FACULTY OF MEDICINE AND HEALTH SCIENCES

DECIPHERING COMPLETE CANCER TRANSCRIPTOMES FROM BULK TO SINGLE CELL LEVEL

Karen Verboom

Promoter: Prof. dr. Frank Speleman
Co-promoters: Prof. dr. Jo Vandesompele and dr. Kaat Durinck

This thesis is submitted as fulfillment of the requirements for the degree of Doctor in Medical Sciences, 2019

Center for Medical Genetics Ghent

Ghent University Hospital, Medical Research Building 1

Corneel Heymanslaan 10, 9000 Gent, Belgium

Karen.Verboom@UGent.be

GHENT UNIVERSITY

Thesis submitted to fulfill the requirements for the degree of Doctor in Medical Sciences

Promoter

Prof. dr. Frank Speleman (Ghent University, Belgium)

Co-promoters

Prof. dr. Jo Vandesompele (Ghent University, Belgium)

dr. Kaat Durinck (Ghent University, Belgium)

Members of the examination committee

Prof. dr. João Pedro Taborda Barata (University of Lisbon, Portugal)

Prof. dr. Kim De Keersmaecker (KU Leuven, Belgium)

Prof. dr. Tim Lammens (Ghent University, Belgium)

Prof. dr. Katleen De Preter (Ghent University, Belgium)

dr. Wim Trypsteen (Ghent University, Belgium)

dr. Julie Morscio (Ghent University, Belgium)

Prof. dr. Jan Gettemans, chairman (Ghent University, Belgium)

The research in this thesis was conducted at the Center for Medical Genetics Ghent, Ghent University, Ghent, Belgium.

This work was supported by Bijzonder onderzoeksfonds (BOF; PhD grant to Karen Verboom), the Fund for Scientific Research Flanders (FWO; project grants) and the GOA-Ugent.

CONTENTS

ABBREVIATIONS

3C	chromosome conformation capture
4C	circular chromosome conformation capture
5C	3C carbon-copy
ALL	acute lymphoblastic leukemia
AML	acute myeloid leukemia
ASO	antisense oligonucleotide
ATAC-seq	assay for transposase accessible chromatin followed by high throughput sequencing
ATP	adenosine triphosphate
B-ALL	B-cell acute lymphoblastic leukemia
BET	bromodomain and extra-terminal motif
bHLH	basic helix-loop-helix
CAGE-seq	single cell cap analysis gene expression sequencing
CAR-T	chimeric antigen receptor T-cell
CD	cluster of differentiation
cDNA	complementary DNA
CEL-seq	cell expression by linear amplification and sequencing
CHART	capture hybridization analysis of RNA targets
ChIP-seq	chromatin immunoprecipitation sequencing
ChIRP	chromatin isolation by RNA purification
circRNA	circular RNA
CITE-seq	cellular indexing of transcriptome and epitope by sequencing
CLL	chronic lymphoblastic leukemia
CML	chronic myeloid leukemia
CNS	central nervous system
CNV	copy number variation
CRISPR	clustered regulatory interspaced short palindromic repeats
CRISPRi	clustered regularly interspaced short palindromic repeats interference
CROP-seq	CRISPR droplet sequencing
CTC	circulating tumor cell
CUT&RUN	cleavage under targets and release using nuclease
DamID	DNA adenine methyltransferase identification
dCas9	death Cas9
dChIRP	domain specific ChIRP
DN	double negative
DNA	deoxyribonucleic acid
DP	double positive
EpCam	epithelial cell adhesion molecule
ERCC	external RNA controls consortium
eRNA	enhancer RNA
ETP	early T-cell progenitor
FACS	fluorescence in situ hybridization
FBS	fetal bovine serum
FFPE	formalin-fixed paraffin-embedded
FISH	fluorescent in situ hybridization
FISSEQ	fluorescent in situ sequencing
Fucci	fluorescence ubiquitination-based cell cycle indicator
G&T-seq	genome and transcriptome sequencing
GEO	gene Expression Omnibus
GLM	generalized linear model

gRNA	guide RNA
GRO-seq	global run-on sequencing
GSEA	gene set enrichment analysis
GSI	gamma secretase inhibitor
HPC	high performance computing
HSC	hematopoietic stem cell
ICN	intracellular NOTCH1
ISP	Immature single positive
IVT	in vitro transcription
lincRNA	long intergenic non-coding RNA
LNA	locked nucleic acid
lncRNA	long non-coding RNA
MAD	median absolute deviation
MDA	multiple displacement amplification
MERFISH	multiplexed error robust FISH
MHC	major histocompatibility complex
miRNA	micro RNA
miscRNA	miscellaneous RNA
mRNA	messenger RNA
ncRNA	non-coding RNA
NGS	next generation sequencing
NK	natural killer
NSCLC	non-small-cell-lung carcinoma
nt	nucleotides
ORF	open reading frame
PacBio	pacific Biosciences
PCA	principal compound analysis
PCR	polymerase chain reaction
PDX	patient derived xenograft
PIM	probabilistic index model
PRC2	polycomb repressive complex
PRO-seq	precision nuclear run-on sequencing
RAP	RNA antisense purification
REAP-seq	RNA expression and protein sequencing
RISC	RNA-induced silencing complex
RNA	ribonucleic acid
RNAi	RNA interference
RNAPII	RNA polymerase II
RNA-seq	RNA sequencing
rRNA	ribosomal RNA
RT-qPCR	reverse transcription quantitative polymerase chain reaction
sci-MET	single cell combinatorial indexing for methylation sequencing
ScISOr-seq	Single-cell isoform RNA sequencing
scLVM	single-cell latent variable model
SCRB-seq	single cell RNA barcoding and sequencing
seqFISH	sequential FISH
siRNA	short interfering RNA
smFISH	single molecule FISH
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
snRNA	small nuclear RNA

sn-seq	single nucleus sequencing
SNV	single nucleotide variation
SP	single positive
STRT-seq	single cell tagged reverse transcription sequencing
T-ALL	T-cell acute lymphoblastic leukemia
TCR	T-cell receptor
TLX1	T-cell leukemia homeobox 1
Treg	regulatory T-cell
tRNA	transfer RNA
tSNE	T-distributed stochastic neighbor embedding
TSS	transcription start site
UMI	unique molecular identifier
WGA	whole genome amplification

SUMMARY

T-cell acute lymphoblastic leukemia (T-ALL) is an aggressive type of blood cancer that results from malignant transformation of precursor T-cells. Several gain-of-function and loss-of-function mutations in oncogenes and tumor suppressor genes have been described, cooperating to fully transform normal developing T-cells into lymphoblasts. Gene expression profiling of T-ALL patients revealed several driver oncogenes, demarcating distinct molecular subgroups. One of these drivers is '*T-cell leukemia homeobox 1*' (*TLX1*), which defines a molecular T-ALL subgroup, characterized by a specific gene expression pattern and a T-cell developmental arrest at the early cortical stage. Although this subgroup is characterized by a favorable prognosis, still a large fraction of the T-ALL patients relapse and current therapies are associated with acute and long-term toxicities. Hence, genetic alterations in T-ALL should be further investigated to develop more effective and less toxic therapies. Therefore, the network downstream of TLX1 has already been studied extensively in terms of protein-coding genes, however, the downstream long non-coding RNA (lncRNA) network, remained thus far unexplored. In this PhD research, I revealed TLX subgroup specific and more specifically, TLX1 regulated lncRNAs.

To identify these specific lncRNAs, I performed polyA[+] and total RNA-seq of TLX1 positive ALL-SIL lymphoblasts upon *TLX1* knockdown and a large primary T-ALL cohort. By further integrating ATAC-seq, H3K4me1, H3K4me3, H3K27ac and TLX1 ChIP-seq of ALL-SIL lymphoblasts, I identified known as well as novel TLX1 regulated and super-enhancer associated lncRNAs. I extended the T-ALL cohort with lncRNA expression data of a progenitor T-cell subset to determine possibly oncogenic lncRNAs. Since lncRNAs may serve as excellent therapeutic targets owing to their tissue specific expression and low abundance, these potentially oncogenic lncRNAs require further functional characterization to validate if these can serve as new therapeutic targets in T-ALL. Since I generated a comprehensive and unique dataset in the T-ALL field that contains extensive unexplored information, I wrote a data descriptor to share the data with the research community, ensuring re-usability of the dataset.

Although these bulk sequencing experiments provide comprehensive information about TLX subgroup specific and TLX1 regulated lncRNAs, average expression profiles are generated, which can mask subtle differences among cells. To unravel this heterogeneity, single cell RNA-seq devices were developed. I first developed a single cell total RNA-seq protocol enabling to capture polyadenylated as well as non-polyadenylated transcripts, since virtually all methods at that time only sequenced polyadenylated transcripts, ignoring the vast non-polyadenylated part of the transcriptome. Using my method, more genes can be identified compared to classic single cell polyA[+] RNA-seq methods. Furthermore, I showed that my method can detect circular RNAs, which lack a polyA tail, and novel genes and recapitulate the expected biological signal after perturbation. Since my protocol works on Fluidigm's C1 and flow cytometry sorted cells, it is widely applicable. However, the throughput of these two devices is still relatively low compared to the latest developed single cell RNA-seq devices that can capture thousands to tens of thousands of cells. Several studies compared these devices with respect to data quality and the ability to distinguish cellular subpopulations, however none of these comparative studies investigated the heterogeneity of the cellular transcriptional response upon chemical perturbation. Therefore, I compared the C1 (Fluidigm), ddSeq (Bio-Rad, Illumina) and Chromium (10x Genomics) in terms of data quality and their ability to detect differentially expressed genes and putative transcriptional heterogeneity. I revealed that despite the lower number of differentially expressed genes in single cell RNA-seq experiments compared to bulk RNA-seq experiments, the biological signal can be detected by gene set enrichment analysis for all single cell devices. Furthermore, I showed that single cell RNA-seq analyses enable to reveal heterogeneity in the

response on nutlin-3 treatment and to identify potentially late-responders or resistant cells, which are hidden in bulk RNA-seq experiments and require further in depth investigation.

In conclusion, I identified a set of TLX1 regulated and TLX subgroup specific lncRNAs, of which some are potentially oncogenic, marking them as highly interesting targets for further in depth characterization. Furthermore, I developed a single cell total RNA-seq protocol that for the first time combines strandedness and effective removal of ribosomal cDNA and enables the detection of both polyadenylated and non-polyadenylated transcripts, including lncRNAs, circRNAs and novel genes. Finally, I performed an in depth evaluation of the C1, ddSeq and Chromium single cell devices and showed that detection of the most abundant genes in single cell experiments is sufficient to faithfully detect biological signal through genes set enrichment analysis and may help to identify potentially late-responders or resistant cells upon compound treatment.

SAMENVATTING

T-cel acute lymfoblastische leukemie (T-ALL) is een agressieve bloedkanker die ontstaat door een maligne transformatie van voorloper T-cellen. Er zijn reeds verschillende *gain-of-function* en *loss-of-function* mutaties in oncogenen en tumor suppressor genen beschreven, die samenwerken om zich normaal ontwikkelende T-cellen te transformeren tot lymfoblasten. Genexpressie profilering van T-ALL patiënten onthulde verschillende *driver* oncogenen die moleculaire subgroepen aflijnen. Een van deze *driver* genen is '*T-cell leukemia homeobox 1*' (*TLX1*), die een moleculaire subgroep afbakent met een specifiek genexpressie patroon en een arrest vertoont in het vroege corticale stadium van de T-cel ontwikkeling. Hoewel deze subgroep gekenmerkt wordt door een relatief goede prognose, hervallen nog steeds veel T-ALL patiënten en leiden de huidige therapieën tot acute en lange termijn toxiciteit. Het is daarom belangrijk om de genetische defecten in T-ALL verder te onderzoeken, om zo effectievere en minder toxische therapieën te kunnen ontwikkelen. Hoewel het eiwit-coderende netwerk *downstream* van TLX1 reeds uitgebreid onderzocht werd, bleef het geassocieerde lange niet-coderende RNA (lncRNA) netwerk tot nu toe grotendeels onbestudeerd. Daarom heb ik in de context van mijn PhD onderzoek gedaan naar TLX subgroep geassocieerde lncRNAs en meer specifiek naar TLX1 gereguleerde lncRNAs.

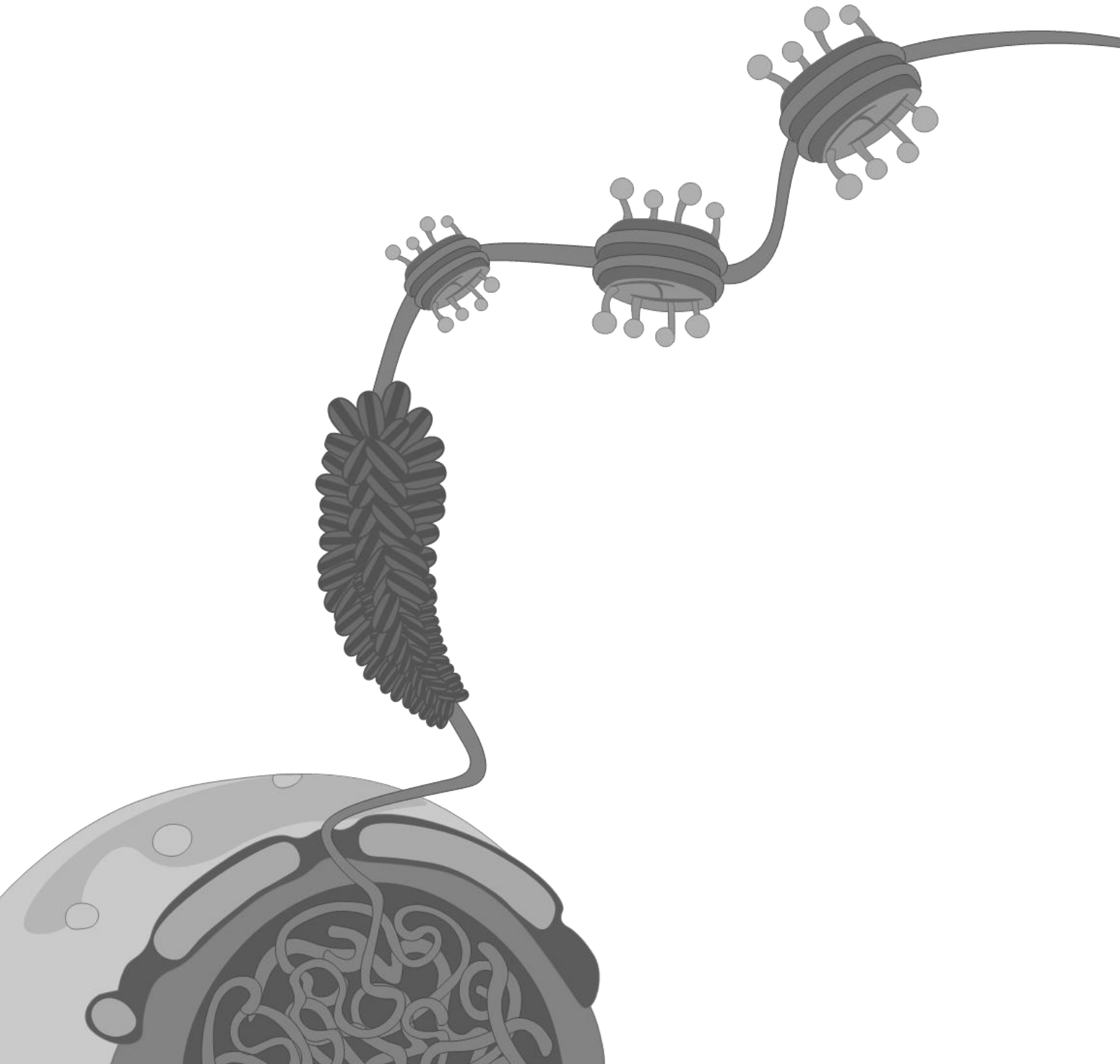
Om deze lncRNAs te identificeren heb ik polyA[+] en total RNA-seq uitgevoerd op TLX1 positieve ALL-SIL lymfoblasten na *knockdown* van *TLX1* en op een grote T-ALL patiënten cohorte. Na integratie van ATAC-seq en H3K4me1, H3K4me3, H3K27ac en TLX1 ChIP-seq data, gegenereerd op ALL-SIL lymfoblasten, heb ik gekende en nieuwe TLX1 gereguleerde en super-enhancer geassocieerde lncRNAs geïdentificeerd. Vervolgens heb ik deze primaire T-ALL dataset uitgebreid met lncRNA expressie data van een voorloper T-cel subset, waardoor ik potentieel oncogene lncRNAs kon identificeren. Aangezien lncRNAs als excellente therapeutische targets kunnen dienen door hun hoge weefsel-specificiteit en lage expressie, zouden de potentieel oncogene lncRNAs verder functioneel onderzocht moeten worden, om na te gaan of deze effectief gebruikt kunnen worden als nieuwe therapeutische targets in T-ALL. Aangezien ik een uitgebreide en unieke dataset in het T-ALL onderzoeksveld heb gegenereerd, die informatie bevat die nog niet onderzocht werd, heb ik een data descriptor geschreven om zo de data te delen met de onderzoekswereld, zodat deze hergebruikt kan worden.

Hoewel deze 'bulk' experimenten uitgebreide informatie bevatten over TLX-subgroep specifieke en TLX1 gereguleerde lncRNAs, werden er gemiddelde expressie profielen gegenereerd, wat kleine verschillen tussen cellen kan maskeren. Om deze heterogeniteit te ontrafelen, werden toestellen voor *single cell* RNA-seq ontwikkeld. Ik heb eerst een *single cell* total RNA-seq protocol ontwikkeld dat zowel gepolyadenyleerde als niet-gepolyadenyleerde transcripten kan detecteren, aangezien zo goed als al de bestaande methoden tot dan toe enkel gepolyadenyleerde transcripten sequeneerden, waarbij het niet-gepolyadenyleerde deel van het transcriptoom dus niet opgepikt werd. Met de methode ontwikkeld tijdens mijn PhD kunnen meer genen gedetecteerd worden in vergelijking met de klassieke *single cell* polyA[+] RNA-seq methodes. Daarenboven toonde ik aan dat deze nieuwe methode nieuwe genen en circulaire RNAs (circRNAs), die geen polyA staart hebben, kan detecteren en het verwachte biologische signaal na perturbatie kan oppikken. Aangezien mijn methode op C1 (Fluidigm) en flow-cytometrie gesorteerde cellen werkt, is de methode wijd toepasbaar. Een nadeel is dat de *throughput* van deze twee toestellen relatief laag is in vergelijking met de laatst ontwikkelde *single cell* RNA-seq methodes, die meer dan tienduizend cellen kunnen isoleren. In verschillende studies werden de kwaliteit van de data en het vermogen van deze methoden om cel subpopulaties te onderscheiden vergeleken, terwijl geen enkele van deze vergelijkende studies de heterogeniteit van de transcriptionele respons op een chemische perturbatie heeft bestudeerd. Daarom heb ik de C1 (Fluidigm), ddSeq (Bio-Rad, Illumina) en Chromium (10x Genomics) vergeleken met betrekking tot de

datakwaliteit en hun vermogen om differentieel geëxpresseerde genen en transcriptionele heterogeniteit te detecteren. Ik heb aangetoond dat ondanks dat er minder differentieel geëxpesseerde genen gedetecteerd worden in *single cell* RNA-seq experimenten in vergelijking met bulk experimenten, het biologisch signaal toch gedetecteerd kan worden met zogenaamde *gene set enrichment* analyse voor de drie *single cell* methodes. Ik heb ook aangetoond dat *single cell* RNA-seq analyses de heterogeniteit in respons op nultin-3 behandeling kunnen onthullen en cellen kunnen detecteren die resistent zijn of vertraagd reageren op de therapie, wat niet mogelijk is met bulk RNA-seq en verder onderzoek vraagt.

Samenvattend heb ik een set van TLX1 gereguleerde en TLX-subgroep specifieke lncRNAs geïdentificeerd, waarvan sommige mogelijks oncogeen zijn, waardoor ze interessante targets vormen voor verdere functionele karakterisatie. Daarnaast ontwikkelde ik ook een *single cell* total RNA-seq protocol dat voor het eerst de informatie van de DNA streng behoudt en zorgt voor een effectieve verwijdering van het ribosomaal cDNA en de detectie toelaat van zowel gepolyadenyleerde als niet-gepolyadenyleerde transcripten, zoals lncRNAs en circRNAs, en nieuwe genen. Ten slotte heb ik een gedetailleerde evaluatie van de C1, ddSeq en Chromium *single cell* toestellen uitgevoerd en aangetoond dat de detectie van de meest abundante genen in *single cell* experimenten voldoende is om het biologisch signaal op te pikken door middel van *gene set enrichment* analyse, wat kan bijdragen tot de detectie van cellen die resistent zijn of vertraagd reageren op therapie.

1. Introduction



1.1. T-cell acute lymphoblastic leukemia

1.1.1. Leukemia

Leukemia is a broad term for blood and bone marrow cancers and is the most common cancer in children younger than 15 years (1). Leukemia can be subdivided in lymphoblastic and myeloid leukemia depending on the type of blood cells that is transformed. In addition, leukemias are further subdivided in an acute form, which develops rapidly and results in the accumulation of immature blood cells in the bone marrow, and a chronic form that develops slowly and leads to an accumulation of more mature blood cells. Based on this distinction, leukemias can be subdivided in four major categories: acute lymphoblastic leukemia (ALL), chronic lymphoblastic leukemia (CLL), acute myeloid leukemia (AML) and chronic myeloid leukemia (CML) (2, 3). CLL, AML and CML most frequently occur in adults, while ALL is mostly diagnosed in children (3). ALL accounts for 80 % of all leukemia cases in children, with a peak incidence between two and five years and is the most frequent cause of death from cancer before the age of 20 (1, 4, 5). ALL can be further subdivided in T-cell acute lymphoblastic leukemia (T-ALL) and B-cell acute lymphoblastic leukemia (B-ALL) with a malignant transformation of T-cell and B-cell progenitors, respectively. T-ALL, studied in this PhD thesis, accounts for 10% - 15% of pediatric and 25% of adult ALL cases and is characterized by an arrest during T-cell development (6, 7).

1.1.2. T-cell development

T-lymphocytes are immune cells with an important role in the adaptive immune response to protect the body from infections. Normal T-cell development is a strictly regulated and hierarchical process in which hematopoietic progenitor cells develop to mature T-cells and involves several steps of proliferation, maturation, differentiation and selection (**Figure 1**). First, the hematopoietic progenitor cells migrate from the bone marrow to the thymus via the bloodstream (8). These early T-cell progenitor (ETP) cells start to proliferate and express the stem cell marker cluster of differentiation 34 (CD34) (9, 10). Since the CD4 and CD8 receptors are not expressed on the membrane of these cells, this stage is called the double negative (DN) stage. The DN stage comprises four consecutive development stages with different expression patterns of *CD44* and *CD25* (11). During the DN stage, cells proliferate and T-cell specification is acquired via high *NOTCH1* expression levels (12). The latter evokes a definitive block in the development into the B-cell lineage and induces initiation of T-cell development, but is however not sufficient for final T-cell commitment as natural killer (NK) cells can still develop from these cells (12, 13). To achieve full T-cell commitment, *GATA3* expression is required to initiate the gene expression program required for T-cell development (14). At this point, the expression levels of *NOTCH1* will determine if a cell further develops into the αβ or γδ lineage. Only a small subset of these cells will undergo rearrangements of the γδ chains under influence of high *NOTCH1* levels, while the majority of the cells will develop into the αβ lineage, which requires a drop in *NOTCH1* expression (15). Completion of the T-cell commitment is marked with CD1a expression (16). To further develop into a mature T-cell, CD4 is expressed during the immature single positive (ISP) stage of human T-cell development, in absence of CD3 expression, which is strongly expressed in mature single positive T-cells (17, 18). Next, T-cell receptor (TCR) genes need to be rearranged to obtain functional TCRs that can detect antigens presented by the major histocompatibility complex (MHC). First, a pre-TCR is generated and lymphocytes that fail to generate a functional pre-TCR undergo apoptosis (β-selection). Next, the lymphocytes express CD4 and CD8 (double positive cells, DP) and the α-chain is rearranged, generating the TCRαβ (8, 10, 12, 19). Only cells with TCRs that interact with intermediate avidity to the MHCs survive (positive selection). Furthermore, cells that interact too strongly with self-antigens are eliminated by apoptosis (negative selection), preventing auto-immunity. Finally, cells further differentiate into the single positive mature stage, expressing

either CD4 or CD8 (single positive cells, SP) (8, 10, 12, 19). Hereafter, naïve T-cells migrate to the bloodstream and peripheral immune organs to perform their function. T-cells are activated by the presentation of foreign antigens by the MHC. After activation, CD8 positive T-cells further differentiate into cytotoxic T-cells that detect and kill infected cells, while CD4 positive cells can differentiate in T-helper cells that stimulate other cells of the immune system. Finally, regulatory T-cells suppress the activity of other lymphocytes to tightly control the immune response (19). Since T-cell differentiation is a highly regulated process including several proliferation, differentiation and selection steps, in which different transcription factors and signal pathways are involved, deregulation of these processes can lead to the development of T-ALL (10, 20).

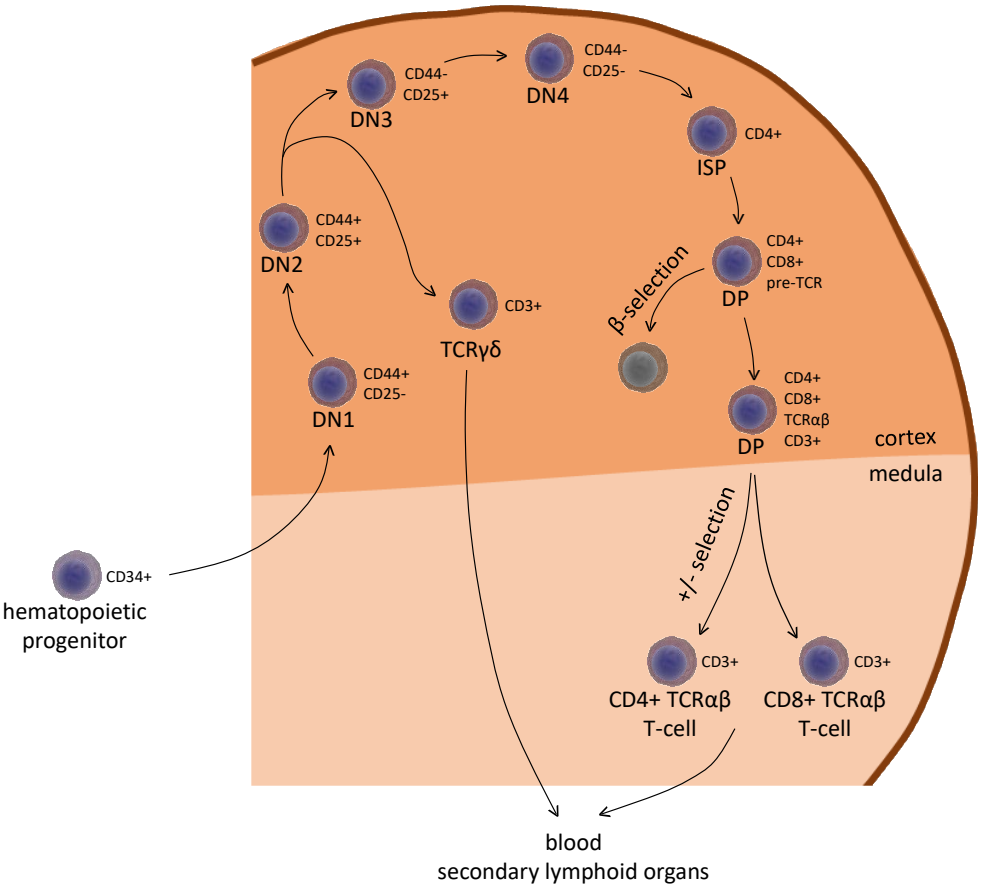


Figure 1: T-cell development in the thymus. Development of hematopoietic progenitors to naïve T-cells in the thymus. Mature CD4+ TCRαβ, CD8+ TCRαβ and TCRγδ T-cells migrate to the blood and secondary lymphoid organs. DN: double negative; ISP: immature single positive; DP: double positive; TCR: T-cell receptor; ISP: immature single positive. Figure adapted from (8) and (38).

1.1.3. **Symptoms, diagnosis and prognosis of T-ALL**

The symptoms displayed by T-ALL patients are non-specific and result from a decrease in normal hematopoiesis due to an increase of lymphoblasts in the bone marrow. This leads to bleedings and bruises due to a reduction in platelet counts, anemia due to a reduction in red blood cells and infections due to neutropenia (4). T-ALL patients often have a large tumor burden and present with mediastinal thymic masses, very high circulating blast cell counts, and infiltration of the central nervous system, resulting in headaches and nausea, at the time of diagnosis (21, 22). Genetic factors, such as Down and Bloom syndrome, are associated with an increased risk for T-ALL development, however, most patients have no inherent factors (4, 23, 24). The current diagnosis is based on cell morphology, immunophenotype and genetics of peripheral blood and bone marrow and a lumbar puncture to

determine potential leukemic infiltration in the CNS is carried out as well. It is for instance possible to distinguish ALL from AML based on cell morphology, while flow cytometry is warranted to distinguish T-ALL from B-ALL (25). The prognosis is mainly determined by clinical and genetic features at diagnosis and by early response to the treatment (4). It is known that children between one and nine years old have the best prognosis, while the prognosis gets worse with increasing age for adults. Children younger than 12 months generally have a bad prognosis. Furthermore, high leukocyte counts also result in a more poor outcome (5). Currently, the survival rate of pediatric T-ALL cases is close to 90 %, however, adult cases still have a poor prognosis (26, 27). Furthermore, current standard therapies give short-term and long-term side-effects and still up to 20 % of the pediatric and 40 % of the adult T-ALL patients relapse, warranting for a more profound insight into the molecular basis of T-ALL, to enable the development of innovative and more effective targeted therapeutic strategies (5, 28).

1.1.4. **Common genetic alterations in T-ALL**

T-ALL is a multistep oncogenic process in which genetic aberrations accumulate over time, resulting in a differentiation arrest during T-cell development and consequently proliferation of immature T-cells. The emergence of next generation sequencing (NGS) technologies has expanded our knowledge on the molecular basis of T-ALL (29–35). Recent studies have shown that on average 10-20 protein-coding alterations are detected in a T-ALL cell, which are assumed to cooperate to fully transform thymocytes into lymphoblasts (29, 31, 33, 36). These genetic alterations affect several biological processes, including cell cycle, self-renewal capacity, TCR signaling, and activation of tyrosine kinases during thymocyte development (37, 38). The cell cycle is a tightly controlled process that comprises several checkpoints to maintain genomic integrity and defects in these checkpoints can result in apoptosis or tumor formation. The most common genetic defect in T-ALL is an inactivation of the *CDKN2A* locus that occurs in up to 90 % of the T-ALL patients. *CDKN2A* encodes both the p14/ARF and p16 proteins involved in pRB1 phosphorylation and TP53 activation, hindering cell cycle entry and TP53 controlled cell cycle inhibition and apoptosis (37, 39, 40). *CDKN2A* is mostly inactivated due to a deletion of the 9p21 locus, in which *CDKN2B/p15* is often co-deleted (37, 40, 41). However, *CDKN2A/B* inactivation may also result from cryptic deletions, inactivating mutations or promoter hypermethylation (37, 40, 42–45). Acquiring self-renewal capacity is one of the hallmarks of cancer and is in T-ALL often acquired through a constitutive activation of NOTCH1, a key factor in T-cell development (46). Upon activation, NOTCH1 is cleaved and intracellular NOTCH1 (ICN1) is translocated to the nucleus to activate target genes (37, 47). A rare translocation (t(7;9)(q34;q34)) activates *NOTCH1* by juxtaposing *NOTCH1* to TCRβ regulatory sequences, but is only found in < 1 % of the T-ALL patients (48). Most frequently, *NOTCH1* is activated by activating mutations in the heterodimerization domain, the PEST domain or both, resulting in an increased stability of ICN or ligand independent activation of the NOTCH1 receptor (46). These activating mutations result in a constitutive active NOTCH1, increasing the transcription of target genes that are involved in several functions, including self-renewal capacity and the regulation of cell cycle, cell growth and cell survival. NOTCH1 regulates the cell cycle through direct upregulation of *Cyclin D3*, *CDK4* and *CDK6,* which contributes to the deregulation of G1/S cell cycle progression and proliferation in T-ALL cells (49). G1/S transition is further promoted by NOTCH1 through upregulation of the cell cycle regulator *SKP2* that negatively regulates the CDK inhibitor *p27/KIP* (50). Identification of NOTCH1 direct target genes revealed that many target genes, including *c-MYC*, *IL7R* and *IGF1R*, are involved in the regulation of cell growth (51–53). NOTCH1 also promotes cell survival via activation of the NF-kB pathway via several mechanisms, including binding on *NF-kB2* and *RELB* and inducing their transcription and by directly increasing the activity of the IKK complex (54, 55). Besides *NOTCH1* mutations, inactivating mutations or deletions in the *FBXW7* tumor suppressor gene also result in an increased NOTCH1 protein stability as *FBXW7* normally degrades NOTCH1 (56). Activating *NOTCH1*

mutations are detected in more than 60 % of the T-ALL cases, marking NOTCH1 as a potential important therapeutic target (57–59). Gamma secretase inhibitors (GSI) inhibit the proteolytic cleavage of NOTCH1 and consequently the constitutive activation of NOTCH1 target genes. Unfortunately, this therapy suffers from gastro-intestinal toxicity and low efficiency (57). However, recently it has been shown that specifically targeting the PSEN1 gamma secretase results in high anti-leukemic activity and avoids gastro-intestinal toxicity, providing optimism for further clinical testing (60). *PHF6* is another gene that is frequently mutated in T-ALL. Inactivating mutations or deletions in PHF6 are identified in 16 % and 38 % of the pediatric and adult T-ALL patients, respectively, and are associated with TLX1 and TLX3 overexpression (31, 32, 35). Moreover, it has recently been suggested that *PHF6* mutations increase the self-renewal capacity of hematopoietic stem cells (HSC), however this requires further investigation (61–64). During T-ALL development, several genes involved in TCR signaling are mutated or targeted for chromosomal rearrangement. For instance, *LCK* is a tyrosine kinase that is highly expressed during T-cell development and can be ectopically expressed in T-ALL due to the t(1;7)(p34;q34), placing *LCK* in the neighborhood of TCRβ (65). Also other genes involved in TCR signaling, such as *RAS* and *PTEN*, are frequently mutated in T-ALL, resulting in proliferation and survival of immature thymocytes (40, 66, 67). Besides *LCK,* also other tyrosine kinases are often activated in T-ALL. The *ABL1* tyrosine kinase is typically fused with *BCR* by the Philadelphia translocation t(9;22) (q34;q11) in chronic myeloid leukemia and B-cell acute lymphoblastic leukemia (68, 69). In contrast, while this fusion is rare in T-ALL, *ABL1* is activated in 6 % of the cases by a *NUP214-ABL1* fusion, resulting in a constitutive active *ABL1*, and consequently in proliferation and survival (70). Other tyrosine kinases, such as *JAK2* and *FLT3*, can also be activated resulting in consecutive tyrosine kinase activity and proliferation of T-cells (71, 72).

