkia A, Jalilvand S, Mohajel N, Afchangi A, Azadmanesh K, *et al.*** Severe acute respiratory syndrome-coronavirus-2 spike (S) protein based vaccine candidates: State of the art and future prospects. *Rev Med Virol* 2020;e2183.

451. **Jones I, Roy P**. Sputnik V COVID-19 vaccine candidate appears safe and effective. *The Lancet* 2021;397:642–643.

452. **Petrovsky N, Aguilar JC**. Vaccine adjuvants: Current state and future trends. *Immunology & Cell Biology* 2004;82:488–496.

453. **Ella R, Reddy S, Blackwelder W, Potdar V, Yadav P, *et al.*** *Efficacy, safety, and lot to lot immunogenicity of an inactivated SARS-CoV-2 vaccine (BBV152): a, double-blind, randomised, controlled phase 3 trial*.

454. **Reuters**. Russia approves its third COVID-19 vaccine, CoviVac. *Reuters*, 20 February 2021. https://www.reuters.com/article/us-health-coronavirus-russia-vaccine-idUSKBN2AK07H (20 February 2021, accessed 26 August 2021).

455. **Butantan Institute**. *Double-Blind, Randomized, Placebo-Controlled Phase III Clinical Trial to Evaluate Efficacy and Safety in Healthcare Professionals of the Adsorbed COVID-19 (Inactivated) Vaccine Manufactured by Sinovac*. Clinical Trial Registration NCT04456595; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04456595 (10 February 2021, accessed 25 August 2021).

456. **Sinovac Research and Development Co., Ltd.** *A Randomized, Double-blind, Placebo-controlled Clinical Trial, to Evaluate Safety and Immunogenicity of Inactivated SARS-CoV-2 Vaccine (Vero Cell), in Healthy Elderly Aged 60 Years and Above*. Clinical Trial Registration NCT04383574; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04383574 (16 August 2021, accessed 25 August 2021).

457. **Al Kaabi N, Zhang Y, Xia S, Yang Y, Al Qahtani MM, *et al.*** Effect of 2 Inactivated SARS-CoV-2 Vaccines on Symptomatic COVID-19 Infection in Adults: A Randomized Clinical Trial. *JAMA* 2021;326:35–45.

458. **Xia S, Zhang Y, Wang Y, Wang H, Yang Y, *et al.*** Safety and immunogenicity of an inactivated SARS-CoV-2 vaccine, BBIBP-CorV: a randomised, double-blind, placebo-controlled, phase 1/2 trial. *The Lancet Infectious Diseases* 2021;21:39–51.

459. **Mallapaty S**. Iran hopes to defeat COVID with home-grown crop of vaccines. *Nature* 2021;596:475–475.

460. **Reuters**. Kazakhstan rolls out its own COVID-19 vaccine. *Reuters*, 27 April 2021. https://www.reuters.com/business/healthcare-pharmaceuticals/kazakhstan-rolls-out-its-own-covid-19-vaccine-2021-04-27/ (27 April 2021, accessed 26 August 2021).

461. **Schiller JT, Lowy DR**. Raising Expectations For Subunit Vaccine. *The Journal of Infectious Diseases* 2015;211:1373–1375.

462. **Novavax**. *A 2-Part, Phase 1/2, Randomized, Observer-Blinded Study To Evaluate The Safety And Immunogenicity Of A SARS-CoV-2 Recombinant Spike Protein Nanoparticle Vaccine (SARS-CoV-2 rS) With Or Without MATRIX-M^{TM} Adjuvant In Healthy Subjects*. Clinical Trial Registration NCT04368988; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04368988 (23 July 2021, accessed 25 August 2021).

463. **Federal Budgetary Research Institution State Research Center of Virology and Biotechnology 'Vector'**. *Simple, Blind, Placebo-controlled, Randomized Study of the Safety, Reactogenicity and Immunogenicity of Vaccine Based on Peptide Antigens for the Prevention of COVID-19 (EpiVacCorona), in Volunteers Aged 18-60 Years (I-II Phase)*. Clinical Trial Registration NCT04527575; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT04527575 (20 February 2021, accessed 25 August 2021).

464. **Yang S, Li Y, Dai L, Wang J, He P, *et al.*** Safety and immunogenicity of a recombinant tandem-repeat dimeric RBD-based protein subunit vaccine (ZF2001) against COVID-19 in adults: two randomised, double-blind, placebo-controlled, phase 1 and 2 trials. *The Lancet Infectious Diseases* 2021;21:1107–1119.
465. **Sanofi Pasteur, a Sanofi Company**. *A Parallel-group, Phase III, Multi-stage, Modified Double-blind, Multi-armed Study to Assess the Efficacy, Safety, and Immunogenicity of Two SARS-CoV-2 Adjuvanted Recombinant Protein Vaccines (Monovalent and Bivalent) for Prevention Against COVID-19 in Adults 18 Years of Age and Older*. Clinical Trial Registration study/NCT04904549; clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/study/NCT04904549 (27 July 2021, accessed 25 August 2021).
466. **European Medicines Agency**. COVID-19 vaccines. *European Medicines Agency*. https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/treatments-vaccines/covid-19-vaccines (2021, accessed 8 August 2022).
467. **Abdel-Moneim AS, Abdelwhab EM, Memish ZA**. Insights into SARS-CoV-2 evolution, potential antivirals, and vaccines. *Virology* 2021;558:1–12.
468. **Olliaro P, Torreele E, Vaillant M**. COVID-19 vaccine efficacy and effectiveness-the elephant (not) in the room. *Lancet Microbe* 2021;2:e279–e280.
469. **Tarighi P, Eftekhari S, Chizari M, Sabernavaei M, Jafari D, *et al.*** A review of potential suggested drugs for coronavirus disease (COVID-19) treatment. *Eur J Pharmacol* 2021;895:173890.
470. **Jiang Y, Yin W, Xu HE**. RNA-dependent RNA polymerase: Structure, mechanism, and drug discovery for COVID-19. *Biochem Biophys Res Commun* 2021;538:47–53.
471. **European Medicines Agency**. Regkirona. *Regkirona*. https://www.ema.europa.eu/en/medicines/human/EPAR/regkirona (2021, accessed 16 November 2021).
472. **European Medicines Agency**. Ronapreve. *Ronapreve*. https://www.ema.europa.eu/en/medicines/human/EPAR/ronapreve (2021, accessed 16 November 2021).
473. **European Medicines Agency**. Veklury. *Veklury*. https://www.ema.europa.eu/en/medicines/human/EPAR/veklury (2020, accessed 16 November 2021).
474. **European Medicines Agency**. Evusheld. *European Medicines Agency*. https://www.ema.europa.eu/en/medicines/human/EPAR/evusheld (2022, accessed 8 August 2022).
475. **European Medicines Agency**. Kineret. *European Medicines Agency*. https://www.ema.europa.eu/en/medicines/human/EPAR/kineret (2018, accessed 8 August 2022).
476. **European Medicines Agency**. Paxlovid. *European Medicines Agency*. https://www.ema.europa.eu/en/medicines/human/EPAR/paxlovid (2022, accessed 8 August 2022).
477. **European Medicines Agency**. RoActemra. *European Medicines Agency*. https://www.ema.europa.eu/en/medicines/human/EPAR/roactemra (2018, accessed 8 August 2022).
478. **European Medicines Agency**. Xevudy. *European Medicines Agency*. https://www.ema.europa.eu/en/medicines/human/EPAR/xevudy (2021, accessed 8 August 2022).
479. **Yang Y, Du L**. Neutralizing antibodies and their cocktails against SARS-CoV-2 Omicron and other circulating variants. *Cell Mol Immunol* 2022;19:962–964.
480. **European Medicines Agency**. COVID-19 treatments. *COVID-19 treatments*. https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/treatments-vaccines/covid-19-treatments (2021, accessed 16 November 2021).
481. **Carter LJ, Garner LV, Smoot JW, Li Y, Zhou Q, *et al.*** Assay Techniques and Test Development for COVID-19 Diagnosis. *ACS Cent Sci* 2020;6:591–605.
482. **Cheng MP, Papenburg J, Desjardins M, Kanjilal S, Quach C, *et al.*** Diagnostic Testing for Severe Acute Respiratory Syndrome–Related Coronavirus 2. *Annals of Internal Medicine* 2020;172:726–734.
483. **Vashist SK**. *In vitro* Diagnostic Assays for COVID-19: Recent Advances and Emerging Trends. *Diagnostics* 2020;10:202.
484. **Kalil AC, Thomas PG**. Influenza virus-related critical illness: pathophysiology and epidemiology. *Crit Care* 2019;23:258.
485. **Ponti G, Maccaferri M, Ruini C, Tomasi A, Ozben T**. Biomarkers associated with COVID-19 disease progression. *Critical Reviews in Clinical Laboratory Sciences* 2020;57:389–399.
486. **Wang D, Hu B, Hu C, Zhu F, Liu X, *et al.*** Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA* 2020;323:1061.
487. **Inui S, Fujikawa A, Jitsu M, Kunishima N, Watanabe S, *et al.*** Chest CT Findings in Cases from the Cruise Ship *Diamond Princess* with Coronavirus Disease (COVID-19). *Radiology: Cardiothoracic Imaging* 2020;2:e200110.
488. **Carlile M, Hurt B, Hsiao A, Hogarth M, Longhurst CA, *et al.*** Deployment of artificial intelligence for radiographic diagnosis of COVID-19 pneumonia in the emergency department. *Journal of the American College of Emergency Physicians Open* 2020;1:1459–1464.

489. **Wang S, Zha Y, Li W, Wu Q, Li X, *et al.*** A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. *Eur Respir J* 2020;56:2000775.

490. **Imran A, Posokhova I, Qureshi HN, Masood U, Riaz MS, *et al.*** AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Informatics in Medicine Unlocked* 2020;20:100378.

491. **Allam M, Cai S, Ganesh S, Venkatesan M, Doodhwala S, *et al.*** COVID-19 Diagnostics, Tools, and Prevention. *Diagnostics* 2020;10:409.

492. **Filipiak W, Ager C, Troppmair J**. Predicting the future from the past: volatile markers for respiratory infections. *Eur Respir J* 2017;49:1700264.

493. **Weissleder R, Lee H, Ko J, Pittet MJ**. COVID-19 diagnostics in context. *Sci Transl Med* 2020;12:eabc1931.

494. **Hou H, Wang T, Zhang B, Luo Y, Mao L, *et al.*** Detection of IgM and IgG antibodies in patients with coronavirus disease 2019. *Clinical & Translational Immunology* 2020;9:e1136.

495. **Abbasi J**. The Promise and Peril of Antibody Testing for COVID-19. *JAMA* 2020;323:1881.

496. **Lisboa Bastos M, Tavaziva G, Abidi SK, Campbell JR, Haraoui L-P, *et al.*** Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. *BMJ* 2020;m2516.

497. **Winter AK, Hegde ST**. The important role of serology for COVID-19 control. *The Lancet Infectious Diseases* 2020;20:758–759.

498. **Chen L, Xiong J, Bao L, Shi Y**. Convalescent plasma as a potential therapy for COVID-19. *The Lancet Infectious Diseases* 2020;20:398–400.

499. **Ladner JT, Henson SN, Boyle AS, Engelbrektson AL, Fink ZW, *et al.*** Epitope-resolved profiling of the SARS-CoV-2 antibody response identifies cross-reactivity with endemic human coronaviruses. *Cell Reports Medicine* 2021;2:100189.

500. **Hicks J, Klumpp-Thomas C, Kalish H, Shunmugavel A, Mehalko J, *et al.*** Serologic Cross-Reactivity of SARS-CoV-2 with Endemic and Seasonal Betacoronaviruses. *J Clin Immunol* 2021;41:906–913.

501. **Li H, Mendelsohn E, Zong C, Zhang W, Hagan E, *et al.*** Human-animal interactions and bat coronavirus spillover potential among rural residents in Southern China. *Biosafety and Health* 2019;1:84–90.

502. **Sun H, Xiao Y, Liu J, Wang D, Li F, *et al.*** Prevalent Eurasian avian-like H1N1 swine influenza virus with 2009 pandemic viral genes facilitating human infection. *Proc Natl Acad Sci USA* 2020;117:17204–17210.

503. **Vemula SV, Zhao J, Liu J, Wang X, Biswas S, *et al.*** Current Approaches for Diagnosis of Influenza Virus Infections in Humans. *Viruses* 2016;8:96–96.

504. **Stephenson I, Heath A, Major D, Newman RW, Hoschler K, *et al.*** Reproducibility of serologic assays for influenza virus A (H5N1). *Emerg Infect Dis* 2009;15:1252–1259.

505. **Haaheim R**. Single-radial-complement-fixation: a new immunodiffusion technique. 2. Assay of the antibody response to the internal antigens (MP and NP) of influenza A virus in human sera after vaccination and infection. *Dev Biol Stand* 1977;39:481–484.

506. **Kontou PI, Braliou GG, Dimou NL, Nikolopoulos G, Bagos PG**. Antibody Tests in Detecting SARS-CoV-2 Infection: A Meta-Analysis. *Diagnostics* 2020;10:319.

507. **Jin Y, Wang M, Zuo Z, Fan C, Ye F, *et al.*** Diagnostic value and dynamic variance of serum antibody in coronavirus disease 2019. *International Journal of Infectious Diseases* 2020;94:49–52.

508. **Infantino M, Grossi V, Lari B, Bambi R, Perri A, *et al.*** Diagnostic accuracy of an automated chemiluminescent immunoassay for anti-SARS-CoV-2 IgM and IgG antibodies: an Italian experience. *J Med Virol* 2020;92:1671–1675.

509. **Li Z, Yi Y, Luo X, Xiong N, Liu Y, *et al.*** Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. *J Med Virol* 2020;92:1518–1524.

510. **Martinaud C, Hejl C, Igert A, Bigaillon C, Bonnet C, *et al.*** Evaluation of the Quotient® MosaiQ^{TM} COVID-19 antibody microarray for the detection of IgG and IgM antibodies to SARS-CoV-2 virus in humans. *Journal of Clinical Virology* 2020;130:104571.

511. **Jiang H, Li Y, Zhang H, Wang W, Yang X, *et al.*** SARS-CoV-2 proteome microarray for global profiling of COVID-19 specific IgG and IgM responses. *Nat Commun* 2020;11:3581.

512. **Shrock E, Fujimura E, Kula T, Timms RT, Lee I-H, *et al.*** Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science* 2020;370:eabd4250.

513. **Nag P, Sadani K, Mukherji S**. Optical Fiber Sensors for Rapid Screening of COVID-19. *Trans Indian Natl Acad Eng* 2020;5:233–236.

514. **Tripathi S, Agrawal A**. Blood Plasma Microfluidic Device: Aiming for the Detection of COVID-19 Antibodies Using an On-Chip ELISA Platform. *Trans Indian Natl Acad Eng* 2020;5:217–220.

515. **Habli Z, Saleh S, Zaraket H, Khraiche ML**. COVID-19 in-vitro Diagnostics: State-of-the-Art and Challenges for Rapid, Scalable, and High-Accuracy Screening. *Front Bioeng Biotechnol* 2021;8:605702.

516. **Poschenrieder A, Thaler M, Junker R, Luppa PB**. Recent advances in immunodiagnostics based on biosensor technologies—from central laboratory to the point of care. *Anal Bioanal Chem* 2019;411:7607–7621.

517. **Krammer F, Smith GJD, Fouchier RAM, Peiris M, Kedzierska K, *et al.*** Influenza. *Nat Rev Dis Primers* 2018;4:3.

518. **Lau SKP, Woo PCY, Wong BHL, Tsoi H-W, Woo GKS, *et al.*** Detection of Severe Acute Respiratory Syndrome (SARS) Coronavirus Nucleocapsid Protein in SARS Patients by Enzyme-Linked Immunosorbent Assay. *J Clin Microbiol* 2004;42:2884–2889.

519. **Chen Y, Chan K-H, Kang Y, Chen H, Luk HK, *et al.*** A sensitive and specific antigen detection assay for Middle East respiratory syndrome coronavirus. *Emerging Microbes & Infections* 2015;4:1–5.