1.1.5. **Molecular subgroups in T-ALL**

T-ALL patients can be subdivided in molecular subgroups based on the overexpression of specific transcription factors (73). These transcription factors are typically activated by four mechanisms: (a) chromosomal translocations involving promoters or enhancers of TCR genes, (b) chromosomal rearrangements with other regulatory sequences, (c) duplications / amplifications and (d) mutations or small insertions generating novel regulatory sequences acting as enhancers (36). The subgroups each have a specific gene expression profile and are characterized by a fixed differentiation arrest during T-cell development (**Figure 2**). Using microarray expression analysis, Ferrando et al. showed that T-cell leukemia homeobox 1 (TLX1,HOX11), TAL-R and LYL1 positive T-ALLs have specific gene expression patterns with a differentiation arrest in the early cortical, late cortical and pro-T-cell stage, respectively (6). Clustering analysis revealed a fourth group, the TLX3 subgroup, that has a similar expression profile as TLX1, but lacks *TLX1* expression (6). As *TLX1* and *TLX3* have a similar gene expression signature and induce T-ALL in a similar way, TLX1+ and TLX3+ patients are often grouped as one subgroup (74). Another study used FISH with probes for TCRβ to identify new translocation partners of TCRβ, known to be often involved in translocations in T-ALL. This revealed *HOXA* as a new translocation partner of TCRβ, placing genes of the HOXA cluster in the vicinity of TCRβ regulatory elements and defining a new homogenous subgroup with an arrest in pre-T-cell stage (75). These four subgroups are described below with a focus on the TLX1/TLX3 subgroup as this subgroup is investigated in more depth in this PhD thesis.

1.1.5.1. TAL-R

The TAL-R subgroup is predominantly characterized by *TAL1* or *LMO2* expression. TAL1 is a basic helix-loop-helix (bHLH) protein and is often ectopically expressed due to juxtaposition to regulatory elements of TCRα/δ or TCRβ by t(1;14)(p32;q11) and t(1;7)(p32;q35), respectively. Later, also a 1p35

deletion, placing *TAL1* under control of the *STIL* promoter, had been identified (37, 76). Furthermore, *TAL1* can also be overexpressed by a noncoding mutation upstream of the *TAL* locus, creating a super-enhancer with a new *MYB* binding site, activating *TAL1* expression (77). GATA3 and RUNX1 form a core regulatory circuit with TAL1 by binding and positively regulating its own and each other's gene expression and contributing to the development of TAL1 positive T-ALL (78). For its function, TAL1 binds to the LMO family, forming a transcriptional complex that inhibits E2A function. E2A is required for the proper expression of *CD4* and *CD5* genes during the early stages of T-cell development and inhibition of *E2A* consequently results in a differentiation arrest (79). *LMO2* can also be ectopically expressed by translocations (t(11;14)(p13;q11) and t(7;11)(q35;p13)) to the TCRα/δ or TCRβ locus or by a deletion (del(11)(p12p13)) resulting in the loss of a negative regulatory element of *LMO2*, thus activating the proximal *LMO2* promoter (80–83).

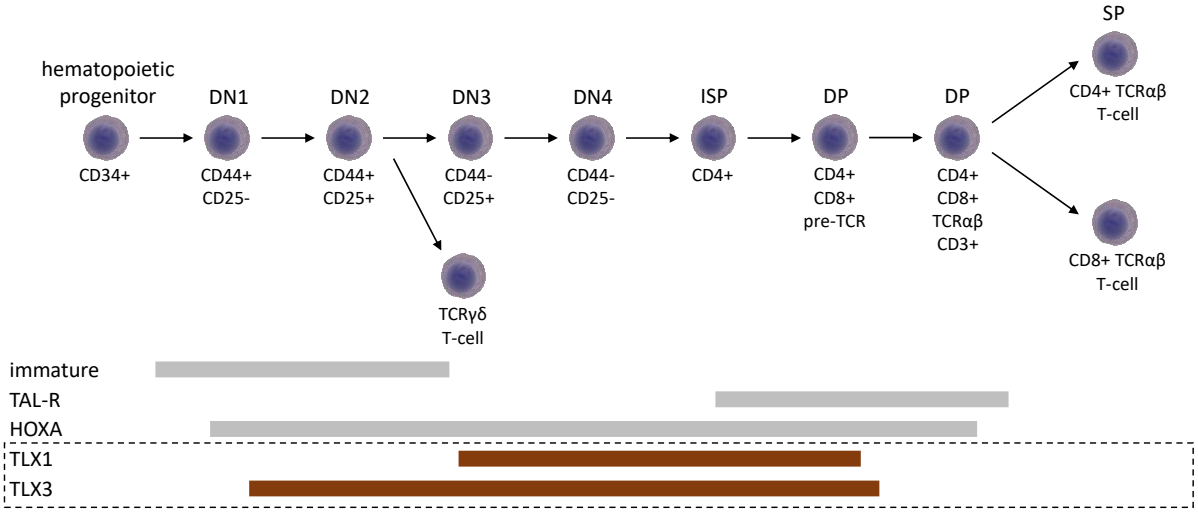


Figure 2: schematic overview of molecular T-ALL subgroups in relation to their T-cell development stage. T-ALL can be subdivided in four subgroups, immature, TAL-R, HOXA and TLX (TLX1/TLX3), each with a specific gene expression signature and an arrest at a specific stage of the T-cell development. DN: double negative; DP: double positive; SP: single positive; CD: cluster of differentiation; TCR: T-cell receptor; ISP: immature single positive. Figure adapted from (6, 448, 449).

1.1.5.2. HOXA

The HOXA subgroup is characterized by overexpression of members of the HOXA cluster, typically due to an inversion on chromosome 7, a PICALM-MLL10 rearrangement, MLL fusions or SET-NUP214 fusions (37, 84). The inversion (inv(7)(p15q35)) juxtaposes the HOXA cluster to the TCRβ enhancer, thereby activating *HOXA* genes (in particular *HOX10* and *HOX11*). In contrast, the PICALM-MLL10 (t(10;11)(p13;q14)) translocation results in the recruitment of hDOT1L, a H3K79 methyl transferase that methylates and activates genes of the HOXA cluster (37, 84–87). Likewise, MLL-MLLT1/ENL and MLL-MLLT10/AF10 fusions also recruit hDOT1L resulting in consecutive methylation and activation of *HOXA* genes. The SET-NUP214 fusion results from a deletion (del(9)(q34.11q34.13)) and acts as a transcriptional co-factor to activate *HOXA* genes (37, 88, 89). As hDOT1L is often involved in the activation of *HOXA* genes, it may be an interesting therapeutic target for this subgroup (37).

1.1.5.3. Immature subgroup

The immature subgroup is characterized by a high expression of *LMO2*, *LYL1*, stem cell marker *CD34*, *BCL2* and a frequent expression of the myeloid markers *CD13* and *CD33*. This subgroup has a differentiation arrest in the early T-cell development (90, 91). Driver genomic rearrangements have not been identified and this subgroup has in general a bad prognosis (6, 37). This can partially be explained by the high expression of BCL2, as BCL2 is an anti-apoptotic protein which can prevent that

drug-induced damage results in cell death. This enables treated cells to survive, which may explain the resistance to chemotherapeutic agents for this subgroup (6, 92). A part of the immature T-ALL patients display a molecular signature corresponding to a T-cell development differentiation arrest at the ETP stage, reflecting cells migrating from the bone marrow to the thymus. This subgroup is characterized by a lack of *CD1a* and *CD8* expression, a weak expression of *CD5* and expression of one or more myeloid and/or stem-cell markers (*CD117, CD34, HLA-DR, CD13, CD33, CD11b* and/or *CD65*). This class also has a higher expression of *CD34, LYL1, LMO1* and *ERG* and is characterized by more copy number alterations compared to the other subgroups (29). Whole genome sequencing revealed alterations in epigenetic regulators (e.g. *DNMT3A*, *IDH2*), signaling factors (e.g. *RAS*, *FLT3*) and genes involved in hematopoietic development (e.g. *GATA*, *RUNX1*). Of interest, 48 % of the ETP-ALL cases have alterations in genes encoding components of the polycomb repressive complex (PRC2), such as *SUZ12* and *EZH2*, disrupting PRC2 mediated gene silencing. Of note, *CDKN2A/B* mutations are less frequently identified in this subgroup (29, 31). This ETP subgroup was initially characterized by a poor outcome, however, recent research suggests that intensified therapies can overcome this poor outcome (29, 38, 93).

1.1.5.4. TLX1/TLX3

T-cell leukemia homeobox 1 (*TLX1, HOX11*) is a homeobox gene that is essential for nervous system and spleen development, as it has been shown that *Hox11*-/- mice show asplenia (94–96). *TLX1* is not expressed in developing thymocytes, whereas ectopic expression of *TLX1* in lymphoblasts, caused by *TLX1* juxtaposition to the regulatory elements of TCRβ or TCRα/δ as a consequence of the t(7;10)(q35;24) and t(10;14)(q24;q11) chromosomal translocations, respectively, disrupts T-cell differentiation and initially leads to a drastic decrease in the absolute number of thymocytes. This indicates that *TLX1* overexpression affects T-cell differentiation, proliferation and survival (97–99). TLX1 is a driver in T-ALL development and is ectopically expressed in 5-10 % of pediatric and 30 % of adult T-ALL patients (73, 95, 100). Besides chromosomal translocations, ectopic *TLX1* expression can also be caused by other genetic defects, such as subtle mutations in *cis* regulatory elements, deregulation of trans-factors that regulate *TLX1* expression and methylation status of the promoter (95). For the latter, it has been shown that demethylation of the proximal promoter of *TLX1* results in *TLX1* expression and that this promoter demethylation is detected in TLX1 positive patients with and without translocations (101). Furthermore, *TLX1* overexpression defines a molecular subgroup with a gene expression profile that is indicative for leukemic arrest at the early cortical stage of T-cell development with corresponding expression of CD1a, CD4 and CD8 surface marker proteins (99, 102). This arrest results from the ETS mediated recruitment of TLX1 to the TCRα enhanceosome, reducing its activity and consequently blocking correct TCR-Jα rearrangements (**Figure 3**). Moreover, increased levels the repressive H3K27me3 mark across the TCRα locus in TLX positive leukemia indicate that these non-rearranged TCRα segments are epigenetically silenced (103, 104). TLX1 positive T-ALL patients are generally associated with a favorable prognosis and have a longer overall survival as compared with the other subgroups. This can be partially explained by the lack of expression of BLC2 and related anti-apoptotic proteins, since most anti-neoplastic drugs act through the apoptosis machinery and their activity is consequently inhibited by these anti-apoptotic proteins. Thus, upregulation of these anti-apoptotic proteins can results in resistance, while the lack of expression provides a better treatment response (6, 73). In mice, it has been shown that TLX1-driven leukemia is characterized by a long latency (> 25 weeks), showing that *TLX1* overexpression is not sufficient for T-cell transformation and that cooperating genetic alterations are required to fully transform T-cells to leukemia cells (105, 106). Therefore, secondary mutations and/or deletions in *NOTCH1*, *WT1*, *PHF6*, *PTEN*, *PTPN2* and *BCL11B* and NUP214-ABL1 rearrangements have a high frequency in TLX1 positive T-

ALL, enabling further malignant T-cell transformation and providing a TLX1 subgroup specific gene expression pattern with genetic alterations rarely found in other subgroups (35, 70, 107–109). Although activating *NOTCH1* mutations are identified in all subgroups, these mutations are more frequently detected in TLX1, TLX3, LMO1/2 positive patients and are a prerequisite to further evolve towards full malignant transformation (31, 98, 105, 110, 111). Of interest, it has been shown that TLX1 downregulates *NOTCH1* and key NOTCH1 target genes (e.g. *NOTCH3* and *IL7R*), and that there is an overlap of NOTCH1, ETS, RUNX1 and TLX1 binding on these NOTCH1 target genes (98, 112, 113). This unique antagonism between two oncogenes might explain the very high incidence of activating *NOTCH1* mutations in TLX1 driven T-ALL, as well as the very long latency of T-ALL development in a TLX1 driven leukemia mouse model (98). Therefore, targeting *NOTCH1* in TLX1 positive T-ALL may be beneficial and it has been shown that using GSIs reduces cell growth in TLX1 positive T-ALL although the effect is only transient, hinting to combination therapies (106).

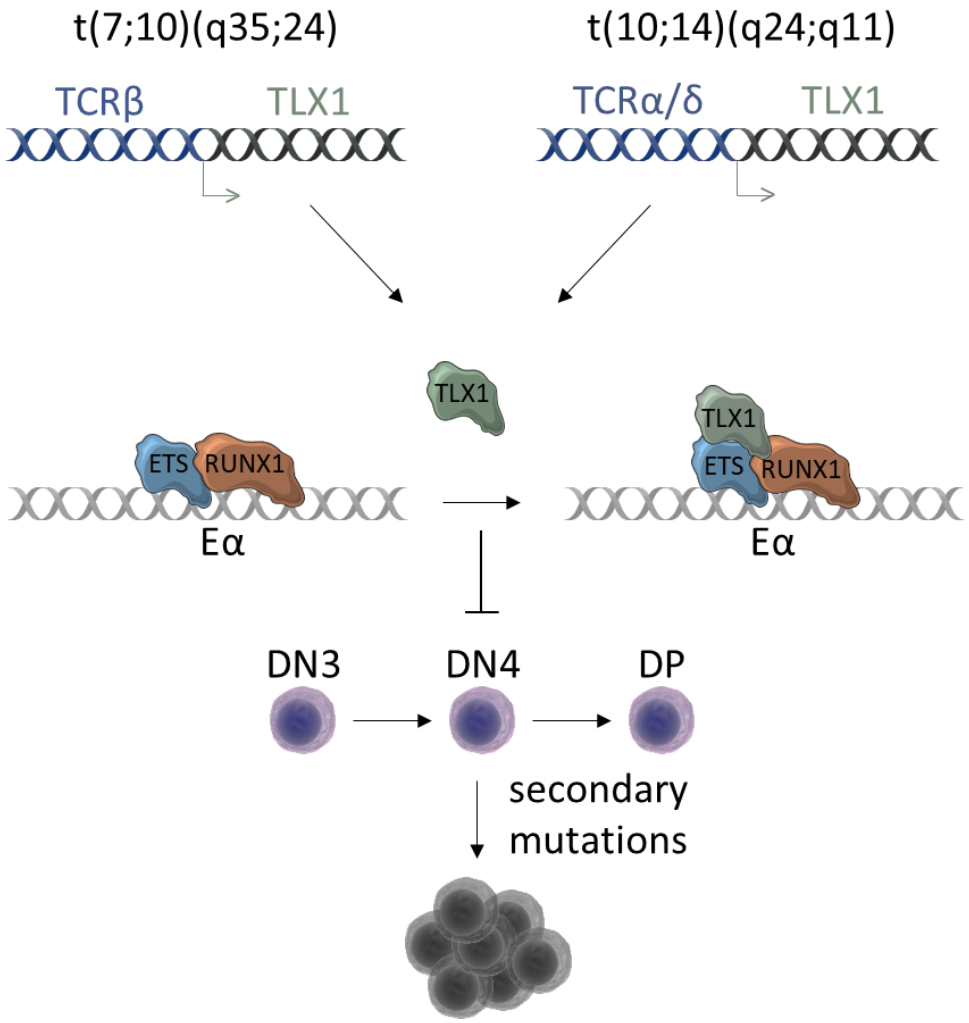


Figure 3: development of TLX1 driven T-ALL. TLX1 is ectopically expressed due to the t(7;10)(q35;q24) or t(10;14)(q24;q11) translocation. Next, TLX1 is recruited by ETS and RUNX to the TCRα enhanceosome, reducing its activity and consequently blocking correct TCR-Jα rearrangements and T-cell differentiation. Secondary mutations cooperate to fully transform immature T-cells in leukemia cells. TCR: T-cell receptor; DN: double positive; DP: double positive.

Overexpression of *TLX1* in immature thymocytes and short interfering RNAs (siRNA) mediated *TLX1* knockdown in TLX1+ ALL-SIL lymphoblasts revealed that TLX1 represses several tumor suppressor genes, including *PTPN2* and *BCL11B* (98, 107). *PTPN2* is often deleted in TLX1 positive T-ALL and is highly expressed in the early cortical stage of T-cell development and consequently most sensitive to

loss at that stage (35, 70, 107–109). *BCL11B c*ontrols early T-cell progenitor differentiation in the thymus, and is often deleted or mutated in TLX1 positive T-ALL (105). Durinck et al. have demonstrated that there is an overlap in the binding of TLX1, RUNX and ETS in the promoters of these TLX1 repressed genes. Furthermore, siRNA mediated knockdown of these three genes showed a common transcriptional expression pattern (98). Recently, it has been shown in mice that the latency of T-ALL development by TLX1 can be shortened by combining *TLX1* expression with the NUP214-ABL1 rearrangement, which is most frequently detected in TLX1 positive T-ALL (70, 114). This combination results in the upregulation of amongst others genes of the JAK-STAT pathway, including *STAT5*. ChIP-seq of STAT5 revealed that STAT5 binding overlaps with TLX1 binding sites throughout the genome and that these binding sites are especially enriched at poised enhancers. Binding of STAT5 and TLX1 can open and activate these enhancers resulting in expression of their target genes. Furthermore, STAT5 and TLX1 bind on the MYC enhancer to activate *MYC* and subsequently reinforce the expression of STAT5 and TLX1 target genes, showing that *MYC* and *STAT5* upregulation contribute to the faster development of TLX1 positive T-ALL (115). In addition to these secondary genetic alterations, de Keersmaecker et al. showed in TLX1 positive mice a high rate of aneuploidy indicating that TLX1 has an influence on the mitotic machinery (105, 108). Indeed, some of the direct TLX1 targets, such as *CHEK1*, a known regulator of the mitotic spindle checkpoint, are downregulated in TLX1 positive T-ALL, resulting in the loss of mitotic checkpoint control and chromosomal missegregation (105). Furthermore, TLX1 has been shown to be involved in G2 to M transition via interaction with the PP2A phosphatase (116). Together, this shows that the development of TLX1 positive T-ALL is a multi-step process in which multiple genetic alterations are required.

TLX3 is another homeobox gene that is highly similar to *TLX1* and is also not expressed in developing thymocytes. Its expression mostly results from the t(5;14)(q35;q32) translocation, juxtaposing *TLX3* to the distal region of *BCL11B*. In addition, translocations of *TLX3* to *CDK6* (t(5;7)(q35;q21)), which is highly expressed during T-cell differentiation and in T-ALL, and to TCRα/δ (t(5;14)(q32;q11)) have been described (117–119). While *TLX1* expression is more prevalent in adults, *TLX3* is expressed in 25 % of the pediatric and 5 % of the adult patients and has a more variable prognosis as compared to TLX1 patients (83, 100, 105, 120). Furthermore, TLX3 patients have a more heterogeneous differentiation arrest, where some cases have an arrest in the immature stage, while others have an arrest in the more mature intermediate αβ/γδ stage, allowing further development into the αβ and γδ lineage (**Figure 2**) (120–122). *TLX1* and *TLX3* share a common gene expression pattern, and 75 % of the regions bound by TLX1 are also bound by TLX3. Furthermore, RUNX1 is a key regulator of TLX1/TLX3 regulatory programs as 50 % of the promoters bound by RUNX1 are also bound by TLX1 and TLX3 (119). As *TLX1* and *TLX3* have similar gene expression signatures, a broad overlap between the regulated genes and induce T-ALL in a similar way, TLX1 and TLX3 patients are mostly grouped in one subgroup. Furthermore, TLX1 and TLX3 positive leukemias have genetic alterations that are rarely found in other T-ALLs such as mutations in *BLC11B*, *PTPN2* and *WT1* and the NUP214-ABL1 fusion (38, 109, 114, 119).

1.1.6. **T-ALL treatment**

The main treatment regimen still applied nowadays for pediatric T-ALL patients is chemotherapy, consisting of three phases: the induction, consolidation and maintenance phase, and typically takes two to three years (**Figure 4**). During the induction phase, one aims to eliminate more than 99 % of the initial leukemic cells to obtain remission and restore normal hematopoiesis. During this phase, asparaginase, vincristine and a glucocorticoid (prednisone or dexamethasone) are given to the patient, in combination with a fourth drug such as anthracycline for high-risk cases. Remission is mostly reached after four to six weeks of treatment in up to 99 % of the pediatric patients. Despite high remission rates, relapse mostly occurs without further treatment, since resistant cells may be selected and

promote further growth. Therefore, an intensive combination chemotherapy is given during the consolidation phase, which typically takes six to eight months. The most commonly used treatment is a combination of asparaginase, vincristine and dexamethasone with or without the addition of anthracyclines, mercaptopurine and methotrexate. Finally, the maintenance phase typically takes 18 to 30 months, with 6'mercaptopurine or thioguanine given daily and methotrexate weekly, often in combination with vincristine and glucocorticoids. In addition, intrathecal chemotherapy is often provided to kill leukemia cells that might be present in the brain and spinal cord. Some patients get radiation therapy, however the role of radiation therapy is currently under debate, as this gives adverse side-effects such as secondary central nervous system (CNS) tumors (5, 7, 123). Unfortunately, up to 20 % of the children relapse and require more aggressive chemotherapy. Nelarabine, a drug licensed for relapsed T-ALL, gives promising results and is the frontline therapy in combination with other chemotherapeutic drugs (124, 125). In addition, allogenic hematopoietic cell transplantation is often provided to these relapsed patients, while this is only given to 5 – 10 % of the children during primary therapy. Since some mutations are frequently detected in relapsed leukemia and promote resistance, it is important to determine these mutations. *CREBBP* mutations are for instance frequently identified in relapsed patients and linked to resistance to glucocorticoids (5, 7, 123).

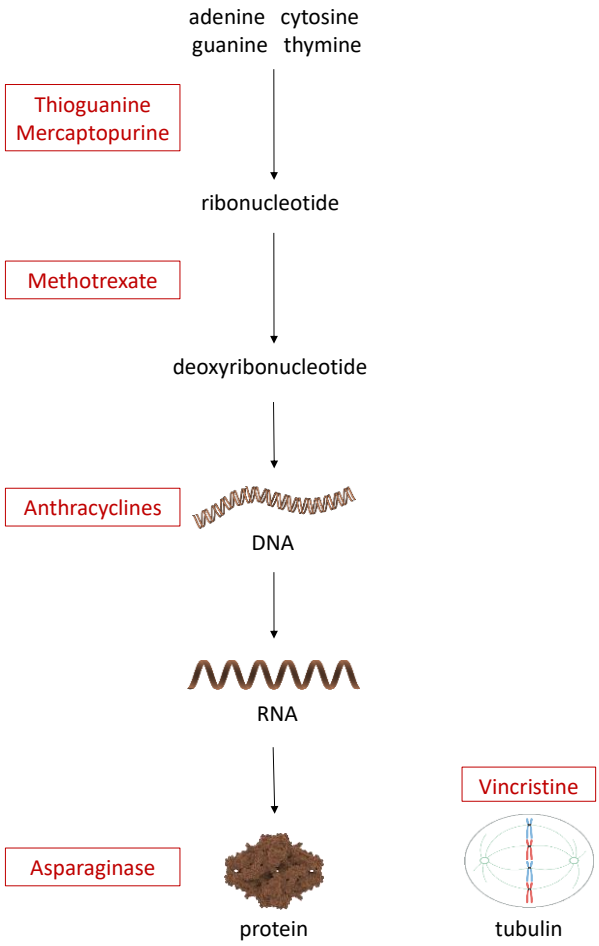


Figure 4: chemotherapeutics used for T-ALL treatment interfere with different stages of DNA and protein synthesis. Chemotherapeutics are highlighted in red. Adapted from (451).

Although the treatment responses are good in pediatric T-ALL, prognosis is much worse for adults. For adolescents and young adults, it has been shown that good results are obtained using pediatric-intense strategies, however, older adults do not tolerate these intensive therapies equally well. These adults

should receive multi-agent chemotherapy based on the pediatric protocols. However, asparaginase is mostly omitted during the induction phase, since this significantly reduces early death and levels of anthracyclines are decreased as this results in extreme bone marrow toxicity. Furthermore, adults often benefit from hematopoietic stem cell transplantation. No consensus exists about the optimal strategy and targeted therapies are an interesting option for these patients (126, 127). Since up to 20 % of the pediatric and 50 % of the adult patients still relapse and current therapies are associated with severe short- and long-term side effects such as osteonecrosis, cardiac dysfunction and CNS toxic effects, targeted therapies are warranted (5, 7, 123, 128). One interesting pathway to target in T-ALL is the NOTCH1 pathway, with very recently promising results for PSEN1 gamma secretase inhibition as described in section 1.1.2. Since monotherapy often results in resistance, also targeted therapies need to be combined to circumvent this resistance (60). The NUP214-ABL fusion, which is most frequently detected in TLX1/TLX3 positive T-ALL, is another target that has promising results using tyrosine kinase inhibitors (129, 130). Despite promising results of chimeric antigen receptor T-cell (CAR-T) therapy for B-cell malignancies, CAR-T therapy for T-ALL remains challenging, as most antigens are shared between leukemic and normal T-cells (131, 132). CD7 is a possible target since this antigen is highly expressed in T-ALL, however, absent in a subpopulation (9 %) of normal T-cells. Furthermore, it has been shown that CD7 knockout has almost no effect on normal T-cell development and T-cell effector function, making it a good target for CAR-T therapy. It has been shown that this therapy has anti-tumor effects *in vitro* and patient derived xenograft (PDX) mouse models, however, this requires further validation (132, 133). Although several potential interesting targets have been identified, recent research has shown that alterations in the non-coding part of the genome also contribute to cancer development and that targeting these non-coding RNAs (ncRNAs) can be beneficial, as demonstrated for SAMMSON in melanoma (134). Since the ncRNA part, and more specifically the long non-coding RNA (lncRNA) part, in T-ALL remained largely unexplored, I investigated the lncRNAome of TLX positive T-ALL in this PhD thesis.

1.2. Long non-coding RNAs

The central dogma of molecular biology states that deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA), which is subsequently translated in proteins (135). Therefore, researchers have investigated the role of protein-coding genes extensively for years, while RNA has just been seen as a mediator molecule between DNA and functional proteins. However, advances in NGS technologies enabled to investigate whole genomes and transcriptomes in a cost efficient and high-throughput manner and revealed that 75 % of the genome is transcribed, while only 3 % is translated into proteins. The genomic regions that are not translated into proteins were previously considered as 'junk DNA', however, a large fraction of these regions is transcribed in ncRNA and this was considered as 'transcriptional noise' (136–139). In recent years, researchers have started to study this vast unexplored part of the human genome and obtained evidence for 1000s of ncRNA, for which crucial biological functions in normal state and diseases are being annotated with increasing speed (139).

1.2.1. Shedding light on the dark matter of the genome

Recent advances in NGS technologies revealed that the fraction of ncRNAs in the genome is much higher as compared to protein-coding genes. These ncRNAs can be subdivided in small (<200 nucleotides (nt) and lncRNAs (>200 nt) based on their length (140). Up to 85 % of the small ncRNAs belong to microRNAs (miRNAs), transfer RNAs (tRNA), small nuclear RNAs (snRNA) and small nucleolar RNAs (snoRNAs), all described to be involved in cancer development (137). MiRNAs are typically 21 nucleotides long and contribute to the regulatory control of about 30 % of the protein-coding genes through binding on their 3'UTR and hindering their function by degradation of the target messenger RNA (mRNA) or inhibiting translation (141, 142). Since these miRNAs are important for the regulation of gene activity, multiple miRNAs have been described to have deregulated expression during cancer development. MIR-19b expression is for instance induced in T-ALL by a translocation placing MIR-19b in the neighborhood of regulatory elements of TCRα/δ (143). tRNAs are involved in transporting amino acids to the ribosomes and changes in tRNA expression can alter protein expression and consequently be involved in the development of diseases. For breast cancer, it has been demonstrated that upregulation of specific tRNAs promotes metastasis, since the upregulation of these tRNAs enhances the translation of genes involved in metastasis (144). SnRNAs are involved in splicing by processing of pre-mRNAs and upregulation of snRNA *U1* revealed that this snRNA regulates genes that are important in cancer development (145). SnoRNAs are encoded intronic regions located within genes and are involved in the regulation of splicing, ribosomal RNA (rRNA) biogenesis and regulation of chromatin structure. Altered expression of these snoRNAs can also play a role in cancer development as shown for *SNORD78* in lung cancer (140, 146). *SNORD78* is upregulated in lung cancer and known to be involved in proliferation and invasion via epithelial-mesenchymal-transition. Furthermore, *SNORD78* is also upregulated in cancer stem-like cells and is important for the self-renewal capacity in these cells, highlighting the importance of *SNORD78* in cancer development (147). According to their function, tRNAs and miRNAs are mostly located in the cytosol while snoRNAs and snRNAs are detected both in the nucleus and cytosol (137, 140). In contrast to these small ncRNAs, lncRNAs are arbitrary defined by transcripts longer than 200 nt that lack an open reading frame (ORF) (< 100 amino acids) (139, 148).

1.2.2. **Characteristics of long non-coding RNAs**

The first identified functional lncRNA, *H19,* was already described in 1990, as an abundant hepatic fetal-specific RNA that is not associated with the translation machinery. While initially an apparent outlier, recent work increasingly shows that many lncRNAs exert critical functions in normal development and diseases (149, 150). LncRNAs are defined by a lack of open reading frame (ORF) (< 100 amino acids) and are longer than 200 nt. LncRNAs are transcribed by Polymerase II, have like protein-coding genes a 5'cap, form complex secondary structures and more than 25 % are spliced (139, 148, 151, 152). In contrast to protein-coding mRNAs, lncRNAs are often located in the nucleus, are less conserved, display higher tissue-specific expression patterns and are in general less abundantly expressed (139, 148, 151, 153). Furthermore, it has been shown that lncRNAs are less stable as compared to protein-coding mRNAs, which is especially the case for nuclear and mono-exonic lncRNAs (154). Initially, microarrays were used for the detection of lncRNA expression levels, thus allowing only assessment of known lncRNAs, for which a probe was designed (155, 156). More recently, RNA sequencing (RNA-seq) became the gold standard, allowing the detection of lncRNAs in an unbiased manner (157). However, classic polyA[+] RNA-seq captures transcripts based on their polyA tail, whereby a large fraction of the lncRNAs remains undetected as some lack a polyA tail (~40 %) (158, 159). Therefore, total RNA-seq protocols are more appropriate for lncRNA research, as these enable to capture both polyadenylated and non-polyadenylated transcripts (139, 148). *NEAT1* is such a non-polyadenylated lncRNA whose expression is elevated in several cancer types, including colorectal cancer and esophageal squamous cell carcinoma, and is associated with metastases and poor overall survival (152, 160). Moreover, *NEAT1* expression can be used as a diagnostic biomarker for colorectal cancer (152). Currently, 56,946 lncRNA genes and 127,802 transcripts have been described (Lncipedia, June 28th 2019) and this number will further increase over the next years (161). LncRNAs can be subdivided in five categories based on their position relative to protein-coding genes: (a) sense lncRNAs overlap with one or more exons at the same strand, (b) antisense lncRNAs overlap with one or more exons on the complementary strand, (c) bidirectional expressed lncRNAs for which the lncRNA and protein-coding gene are located on the opposite strand with their transcription start site (TSS) on less than 1 kb from each other, (d) intronic lncRNAs that are located in an intron and (e) intergenic lncRNAs (lincRNA) located between two genes (**Figure 5**) (162).

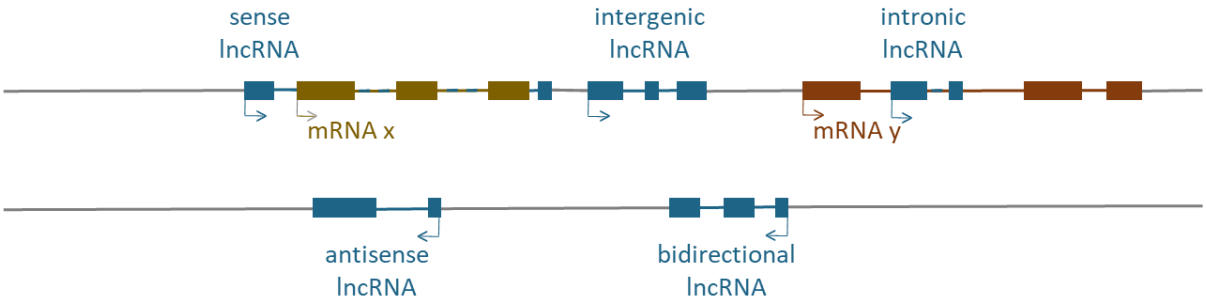


Figure 5: categories of lncRNAs. LncRNAs can be subdivided in five subgroups based on their location relative to nearby protein-coding genes. lncRNA: long non-coding RNA; mRNA: messenger RNA. Figure adapted from (162).