520. **Smithgall MC, Dowlatshahi M, Spitalnik SL, Hod EA, Rai AJ**. Types of Assays for SARS-CoV-2 Testing: A Review. *Laboratory Medicine* 2020;51:e59–e65.

521. **Wu F, Zhao S, Yu B, Chen Y-M, Wang W, *et al.*** A new coronavirus associated with human respiratory disease in China. *Nature* 2020;579:265–269.

522. **Murugan D, Bhatia H, Sai VVR, Satija J**. P-FAB: A Fiber-Optic Biosensor Device for Rapid Detection of COVID-19. *Trans Indian Natl Acad Eng* 2020;5:211–215.

523. **Seo G, Lee G, Kim MJ, Baek S-H, Choi M, *et al.*** Rapid Detection of COVID-19 Causative Virus (SARS-CoV-2) in Human Nasopharyngeal Swab Specimens Using Field-Effect Transistor-Based Biosensor. *ACS Nano* 2020;14:5135–5142.

524. **Song Y, Song J, Wei X, Huang M, Sun M, *et al.*** Discovery of Aptamers Targeting the Receptor-Binding Domain of the SARS-CoV-2 Spike Glycoprotein. *Anal Chem* 2020;92:9895–9900.

525. **Bustin SA, Nolan T**. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J Biomol Tech* 2004;15:155–166.

526. **Premraj A, Aleyas AG, Nautiyal B, Rasool TJ**. Nucleic Acid and Immunological Diagnostics for SARS-CoV-2: Processes, Platforms and Pitfalls. *Diagnostics* 2020;10:866.

527. **Afzal A**. Molecular diagnostic technologies for COVID-19: Limitations and challenges. *Journal of Advanced Research* 2020;26:149–159.

528. **Artesi M, Bontems S, Göbbels P, Franckh M, Maes P, *et al.*** A Recurrent Mutation at Position 26340 of SARS-CoV-2 Is Associated with Failure of the E Gene Quantitative Reverse Transcription-PCR Utilized in a Commercial Dual-Target Diagnostic Assay. *J Clin Microbiol*;58. Epub ahead of print 22 September 2020. DOI: 10.1128/JCM.01598-20.

529. **Khan KA, Cheung P**. Presence of mismatches between diagnostic PCR assays and coronavirus SARS-CoV-2 genome. *R Soc open sci* 2020;7:200636.

530. **Wu Z, Harrich D, Li Z, Hu D, Li D**. The unique features of SARS-CoV-2 transmission: Comparison with SARS-CoV, MERS-CoV and 2009 H1N1 pandemic influenza virus. *Rev Med Virol*;31. Epub ahead of print March 2021. DOI: 10.1002/rmv.2171.

531. **Cheng H-Y, Jian S-W, Liu D-P, Ng T-C, Huang W-T, *et al.*** Contact Tracing Assessment of COVID-19 Transmission Dynamics in Taiwan and Risk at Different Exposure Periods Before and After Symptom Onset. *JAMA Intern Med* 2020;180:1156.

532. **Kucirka LM, Lauer SA, Laeyendecker O, Boon D, Lessler J**. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure. *Annals of Internal Medicine* 2020;173:262–267.

533. **Tang Y-W, Schmitz JE, Persing DH, Stratton CW**. Laboratory Diagnosis of COVID-19: Current Issues and Challenges. *J Clin Microbiol*;58. Epub ahead of print 26 May 2020. DOI: 10.1128/JCM.00512-20.

534. **Sheridan C**. Fast, portable tests come online to curb coronavirus pandemic. *Nat Biotechnol* 2020;38:515–518.

535. **Vasudevan HN, Xu P, Servellita V, Miller S, Liu L, *et al.*** Digital droplet PCR accurately quantifies SARS-CoV-2 viral load from crude lysate without nucleic acid purification. *Sci Rep* 2021;11:780.

536. **Pinheiro LB, Coleman VA, Hindson CM, Herrmann J, Hindson BJ, *et al.*** Evaluation of a Droplet Digital Polymerase Chain Reaction Format for DNA Copy Number Quantification. *Anal Chem* 2012;84:1003–1011.

537. **Campomenosi P, Gini E, Noonan DM, Poli A, D'Antona P, *et al.*** A comparison between quantitative PCR and droplet digital PCR technologies for circulating microRNA quantification in human lung cancer. *BMC Biotechnol* 2016;16:60.

538. **Quyen TL, Ngo TA, Bang DD, Madsen M, Wolff A**. Classification of Multiple DNA Dyes Based on Inhibition Effects on Real-Time Loop-Mediated Isothermal Amplification (LAMP): Prospect for Point of Care Setting. *Front Microbiol* 2019;10:2234.

539. **Abe T, Segawa Y, Watanabe H, Yotoriyama T, Kai S, *et al.*** Point-of-care testing system enabling 30 min detection of influenza genes. *Lab Chip* 2011;11:1166–1167.

540. **Notomi T**. Loop-mediated isothermal amplification of DNA. *Nucleic Acids Research* 2000;28:63e–663.

541. **Konwar AN, Borse V**. Current status of point-of-care diagnostic devices in the Indian healthcare system with an update on COVID-19 pandemic. *Sensors International* 2020;1:100015.

542. **Courtney SJ, Stromberg ZR, Kubicek-Sutherland JZ**. Nucleic Acid-Based Sensing Techniques for Diagnostics and Surveillance of Influenza. *Biosensors* 2021;11:47.

543. **Ali Z, Aman R, Mahas A, Rao GS, Tehseen M, *et al.*** iSCAN: An RT-LAMP-coupled CRISPR-Cas12 module for rapid, sensitive detection of SARS-CoV-2. *Virus Research* 2020;288:198129.

544. **Bhattacharyya RP, Thakku SG, Hung DT**. Harnessing CRISPR Effectors for Infectious Disease Diagnostics. *ACS Infect Dis* 2018;4:1278–1282.

545. **Hou T, Zeng W, Yang M, Chen W, Ren L, *et al.*** Development and evaluation of a rapid CRISPR-based diagnostic for COVID-19. *PLoS Pathog* 2020;16:e1008705.

546. **Ishino Y, Krupovic M, Forterre P**. History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *J Bacteriol*;200. Epub ahead of print April 2018. DOI: 10.1128/JB.00580-17.

547. **Xiang X, Qian K, Zhang Z, Lin F, Xie Y, *et al.*** CRISPR-cas systems based molecular diagnostic tool for infectious diseases and emerging 2019 novel coronavirus (COVID-19) pneumonia. *Journal of Drug Targeting* 2020;28:727–731.

548. **Sakurai A, Shibasaki F**. Updated Values for Molecular Diagnosis for Highly Pathogenic Avian Influenza Virus. *Viruses* 2012;4:1235–1257.

549. **Moore C, Hibbitts S, Owen N, Corden SA, Harrison G, *et al.*** Development and evaluation of a real-time nucleic acid sequence based amplification assay for rapid detection of influenza A. *J Med Virol* 2004;74:619–628.

550. **Wu L-T, Curran MD, Ellis JS, Parmar S, Ritchie AV, *et al.*** Nucleic Acid Dipstick Test for Molecular Diagnosis of Pandemic H1N1. *Journal of Clinical Microbiology* 2010;48:3608–3613.

551. **da Silva SJR, Silva CTA da, Guarines KM, Mendes RPG, Pardee K, *et al.*** Clinical and Laboratory Diagnosis of SARS-CoV-2, the Virus Causing COVID-19. *ACS Infect Dis* 2020;6:2319–2336.

552. **Watson JD, Crick FHC**. The Structure of Dna. *Cold Spring Harb Symp Quant Biol* 1953;18:123–131.

553. **Hutchison CA III**. DNA sequencing: bench to bedside and beyond †. *Nucleic Acids Research* 2007;35:6227–6237.

554. **Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, *et al.*** Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 1995;269:496–512.

555. **Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, *et al.*** The minimal gene complement of Mycoplasma genitalium. *Science* 1995;270:397–403.

556. **Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, *et al.*** The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res* 1981;9:2871–2888.

557. **Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, *et al.*** Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.

558. **Liu P, Fang X, Feng Z, Guo Y-M, Peng R-J, *et al.*** Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. *J Virol* 2011;85:11291–11299.

559. **Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, *et al.*** Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004;304:66–74.

560. **Mulcahy-O'Grady H, Workentine ML**. The Challenge and Potential of Metagenomics in the Clinic. *Front Immunol* 2016;7:29.

561. **Calvet G, Aguiar RS, Melo ASO, Sampaio SA, de Filippis I, *et al.*** Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study. *Lancet Infect Dis* 2016;16:653–660.

562. **Lei H, Li T, Li B, Tsai S, Biggar RJ, *et al.*** Epstein-Barr virus from Burkitt Lymphoma biopsies from Africa and South America share novel LMP-1 promoter and gene variations. *Sci Rep* 2015;5:16706.

563. **Thomson E, Ip CLC, Badhan A, Christiansen MT, Adamson W, *et al.*** Comparison of Next-Generation Sequencing Technologies for Comprehensive Assessment of Full-Length Hepatitis C Viral Genomes. *J Clin Microbiol* 2016;54:2470–2484.

564. **Cotten M, Petrova V, Phan MVT, Rabaa MA, Watson SJ, *et al.*** Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical setting. *J Virol* 2014;88:11056–11069.

565. **Brown JR, Roy S, Ruis C, Yara Romero E, Shah D, *et al.*** Norovirus Whole-Genome Sequencing by SureSelect Target Enrichment: a Robust and Sensitive Method. *J Clin Microbiol* 2016;54:2530–2537.

566. **Depledge DP, Kundu S, Jensen NJ, Gray ER, Jones M, *et al.*** Deep sequencing of viral genomes provides insight into the evolution and pathogenesis of varicella zoster virus and its vaccine in humans. *Mol Biol Evol* 2014;31:397–409.

567. **Tsangaras K, Wales N, Sicheritz-Pontén T, Rasmussen S, Michaux J, *et al.*** Hybridization capture using short PCR products enriches small genomes by capturing flanking sequences (CapFlank). *PLoS One* 2014;9:e109101.

568. **Wylie TN, Wylie KM, Herter BN, Storch GA**. Enhanced virome sequencing using targeted sequence capture. *Genome Res* 2015;25:1910–1920.

569. **Depledge DP, Palser AL, Watson SJ, Lai IY-C, Gray ER, *et al.*** Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS One* 2011;6:e27805.

570. **Sanger F, Nicklen S, Coulson AR**. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977;74:5463–5467.

571. **Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, *et al.*** Fluorescence detection in automated DNA sequence analysis. *Nature* 1986;321:674–679.

572. **Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, *et al.*** Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 1988;239:487–491.

573. **Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, *et al.*** Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 1985;230:1350–1354.

574. **Cohen SN, Chang AC, Boyer HW, Helling RB**. Construction of biologically functional bacterial plasmids *in vitro*. *Proc Natl Acad Sci U S A* 1973;70:3240–3244.

575. **Jackson DA, Symons RH, Berg P**. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proc Natl Acad Sci U S A* 1972;69:2904–2909.

576. **Chen C-Y**. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. *Front Microbiol* 2014;5:305.

577. **Adams JU**. DNA Sequencing Technologies. *Nature Education*. http://www.nature.com/scitable/topicpage/dna-sequencing-technologies-690 (2008, accessed 22 September 2021).

578. **Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, *et al.*** The sequence of the human genome. *Science* 2001;291:1304–1351.

579. **Bragstad K, Nielsen LP, Fomsgaard A**. The evolution of human influenza A viruses from 1999 to 2006: a complete genome study. *Virol J* 2008;5:40.

580. **Perales C, Chen Q, Soria ME, Gregori J, Garcia-Cehic D, *et al.*** Baseline hepatitis C virus resistance-associated substitutions present at frequencies lower than 15% may be clinically significant. *Infect Drug Resist* 2018;11:2207–2210.

581. **Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P**. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* 1996;242:84–89.

582. **Muir P, Li S, Lou S, Wang D, Spakowicz DJ, *et al.*** The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol* 2016;17:53.

583. **Gupta AK, Gupta UD**. Chapter 19 - Next Generation Sequencing and Its Applications. In: Verma AS, Singh A (editors). *Animal Biotechnology*. San Diego: Academic Press. pp. 345–367.

584. **Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, *et al.*** Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53–59.

585. **Heather JM, Chain B**. The sequence of sequencers: The history of sequencing DNA. *Genomics* 2016;107:1–8.

586. **Voelkerding KV, Dames SA, Durtschi JD**. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry* 2009;55:641–658.

587. **Ghedin E, Fitch A, Boyne A, Griesemer S, DePasse J, *et al.*** Mixed infection and the genesis of influenza virus diversity. *J Virol* 2009;83:8832–8841.

588. **Bleidorn C**. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity* 2016;14:1–8.

589. **Rhoads A, Au KF**. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* 2015;13:278–289.

590. **Schadt EE, Turner S, Kasarskis A**. A window into third-generation sequencing. *Human Molecular Genetics* 2010;19:227–240.

591. **Kumar KR, Cowley MJ, Davis RL**. Next-Generation Sequencing and Emerging Technologies. *Semin Thromb Hemost* 2019;45:661–673.

592. **Quail MA, Smith M, Coupland P, Otto TD, Harris SR, *et al.*** A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012;13:341.

593. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. https://www.science.org/lookup/doi/10.1126/science.1079700 (accessed 23 September 2021).

594. **Eid J, Fehr A, Gray J, Luong K, Lyle J, *et al.*** Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* 2009;323:133–138.

595. **Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, *et al.*** Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;37:1155–1162.

596. **Hu T, Chitnis N, Monos D, Dinh A**. Next-generation sequencing technologies: An overview. *Human Immunology*. Epub ahead of print 19 March 2021. DOI: 10.1016/j.humimm.2021.02.012.

597. **Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, *et al.*** Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotech* 2009;4:265–270.

598. **Lu H, Giordano F, Ning Z**. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics* 2016;14:265–279.

599. **Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, *et al.*** Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 2020;585:79–84.

600. **Jain M, Olsen HE, Paten B, Akeson M**. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* 2016;17:239.

601. **Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, *et al.*** Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* 2020;38:1044–1053.

602. **Edwards HS, Krishnakumar R, Sinha A, Bird SW, Patel KD, *et al.*** Real-Time Selective Sequencing with RUBRIC: Read Until with Basecall and Reference-Informed Criteria. *Sci Rep* 2019;9:11475.

603. **Beerenwinkel N, Zagordi O**. Ultra-deep sequencing for the analysis of viral populations. *Curr Opin Virol* 2011;1:413–418.

604. **Abayasingam A, Leung P, Eltahla A, Bull RA, Luciani F, *et al.*** Genomic characterization of hepatitis C virus transmitted founder variants with deep sequencing. *Infect Genet Evol* 2019;71:36–41.

605. **Fabbro CD, Scalabrin S, Morgante M, Giorgi FM**. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLOS ONE* 2013;8:e85024.

606. **Wright ES, Vetsigian KH**. Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics* 2016;17:876.

607. **Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J**. Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys. *PLOS ONE* 2014;9:e94249.

608. **Andrews S**. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010, accessed 24 September 2021).

609. **Broad Institute**. Picard Tools. https://broadinstitute.github.io/picard/ (2009, accessed 24 September 2021).

610. **Bolger AM, Lohse M, Usadel B**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.

611. **Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, *et al.*** SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience*;7. Epub ahead of print 1 January 2018. DOI: 10.1093/gigascience/gix120.

612. **Martin M**. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 2011;17:10–12.

613. **Blawid R, Silva J m. f., Nagata T**. Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline. *Annals of Applied Biology* 2017;170:301–314.

614. **Chikhi R, Medvedev P**. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 2014;30:31–37.

615. **ECDC**. *Sequencing of SARS-CoV-2: first update (18 January 2021)*. 2021.

616. **Lander ES, Waterman MS**. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 1988;2:231–239.

617. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al.*** SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.

618. **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, *et al.*** Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;29:644–652.

619. **Robertson G, Schein J, Chiu R, Corbett R, Field M, *et al.*** De novo assembly and analysis of RNA-seq data. *Nat Methods* 2010;7:909–912.

620. **Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, *et al.*** MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 2016;102:3–11.

621. **Meleshko D, Hajirasouliha I, Korobeynikov A**. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. *Bioinformatics* 2021;btab597.

622. **Liao Y-C, Cheng H-W, Wu H-C, Kuo S-C, Lauderdale T-LY, *et al.*** Completing Circular Bacterial Genomes With Assembly Complexity by Using a Sampling Strategy From a Single MinION Run With Barcoding. *Front Microbiol* 2019;10:2068.

623. **Ferragina P, Manzini G**. Indexing compressed text. *J ACM* 2005;52:552–581.

624. **Li H**. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:13033997 [q-bio]*. http://arxiv.org/abs/1303.3997 (2013, accessed 1 October 2021).
625. **Langmead B, Salzberg SL**. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
626. **Fonseca NA, Rung J, Brazma A, Marioni JC**. Tools for mapping high-throughput sequencing data. *Bioinformatics* 2012;28:3169–3177.
627. **Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, *et al.*** Twelve years of SAMtools and BCFtools. *GigaScience*;10. Epub ahead of print February 2021. DOI: 10.1093/gigascience/giab008.
628. **Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, *et al.*** Integrative genomics viewer. *Nat Biotechnol* 2011;29:24–26.
629. **Auwera GAV der, O'Connor BD**. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Inc.; 2020.
630. **Lischer HEL, Shimizu KK**. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics* 2017;18:474.
631. **Gwinn M, MacCannell DR, Khabbaz RF**. Integrating Advanced Molecular Technologies into Public Health. *Journal of Clinical Microbiology* 2017;55:703–714.
632. **Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, *et al.*** Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clin Microbiol Rev* 2017;30:1015–1063.
633. **Baele G, Dellicour S, Suchard MA, Lemey P, Vrancken B**. Recent advances in computational phylodynamics. *Current Opinion in Virology* 2018;31:24–32.
634. **Baele G, Suchard MA, Rambaut A, Lemey P**. Emerging Concepts of Data Integration in Pathogen Phylodynamics. *Systematic Biology* 2017;66:e47–e65.
635. **Mount DW**. *Bioinformatics: Sequence and Genome Analysis*. CSHL Press; 2004.
636. **Ahrenfeldt J, Skaarup C, Hasman H, Pedersen AG, Aarestrup FM, *et al.*** Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. *BMC Genomics* 2017;18:19.
637. **Posada D, Buckley TR**. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol* 2004;53:793–808.
638. **Bos DH, Posada D**. Using models of nucleotide evolution to build phylogenetic trees. *Dev Comp Immunol* 2005;29:211–227.
639. **Stamatakis A, Ludwig T, Meier H**. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 2005;21:456–463.
640. **Guindon S, Lethiec F, Duroux P, Gascuel O**. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* 2005;33:W557-559.
641. **Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ**. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.
642. **Ronquist F, Huelsenbeck JP**. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003;19:1572–1574.
643. **Drummond AJ, Rambaut A**. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007;7:214.
644. **Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, *et al.*** BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology* 2014;10:e1003537.
645. **Lartillot N, Lepage T, Blanquart S**. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 2009;25:2286–2288.
646. **Rambaut A**. FigTree. http://tree.bio.ed.ac.uk/software/figtree/ (2018, accessed 4 October 2021).
647. **Daniel H. H, Celine S**. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology*;61. Epub ahead of print 1 December 2012. DOI: 10.1093/sysbio/sys062.
648. **Yu G, Lam TT-Y, Zhu H, Guan Y**. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol* 2018;35:3041–3043.
649. **Subramanian B, Gao S, Lercher MJ, Hu S, Chen W-H**. Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res* 2019;47:W270–W275.
650. **Letunic I, Bork P**. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
651. **Sagulenko P, Puller V, Neher RA**. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*;4. Epub ahead of print 1 January 2018. DOI: 10.1093/ve/vex042.
652. **Salk JJ, Schmitt MW, Loeb LA**. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat Rev Genet* 2018;19:269–285.

653. **Huber M, Metzner KJ, Geissberger FD, Shah C, Leemann C, *et al.*** MinVar: A rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing. *J Virol Methods* 2017;240:7–13.

654. **McElroy K, Zagordi O, Bull R, Luciani F, Beerenwinkel N**. Accurate single nucleotide variant detection in viral populations by combining probabilistic clustering with a statistical test of strand bias. *BMC Genomics* 2013;14:501.

655. **Verbist BMP, Thys K, Reumers J, Wetzels Y, Van der Borght K, *et al.*** VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics* 2015;31:94–101.

656. **Zagordi O, Klein R, Däumer M, Beerenwinkel N**. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* 2010;38:7400–7409.

657. **Marinier E, Enns E, Tran C, Fogel M, Peters C, *et al.*** quasitools: A Collection of Tools for Viral Quasispecies Analysis. *bioRxiv [Preprint]*. Epub ahead of print 13 August 2019. DOI: 10.1101/733238.

658. **Posada-Cespedes S, Seifert D, Beerenwinkel N**. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Research* 2017;239:17–32.

659. **Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, *et al.*** LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 2012;40:11189–11201.

660. **Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, *et al.*** VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 2009;25:2283–2285.

661. **Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, *et al.*** Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun* 2012;3:811.

662. **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, *et al.*** The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.

663. **McCrone JT, Lauring AS, Lauring S**. Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling. *Journal of Virology* 2016;90:6884–6895.

664. **Acevedo A, Brodsky L, Andino R**. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* 2014;505:686–690.

665. **Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R**. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci U S A* 2011;108:20166–20171.

666. **Rosseel T, Van Borm S, Vandenbussche F, Hoffmann B, van den Berg T, *et al.*** The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing. *PLoS One* 2013;8:e76144.

667. **Kugelman JR, Kugelman-Tonos J, Ladner JT, Pettit J, Keeton CM, *et al.*** Emergence of Ebola Virus Escape Variants in Infected Nonhuman Primates Treated with the MB-003 Antibody Cocktail. *Cell Rep* 2015;12:2111–2120.

668. **Houldcroft CJ, Beale MA, Breuer J**. Clinical and biological insights from viral genome sequencing. *Nat Rev Microbiol* 2017;15:183–192.

669. **Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N**. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 2011;12:119.

670. **Prosperi MCF, Salemi M**. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics* 2012;28:132–133.

671. **Szpara ML, Parsons L, Enquist LW**. Sequence variability in clinical and laboratory isolates of herpes simplex virus 1 reveals new mutations. *J Virol* 2010;84:5303–5313.

672. **Victoria JG, Wang C, Jones MS, Jaing C, McLoughlin K, *et al.*** Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. *J Virol* 2010;84:6033–6040.

673. **Neverov A, Chumakov K**. Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines. *PNAS* 2010;107:20063–20068.

674. **Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, *et al.*** Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill* 2020;25:2000045.

675. **Chiara M, D'Erchia AM, Gissi C, Manzari C, Parisi A, *et al.*** Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. *Briefings in Bioinformatics* 2021;22:616–630.

676. **Baker M**. Next-generation sequencing: adjusting to data overload. *Nat Methods* 2010;7:495–499.

677. **Thézé J, Li T, Plessis L du, Bouquet J, Kraemer MUG, *et al.*** Genomic Epidemiology Reconstructs the Introduction and Spread of Zika Virus in Central America and Mexico. *Cell Host & Microbe* 2018;23:855-864.e7.

678. **Araujo THA, Souza-Brito LI, Libin P, Deforche K, Edwards D, *et al.*** A public HTLV-1 molecular epidemiology database for sequence management and data mining. *PLoS One* 2012;7:e42123.

679. **Gargis AS, Kalman L, Lubin IM**. Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. *J Clin Microbiol* 2016;54:2857–2865.
680. **Leung P, Eltahla AA, Lloyd AR, Bull RA, Luciani F**. Understanding the complex evolution of rapidly mutating viruses with deep sequencing: Beyond the analysis of viral diversity. *Virus Research* 2017;239:43–54.
681. **Datta S, Budhauliya R, Das B, Chatterjee S, Vanlalhmuaka, *et al.*** Next-generation sequencing in clinical virology: Discovery of new viruses. *World J Virol* 2015;4:265–276.
682. **Chan M, Leung A, Hisanaga T, Pickering B, Griffin BD, *et al.*** H7N9 Influenza Virus Containing a Polybasic HA Cleavage Site Requires Minimal Host Adaptation to Obtain a Highly Pathogenic Disease Phenotype in Mice. *Viruses* 2020;12:65.
683. **Liu WJ, Li J, Zou R, Pan J, Jin T, *et al.*** Dynamic PB2-E627K substitution of influenza H7N9 virus indicates the *in vivo* genetic tuning and rapid host adaptation. *Proc Natl Acad Sci U S A* 2020;117:23807–23814.
684. **Domingo E**. Quasispecies: from molecular Darwinism to viral diseases. *Contrib Sci* 2009;5:161–168.
685. **Voeten JT, Bestebroer TM, Nieuwkoop NJ, Fouchier R a, Osterhaus a D, *et al.*** Antigenic drift in the influenza A virus (H3N2) nucleoprotein and escape from recognition by cytotoxic T lymphocytes. *Journal of Virology* 2000;74:6800–7.
686. **McKimm-Breschkin JL, Sahasrabudhe A, Blick TJ, McDonald M, Colman PM, *et al.*** Mutations in a conserved residue in the influenza virus neuraminidase active site decreases sensitivity to Neu5Ac2en-derived inhibitors. *JVirol* 1998;72:2456–2462.
687. **Demicheli V, Jefferson T, Pietrantonj C, Ferroni E, Thorning RE, *et al.*** Vaccines for preventing influenza in the elderly. *Cochrane Database of Systematic Reviews*. Epub ahead of print 2018. DOI: 10.1002/14651858.cd004876.pub4.
688. **Brown EG, Liu H, Kit LC, Baird S, Nesrallah M**. Pattern of mutation in the genome of influenza A virus on adaptation to increased virulence in the mouse lung: Identification of functional themes. *Proceedings of the National Academy of Sciences* 2002;98:6883–6888.
689. **Woo HJ, Reifman J**. Quantitative Modeling of Virus Evolutionary Dynamics and Adaptation in Serial Passages Using Empirically Inferred Fitness Landscapes. *Journal of Virology* 2014;88:1039–1050.
690. **Choi WY, Shin JY, Jeong HE, Jeong MJ, Kim SJ, *et al.*** Generation and Characterization of Recombinant Influenza A(H1N1) Viruses Resistant to Neuraminidase Inhibitors. *Osong Public Health and Research Perspectives* 2013;4:323–328.
691. **Okomo-adhiambo M, Sheu TG, Gubareva L V.** Assays for monitoring susceptibility of influenza viruses to neuraminidase inhibitors. *Influenza and other Respiratory Viruses* 2013;7:44–49.
692. **Victoria X, Blades N, Ding J, Sultana R, Parmigiani G**. Estimation of sequencing error rates in short reads. *BMC Bioinformatics* 2012;13:185.
693. **Goldfeder RL, Wall DP, Khoury MJ, Ioannidis JPA, Ashley EA**. Human Genome Sequencing at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis. *American Journal of Epidemiology* 2017;186:1000–1009.
694. **Illumina**. Targeted next-generation sequencing versus qPCR and Sanger sequencing. https://www.illumina.com/content/dam/illumina-marketing/documents/products/other/infographic-targeted-ngs-vs-sanger-qpcr.pdf (2019, accessed 1 September 2019).
695. **World Health Organization**. *Whole genome sequencing for foodborne disease surveillance*. 9789241513869; 2018.
696. **Wang R, Taubenberger JK**. Methods for Molecular Surveillance of Influenza. *Expert review of antiinfective therapy* 2010;8:517–527.
697. **Nguyen HT, Fry AM, Gubareva L V.** Neuraminidase inhibitor resistance in influenza viruses and laboratory testing methods. *Antiviral Therapy* 2012;17:159–173.
698. **World Health Organization**. *Monitoring drug resistance in influenza viruses*. 2010.
699. **Pandey RV, Pabinger S, Kriegner A, Weinhäusel A**. ClinQC: A tool for quality control and cleaning of Sanger and NGS data in clinical research. *BMC Bioinformatics*;17. Epub ahead of print 2016. DOI: 10.1186/s12859-016-0915-y.
700. **Slatko BE, Gardner AF, Ausubel FM**. Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology* 2018;122:e59.
701. **Hutchinson EC**. Influenza Virus. *Trends in Microbiology* 2018;26:809–810.
702. **Patel N, Ferns BR, Nastouli E, Kozlakidis Z, Kellam P, *et al.*** Cost analysis of standard Sanger sequencing versus next generation sequencing in the ICONIC study. *The Lancet* 2016;388:S86.
703. **Arsenic R, Treue D, Lehmann A, Hummel M, Dietel M, *et al.*** Comparison of targeted next-generation sequencing and Sanger sequencing for the detection of PIK3CA mutations in breast cancer. *BMC Clinical Pathology* 2015;15:1–9.

704. **Tsiatis AC, Norris-Kirby A, Rich RG, Hafez MJ, Gocke CD, *et al.*** Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: Diagnostic and clinical implications. *Journal of Molecular Diagnostics* 2010;12:425–432.
705. **Altimari A, De Biase D, De Maglio G, Gruppioni E, Capizzi E, *et al.*** 454 next generation-sequencing outperforms allele-specific PCR, sanger sequencing, and pyrosequencing for routine KRAS mutation analysis of formalin-fixed, paraffin-embedded samples. *OncoTargets and Therapy* 2013;6:1057–1064.
706. **Vernikos G, Medini D, Riley DR, Tettelin H**. Ten years of pan-genome analyses. *Current Opinion in Microbiology* 2015;23:148–154.
707. **Bright RA, Medina MJ, Xu X, Perez-Oronoz G, Wallis TR, *et al.*** Incidence of adamantane resistance among influenza A (H3N2) viruses isolated worldwide from 1994 to 2005: A cause for concern. *Lancet* 2005;366:1175–1181.
708. **Kchouk M, Gibrat JF, Elloumi M**. Generations of Sequencing Technologies: From First to Next Generation. *Biology and Medicine*;09. Epub ahead of print 2017. DOI: 10.4172/0974-8369.1000395.
709. **Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J**. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian Journal of Microbiology* 2016;56:394–404.
710. **Van den Hoecke S, Verhelst J, Vuylsteke M, Saelens X**. Analysis of the genetic diversity of influenza A viruses using next-generation DNA sequencing. *BMC Genomics* 2015;16:1–23.
711. **Goodwin S, McPherson JD, McCombie WR**. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 2016;17:333–351.
712. **Braslavsky I, Hebert B, Kartalov E, Quake SR**. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences* 2003;100:3960–3964.
713. **Dark MJ**. Whole-genome sequencing in bacteriology: state of the art. *Infection and Drug Resistance* 2013;6:115–123.
714. **Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB**. The real cost of sequencing: Higher than you think! *Genome Biology*;12. Epub ahead of print 2011. DOI: 10.1186/gb-2011-12-8-125.
715. **Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, *et al.*** Application of next generation sequencing in clinical microbiology and infection prevention. *Journal of Biotechnology* 2017;243:16–24.
716. **Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, *et al.*** Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology* 2015;33:623–630.
717. **Fuentes-Pardo AP, Ruzzante DE**. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology* 2017;26:5369–5406.
718. **van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C**. Ten years of next-generation sequencing technology. *Trends in genetics : TIG* 2014;30:418–26.
719. **Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, *et al.*** Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Human Cell* 2017;30:149–161.
720. **Poon LLM, Song T, Rosenfeld R, Lin X, Rogers MB, *et al.*** Quantifying influenza virus diversity and transmission in humans. *Nature Genetics* 2016;48:195–200.
721. **Artyomenko A, Wu NC, Mangul S, Eskin E, Sun R, *et al.*** Long single-molecule reads can resolve the complexity of the influenza virus composed of rare, closely related mutant variants. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2016;9649:164–175.
722. **Wang J, Moore NE, Deng YM, Eccles DA, Hall RJ**. MinION nanopore sequencing of an influenza genome. *Frontiers in Microbiology* 2015;6:1–7.
723. **Keller MW, Rambo-Martin BL, Wilson MM, Ridenour CA, Shepard SS, *et al.*** Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. *Scientific reports* 2018;8:14408.
724. **Cauldwell A V., Long JS, Moncorge O, Barclay WS, Zhao Z, *et al.*** Segregation of virulent influenza A(H1N1) variants in the lower respiratory tract of critically ill patients during the 2010-2011 seasonal epidemic. *PLoS ONE* 2014;88:1–5.
725. **Xu Y, Lewandowski K, Lumley S, Pullan S, Vipond R, *et al.*** Detection of viral pathogens with multiplex nanopore MinION sequencing: Be careful with cross-Talk. *Frontiers in Microbiology* 2018;9:1–7.
726. **Eckert SE, Chan JZ-M, Houniet D, Breuer J, Speight G**. Enrichment by hybridisation of long DNA fragments for Nanopore sequencing. *Microbial Genomics*;2. Epub ahead of print 2016. DOI: 10.1099/mgen.0.000087.
727. **Fischer N, Indenbirken D, Meyer T, Lütgehetmann M, Lellek H, *et al.*** Evaluation of unbiased next-generation sequencing of RNA (RNA-seq) as a diagnostic method in influenza virus-positive respiratory samples. *Journal of Clinical Microbiology* 2015;53:2238–2250.
728. **Ali R, Blackburn RM, Kozlakidis Z**. Next-Generation Sequencing and Influenza Virus: A Short Review of the Published Implementation Attempts. *HAYATI Journal of Biosciences* 2016;23:155–159.