In 2010, a new class of lncRNAs, the enhancer RNAs (eRNA), were described for the first time. These eRNAs are typically 20-2000 nt long and transcribed from active enhancer regions that are characterized by open chromatin and high levels of H3K4me1 and H3K27ac (163, 164). These enhancers are often bound by RNA polymerase II (RNAPII) resulting in bidirectional transcription of eRNAs from these enhancers (164). However, single cell cap analysis gene expression sequencing (CAGE-seq) analysis has revealed that eRNAs are mostly unidirectionally transcribed, as some are transcribed from the sense strand, while others from the antisense strand (165). These eRNAs are

expressed at lower levels than standard lncRNAs, rapidly degraded by exosomes, predominantly non-polyadenylated, and are mostly non-spliced (164, 166, 167). eRNAs have been shown to be involved in transcription regulation via several mechanisms (164). eRNAs can alter gene expression by modifying chromatin structure or by acting as a scaffold for transcription factors. eRNAs can for instance serve as a decoy for *NELF*, which is normally involved in RNAPII pausing, promoting RNAPII elongation (168, 169). Furthermore, these eRNAs can be involved in looping, bringing enhancers and promoters in each other neighborhood and promoting the expression of the target gene (170, 171). This has for instance been shown for *ARIEL*, which is required for growth and survival in T-ALL. *ARIEL* is an eRNA that is significantly higher expressed in the TAL-R T-ALL subgroup compared to the other subgroups and is transcribed from the *ARID5B* enhancer. Furthermore, *ARIEL* is a direct target of TAL1 and activates the TAL-R induced program by activating *ARID5B* expression via enhancer-promoter interactions (172). Currently, only a few eRNAs have been functionally investigated, hinting to new studies to functionally characterize more eRNAs (171). Besides these eRNAs, yet another subtype of lncRNAs, the circular RNAs (circRNAs), have recently been identified as functional RNA molecules, previously thought to be the result of erroneous splicing (173). These circRNAs are not detected using the classic polyA[+] protocols as circRNAs lack a polyA tail (174). CircRNAs are generated by backsplicing whereby a 5' splice site donor and 3' splice site acceptor are ligated, by lariat formation (circularization of an intron) or through exon skipping (175). These circRNAs can consequently contain exons, introns or both, show evolutionary conservation, are expressed at low levels and are tissue-specific (174, 176–178). Notably, up to 20 % of all genes have been shown to produce circRNAs (174, 178). CircRNAs can regulate transcription of other genes via positive regulation of RNA polymerase II or via sponging miRNAs from their target through binding of the miRNA on their miRNA binding sites and post-transcriptional via interfering with splicing (140). Like classic lncRNAs, several circRNAs have been shown to have important roles in cancer. For instance, circRNA *PVT1* inhibits apoptosis and induces cell proliferation in lung cancer by sponging *miRNA-497*, which normally suppresses the anti-apoptotic protein BCL2 (179).

1.2.3. **Mechanisms of action of long non-coding RNAs**

As mentioned above, recent work has revealed a broad potential functionality for lncRNAs, including the regulation of chromatin structure, transcriptional regulation, post-transcriptional regulation and regulation of protein synthesis. Typically, lncRNAs carry out their function as guide, decoy or scaffold for binding partners (**Figure 6**). LncRNAs often influence the transcription of genes through epigenetic control (e.g. modifying chromatin structure). Chromatin is organized in nucleosomes that consist of DNA wrapped around histones. These histones contain two copies of the H3, H4, H2A and H2B proteins, which can be subject to histone modifications at their N-terminal tails extruding from the nucleosome core, instructing the formation of euchromatin (low nucleosomal density) and heterochromatin (high nucleosomal density) regions (180, 181). The latter is mainly marked by heterochromatin, H3K9 methylation, H3K27me3 and H4K20me3 and associated with repression of gene expression, while active genes are characterized by H3K4me3 on their promoter and high levels of H3K36me3 and H3K79me3 in the gene body (181, 182). LncRNAs can alter chromatin structure by binding and recruiting chromatin modifying complexes to target loci in order to activate or repress the expression of target genes. Up to 24 % of the lincRNAs interact with the chromatin modifying complex PRC2, which alters chromatin structure by trimethylating H3K27 and thereby repressing the expression of target genes (183). *HOTAIR* is such a lncRNA that binds PRC2. *HOTAIR* is located in the *HOXC* locus and promotes metastases in amongst others colorectal, gastric and breast cancer. Binding of *HOTAIR* with PRC2 and consecutive recruitment to target loci results in a genome-wide increase of the H3K27me3 repressive chromatin mark, including the *HOXD* locus, promoting invasiveness (184–187).

LncRNAs, such as *ANRIL* also regulate chromatin structure, however by serving as a scaffold for PRC1 and PRC2, resulting in recruitment to and inhibition of the INK4b/ARF/INK4a locus (188). Besides inhibiting gene expression, lncRNAs can also activate genes by binding and recruiting chromatin modifying complexes that activate gene expression (189, 190). This is the case for *HOTTIP*, a lncRNA that drives *HOXA* expression. *HOTTIP* is placed in the neighborhood of its target genes by chromosomal looping and drives H3K4 trimethylation and gene expression of the *HOXA* locus by binding and recruiting WDR5 and the MLL4 methyl transferase complex (191, 192). Besides binding with chromatin modifying complexes, lncRNAs can also bind transcription factors and inhibit their function. This is the case for *PANDA*, which is transcribed antisense to *CDKN1A* and is induced upon DNA damage. *PANDA* inhibits the expression of pro-apoptotic proteins by binding transcription factor NF-YA and prohibiting this transcription factor to bind and activate the pro-apoptotic genes (193). *GAS5* is another lncRNA that inhibits the expression of target genes. *GAS5* inhibits the function of the glucocorticoid receptor by binding with its DNA binding domain. Consequently, *GAS5* competes with the receptor ligands to bind and consequently inhibits its anti-apoptotic effect (194). Finally, lncRNAs can also serve as miRNA sponges (competitive endogenous RNA), sequestering miRNAs from their targets. *HULC* is such a lncRNA that is upregulated in hepatocellular carcinoma and binds with *miRNA-372*, prohibiting to perform its function. Moreover, downregulation of *miRNA-372* has been shown to be linked with poor prognosis in hepatocellular carcinoma, showing the importance of *HULC* for the prediction of prognosis in hepatocellular carcinoma (195).

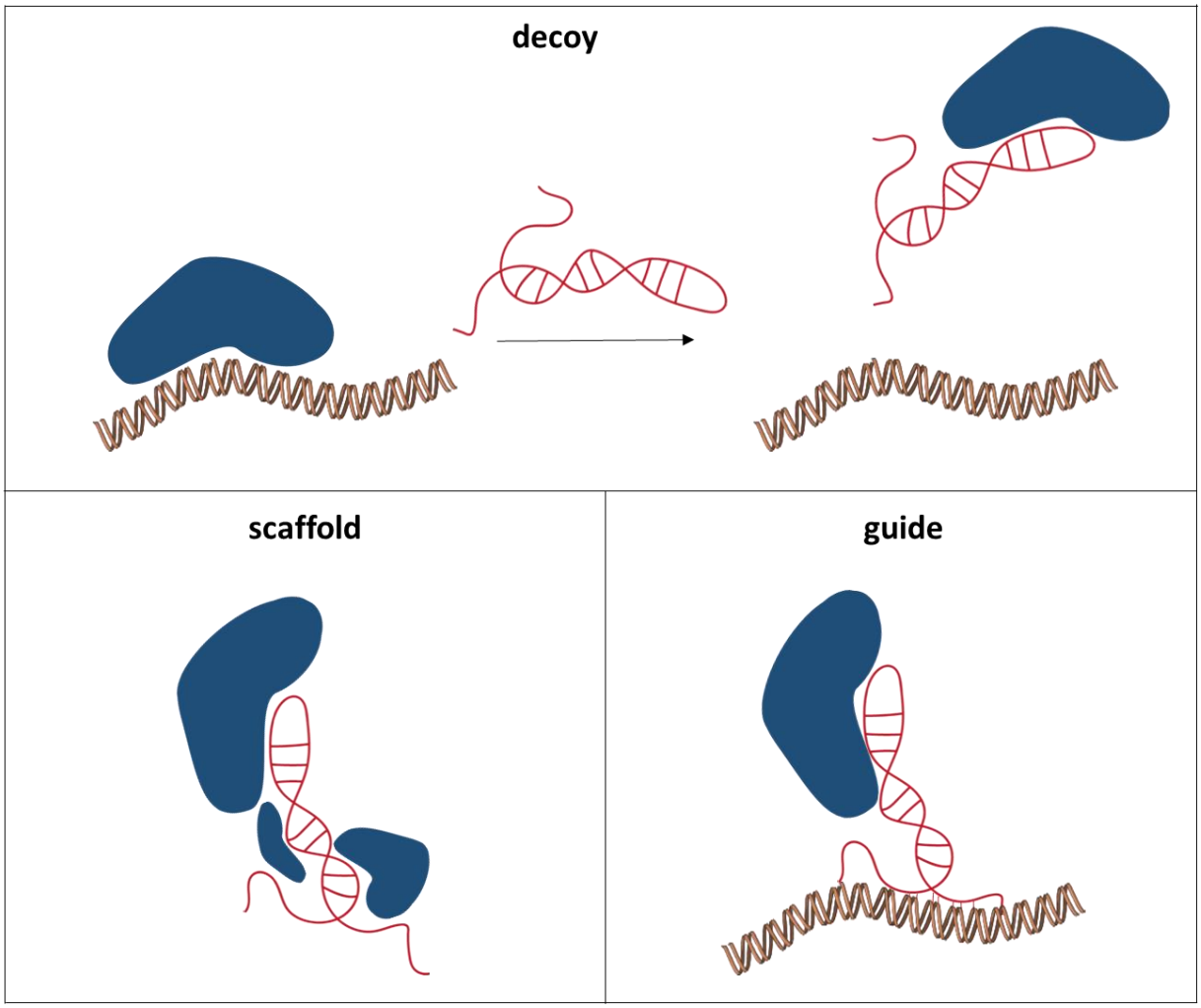


Figure 6: lncRNA mechanisms of action. LncRNAs can act as decoy that hinder binding of DNA-binding proteins with DNA, as scaffolds that bring proteins into a complex or as guides to recruit proteins to the DNA. Figure adapted from (450).

After transcription at a given locus is finished, transcripts are spliced and further processed to increase the complexity of gene expression and generate a higher protein diversity. Several lncRNAs have been described to interfere with these processes. One of the most known cancer related lncRNAs, *MALAT1*, is known to regulate alternative splicing by binding pre-mRNA splicing factors. Depletion of *MALAT1* results in an altered distribution of splicing factors, leading to mis-localisation and altered alternative splicing of several endogenous pre-mRNAs. Furthermore, higher levels of dephosphorylated splicing factors are identified, which influences tracking of the splicing factors to TSSs. This shows that *MALAT1* can influence alternative splicing by interacting with a specific set of splicing factors (196). LncRNAs can also inhibit translation, which has been shown for *lincRNA-p21*, which is unstable when bound with *HuR*. Decreased *HuR* levels stabilize *lincRNA-p21* and this enables association with its target genes, repressing their translation (197). Finally, lncRNAs can also inhibit signal transduction by hiding phosphorylation sites. *NKILA* binds NFkB/IkB, masking its phosphorylation sites. Decreased levels of *NKILA* results in phosphorylation and activation of target genes and is associated with cancer metastasis (198).

LncRNAs can regulate neighboring genes (*cis*) or genes on a distance (> 1mB) or on other chromosomes (*trans*) (139). Of interest, the lncRNA transcript itself is not always necessary. Sometimes, only the act of transcription itself has an influence, while the production of the transcript is not required for the function. This is for instance the case for *Lockd* in mice, which regulates the effect of *CDKN1B/p27*. By deleting the *Lockd* locus, *CDKN1A/p21* transcription is considerably reduced, while producing a shorter *Lockd* transcript by insertion of a polyadenylation signal has no effect on *CDKN1A* expression (199). Another example is *Airn* that is involved in silencing of imprinted gene clusters, including *Igf2r*. The authors have shown that the transcriptional product is not required for silencing, but that conversely the act of transcription is needed as *Airn* transcription interferes with binding of RNAPII on the *Igf2r* promoter. This demonstrates that *Igf2r* silencing depends on transcriptional interference by *Airn* (200).

Despite the fact that lncRNAs have been investigated intensively over the last years, characterizing their function remains challenging (150). The guilt-by-association approach can be used to predict their functions based on correlations with the expression of protein-coding genes. Subsequently, these correlations can be used for gene set enrichment analysis (GSEA) to predict possible functions of the lncRNAs, however, wet lab validation is still required (201–203). Further research will be needed to elucidate the function of the lncRNAs that are currently reported in literature, but not yet functionally characterized.

1.2.4. **Long non-coding RNAs in cancer development**

Cancer is one of the leading causes of death in the developed countries and is a heterogeneous group of diseases arising through malignant transformation of various cell types. Typically, this process proceeds stepwise through accumulation and selection of several genetic cooperative alterations towards a fully transformed cell. Besides DNA copy number variations, base pair variants, small insertions and deletions and changes in epigenetic modifications, also modifications in non-coding regions can contribute to the development of cancer. This underscores the need to further investigate the noncoding part of the genome as current studies mostly focused on the protein-coding part (36, 77, 204). As lncRNA expression is more cell type and cancer-type specific compared to protein-coding genes, lncRNAs can be used to discern cancer (sub)types (155, 205). A large scale study investigating lncRNAs in 5860 tumors of 13 cancer types revealed that 13 % of the lncRNAs contained gain or losses that occur in at least 25 % of the samples of a specific cancer type (205). Furthermore, the authors revealed that base pair variants are often located in lncRNAs or their regulatory elements altering their

expression. Base pair variants can for instance create a miRNA binding site, which reduces the lncRNA activity after binding of the miRNA and can contribute to tumorigenesis (204, 206). Low expression of *NBAT-1*, a lncRNA correlated with poor survival in neuroblastoma, is often obtained by promoter hypermethylation or a base pair variant (207). Furthermore, a base pair variant in a regulatory element can also alter the expression of the cancer associated lncRNA, as shown for *PCAT1*. A base pair variant in its enhancer increases the binding of specific transcription factors upregulating *PCAT1* expression and consequently promoting prostate cell proliferation and tumor growth (204, 208).

In 2000, Hanahan and Weinberg proposed six hallmarks that a cell should acquire to become a cancer cell (**Figure 7**) (209, 210). First, a cell should sustain **proliferative signaling** allowing to maintain growth in the absence of external stimuli. The growth of normal cells is tightly regulated, while cancer cells have accumulated genetic defects through which this growth control is lost. *PCAT1* is overexpressed in several cancer types and was first discovered by transcriptome analysis of prostate cancer and shown to be involved in maintaining proliferative signal. This has been demonstrated by *PCAT1* knockdown, where after genes involved in cell cycle and mitosis were downregulated, suggesting that these genes are highly active in *PCAT1* overexpressing tumors (209, 211). In addition, *PCAT1* upregulates cMYC protein levels, further enhancing cell proliferation and transformation. *PCAT1* stabilizes cMYC protein post-transcriptionally by inhibiting *miR-34a*, which normally targets and degrades cMYC, resulting in cell proliferation (212). Second, a cancer cell should **evade the activity of growth suppressors** that negatively regulate cell proliferation. Tumor suppressor genes, such as *p15*, *p53*, *RB1* and *PTEN* normally reduce cell growth by inducing cell cycle arrest, senescence or apoptosis and their expression is often deregulated in cancer cells. *ANRIL* represses transcription of the tumor suppressor gene *p15*, which is involved in cell cycle regulation, by recruiting SUZ12, a PRC2 component, to the *p15* locus. This results in H3K27me3 and silencing of *p15*, leading to a loss of cell cycle control (188, 209). Third, a cancer cell should acquire **unlimited replicative potential** as a normal cell can only undergo a limited number of cell cycles due to telomere shortening that results in cellular senescence or apoptosis. 90 % of the cancer cells circumvent this by activation of the telomerase enzyme that add new telomeric repeats at the end of the chromosomes. *TERRA* is a lncRNA that is transcribed from the telomeres and that negatively regulates telomerase by binding to this enzyme due to sequence complementary and therefore competes with the telomeric DNA. In cancer, low *TERRA* levels are often observed due to hypermethylation of the telomeres, resulting in low *TERRA* expression and an increase in telomeric lengthening, providing unlimited replicative potential to the cancer cell (209, 213–215). Fourth, cancer cells should be able to **invade in other tissues and to form metastasis**, since most patients die due to these distant metastases. *MALAT1* is one of the most abundant lncRNAs and is significantly associated with metastasis and survival in several cancer types, including non-small-cell lung carcinoma (NSCLC), epithelial ovarian cancer and osteosarcoma, elucidating that *MALAT1* can be used as prognostic marker in these cancer types (209, 216–218). *MALAT1* positively regulates invasion and metastasis by promoting cell motility through transcriptional and post-transcriptional regulation of motility-related genes (219). Furthermore, *MALAT1* can also promote metastasis by binding and recruiting SUZ12 of the PRC2 complex to E-cadherin, a cell adhesion molecule, resulting in downregulation of E-cadherin and consequently in less cell adhesion, promoting invasion (220). Fifth, cancer cells should induce **angiogenesis** in order to secure nutrients and oxygen supply to enable further growth. A natural antisense transcript*, αHIF*, complementary to the 3'UTR of *HIFα*, negatively regulates *HIFα,* which is a key regulator of angiogenesis. *αHIF* can bind on the 3'UTR of *HIFα* and degrade its mRNA disrupting the strict regulation of angiogenesis (209, 221, 222). Finally, a cancer cell must be **resistant to cell death**. *PCGEM1* is highly expressed in prostate cancer and this high expression results in a decrease of apoptosis after doxorubicin treatment. This can be explained by a decrease in TP53 stability upon *PCGEM1* overexpression and consequently a downregulation of *p21,* resulting in a

androgen dependent inhibition of apoptosis upon doxorubicin treatment (209, 223). Later, two hallmarks were added: reprogramming of energy metabolism and evading immune destruction (224). To maintain the rapid growth of cancer cells, these cells should **reprogram their metabolism**. Cancer cells have for example a higher glucose uptake and several lncRNAs are involved in this process. *Lnc-IGFBP4-1* is significantly higher expressed in lung cancer tissues as compared to normal lung tissue and is associated with metastasis. Overexpression of *lnc-IGFBP4-1* results in an increase of Adenosine triphosphate (ATP) production and a decrease in the expression of enzymes involved in glucose homeostasis, resulting in increased glucose uptake (225). Cancer cells should also **evade the immune system**. *Lnc-EGFR* is upregulated in regulatory T-cells (Treg) cells in hepatocellular carcinoma and is involved in this evading by inducing differentiation and inhibition of cytotoxic activity in Treg cells in hepatocellular carcinoma, stimulating tumor growth (226). To acquire these hallmarks, a cell should accumulate several genetic alterations. Since cells have several mechanism to maintain genome integrity, genetic alterations in components of the genomic maintenance machinery should be acquired to allow the accumulation of multiple genetic alterations required for the acquisition of the hallmarks (224). *GUARDIN* is a lncRNA with a key role in the maintenance of genomic integrity by serving as a scaffold for BRCA1 and BARD1, which is required for BRCA1 stabilization, and consequently promotes BRCA1 mediated DNA repair required for DNA integrity (227).

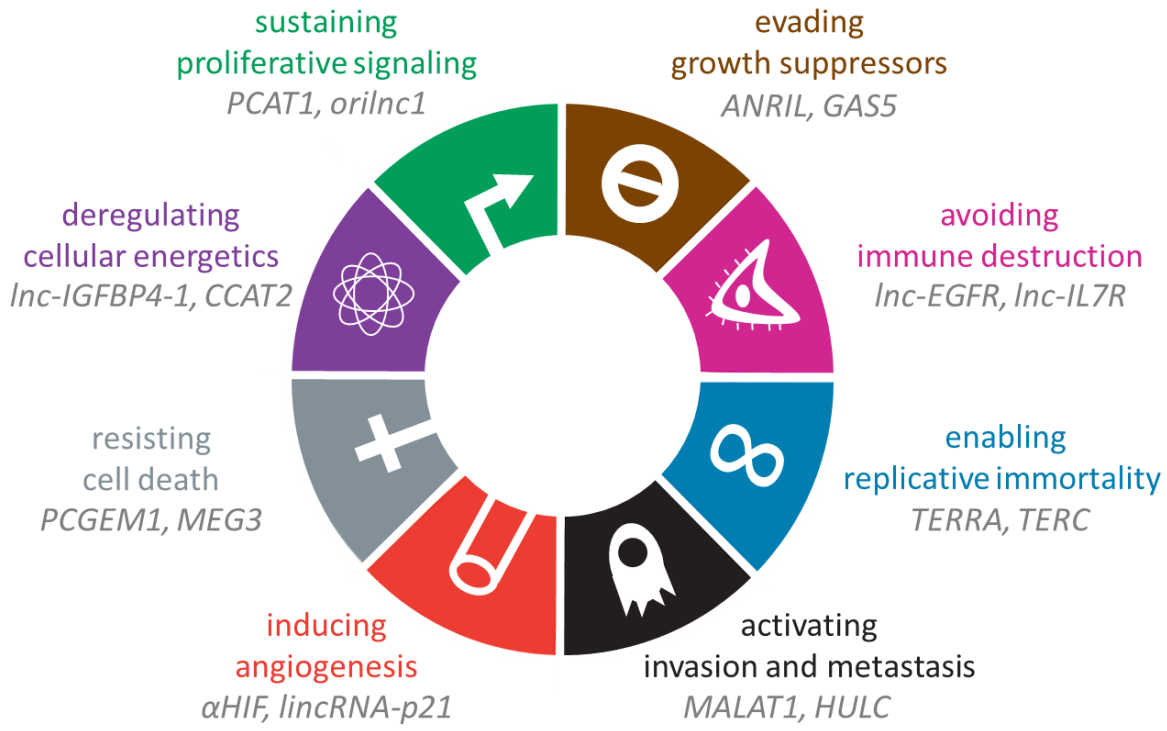


Figure 7: lncRNAs associated with the hallmarks of cancer. LncRNAs have been shown to involved in cancer development and maintenance and are associated with the hallmarks of cancer. Figure adapted from (224).

1.2.4.1. Long non-coding RNAs in T-ALL

Over the last years, several lncRNAs involved in T-ALL development have been described. The first lncRNA studies in T-ALL investigated the effect of NOTCH1 on lncRNAs. Trimarchi et al. have shown that *LUNAR-1* is a NOTCH1 regulated lncRNA that is overexpressed in T-ALL patients harboring a *NOTCH1* mutation and is downregulated upon *NOTCH1* inhibition. HiC and chromosome conformation capture (3C) sequencing revealed an interaction between the promoter of *LUNAR1* and an *IGF1R*

enhancer. Binding of NOTCH1 on this *IGF1R* enhancer activates *LUNAR1*, which subsequently recruits factors such as mediator and RNAPII to obtain full activation of the *IGF1R* promoter. Furthermore, downregulation of *LUNAR1* results in downregulation of *IGF1R* and reduced cell growth, indicating that *LUNAR1* promotes IGF signaling through IGF1R regulation and is consequently required for T-ALL growth (228). The host lab revealed an additional set of lncRNAs, including *LUNAR1*, regulated by NOTCH1. Most of these NOTCH1 regulated lncRNAs are bound by ICN1, but also by MED1 and BRD4 and occupied with H3K27ac, hinting to potential enhancer activity of these loci (156). In contrast to these lncRNAs that are regulated by *NOTCH1*, *NALT1* is located 400 bp upstream of *NOTCH1* and acts as a cis-regulatory activator of NOTCH1. *NALT1* is higher expressed in T-ALL patients compared to normal controls and inhibition resulted in reduced T-ALL growth (229). LncRNAs can also be used to discriminate the molecular subgroups in T-ALL. To explore this, the host lab performed microarray transcriptome profiling with probes covering all protein-coding genes and 13,000 lncRNAs on 64 T-ALL patients (15 immature, 17 TLX1/3, 25 TAL-R, and 7 HOXA), revealing a subset of specific lncRNAs for each subgroup. Some lncRNAs were not expressed in the normal thymocytes implying that these may be oncogenic lncRNAs, whereas other lncRNAs were higher expressed in thymocytes compared to T-ALL patients, hinting to a putative tumor suppressor role of these lncRNAs (155). A specific study focusing on TAL1 positive T-ALL revealed 57 lncRNAs that are bound by and downregulated after knockdown of *TAL1*. Furthermore, most of these lncRNAs were also downregulated upon knockdown of TAL1 regulatory partners (RUNX1, GATA3 and MYB), indicating that TAL1 regulates these lncRNAs with its cofactors. The authors revealed two lncRNAs that are regulated by TAL1 and associated with superenhancers in TAL1 positive leukemia, but not in the thymus, implying that these might have oncogenic activity. In addition, one of these lncRNAs is expressed in hematopoietic stem cells and early progenitors, but absent in more mature stages, showing that this lncRNA is downregulated during T-cell development and that aberrant expression of this lncRNAs can contribute to T-ALL development (230). In this PhD research, I revealed lncRNAs regulated by TLX1 and specific to the TLX subgroup of T-ALL patients (**Paper 1 & 2**).

1.2.5. **Long non-coding RNAs: new opportunities for specific cancer treatments**

LncRNAs are excellent therapeutic targets as these are often expressed in a cell-type specific manner, implying that targeting them could lead to less toxic side-effects on normal cells. In addition, lncRNAs have a lower expression, indicating that lower doses of drug may be sufficient, which will cause less toxicity (150). This has been shown for the *SAMMSON* lncRNA, which has melanoma specific expression and is expressed in more than 90 % of the primary and metastatic skin cutaneous melanomas, while only marginally in normal melanocytes. Furthermore, *SAMMSON* is induced in the transit from an immortalized to a fully transformed cell stage, making it a putative biomarker for malignant melanoma. Furthermore, knockdown of *SAMMSON* reduces cell growth and activates the mitochondrial apoptosis pathway, highlighting that *SAMMSON* may be an interesting therapeutic target. In addition, the authors showed in a xenograft experiment that BRAF resistant patients still require *SAMMSON* expression and that the combination of BRAF inhibition and *SAMMSON* knockdown results in increased apoptosis. Since *SAMMSON* is expressed in more than 90 % of the melanomas and barely in melanocytes and other cancer types, *SAMMSON* is an ideal therapeutic target that will be specific and cause less off-target effects (231).

Oncogenic lncRNAs can be downregulated using siRNA, antisense oligonucleotides (ASO), ribozymes, clustered regularly interspaced short palindromic repeats interference (CRISPRi) or small molecules (232). SiRNAs are double stranded RNA molecules of 21-23 nt and have a 3’ dinucleotide overhang. SiRNAs are incorporated in the RNA-induced silencing complex (RISC) and bind with their target to degrade it (153). These siRNAs can be used to target lncRNAs, however, secondary structures can

prohibit binding of the siRNA to the lncRNA. Furthermore, most lncRNAs are located in the nucleus, while the RNA interference (RNAi) machinery is mainly located in the cytoplasm, making lncNRAs less accessible to siRNAs (150). In contrast, ASOs use RNAseH, which is active in the nucleus, to degrade the lncRNA. ASOs are single stranded DNA molecules that bind RNA with a complementary sequence. This hetero DNA-RNA duplex is subsequently recognized and degraded in the nucleus by RNAseH. ASOs can also be designed to bind on 3' or 5' splice junctions to alter splicing and isoform production (150, 232–234). Furthermore, modified ASOs that do not activate RNASeH can also be used to prevent secondary structure formation or to cause steric hindrance, inhibiting RNA-protein binding (232). To increase their efficiency and decrease their off-target effects and degradation, several modifications can be applied. The 2' sugar position is often modified with 2'-O-Me or 2'-O-(2-methoxyethyl) modifications. However, not all 2' sugar position may be modified as this inhibits RNAseH activity, thus ASOs must contain a normal central section flanked by 2' modified regions. This provides a higher binding affinity and reduces nuclease degradation and immunogenicity (150, 233, 235, 236). Locked nucleic acid (LNA) modifications can also be used at the flanking regions, improving RNA binding affinity, reducing immunogenicity and promoting binding with proteins such as albumin resulting in decreased renal clearance. These LNA modifications contain a phosphothioester binding providing resistance to enzymes and the 2' and 4' position of the ribose are bound ('locked') with a methylene bridge (233, 236–239). Currently, several ASOs and siRNAs for miRNAs and mRNAs are in clinical trials, while pre-clinical tests for lncRNAs are more challenging due to the low conservation (150, 235). Ribozymes are self-cleaving, 30 nt long RNA molecules that bind complementary sequences and cleave the flanking regions. These ribozymes are highly sequence specific and are sensitive to single nucleotide mismatches (mutant vs normal), resulting in minimal off-target effects (150, 153, 240). Finally, also the CRISPR technology can be used to target lncRNAs. Although this approach is well-established for protein-coding genes, a single cut in a non-coding gene does often not generate a knockout. However, two guide RNAs can be used for lncRNAs to generate a large deletion of the lncRNA (241). This may be hindered by the fact that lncRNAs often overlap with enhancer regions or protein-coding genes whereby the effect can be due to a deletion in the enhancer or protein-coding gene or due to deletion of the lncRNA (242). To circumvent this, CRISPRi can be used. Therefore, a guide RNA guides the inactive death Cas9 (dCas9) linked to a repressor to the promoter of the lncRNAs, repressing the lncRNA (243, 244). As it is known that lncRNAs can also act by binding with proteins, their function can also be inhibited using small molecules that interfere with this binding (232).

One of the main challenges of RNA therapeutics is the efficient delivery to the target. Therefore, RNA therapeutics should cope with several intra- and extracellular barriers such as extravasation from the bloodstream to the target tissue, penetration through cell membranes and escape from endosomes and the immune system. First, RNA therapeutics have to migrate through the endothelial pores, which are very tight. Since delivery of free RNA to the target tissue is challenging, carriers can be used. The advantage of a carrier is that the RNA is protected from enzymes and that tissue specificity can be improved by for instance attaching a specific antibody to the carrier (235). Viral vectors have the advantage that these are highly effective, although have a high immunogenicity, while non-viral vectors, such as lipid-based and polymeric vectors are mostly less effective, but have a lower immunogenicity and can be produced at high-throughput and lower costs (150). Next, RNA therapeutics need to pass the cell membrane, which is difficult as membranes repulse the negative loaded RNA therapeutics (150). To facilitate the uptake, carriers are often positively loaded and cell penetration peptides can also improve internalization (150, 245). After internalization, cells must escape from endosomes before these fuse with lysosomes, which can be facilitated by the use of endosomolytic agents (150, 246). Finally, RNA therapeutics must also circumvent immune activation and enzymatic degradation. To reduce enzymatic degradation and immunogenicity, chemical

modifications can be applied as described above (150, 235, 238, 247, 248). The first RNA therapeutics showed poor performance due to the aforementioned limitations, but several second generation RNA therapeutic are now in clinical phases, with promising results (235). Besides therapeutic targets, lncRNAs can also serve as excellent biomarkers since the expression of several lncRNAs has been associated with disease severity and progression. For *HULC* it has been shown that its expression is significantly higher in hepatocellular carcinoma compared to the normal liver tissue and that its expression levels are associated with the tumor grade, marking *HULC* as a high potential biomarker (249). Of interest, *PCA3* is the first FDA-approved lncRNA biomarker for prostate cancer as *PCA3* is expressed in 95 % of all prostate cancer cells and expression levels in the urine are predictive for a positive biopsy (250).

1.3. Single cell omics

1.3.1. Limitations of bulk RNA sequencing are circumvented by single cell RNA sequencing

In 1977, Sanger et al. developed the first widely used DNA sequencing method that has been used extensively for the following 30 years. This method enabled to sequence the genomes of several species, including the first human genome in 2003 (251, 252). In 2005, second generation or NGS technologies, such as 454 and Illumina (former Solexa) sequencing, were developed, allowing to investigate whole genomes in a cost efficient and high-throughput manner (252, 253). In addition, these innovative approaches enable to sequence multiple samples in parallel and generate high coverage data, and are therefore currently widely used. One limitation of these NGS methods is that relatively short stretches of DNA (36-400 bp) are sequenced (i.e. reads) (2,4). Therefore, third generation sequencing technologies, such as PacBio's single-molecule real-time sequencing and Nanopore sequencing, emerged in 2011 to perform long read (up to 100 kb) sequencing, enabling the detection of genomic variants and repeats at high resolution (252, 254, 255). Since the cost of massively parallel sequencing dropped over recent years, the genome sequences of several species are now publicly available. These sequencing efforts contributed to the current knowledge of molecular mechanisms involved in normal development and diseases. These second and third generation sequencing methods allow not only to retrieve sequences from entire genomes, but can also be used to unravel the cellular transcriptome, referred to as RNA-seq. Transcriptome profiling by means of RNA-seq offers multiple advantages over the former use of microarrays: (1) a higher sensitivity for low abundance genes, (2) gene expression can be investigated without the need of probes and prior knowledge enabling to detect novel genes, (3) it is based on sequencing instead of hybridization (with inherent higher specificity) and (4) alternative splice forms and other structural features can be detected (253, 256–258). Originally, transcriptome analysis was conducted at the average level of a (heterogeneous) cell population ('bulk RNA-seq'), and consequently masked subtle differences among cells. While it is clear that tissues are composed of different cell types, more recent research has shown that the transcriptomes of closely related cells can also show remarkable heterogeneity, which cannot be detected at the population level. Furthermore, some genes can be low abundant in one subpopulation, while highly expressed in a second subpopulation, resulting in a moderate average expression of the gene by performing bulk RNA-sequencing and consequently hiding this heterogeneity (**Figure 8**). This molecular heterogeneity can be partially explained by amongst others differences in cell cycle stages across populations as well as the phenomenon of transcriptional bursting (259). Transcriptional bursts are short intervals of transcription followed by a period of transcriptional silence resulting in gene specific temporal expression patterns that differ among cells (260–262). Fluorescence in situ hybridization (FISH) experiments have shown that the expression between similar cells can differ by a 1000 fold (263, 264). As bulk RNA-seq methods generate average expression profiles, subpopulations and rare cell types cannot be detected (253, 256, 265–268). The landscaping of transcriptional heterogeneity is important in various fields, such as cancer (267, 269), embryonic development (270–272), and immune response (273), whereby single cell analysis methods are warranted. The development of adequate single cell RNA-seq methods was challenging as a mammalian cell typically contains 10-20 pg of RNA and previously developed RNA-seq methods required thousands to millions of cells as input to generate high quality data (253, 256). Over the last years, new bulk RNA-seq kits were developed lowering the required number of cells as input. Nevertheless, low input reduces the reverse transcription efficiency of low abundant genes resulting in a biased detection towards highly expressed genes and requires incorporation of pre-amplification

to obtain sufficient amounts of amplified cDNA to sequence, causing amplification bias (274, 275). Although these low input methods allow the detection of thousands of transcripts, a substantial level of technical noise is generated and hides subtle biological differences between samples. Therefore, further improvements were made to enable transcriptome sequencing of a single cell by lowering the reaction volumes and enhancing the efficiency of the required enzymes (275).

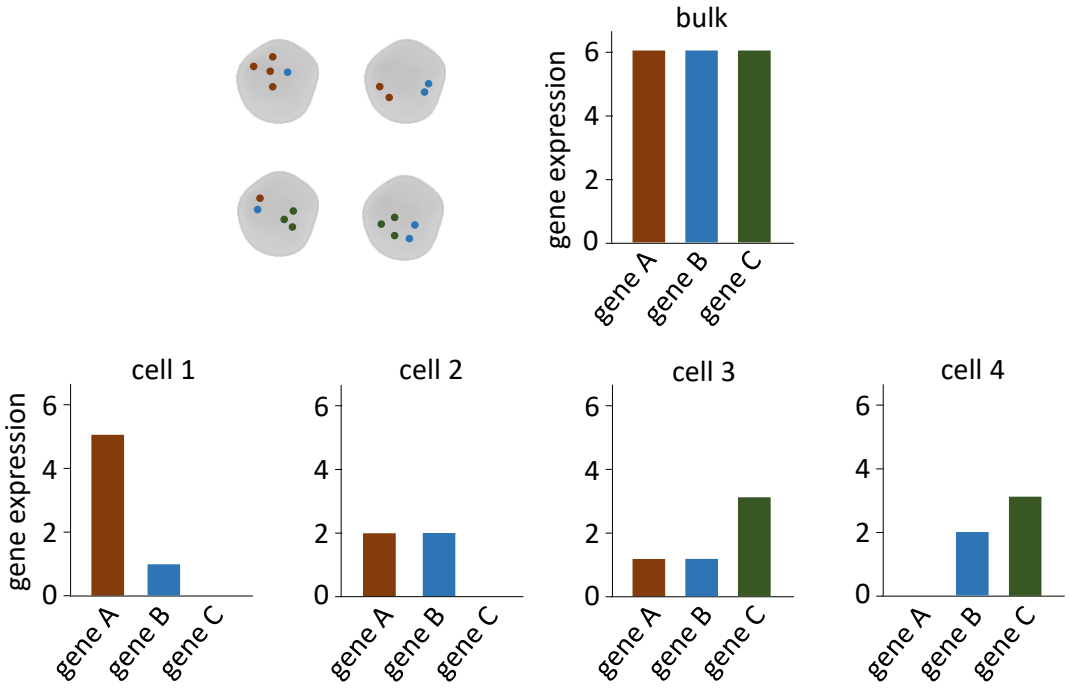


Figure 8: bulk RNA sequencing masks cellular heterogeneity. By performing bulk RNA sequencing, average gene expression profiles are generated, whereby every cell seems to have an equal expression of gene A, B and C. In contrast single cell RNA sequencing reveals that the expression of these genes differs among cells.

In the first single cell experiments, only a limited number of genes could be investigated in a few cells, whereas bulk sequencing experiments enabled to investigate thousands of genes, but required thousands to millions of cells as input, resulting in average profiles. Recently, these two fields merged together due to improvements in molecular methods and analyses allowing to perform high-throughput molecular analyses at the single cell level (**Figure 9**) (276). This encouraged researchers in 2016 to start a large scale international collaboration (Human Cell Atlas) to sequence all cell types in the human body at the single cell level to generate a reference map. This will allow to link molecular profiles with cellular locations throughout the body and provide deeper insights in cell development, cell-cell interactions and pathways in healthy tissues. As diseases emerge as a consequence of rewired homeostasis, this initiative is expected to gain deeper insights in the molecular basis of various disease types, facilitating the development of new treatments (277, 278). This atlas will be an unseen source of information to understand normal development as well as diseases.

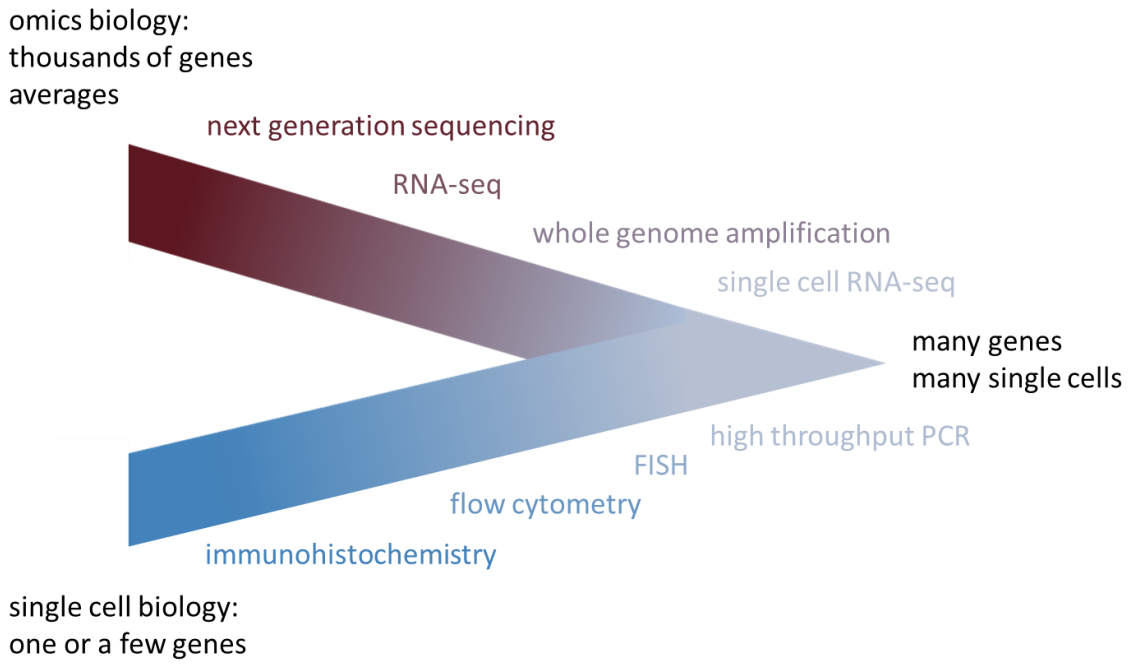


Figure 9: convergence of bulk next generation sequencing technologies and single cell methods. Single cell experiments evolved from a few cells and genes to many cells and genes per experiment. FISH: fluorescence in situ hybridization; PCR: polymerase chain reaction. Figure adapted from (276).

1.3.2. From one up to thousands of single cells

In the first single cell experiments, only a few transcripts in a limited number of cells could be investigated using FISH (279, 280). To increase the number of genes and cells that could be investigated in a single experiment, single cell reverse transcription quantitative polymerase chain reaction (RT-qPCR) emerged. Since a single cell only contains 10-20 pg of RNA, this method comprises a poly(dT) pre-amplification step of the polyadenylated transcripts to obtain detectable numbers of molecules (281). The hands-on time and costs for these experiments could be reduced by automated devices such as Fluidigm's Biomark HD system. This considerably improved the throughput, although still only a fraction of all genes in a cell, which are selected upfront based on scientific hypotheses, can be investigated. To get a more global view of a cell's transcriptome, microarray and sequencing-based methods emerged (276, 281). Using existing or custom made microarrays, which contain probes for genes of interest for a specific application, thousands of genes could be investigated in parallel. Since this was still limited to sets of known genes and the sensitivity and throughput were low, single cell RNA-seq methods were developed enabling the study of a cellular transcriptome in an unbiased way, allowing to detect both known and novel genes (282). In 2009, Tang et al. published the first single cell RNA-seq protocol in which cells were picked manually and transcripts reverse transcribed using a poly(dT) primer with an anchor sequence (**Figure 10**). Next, the cDNA was polyadenylated and a second poly(dT) primer was used to synthesize the second strand (283). As this protocol is labor intensive and consequently only possible for a limited number of cells, other methods rapidly emerged during the last decade (**Table 1**). In 2011, methods using early multiplexing (using a cell specific barcode to discriminate individual cells) were introduced enabling to pool cells at an early stage and increasing the throughput (**Figure 10**). Consequently, cells can be further processed in a single tube and be treated as a single sample reducing the hands-on time and costs (276). The single cell tagged reverse

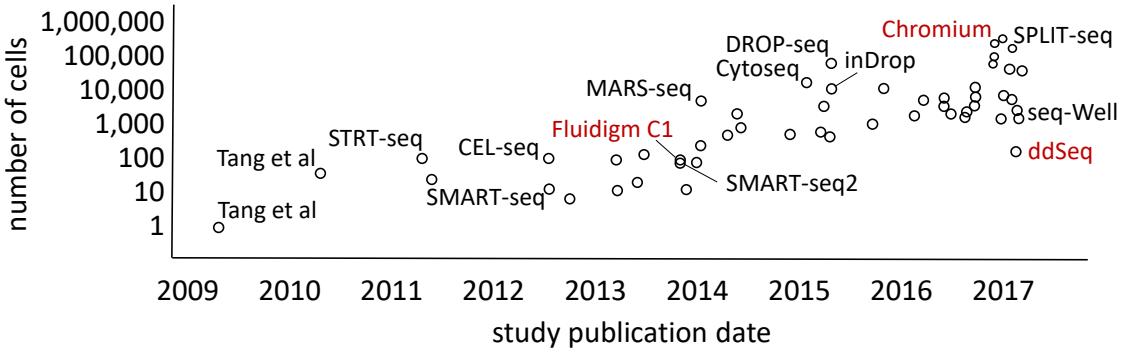


Figure 10: the number of single cells per experiment drastically increased the last decade. In the first experiments, cells were manually picked. The emergence of microfluidic devices enabled to isolate hundreds of cells in parallel and even thousands of cells by sample multiplexing. Droplet-based and nanowell technologies now enable to isolate tens of thousands of single cells in a single experiment. The devices used in this PhD thesis are highlighted in red. Figure adapted from (452).

transcription sequencing (STRT-seq) and single cell RNA barcoding and sequencing (SCRB-seq) methods use this early multiplexing by adding cell specific barcodes at the step of RT. Consequently, a lot more cells can be processed as cells can already be pooled in a single tube following cDNA synthesis (284, 285). These methods use the template switching mechanism that adds non-templated nucleotides to the 3' end of the first cDNA strand, used by the second primer for second strand synthesis and thereby eliminating the need to add a polyA tail after RT (284–286). While the STRT-seq method gives some 5' end bias and the SCRB-seq method 3' end bias, the SMART-seq method generates read coverage across the whole transcript. This whole transcript body coverage expands the spectrum of applications beyond gene expression profiling as this method can be used for the detection of fusion transcripts, single nucleotide variants (SNV), gene copy number analysis and alternative splicing events. A drawback of this method is the lack of an early cell barcoding step, whereby cell pooling is only possible at a later stage compared to methods with early barcoding, making the method more labor intensive (274, 287, 288). However, this issue can be solved by using automated liquid handling devices (289). The aforementioned methods are all based on a polymerase chain reaction (PCR) based amplification step to obtain sufficient cDNA to sequence, which also leads to a considerable amplification bias. Cell expression by linear amplification and sequencing (CEL-seq) reduces this bias through cost and time efficient linear *in vitro* transcription (IVT) in a single cell experiment (**Table 2**). In brief, transcripts are captured using a poly(dT) primer containing a cell specific barcode and a T7 promoter. After cDNA synthesis, the cells are pooled and a single round of IVT is carried out on the pool of cells reducing reagents costs and enabling to analyze many cells in parallel. Finally, the samples are fragmented and the 3' end of the transcripts are selected and converted to sequencing libraries. The sequencing coverage required for this type of libraries is low as only the 3' ends are sequenced further reducing the costs (290). The throughput of the CEL-seq protocol was increased considerably by MARS-seq, in which cells are sorted in 384 well plates and processed automatically increasing the throughput and reproducibility (291, 292). To increase the sensitivity and accuracy and to reduce the noise, time and costs of single cell experiments, optimized versions of STRT-seq, SMART-seq and CEL-seq have consecutively been developed (293–295). In addition, unique molecular identifiers (UMI) were added to the primer sequence of the CEL-seq protocol to count the transcripts more precisely (294). In these first single cell experiments, cells were manually picked limiting the number of cells that could be processed in parallel. By automatically sorting cells in 96 or 384 well plates, the throughput increased,

but the experiments remained expensive due to the large reaction volumes in these wells. To accommodate this, microfluidic devices using microfluidic channels and pressure-controlled valves were developed, in which cells are captured and processed in nanoliter volume reaction chambers reducing the reagents costs (**Figure 10, 11A, Table 1-2**). These valves separate the reaction chambers, to which different reagents can be added sequentially (273, 296, 297). Of note, it has been shown that the sensitivity of single cell experiments in reduced reaction volumes increases compared to higher reaction volumes and that microfluidic methods such as Fluidigm's C1 have a higher sensitivity compared to manual methods (294, 298). As these microfluidic devices are semi-automated, equal amounts of reagents, incubation times and mixing steps are obtained, reducing the technical variation due to human handling, which can be considerable in single cell experiments (273, 296). Another advantage of these microfluidic devices is that closed reaction chambers are used, reducing the risk of contamination. This contamination can be high in open bench-top experiments as the input of single cell RNA-seq experiments is low, whereby any contaminant molecule will be co-amplified (273). Several methods, including STRT-seq, SMART-seq and CEL-seq have been modified to work on plate sorted or microfluidic isolated single cells (294, 295, 299). Using these plate-based or microfluidic-based methods, still only hundreds of single cells can be isolated and capture efficiencies are rather low, making these methods less suitable to detect rare cell types or to analyze clinical samples (300).

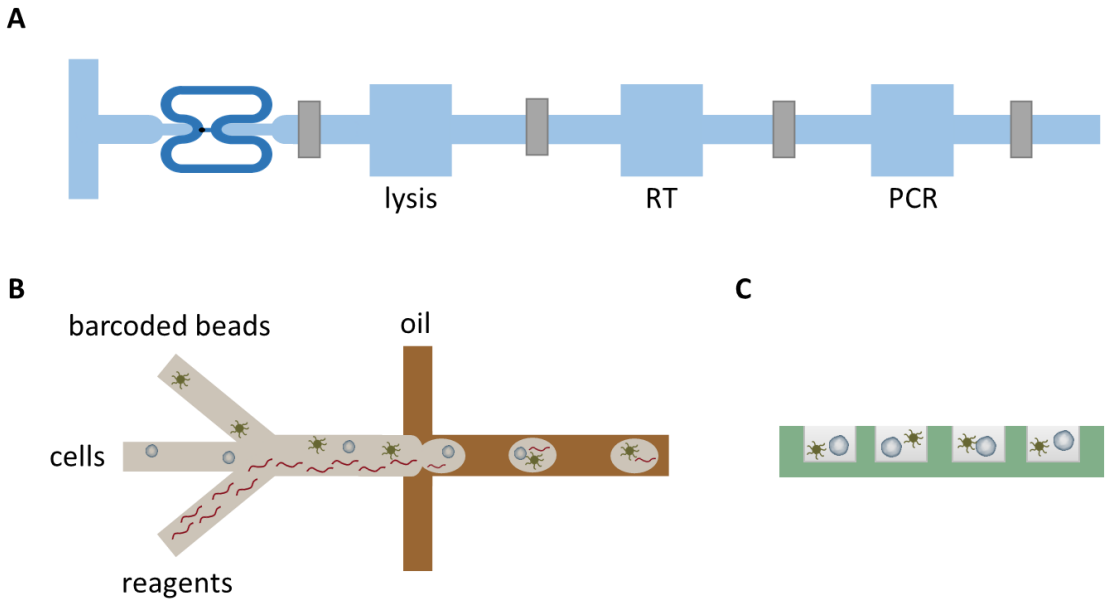


Figure 11: overview of valve-based microfluidic devices, droplet-based devices and nanowells to isolate single cells. (A) In valve-based microfluidic devices, such as the C1, single cells are captured in a specific chamber and cDNA synthesis and amplification occur in subsequent reaction chambers, separated by pressure controlled valves. (B) In droplet-based devices, single cells are isolated together with barcoded beads in aqueous droplets surrounded by an oil phase. Reagents for lysis and reverse transcription are also included in the droplets, whereas amplification occurs in tubes, after pooling of the cells. (C) In nanowells, cells and barcoded beads are isolated by gravity. After lysis, cells can be pooled. RT: reverse transcription, PCR: polymerase chain reaction. Figure adapted from (297).

In 2015, the first microfluidic method generating droplets enabled to isolate thousands of cells in a couple of minutes came to the market. In this type of experiments, cells are captured in oil encapsulated aqueous droplets that form pico- to nanoliter reaction chambers in which cell lysis and RT take place (**Figure 10, 11B, Table 1-2**) (301). Besides single cells, also beads are encapsulated in the droplets. The beads are coated with long oligonucleotides consisting of a universal primer that is identical for every bead, a cell specific barcode that links the transcripts to the cell of origin, a UMI that

is specific for each molecule, and a poly(dT) stretch to reverse transcribe polyadenylated transcripts. The number of cell specific barcodes determines the number of cells that can be captured as too few barcodes gives artificial doublets (279, 300, 302). As cell barcoding occurs during RT, the thousands of cells can already be pooled in one tube afterwards reducing the reagents needed and subsequently the costs of a single cell experiment. One disadvantage of the current droplet-based methods such as Drop-seq, InDrop and Chromium, is that the capture efficiencies of the transcripts are low, whereby only the most abundant genes are detected (279). In addition, the number of cells that are captured in a droplet together with a bead is low, with subsequent loss of many input cells. The cell capture efficiency can be improved by increasing the concentration of cells, although this will also result in higher doublet rates. Therefore, the concentration of loaded cells is a trade-off between the cell capture efficiency and the percentage of cell doublets (296). Due to the attractive features of droplet-based single cell methods such as a high-throughput and low costs, several commercial and in-house made droplet-methods have been developed. Despite the similarity between the droplet-based methods, there are also some substantial differences. On the one hand, the material composing the beads has an impact on the capture efficiency. Drop-seq uses brittle resin while other droplet-based methods such as Chromium and InDrop use hydrogel beads resulting in a higher bead capture efficiency as these hydrogel beads are more flexible. Furthermore, this also has a positive effect on the mRNA capture efficiency as primers can be immobilized throughout the bead for hydrogel beads, whereas this can only occur on the surface for the resin beads. On the other hand, the time point when the RT step is carried out has also an effect on the mRNA capture efficiency. For some methods such as drop-seq, droplets are already broken and pooled in tubes before the RT, while for other methods such as Chromium and InDrop, the RT reaction is carried out in the droplets, which is more efficient as it has been shown that the yield is higher when reactions are performed in small volumes (273, 300, 303). Despite the advantages of the droplet-based methods, expensive instrumentation is often required and visualization is not possible. Therefore, high-throughput methods using nanowell-based cell dispensing devices, such as the ICELL8, or nano- to picoliter wells, such as cyto-seq, micro-well and seq-well, have been recently developed. The first dispenses cells in a microchip, while for the latter cells are randomly distributed over the wells, enabling loading by gravity (**Figure 10, 11C, Table 1-2**). The ICELL8 contains preprinted oligonucleotides containing a cell specific barcode, UMI, poly(dT) and universal sequence, while these oligonucleotides are coated on beads and added to the wells for the nano- to picoliter wells. After lysis and RT, the samples can be pooled and subsequent steps can be carried out in a single tube. Advantages over the droplet-based method are that cells can be microscopically visualized and that these methods are more flexible as multiple protocols can be used for processing the cells after cell lysis (295, 304–307). In addition, the nano- to picoliter wells have the extra advantage that no expensive instrumentation is needed (307).

1.3.3. **What's in a cell: from cell isolation to RNA sequencing**

In general, all single cell RNA-seq methods consist of the same six steps: single cell isolation, cell lysis, reverse transcription, amplification, library prep and sequencing. The protocols used for these six steps can differ between single cell RNA-sequencing methods and help to determine the method warranted for a specific research question.

Table 1: characteristics of the most commonly used microfluidic, droplet-based and nanowell devices. pA: polyA; ATAC: assay for transposase accessible chromatin followed by high throughput sequencing; CAGE-seq: cap analysis gene expression sequencing; HT: high throughput.

		applications	number of cells	capture efficiency (%)	microscopically visible?	cell size dependent?	reference
microfluidic chips	**C1**	pA[+] RNA seq total RNA-seq DNA-seq ATAC-seq CAGE-seq	96 800 (HT chip)	6	yes	yes*	(165, 308–311)
droplet-based	**Chromium**	pA[+] RNA-seq DNA-seq ATAC-seq	8 x 10,000	50	no	no	(300, 311)
	ddSeq	pA[+] RNA-seq ATAC-seq	4 x 300	3-5	no	no	(312)
	Drop-seq	pA[+] RNA-seq	>1000	25	no	no	(300, 302)
	inDrop	pA[+] RNA-seq	>1000	20	no	no	(279, 300)
nanowells	**Seq-Well**	pA[+] RNA-seq	>1000	80	yes	no	(305)
	Microwell-seq	pA[+] RNA-seq	5000-10,000	10	yes	no	(306)
	Cyto-seq	pA[+] RNA-seq	>1000	10	yes	no	(304)

*3 types of chips exist: small (5-10 µm), medium (10-17 µm), large (17-25 µm)

1.3.3.1. Single cell isolation

The first crucial step in a single cell experiment is the generation of a single cell suspension. Whereas this is relatively easy for blood cells or suspension cell lines, making a single cell suspension of adherent cells or tissues that are highly interconnected is more challenging. Selecting the appropriate enzymatic step to create a single cell suspension is important as this can have a substantial influence on the cell viability and cellular transcriptome (281). Once a single cell suspension is obtained, single cells can be isolated by using mouth pipetting, FACS sorting, microfluidic devices with pressure-controlled valves, droplet-based devices, nanowell-based cell dispensing or microwells as described in section 1.3.2. Although mouth pipetting is a very laborious and a low-throughput method, it has the advantage that cells of interest can be visually selected and cell loss is minimal compared to current microfluidic, nanowell-based cell dispensing and droplet-based methods, where only a small fraction of the cells

Table 2: characteristics of the top 15 cited single cell polyA[+] RNA-seq methods in Web of Science and four available single cell total RNA-seq methods. pA: polyA; rRNA: ribosomal RNA; UMI: unique molecular identifier; PCR: polymerase chain reaction; IVT: in vitro transcription.

	SMARTer-seq	MATQ-seq	RamDA-seq	SUPeR-seq	Quartz-seq2	STRT-Seq-2i	SCRB-seq	Microwell-seq	Seq-Well	Cyto-seq	CEL-seq2	Quartz-seq2	STRT-seq	CEL-seq	SMART-seq/ SMART-seq v2	MARS-seq	inDrop	Tang	Drop-seq
stranded	+	-	-	-	-	+	+	-	-	-	+	-	+	+	-	-	-	-	-
pA/total	total	total	total	total	pA	pA	pA	pA	pA	pA	pA	pA	pA	pA	pA	pA	pA	pA	pA
rRNA <5 %	+	NA	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
amplification method	PCR	PCR	NA*	PCR	PCR	PCR	PCR	PCR	PCR	PCR	IVT	PCR	PCR	IVT	PCR	IVT	IVT	PCR	PCR
full gene body coverage	+	+	+	+	+	-	+	-	-	-	-	+	-	-	+	-	-	+	-
UMI	-	+	-	-	-	+	+	+	+	+	+	+	-	-	-	+	+	-	+
early multiplexing	-	-	-	-	-	+	+	+	+	+	+	+	+	+	-	+	+	-	+
reference	(313)	(314)	(315)	(316)	(317) (295) (295) (295)	(295)	(285)	(306)	(318)	(304)	(319)	(317)	(284)	(290)	(287, 320)	(321)	(279)	(322)	(302)

*strand displacement amplification

are captured. In contrast, thousands of cells and even ten thousands of cells can be captured using these nanowell-based cell dispensing devices and droplet- and microwell-based methods, respectively, substantially increasing the throughput of single cell experiments (270, 281, 307, 323). An advantage of FACS sorting over the other described methods is that specific subpopulations can be isolated based on surface markers. In addition, forward and side scatter can be used to remove cell doublets and give information about granularity and cell fate on top of the cellular transcriptome profile. A disadvantage is that the reactions are carried out in microliter volumes, which increases the reagents costs (281). To reduce these reagent volumes and costs, microfluidic devices such as the C1 (Fluidigm) were developed, in which cells are captured in nanoliter capture sites, allowing automatization and parallelization, and decreasing the hands-on time and the technical variability due to reduced pipetting steps (273, 281, 296, 324). A disadvantage is the low-throughput and capture rate, the latter especially hampered with non-spherical cells. The C1 specifically requires to select a chip based on the size of the cells (small 5-10, medium 10-17, large 17-25 µm), making it impossible to capture heterogeneous cell populations consisting of different cell sizes. Despite the reduction in reagent volumes, the costs remain high owing to the instrumentation and chips needed for these experiments (281). To circumvent most of these limitations, droplet-based methods emerged in which cells are captured in aqueous droplets surrounded by an oil phase. This has the advantage that cells are captured independently of their cell size and fate and the throughput is much higher (up to tens of thousands of cells). A disadvantage compared to the aforementioned methods is that the cells cannot be microscopically visualized and only 3' ends of transcripts are sequenced. To visualize cells, nanowell-based cell dispensing methods such as the ICELL8 and microwell-based methods have been introduced in which also thousands of cells can be captured due to dispensing or gravity, and microscopically visualized (295, 305, 306, 304, 307). In addition, microwell-based methods have the extra advantage that no specific devices are needed, making these methods cheaper in general (295, 305, 306, 304). The method one should use depends on: (1) the number and abundance of the cells e.g. rare cell types require a large number of cells, (2) the type of starting material e.g. the use of primary tissue warrant a high cell capture efficiency and (3) the research question e.g. full length coverage sequencing methods are needed for mutation and splicing analyses (279, 288).

1.3.3.2. Cell lysis

In the second step, cells are lysed with a buffer to disrupt the cell membrane for efficient mRNA capture. The crude lysate should not contain inhibitors that interfere with the subsequent RT reaction (281, 287, 293). In this step, RNA spike-in molecules can be added as workflow controls. external RNA controls consortium (ERCC) spikes are commonly used and consist of a set of 92 synthetic RNA molecules that differ in length, GC content and concentration. These spikes are added in equal amounts to the cells during lysis, thus undergoing the same steps as the endogenous RNA molecules and can be used for absolute quantification of the number of molecules and for correcting technical noise (281).

1.3.3.3. Reverse transcription

RT is typically carried out using a poly(dT) primer that binds to the end of polyadenylated transcripts and is used to initiate the RT reaction (281, 293, 325). A substantial part of the human transcriptome, including circRNAs, eRNAs, histone RNAs, and a sizable fraction of long lncRNAs, is not polyadenylated and therefore not quantified using these classic methods (326–328). Therefore, we and others

developed new methods using random primers, enabling to capture both polyadenylated and non-polyadenylated transcripts (**Paper 3, Table 2**) (313, 315, 316). Second strand cDNA synthesis can be done using template switching or polyA tailing, the latter based on the addition of a polyA tail at the 3' end of the first strand cDNA, which can subsequently be bound by the second poly(dT) primer for second strand cDNA synthesis. This method results in reasonable 3' end bias which is partially solved by the template switching mechanism. Here, the enzyme adds some non-templated extra nucleotides to the 3' end of the first strand cDNA, which are bound by the second oligo for second strand cDNA synthesis (293, 323, 329). In single cell experiments, an average of only 10 % of the transcripts are transcribed into cDNA (281, 330). Several efforts have been done to increase the capture efficiency by changing the lysis buffers or enzymes (287, 294).

1.3.3.4. Amplification

Since a single cell only contains 10-20 pg of RNA, a pre-amplification step is needed by PCR or IVT as described in section 1.3.2. To reduce the amplification bias, the incorporation of UMIs is currently implemented in most used methods (**Table 2**) (279, 330–332). These UMIs can be used for absolute quantification of transcripts and to reduce the technical noise by up to 50 %. Therefore, methods without UMIs, such as SMART-seq, have inherently more amplification noise compared to UMI based methods (331).

1.3.3.5. Library preparation

To generate sequencing ready libraries, bulk RNA-seq library prep protocols were modified for use at the single cell level. For methods without early cell barcoding, such as SMART-seq, cell specific barcodes are added at this point enabling the pooling of cells in one tube for further processing. For single cell methods in which the cell already obtains a cell specific barcode introduced during the RT step, the library prep can be conducted in a single tube, reducing the costs and sample handling time (281, 325).

1.3.3.6. Sequencing

The sequencing depth determines to some extent the number of genes that are quantified per cell and depends on the chosen method, the cell type and the research question. By using 3' end or 5' end counting methods or full length coverage methods used for gene expression profiling, fewer reads are needed compared to full length coverage methods that will be used for splicing or mutation analysis. As the sequencing cost considerably contributes to the costs of a single cell experiment, single cell experiments are often a trade-off between the number of cells analyzed and sequencing depth as sequencing of many cells is expensive. For high-throughput experiments, typically 10,000 to 100,000 reads per cell are generated, while for lower throughput experiments mostly on average 1 million reads are generated (332). To get a first view on your cell population, a high-throughput experiment with shallow sequencing can be performed. If the results seem useful, a more in-depth analysis on fewer cells can be performed in a new experiment to obtain a more complete view of single cell transcriptomes (333). The optimal sequencing depth depends on the research question. It has been shown that for the discrimination of different cell types shallow whole transcriptome sequencing (20,000 – 500,000 reads per cell) is sufficient, quantifying the top abundant genes. In contrast, studies that investigate low abundant genes, subtle differences in gene expression among cell states or transcriptional heterogeneity require deeper sequencing (279, 334, 335). To evaluate the ability of

single cell sequencing devices in terms of this transcriptional heterogeneity, I performed the same perturbation experiment on C1 (Fluidigm), ddSeq (Bio-Rad, Illumina) and Chromium (10X genomics) and evaluated these devices with a focus on the detection of differentially expressed genes (**Paper 4**).

1.3.4. **Analysis of (single cell) sequencing data**

Advances in next generation sequencing technologies and the reduction in costs over the years now enable to generate millions to billions of reads per sequencing run and unprecedented amounts of various types of sequencing sequencing data are generated everyday (257, 336). This enormous amount of sequencing data poses bio-informatics challenges in terms of data storage and the development of new computational tools to process these data. Typically, a sequencing analysis starts with the alignment of the data to a reference genome, which can take several days depending on the size of the data. Next, several tools are required for further downstream analyses that are computational intensive. The use of high performance computing (HPC) can considerably speed up these processes that require a lot of memory (336). Currently, multiple tools are integrated in bioinformatics pipelines, enabling automatization and processing of large datasets (337, 338). Furthermore, web-based user-friendly tools, such as Galaxy, also emerged to facilitate the analysis and interpretation of sequencing data by biologists without program skills. A drawback of these tools is that users do not know what the intermediate steps are and can only adapt some parameters (338). It should be noted that the bioinformatics time and storage infrastructure make up a considerable part of the current sequencing costs (337). This is a problem that is even more important in clinics, as huge data storage equipment is required to store raw data of patients properly, since these should be available for new analyses with improved algorithms during time (257).

The bioinformatics challenges associated with the huge amounts of sequencing data are further extended with the development of single cell sequencing methods during the last decade. Analysis of these single cell sequencing data is computationally harder as often thousands of cells are investigated in parallel, increasing the amount of data generated and slowing down the pace of the analyses (339). Although new single cell sequencing methods are emerging rapidly, the number of data analysis pipelines is still limited and often specific for a single cell device (340). Currently, no gold standard exists for the analysis of single cell data, however, the first benchmarking studies comparing different tools are emerging (341–343). First, reads should be demultiplexed to assign reads to the appropriate cell. This is usually done using raw data analysis pipelines such as Cell Ranger and indrops. These pipelines also perform quality assessment, alignment and quantification of the reads (279, 344, 345). The quality of the cells is mostly determined based on the number of genes and reads per cell since too few genes or reads per cell may point to dying cells, while too many genes and reads may indicate cell doublets. A more reliable way to remove doublets is by using one of the recently developed tools such as DoubletFinder (346). To speed up the computational processes, genes that are only expressed in a few cells are typically removed (345, 347, 348). In the next step, ambient (extracellular) RNA that contributes to the noise in single cell experiments and disturbs downstream analyses, needs to be removed. To quantify and correct for the presence of ambient RNA, SoupX was recently developed (349).

In contrast to bulk sequencing methods, single cell sequencing data are characterized by zero-inflated counts due to dropouts or transcriptional bursting. Several tools such as ZIFA, scImpute and drImpute have been developed to account for these dropouts (350–353). To further account for these dropouts, effective normalization is warranted. Spike-in normalization is possible based on the assumption that every cell gets an equal amount of ERCC spikes as these spikes are added during lysis and this should consequently result in the same number of spike-in reads per cell (354). However, their utility is still

under debate for several reasons. First, the capture rate of these synthetic RNA molecules deviates from endogenous RNA molecules as these ERCC spikes generally have shorter polyA tails (20-26 bp vs 250 bp on average) and therefore likely less efficiently captured (333, 355, 356). Second, these synthetic RNA molecules are not bound by mRNA binding proteins and do not make secondary structures and are therefore more easily captured. Third, these ERCC spikes lack a 5' cap, whereby the template switching mechanism is less efficient than for endogenous molecules (333, 335, 357). Therefore, ERCC spikes are not perfect for normalization and the use of endogenous reference genes provides an alternative way to normalize. However, these reference genes are not always expressed at similar levels in all subpopulations, further complicating the normalization of single cell data (358). As these normalization methods are still under debate, new normalization methods should be developed.