729. **Ghedin E, Sengamalay NA, Shumway M, Zaborsky J, Feldblyum T, *et al.*** Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* 2005;437:1162–1166.

730. **Head MG, Fitchett JR, Nageshwaran V, Kumari N, Hayward A, *et al.*** Research Investments in Global Health: A Systematic Analysis of UK Infectious Disease Research Funding and Global Health Metrics, 1997-2013. *EBioMedicine* 2016;3:180–190.

731. **McGinnis J, Laplante J, Shudt M, George KS**. Next generation sequencing for whole genome analysis and surveillance of influenza A viruses. *Journal of Clinical Virology* 2016;79:44–50.

732. **Goldhill DH, Aartjan JW, Fletcher RA, Langat P, Zambon M, *et al.*** The mechanism of resistance to favipiravir in influenza. 2018;115:11613–11618.

733. **Samson M, Abed Y, Desrochers FM, Hamilton S, Luttick A, *et al.*** Characterization of drug-resistant influenza virus A(H1N1) and A(H3N2) variants selected *in vitro* with laninamivir. *Antimicrobial Agents and Chemotherapy* 2014;58:5220–5228.

734. **Capobianchi MR, Giombini E, Rozera G**. Next-generation sequencing technology in clinical virology. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 2013;19:15–22.

735. **Trebbien R, Pedersen SS, Vorborg K, Franck KT, Fischer TK**. Development of oseltamivir and zanamivir resistance in influenza a(H1N1)pdm09 virus, Denmark, 2014. *Eurosurveillance* 2017;22:1–8.

736. **Pichon M, Picard C, Simon B, Gaymard A, Renard C, *et al.*** Clinical management and viral genomic diversity analysis of a child's influenza A(H1N1)pdm09 infection in the context of a severe combined immunodeficiency. *Antiviral Research* 2018;160:1–9.

737. **Bogaerts B, Winand R, Fu Q, Van Braekel J, Ceyssens P-J, *et al.*** Validation of a Bioinformatics Workflow for Routine Analysis of Whole-Genome Sequencing Data and Related Challenges for Pathogen Typing in a European National Reference Center: Neisseria meningitidis as a Proof-of-Concept. *Frontiers in Microbiology*;10. Epub ahead of print March 2019. DOI: 10.3389/fmicb.2019.00362.

738. **Chowell G, Bertozzi SM, Colchero MA, Lopez-Gatell H, Alpuche-Aranda C, *et al.*** Severe Respiratory Disease Concurrent with the Circulation of H1N1 Influenza. *New England Journal of Medicine* 2009;361:674–679.

739. **Webster RG, Bean WJ, Gorman OT, Chambers TM**. Evolution and Ecology of Influenza A Viruses. *Microbiological Reviews* 1992;56:152–179.

740. **Mosnier A, Caini S, Daviaud I, Nauleau E, Bui TT, *et al.*** Clinical Characteristics Are Similar across Type A and B Influenza Virus Infections. *PLoS ONE* 2015;10:1–13.

741. **Kosik I, Yewdell JW**. Influenza Hemagglutinin and Neuraminidase: Yin⁻Yang Proteins Coevolving to Thwart Immunity. *Viruses*;11. Epub ahead of print 2019. DOI: 10.3390/v11040346.

742. **World Health Organization Regional Office for Europe**. *Influenza surveillance country profiles*. http://www.euro.who.int/en/health-topics/communicable-diseases/influenza/surveillance-and-lab-network/influenza-surveillance-country-profiles (2019).

743. **European Centre for Disease Prevention and Control**. *Influenza virus characterisation guidelines for the northern hemisphere influenza season 2018-2019 (technical note)*. 2018.

744. **European Centre for Disease Prevention and Control**. *Influenza virus characterisation, summary Europe, May 2018*. Stockholm; 2014.

745. **Baillie GJ, Galiano M, Agapow P-M, Myers R, Chiam R, *et al.*** Evolutionary dynamics of local pandemic H1N1/2009 influenza virus lineages revealed by whole-genome analysis. *Journal of virology* 2012;86:11–8.

746. **Watson SJ, Langat P, Reid SM, Lam TT-Y, Cotten M, *et al.*** Molecular Epidemiology and Evolution of Influenza Viruses Circulating within European Swine between 2009 and 2013. *Journal of virology* 2015;89:9920–31.

747. **Langat P, Raghwani J, Dudas G, Bowden TA, Edwards S, *et al.*** Genome-wide evolutionary dynamics of influenza B viruses on a global scale. *PLoS pathogens* 2017;13:e1006749.

748. **Berry IM, Melendrez MC, Li T, Hawksworth AW, Brice GT, *et al.*** Frequency of influenza H3N2 intra-subtype reassortment: attributes and implications of reassortant spread. *BMC Biology* 2016;14:117.

749. **Van Poelvoorde LAE, Saelens X, Thomas I, Roosens NH**. Next-Generation Sequencing: An Eye-Opener for the Surveillance of Antiviral Resistance in Influenza. *Trends in Biotechnology*. Epub ahead of print December 2019. DOI: 10.1016/j.tibtech.2019.09.009.

750. **Belanov SS, Bychkov D, Benner C, Ripatti S, Ojala T, *et al.*** Genome-Wide Analysis of Evolutionary Markers of Human Influenza A(H1N1)pdm09 and A(H3N2) Viruses May Guide Selection of Vaccine Strain Candidates. *Genome Biology and Evolution* 2015;7:3472–3483.

751. **Nachbagauer R, Palese P**. Is a Universal Influenza Virus Vaccine Possible? *Annual Review of Medicine* 2020;71:315–327.

752. **Shu Y, McCauley J**. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* 2017;22:30494.

753. **Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, *et al.*** Integrating influenza antigenic dynamics with molecular evolution. *eLife*;3. Epub ahead of print 4 February 2014. DOI: 10.7554/eLife.01914.

754. **Simon B, Pichon M, Valette M, Burfin G, Richard M, *et al.*** Whole Genome Sequencing of A(H3N2) Influenza Viruses Reveals Variants Associated with Severity during the 2016–2017 Season. *Viruses* 2019;11:108.

755. **Blackburne BP, Hay AJ, Goldstein RA**. Changing selective pressure during antigenic changes in human influenza H3. *PLoS Pathogens*;4. Epub ahead of print 2008. DOI: 10.1371/journal.ppat.1000058.

756. **Bouvier NM, Palese P**. The biology of influenza viruses. *Vaccine* 2008;26 Suppl 4:D49-53.

757. **Allen JD, Ross TM**. H3N2 influenza viruses in humans: Viral mechanisms, evolution, and evaluation. *Human Vaccines & Immunotherapeutics* 2018;14:1840–1847.

758. **Wiman Å, Enkirch T, Carnahan A, Böttiger B, Hagey TS, *et al.*** Novel influenza A(H1N2) seasonal reassortant identified in a patient sample, Sweden, January 2019. *Eurosurveillance*;24. Epub ahead of print 28 February 2019. DOI: 10.2807/1560-7917.ES.2019.24.9.1900124.

759. **White MC, Lowen AC**. Implications of segment mismatch for influenza A virus evolution. *Journal of General Virology* 2018;99:3–16.

760. **Westgeest KB, Russell CA, Lin X, Spronken MIJ, Bestebroer TM, *et al.*** Genomewide Analysis of Reassortment and Evolution of Human Influenza A(H3N2) Viruses Circulating between 1968 and 2011. *Journal of Virology* 2014;88:2844–2857.

761. **Neverov AD, Lezhnina K V., Kondrashov AS, Bazykin GA**. Intrasubtype Reassortments Cause Adaptive Amino Acid Replacements in H3N2 Influenza Genes. *PLoS Genetics* 2014;10:e1004037.

762. **Rossen JWA, Friedrich AW, Moran-Gilad J**. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clinical Microbiology and Infection* 2018;24:355–360.

763. **Adlhoch C, Snacken R, Melidou A, Ionescu S, Penttinen P**. Dominant influenza A(H3N2) and B/Yamagata virus circulation in EU/EEA, 2016/17 and 2017/18 seasons, respectively. *Eurosurveillance*;23. Epub ahead of print 29 March 2018. DOI: 10.2807/1560-7917.ES.2018.23.13.18-00146.

764. **Thomas I, Barbezange C, Hombrouck A, Gucht S Van, Weyckmans J, *et al.*** Virological Surveillance of Influenza in Belgium; Season 2016-2017. *Sciensano Influenza Report* 2017;1–33.

765. **Leinonen R, Sugawara H, Shumway M**. The Sequence Read Archive. *Nucleic Acids Research* 2011;39:D19–D21.

766. **Brister JR, Ako-adjei D, Bao Y, Blinkova O**. NCBI Viral Genomes Resource. *Nucleic Acids Research* 2015;43:D571–D577.

767. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.*** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.

768. **Madeira F, Park Y mi, Lee J, Buso N, Gur T, *et al.*** The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research* 2019;47:W636–W641.

769. **European Centre for Disease Prevention and Control**. *Influenza Virus Characterization Reports*. https://www.ecdc.europa.eu/en/seasonal-influenza/surveillance-and-disease-data/influenza-virus-characterisation.

770. **Kumar S, Stecher G, Tamura K**. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* 2016;33:1870–1874.

771. **Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, *et al.*** Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;23:2947–2948.

772. **Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, *et al.*** Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*;4. Epub ahead of print 1 January 2018. DOI: 10.1093/ve/vey016.

773. **Nagarajan N, Kingsford C**. GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic Acids Research* 2011;39:e34–e34.

774. **Benjamini Y, Hochberg Y**. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995;57:289–300.

775. **Huelsenbeck JP, Rannala B**. Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. *Systematic Biology* 2004;53:904–913.

776. **Nascimento FF, Reis M dos, Yang Z**. A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution* 2017;1:1446–1454.

777. **Adlhoch C, Pebody R**. What to expect for the influenza season 2020/21 with the ongoing COVID-19 pandemic in the World Health Organization European Region. *Eurosurveillance*;25. Epub ahead of print 22 October 2020. DOI: 10.2807/1560-7917.ES.2020.25.42.2001816.

778. **Prakash N, Devangi P, Madhuuri K, Khushbu P, Deepali P**. Phylogenetic Analysis of H1N1 Swine Flu Virus Isolated In India. *Journal of Antivirals & Antiretrovirals*;03. Epub ahead of print 2011. DOI: 10.4172/jaa.1000028.

779. **Tewawong N, Suwannakarn K, Prachayangprecha S, Korkong S, Vichiwattana P, *et al.*** Molecular Epidemiology and Phylogenetic Analyses of Influenza B Virus in Thailand during 2010 to 2014. *PLOS ONE* 2015;10:e0116302.
780. **Goldstein EJ, Harvey WT, Wilkie GS, Shepherd SJ, MacLean AR, *et al.*** Integrating patient and whole-genome sequencing data to provide insights into the epidemiology of seasonal influenza A(H3N2) viruses. *Microbial Genomics*;4. Epub ahead of print 1 January 2018. DOI: 10.1099/mgen.0.000137.
781. **Kim J Il, Lee I, Park S, Bae J-Y, Yoo K, *et al.*** Reassortment compatibility between PB1, PB2, and HA genes of the two influenza B virus lineages in mammalian cells. *Scientific Reports* 2016;6:27480.
782. **Drummond AJ, Suchard MA, Xie D, Rambaut A**. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 2012;29:1969–1973.
783. **Liu D, Shi W, Shi Y, Wang D, Xiao H, *et al.*** Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. *The Lancet* 2013;381:1926–1932.
784. **Hedge J, Lycett SJ, Rambaut A**. Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biology Letters* 2013;9:20130331.
785. **Nelson MI, Simonsen L, Viboud C, Miller MA, Taylor J, *et al.*** Stochastic Processes Are Key Determinants of Short-Term Evolution in Influenza A Virus. *PLoS Pathogens* 2006;2:e125.
786. **Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, *et al.*** The genomic and epidemiological dynamics of human influenza A virus. *Nature* 2008;453:615–619.
787. **Rabadan R, Levine AJ, Krasnitz M**. Non-random reassortment in human influenza A viruses. *Influenza and Other Respiratory Viruses* 2008;2:9–22.
788. **Vaughan TG, Welch D, Drummond AJ, Biggs PJ, George T, *et al.*** Inferring Ancestral Recombination Graphs from Bacterial Genomic Data. *Genetics* 2017;205:857–870.
789. **Nelson MI, Viboud C, Simonsen L, Bennett RT, Griesemer SB, *et al.*** Multiple Reassortment Events in the Evolutionary History of H1N1 Influenza A Virus Since 1918. *PLoS Pathogens* 2008;4:e1000012.
790. **Gatherer D**. The 2009 H1N1 influenza outbreak in its historical context. *J Clin Virol* 2009;45:174–178.
791. **Zhou B, Donnelly ME, Scholes DT, St. George K, Hatta M, *et al.*** Single-Reaction Genomic Amplification Accelerates Sequencing and Vaccine Production for Classical and Swine Origin Human Influenza A Viruses. *Journal of Virology* 2009;83:10309–10313.
792. **Van Poelvoorde LAE, Bogaerts B, Fu Q, De Keersmaecker SCJ, Thomas I, *et al.*** Whole-genome-based phylogenomic analysis of the Belgian 2016–2017 influenza A(H3N2) outbreak season allows improved surveillance. *Microbial Genomics*;7. Epub ahead of print 3 September 2021. DOI: 10.1099/mgen.0.000643.
793. **Van Goethem N, Robert A, Bossuyt N, Van Poelvoorde LAE, Quoilin S, *et al.*** Evaluation of the added value of viral genomic information for predicting severity of influenza infection. *BMC Infect Dis* 2021;21:785.
794. **Wedde M, Wählisch S, Wolff T, Schweiger B**. Predominance of HA-222D/G Polymorphism in Influenza A(H1N1)pdm09 Viruses Associated with Fatal and Severe Outcomes Recently Circulating in Germany. *PLoS ONE* 2013;8:e57059.
795. **Abed Y, Baz M, Boivin G**. Impact of neuraminidase mutations conferring influenza resistance to neuraminidase inhibitors in the N1 and N2 genetic backgrounds. *Antiviral therapy* 2006;11:971–6.
796. **Chandler CH, Chari S, Dworkin I**. Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. *Trends in Genetics* 2013;29:358–366.
797. **Lai S, Qin Y, Cowling BJ, Ren X, Wardrop NA, *et al.*** Global epidemiology of avian influenza A H5N1 virus infection in humans, 1997–2015: a systematic review of individual case data. *The Lancet Infectious Diseases* 2016;16:e108–e118.
798. **Hung IFN, To KKW, Lee C, Lin C, Chan JFW, *et al.*** Effect of Clinical and Virological Parameters on the Level of Neutralizing Antibody against Pandemic Influenza A Virus H1N1 2009. *Clinical Infectious Diseases* 2010;51:274–279.
799. **Brittain-Long R, Nord S, Olofsson S, Westin J, Anderson L-M, *et al.*** Multiplex real-time PCR for detection of respiratory tract infections. *Journal of Clinical Virology* 2008;41:53–56.
800. **Hombrouck A, Sabbe M, Casteren V, Wuillaume F, Hue D, *et al.*** Viral aetiology of influenza-like illness in Belgium during the influenza A(H1N1)2009 pandemic. *European Journal of Clinical Microbiology & Infectious Diseases* 2012;31:999–1007.
801. **Stothard P**. The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences. *BioTechniques* 2000;28:1102–1104.
802. **European Centre for Disease Prevention and Control**. *Influenza in Europe, summary of the season 2016–17*. https://www.ecdc.europa.eu/en/seasonal-influenza/season-2016-17 (2020).
803. **Wang L, Wu A, Wang YE, Quanquin N, Li C, *et al.*** Functional Genomics Reveals Linkers Critical for Influenza Virus Polymerase. *Journal of Virology* 2016;90:2938–2947.

804. **Andrés C, Gimferrer L, Codina MG, Campins M, Almirante B, *et al.*** Full Genome Sequence Analysis Of Influenza H1PDM09 And H3N2 Viruses Related To Severe Respiratory Illness At A Tertiary University Hospital From 2012 To 2015 In Catalonia, Spain. 2015. p. 2015.

805. **Cai M, Zhong R, Qin C, Yu Z, Wen X, *et al.*** The R251K Substitution in Viral Protein PB2 Increases Viral Replication and Pathogenicity of Eurasian Avian-like H1N1 Swine Influenza Viruses. *Viruses* 2020;12:52.

806. **Koçer ZA, Fan Y, Huether R, Obenauer J, Webby RJ, *et al.*** Survival analysis of infected mice reveals pathogenic variations in the genome of avian H1N1 viruses. *Scientific Reports* 2015;4:7455.

807. **Tzarum N, de Vries RP, Zhu X, Yu W, McBride R, *et al.*** Structure and Receptor Binding of the Hemagglutinin from a Human H6N1 Influenza Virus. *Cell Host & Microbe* 2015;17:369–376.

808. **Jorquera PA, Mishin VP, Chesnokov A, Nguyen HT, Mann B, *et al.*** Insights into the antigenic advancement of influenza A(H3N2) viruses, 2011–2018. *Scientific Reports* 2019;9:2676.

809. **Trebbien R, Fischer TK, Krause TG, Nielsen L, Nielsen XC, *et al.*** Changes in genetically drifted H3N2 influenza A viruses and vaccine effectiveness in adults 65 years and older during the 2016/17 season in Denmark. *Journal of Clinical Virology* 2017;94:1–7.

810. **Hirst GK**. ADSORPTION OF INFLUENZA HEMAGGLUTININS AND VIRUS BY RED BLOOD CELLS. *The Journal of experimental medicine* 1942;76:195–209.

811. **Gottschalk A**. Neuraminidase: the specific enzyme of influenza virus and Vibrio cholerae. *Biochimica et Biophysica Acta* 1957;23:645–646.

812. **Palese P, Compans RW**. Inhibition of Influenza Virus Replication in Tissue Culture by 2-deoxy-2,3-dehydro-N-trifluoroacetylneuraminic acid (FANA): Mechanism of Action. *Journal of General Virology* 1976;33:159–163.

813. **Russell RJ, Haire LF, Stevens DJ, Collins PJ, Lin YP, *et al.*** The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* 2006;443:45–49.

814. **Webster RG, Laver WG**. Preparation and properties of antibody directed specifically against the neuraminidase of influenza virus. *Journal of immunology (Baltimore, Md : 1950)* 1967;99:49–55.

815. **Varghese JN, Laver WG, Colman PM**. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature* 1983;303:35–40.

816. **Ping J, Keleta L, Forbes NE, Dankar S, Stecho W, *et al.*** Genomic and Protein Structural Maps of Adaptive Evolution of Human Influenza A Virus to Increased Virulence in the Mouse. *PLoS ONE* 2011;6:e21740.

817. **Nilsson BE**. *Viral and Host Factors Regulating Influenza Virus Replication*. Oxford University; 2017.

818. **Tamuri AU, dos Reis M, Hay AJ, Goldstein RA**. Identifying Changes in Selective Constraints: Host Shifts in Influenza. *PLoS Computational Biology* 2009;5:e1000564.

819. **Wan H, Gao J, Yang H, Yang S, Harvey R, *et al.*** The neuraminidase of A(H3N2) influenza viruses circulating since 2016 is antigenically distinct from the A/Hong Kong/4801/2014 vaccine strain. *Nature Microbiology* 2019;4:2216–2225.

820. **Farooqui A, Leon AJ, Lei Y, Wang P, Huang J, *et al.*** Heterogeneous virulence of pandemic 2009 influenza H1N1 virus in mice. *Virology Journal* 2012;9:104.

821. **Wei D, Yu D-M, Wang M, Zhang D, Cheng Q, *et al.*** Genome-wide characterization of the seasonal H3N2 virus in Shanghai reveals natural temperature-sensitive strains conferred by the I668V mutation in the PA subunit. *Emerging Microbes & Infections* 2018;7:1–15.

822. **Fodor E**. The RNA polymerase of influenza A virus: mechanisms of viral transcription and replication. *Acta virologica* 2013;57:113–122.

823. **Guilligay D, Tarendeau F, Resa-Infante P, Coloma R, Crepin T, *et al.*** The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nature Structural & Molecular Biology* 2008;15:500–506.

824. **Lee C-YY, An S-HH, Kim I, Go D-MM, Kim D-YY, *et al.*** Prerequisites for the acquisition of mammalian pathogenicity by influenza A virus with a prototypic avian PB2 gene. *Scientific Reports* 2017;7:10205.

825. **Eshaghi A, Duvvuri VR, Li A, Patel SN, Bastien N, *et al.*** Genetic characterization of seasonal influenza A (H3N2) viruses in Ontario during 2010-2011 influenza season: high prevalence of mutations at antigenic sites. *Influenza and Other Respiratory Viruses* 2014;8:250–257.

826. **Kawakami C, Yamayoshi S, Akimoto M, Nakamura K, Miura H, *et al.*** Genetic and antigenic characterisation of influenza A(H3N2) viruses isolated in Yokohama during the 2016/17 and 2017/18 influenza seasons. *Eurosurveillance*;24. Epub ahead of print 7 February 2019. DOI: 10.2807/1560-7917.ES.2019.24.6.1800467.

827. **Galiano M, Johnson BF, Myers R, Ellis J, Daniels R, *et al.*** Fatal Cases of Influenza A(H3N2) in Children: Insights from Whole Genome Sequence Analysis. *PLoS ONE* 2012;7:e33166.

828. **Lyons D, Lauring A**. Mutation and Epistasis in Influenza Virus Evolution. *Viruses* 2018;10:407.

829. **Visher E, Whitefield SE, McCrone JT, Fitzsimmons W, Lauring AS**. The Mutational Robustness of Influenza A Virus. *PLOS Pathogens* 2016;12:e1005856.

830. **Milián E, Kamen AA**. Current and Emerging Cell Culture Manufacturing Technologies for Influenza Vaccines. *BioMed Research International* 2015;2015:1–11.

831. **European Medicines Agency**. EU recommendations for 2017/2018 seasonal flu vaccine composition. *European Medicines Agency*. https://www.ema.europa.eu/en/news/eu-recommendations-20172018-seasonal-flu-vaccine-composition (2018, accessed 7 March 2022).

832. **Tsou TP, Su CP, Huang WT, Yang JR, Liu MT**. Influenza a(H3n2) virus variants and patient characteristics during a summer influenza epidemic in Taiwan, 2017. *Eurosurveillance* 2017;22:1–6.

833. **Glatman-Freedman A, Drori Y, Beni SA, Friedman N, Pando R, *et al.*** Genetic divergence of Influenza A(H3N2) amino acid substitutions mark the beginning of the 2016–2017 winter season in Israel. *Journal of Clinical Virology* 2017;93:71–75.

834. **Nastouli E, Kellam P, Harvala H, Frampton D, Pillay D, *et al.*** Emergence of a novel subclade of influenza A(H3N2) virus in London, December 2016 to January 2017. *Eurosurveillance* 2017;22:1–6.

835. **Skowronski DM, Chambers C, Sabaiduc S, Dickinson JA, Winter A, *et al.*** Interim estimates of 2016/17 vaccine effectiveness against influenza A(H3N2), Canada, January 2017. *Eurosurveillance* 2017;22:1–8.

836. **Melidou A, Gioula G, Exindari M, Ioannou E, Gkolfinpoulou K, *et al.*** Influenza A(H3N2) genetic variants in vaccinated patients in northern Greece.pdf. 2017;29–32.

837. **Gilbertson DT, Rothman KJ, Chertow GM, Bradbury BD, Brookhart MA, *et al.*** Excess Deaths Attributable to Influenza-Like Illness in the ESRD Population. *Journal of the American Society of Nephrology* 2019;30:346–353.

838. **Choudhury D, Luna-Salazar C**. Preventive health care in chronic kidney disease and end-stage renal disease. *Nature Clinical Practice Nephrology* 2008;4:194–206.

839. **Kausz A, Pahari D**. The Value of Vaccination in Chronic Kidney Disease. *Seminars in Dialysis* 2004;17:9–11.

840. **Demirjian SG, Raina R, Bhimraj A, Navaneethan SD, Gordon SM, *et al.*** 2009 Influenza A Infection and Acute Kidney Injury: Incidence, Risk Factors, and Complications. *American Journal of Nephrology* 2011;34:1–8.

841. **World Health Organization**. Summary of neuraminidase amino acid substitutions associated with reduced inhibition by neuraminidase inhibitors. 2016;1–9.

842. **Maurer-Stroh S, Lee RTC, Limviphuvadh V, Ma J, Sirota FL, *et al.*** FluServer. http://flusurver.bii.a-star.edu.sg (2020, accessed 21 February 2020).

843. **Velazquez A, Bustria M, Ouyang Y, Moshiri N**. An analysis of clinical and geographical metadata of over 75,000 records in the GISAID COVID-19 database. *medRxiv [Preprint]*. Epub ahead of print 23 September 2020. DOI: 10.1101/2020.09.22.20199497.

844. **Nowak MA**. What is a quasispecies? *Trends in Ecology & Evolution* 1992;7:118–121.

845. **Andino R, Domingo E**. Viral quasispecies. *Virology* 2015;479–480:46–51.

846. **Webster RG, Laver WG, Air GM, Schild GC**. Molecular mechanisms of variation in influenza viruses. *Nature* 1982;296:115–121.

847. **Hurt AC, Barr IG**. Influenza viruses with reduced sensitivity to the neuraminidase inhibitor drugs in untreated young children. *Communicable diseases intelligence quarterly report* 2008;32:57—62.

848. **Forns X, Purcell RH, Bukh J**. Quasispecies in viral persistence and pathogenesis of hepatitis C virus. *Trends in Microbiology* 1999;7:402–410.

849. **Boivin G**. Detection and management of antiviral resistance for influenza viruses. *Influenza and other Respiratory Viruses* 2013;7:18–23.

850. **Xu Y, Lewandowski K, Downs LO, Kavanagh J, Hender T, *et al.*** Nanopore metagenomic sequencing of influenza virus directly from respiratory samples: diagnosis, drug resistance and nosocomial transmission, United Kingdom, 2018/19 influenza season. *Euro Surveill* 2021;26:2000004.

851. **Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GCM, *et al.*** Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. *Science* 2013;342:976–979.

852. **Posada-Céspedes S, Seifert D, Topolsky I, Jablonski KP, Metzner KJ, *et al.*** V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics* 2021;37:1673–1680.

853. **Tonkin-Hill G, Martincorena I, Amato R, Lawson AR, Gerstung M, *et al.*** Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* 2021;10:e66857.

854. **Łuksza M, Lässig M**. A predictive fitness model for influenza. *Nature* 2014;507:57–61.

855. **Pompei S, Loreto V, Tria F**. Phylogenetic Properties of RNA Viruses. *PLoS ONE* 2012;7:e44849.

856. **Varble A, Albrecht RA, Backes S, Crumiller M, Bouvier NM, *et al.*** Influenza A Virus Transmission Bottlenecks Are Defined by Infection Route and Recipient Host. *Cell Host & Microbe* 2014;16:691–700.

857. **Xue KS, Stevens-Ayers T, Campbell AP, Englund JA, Pergam SA, *et al.*** Parallel evolution of influenza across multiple spatiotemporal scales. *eLife*;6. Epub ahead of print 2017. DOI: 10.7554/eLife.26875.

858. **Rogers MB, Song T, Sebra R, Greenbaum BD, Hamelin M-E, *et al.*** Intrahost dynamics of antiviral resistance in influenza A virus reflect complex patterns of segment linkage, reassortment, and natural selection. *mBio*;6. Epub ahead of print 7 April 2015. DOI: 10.1128/mBio.02464-14.

859. **Ghedin E, Laplante J, DePasse J, Wentworth DE, Santos RP, *et al.*** Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. *The Journal of infectious diseases* 2011;203:168–74.

860. **Xue KS, Bloom JD**. Linking influenza virus evolution within and between human hosts. *Virus Evolution*;6. Epub ahead of print 1 January 2020. DOI: 10.1093/ve/veaa010.

861. **Dinis JM, Florek NW, Fatola OO, Moncla LH, Mutschler JP, *et al.*** Deep Sequencing Reveals Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans. *Journal of virology* 2016;90:3355–65.

862. **Debbink K, McCrone JT, Petrie JG, Truscon R, Johnson E, *et al.*** Vaccination has minimal impact on the intrahost diversity of H3N2 influenza viruses. *PLOS Pathogens* 2017;13:e1006194.

863. **Kundu S, Lockwood J, Depledge DP, Chaudhry Y, Aston A, *et al.*** Next-Generation Whole Genome Sequencing Identifies the Direction of Norovirus Transmission in Linked Patients. *Clinical Infectious Diseases* 2013;57:407–414.

864. **Robasky K, Lewis NE, Church GM**. The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics* 2014;15:56–62.

865. **Van Poelvoorde LAE, Vanneste K, De Keersmaecker SCJ, Thomas I, Van Goethem N, *et al.*** Whole-genome viral sequence analysis reveals mutations associated with influenza patient data. *Frontiers in Microbiology (Accepted)*.

866. **Thomas I, Barbezange C, Hombrouck A, Gucht SV, Weyckmans J, *et al.*** Virological Surveillance of Influenza in Belgium; Season 2016-2017. *Sciensano Influenza Report* 2017;1–33.

867. **Hunter JD**. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* 2007;9:90–95.

868. **World Health Organization**. *CDC protocol of realtime RTPCR for influenza A(H1N1)*. https://cdn.who.int/media/docs/default-source/influenza/molecular-detention-of-influenza-viruses/protocols_influenza_virus_detection_feb_2021.pdf?sfvrsn=df7d268a_5 (28 April 2009).

869. **Invitrogen**. *Advanced analysis with the SuperScript IV One-Step RT-PCR System*. https://assets.thermofisher.com/TFS-Assets/BID/Reference-Materials/advanced-analysis-superscript-iv-one-step-rt-pcr-system-white-paper.pdf (2018, accessed 18 January 2022).

870. **Broad Institute**. Data pre-processing for variant discovery. *GATK*. https://gatk.broadinstitute.org/hc/en-us/articles/360035535912-Data-pre-processing-for-variant-discovery (2022, accessed 14 January 2022).

871. **Marx V**. How to deduplicate PCR. *Nat Methods* 2017;14:473–476.