After these steps, single cell sequencing data are typically used to identify new subpopulations and their markers (1), to perform trajectory analysis (2) or to perform differential gene expression analysis (3). First, dimensionality reduction algorithms such as principal compound analysis (PCA) and T-distributed stochastic neighbor embedding (tSNE) are often used to identify subpopulations in a sample. These methods enable to visualize different clusters in the dataset. Based on differential gene expression analysis between the clusters, marker genes can be identified (345, 347). Large sequencing projects such as the Human Cell Atlas facilitate the determination of cell types in the clusters. Recently, tools have been developed to automatically annotate clusters (359). Second, trajectory analyses order cells along a trajectory to reconstruct differentiation processes. A recent study evaluated various trajectory tools and provided guidelines for specific applications (342). Third, performing differential gene expression analysis at the single cell level is interesting as this shows how each individual reacts on a specific treatment or perturbation or to identify marker genes that differ subpopulations. A comparison study has shown that methods for bulk differential gene expression analysis work equally well compared to single cell specific methods and that EdgeR is the best method for single cell differential gene expression analysis (343, 345). However, due to the increasing number of cells per experiment, run-time has to be taken into account for the decision of the method. Therefore, single cell specific methods such as MAST are sometime more appropriate (360).

The last decade, plenty of tools have been developed, however, this field is still in its infancy and no gold standard exists yet. As the field is developing rapidly, the number of data analysis tools increases, but also the size of datasets keeps increasing, posing further computational challenges for the run-times of the tools. scRNA-tools.org gives a nice overview of the currently existing single cell analysis tools and further comparative studies will help to decide which tools to use for a specific application.

1.3.5. **Single cell genomics, epigenomics and proteomics add extra layers of information**

To extend our understanding of the regulation of genes and cellular processes in single cells, genetic, epigenetic and proteomic data are required. Despite the rapid increase in the number of new single cell RNA-seq methods, the development of single cell DNA and epigenetic sequencing methods has been more challenging as a cell only contains two copies of the chromosomes in contrast to the thousands of copies of several mRNAs (361). Nevertheless, the equimolar nature of genes at the DNA level is beneficial as only a limited number of reads are needed to examine a single cell's (epi)genome (297). Likewise, investigating protein expression at the single cell level has also been challenging due to heterogeneous expression, the difficulty to amplify proteins and the lack of powerful tools for proteomics (**Figure 12**) (324).

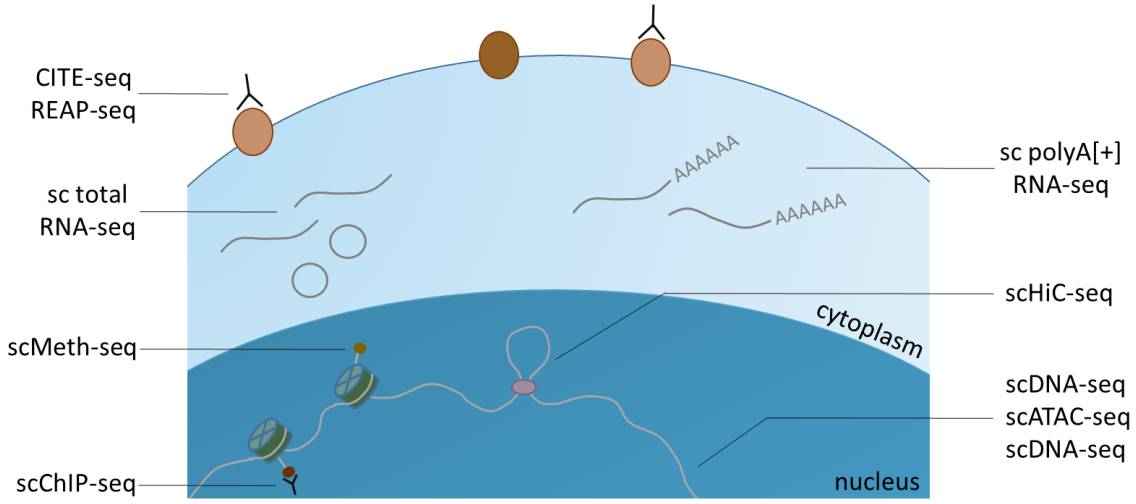


Figure 12: single cell sequencing methods. Single cell sequencing methods have been developed to investigate DNA, RNA, epigenetic marks and proteins of single cells. Currently, several methods combine some of these layers to obtain complementary information from the same single cell. Sc: single cell; meth: methylation; ChIP: chromatin immune precipitation; ATAC: assay for transposase accessible chromatin; CITE-seq: cellular indexing of transcriptome and epitope by sequencing; REAP: RNA expression and protein sequencing. Figure adapted from (353).

1.3.5.1. Single cell genomics

Single cell DNA sequencing data provides insights in the heterogeneity of SNV and copy number variations (CNV) across cells. To obtain sufficient genomic coverage, whole genome amplification (WGA) methods such as PCR, multiple displacement amplification (MDA) or a combination of these two are required to amplify the DNA molecules (362–364). The PCR-based methods yield a uniform amplification, but sparse genomic coverage, while MDA-based methods generate a less uniform amplification, but better genomic coverage. Consequently, PCR-based methods are more suitable for CNV detection, whereas MDA methods are warranted for single nucleotide polymorphism (SNP) calling. More recently, these two methods have been combined to obtain a uniform amplification and high coverage (363–366). After whole genome amplification, sequencing can be performed. Since the actual sequencing costs consume a considerable fraction of the costs of these experiments, using whole exome sequencing or targeted approaches, where only a panel of specific genes are enriched, can reduce the required number of sequencing reads per cell and consequently the costs (367). Recently, these targeted DNA sequencing approaches were combined with droplet-based methods, significantly increasing the throughput of single cell DNA-seq experiments from a hundred cells to thousands of cells (368).

1.3.5.2. Single cell epigenomics

The regulation of gene expression has been studied extensively by integrating measurements of DNA methylation, chromatin accessibility, histone post-translational modifications and 3D conformation. Investigating transcription factor binding on these regions can give further insights in how the regulatory regions impact the expression of target genes. In order to identify direct interactions between these regulatory regions and their targets, chromosome conformation capture methods can be used. Similar as to RNA-seq experiments, all these methods were originally developed and used in bulk cell populations, resulting in average signals and masking the regulatory heterogeneity. Studying these epigenetic layers at the single cell level can depict regulatory heterogeneity and can define

regulatory regions in new cell types. This contributes to a more complete view of the characteristics of cell populations as epigenetic and transcriptomic heterogeneity can differ and provide complementary information. Regulatory heterogeneity can for instance originate from regions that are poised, repressed or primed, while this will have no or little effect on gene expression and consequently will not give transcriptional heterogeneity. Thus, cells can show epigenetic heterogeneity with no or little effect on gene expression or the other way around, emphasizing the importance to add single cell epigenomic data as an extra layer of information (369–371).

Single cell DNA methylation

Single cell methylation sequencing was the first single cell epigenomic method developed (372). DNA methylation mostly occurs at CpG dinucleotides and is an epigenetic modification involved in silencing of gene expression (373). Methylation plays an important role in biological processes such as cell differentiation, genomic imprinting and cancer development (361, 372, 373). As a mammalian cell only contains two copies of DNA molecules, the biggest challenge for single cell methylation experiments is to retain these copies. Therefore, bulk methylation sequencing protocols were optimized with less purification steps to reduce possible loss of DNA (361). Most of the methods, such as single cell bisulfite sequencing, are low throughput methods that capture up to 48 % of the genome of a cell (374). Recently, new methods, such as single cell combinatorial indexing for methylation sequencing (sci-MET) emerged, increasing the throughput up to thousands of cells, but simultaneously reducing the percentage of the genome that is captured to 1 - 5 %, underscoring the need for further improvement (375, 376). These single cell methylation methods can be used to separate subpopulations based on their methylation profile as regulatory elements with cell type specific activity can be identified (373, 376).

Single cell open chromatin mapping

Identifying regions of open chromatin contributes to the detection of regulatory elements involved in control of gene expression (310). As regulatory regions are often cell type specific, chromatin states are more suitable to distinguish cell types compared to single cell RNA-seq (377). Moreover, changes in chromatin states can precede changes at the RNA level, whereby initial changes may not be noticeable by only performing single cell RNA-seq (377, 378). Current methods to measure chromatin states are based on the fact that open chromatin regions are more accessible to enzymes that can fragment these regions. Single cell DNAse-seq maps open chromatin regions by the use of DNAseI. Since barcodes can only be added, and cells consequently be pooled, during library prep, the protocol is labor intensive and only possible for a limited number of cells (373). In contrast, assay for transposase accessible chromatin followed by high-throughput sequencing (ATAC-seq) uses a Tn5 transposase that fragments and adds adaptor sequences to accessible chromatin regions, enabling to pool cells after this step (310, 373). The method has been modified to perform ATAC-seq at the single cell level at relatively low throughput (<100 cells) using microfluidic chips or at higher throughput (>1000 cells) using combinatorial indexing in microtiter plate wells. The latter method distributes pools of cells in a plate where a first specific barcode is incorporated. Next, these cells are distributed in new pools in a new plate, where a second barcode is incorporated. By combining the first and second barcode, most of the cells obtain a unique combination, enabling to pool thousands of cells without the need to isolate cells physically (379). Despite the higher throughput of this method, the detection rate is lower compared to microfluidic devices as reactions are less efficient in these larger reaction volumes, which has also been shown for single cell RNA-seq (263, 294, 373). Using single cell ATAC-

seq, only up to 10 % of the promoters are captured (380). Recently, protocols for commercial devices such as the ddSeq single cell isolator and Chromium have also been released.

Single cell histone modification and protein-DNA interactions

Bulk chromatin immunoprecipitation sequencing (ChIP-seq) identifies histone marks or binding of transcription factors and is often investigated as binding of transcription factors drives gene expression, and binding on inappropriate places can contribute to disease development (381). ChIP-seq has the limitation that a lot of input is required and the success rate is mainly dependent on the quality of the antibody. Non-specific antibodies give off-target effects, of which the number increases using low input samples as a consequence of the lower amount of the target of the antibody available (381, 382). The single cell Drop-ChIP method circumvents this limitation by first labeling the chromatin of single cells in droplets followed by chromatin pooling. Therefore, the amount of the target of the antibody increases, thus decreasing off-target effects and noise. Despite the fact that the detection rate is only about 5 % due to low coverage and only a thousand promoters/enhancers can be identified, this is sufficient to separate different cell types. Currently, single cell ChIP-seq has only been performed for abundant histone marks and should be further optimized to be also applicable for transcription factors (382). Cleavage under targets and release using nuclease (CUT&RUN) is another antibody-based method to detect transcription factor binding and circumvents several limitations of ChIP-seq. CUT&RUN uses transcription factor specific antibodies to tether MNAse that specifically cleaves the DNA at binding sites, reducing the background and consequently the sequencing costs, which are both high in ChIP-seq experiments. Another advantage of CUT&RUN compared to ChIP-seq is that the antibodies bind in an intact cell, providing information of binding in a cell's natural state (381, 383, 384). This protocol has been optimized to be able to perform CUT&RUN at the single cell level (384). One of the limitations of ChIP-seq and CUT&RUN remains that a specific antibody is required, which it not always available. Therefore, another approach to study protein-DNA interactions at the single cell level is the DNA adenine methyltransferase identification (DamID) method. By using single cell DamID, regions that interact with the nuclear lamina were identified. This is obtained by fusing Lamin B1 with a DAM that methylates all lamin B1 interacting loci. After fragmentation with an enzyme specific to DAM methylated sequences, these loci can be sequenced (385, 386). Bulk DamID protocols have been optimized to study other DNA-protein interactions and can potentially also be used at the single cell level (373, 386).

Single cell chromatin maps

In addition to the epigenetic changes described above, also chromosome conformation can contribute to the regulation of gene expression. To investigate chromosome structure, 3C sequencing has been introduced in 2002, which can capture interactions between known interacting regions (387, 388). 3C has the limitation that only interactions between regions that are in close proximity can be detected and that prior knowledge of the interacting regions is required (387). To circumvent this, several other conformation capture methods have been developed and are based on the 3C method. Circular chromosome conformation capture (4C) enables to detect longer distance interactions and genome-wide interactions of a region of interest, while 3C carbon-copy (5C) can be used to map multiple known interacting pairs (388–390). Finally, HiC was developed enabling to investigate global genome-wide chromosome conformation by identifying direct contact points throughout the whole genome (388, 391). Up to now, only efforts have been done to perform HiC sequencing at the single cell level, where up to 1,900,000 contact points per cell can be identified (392–394). The throughput of these first single cell HiC experiments was low, but has been drastically improved to thousands of cells by implementing

combinatorial indexing (395). These single cell HiC methods have demonstrated that the domain structure of chromosomes at the megabase scale is relatively stable amongst cells and conserved across cell types, while chromosome structure at larger scale (interdomain and trans-chromosomal) can considerably differ across cells (393, 394, 396). Moreover, it has been shown that the cell cycle is a major contributor to this variability, which is in concordance with the (de)condensation of chromosomes during the cell cycle and which is masked in bulk HiC experiments (397).

1.3.5.3. Single cell proteomics

As proteomics data form a bridge between genomic data and cellular functions, also single cell proteomic methods are warranted (324). The first methods described used fluorescent proteins and FACS analyses with fluorescent labeled antibodies to identify and quantify proteins at the single cell level. One of the main drawbacks of these methods is that only few proteins can be detected due to the limited number of visual markers. Mass cytometry circumvented this limitation by using isotopes conjugated to the antibodies, increasing the number of proteins that can be detected simultaneously (398–400). Single cell western blot devices, such as the Milo (ProteinSimple), have the advantage that molecular mass and antibody binding are combined, enabling the detection of isoforms; the limitation is that only few proteins can be studied per cell (399). To further increase the number of proteins, more recent methods benefit from single cell sequencing methods. Here, antibodies are linked to an antibody specific oligo whereby the protein signal is transformed into a nucleotide sequence that can be sequenced (324, 401). By combining an antibody specific and cell specific barcode, proteins can be investigated at the single cell level in a high-throughput manner (>10,000 cells). Moreover, by the incorporation of UMIs, absolute protein quantification became possible (324, 398). In contrast to the fluorescent-based methods that suffer from overlap of fluorescent labels, this method has theoretically unlimited multiplexing potential. Moreover, the sensitivity is higher as barcode sequences can be amplified at low levels (324). Given that a cell contains more than 20,000 proteins, detecting all proteins in a cell is not feasible, but also not required as most of the heterogeneity can be captured based on the combination of a subset of proteins (324).

1.3.5.4. Integrative analysis of transcriptome, (epi)genome and proteome layers at the single cell level

As the genotype-phenotype relation in a cell depends on several layers, new methods to investigate several layers in parallel in the same cell are required (373, 402). Investigating these layers in different cells of the same cell type gives some biases as cells can for instance be in a different cell cycle stage. To circumvent these differences between similar cells, methods sifting through these various layers within one cell have been developed over the recent years (401).

To investigate the transcriptome and (epi)genome within one cell, gDNA and RNA need to be separated (**Figure 13**). The first method to separate gDNA and RNA is pre-amplification of the gDNA and RNA followed by dividing the sample for DNA and RNA sequencing. The second method physically separates the gDNA and RNA by the use of a membrane specific lysis buffer that only ruptures the cell membrane and maintains the nucleus. Next, the cytosol, containing most of the RNA molecules, and the nucleus, containing the gDNA, are separated for subsequent processing (401–403). The third method uses magnetic beads coated with oligo(dT) primers that only capture the RNA molecules, separating them from DNA (401, 402, 404). After separation, the RNA and gDNA fractions can be used to perform single cell RNA and gDNA and/or epigenome sequencing with a method of choice (373, 405). Combining

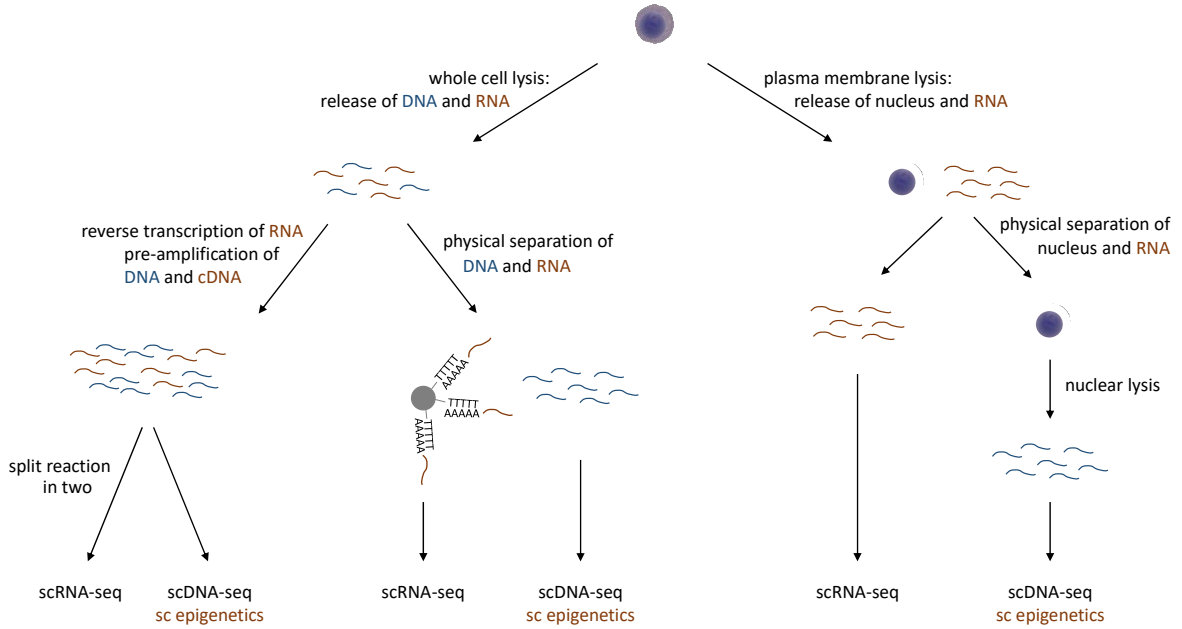


Figure 13: approaches for DNA or epigenetics sequencing and RNA sequencing of the same single cell. For the first two methods, the whole cell is lysed. Subsequently, the RNA is reverse transcribed, followed by pre-amplification of the DNA and cDNA for the first method. Next, the reaction is split in two for single cell DNA or epigenetics sequencing and RNA sequencing. For the second method, the DNA and RNA are physically separated by binding of the RNA to oligo(dT) coated magnetic beads. The third method only lyses the plasma membrane where after cytoplasmic RNA and nuclei are separated. The DNA isolated after nuclear lysis can be used for single cell DNA or epigenetics sequencing. Sc: single cell. Figure adapted from (401).

single cell genomics and transcriptomics data by for instance genome and transcriptome sequencing (G&T-seq) depicted a strong correlation between the copy number of a gene and the expression within one cell. Moreover, it has been shown that genes with low copy numbers have a noisier expression, indicating that CNVs contribute to the variability in gene expression between cells (401, 404, 406). Combining these two layers also more accurately predicts SNVs as the SNVs detected by genomic methods can be validated by RNA-seq if the allele carrying the SNV is expressed. The detection of SNVs is often used to study cell lineages and cancer development for which transcriptomic data of the same cells can provide additional information about the cell state of these cells (401). Besides genome analyses, also open chromatin or methylome sequencing can be conducted on the gDNA fraction after separation of gDNA and RNA, providing insight into the relation between the epigenome and transcriptome. Profiling the open chromatin and transcriptome in a single cell has shown that DNAse hypersensitivity sites of highly active genes are detected in most of the cells, whereas DNAse hypersensitivity sites of lower expressed genes are detected in less cells. Moreover, the DNAse hypersensitivity sites that are only detected in a few cells show more variation in the expression pattern of the associated genes. Thus, determining chromatin states and gene expression profiles in the same cells allows to link chromatin state with gene expression (407). Methylome and transcriptome sequencing data can obtain other, non-redundant information about the cell state. In addition, positive as well as negative correlations between the transcriptome and methylome data were identified underscoring the complex relation between the transcriptome and methylome, especially at distal regulatory regions (401, 408, 409). It has been shown for hypomethylated promoters that half of the genes show an expression pattern that is consistent across all cells, however the other half of genes are highly expressed in some cells while low in other cells. This is masked by

bulk sequencing and depicts that other factors than methylation contribute to the regulation of gene expression (410).

Besides the combination of single cell transcriptome and (epi)genome sequencing within one cell, also several epigenetic layers have been combined. By using single cell ATAC-seq or DNAse-seq, regions of open chromatin can be identified, while undetected regions are assumed to be closed. As these regions can also be regions of undetected open chromatin (false positives), an extra layer of information should be added to distinguish between closed chromatin regions and false positives. This can be done by combining the chromatin state with methylome sequencing within one cell for which several methods have been developed (373, 411). Also, cut and tag, a method that combines the detection of histone marks or transcription factors with accessible chromatin has been developed by combining specific antibodies and a hyperactive Tn5 transposase that integrates sequencing adaptors in the open chromatin. This method produces single cell ChIP signals with a low signal to noise ratio and maps the open chromatin of the same cells for thousands of single cells in parallel (412). To link protein expression with RNA expression, new methods that benefit from the antibodies linked to oligos as described in section 1.3.5.3, arose. These antibodies coupled to oligos allowed to develop methods that can simultaneously measure protein and RNA expression in single cells by sequencing the RNA and oligos in parallel. This can be done by cellular indexing of transcriptome and epitope by sequencing (CITE-seq) or RNA expression and protein sequencing (REAP-seq). In these methods, the antibody is coupled to an antibody specific barcode with a polyA tail. Subsequently, these polyA tailed barcodes are captured together with the endogenous polyadenylated transcripts by the oligo(dT) primer. After the addition of a cellular barcode, the transcripts and proteins of interest can be investigated at the single cell level. Both methods are compatible with commercially available systems (413, 414). As the correlation between mRNA and protein expression is sometimes low, the measurement of proteins and RNA transcripts in the same cell can give complementary information to better characterize subpopulations (414, 415). Remarkably, low abundant proteins are more easily to identify compared to low abundant mRNA molecules in these libraries as proteins have in general longer half-lives and have on average 1000-fold higher copy numbers per cell compared to the mRNA copy numbers (297, 414, 415).

Over the last years, the number of layers that can be analyzed simultaneously in one cell has further evolved. It is now possible to combine genetic, epigenetic and transcriptomic data from the same cells. scTRIO-seq combines CNV, methylome and transcriptome sequencing, while COOL-seq even combines four layers by integrating the chromatin state, methylome sequencing, CNVs and the ploidy of cells (403, 411). In addition, the capture efficiency is much higher compared to the current methods that only measure one epignomic layer, as the methyl and chromatin profile for up to 70 % of the RefSeq genes can be determined using COOL-seq (411). Furthermore, these multi-omics methods can in principle be combined with tens of surface markers by using FACS sorting to isolate single cells adding an additional layer of information (380). These methods will give better insights into the complexity of a single cell and will allow to determine relations between these different layers at the single cell level.

1.3.6. **Deciphering cancer: one cell at a time**

It has been known for decades that tumors comprise cellular and molecular heterogeneity and consist of a mixture of tumor cells and normal cells, including fibroblasts, lymphocytes and endothelial cells (416–420). Moreover, these cells interact with each other and with the tumor microenvironment, further complicating tumor development and maintenance (421). Therefore, taking one biopsy is not

representative for the whole tumor as different sites have other compositions of cells and do possibly not comprise all subclones of the tumor. Identification of these subclones is important as some of these clones may be able to develop drug resistance. Consequently, identification of these subclones can help to provide a better prediction of the prognosis and to develop better therapies by combining drugs against specific subclones (**Figure 14**) (416, 422, 423). This heterogeneity can be partially characterized by performing bulk sequencing of different time points and tumor regions (353, 362, 418, 424–426). Nevertheless, rare cell types and intra-tumor heterogeneity within a sample remain masked using these bulk sequencing methods and this can only partially be solved by performing deconvolution based on known cell types (362, 418, 425, 426). To fully capture the heterogeneity within a tumor and to understand the development of cancer, single cancer cells have been investigated for many years using microscopy and fluorescent-based methods, including FISH. However, generating genome-wide mutation and transcriptome profiles of multiple individual cancer cells has only been possible the last decade (427). As several cancer mutations are known based on the bulk sequencing experiment, targeted sequencing approaches using gene panels are often used to reduce the costs of these single cell sequencing experiments (418, 428). Investigating mutations and CNVs over time in a tumor at single cell resolution allows to identify subclones and the order in which the mutations, CNVs and subclones emerge. Based on these data, phylogenetic trees can be constructed to unravel the clonal evolution of a tumor and to gain insights in the origin of the cancer (362, 429). This is illustrated for T-ALL, for which the type of progenitor cells in which the first mutations occur varies across patients and mutations occur at specific timepoints during development. Performing single cell DNA sequencing on multipotent and myeloid progenitor cells of T-ALL patients revealed oncogenic mutations in these progenitor cells that were also found at diagnosis in some patients, while these mutations were barely found in the progenitors of other patients. This indicates that mutations start to accumulate in the multipotent progenitor cells for some T-ALL patients, whereas only in the lymphoid lineage for other patients (430). This also highlights why autologous stem cell transplantation can lead to relapse as the bone marrow can still contain some progenitors that contain these oncogenic mutations if these are not eradicated before treatment (430, 431). Single cell DNA sequencing over time also revealed that the highly abundant *NOTCH1* mutations only occur at a late stage of T-ALL development. Therefore, *NOTCH1* mutations are probably not present in all clones, which has to be taken into account for the treatment of these patients (430). Currently, most cancer treatments suffer from resistance as the penetration of the drug in the tumor is not uniform and not all cells are sensitive to the therapy due to intra-tumor heterogeneity. The small number of tumor cells that remain present in the patient upon treatment are called minimal residual disease cells and can lead to clinical relapse (267, 432, 433). Identifying these subpopulations of cells that do not react on

the drug using single cell sequencing methods can help to develop new therapies that specifically target these cells to get complete remission (432).

The importance of identifying these rare resistant subpopulations has been illustrated for melanoma, for which treatment with MAPK inhibitors results in distinct transcriptional subpopulations of which one subpopulation has stem-like properties and has been shown to be the driver for relapse. As this subpopulation is only present at 0.58 % of the cells before treatment, this subpopulation can only be detected using single cell sequencing methods. Targeting these resistant cells in combination with MAPK inhibition results in a longer median progression free survival underlying the importance to identify and attack these specific drug resistant subpopulation (267). In addition, treatment of melanoma with RAF/MEK inhibitors also leads to resistance in melanomas with a high expression of *AXL*. Bulk RNA-seq divides melanomas in AXL high and low tumors to predict if the tumor will react on the drug. However, by performing single cell RNA-seq of these tumors, it became visible that in AXL low tumors, also a subclone with high *AXL* expression can be present, hidden by bulk RNA-seq (434). Pre-existing drug resistant clones have also been identified for breast cancer based on CNV profiles before treatment, while transcriptional changes are only acquired upon treatment. Noticeably, this also depicts the need to combine multiple layers as the drug resistant clone could only be detected at the genome level prior to treatment, while not on the transcriptome level (435). Conversely, genetically similar cells can also consist of different transcriptional cell states that can result in drug resistance (401). The presence of drug resistant clones before treatment emphasizes the need to detect these clones prior treating the patients to predict whether the patient will respond and if other therapies should be considered to circumvent the resistance (353, 435, 436). Besides the tumor itself,

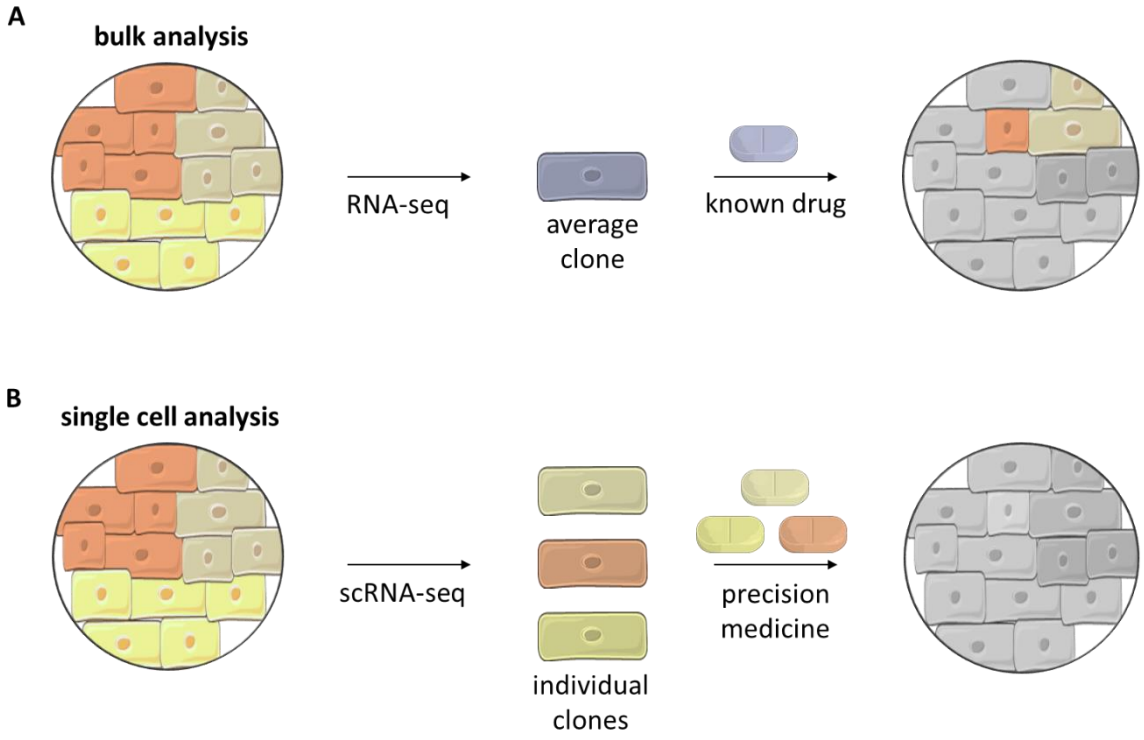


Figure 14: the role of single cell RNA sequencing in cancer treatment. (A) Bulk RNA sequencing generates an average expression profile of the tumor, hiding possibly resistant subpopulations. Based on the average expression profile, the patient will receive a drug, potentially not targeting all clones. (B) By performing single cell RNA sequencing, the expression profile for each cell is obtained, enabling to characterize subpopulations and combine drugs to target all clones. scRNA-seq: single cell RNA-seq. Figure adapted from (423)

the immune cells and microenvironment can also contribute to drug resistance (437, 438). For instance, it is known that tumors containing exhausted T-cells, which is a state of T-cell dysfunction, are not responsive on immunotherapy by checkpoint inhibition (437, 439). However, the presence of a newly identified type of T-cells that precedes the exhausted T-cells (pre-exhausted T-cells) has been shown to have a better prognosis compared to the presence of exhausted T-cells in lung cancer patients. Therefore, single cell sequencing of the T-cell composition of cancer patients can contribute to the prediction of the prognosis and the decision whether or not to give immunotherapy to a patient (437, 439).

To be able to identify these different subpopulations within a tumor using single cell sequencing methods, solid tumors require an invasive biopsy with low yields and enzymatic digestion. Therefore, analyzing circulating tumor cells (CTCs) is currently gaining interest as these cells are shed in the blood stream and can be isolated in a non-invasive way from blood, permitting to isolate them at several timepoints providing a better follow-up of disease progression and treatment (362, 363, 427). One drawback is that these CTCs are extremely rare in blood (1 in 10^6), requiring specialized isolation methods (353, 418, 440). More than halve of the mutations found in the primary and metastatic tumor can also be found in CTCs making them suitable for the analysis of several cancer types (427, 441). In addition, for colon cancer it has been shown that up to 85 % of the CTC specific mutations can also be identified in the primary tumor using ultradeep sequencing, depicting that the mutations found in CTCs are real (442). The importance of sequencing single CTCs has been depicted for several cancer types. First, sequencing profiles of CTCs can be used identify cancer subtypes. This has been illustrated for breast cancer, for which the CNV profiles of single CTCs can be used for the clinical classification of breast cancer patients (362, 443). Second, the therapy response can be predicted based on single cell CTCs' sequencing profiles as illustrated for lung cancer and prostate cancer. For lung cancer, it has been shown that the CTCs' CNV profiles can determine whether a patient will respond to chemotherapy or not while for prostate cancer mutations and splice variants in the androgen receptor are suggestive for resistance to anti-androgen therapies (362, 444–446). The response on these anti-androgen therapies is heterogeneous, which is reflected in the CTCs' profiles. These androgen receptor mutations and splice variants are only detected in CTCs of resistant prostate cancer patients, while not in primary localized prostate cancer patients, showing the value of identifying these mutations and splice variants in CTCs (446). Third, CTCs can also be used for treatment follow-up as shown for prostate cancer. Here, a new drug resistant clone can be identified during relapse, underscoring the need to identify this clone during treatment follow-up and consequently adapt the treatment (447).

The number of single cell sequencing methods has drastically increased over the last years and has promising clinical applications for cancer patients as often too little material is available for bulk analyses and as these bulk analyses can hide cells of interest. Moreover, single cell sequencing methods can detect rare resistant cells and tumor heterogeneity resulting in a better prediction of treatment response. Furthermore, analyzing CTC is promising as this eliminates the need for invasive biopsies and enables a better follow-up of disease progression and treatment over time. As the costs of these single cell experiments and the sequencing cost are dropping over the last years, single cell sequencing analyses will probably be feasible in the near future for clinical application (420).