872. **Kassahn KS, Holmes O, Nones K, Patch A-M, Miller DK, *et al.*** Somatic Point Mutation Calling in Low Cellularity Tumors. *PLOS ONE* 2013;8:e74380.

873. **Tian S, Yan H, Kalmbach M, Slager SL**. Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics* 2016;17:403.

874. **van Rossum G**. Python tutorial, May 1995. *CWI Report CS-R9526* 1995;1–65.

875. **Kenmoe S, Tchendjou P, Moyo Tetang S, Mossus T, Njankouo Ripa M, *et al.*** Evaluating the performance of a rapid antigen test for the detection of influenza virus in clinical specimens from children in Cameroon. *Influenza and Other Respiratory Viruses* 2014;8:131–134.

876. **Caselton DL, Arunga G, Emukule G, Muthoka P, Mayieka L, *et al.*** Does the length of specimen storage affect influenza testing results by real-time reverse transcription-polymerase chain reaction? An analysis of influenza surveillance specimens, 2008 to 2010. *Eurosurveillance*;19. Epub ahead of print 11 September 2014. DOI: 10.2807/1560-7917.ES2014.19.36.20893.

877. **Bouscambert Duchamp M, Casalegno JS, Gillet Y, Frobert E, Bernard E, *et al.*** Pandemic A(H1N1)2009 influenza virus detection by real time RT-PCR : is viral quantification useful? *Clinical Microbiology and Infection* 2010;16:317–321.

878. **Chen H, Cohen P, Chen S**. How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics - Simulation and Computation* 2010;39:860–864.

879. **Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, *et al.*** VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research* 2016;44:e108.

880. **Garrison E, Marth G**. Haplotype-based variant detection from short-read sequencing. *arXiv:12073907 [q-bio]*. http://arxiv.org/abs/1207.3907 (2012, accessed 18 January 2022).

881. **Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, *et al.*** VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;22:568–576.

882. **Mohammed KS, Kibinge N, Prins P, Agoti CN, Cotten M, *et al.*** Evaluating the performance of tools used to call minority variants from whole genome short-read data. Epub ahead of print 13 September 2018. DOI: 10.12688/wellcomeopenres.13538.2.

883. **Deng Z-L, Dhingra A, Fritz A, Götting J, Münch PC, *et al.*** Evaluating assembly and variant calling software for strain-resolved analysis of large DNA viruses. *Briefings in Bioinformatics* 2021;22:bbaa123.

884. **Stead LF, Sutton KM, Taylor GR, Quirke P, Rabbitts P**. Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: Applications in tumor subclone resolution. *Human Mutation* 2013;34:1432–1438.

885. **Gelbart M, Harari S, Ben-Ari Y, Kustin T, Wolf D, *et al.*** Drivers of within-host genetic diversity in acute infections of viruses. *PLOS Pathogens* 2020;16:e1009029.

886. **Orton RJ, Wright CF, Morelli MJ, King DJ, Paton DJ, *et al.*** Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. *BMC Genomics* 2015;16:229.

887. **King DJ, Freimanis G, Lasecka-Dykes L, Asfor A, Ribeca P, *et al.*** A Systematic Evaluation of High-Throughput Sequencing Approaches to Identify Low-Frequency Single Nucleotide Variants in Viral Populations. *Viruses* 2020;12:1187.

888. **Honce R, Schultz-Cherry S**. They are what you eat: Shaping of viral populations through nutrition and consequences for virulence. *PLOS Pathogens* 2020;16:e1008711.

889. **Lin S-R, Yang T-Y, Peng C-Y, Lin Y-Y, Dai C-Y, *et al.*** Whole genome deep sequencing analysis of viral quasispecies diversity and evolution in HBeAg seroconverters. *JHEP Reports* 2021;3:100254.

890. **Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, *et al.*** Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine* 2020;26:1037–1040.

891. **Wurtzer S, Marechal V, Mouchel J, Maday Y, Teyssou R, *et al.*** Evaluation of lockdown effect on SARS-CoV-2 dynamics through viral genome quantification in waste water, Greater Paris, France, 5 March to 23 April 2020. *Eurosurveillance*;25. Epub ahead of print 17 December 2020. DOI: 10.2807/1560-7917.ES.2020.25.50.2000776.

892. **Emami A, Javanmardi F, Pirbonyeh N, Akbari A**. Prevalence of Underlying Diseases in Hospitalized Patients with COVID-19: a Systematic Review and Meta-Analysis. *Archives of academic emergency medicine* 2020;8:e35.

893. **Machado BAS, Hodel KVS, Barbosa-Júnior VG, Soares MBP, Badaró R**. The Main Molecular and Serological Methods for Diagnosing COVID-19: An Overview Based on the Literature. *Viruses* 2020;13:40.

894. **Lu N, Cheng KW, Qamar N, Huang KC, Johnson JA**. Weathering COVID-19 storm: Successful control measures of five Asian countries. *American Journal of Infection Control* 2020;48:851–852.

895. **Cha V**. Asia's COVID-19 Lessons for the West: Public Goods, Privacy, and Social Tagging. *Washington Quarterly* 2020;43:33–50.

896. **Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, *et al.*** Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution* 2020;6:1–8.

897. **SAGE-EMG S-B Transmission Group**. Mitigations to Reduce Transmission of the new variant SARS-CoV-2 virus. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/948607/s0995-mitigations-to-reduce-transmission-of-the-new-variant.pdf (2020, accessed 2 July 2021).

898. **Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, *et al.*** Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host & Microbe* 2021;29:44-57.e9.

899. **Gand M, Vanneste K, Thomas I, Van Gucht S, Capron A, *et al.*** Use of Whole Genome Sequencing Data for a First *in silico* Specificity Evaluation of the RT-qPCR Assays Used for SARS-CoV-2 Detection. *International Journal of Molecular Sciences* 2020;21:5585.

900. **Han MS, Byun J-H, Cho Y, Rim JH**. RT-PCR for SARS-CoV-2: quantitative versus qualitative. *The Lancet Infectious Diseases* 2021;21:165.

901. **Girones R, Ferrús MA, Alonso JL, Rodriguez-Manzano J, Calgua B, *et al.*** Molecular detection of pathogens in water – The pros and cons of molecular techniques. *Water Research* 2010;44:4325–4339.

902. **Whale AS, von der Heide EK, Kohlenberg M, Brinckmann A, Baedker S, *et al.*** Digital PCR can augment the interpretation of RT-qPCR Cq values for SARS-CoV-2 diagnostics. *Methods*. Epub ahead of print 26 August 2021. DOI: 10.1016/j.ymeth.2021.08.006.

903. **Liu X, Feng J, Zhang Q, Guo D, Zhang L, *et al.*** Analytical comparisons of SARS-COV-2 detection by qRT-PCR and ddPCR with multiple primer/probe sets. *Emerging Microbes & Infections* 2020;9:1175–1179.

904. **Suo T, Liu X, Feng J, Guo M, Hu W, *et al.*** ddPCR: a more accurate tool for SARS-CoV-2 detection in low viral load specimens. *Emerging Microbes & Infections* 2020;9:1259–1268.

905. **Vogelstein B, Kinzler KW**. Digital PCR. *Proceedings of the National Academy of Sciences* 1999;96:9236–9241.

906. **Sanders R, Huggett JF, Bushell CA, Cowen S, Scott DJ, *et al.*** Evaluation of Digital PCR for Absolute DNA Quantification. *Analytical Chemistry* 2011;83:6474–6484.

907. **Whale AS, Huggett JF, Cowen S, Speirs V, Shaw J, *et al.*** Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. *Nucleic Acids Research* 2012;40:e82–e82.

908. **Sanders R, Mason DJ, Foy CA, Huggett JF**. Evaluation of Digital PCR for Absolute RNA Quantification. *PLoS ONE* 2013;8:e75296.

909. **Hindson CM, Chevillet JR, Briggs HA, Gallichotte EN, Ruf IK, *et al.*** Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nature Methods* 2013;10:1003–1005.

910. **Taylor SC, Carbonneau J, Shelton DN, Boivin G**. Optimization of Droplet Digital PCR from RNA and DNA extracts with direct comparison to RT-qPCR: Clinical implications for quantification of Oseltamivir-resistant subpopulations. *Journal of Virological Methods* 2015;224:58–66.

911. **McDermott GP, Do D, Litterst CM, Maar D, Hindson CM, *et al.*** Multiplexed Target Detection Using DNA-Binding Dye Chemistry in Droplet Digital PCR. *Analytical Chemistry* 2013;85:11619–11627.

912. **de Kock R, Baselmans M, Scharnhorst V, Deiman B**. Sensitive detection and quantification of SARS-CoV-2 by multiplex droplet digital RT-PCR. *Eur J Clin Microbiol Infect Dis* 2021;40:807–813.

913. **Alteri C, Cento V, Antonello M, Colagrossi L, Merli M, *et al.*** Detection and quantification of SARS-CoV-2 by droplet digital PCR in real-time PCR negative nasopharyngeal swabs from suspected COVID-19 patients. *PLOS ONE* 2020;15:e0236311.

914. **Deiana M, Mori A, Piubelli C, Scarso S, Favarato M, *et al.*** Assessment of the direct quantitation of SARS-CoV-2 by droplet digital PCR. *Scientific Reports* 2020;10:18764.

915. **Heijnen L, Elsinga G, de Graaf M, Molenkamp R, Koopmans MPG, *et al.*** Droplet Digital RT-PCR to detect SARS-CoV-2 variants of concern in wastewater. *medRxiv [Preprint]* 2021;2021.03.25.21254324.

916. **Gonzalez R, Larson A, Thompson H, Carter E, Cassi XF**. Redesigning SARS-CoV-2 clinical RT-qPCR assays for wastewater RT-ddPCR. *medRxiv [Preprint]* 2021;2021.03.02.21252754.

917. **D'Aoust PM, Mercier E, Montpetit D, Jia J-J, Alexandrov I, *et al.*** Quantitative analysis of SARS-CoV-2 RNA from wastewater solids in communities with low COVID-19 incidence and prevalence. *Water Research* 2021;188:116560.

918. **Kinloch NN, Ritchie G, Dong W, Cobarrubias KD, Sudderuddin H, *et al.*** SARS-CoV-2 RNA Quantification Using Droplet Digital RT-PCR. *The Journal of Molecular Diagnostics* 2021;23:907–919.

919. **Rački N, Morisset D, Gutierrez-Aguirre I, Ravnikar M**. One-step RT-droplet digital PCR: a breakthrough in the quantification of waterborne RNA viruses. *Analytical and Bioanalytical Chemistry* 2014;406:661–667.

920. **World Health Organization**. World Health Organization (WHO) Molecular Assays to Diagnose COVID-19: Summary Table of Available Protocols.

921. **Lu R, Zhao X, Li J, Niu P, Yang B, *et al.*** Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* 2020;395:565–574.

922. **Chan JF-W, Yip CC-Y, To KK-W, Tang TH-C, Wong SC-Y, *et al.*** Improved Molecular Diagnosis of COVID-19 by the Novel, Highly Sensitive and Specific COVID-19-RdRp/Hel Real-Time Reverse Transcription-PCR Assay Validated *In vitro* and with Clinical Specimens. *Journal of Clinical Microbiology*;58. Epub ahead of print 4 March 2020. DOI: 10.1128/JCM.00310-20.

923. **Gand M, Vanneste K, Thomas I, Van Gucht S, Capron A, *et al.*** Deepening of *In silico* Evaluation of SARS-CoV-2 Detection RT-qPCR Assays in the Context of New Variants. *Genes* 2021;12:565.

924. **Vanneste K, Garlant L, Broeders S, Van Gucht S, Roosens NH**. Application of whole genome data for *in silico* evaluation of primers and probes routinely employed for the detection of viral species by RT-qPCR using dengue virus as a case study. *BMC Bioinformatics* 2018;19:1–18.

925. **Lefever S, Pattyn F, Hellemans J, Vandesompele J**. Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays. *Clinical chemistry* 2013;59:1470–1480.

926. **Whiley DM, Sloots TP**. Sequence variation in primer targets affects the accuracy of viral quantitative PCR. *Journal of Clinical Virology* 2005;34:104–107.

927. **Fraiture M-A, Deckers M, Papazova N, Roosens NHC**. Detection strategy targeting a chloramphenicol resistance gene from genetically modified bacteria in food and feed products. *Food Control* 2020;108:106873.

928. **Uhlig S, Frost K, Colson B, Simon K, Mäde D, *et al.*** Validation of qualitative PCR methods on the basis of mathematical–statistical modelling of the probability of detection. *Accreditation and Quality Assurance* 2015;20:75–83.

929. **Federaal Agentschap voor Geneesmiddelen en Gezondheidsproducten**. Compendium biobanken. https://www.fagg-afmps.be/nl/MENSELIJK_gebruik/gezondheidsproducten/menselijk_lichaamsmateriaal/menselijk_lichaamsmateriaal_0 (accessed 3 June 2021).

930. **Phan T**. Genetic diversity and evolution of SARS-CoV-2. *Infection, Genetics and Evolution* 2020;81:104260.

931. **Peñarrubia L, Ruiz M, Porco R, Rao SN, Juanola-Falgarona M, *et al.*** Multiple assays in a real-time RT-PCR SARS-CoV-2 panel can mitigate the risk of loss of sensitivity by new genomic variants during the COVID-19 outbreak. *International Journal of Infectious Diseases* 2020;97:225–229.

932. **Telwatte S, Kumar N, Vallejo-Gracia A, Kumar GR, Lu CM, *et al.*** Novel RT-ddPCR assays for simultaneous quantification of multiple noncoding and coding regions of SARS-CoV-2 RNA. *Journal of Virological Methods* 2021;292:114115.

933. **Telwatte S, Martin HA, Marczak R, Fozouni P, Vallejo-Gracia A, *et al.*** Novel RT-ddPCR assays for measuring the levels of subgenomic and genomic SARS-CoV-2 transcripts. *Methods*. Epub ahead of print 18 April 2021. DOI: 10.1016/j.ymeth.2021.04.011.

934. **Pezzi L, Charrel RN, Ninove L, Nougairede A, Molle G, *et al.*** Development and Evaluation of a duo SARS-CoV-2 RT-qPCR Assay Combining Two Assays Approved by the World Health Organization Targeting the Envelope and the RNA-Dependant RNA Polymerase (RdRp) Coding Regions. *Viruses* 2020;12:686.

935. **Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, *et al.*** The Architecture of SARS-CoV-2 Transcriptome. *Cell* 2020;181:914-921.e10.

936. **Leclerc QJ, Fuller NM, Knight LE, Funk S, Knight GM**. What settings have been linked to SARS-CoV-2 transmission clusters? *Wellcome Open Research* 2020;5:83.

937. **Azuma K, Yanagi U, Kagi N, Kim H, Ogata M, *et al.*** Environmental factors involved in SARS-CoV-2 transmission: effect and role of indoor environmental quality in the strategy for COVID-19 infection control. *Environmental Health and Preventive Medicine* 2020;25:66.

938. **Bayle C, Cantin D, Vidal J-S, Sourdeau E, Slama L, *et al.*** Asymptomatic SARS COV-2 carriers among nursing home staff: A source of contamination for residents? *Infectious Diseases Now* 2021;51:197–200.

939. **Contreras S, Dehning J, Loidolt M, Zierenberg J, Spitzner FP, *et al.*** The challenges of containing SARS-CoV-2 via test-trace-and-isolate. *Nature Communications* 2021;12:378.

940. **Zhang W, Du R-H, Li B, Zheng X-S, Yang X-L, *et al.*** Molecular and serological investigation of 2019-nCoV infected patients: implication of multiple shedding routes. *Emerging microbes & infections* 2020;9:386–389.

941. **Wu Y, Guo C, Tang L, Hong Z, Zhou J, *et al.*** Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. *The lancet Gastroenterology & hepatology* 2020;5:434–435.

942. **European Commission**. *Commission Recommendation of 17.3.2021 on a common approach to establish a systematic surveillance of SARS-CoV-2 and its variants in wastewaters in the EU*. 2021.

943. **Thompson JR, Nancharaiah Y V, Gu X, Lee WL, Rajal VB, *et al.*** Making waves: Wastewater surveillance of SARS-CoV-2 for population-based health management. *Water Research* 2020;184:116181.

944. **Panchal D, Prakash O, Bobde P, Pal S**. SARS-CoV-2: sewage surveillance as an early warning system and challenges in developing countries. *Environmental Science and Pollution Research*. Epub ahead of print 17 March 2021. DOI: 10.1007/s11356-021-13170-8.

945. **Gómez CE, Perdiguero B, Esteban M**. Emerging SARS-CoV-2 Variants and Impact in Global Vaccination Programs against SARS-CoV-2/COVID-19. *Vaccines* 2021;9:243.

946. **Boni MF**. Vaccination and antigenic drift in influenza. *Vaccine* 2008;26:C8–C14.

947. **Bal A, Destras G, Gaymard A, Stefic K, Marlet J, *et al.*** Two-step strategy for the identification of SARS-CoV-2 variant of concern 202012/01 and other variants with spike deletion H69–V70, France, August to December 2020. *Eurosurveillance* 2021;26:1–5.

948. **Charre C, Ginevra C, Sabatier M, Regue H, Destras G, *et al.*** Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evolution*;6. Epub ahead of print 1 July 2020. DOI: 10.1093/ve/veaa075.

949. **Hartley PD, Tillett RL, AuCoin DP, Sevinsky JR, Xu Y, *et al.*** Genomic surveillance of Nevada patients revealed prevalence of unique SARS-CoV-2 variants bearing mutations in the RdRp gene. *Journal of Genetics and Genomics*. Epub ahead of print February 2021. DOI: 10.1016/j.jgg.2021.01.004.

950. **Firestone MJ, Lorentz AJ, Meyer S, Wang, X, Como-Sabetti K, *et al.*** First Identified Cases of SARS-CoV-2 Variant P.1 in the United States — Minnesota, January 2021. *MMWR Morbidity and Mortality Weekly Report* 2021;70:346–347.

951. **Lin J, Tang C, Wei H, Du B, Chen C, *et al.*** Genomic monitoring of SARS-CoV-2 uncovers an Nsp1 deletion variant that modulates type I interferon response. *Cell Host & Microbe* 2021;29:489-502.e8.

952. **Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, *et al.*** SARS-CoV-2 within-host diversity and transmission. *Science* 2021;372:eabg0821.

953. **Siqueira JD, Goes LR, Alves BM, de Carvalho PS, Cicala C, *et al.*** SARS-CoV-2 genomic analyses in cancer patients reveal elevated intrahost genetic diversity. *Virus Evol* 2021;7:veab013.

954. **Karim F, Moosa MYS, Gosnell BI, Cele S, Giandhari J, *et al.*** Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection. *medRxiv [Preprint]* 2021;2021.06.03.21258228.

955. **Bar-Or I, Weil M, Indenbaum V, Bucris E, Bar-Ilan D, *et al.*** Detection of SARS-CoV-2 variants by genomic analysis of wastewater samples in Israel. *Science of The Total Environment* 2021;789:148002.

956. **Crits-Christoph A, Kantor RS, Olm MR, Whitney ON, Al-Shayeb B, *et al.*** Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. *mBio*;12. Epub ahead of print 19 January 2021. DOI: 10.1128/mBio.02703-20.

957. **Sharif S, Ikram A, Khurshid A, Salman M, Mehmood N, *et al.*** Detection of SARs-CoV-2 in wastewater using the existing environmental surveillance network: A potential supplementary system for monitoring COVID-19 transmission. *PLOS ONE* 2021;16:e0249568.

958. **Jahn K, Dreifuss D, Topolsky I, Kull A, Ganesanandamoorthy P, *et al.*** Detection of SARS-CoV-2 variants in Switzerland by genomic analysis of wastewater samples. *medRxiv* 2021;2021.01.08.21249379.

959. **Rios G, Lacoux C, Leclercq V, Diamant A, Lebrigand K, *et al.*** Monitoring SARS-CoV-2 variants alterations in Nice neighborhoods by wastewater nanopore sequencing. *medRxiv [Preprint]*. Epub ahead of print 9 July 2021. DOI: 10.1101/2021.07.09.21257475.

960. **Isakov O, Bordería A V., Golan D, Hamenahem A, Celniker G, *et al.*** Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. *Bioinformatics* 2015;31:2141–2150.

961. **Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, *et al.*** Highly Sensitive and Specific Detection of Rare Variants in Mixed Viral Populations from Massively Parallel Sequence Data. *PLoS Computational Biology* 2012;8:e1002417.

962. **Bushnell B**. BBMap. https://sourceforge.net/projects/bbmap/ (accessed 29 March 2021).

963. **Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, *et al.*** Sustainable data analysis with Snakemake. *F1000Research* 2021;10:33.

964. **Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, *et al.*** *Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations*. 2020.

965. **Public Health England**. Variants of concern or under investigation. https://www.gov.uk/government/publications/covid-19-variants-genomically-confirmed-case-numbers/variants-distribution-of-cases-data (2021, accessed 4 March 2021).

966. **Mishra S, Mindermann S, Sharma M, Whittaker C, Mellan TA, *et al.*** Changing composition of SARS-CoV-2 lineages and rise of Delta variant in England. *EClinicalMedicine* 2021;39:101064.

967. **Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, *et al.*** Array programming with NumPy. *Nature* 2020;585:357–362.

968. **Lindenbaum P**. JVarkit: java-based utilities for Bioinformatics.

969. **Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, *et al.*** Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods* 2015;12:623–630.

970. **Quinlan AR, Hall IM**. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–842.

971. **Sievert C**. Interactive Web-Based Data Visualization with R, plotly, and shiny.

972. **Saawarn B, Hait S**. Occurrence, fate and removal of SARS-CoV-2 in wastewater: Current knowledge and future perspectives. *Journal of environmental chemical engineering* 2021;9:104870.

973. **Eliseev A, Gibson KM, Avdeyev P, Novik D, Bendall ML, *et al.*** Evaluation of haplotype callers for next-generation sequencing of viruses. *Infection, Genetics and Evolution* 2020;82:104277.

974. **Elwert F, Winship C**. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology* 2014;40:31–53.

975. **Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, *et al.*** Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology* 2021;50:620–632.

976. **Boogaerts T, Van den Bogaert S, Van Poelvoorde LAE, El Masri D, De Roeck N, *et al.*** Optimization and Application of a Multiplex Digital PCR Assay for the Detection of SARS-CoV-2 Variants of Concern in Belgian Influent Wastewater. *Viruses* 2022;14:610.

977. **Bouki C, Venieri D, Diamadopoulos E**. Detection and fate of antibiotic resistant bacteria in wastewater treatment plants: A review. *Ecotoxicology and Environmental Safety* 2013;91:1–9.

978. **García-Aljaro C, Blanch AR, Campos C, Jofre J, Lucena F**. Pathogens, faecal indicators and human-specific microbial source-tracking markers in sewage. *J Appl Microbiol* 2019;126:701–717.

979. **Manaia CM, Rocha J, Scaccia N, Marano R, Radu E, *et al.*** Antibiotic resistance in wastewater treatment plants: Tackling the black box. *Environment International* 2018;115:312–324.

980. **Aguiar-Oliveira M de L, Campos A, R. Matos A, Rigotto C, Sotero-Martins A, *et al.*** Wastewater-Based Epidemiology (WBE) and Viral Detection in Polluted Surface Water: A Valuable Tool for COVID-19 Surveillance—A Brief Review. *IJERPH* 2020;17:9251.

981. **Amoah ID, Kumari S, Bux F**. Coronaviruses in wastewater processes: Source, fate and potential risks. *Environment International* 2020;143:105962.

982. **Kitajima M, Ahmed W, Bibby K, Carducci A, Gerba CP, *et al.*** SARS-CoV-2 in wastewater: State of the knowledge and research needs. *Science of The Total Environment* 2020;739:139076.

983. **Lyu J, Yang L, Zhang L, Ye B, Wang L**. Antibiotics in soil and water in China–a systematic review and source analysis. *Environmental Pollution* 2020;266:115147.

984. **Sims N, Kasprzyk-Hordern B**. Future perspectives of wastewater-based epidemiology: Monitoring infectious disease spread and resistance to the community level. *Environment International* 2020;139:105689.

985. **Wang J, Chu L, Wojnárovits L, Takács E**. Occurrence and fate of antibiotics, antibiotic resistant genes (ARGs) and antibiotic resistant bacteria (ARB) in municipal wastewater treatment plant: An overview. *Science of The Total Environment* 2020;744:140997.

986. **Ahmad J, Ahmad M, Usman ARA, Al-Wabel MI**. Prevalence of human pathogenic viruses in wastewater: A potential transmission risk as well as an effective tool for early outbreak detection for COVID-19. *Journal of Environmental Management* 2021;298:113486.

987. **Anand U, Li X, Sunita K, Lokhandwala S, Gautam P, *et al.*** SARS-CoV-2 and other pathogens in municipal wastewater, landfill leachate, and solid waste: A review about virus surveillance, infectivity, and inactivation. *Environmental Research* 2022;203:111839.

988. **Buonerba A, Corpuz MVA, Ballesteros F, Choo K-H, Hasan SW, *et al.*** Coronavirus in water media: Analysis, fate, disinfection and epidemiological applications. *Journal of Hazardous Materials* 2021;415:125580.

989. **Hamouda M, Mustafa F, Maraqa M, Rizvi T, Aly Hassan A**. Wastewater surveillance for SARS-CoV-2: Lessons learnt from recent studies to define future applications. *Science of The Total Environment* 2021;759:143493.

990. **Mohapatra S, Menon NG, Mohapatra G, Pisharody L, Pattnaik A, *et al.*** The novel SARS-CoV-2 pandemic: Possible environmental transmission, detection, persistence and fate during wastewater and water treatment. *Science of The Total Environment* 2021;765:142746.

991. **Nguyen AQ, Vu HP, Nguyen LN, Wang Q, Djordjevic SP, *et al.*** Monitoring antibiotic resistance genes in wastewater treatment: Current strategies and future challenges. *Science of The Total Environment* 2021;783:146964.

992. **Patel M, Chaubey AK, Pittman CU, Mlsna T, Mohan D**. Coronavirus (SARS-CoV-2) in the environment: Occurrence, persistence, analysis in aquatic systems and possible management. *Science of The Total Environment* 2021;765:142698.

993. **Silverman AI, Boehm AB**. Systematic Review and Meta-Analysis of the Persistence of Enveloped Viruses in Environmental Waters and Wastewater in the Absence of Disinfectants. *Environ Sci Technol* 2021;55:14480–14493.

994. **Chau KK, Barker L, Budgell EP, Vihta KD, Sims N, *et al.*** Systematic review of wastewater surveillance of antimicrobial resistance in human populations. *Environment International* 2022;162:107171.

995. **Gholipour S, Ghalhari MR, Nikaeen M, Rabbani D, Pakzad P, *et al.*** Occurrence of viruses in sewage sludge: A systematic review. *Science of The Total Environment* 2022;824:153886.

996. **Holton E, Sims N, Jagadeesan K, Standerwick R, Kasprzyk-Hordern B**. Quantifying community-wide antimicrobials usage via wastewater-based epidemiology. *Journal of Hazardous Materials* 2022;436:129001.

997. **Shah S, Gwee SXW, Ng JQX, Lau N, Koh J, *et al.*** Wastewater surveillance to infer COVID-19 transmission: A systematic review. *Science of The Total Environment* 2022;804:150060.

998. **Twigg C, Wenk J**. Review and Meta-Analysis: SARS-CoV-2 and Enveloped Virus Detection in Feces and Wastewater. *ChemBioEng Reviews* 2022;9:129–145.

999. **Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, *et al.*** Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* 2020;21:30.

1000. **Gamaarachchi H, Samarakoon H, Jenner SP, Ferguson JM, Amos TG, *et al.*** Fast nanopore sequencing data analysis with SLOW5. *Nat Biotechnol* 2022;40:1026–1029.

1001. **Lewandowski K, Xu Y, Pullan ST, Lumley SF, Foster D, *et al.*** Metagenomic Nanopore Sequencing of Influenza Virus Direct from Clinical Respiratory Samples. *J Clin Microbiol* 2019;58:e00963-19.

1002. **King J, Harder T, Beer M, Pohlmann A**. Rapid multiplex MinION nanopore sequencing workflow for Influenza A viruses. *BMC Infectious Diseases* 2020;20:648.

1003. **Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, *et al.*** Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun* 2020;11:6272.

1004. **Wang Y, Chen D, Zhu C, Zhao Z, Gao S, *et al.*** Genetic Surveillance of Five SARS-CoV-2 Clinical Samples in Henan Province Using Nanopore Sequencing. *Frontiers in Immunology*;13. https://www.frontiersin.org/articles/10.3389/fimmu.2022.814806 (2022, accessed 4 August 2022).

1005. **Töpfer A, Marschall T, Bull RA, Luciani F, Schönhuth A, *et al.*** Viral Quasispecies Assembly via Maximal Clique Enumeration. *PLOS Computational Biology* 2014;10:e1003515.

1006. **Mangul S, Wu NC, Mancuso N, Zelikovsky A, Sun R, *et al.*** Accurate viral population assembly from ultra-deep sequencing data. *Bioinformatics* 2014;30:i329–i337.

1007. **Prosperi MC, Prosperi L, Bruselles A, Abbate I, Rozera G, *et al.*** Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics* 2011;12:5.

1008. **Ferdinands JM, Thompson MG, Blanton L, Spencer S, Grant L, *et al.*** Does influenza vaccination attenuate the severity of breakthrough infections? A narrative review and recommendations for further research. *Vaccine* 2021;39:3678–3695.

1009. **Oude Munnink BB, Worp N, Nieuwenhuijse DF, Sikkema RS, Haagmans B, *et al.*** The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat Med* 2021;27:1518–1524.

1010. **Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, *et al.*** GenBank. *Nucleic Acids Research* 2019;47:D94–D99.

1011. **Furuse Y, Suzuki A, Oshitani H**. Large-scale sequence analysis of M gene of influenza A viruses from different species: Mechanisms for emergence and spread of amantadine resistance. *Antimicrobial Agents and Chemotherapy* 2009;53:4457–4463.

1012. **Sciensano**. National infrastructure for genomic-epidemiologic surveillance of infectious diseases. *sciensano.be*. https://www.sciensano.be/en/projects/national-infrastructure-genomic-epidemiologic-surveillance-infectious-diseases (accessed 4 July 2022).

1013. **Salipante SJ, Sengupta DJ, Cummings LA, Robinson A, Kurosawa K, *et al.*** Whole genome sequencing indicates Corynebacterium jeikeium comprises 4 separate genomospecies and identifies a dominant genomospecies among clinical isolates. *Int J Med Microbiol* 2014;304:1001–1010.

1014. **Salipante SJ, Hoogestraat DR, Abbott AN, SenGupta DJ, Cummings LA, *et al.*** Coinfection of Fusobacterium nucleatum and Actinomyces israelii in mastoiditis diagnosed by next-generation DNA sequencing. *J Clin Microbiol* 2014;52:1789–1792.

1015. **Afshinnekoo E, Chou C, Alexander N, Ahsanuddin S, Schuetz AN, *et al.*** Precision Metagenomics: Rapid Metagenomic Analyses for Infectious Disease Diagnostics and Public Health Surveillance. *J Biomol Tech* 2017;28:40–45.

1016. **Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, *et al.*** Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010;7:111–118.

1017. **Andermann T, Torres Jiménez MF, Matos-Maraví P, Batista R, Blanco-Pastor JL, *et al.*** A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project. *Frontiers in Genetics*;10. https://www.frontiersin.org/articles/10.3389/fgene.2019.01407 (2020, accessed 11 August 2022).

1018. **Dacheux L, Cervantes-Gonzalez M, Guigon G, Thiberge J-M, Vandenbogaert M, *et al.*** A preliminary study of viral metagenomics of French bat species in contact with humans: identification of new mammalian viruses. *PLoS One* 2014;9:e87194.