1.4. References

1. Redaelli,A., Laskin,B.L., Stephens,J.M., Botteman,M.F. and Pashos,C.L. (2005) A systematic literature review of the clinical and epidemiological burden of acute lymphoblastic leukaemia

(ALL). *Eur. J. Cancer Care (Engl).*, **14**, 53–62.
2. National Cancer Institute (NCI) U.S. Department of health and human services (2013) What You Need To Know About Leukemia.
3. Rodriguez-Abreu,D., Bordoni,A. and Zucca,E. (2007) Epidemiology of hematological malignancies. *Ann. Oncol.*, **18**, 3–8.
4. Hunger,S.P. and Mullighan,C.G. (2015) Acute Lymphoblastic Leukemia in Children. *N. Engl. J. Med.*, **373**, 1541–1552.
5. Pui,C.-H., Robison,L.L. and Look,A.T. (2008) Acute lymphoblastic leukaemia. *Lancet*, **371**, 1030–1043.
6. Ferrando,A.A., Neuberg,D.S., Staunton,J., Loh,M.L., Huard,C., Raimondi,S.C., Behm,F.G., Pui,C.-H., Downing,J.R., Gilliland,D.G., *et al.* (2002) Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell*, **1**, 75–87.
7. Hunger,S.P. and Mullighan,C.G. (2015) Acute Lymphoblastic Leukemia in Children. *N. Engl. J. Med.*, **373**, 1541–1552.
8. Germain,R.N. (2002) T-cell development and the CD4–CD8 lineage decision. *Nat. Rev. Immunol.*, **2**, 309–322.
9. Bell,J.J. and Bhandoola,A. (2008) The earliest thymic progenitors for T cells possess myeloid lineage potential. *Nature*, **452**, 764–767.
10. Koch,U. and Radtke,F. (2011) Mechanisms of T cell development and transformation. *Annu. Rev. Cell Dev. Biol.*, **27**, 539–62.
11. Godfrey,D.I., Kennedy,J., Suda,T. and Zlotnik,A. (1993) A developmental pathway involving four phenotypically and functionally distinct subsets of CD3-CD4-CD8- triple-negative adult mouse thymocytes defined by CD44 and CD25 expression. *J. Immunol.*, **150**, 4244–52.
12. Van Walle,I. De, De Smet,G., De Smedt,M., Vandekerckhove,B., Leclercq,G., Plum,J. and Taghon,T. (2009) An early decrease in Notch activation is required for human TCR-αβ lineage differentiation at the expense of TCR-γδ T cells. *Blood*, **113**, 2988–2998.
13. Pui,J.C., Allman,D., Xu,L., DeRocco,S., Karnell,F.G., Bakkour,S., Lee,J.Y., Kadesch,T., Hardy,R.R., Aster,J.C., *et al.* (1999) Notch1 Expression in Early Lymphopoiesis Influences B versus T Lineage Determination. *Immunity*, **11**, 299–308.
14. Van de Walle,I., Dolens,A.-C., Durinck,K., De Mulder,K., Van Loocke,W., Damle,S., Waegemans,E., De Medts,J., Velghe,I., De Smedt,M., *et al.* (2016) GATA3 induces human T-cell commitment by restraining Notch activity and repressing NK-cell fate. *Nat. Commun.*, **7**, 11171.
15. García-Peydró,M., de Yébenes,V.G., Toribio,M.L., Bakker,A.Q., van Gastel-Mol,E.J., Wolvers-Tettero,I.L., van Dongen,J.J. and Spits,H. (2003) Sustained Notch1 signaling instructs the earliest human intrathymic precursors to adopt a gammadelta T-cell fate in fetal thymus organ culture. *Blood*, **102**, 2444–51.
16. Blom,B. and Spits,H. (2005) Development of Human Lymphoid Cells. *Annu. Rev. Immunol.*, **24**, 287–320.
17. Takeuchi,Y., Fujii,Y., Okumura,M., Inada,K., Nakahara,K. and Matsuda,H. (1993) Characterization of CD4+ Single Positive Cells That Lack CD3 in the Human Thymus. *Cell. Immunol.*, **151**, 481–490.
18. Weerkamp,F., Pike-Overzet,K. and Staal,F.J.T. (2006) T-sing progenitors to commit. *Trends Immunol.*, **27**, 125–131.
19. *Murphy K. (2012) Janeway's immunobiology. 8th edition. London: Garland Science. p291-316.*
20. Yui,M.A. and Rothenberg,E. V (2014) Developmental gene networks: a triathlon on the course to T cell identity. *Nat. Rev. Immunol.*, **14**, 529–45.
21. Silverman,L.B., Gelber,R.D., Dalton,V.K., Asselin,B.L., Barr,R.D., Clavell,L.A., Hurwitz,C.A., Moghrabi,A., Samson,Y., Schorin,M.A., *et al.* (2001) Improved outcome for children with acute lymphoblastic leukemia: Results of Dana-Farber Consortium Protocol 91-01. *Blood*, **97**, 1211–1218.
22. Schrappe,M., Reiter,A., Ludwig,W.D., Harbott,J., Zimmermann,M., Hiddemann,W., Niemeyer,C., Henze,G., Feldges,A., Zintl,F., *et al.* (2000) Improved outcome in childhood acute lymphoblastic leukemia despite reduced use of anthracyclines and cranial radiotherapy: results of trial ALL-BFM

90. German-Austrian-Swiss ALL-BFM Study Group. *Blood*, **95**, 3310–22.
23. Buitenkamp,T.D., Izraeli,S., Zimmermann,M., Forestier,E., Heerema,N.A., van den Heuvel-Eibrink,M.M., Pieters,R., Korbijn,C.M., Silverman,L.B., Schmiegelow,K., *et al.* (2014) Acute lymphoblastic leukemia in children with Down syndrome: a retrospective analysis from the Ponte di Legno study group. *Blood*, **123**, 70–7.
24. Horton,T. and Steuber,P. (2013) Overview of the presentation and diagnosis of acute lymphoblastic leukemia in children. *UpToDate*.
25. Chiaretti,S., Zini,G. and Bassan,R. (2014) Diagnosis and subclassification of acute lymphoblastic leukemia. *Mediterr. J. Hematol. Infect. Dis.*, **6**, e2014073.
26. Pui,C.-H., Yang,J.J., Hunger,S.P., Pieters,R., Schrappe,M., Biondi,A., Vora,A., Baruchel,A., Silverman,L.B., Schmiegelow,K., *et al.* (2015) Childhood Acute Lymphoblastic Leukemia: Progress Through Collaboration. *J. Clin. Oncol.*, **33**, 2938–2948.
27. Bassan,R. and Hoelzer,D. (2011) Modern therapy of acute lymphoblastic leukemia. *J. Clin. Oncol.*, **29**, 532–43.
28. T.B.,H., R.B.,M. and G.H.,R. (2009) Late effects in long-term survivors after treatment for childhood acute leukemia. *Clin. Pediatr. (Phila).*, **48**, 601–608.
29. Zhang,J., Ding,L., Holmfeldt,L., Wu,G., Heatley,S.L., Payne-Turner,D., Easton,J., Chen,X., Wang,J., Rusch,M., *et al.* (2012) The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature*, **481**, 157–163.
30. Mullighan,C.G. (2013) Genomic characterization of childhood acute lymphoblastic leukemia. *Semin. Hematol.*, **50**, 314–24.
31. Liu,Y., Easton,J., Shao,Y., Maciaszek,J., Wang,Z., Wilkinson,M.R., McCastlain,K., Edmonson,M., Pounds,S.B., Shi,L., *et al.* (2017) The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat. Genet.*, **49**, 1211–1218.
32. Spinella,J.-F., Cassart,P., Richer,C., Saillour,V., Ouimet,M., Langlois,S., St-Onge,P., Sontag,T., Healy,J., Minden,M.D., *et al.* (2016) Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations. *Oncotarget*, **7**, 65485–65503.
33. De Keersmaecker,K., Atak,Z.K., Li,N., Vicente,C., Patchett,S., Girardi,T., Gianfelici,V., Geerdens,E., Clappier,E., Porcu,M., *et al.* (2013) Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat. Genet.*, **45**, 186–190.
34. Kalender Atak,Z., Gianfelici,V., Hulselmans,G., De Keersmaecker,K., Devasia,A.G., Geerdens,E., Mentens,N., Chiaretti,S., Durinck,K., Uyttebroeck,A., *et al.* (2013) Comprehensive Analysis of Transcriptome Variation Uncovers Known and Novel Driver Events in T-Cell Acute Lymphoblastic Leukemia. *PLoS Genet.*, **9**, e1003997.
35. Van Vlierberghe,P., Palomero,T., Khiabanian,H., Van Der Meulen,J., Castillo,M., Van Roy,N., De Moerloose,B., Philippé,J., González-García,S., Toribio,M.L., *et al.* (2010) PHF6 mutations in T-cell acute lymphoblastic leukemia. *Nat. Genet.*, **42**, 338–342.
36. Girardi,T., Vicente,C., Cools,J. and De Keersmaecker,K. (2017) The genetics and molecular biology of T-ALL. *Blood*, **129**, 1113–1123.
37. Van Vlierberghe,P., Pieters,R., Beverloo,H.B. and Meijerink,J.P.P. (2008) Molecular-genetic insights in paediatric T-cell acute lymphoblastic leukaemia. *Br. J. Haematol.*, **143**, 153–168.
38. Belver,L. and Ferrando,A. (2016) The genetics and mechanisms of T cell acute lymphoblastic leukaemia. *Nat. Rev. Cancer*, **16**, 494–507.
39. Hebert,J., Cayuela,J.M., Berkeley,J. and Sigaux,F. (1994) Candidate tumor-suppressor genes MTS1 (p16INK4A) and MTS2 (p15INK4B) display frequent homozygous deletions in primary cells from T- but not from B-cell lineage acute lymphoblastic leukemias. *Blood*, **84**, 4038–44.
40. De Keersmaecker,K., Marynen,P. and Cools,J. (2005) Genetic insights in the pathogenesis of T-cell acute lymphoblastic leukemia. *Haematologica*, **90**, 1116–27.
41. Bertin,R., Acquaviva,C., Mirebeau,D., Guidal-Giroux,C., Vilmer,E. and Cavé,H. (2003) CDKN2A, CDKN2B, and MTAP gene dosage permits precise characterization of mono- and bi-allelic 9p21 deletions in childhood acute lymphoblastic leukemia. *Genes Chromosom. Cancer*, **37**, 44–57.

42. Okamoto,A., Demetrick,D.J., Spillare,E.A., Hagiwara,K., Hussain,S.P., Bennett,W.P., Forrester,K., Gerwin,B., Serrano,M. and Beach,D.H. (2006) Mutations and altered expression of p16INK4 in human cancer. *Proc. Natl. Acad. Sci.*, **91**, 11045–11049.
43. Merlo,A., Herman,J.G., Mao,L., Lee,D.J., Gabrielson,E., Burger,P.C., Baylin,S.B. and Sidransky,D. (1995) 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat. Med.*, **1**, 686–92.
44. Okamoto,A., Demetrick,D.J., Spillare,E.A., Hagiwara,K., Hussain,S.P., Bennett,W.P., Forrester,K., Gerwin,B., Serrano,M., Beach,D.H., *et al.* (1994) Mutations and altered expression of p16INK4 in human cancer. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 11045.
45. Herman,J.G., Jen,J., Merlo,A. and Baylin,S.B. (1996) Hypermethylation-associated inactivation indicates a tumor suppressor role for p15INK4B. *Cancer Res.*, **56**, 722–7.
46. Weng,A.P., Ferrando,A.A., Lee,W., Morris,J.P., Silverman,L.B., Sanchez-Irizarry,C., Blacklow,S.C., Look,A.T. and Aster,J.C. (2004) Activating Mutations of NOTCH1 in Human T Cell Acute Lymphoblastic Leukemia. *Science (80-.).*, **306**, 269–271.
47. Palomero,T., McKenna,K., O-Neil,J., Galinsky,I., Stone,R., Suzukawa,K., Stiakaki,E., Kalmanti,M., Fox,E.A., Caligiuri,M.A., *et al.* (2006) Activating mutations in NOTCH1 in acute myeloid leukemia and lineage switch leukemias. *Leukemia*, **20**, 1963–1966.
48. Ellisen,L.W., Bird,J., West,D.C., Soreng,A.L., Reynolds,T.C., Smith,S.D. and Sklar,J. (1991) TAN-1, the human homolog of the Drosophila Notch gene, is broken by chromosomal translocations in T lymphoblastic neoplasms. *Cell*, **66**, 649–661.
49. Joshi,I., Minter,L.M., Telfer,J., Demarest,R.M., Capobianco,A.J., Aster,J.C., Sicinski,P., Fauq,A., Golde,T.E. and Osborne,B.A. (2009) Notch signaling mediates G 1/S cell-cycle progression in T cells via cyclin D3 and its dependent kinases. *Blood*, **113**, 1689–1698.
50. Dohda,T., Maljukova,A., Liu,L., Heyman,M., Grandér,D., Brodin,D., Sangfelt,O. and Lendahl,U. (2007) Notch signaling induces SKP2 expression and promotes reduction of p27Kip1 in T-cell acute lymphoblastic leukemia cell lines. *Exp. Cell Res.*, **313**, 3141–3152.
51. Palomero,T., Lim,W.K., Odom,D.T., Sulis,M.L., Real,P.J., Margolin,A., Barnes,K.C., O'Neil,J., Neuberg,D., Weng,A.P., *et al.* (2006) NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. *Proc. Natl. Acad. Sci.*, **103**, 18261–18266.
52. Medyouf,H., Gusscott,S., Wang,H., Tseng,J.-C., Wai,C., Nemirovsky,O., Trumpp,A., Pflumio,F., Carboni,J., Gottardis,M., *et al.* (2011) High-level IGF1R expression is required for leukemia-initiating cell activity in T-ALL and is supported by Notch signaling. *J. Exp. Med.*, **208**, 1809–1822.
53. González-García,S., García-Peydró,M., Martín-Gayo,E., Ballestar,E., Esteller,M., Bornstein,R., de la Pompa,J.L., Ferrando,A.A. and Toribio,M.L. (2009) CSL–MAML-dependent Notch1 signaling controls T lineage–specific IL-7Rα gene expression in early human thymopoiesis and leukemia. *J. Exp. Med.*, **206**, 1633–1633.
54. Vilimas,T., Mascarenhas,J., Palomero,T., Mandal,M., Buonamici,S., Meng,F., Thompson,B., Spaulding,C., Macaroun,S., Alegre,M.L., *et al.* (2007) Targeting the NF-κB signaling pathway in Notch1-induced T-cell leukemia. *Nat. Med.*, **13**, 70–77.
55. McCarter,A.C., Wang,Q. and Chiang,M. (2018) Notch in Leukemia. In. Springer, Cham, pp. 355–394.
56. Thompson,B.J., Buonamici,S., Sulis,M.L., Palomero,T., Vilimas,T., Basso,G., Ferrando,A. and Aifantis,I. (2007) The SCFFBW7 ubiquitin ligase complex as a tumor suppressor in T cell leukemia. *J. Exp. Med.*, **204**, 1825–35.
57. Milano,J., McKay,J., Dagenais,C., Foster-Brown,L., Pognan,F., Gadient,R., Jacobs,R.T., Zacco,A., Greenberg,B. and Ciaccio,P.J. (2004) Modulation of Notch Processing by γ-Secretase Inhibitors Causes Intestinal Goblet Cell Metaplasia and Induction of Genes Known to Specify Gut Secretory Lineage Differentiation. *Toxicol. Sci.*, **82**, 341–358.
58. Breit,S., Stanulla,M., Flohr,T., Schrappe,M., Ludwig,W.-D., Tolle,G., Happich,M., Muckenthaler,M.U. and Kulozik,A.E. (2006) Activating NOTCH1 mutations predict favorable early treatment response and long-term outcome in childhood precursor T-cell lymphoblastic

leukemia. *Blood*, **108**, 1151–7.
59. Van Vlierberghe,P., Ambesi-Impiombato,A., De Keersmaecker,K., Hadler,M., Paietta,E., Tallman,M.S., Rowe,J.M., Forne,C., Rue,M. and Ferrando,A.A. (2013) Prognostic relevance of integrated genetic profiling in adult T-cell acute lymphoblastic leukemia. *Blood*, **122**, 74–82.
60. Habets,R.A., de Bock,C.E., Serneels,L., Lodewijckx,I., Verbeke,D., Nittner,D., Narlawar,R., Demeyer,S., Dooley,J., Liston,A., *et al.* (2019) Safe targeting of T cell acute lymphoblastic leukemia by pathology-specific NOTCH inhibition. *Sci. Transl. Med.*, **11**, eaau6246.
61. Wendorff,A.A., Quinn,S.A., Rashkovan,M., Madubata,C.J., Ambesi-Impiombato,A., Litzow,M.R., Tallman,M.S., Paietta,E., Paganin,M., Basso,G., *et al.* (2019) Phf6 loss enhances HSC self-renewal driving tumor initiation and leukemia stem cell activity in T-All. *Cancer Discov.*, **9**, 436–451.
62. McRae,H.M., Garnham,A.L., Hu,Y., Witkowski,M.T., Corbett,M.A., Dixon,M.P., May,R.E., Sheikh,B.N., Chiang,W., Kueh,A.J., *et al.* (2019) PHF6 regulates hematopoietic stem and progenitor cells and its loss synergizes with expression of TLX3 to cause leukemia. *Blood*, **133**, 1729–1741.
63. Ntziachristos,P. (2019) PHF6: it is written in the stem cells. *Blood*, **133**, 2461–2462.
64. Trowbridge,J.J. (2019) Context-specific tumor suppression by PHF6. *Blood*, **133**, 1698–1700.
65. Tycko,B. (2004) Chromosomal translocations joining LCK and TCRB loci in human T cell leukemia. *J. Exp. Med.*, **174**, 867–873.
66. Yokota,S., Nakao,M., Horiike,S., Seriu,T., Iwai,T., Kaneko,H., Azuma,H., Oka,T., Takeda,T., Watanabe,A., *et al.* (1998) Mutational analysis of the N-ras gene in acute lymphoblastic leukemia: a study of 125 Japanese pediatric cases. *Int. J. Hematol.*, **67**, 379–87.
67. Palomero,T., Sulis,M.L., Cortina,M., Real,P.J., Barnes,K., Ciofani,M., Caparros,E., Buteau,J., Brown,K., Perkins,S.L., *et al.* (2007) Mutational loss of PTEN induces resistance to NOTCH1 inhibition in T-cell leukemia. *Nat. Med.*, **13**, 1203–10.
68. De Klein,A., Hagemeijer,A., Bartram,C.R., Houwen,R., Hoefsloot,L., Carbonell,F., Chan,L., Barnett,M., Greaves,M. and Kleihauer,E. (1986) bcr rearrangement and translocation of the c-abl oncogene in Philadelphia positive acute lymphoblastic leukemia. *Blood*, **68**, 1369–75.
69. Klein,A. de, Kessel,A.G. van, Grosveld,G., Bartram,C.R., Hagemeijer,A., Bootsma,D., Spurr,N.K., Heisterkamp,N., Groffen,J. and Stephenson,J.R. (1982) A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature*, **300**, 765–767.
70. Graux,C., Cools,J., Melotte,C., Quentmeier,H., Ferrando,A., Levine,R., Vermeesch,J.R., Stul,M., Dutta,B., Boeckx,N., *et al.* (2004) Fusion of NUP214 to ABL1 on amplified episomes in T-cell acute lymphoblastic leukemia. *Nat. Genet.*, **36**, 1084–1089.
71. Lacronique,V., Boureux,A., Valle,V.D., Poirel,H., Quang,C.T., Mauchauffé,M., Berthou,C., Lessard,M., Berger,R., Ghysdael,J., *et al.* (1997) A TEL-JAK2 fusion protein with constitutive kinase activity in human leukemia. *Science*, **278**, 1309–12.
72. Paietta,E., Ferrando,A.A., Neuberg,D., Bennett,J.M., Racevskis,J., Lazarus,H., Dewald,G., Rowe,J.M., Wiernik,P.H., Tallman,M.S., *et al.* (2004) Activating FLT3 mutations in CD117/KIT(+) T-cell acute lymphoblastic leukemias. *Blood*, **104**, 558–60.
73. Ferrando,A.A., Neuberg,D.S., Dodge,R.K., Paietta,E., Larson,R.A., Wiernik,P.H., Rowe,J.M., Caligiuri,M.A., Bloomfield,C.D. and Look,A.T. (2004) Prognostic importance of TLX1 (HOX11) oncogene expression in adults with T-cell acute lymphoblastic leukaemia. *Lancet*, **363**, 535–536.
74. Della Gatta,G., Palomero,T., Perez-Garcia,A., Ambesi-Impiombato,A., Bansal,M., Carpenter,Z.W., De Keersmaecker,K., Sole,X., Xu,L., Paietta,E., *et al.* (2012) Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. *Nat. Med.*, **18**, 436–40.
75. Soulier,J., Clappier,E., Cayuela,J.M., Regnault,A., García-Peydró,M., Dombret,H., Baruchel,A., Toribio,M.L. and Sigaux,F. (2005) HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood*, **106**, 274–286.
76. Baer,R. (1993) TAL1, TAL2 and LYL1: a family of basic helix-loop-helix proteins implicated in T cell acute leukaemia. *Semin. Cancer Biol.*, **4**, 341–7.
77. Mansour,M.R., Abraham,B.J., Anders,L., Berezovskaya,A., Gutierrez,A., Durbin,A.D., Etchin,J., Lawton,L., Sallan,S.E., Silverman,L.B., *et al.* (2014) Oncogene regulation. An oncogenic super-

enhancer formed through somatic mutation of a noncoding intergenic element. *Science*, **346**, 1373–7.
78. Sanda,T., Lawton,L.N., Barrasa,M.I., Fan,Z.P., Kohlhammer,H., Gutierrez,A., Ma,W., Tatarek,J., Ahn,Y., Kelliher,M.A., *et al.* (2012) Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell*, **22**, 209–21.
79. Yan,W., Young,A.Z., Soares,V.C., Kelley,R., Benezra,R. and Zhuang,Y. (2015) High incidence of T-cell tumors in E2A-null mice and E2A/Id1 double-knockout mice. *Mol. Cell. Biol.*, **17**, 7317–7327.
80. Royer-Pokora,B., Loos,U. and Ludwig,W.D. (1991) TTG-2, a new gene encoding a cysteine-rich protein with the LIM motif, is overexpressed in acute T-cell leukaemia with the t(11;14)(p13;q11). *Oncogene*, **6**, 1887–93.
81. Sánchez-García,I. and Rabbitts,T.H. (1993) LIM domain proteins in leukaemia and development. *Semin. Cancer Biol.*, **4**, 349–58.
82. Van Vlierberghe,P., van Grotel,M., Beverloo,H.B., Lee,C., Helgason,T., Buijs-Gladdines,J., Passier,M., van Wering,E.R., Veerman,A.J.P., Kamps,W.A., *et al.* (2006) The cryptic chromosomal deletion del(11)(p12p13) as a new activation mechanism of LMO2 in pediatric T-cell acute lymphoblastic leukemia. *Blood*, **108**, 3520–9.
83. Bernard,O., Busson-LeConiat,M., Ballerini,P., Mauchauffé,M., Della Valle,V., Monni,R., Khac,F.N., Mercher,T., Penard-Lacronique,V., Pasturaud,P., *et al.* (2001) A new recurrent and specific cryptic translocation, t(5;14)(q35;q32), is associated with expression of the Hox11L2 gene in T acute lymphoblastic leukemia. *Leukemia*, **15**, 1495–1504.
84. Soulier,J., Clappier,E., Cayuela,J.-M., Regnault,A., García-Peydró,M., Dombret,H., Baruchel,A., Toribio,M.-L. and Sigaux,F. (2005) HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood*, **106**, 274–86.
85. Speleman,F., Cauwelier,B., Dastugue,N., Cools,J., Verhasselt,B., Poppe,B., Van Roy,N., Vandesompele,J., Graux,C., Uyttebroeck,A., *et al.* (2005) A new recurrent inversion, inv(7)(p15q34), leads to transcriptional activation of HOXA10 and HOXA11 in a subset of T-cell acute lymphoblastic leukemias. *Leukemia*, **19**, 358–366.
86. Carlson,K., Vignon,C., Bohlander,S., Martinez-Climent,J., Beau,M.M. Le and Rowley,J. (2000) Identification and molecular characterization of CALM/AF10fusion products in T cell acute lymphoblastic leukemia and acute myeloid leukemia. *Leukemia*, **14**, 100–104.
87. Okada,Y., Jiang,Q., Lemieux,M., Jeannotte,L., Su,L. and Zhang,Y. (2006) Leukaemic transformation by CALM-AF10 involves upregulation of Hoxa5 by hDOT1L. *Nat. Cell Biol.*, **8**, 1017–1024.
88. Okada,Y., Feng,Q., Lin,Y., Jiang,Q., Li,Y., Coffield,V.M., Su,L., Xu,G. and Zhang,Y. (2005) hDOT1L links histone methylation to leukemogenesis. *Cell*, **121**, 167–78.
89. Mueller,D., Bach,C., Zeisig,D., Garcia-Cuellar,M.-P., Monroe,S., Sreekumar,A., Zhou,R., Nesvizhskii,A., Chinnaiyan,A., Hess,J.L., *et al.* (2007) A role for the MLL fusion partner ENL in transcriptional elongation and chromatin modification. *Blood*, **110**, 4445–54.
90. Soulier,J., Clappier,E., Cayuela,J.-M., Regnault,A., García-Peydró,M., Dombret,H., Baruchel,A., Toribio,M.-L. and Sigaux,F. (2005) HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood*, **106**, 274–86.
91. Herblot,S., Steff,A.-M., Hugo,P., Aplan,P.D. and Hoang,T. (2000) SCL and LMO1 alter thymocyte differentiation: inhibition of E2A-HEB function and pre-Tα chain expression. *Nat. Immunol.*, **1**, 138–144.
92. Reed,J.C. (1995) Regulation of apoptosis by bcl-2 family proteins and its role in cancer and chemoresistance. *Curr. Opin. Oncol.*, **7**, 541–6.
93. Patrick,K., Wade,R., Goulden,N., Mitchell,C., Moorman,A. V., Rowntree,C., Jenkinson,S., Hough,R. and Vora,A. (2014) Outcome for children and young people with Early T-cell precursor acute lymphoblastic leukaemia treated on a contemporary protocol, UKALL 2003. *Br. J. Haematol.*, **166**, 421–424.
94. Dear,T.N., Colledge,W.H., Carlton,M.B., Lavenir,I., Larson,T., Smith,A.J., Warren,A.J., Evans,M.J., Sofroniew,M. V and Rabbitts,T.H. (1995) The Hox11 gene is essential for cell survival during spleen development. *Development*, **121**, 2909–2915.

95. Kees,U.R., Heerema,N.A., Kumar,R., Watt,P.M., Baker,D.L., La,M.K., Uckun,F.M. and Sather,H.N. (2003) Expression of HOX11 in childhood T-lineage acute lymphoblastic leukaemia can occur in the absence of cytogenetic aberration at 10q24: a study from the Children's Cancer Group (CCG). *Leukemia*, **17**, 887–893.
96. Guo,Z., Zhao,C., Huang,M., Huang,T., Fan,M., Xie,Z., Chen,Y., Zhao,X., Xia,G., Geng,J., *et al.* (2012) Tlx1/3 and Ptf1a Control the Expression of Distinct Sets of Transmitter and Peptide Receptor Genes in the Developing Dorsal Spinal Cord. *J. Neurosci.*, **32**, 8509–8520.
97. Owens,B.M., Hawley,T.S., Spain,L.M., Kerkel,K.A. and Hawley,R.G. (2006) TLX1/HOX11-mediated disruption of primary thymocyte differentiation prior to the CD4+CD8+ double-positive stage. *Br. J. Haematol.*, **132**, 216–29.
98. Durinck,K., Van Loocke,W., Van der Meulen,J., Van de Walle,I., Ongenaert,M., Rondou,P., Wallaert,A., de Bock,C.E., Van Roy,N., Poppe,B., *et al.* (2015) Characterization of the genome-wide TLX1 binding profile in T-cell acute lymphoblastic leukemia. *Leukemia*, **29**, 2317–2327.
99. Bergeron,J., Clappier,E., Radford,I., Buzyn,A., Millien,C., Soler,G., Ballerini,P., Thomas,X., Soulier,J., Dombret,H., *et al.* (2007) Prognostic and oncogenic relevance of TLX1/HOX11 expression level in T-ALLs. *Blood*, **110**, 2324–2330.
100. Berger,R., Dastugue,N., Busson,M., van den Akker,J., Pérot,C., Ballerini,P., Hagemeijer,A., Michaux,L., Charrin,C., Pages,M.P., *et al.* (2003) t(5;14)/HOX11L2-positive T-cell acute lymphoblastic leukemia. A collaborative study of the Groupe Français de Cytogénétique Hématologique (GFCH). *Leukemia*, **17**, 1851–1857.
101. Watt,P.M., Kumar,R. and Kees,U.R. (2000) Promoter demethylation accompanies reactivation of the HOX11 proto-oncogene in leukemia. *Genes, Chromosom. Cancer*, **29**, 371–377.
102. Riz,I. and Hawley,R.G. (2005) G1/S transcriptional networks modulated by the HOX11/TLX1 oncogene of T-cell acute lymphoblastic leukemia. *Oncogene*, **24**, 5561–75.
103. Dadi,S., Le Noir,S., Payet-Bornet,D., Lhermitte,L., Zacarias-Cabeza,J., Bergeron,J., Villar??se,P., Vachez,E., Dik,W.A., Millien,C., *et al.* (2012) TLX Homeodomain Oncogenes Mediate T Cell Maturation Arrest in T-ALL via Interaction with ETS1 and Suppression of TCR?? Gene Expression. *Cancer Cell*, **21**, 563–576.
104. King,B., Ntziachristos,P. and Aifantis,I. (2012) Hijacking T cell differentiation: new insights in TLX function in T-ALL. *Cancer Cell*, **21**, 453–5.
105. De Keersmaecker,K., Real,P.J., Gatta,G. Della, Palomero,T., Sulis,M.L., Tosello,V., Van Vlierberghe,P., Barnes,K., Castillo,M., Sole,X., *et al.* (2010) The TLX1 oncogene drives aneuploidy in T cell transformation. *Nat. Med.*, **16**, 1321–1327.
106. Rakowski,L.A., Lehotzky,E.A. and Chiang,M.Y. (2011) Transient responses to NOTCH and TLX1/HOX11 inhibition in T-cell acute lymphoblastic leukemia/lymphoma. *PLoS One*, **6**.
107. Kleppe,M., Soulier,J., Asnafi,V., Mentens,N., Hornakova,T., Knoops,L., Sigaux,F., Meijerink,J.P., Vandenberghe,P., Tartaglia,M., *et al.* (2013) lymphoblastic leukemia PTPN2 negatively regulates oncogenic JAK1 in T-cell acute lymphoblastic leukemia. **117**, 7090–7098.
108. De Keersmaecker,K. and Ferrando,A.A. (2011) TLX1-Induced T-cell Acute Lymphoblastic Leukemia. *Clin. Cancer Res.*, **17**, 6381–6386.
109. Tosello,V., Mansour,M.R., Barnes,K., Paganin,M., Sulis,M.L., Jenkinson,S., Allen,C.G., Gale,R.E., Linch,D.C., Palomero,T., *et al.* (2009) WT1 mutations in T-ALL. *Blood*, **114**, 1038–1045.
110. Zuurbier,L., Homminga,I., Calvert,V., Winkel,M. te, Buijs-Gladdines,J.G.C.A.M., Kooi,C., Smits,W.K., Sonneveld,E., Veerman,A.J.P., Kamps,W.A., *et al.* (2010) NOTCH1 and/or FBXW7 mutations predict for initial good prednisone response but not for improved outcome in pediatric T-cell acute lymphoblastic leukemia patients treated on DCOG or COALL protocols. *Leukemia*, **24**, 2014–2022.
111. Asnafi,V., Buzyn,A., Le Noir,S., Baleydier,F., Simon,A., Beldjord,K., Reman,O., Witz,F., Fagot,T., Tavernier,E., *et al.* (2009) NOTCH1/FBXW7 mutation identifies a large subgroup with favorable outcome in adult T-cell acute lymphoblastic leukemia (T-ALL): a Group for Research on Adult Acute Lymphoblastic Leukemia (GRAALL) study. *Blood*, **113**, 3918–24.
112. Wang,H., Zang,C., Taing,L., Arnett,K.L., Wong,Y.J., Pear,W.S., Blacklow,S.C., Liu,X.S. and Aster,J.C.