1019. **Wu Z, Yang L, Ren X, He G, Zhang J, *et al.*** Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. *ISME J* 2016;10:609–620.

1020. **Moore NE, Wang J, Hewitt J, Croucher D, Williamson DA, *et al.*** Metagenomic analysis of viruses in feces from unsolved outbreaks of gastroenteritis in humans. *J Clin Microbiol* 2015;53:15–21.

1021. **Deng L, Silins R, Castro-Mejía JL, Kot W, Jessen L, *et al.*** A Protocol for Extraction of Infective Viromes Suitable for Metagenomics Sequencing from Low Volume Fecal Samples. *Viruses* 2019;11:667.

1022. **Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, *et al.*** Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res* 2019;29:831–842.

1023. **Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, *et al.*** Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N Engl J Med* 2019;380:2327–2340.

1024. **Rascovan N, Duraisamy R, Desnues C**. Metagenomics and the Human Virome in Asymptomatic Individuals. *Annu Rev Microbiol* 2016;70:125–141.

1025. **Law J, Jovel J, Patterson J, Ford G, O'keefe S, *et al.*** Identification of Hepatotropic Viruses from Plasma Using Deep Sequencing: A Next Generation Diagnostic Tool. *PLOS ONE* 2013;8:e60595.

1026. **Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, *et al.*** Characterization of the Viral Microbiome in Patients with Severe Lower Respiratory Tract Infections, Using Metagenomic Sequencing. *PLOS ONE* 2012;7:e30875.

1027. **Takeuchi S, Kawada J, Horiba K, Okuno Y, Okumura T, *et al.*** Metagenomic analysis using next-generation sequencing of pathogens in bronchoalveolar lavage fluid from pediatric patients with respiratory failure. *Sci Rep* 2019;9:12909.

1028. **Li Y, Fu X, Ma J, Zhang J, Hu Y, *et al.*** Altered respiratory virome and serum cytokine profile associated with recurrent respiratory tract infections in children. *Nat Commun* 2019;10:2288.

1029. **van den Munckhof EHA, de Koning MNC, Quint WGV, van Doorn L-J, Leverstein-van Hall MA**. Evaluation of a stepwise approach using microbiota analysis, species-specific qPCRs and culture for the diagnosis of lower respiratory tract infections. *Eur J Clin Microbiol Infect Dis* 2019;38:747–754.

1030. **Kohl C, Brinkmann A, Dabrowski PW, Radonić A, Nitsche A, *et al.*** Protocol for Metagenomic Virus Detection in Clinical Specimens1. *Emerg Infect Dis* 2015;21:48–57.

1031. **Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, *et al.*** Raw sewage harbors diverse viral populations. *mBio* 2011;2:e00180-11.

1032. **Martínez-Puchol S, Rusiñol M, Fernández-Cassi X, Timoneda N, Itarte M, *et al.*** Characterisation of the sewage virome: comparison of NGS tools and occurrence of significant pathogens. *Sci Total Environ* 2020;713:136604–136604.

1033. **Fernandez-Cassi X, Timoneda N, Martínez-Puchol S, Rusiñol M, Rodriguez-Manzano J, *et al.*** Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance. *Science of The Total Environment* 2018;618:870–880.

1034. **Adriaenssens EM, Farkas K, Harrison C, Jones DL, Allison HE, *et al.*** Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. *mSystems* 2018;3:e00025-18.

1035. **Guerrero-Latorre L, Romero B, Bonifaz E, Timoneda N, Rusiñol M, *et al.*** Quito's virome: Metagenomic analysis of viral diversity in urban streams of Ecuador's capital city. *Sci Total Environ* 2018;645:1334–1343.

1036. **Sievers A, Bosiek K, Bisch M, Dreessen C, Riedel J, *et al.*** K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. *Genes (Basel)* 2017;8:E122.

1037. **Kinsella CM, Deijs M, van der Hoek L**. Enhanced bioinformatic profiling of VIDISCA libraries for virus detection and discovery. *Virus Research* 2019;263:21–26.

1038. **Hayes S, Mahony J, Nauta A, Van Sinderen D**. Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches. *Viruses* 2017;9:127.

1039. **Cantalupo PG, Pipas JM**. Detecting viral sequences in NGS data. *Current Opinion in Virology* 2019;39:41–48.

1040. **Gray GC, Robie ER, Studstill CJ, Nunn CL**. Mitigating Future Respiratory Virus Pandemics: New Threats and Approaches to Consider. *Viruses* 2021;13:637.

1041. **Lloyd-Smith JO, George D, Pepin KM, Pitzer VE, Pulliam JRC, *et al.*** Epidemic dynamics at the human-animal interface. *Science* 2009;326:1362–1367.

1042. **Brucker MC**. Novel Viruses, Zoonotic Infections, and Travel Health. *Nursing for Women's Health* 2020;24:65–66.

1043. **Nordahl Petersen T, Rasmussen S, Hasman H, Carøe C, Bælum J, *et al.*** Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance. *Sci Rep* 2015;5:11444.

1044. **Nieuwenhuijse DF, Oude Munnink BB, Phan MVT, Munk P, Venkatakrishnan S, *et al.*** Setting a baseline for global urban virome surveillance in sewage. *Sci Rep* 2020;10:13748.

1045. **Danko D, Bezdan D, Afshin EE, Ahsanuddin S, Bhattacharya C, *et al.*** A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* 2021;184:3376-3393.e17.

1046. **Dean FB, Nelson JR, Giesler TL, Lasken RS**. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* 2001;11:1095–1099.

1047. **Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, *et al.*** The marine viromes of four oceanic regions. *PLoS Biol* 2006;4:e368.

1048. **Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, *et al.*** Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* 2002;99:14250–14255.

1049. **Froussard P**. A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Res* 1992;20:2900.

1050. **Djikeng A, Halpin R, Kuzmickas R, Depasse J, Feldblyum J, *et al.*** Viral genome sequencing by random priming methods. *BMC Genomics* 2008;9:5.

1051. **Gauthier NPG, Nelson C, Bonsall MB, Locher K, Charles M, *et al.*** Nanopore metagenomic sequencing for detection and characterization of SARS-CoV-2 in clinical samples. *PLoS One* 2021;16:e0259712.

1052. **Liu B, Shao N, Wang J, Zhou S, Su H, *et al.*** An Optimized Metagenomic Approach for Virome Detection of Clinical Pharyngeal Samples With Respiratory Infection. *Frontiers in Microbiology*;11. https://www.frontiersin.org/articles/10.3389/fmicb.2020.01552 (2020, accessed 16 August 2022).

1053. **Gu X, Yang Y, Mao F, Lee WL, Armas F, *et al.*** A comparative study of flow cytometry-sorted communities and shotgun viral metagenomics in a Singapore municipal wastewater treatment plant. *iMeta*;n/a:e39.

1054. **Wang H, Neyvaldt J, Enache L, Sikora P, Mattsson A, *et al.*** Variations among Viruses in Influent Water and Effluent Water at a Wastewater Plant over One Year as Assessed by Quantitative PCR and Metagenomics. *Appl Environ Microbiol* 2020;86:e02073-20.

1055. **Li Y, He X-Z, Li M-H, Li B, Yang M-J, *et al.*** Comparison of third-generation sequencing approaches to identify viral pathogens under public health emergency conditions. *Virus Genes* 2020;56:288–297.

1056. **Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, *et al.*** Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med* 2015;7:99.

1057. Oxford Nanopore releases a novel protocol for the rapid metagenomic characterisation of DNA and RNA viruses. *Oxford Nanopore Technologies*. http://nanoporetech.com/about-us/news/oxford-nanopore-releases-novel-protocol-rapid-metagenomic-characterisation-dna-and (accessed 10 August 2022).

1058. **Simner PJ, Miller S, Carroll KC**. Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases. *Clinical Infectious Diseases* 2018;66:778–788.

1059. **Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, *et al.*** Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat Microbiol* 2019;4:663–674.

1060. **Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G, *et al.*** Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. *Arch Pathol Lab Med* 2017;141:776–786.

1061. **Bharucha T, Oeser C, Balloux F, Brown JR, Carbo EC, *et al.*** STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies. *Lancet Infect Dis* 2020;20:e251–e260.

1062. **Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, *et al.*** Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat Methods* 2017;14:1063–1071.

1063. **Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, *et al.*** Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology*;7. https://www.frontiersin.org/articles/10.3389/fmicb.2016.00459 (2016, accessed 4 July 2022).

1064. **Bokma J, Vereecke N, Pas ML, Chantillon L, Vahl M, *et al.*** Evaluation of Nanopore Sequencing as a Diagnostic Tool for the Rapid Identification of Mycoplasma bovis from Individual and Pooled Respiratory Tract Samples. *J Clin Microbiol* 2021;59:e0111021.

1065. **Vereecke N, Kvisgaard LK, Baele G, Boone C, Kunze M, *et al.*** Molecular epidemiology of Porcine Parvovirus Type 1 (PPV1) and the reactivity of vaccine-induced antisera against historical and current PPV1 strains. *Virus Evolution* 2022;8:veac053.

1066. **Theuns S, Vanmechelen B, Bernaert Q, Deboutte W, Vandenhole M, *et al.*** Nanopore sequencing as a revolutionary diagnostic tool for porcine viral enteric disease complexes identifies porcine kobuvirus as an important enteric virus. *Sci Rep* 2018;8:9830.

ACADEMIC CV

Personal information

Name: Laura Van Poelvoorde

Date of Birth: 08/06/1994

Place of Birth: Knokke, Belgium

Education

2017 – 2023 **Doctor of Science, Biochemistry and Biotechnology**

Ghent University – Gent, Belgium

VIB – Gent, Belgium

Sciensano – Brussels, Belgium

2013 – 2017 **Master of Science, Industrial engineering: Biochemistry**

Ghent University – Gent, Belgium

Graduated magna cum laude

2008 – 2012 **Science-Mathematics**

Sint-Jozefslyceum – Knokke, Belgium

Publications

1. Laura A. E. **Van Poelvoorde**, Thomas Delcourt, Marnik Vuylsteke, Sigrid C. J. De Keersmaecker, Isabelle Thomas, Steven Van Gucht, Xavier Saelens, Nancy Roosens, Kevin Vanneste (2022) A General Approach to Identify Low-Frequency Variants within Influenza Samples collected during Routine Surveillance. Microbial Genomics. **8**. https://doi.org/10.1099/mgen.0.000867

2. Laura A. E. **Van Poelvoorde**, Kevin Vanneste, Sigrid C. J. De Keersmaecker, Isabelle Thomas, Nina Van Goethem, Steven Van Gucht, Xavier Saelens and Nancy Roosens (2022) Whole-Genome Sequence Approach and Phylogenomic Stratification Improve the Association Analysis of Mutations with Patient Data in Influenza Surveillance. Frontiers in Microbiology. **13**. https://doi.org/10.3389/fmicb.2022.809887

3. Tim Boogaerts, Siel Van den Bogaert, Laura A. E. **Van Poelvoorde**, Diala El Masri, Naomi De Roeck, Nancy H. C. Roosens, Marie Lesenfants, Lies Lahousse, Koenraad Van Hoorde, Alexander L. N. van Nuijs and Peter Delputte (2022) Optimization and Application of a Multiplex Digital PCR Assay for the Detection of SARS-CoV-2 Variants of Concern in Belgian Influent Wastewater. Viruses. **14**. https://doi.org/10.3390/v14030610

4. Laura A. E. **Van Poelvoorde**, Mathieu Gand, Marie-Alice Fraiture, Sigrid C. J. De Keersmaecker, Bavo Verhaegen, Koenraad Van Hoorde, Ann Brigitte Cay, Nadège Balmelle, Philippe Herman and Nancy H.C. Roosens (2021) Strategy to Develop and Evaluate a Multiplex RT-ddPCR in Response to SARS-CoV-2 Genomic Evolution. Current issues in Molecular Biology. **43**. https://doi.org/10.3390/cimb43030134

5. Laura A. E. **Van Poelvoorde**¶, Thomas Delcourt¶, Wim Coucke, Phillippe Herman, Sigrid C.J. De Keersmaecker, Xavier Saelens, Nancy H.C. Roosens¶ and Kevin Vanneste¶ (2021) Strategy and Performance Evaluation of Low-Frequency Variant Calling for SARS-CoV-2 Using Targeted Deep Illumina Sequencing. Frontiers in Microbiology. **12**. https://doi.org/10.3389/fmicb.2021.747458

6. Nina Van Goethem, Annie Robert, Nathalie Bossuyt, Laura A. E. **Van Poelvoorde**, Sophie Quoilin, Sigrid C. J. De Keersmaecker, Brecht Devleesschauwer, Isabelle Thomas, Kevin Vanneste, Nancy H. C. Roosens & Herman Van Oyen (2021) Evaluation of the added value of viral genomic information for predicting severity of influenza infection. BMC Infectious Diseases. **21**. https://doi.org/10.1186/s12879-021-06510-z

7. Laura A. E. **Van Poelvoorde**, Bert Bogaerts, Qiang Fu, Sigrid C.J. De Keersmaecker, Isabelle Thomas, Nina Van Goethem, Steven Van Gucht, Raf Winand, Xavier Saelens, Nancy Roosens¶, Kevin Vanneste¶. (2021) Whole-genome-based phylogenomic analysis of the Belgian 2016–2017 influenza A(H3N2) outbreak season allows improved surveillance. *Microbial Genomics*. **7(9)**. https://doi.org/10.1099/mgen.0.000643

8. Laura A. E. **Van Poelvoorde**, Xavier Saelens, Isabelle Thomas, Nancy H. Roosens. (2020) Next-Generation Sequencing: An Eye-Opener for the Surveillance of Antiviral Resistance in Influenza. *Trends in Biotechnology*. **38(4)**. https://doi.org/10.1016/j.tibtech.2019.09.009

¶ equally contributed

Selected meetings

- **Options XI (ISIRV) (2022) –** Belfast, United Kingdom
  Poster presentation: Exploring the added value of Next Generation Sequencing data in influenza surveillance
- **EU Sewage Sentinel System for SARS-CoV-2: Town Hall VI (2021)** – Online
  Poster presentation: Strategy and performance evaluation of low-frequency variant calling for SARS-CoV-2 in wastewater using targeted deep Illumina sequencing
- **ESWI Conference (2020) –** Online
  Poster presentation: Whole genome-based phylogenomic analysis of the Belgian 2016-2017 influenza A(H3N2) outbreak season allows improved surveillance
- **Options X (ISIRV) (2019) –** Singapore
  Poster presentation: Use of whole genome sequencing to improve the influenza surveillance in Belgium during the 2016-2017 pilot season
- **BELVIR Conference (2018) –** Brussels, Belgium
  Oral “shotgun” presentation: Monitoring of influenza: Whole-Genome Sequencing to provide insights into disease severity
- **6th Influenza Meeting (2018) –** Münster, Germany
  Poster presentation: Monitoring of influenza: Whole-Genome Sequencing to provide insights into disease severity

Selected training – Specialist courses

- **Introduction to Git and Github** – Leuven, Belgium
  VIB research training course
- **Basic Statistics in R part II** – online
  VIB research training course
- **Basic Statistics in R part I** – online
  VIB research training course
- **Gentle hands-on introduction to Python programming** – online
  VIB research training course
- **Initiation to Linux command line** – Ghent, Belgium
  VIB research training course
- **Introduction to Protein Structure analysis** – Ghent, Belgium
  VIB research training course
- **Hands-on introduction to NGS** – Leuven, Belgium
  VIB research training course

Selected training – Transferable skills courses

- **How to Get Published** – Online
  Ghent University – Doctoral Schools course
- **Creating effective research posters** – Online
  Ghent University – Doctoral Schools course
- **Effective Scientific Communication** – Online
  Ghent University – Doctoral Schools course
- **Effective Graphical Displays** – Ghent, Belgium
  Ghent University – Doctoral Schools course

Involvement in organizational tasks in the laboratory/ department/faculty/university

- Redaction of standard operating procedures
- Maintenance and management of laboratory instruments

Involvement in training of young scientists, students and technicians

- Providing training in developed protocols to scientists within Sciensano
- Supervision of an intern for three months (2022)