(2013) NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. *Proc. Natl. Acad. Sci.*, **111**, 705–710.
113. Wang,H., Zou,J., Zhao,B., Johannsen,E., Ashworth,T., Wong,H., Pear,W.S., Schug,J., Blacklow,S.C., Arnett,K.L., *et al.* (2011) Genome-wide analysis reveals conserved and divergent features of Notch1/RBPJ binding in human and murine T-lymphoblastic leukemia cells. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 14908–13.
114. Kleppe,M., Lahortiga,I., El Chaar,T., De Keersmaecker,K., Mentens,N., Graux,C., Van Roosbroeck,K., Ferrando,A.A., Langerak,A.W., Meijerink,J.P.P., *et al.* (2010) Deletion of the protein tyrosine phosphatase gene PTPN2 in T-cell acute lymphoblastic leukemia. *Nat. Genet.*, **42**, 530–535.
115. Vanden Bempt,M., Demeyer,S., Broux,M., De Bie,J., Bornschein,S., Mentens,N., Vandepoel,R., Geerdens,E., Radaelli,E., Bornhauser,B.C., *et al.* (2018) Cooperative Enhancer Activation by TLX1 and STAT5 Drives Development of NUP214-ABL1/TLX1-Positive T Cell Acute Lymphoblastic Leukemia. *Cancer Cell*, **34**, 271–285.e7.
116. Kawabe,T., Muslin,A.J. and Korsmeyer,S.J. (1997) HOX11 interacts with protein phosphatases PP2A and PP1 and disrupts a G2/M cell-cycle checkpoint. *Nature*, **385**, 454–458.
117. Su,X.Y., Busson,M., Della Valle,V., Ballerini,P., Dastugue,N., Talmant,P., Ferrando,A.A., Baudry-Bluteau,D., Romana,S., Berger,R., *et al.* (2004) Various types of rearrangements target TLX3 locus in T-cell acute lymphoblastic leukemia. *Genes Chromosom. Cancer*, **41**, 243–249.
118. Hansen-Hagge,T., Schäfer,M., Kiyoi,H., Morris,S., Whitlock,J., Koch,P., Bohlmann,I., Mahotka,C., Bartram,C. and Janssen,J. (2002) Disruption of the RanBP17/Hox11L2 region by recombination with the TCRδ locus in acute lymphoblastic leukemias with t(5;14)(q34;q11). *Leukemia*, **16**, 2205–2212.
119. Della Gatta,G., Palomero,T., Perez-Garcia,A., Ambesi-Impiombato,A., Bansal,M., Carpenter,Z.W., De Keersmaecker,K., Sole,X., Xu,L., Paietta,E., *et al.* (2012) Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. *Nat. Med.*, **18**, 436–440.
120. Van Grotel,M., Meijerink,J.P.P., Beverloo,H.B., Langerak,A.W., Buys-Gladdines,J.G.C.A.M., Schneider,P., Poulsen,T.S., den Boer,M.L., Horstmann,M., Kamps,W.A., *et al.* (2006) The outcome of molecular-cytogenetic subgroups in pediatric T-cell acute lymphoblastic leukemia: a retrospective study of patients treated according to DCOG or COALL protocols. *Haematologica*, **91**, 1212–21.
121. Van Grotel,M., Meijerink,J.P.P., van Wering,E.R., Langerak,A.W., Beverloo,H.B., Buijs-Gladdines,J.G.C.A.M., Burger,N.B., Passier,M., van Lieshout,E.M., Kamps,W.A., *et al.* (2008) Prognostic significance of molecular-cytogenetic abnormalities in pediatric T-ALL is not explained by immunophenotypic differences. *Leukemia*, **22**, 124–131.
122. Asnafi,V., Beldjord,K., Garand,R., Millien,C., Bahloul,M., LeTutour,P., Douay,L., Valensi,F. and Macintyre,E. (2004) IgH DJ rearrangements within T-ALL correlate with cCD79a expression, an immature/TCRγδ phenotype and absence of IL7Rα/CD127 expression. *Leukemia*, **18**, 1997–2001.
123. Locatelli,F., Schrappe,M., Bernardo,M.E. and Rutella,S. (2012) How i treat relapsed childhood acute lymphoblastic leukemia. *Blood*, **120**, 2807–2816.
124. Berg,S.L., Blaney,S.M., Devidas,M., Lampkin,T.A., Murgo,A., Bernstein,M., Billett,A., Kurtzberg,J., Reaman,G., Gaynon,P., *et al.* (2005) Phase II Study of Nelarabine (compound 506U78) in Children and Young Adults With Refractory T-Cell Malignancies: A Report From the Children's Oncology Group. *J. Clin. Oncol.*, **23**, 3376–3382.
125. Hefazi,M. and Litzow,M.R. (2018) Recent Advances in the Biology and Treatment of T Cell Acute Lymphoblastic Leukemia. *Curr. Hematol. Malig. Rep.*, **13**, 265–274.
126. Nicola,G. (2013) How I Treat How I treat older patients with ALL. *Blood*, **122**, 1366–1375.
127. Rowe,J.M., Goldstone,A.H. and Dc,W. (2012) How I treat acute lymphocytic leukemia in adults How I treat How I treat acute lymphocytic leukemia in adults. **126**, 2268–2275.
128. Oriol,A., Vives,S., Hernández-Rivas,J.M., Tormo,M., Heras,I., Rivas,C., Bethencourt,C., Moscardó,F., Bueno,J., Grande,C., *et al.* (2010) Outcome after relapse of acute lymphoblastic leukemia in adult patients included in four consecutive risk-adapted trials by the PETHEMA study

group. *Haematologica*, **95**, 589–596.
129. Quintás-Cardama,A., Tong,W., Manshouri,T., Vega,F., Lennon,P.A., Cools,J., Gilliland,D.G., Lee,F., Cortes,J., Kantarjian,H., *et al.* (2008) Activity of tyrosine kinase inhibitors against human NUP214-ABL1-positive T cell malignancies. *Leukemia*, **22**, 1117–1124.
130. Can one target T-cell ALL.pdf.
131. Brentjens,R.J., Davila,M.L., Riviere,I., Park,J., Wang,X., Cowell,L.G., Bartido,S., Stefanski,J., Taylor,C., Olszewska,M., *et al.* (2013) CD19-Targeted T Cells Rapidly Induce Molecular Remissions in Adults with Chemotherapy-Refractory Acute Lymphoblastic Leukemia. *Sci. Transl. Med.*, **5**, 177ra38-177ra38.
132. You,F., Wang,Y., Jiang,L., Zhu,X., Chen,D., Yuan,L., An,G., Meng,H. and Yang,L. (2019) A novel CD7 chimeric antigen receptor-modified NK-92MI cell line targeting T-cell acute lymphoblastic leukemia. *Am. J. Cancer Res.*, **9**, 64–78.
133. Gomes-Silva,D., Srinivasan,M., Sharma,S., Lee,C.M., Wagner,D.L., Davis,T.H., Rouce,R.H., Bao,G., Brenner,M.K. and Mamonkin,M. (2017) CD7-edited T cells expressing a CD7-specific CAR for the therapy of T-cell malignancies. *Blood*, **130**, 285–296.
134. Leucci,E., Vendramin,R., Spinazzi,M., Laurette,P., Fiers,M., Wouters,J., Radaelli,E., Eyckerman,S., Leonelli,C., Vanderheyden,K., *et al.* (2016) Melanoma addiction to the long non-coding RNA SAMMSON. *Nature*, **531**, 518–522.
135. Crick,F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–3.
136. Ling,H., Vincent,K., Pichler,M., Fodde,R., Berindan-Neagoe,I., Slack,F.J. and Calin,G.A. (2015) Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene*, **34**, 5003–5011.
137. Djebali,S., Davis,C.A., Merkel,A. and Gingeras,T.R. (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
138. ENCODE Project Consortium,T.E.P., Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigó,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
139. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G., *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775.
140. Braicu,C., Zimta,A., Harangus,A., Iurca,I., Irimie,A., Coza,O. and Berindan-neagoe,I. (2019) The Function of Non-Coding RNAs in Lung Cancer Tumorigenesis. **3**, 1–18.
141. Filipowicz,W., Bhattacharyya,S.N. and Sonenberg,N. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.*, **9**, 102–114.
142. Ha,M. and Kim,V.N. (2014) Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, **15**, 509–524.
143. Mavrakis,K.J., Wolfe,A.L., Oricchio,E., Palomero,T., Keersmaecker,D., Mcjunkin,K., Zuber,J., James,T., Khan,A.A., Leslie,C.S., *et al.* (2010) Genome-wide RNAi screen identifies miR-19 targets in Notch-induced acute T-cell leukaemia (T-ALL). *Cell*, **12**, 372–379.
144. Goodarzi,H., Nguyen,H.C.B., Zhang,S., Dill,B.D., Molina,H. and Tavazoie,S.F. (2016) Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression. *Cell*, **165**, 1416–1427.
145. Bohnsack,M.T. and Sloan,K.E. (2018) Modifications in small nuclear RNAs and their roles in spliceosome assembly and function. *Biol. Chem.*, **399**, 1265–1276.
146. Dupuis-Sandoval,F., Poirier,M. and Scott,M.S. (2015) The emerging landscape of small nucleolar RNAs in cell biology. *Wiley Interdiscip. Rev. RNA*, **6**, 381–97.
147. Zheng,D., Zhang,J., Ni,J., Luo,J., Wang,J., Tang,L., Zhang,L., Wang,L., Xu,J., Su,B., *et al.* (2015) Small nucleolar RNA 78 promotes the tumorigenesis in non-small cell lung cancer. *J. Exp. Clin. Cancer Res.*, **34**, 49.
148. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.

149. Brannan,C.I., Dees,E.C., Ingram,R.S. and Tilghman,S.M. (1990) The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.*, **10**, 28–36.
150. Slaby,O., Laga,R. and Sedlacek,O. (2017) Therapeutic targeting of non-coding RNAs in cancer. *Biochem. J.*, **474**, 4219–4251.
151. Johnsson,P., Lipovich,L., Grandér,D. and Morris,K. V (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta*, **1840**, 1063–71.
152. Wu,Y., Yang,L., Zhao,J., Li,C., Nie,J., Liu,F., Zhuo,C., Zheng,Y., Li,B., Wang,Z., *et al.* (2015) Nuclear-enriched abundant transcript 1 as a diagnostic and prognostic biomarker in colorectal cancer. *Mol. Cancer*, **14**.
153. Prabhakar,B., Zhong,X.-B. and Rasmussen,T.P. (2017) Exploiting Long Noncoding RNAs as Pharmacological Targets to Modulate Epigenetic Diseases. *Yale J. Biol. Med.*, **90**, 73–86.
154. Clark,M.B., Johnston,R.L., Inostroza-Ponta,M., Fox,A.H., Fortini,E., Moscato,P., Dinger,M.E. and Mattick,J.S. (2012) Genome-wide analysis of long noncoding RNA stability. *Genome Res.*, **22**, 885.
155. Wallaert,A., Durinck,K., Van Loocke,W., Van de Walle,I., Matthijssens,F., Volders,P.J., Avila Cobos,F., Rombaut,D., Rondou,P., Mestdagh,P., *et al.* (2016) Long noncoding RNA signatures define oncogenic subtypes in T-cell acute lymphoblastic leukemia. *Leukemia*, **30**, 1927–1930.
156. Durinck,K., Wallaert,A., Van de Walle,I., Van Loocke,W., Volders,P.J., Vanhauwaert,S., Geerdens,E., Benoit,Y., Van Roy,N., Poppe,B., *et al.* (2014) The notch driven long non-coding RNA repertoire in T-cell acute lymphoblastic leukemia. *Haematologica*, **99**, 1808–1816.
157. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
158. Lee,C. and Kikyo,N. (2012) Strategies to identify long noncoding RNAs involved in gene regulation. *Cell Biosci.*, **2**, 37.
159. Yang,L., Duff,M.O., Graveley,B.R., Carmichael,G.G. and Chen,L.-L. Genomewide characterization of non-polyadenylated RNAs. 10.1186/gb-2011-12-2-r16.
160. Chen,X., Kong,J., Ma,Z., Gao,S. and Feng,X. (2015) Up regulation of the long non-coding RNA NEAT1 promotes esophageal squamous cell carcinoma cell progression and correlates with poor prognosis. *Am. J. Cancer Res.*, **5**, 2808.
161. Volders,P.J., Anckaert,J., Verheggen,K., Nuytens,J., Martens,L., Mestdagh,P. and Vandesompele,J. (2019) Lncipedia 5: Towards a reference set of human long non-coding rnas. *Nucleic Acids Res.*, **47**, D135–D139.
162. Rinn,J.L. and Chang,H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–66.
163. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W., *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–12.
164. Kim,T.-K., Hemberg,M., Gray,J.M., Costa,A.M., Bear,D.M., Wu,J., Harmin,D.A., Laptewicz,M., Barbara-Haley,K., Kuersten,S., *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–7.
165. Kouno,T., Moody,J., Kwon,A.T.J., Shibayama,Y., Kato,S., Huang,Y., Böttcher,M., Motakis,E., Mendez,M., Severin,J., *et al.* (2019) C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nat. Commun.*, **10**.
166. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T., *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–61.
167. Wang,J., Ren,Q., Hua,L., Chen,J., Zhang,J., Bai,H., Li,H., Xu,B., Shi,Z., Cao,H., *et al.* (2019) Comprehensive Analysis of Differentially Expressed mRNA, lncRNA and circRNA and Their ceRNA Networks in the Longissimus Dorsi Muscle of Two Different Pig Breeds. *Int. J. Mol. Sci.*, **20**, 1107.
168. Schaukowitch,K., Joo,J.-Y., Liu,X., Watts,J.K., Martinez,C. and Kim,T.-K. (2014) Enhancer RNA facilitates NELF release from immediate early genes. *Mol. Cell*, **56**, 29–42.
169. Mousavi,K., Zare,H., Dell'orso,S., Grontved,L., Gutierrez-Cruz,G., Derfoul,A., Hager,G.L. and Sartorelli,V. (2013) eRNAs promote transcription by establishing chromatin accessibility at

defined genomic loci. *Mol. Cell*, **51**, 606–17.
170. Li,W., Notani,D., Ma,Q., Tanasa,B., Nunez,E., Chen,A.Y., Merkurjev,D., Zhang,J., Ohgi,K., Song,X., *et al.* (2013) Functional Importance of eRNAs for Estrogen-dependent Transcriptional Activation Events. *Nature*, **498**, 516.
171. De Lara,J.C.-F., Arzate-Mejía,R.G. and Recillas-Targa,F. (2019) Enhancer RNAs: Insights Into Their Biological Role. *Epigenetics insights*, **12**, 2516865719846093.
172. Tan,S.H., Leong,W.Z., Ngoc,P.C.T., Tan,T.K., Bertulfo,F.C., Lim,M.C., An,O., Li,Z., Yeoh,A.E.J., Fullwood,M.J., *et al.* (2019) The enhancer RNA ARIEL activates the oncogenic transcriptional program in T-cell acute lymphoblastic leukemia. *Blood*, 10.1182/blood.2018874503.
173. Cocquerelle,C., Mascrez,B., Hétuin,D. and Bailleul,B. (1993) Mis-splicing yields circular RNA molecules. *FASEB J.*, **7**, 155–60.
174. Huang,S., Yang,B., Chen,B.J., Bliim,N., Ueberham,U., Arendt,T. and Janitz,M. (2017) The emerging role of circular RNAs in transcriptome regulation. *Genomics*, **109**, 401–407.
175. Yang,Z., Xie,L., Han,L., Qu,X., Yang,Y., Zhang,Y., He,Z., Wang,Y. and Li,J. (2017) Circular RNAs: Regulators of Cancer-Related Signaling Pathways and Potential Diagnostic Biomarkers for Human Cancers. *Theranostics*, **7**, 3106–3117.
176. Hansen,T.B., Jensen,T.I., Clausen,B.H., Bramsen,J.B., Finsen,B., Damgaard,C.K. and Kjems,J. (2013) Natural RNA circles function as efficient microRNA sponges. *Nature*, **495**, 384–388.
177. Barrett,S.P. and Salzman,J. (2016) Circular RNAs: analysis, expression and potential functions. *Development*, **143**, 1838–47.
178. Braicu,C., Zimta,A.-A., Gulei,D., Olariu,A. and Berindan-Neagoe,I. (2019) Comprehensive analysis of circular RNAs in pathological states: biogenesis, cellular regulation, and therapeutic relevance. *Cell. Mol. Life Sci.*, **76**, 1559–1577.
179. Qin,S., Zhao,Y., Lim,G., Lin,H., Zhang,X. and Zhang,X. (2019) Circular RNA PVT1 acts as a competing endogenous RNA for miR-497 in promoting non-small cell lung cancer progression. *Biomed. Pharmacother.*, **111**, 244–250.
180. Luger,K., Richmond,R.K., Sargent,D.F., Richmond,T.J. and A,A.W. (1997) Crystal structure of the nucleosome core particle at 2 . 8 A resolution. *Nature*, **389**, 251–260.
181. Gates,L.A., Foulds,C.E. and O'Malley,B.W. (2017) Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle. *Trends Biochem. Sci.*, **42**, 977–989.
182. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
183. Khalil,A.M., Guttman,M., Huarte,M., Garber,M., Raj,A., Rivea Morales,D., Thomas,K., Presser,A., Bernstein,B.E., van Oudenaarden,A., *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci.*, **106**, 11667–11672.
184. Gupta,R.A., Shah,N., Wang,K.C., Kim,J., Horlings,H.M., Wong,D.J., Tsai,M.-C., Hung,T., Argani,P., Rinn,J.L., *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.
185. Rinn,J.L., Kertesz,M., Wang,J.K., Squazzo,S.L., Xu,X., Brugmann,S.A., Goodnough,L.H., Helms,J.A., Farnham,P.J., Segal,E., *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, **129**, 1311–23.
186. Endo,H., Shiroki,T., Nakagawa,T., Yokoyama,M., Tamai,K., Yamanami,H., Fujiya,T., Sato,I., Yamaguchi,K., Tanaka,N., *et al.* (2013) Enhanced Expression of Long Non-Coding RNA HOTAIR Is Associated with the Development of Gastric Cancer. *PLoS One*, **8**, e77070.
187. Kogo,R., Shimamura,T., Mimori,K., Kawahara,K., Imoto,S., Sudo,T., Tanaka,F., Shibata,K., Suzuki,A., Komune,S., *et al.* (2011) Long Noncoding RNA HOTAIR Regulates Polycomb-Dependent Chromatin Modification and Is Associated with Poor Prognosis in Colorectal Cancers. *Cancer Res.*, **71**, 6320–6326.
188. Badalà,F., Nouri-mahdavi,K. and Raoof,D.A. (2008) Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15INK4B tumor suppressor geLeung A, Trac C, Jin W,

et al. Novel Long Non-Coding RNAs Are Regulated by Angiotensin II in Vascular Smooth Muscle Cells[J]. Circulation Resear. *Computer (Long. Beach. Calif).*, **144**, 724–732.
189. Marchese,F.P., Raimondi,I. and Huarte,M. (2017) The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.*, **18**, 206.
190. Grote,P., Wittler,L., Hendrix,D., Koch,F., Währisch,S., Beisaw,A., Macura,K., Bläss,G., Kellis,M., Werber,M., *et al.* (2013) The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev. Cell*, **24**, 206–14.
191. Yang,Y.W., Flynn,R.A., Chen,Y., Qu,K., Wan,B., Wang,K.C., Lei,M. and Chang,H.Y. (2014) Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. *Elife*, **3**, e02046.
192. Wang,K.C., Yang,Y.W., Liu,B., Sanyal,A., Corces-Zimmerman,R., Chen,Y., Lajoie,B.R., Protacio,A., Flynn,R.A., Gupta,R.A., *et al.* (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, **472**, 120–124.
193. Hung,T., Wang,Y., Lin,M.F., Koegel,A.K., Kotake,Y., Grant,G.D., Horlings,H.M., Shah,N., Umbricht,C., Wang,P., *et al.* (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat. Genet.*, **43**, 621–9.
194. Kino,T., Hurt,D.E., Ichijo,T., Nader,N. and Chrousos,G.P. (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal.*, **3**, ra8.
195. Yu,X., Zheng,H., Chan,M.T.V. and Wu,W.K.K. (2017) HULC: an oncogenic long non-coding RNA in human cancer. *J. Cell. Mol. Med.*, **21**, 410–417.
196. Tripathi,V., Ellis,J.D., Shen,Z., Song,D.Y., Pan,Q., Watt,A.T., Freier,S.M., Bennett,C.F., Sharma,A., Bubulya,P.A., *et al.* (2010) The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. *Mol. Cell*, **39**, 925–938.
197. Yoon,J.-H., Abdelmohsen,K., Srikantan,S., Yang,X., Martindale,J.L., De,S., Huarte,M., Zhan,M., Becker,K.G. and Gorospe,M. (2012) LincRNA-p21 suppresses target mRNA translation. *Mol. Cell*, **47**, 648–55.
198. Liu,B., Sun,L., Liu,Q., Gong,C., Yao,Y., Lv,X., Lin,L., Yao,H., Su,F., Li,D., *et al.* (2015) A Cytoplasmic NF-κB Interacting Long Noncoding RNA Blocks IκB Phosphorylation and Suppresses Breast Cancer Metastasis. *Cancer Cell*, **27**, 370–381.
199. Paralkar,V.R., Taborda,C.C., Huang,P., Yao,Y., Kossenkov,A.V., Prasad,R., Luan,J., Davies,J.O.J., Hughes,J.R., Hardison,R.C., *et al.* (2016) Unlinking an lncRNA from Its Associated cis Element. *Mol. Cell*, **62**, 104–110.
200. Latos,P.A., Pauler,F.M., Koerner,M. V., Senergin,H.B., Hudson,Q.J., Stocsits,R.R., Allhoff,W., Stricker,S.H., Klement,R.M., Warczok,K.E., *et al.* (2012) Airn Transcriptional Overlap, But Not Its lncRNA Products, Induces Imprinted Igf2r Silencing. *Science (80-.).*, **338**, 1469–1472.
201. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–50.
202. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P., *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
203. Huarte,M., Guttman,M., Feldser,D., Garber,M., Koziol,M.J., Kenzelmann-Broz,D., Khalil,A.M., Zuk,O., Amit,I., Rabani,M., *et al.* (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, **142**, 409–19.
204. Hu,X., Sood,A.K., Dang,C. V and Zhang,L. (2018) The role of long noncoding RNAs in cancer: the dark matter matters. *Curr. Opin. Genet. Dev.*, **48**, 8–15.
205. Yan,X., Hu,Z., Feng,Y., Hu,X., Yuan,J., Zhao,S.D., Zhang,Y., Yang,L., Shan,W., He,Q., *et al.* (2015) Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell*, **28**, 529–540.
206. Zheng,J., Huang,X., Tan,W., Yu,D., Du,Z., Chang,J., Wei,L., Han,Y., Wang,C., Che,X., *et al.* (2016) Pancreatic cancer risk variant in LINC00673 creates a miR-1231 binding site and interferes with

PTPN11 degradation. *Nat. Genet.*, **48**, 747–757.
207. Pandey,G.K., Mitra,S., Subhash,S., Hertwig,F., Kanduri,M., Mishra,K., Fransson,S., Ganeshram,A., Mondal,T., Bandaru,S., *et al.* (2014) The risk-associated long noncoding RNA NBAT-1 controls neuroblastoma progression by regulating cell proliferation and neuronal differentiation. *Cancer Cell*, **26**, 722–37.
208. Guo,H., Ahmed,M., Zhang,F., Yao,C.Q., Li,S., Liang,Y., Hua,J., Soares,F., Sun,Y., Langstein,J., *et al.* (2016) Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer. *Nat. Genet.*, **48**, 1142–1150.
209. Gutschner,T. and Diederichs,S. (2012) The hallmarks of cancer: A long non-coding RNA point of view. *RNA Biol.*, **9**, 703.
210. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
211. Prensner,J.R., Iyer,M.K., Balbin,O.A., Dhanasekaran,S.M., Cao,Q., Brenner,J.C., Laxman,B., Asangani,I.A., Grasso,C.S., Kominsky,H.D., *et al.* (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.*, **29**, 742–749.
212. Prensner,J.R., Chen,W., Han,S., Iyer,M.K., Cao,Q., Kothari,V., Evans,J.R., Knudsen,K.E., Paulsen,M.T., Ljungman,M., *et al.* (2014) The Long Non-Coding RNA PCAT-1 Promotes Prostate Cancer Cell Proliferation through cMyc. *Neoplasia*, **16**, 900–908.
213. Redon,S., Reichenbach,P. and Lingner,J. (2010) The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. *Nucleic Acids Res.*, **38**, 5797–5806.
214. Oliva-Rico,D. and Herrera,L.A. (2017) Regulated expression of the lncRNA TERRA and its impact on telomere biology. *Mech. Ageing Dev.*, **167**, 16–23.
215. Sampl,S., Pramhas,S., Stern,C., Preusser,M., Marosi,C. and Holzmann,K. (2012) Expression of telomeres in astrocytoma WHO grade 2 to 4: TERRA level correlates with telomere length, telomerase activity, and advanced clinical grade. *Transl. Oncol.*, **5**, 56–65.
216. Ji,P., Diederichs,S., Wang,W., Böing,S., Metzger,R., Schneider,P.M., Tidow,N., Brandt,B., Buerger,H., Bulk,E., *et al.* (2003) MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, **22**, 8031–8041.
217. Huo,Y., Li,Q., Wang,X., Jiao,X., Zheng,J., Li,Z. and Pan,X. (2017) MALAT1 predicts poor survival in osteosarcoma patients and promotes cell metastasis through associating with EZH2. *Oncotarget*, **8**, 46993–47006.
218. Jin,Y., Feng,S.-J., Qiu,S., Shao,N. and Zheng,J.-H. (2017) LncRNA MALAT1 promotes proliferation and metastasis in epithelial ovarian cancer via the PI3K-AKT pathway. *Eur. Rev. Med. Pharmacol. Sci.*, **21**, 3176–3184.
219. Tano,K., Mizuno,R., Okada,T., Rakwal,R., Shibato,J., Masuo,Y., Ijiri,K. and Akimitsu,N. (2010) MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. *FEBS Lett.*, **584**, 4575–4580.
220. Fan,Y., Shen,B., Tan,M., Mu,X., Qin,Y., Zhang,F. and Liu,Y. (2014) TGF- -Induced Upregulation of malat1 Promotes Bladder Cancer Metastasis by Associating with suz12. *Clin. Cancer Res.*, **20**, 1531–1541.
221. Uchida,T., Rossignol,F., Matthay,M.A., Mounier,R., Couette,S., Clottes,E. and Clerici,C. (2004) Prolonged hypoxia differentially regulates hypoxia-inducible factor (HIF)-1alpha and HIF-2alpha expression in lung epithelial cells: implication of natural antisense HIF-1alpha. *J. Biol. Chem.*, **279**, 14871–8.
222. Rossignol,F., Vaché,C. and Clottes,E. (2002) Natural antisense transcripts of hypoxia-inducible factor 1alpha are detected in different normal and tumour human tissues. *Gene*, **299**, 135–140.
223. Fu,X., Ravindranath,L., Tran,N., Petrovics,G. and Srivastava,S. (2006) Regulation of Apoptosis by a Prostate-Specific and Prostate Cancer-Associated Noncoding Gene, *PCGEM1*. *DNA Cell Biol.*, **25**, 135–141.
224. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of Cancer: The Next Generation. *Cell*, **144**, 646–674.
225. Yang,B., Zhang,L., Cao,Y., Chen,S., Cao,J., Wu,D., Chen,J., Xiong,H., Pan,Z., Qiu,F., *et al.* (2017)

Overexpression of lncRNA IGFBP4–1 reprograms energy metabolism to promote lung cancer progression. *Mol. Cancer*, **16**, 154.
226. Jiang,R., Tang,J., Chen,Y., Deng,L., Ji,J., Xie,Y., Wang,K., Jia,W., Chu,W.-M. and Sun,B. (2017) The long noncoding RNA lnc-EGFR stimulates T-regulatory cells differentiation thus promoting hepatocellular carcinoma immune evasion. *Nat. Commun.*, **8**, 15129.
227. Hu,W.L., Jin,L., Xu,A., Wang,Y.F., Thorne,R.F., Zhang,X.D. and Wu,M. (2018) GUARDIN is a p53-responsive long non-coding RNA that is essential for genomic stability. *Nat. Cell Biol.*, **20**, 492–502.
228. Trimarchi,T., Bilal,E., Ntziachristos,P., Fabbri,G., Dalla-Favera,R., Tsirigos,A. and Aifantis,I. (2014) Genome-wide mapping and characterization of a Notch-regulated long non-coding RNAs in acute leukemia. *Cell*, **158**, 593.
229. Wang,Y., Wu,P., Lin,R., Rong,L., Xue,Y. and Fang,Y. (2015) LncRNA NALT interaction with NOTCH1 promoted cell proliferation in pediatric T cell acute lymphoblastic leukemia. *Sci. Rep.*, **5**, 1–10.
230. Cao Thi Ngoc,P., Hao Tan,S., King Tan,T., Min Chan,M., Li,Z., J Yeoh,A.E., Tenen,D.G. and Sanda,T. Identification of novel lncRNAs regulated by the TAL1 complex in T- cell acute lymphoblastic leukemia. *Leukemia*, 10.1038/s41375-018-0110-4.
231. Leucci,E., Vendramin,R., Spinazzi,M., Laurette,P., Fiers,M., Wouters,J., Radaelli,E., Eyckerman,S., Leonelli,C., Vanderheyden,K., *et al.* (2016) Melanoma addiction to the long non-coding RNA SAMMSON. *Nature*, **531**, 518–522.
232. Arun,G., Diermeier,S.D. and Spector,D.L. (2018) Therapeutic Targeting of Long Non-Coding RNAs in Cancer. *Trends Mol. Med.*, **24**, 257–277.
233. Juliano,R.L. (2016) The delivery of therapeutic oligonucleotides. *Nucleic Acids Res.*, **44**, 6518–6548.
234. Bergmann,J.H. and Spector,D.L. (2014) Long non-coding RNAs: Modulators of nuclear structure and function. *Curr. Opin. Cell Biol.*, **0**, 10.
235. Burnett,J.C. and Rossi,J.J. (2012) RNA-based therapeutics: current progress and future prospects. *Chem. Biol.*, **19**, 60–71.
236. Shukla,S., Sumaria,C.S. and Pradeepkumar,P.I. (2010) Exploring Chemical Modifications for siRNA Therapeutics: A Structural and Functional Outlook. *ChemMedChem*, **5**, 328–349.
237. Koshkin,A.A., Singh,S.K., Nielsen,P., Rajwanshi,V.K., Kumar,R., Meldgaard,M., Olsen,C.E. and Wengel,J. (1998) LNA (Locked Nucleic Acids): Synthesis of the adenine, cytosine, guanine, 5-methylcytosine, thymine and uracil bicyclonucleoside monomers, oligomerisation, and unprecedented nucleic acid recognition. *Tetrahedron*, **54**, 3607–3630.
238. Eder,P.S., DeVine,R.J., Dagle,J.M. and Walder,J.A. (1991) Substrate specificity and kinetics of degradation of antisense oligonucleotides by a 3' exonuclease in plasma. *Antisense Res. Dev.*, **1**, 141–51.
239. Eckstein,F. (2014) Phosphorothioates, Essential Components of Therapeutic Oligonucleotides. *Nucleic Acid Ther.*, **24**, 374–387.
240. Phylactou,L.A., Tsipouras,P. and Kilpatrick,M.W. (1998) Hammerhead ribozymes targeted to the FBN1 mRNA can discriminate a single base mismatch between ribozyme and target. *Biochem. Biophys. Res. Commun.*, **249**, 804–810.
241. Han,J., Zhang,J., Chen,L., Shen,B., Zhou,J., Hu,B., Du,Y., Tate,P.H., Huang,X. and Zhang,W. (2014) Efficient in vivo deletion of a large imprinted lncRNA by CRISPR/Cas9. *RNA Biol.*, **11**, 829–35.
242. Groff,A.F., Sanchez-Gomez,D.B., Soruco,M.M.L., Gerhardinger,C., Barutcu,A.R., Li,E., Elcavage,L., Plana,O., Sanchez,L.V., Lee,J.C., *et al.* (2016) In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements. *Cell Rep.*, **16**, 2178–2186.
243. Thakore,P.I., D'Ippolito,A.M., Song,L., Safi,A., Shivakumar,N.K., Kabadi,A.M., Reddy,T.E., Crawford,G.E. and Gersbach,C.A. (2015) Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods*, **12**, 1143–9.
244. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell*, **152**, 1173–1183.

245. Abushahba,M.F.N., Mohammad,H., Thangamani,S., Hussein,A.A.A. and Seleem,M.N. (2016) Impact of different cell penetrating peptides on the efficacy of antisense therapeutics for targeting intracellular pathogens. *Sci. Rep.*, **6**, 20832.
246. Yang,B., Ming,X., Cao,C., Laing,B., Yuan,A., Porter,M.A., Hull-Ryde,E.A., Maddry,J., Suto,M., Janzen,W.P., *et al.* (2015) High-throughput screening identifies small molecules that enhance the pharmacological effects of oligonucleotides. *Nucleic Acids Res.*, **43**, 1987–1996.
247. Karikó,K., Bhuyan,P., Capodici,J. and Weissman,D. (2004) Small Interfering RNAs Mediate Sequence-Independent Gene Suppression and Induce Immune Activation by Signaling through Toll-Like Receptor 3. *J. Immunol.*, **172**, 6545–6549.
248. Heil,F., Hemmi,H., Hochrein,H., Ampenberger,F., Akira,S., Lipford,G., Wagner,H. and Bauer,S. (2004) Species-Specific Recognition of Single-Stranded RNA via Toll-like Receptor 7 and 8 Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/3836571 . *Science (80-.).*, **303**, 1526–1529.
249. Xie,H., Ma,H. and Zhou,D. (2013) Plasma HULC as a Promising Novel Biomarker for the Detection of Hepatocellular Carcinoma. *Biomed Res. Int.*, **2013**.
250. Schalken,J., Dijkstra,S., Baskin-Bey,E. and Oort,I. van (2014) Potential utility of cancer-specific biomarkers for assessing response to hormonal treatments in metastatic prostate cancer. *Ther. Adv. Urol.*, **6**, 245.
251. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, **74**, 5463–7.
252. Kchouk,M., Gibrat,J.F. and Elloumi,M. (2017) Generations of Sequencing Technologies: From First to Next Generation. *Biol. Med.*, **9**, 1–8.
253. Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.
254. Besser,J., Carleton,H.A., Gerner-Smidt,P., Lindsey,R.L. and Trees,E. (2018) Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin. Microbiol. Infect.*, **24**, 335–341.
255. Van Dijk,E.L., Jaszczyszyn,Y., Naquin,D. and Thermes,C. (2018) The Third Revolution in Sequencing Technology. *Trends Genet.*, **34**, 666–681.
256. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
257. Xuan,J., Yu,Y., Qing,T., Guo,L. and Shi,L. (2013) Next-generation sequencing in the clinic: promises and challenges. *Cancer Lett.*, **340**, 284–95.
258. Rai,M.F., Tycksen,E.D., Sandell,L.J. and Brophy,R.H. (2018) Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears. *J. Orthop. Res.*, **36**, 484–497.
259. Buettner,F., Natarajan,K.N., Casale,F.P., Proserpio,V., Scialdone,A., Theis,F.J., Teichmann,S.A., Marioni,J.C. and Stegle,O. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.
260. Pedraza,J.M., Paulsson,J., Gatfield,D., Schneider,K., Schibler,U. and Naef,F. (2008) Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, **319**, 339–43.
261. Kalisky,T., Blainey,P. and Quake,S.R. (2011) Genomic Analysis at the Single-Cell Level. *Annu. Rev. Genet.*, **45**, 431–445.
262. Raj,A. and van Oudenaarden,A. (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, **135**, 216–26.
263. Wu,A.R., Neff,N.F., Kalisky,T., Dalerba,P., Treutlein,B., Rothenberg,M.E., Mburu,F.M., Mantalas,G.L., Sim,S., Clarke,M.F., *et al.* (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, **11**, 41–46.
264. Raj,A., Peskin,C.S., Tranchina,D., Vargas,D.Y. and Tyagi,S. (2006) Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biol.*, **4**, e309.
265. Grün,D., Lyubimova,A., Kester,L., Wiebrands,K., Basak,O., Sasaki,N., Clevers,H. and van Oudenaarden,A. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types.

Nature, **525**, 251–255.
266. Artegiani,B., Lyubimova,A., Muraro,M., van Es,J.H., van Oudenaarden,A. and Clevers,H. (2017) A Single-Cell RNA Sequencing Study Reveals Cellular and Molecular Dynamics of the Hippocampal Neurogenic Niche. *Cell Rep.*, **21**, 3271–3284.
267. Rambow,F., Rogiers,A., Marin-Bejar,O., Aibar,S., Femel,J., Dewaele,M., Karras,P., Brown,D., Chang,Y.H., Debiec-Rychter,M., *et al.* (2018) Toward Minimal Residual Disease-Directed Therapy in Melanoma. *Cell*, **174**, 843–855.e19.
268. Villani,A.-C., Satija,R., Reynolds,G., Sarkizova,S., Shekhar,K., Fletcher,J., Griesbeck,M., Butler,A., Zheng,S., Lazo,S., *et al.* (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, **356**, eaah4573.
269. Venteicher,A.S., Tirosh,I., Hebert,C., Yizhak,K., Neftel,C., Filbin,M.G., Hovestadt,V., Escalante,L.E., Shaw,M.L., Rodman,C., *et al.* (2017) Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science*, **355**, eaai8478.
270. Tang,F., Barbacioru,C., Bao,S., Lee,C., Nordman,E., Wang,X., Lao,K. and Surani,M.A. (2010) Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis. *Cell Stem Cell*, **6**, 468–478.
271. Semrau,S., Goldmann,J.E., Soumillon,M., Mikkelsen,T.S., Jaenisch,R. and van Oudenaarden,A. (2017) Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat. Commun.*, **8**, 1096.
272. Yan,L., Yang,M., Guo,H., Yang,L., Wu,J., Li,R., Liu,P., Lian,Y., Zheng,X., Yan,J., *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, **20**, 1131–1139.
273. Streets,A.M., Zhang,X., Cao,C., Pang,Y., Wu,X., Xiong,L., Yang,L., Fu,Y., Zhao,L., Tang,F., *et al.* (2014) Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 7048–53.
274. Ramsköld,D., Luo,S., Wang,Y.-C., Li,R., Deng,Q., Faridani,O.R., Daniels,G.A., Khrebtukova,I., Loring,J.F., Laurent,L.C., *et al.* (2012) Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells HHS Public Access. *Nat Biotechnol*, **30**, 777–782.
275. Bhargava,V., Head,S.R., Ordoukhanian,P., Mercola,M. and Subramaniam,S. (2015) Technical Variations in Low-Input RNA-seq Methodologies. *Sci. Rep.*, **4**, 3678.
276. Junker,J.P. and van Oudenaarden,A. (2014) Every Cell Is Special: Genome-wide Studies Add a New Dimension to Single-Cell Biology. *Cell*, **157**, 8–11.
277. Paper,W., Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., *et al.* (2017) The human cell atlas. *Elife*, **6**, 1–30.
278. Rozenblatt-Rosen,O., Stubbington,M.J.T., Regev,A. and Teichmann,S.A. (2017) The Human Cell Atlas: from vision to reality. *Nature*, **550**, 451–453.
279. Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.
280. Levsky,J.M., Shenoy,S.M., Pezo,R.C. and Singer,R.H. (2002) Single-cell gene expression profiling. *Science*, **297**, 836–40.
281. Kolodziejczyk,A.A. and Lönnberg,T. (2018) Global and targeted approaches to single-cell transcriptome characterization. *Brief. Funct. Genomics*, **17**, 209–219.
282. Chiang,M.-K. and Melton,D.A. (2003) Single-Cell Transcript Analysis of Pancreas Development. *Dev. Cell*, **4**, 383–393.
283. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A., *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. **6**.
284. Islam,S., Kjällquist,U., Moliner,A., Zajac,P., Fan,J.-B., Lönnerberg,P. and Linnarsson,S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, **21**, 1160–7.
285. Soumillon,M., Cacchiarelli,D., Semrau,S., van Oudenaarden,A. and Mikkelsen,T.S. (2014) Characterization of directed differentiation by high-throughput single-cell RNA-Seq - SI. *bioRxiv*,

10.1101/003236.
286. Schmidt,W.M. and Mueller,M.W. (1999) a highly sensitive method for 5′ CAP-dependent enrichment.pdf. **27**, 2–5.
287. Picelli,S., Faridani,O.R., Björklund,Å.K., Winberg,G., Sagasser,S. and Sandberg,R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.
288. Schroth,G.P., Gertz,J., Myers,R.M., Williams,B.A., McCue,K., Marinov,G.K. and Wold,B.J. (2013) From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.*, **24**, 496–510.
289. Mora-Castilla,S., To,C., Vaezeslami,S., Morey,R., Srinivasan,S., Dumdie,J.N., Cook-Andersen,H., Jenkins,J. and Laurent,L.C. (2016) Miniaturization Technologies for Efficient Single-Cell Library Preparation for Next-Generation Sequencing. *J. Lab. Autom.*, **21**, 557–67.
290. Hashimshony,T., Wagner,F., Sher,N. and Yanai,I. (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, **2**, 666–73.
291. Meador,J.P., Lech,J.J., Rice,S.D., Hose,J.E., Short,J.W., Rice,S.D., Collier,T.K., Scholz,N.L., Collier,T.K., Scholz,N.L., *et al.* (2014) Massively Parallel Single-Cell. *Science (80-.).*, **343**, **Issue**, 776–779.
292. Ziegenhain,C., Vieth,B., Parekh,S., Reinius,B., Guillaumet-Adkins,A., Smets,M., Leonhardt,H., Heyn,H., Hellmann,I. and Enard,W. (2017) Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell*, **65**, 631–643.e4.
293. Picelli,S., Björklund,Å.K., Faridani,O.R., Sagasser,S., Winberg,G. and Sandberg,R. smart-seq2 for sensitive full-length transcriptome profiling in single cells. 10.1038/nMeth.2639.
294. Hashimshony,T., Senderovich,N., Avital,G., Klochendler,A., de Leeuw,Y., Anavy,L., Gennert,D., Li,S., Livak,K.J., Rozenblatt-Rosen,O., *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, **17**, 77.
295. Hochgerner,H., Lönnerberg,P., Hodge,R., Mikes,J., Heskol,A., Hubschle,H., Lin,P., Picelli,S., La Manno,G., Ratz,M., *et al.* (2017) STRT-seq-2i: dual-index 5′ single cell and nucleus RNA-seq on an addressable microwell array. *Sci. Rep.*, **7**, 16327.
296. Baran-Gale,J., Chandra,T. and Kirschner,K. (2018) Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics*, **17**, 233–239.
297. Prakadan,S.M., Shalek,A.K. and Weitz,D.A. (2017) Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.*, **18**, 345–361.
298. Wu,A.R., Neff,N.F., Kalisky,T., Dalerba,P., Treutlein,B., Rothenberg,M.E., Mburu,F.M., Mantalas,G.L., Sim,S., Clarke,M.F., *et al.* (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, **11**, 41–6.
299. Shalek,A.K., Satija,R., Shuga,J., Trombetta,J.J., Gennert,D., Lu,D., Chen,P., Gertner,R.S., Gaublomme,J.T., Yosef,N., *et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, **510**, 363–369.
300. Zhang,X., Li,T., Liu,F., Chen,Y., Yao,J., Li,Z., Huang,Y. and Wang,J. (2018) Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol. Cell*, 10.1016/J.MOLCEL.2018.10.020.
301. Zilionis,R., Nainys,J., Veres,A., Savova,V., Zemmour,D., Klein,A.M. and Mazutis,L. (2017) Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.*, **12**, 44–73.
302. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M., *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.
303. Marcy,Y., Ishoey,T., Lasken,R.S., Stockwell,T.B., Walenz,B.P., Halpern,A.L., Beeson,K.Y., Goldberg,S.M.D. and Quake,S.R. (2007) Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet.*, **3**, 1702–1708.
304. Fan,H.C., Fu,G.K. and Fodor,S.P.A. (2015) Combinatorial labeling of single cells for gene expression cytometry. *Science (80-.).*, **347**.
305. Gierahn,T.M., Wadsworth,M.H., Hughes,T.K., Bryson,B.D., Butler,A., Satija,R., Fortune,S., Love,J.C. and Shalek,A.K. (2017) Seq-Well: portable, low-cost RNA sequencing of single cells at

high throughput. *Nat. Methods*, 10.1038/nmeth.4179.
306. Han,X., Wang,R., Zhou,Y., Fei,L., Sun,H., Lai,S., Saadatpour,A., Zhou,Z., Chen,H., Ye,F., *et al.* (2018) Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, **172**, 1091–1107.e17.
307. Goldstein,L.D., Chen,Y.J.J., Dunne,J., Mir,A., Hubschle,H., Guillory,J., Yuan,W., Zhang,J., Stinson,J., Jaiswal,B., *et al.* (2017) Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics*, **18**, 1–10.
308. Verboom,K., Everaert,C., Bolduc,N., Livak,K.J., Yigit,N., Rombaut,D., Anckaert,J., Veno,M.T., Kjems,J., Speleman,F., *et al.* (2018) SMARTer single cell total RNA sequencing. *bioRxiv*, 10.1101/430090.
309. Hayashi,T., Ozaki,H., Sasagawa,Y., Umeda,M., Danno,H. and Nikaido,I. (2018) Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.*, **9**, 619.
310. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.
311. Wang,Y.J., Schug,J., Lin,J., Wang,Z. and Kossenkov,A. (2019) Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues.
312. Kaihara,K. and Agresti,J. (2018) Implementing the Drop-Seq Protocol on Bio-Rad â€™ s ddSEQ Single-Cell Isolator. *BioRad Bull.*
313. Verboom,K., Everaert,C., Bolduc,N., Livak,K.J., Yigit,N., Rombaut,D., Anckaert,J., Lee,S., Venø,M.T., Kjems,J., *et al.* (2019) SMARTer single cell total RNA sequencing. *Nucleic Acids Res.*, 10.1093/nar/gkz535.
314. Sheng,K., Cao,W., Niu,Y., Deng,Q. and Zong,C. (2017) Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods*, **14**.
315. Hayashi,T., Ozaki,H., Sasagawa,Y., Umeda,M., Danno,H. and Nikaido,I. (2018) Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.*, **9**, 619.
316. Fan,X., Zhang,X., Wu,X., Guo,H., Hu,Y., Tang,F. and Huang,Y. (2015) Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.*, **16**, 148.
317. Nikaido,I., Ebisawa,M., Sasagawa,Y., Kurisaki,A., Danno,H., Tanaka,K., Hayashi,T. and Takada,H. (2018) Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.*, **19**, 1–24.
318. Gierahn,T.M., Wadsworth,M.H., Hughes,T.K., Bryson,B.D., Butler,A., Satija,R., Fortune,S., Love,J.C. and Shalek,A.K. (2017) Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods*, **14**, 395–398.
319. Hashimshony,T., Senderovich,N., Avital,G., Klochendler,A., de Leeuw,Y., Anavy,L., Gennert,D., Li,S., Livak,K.J., Rozenblatt-Rosen,O., *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, **17**, 77.
320. Fish,R.N., Bostick,M., Lehman,A. and Farmer,A. (2016) Transcriptome analysis at the single-cell level using SMART technology. *Curr. Protoc. Mol. Biol.*, **2016**, 1–24.
321. Jaitin,D.A., Kenigsberg,E., Keren-Shaul,H., Elefant,N., Paul,F., Zaretsky,I., Mildner,A., Cohen,N., Jung,S., Tanay,A., *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, **343**, 776–9.
322. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A., *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods 2009 65*, **6**, 377.
323. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A., *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.
324. Liu,Y., Chen,X., Zhang,Y. and Liu,J. (2019) Advancing single-cell proteomics and metabolomics

with microfluidic technologies. *Analyst*, **144**, 846–858.
325. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M., *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.
326. Lai,F., Gardini,A., Zhang,A. and Shiekhattar,R. (2015) Integrator mediates the biogenesis of enhancer RNAs. *Nature*, **525**, 399–403.
327. Yang,L., Duff,M.O., Graveley,B.R., Carmichael,G.G. and Chen,L.-L. (2011) Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.*, **12**, R16.
328. Jeck,W.R. and Sharpless,N.E. (2014) Detecting and characterizing circular RNAs. *Nat. Biotechnol.*, **32**, 453–61.
329. Ramsköld,D., Luo,S., Wang,Y.-C., Li,R., Deng,Q., Faridani,O.R., Daniels,G.A., Khrebtukova,I., Loring,J.F., Laurent,L.C., *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–782.
330. Islam,S., Zeisel,A., Joost,S., La Manno,G., Zajac,P., Kasper,M., Lönnerberg,P. and Linnarsson,S. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, **11**, 163–166.
331. Grün,D., Kester,L. and van Oudenaarden,A. (2014) Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**, 637–640.
332. Zhang,X., Li,T., Liu,F., Chen,Y., Li,Z., Huang,Y. and Wang,J. (2018) Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. *bioRxiv*, 10.1101/313130.
333. Baran-Gale,J., Chandra,T. and Kirschner,K. (2018) Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics*, **17**, 233–239.
334. Pollen,A.A., Nowakowski,T.J., Shuga,J., Wang,X., Leyrat,A.A., Lui,J.H., Li,N., Szpankowski,L., Fowler,B., Chen,P., *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, **32**, 1053–1058.
335. Svensson,V., Natarajan,K.N., Ly,L.-H., Miragaia,R.J., Labalette,C., Macaulay,I.C., Cvejic,A. and Teichmann,S.A. (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods*, 10.1038/nmeth.4220.
336. Yin,Z., Lan,H., Tan,G., Lu,M., Vasilakos,A. V and Liu,W. (2017) Computing Platforms for Big Biological Data Analytics: Perspectives and Challenges. *Comput. Struct. Biotechnol. J.*, **15**, 403–411.
337. Spjuth,O., Bongcam-Rudloff,E., Dahlberg,J., Dahlö,M., Kallio,A., Pireddu,L., Vezzi,F. and Korpelainen,E. (2016) Recommendations on E-infrastructures for next-generation sequencing. *Gigascience*, **5**, 1–9.
338. De Brevern,A.G., Meyniel,J.-P., Fairhead,C., Neuvéglise,C. and Malpertuy,A. (2015) Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies. *Biomed Res. Int.*, **2015**, 904541.
339. Yu,P. and Lin,W. (2016) Single-cell Transcriptome Study as Big Data. *Genomics. Proteomics Bioinformatics*, **14**, 21–30.
340. Hwang,B., Lee,J.H. and Bang,D. (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.*, **50**.
341. Tian,L., Dong,X., Freytag,S., Lê Cao,K.A., Su,S., JalalAbadi,A., Amann-Zalcenstein,D., Weber,T.S., Seidi,A., Jabbari,J.S., *et al.* (2019) Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods*, **16**, 479–487.
342. Saelens,W., Cannoodt,R., Todorov,H. and Saeys,Y. (2018) A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. 10.1101/276907.
343. Soneson,C. and Robinson,M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, **15**, 255–261.
344. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J., *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.

345. Luecken,M.D. and Theis,F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, **15**, e8746.
346. McGinnis,C.S., Murrow,L.M. and Gartner,Z.J. (2019) DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.*, 10.1016/j.cels.2019.03.003.
347. AlJanahi,A.A., Danielsen,M. and Dunbar,C.E. (2018) An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Mol. Ther. Methods Clin. Dev.*, **10**, 189–196.
348. Lun,A.T.L., McCarthy,D.J. and Marioni,J.C. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data. *F1000Research*, **5**, 2122.
349. Young,M.D. and Behjati,S. (2018) SoupX removes ambient RNA contamination from droplet based single cell RNA sequenc-ing data. 10.1101/303727.
350. Li,W.V. and Li,J.J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, **9**, 997.
351. Gong,W., Kwak,I.-Y., Pota,P., Koyano-Nakagawa,N. and Garry,D.J. (2018) DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, **19**, 220.
352. Pierson,E. and Yau,C. (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, **16**, 241.
353. Ren,X., Kang,B. and Zhang,Z. (2018) Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biol.*, **19**.
354. L. Lun,A.T., Bach,K. and Marioni,J.C. (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 75.
355. Svensson,V., Natarajan,K.N., Ly,L., Miragaia,R.J., Labalette,C., Macaulay,I.C., Cvejic,A. and Teichmann,S.A. (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat. Publ. Gr.*, **14**, 381–387.
356. Kapteyn,J., He,R., McDowell,E.T. and Gang,D.R. (2010) Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. *BMC Genomics*, **11**, 413.
357. Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann and Wolfgang Enard7 (2017) Comparative Analysis of Single-Cell RNA Sequencing Methods: Molecular Cell. 10.1016/j.molcel.2017.01.023.
358. Barber,R.D., Harmer,D.W., Coleman,R.A. and Clark,B.J. (2005) GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol. Genomics*, **21**, 389–395.
359. Ma,F. and Pellegrini,M. (2019) Automated identification of Cell Types in Single Cell RNA Sequencing. *bioRxiv*, 10.1101/532093.
360. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M., *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
361. Guo,H., Zhu,P., Guo,F., Li,X., Wu,X., Fan,X., Wen,L. and Tang,F. (2015) Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nat. Protoc.*, **10**, 645–659.
362. Van den Bos,H., Bakker,B., Spierings,D.C.J., Lansdorp,P.M. and Foijer,F. (2018) Single-cell sequencing to quantify genomic integrity in cancer. *Int. J. Biochem. Cell Biol.*, **94**, 146–150.
363. X ie,X.S., Ma,F., Chapman,A., Lu,S. and Huang,L. (2015) Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. *Annu. Rev. Genomics Hum. Genet.*, **16**, 79–102.
364. Grün,D. and van Oudenaarden,A. (2015) Design and Analysis of Single-Cell Sequencing Experiments. *Cell*, **163**, 799–810.
365. Telenius,H., Carter,N.P., Bebb,C.E., Nordenskjo¨ld,M., Ponder,B.A.J. and Tunnacliffe,A. (1992) Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics*, **13**, 718–725.

366. Shpunt,A., Stamatoyannopoulos,J.A., Sunyaev,S.R., Lee,C., Ng,S.B., Nickerson,D.A., Shendure,J., Lord,C., Dagli,A., Battaglia,A., *et al.* (2012) Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell.
367. Leung,M.L., Wang,Y., Kim,C., Gao,R., Jiang,J., Sei,E. and Navin,N.E. (2016) Highly multiplexed targeted DNA sequencing from single nuclei. *Nat. Protoc.*, **11**, 214–235.
368. Jones,K.W., Geis,J.A., Gokhale,K., Pellegrino,M., Sciambi,A., Oldham,W., Eastburn,D.J., Futreal,P.A., Matthews,J., Treusch,S., *et al.* (2018) High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Res.*, **28**, 1345–1352.
369. Shema,E., Bernstein,B.E. and Buenrostro,J.D. (2019) Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat. Genet.*, **51**, 19–25.
370. Yu,V.W.C., Yusuf,R.Z., Oki,T., Wu,J., Saez,B., Wang,X., Cook,C., Baryawno,N., Ziller,M.J., Lee,E., *et al.* (2016) Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. *Cell*, **167**, 1310–1322.e17.
371. John,S., Sabo,P.J., Thurman,R.E., Sung,M.-H., Biddie,S.C., Johnson,T.A., Hager,G.L. and Stamatoyannopoulos,J.A. (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.*, **43**, 264–268.
372. Zhu,P., Wen,L., Wu,X., Tang,F., Guo,H. and Li,X. (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.*, **23**, 2126–2135.
373. Wen,L. and Tang,F. (2018) Single cell epigenome sequencing technologies. *Mol. Aspects Med.*, **59**, 62–69.
374. Smallwood,S.A., Lee,H.J., Angermueller,C., Krueger,F., Saadeh,H., Peat,J., Andrews,S.R., Stegle,O., Reik,W. and Kelsey,G. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods*, **11**, 817–820.
375. Mulqueen,R.M., Pokholok,D., Norberg,S.J., Torkenczy,K.A., Fields,A.J., Sun,D., Sinnamon,J.R., Shendure,J., Trapnell,C., O'Roak,B.J., *et al.* (2018) Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.*, **36**, 428–431.
376. Luo,C., Keown,C.L., Kurihara,L., Zhou,J., He,Y., Li,J., Castanon,R., Lucero,J., Nery,J.R., Sandoval,J.P., *et al.* (2017) Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, **357**, 600–604.
377. Ruscio,A. Di, Ebralidze,A.K., Benoukraf,T., Goff,L.A., Terragni,J., Figueroa,M.E., De,L.L., Pontes,F., Alberich-jorda,M., Zhang,P., *et al.* (2014) Lineage-specific and single cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.*, **503**, 371–376.
378. Tanay,A. and Regev,A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature*, **541**, 331–338.
379. Cusanovich,D.A., Daza,R., Adey,A., Pliner,H.A., Christiansen,L., Gunderson,K.L., Steemers,F.J., Trapnell,C. and Shendure,J. (2015) Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, **348**, 910–4.
380. Kelsey,G., Stegle,O. and Reik,W. (2017) Single-cell epigenomics: Recording the past and predicting the future. *Science*, **358**, 69–75.
381. Skene,P.J. and Henikoff,S. (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*, **6**.
382. Rotem,A., Ram,O., Shoresh,N., Sperling,R.A., Goren,A., Weitz,D.A. and Bernstein,B.E. (2015) Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, **33**, 1165–1172.
383. Skene,P.J., Henikoff,J.G. and Henikoff,S. (2018) Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat. Protoc.*, **13**, 1006–1019.
384. Hainer,S.J., Bošković,A., Rando,O.J. and Fazzio,T.G. (2018) Profiling of pluripotency factors in individual stem cells and early embryos. *bioRxiv*, 10.1101/286351.
385. Kind,J., Pagie,L., de Vries,S.S., Nahidiazar,L., Dey,S.S., Bienko,M., Zhan,Y., Lajoie,B., de Graaf,C.A., Amendola,M., *et al.* (2015) Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell*, **163**, 134–147.

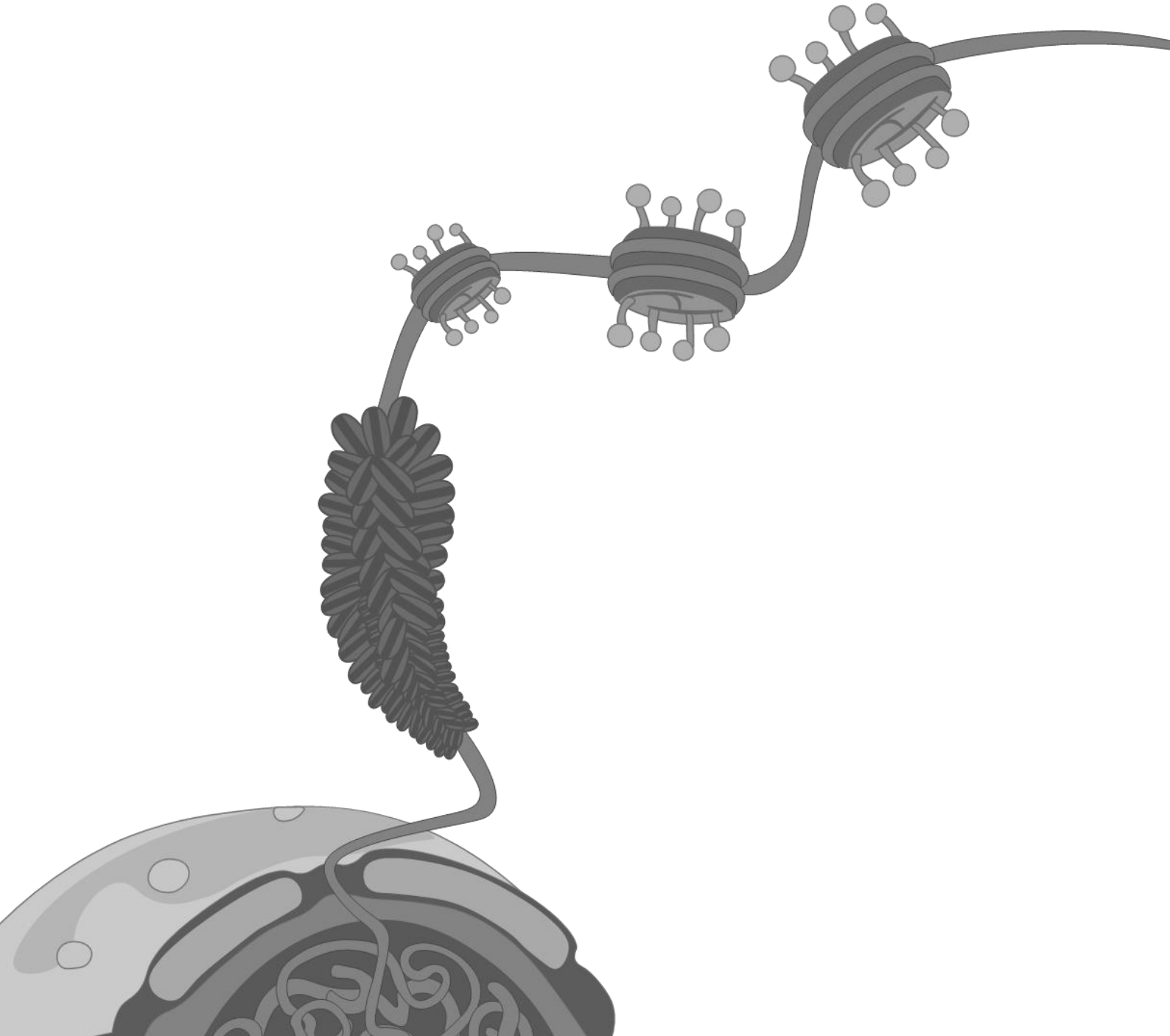
386. Aughey,G.N. and Southall,T.D. (2016) Dam it's good! DamID profiling of protein-DNA interactions. *Wiley Interdiscip. Rev. Dev. Biol.*, **5**, 25–37.
387. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–11.
388. Jia,R., Chai,P., Zhang,H. and Fan,X. (2017) Novel insights into chromosomal conformations in cancer. *Mol. Cancer*, **16**, 173.
389. Zhao,Z., Tavoosidana,G., Sjölinder,M., Göndör,A., Mariano,P., Wang,S., Kanduri,C., Lezcano,M., Singh Sandhu,K., Singh,U., *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.*, **38**, 1341–1347.
390. Ferraiuolo,M.A., Sanyal,A., Naumova,N., Dekker,J. and Dostie,J. (2012) Mapping chromatin interactions with 5C technology: 5C; a quantitative approach to capturing chromatin conformation over large genomic distances. *Methods*, **58**, 255–67.
391. Sexton,T., Schober,H., Fraser,P. and Gasser,S.M. (2007) Gene regulation through nuclear organization. *Nat. Struct. Mol. Biol.*, **14**, 1049–1055.
392. Flyamer,I.M., Gassler,J., Imakaev,M., Brandão,H.B., Ulianov,S. V., Abdennur,N., Razin,S. V., Mirny,L.A. and Tachibana-Konwalski,K. (2017) Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, **544**, 110–114.
393. Nagano,T., Lubling,Y., Stevens,T.J., Schoenfelder,S., Yaffe,E., Dean,W., Laue,E.D., Tanay,A. and Fraser,P. (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, 59–64.
394. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S., *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, **159**, 1665–1680.
395. Ramani,V., Deng,X., Qiu,R., Gunderson,K.L., Steemers,F.J., Disteche,C.M., Noble,W.S., Duan,Z. and Shendure,J. (2017) Massively multiplex single-cell Hi-C. *Nat. Methods*, **14**, 263–266.
396. Nagano,T., Lubling,Y., Yaffe,E., Wingett,S.W., Dean,W., Tanay,A. and Fraser,P. (2015) Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat. Protoc.*, **10**, 1986–2003.
397. Nagano,T., Lubling,Y., Várnai,C., Dudley,C., Leung,W., Baran,Y., Mendelson Cohen,N., Wingett,S., Fraser,P. and Tanay,A. (2017) Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, **547**, 61–67.
398. Shahi,P., Kim,S.C., Haliburton,J.R., Gartner,Z.J. and Abate,A.R. (2017) Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci. Rep.*, **7**, 44447.
399. Hughes,A.J., Spelke,D.P., Xu,Z., Kang,C.-C., Schaffer,D. V and Herr,A.E. (2014) Single-cell western blotting. *Nat. Methods*, **11**, 749–755.
400. Budnik,B., Levy,E., Harmange,G. and Slavov,N. (2018) SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.*, **19**, 161.
401. Macaulay,I.C., Ponting,C.P. and Voet,T. (2017) Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends Genet.*, **33**, 155–168.
402. Hu,Y., An,Q., Sheu,K., Trejo,B., Fan,S. and Guo,Y. (2018) Single Cell Multi-Omics Technology: Methodology and Application. *Front. cell Dev. Biol.*, **6**, 28.
403. Hou,Y., Guo,H., Cao,C., Li,X., Hu,B., Zhu,P., Wu,X., Wen,L., Tang,F., Huang,Y., *et al.* (2016) Single-cell triple omics sequencing reveals genetic, epigenetic and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.*, **26**, 304–319.
404. Macaulay,I.C., Haerty,W., Kumar,P., Li,Y.I., Hu,T.X., Teng,M.J., Goolam,M., Saurat,N., Coupland,P., Shirley,L.M., *et al.* (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods*, **12**, 519–522.
405. Clark,S.J., Argelaguet,R., Kapourani,C.-A., Stubbs,T.M., Lee,H.J., Alda-Catalinas,C., Krueger,F., Sanguinetti,G., Kelsey,G., Marioni,J.C., *et al.* (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.*, **9**, 781.

406. Dey,S.S., Kester,L., Spanjaard,B., Bienko,M. and van Oudenaarden,A. (2015) Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.*, **33**, 285–289.
407. Jin,W., Tang,Q., Wan,M., Cui,K., Zhang,Y., Ren,G., Ni,B., Sklar,J., Przytycka,T.M., Childs,R., *et al.* (2015) Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature*, **528**, 142–6.
408. Angermueller,C., Clark,S.J., Lee,H.J., Macaulay,I.C., Teng,M.J., Hu,T.X., Krueger,F., Smallwood,S.A., Ponting,C.P., Voet,T., *et al.* (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, **13**, 229–232.
409. Luo,C., Keown,C.L., Kurihara,L., Zhou,J., He,Y., Li,J., Castanon,R., Lucero,J., Nery,J.R., Sandoval,J.P., *et al.* (2017) Elements in Mammalian Cortex. *Science (80-.).*, **604**, 600–604.
410. Zhu,X., Fan,G., Hu,Y., Hu,G., Huang,K., Wang,C.-Y., An,Q., Xue,Z., Xue,J. and Du,G. (2016) Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.*, **17**, 1–11.
411. Guo,F., Li,L., Li,J., Wu,X., Hu,B., Zhu,P., Wen,L. and Tang,F. (2017) Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.*, **27**, 967–988.
412. Kaya-Okur,H.S., Wu,S.J., Codomo,C.A., Pledger,E.S., Bryson,T.D., Henikoff,J.G., Ahmad,K. and Henikoff,S. (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. 10.1101/568915.
413. Stoeckius,M., Hafemeister,C., Stephenson,W., Houck-Loomis,B., Chattopadhyay,P.K., Swerdlow,H., Satija,R. and Smibert,P. (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, **14**, 865–868.
414. Peterson,V.M., Zhang,K.X., Kumar,N., Wong,J., Li,L., Wilson,D.C., Moore,R., McClanahan,T.K., Sadekova,S. and Klappenbach,J.A. (2017) Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.*, **35**, 936–939.
415. Levy,E. and Slavov,N. (2018) Single cell protein analysis for systems biology. *Essays Biochem.*, **62**, 595–605.
416. Garraway,L.A. and Lander,E.S. (2013) Lessons from the Cancer Genome. *Cell*, **153**, 17–37.
417. Nowell,P.C. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23–8.
418. Van Loo,P. and Voet,T. (2014) Single cell analysis of cancer genomes. *Curr. Opin. Genet. Dev.*, **24**, 82–91.
419. Nicholas,M.N., Jude,K., Jennifer,T., Peter,A., Linda,R., Jeanne,M., Kerry,C., Asya,S., Dan,L., Diane,E., *et al.* (2011) Tumor Evolution Inferred by Single Cell Sequencing. *Nature*, **472**, 90–94.
420. Navin,N. and Hicks,J. (2011) Future medical applications of single-cell sequencing in cancer. *Genome Med.*, **3**, 31.
421. Baslan,T. and Hicks,J. (2017) Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat. Rev. Cancer*, **17**, 557–569.
422. Landau,D.A., Carter,S.L., Stojanov,P., McKenna,A., Stevenson,K., Lawrence,M.S., Sougnez,C., Stewart,C., Sivachenko,A., Wang,L., *et al.* (2013) Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell*, **152**, 714–726.
423. Shalek,A.K. and Benson,M. (2017) Single-cell analyses to tailor treatments. *Sci. Transl. Med.*, **9**.
424. Racle,J., de Jonge,K., Baumgaertner,P., Speiser,D.E. and Gfeller,D. (2017) Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife*, **6**.
425. Ding,L., Ley,T.J., Larson,D.E., Miller,C.A., Koboldt,D.C., Welch,J.S., Ritchey,J.K., Young,M.A., Lamprecht,T., McLellan,M.D., *et al.* (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**, 506–510.
426. Gerlinger,M., Rowan,A.J., Horswell,S., Larkin,J., Endesfelder,D., Gronroos,E., Martinez,P., Matthews,N., Stewart,A., Tarpey,P., *et al.* (2012) Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.*, **366**, 883–892.
427. Navin,N.E. (2014) Cancer genomics: one cell at a time. *Genome Biol.*, **15**, 452.
428. Gasch,C., Bauernhofer,T., Pichler,M., Langer-Freitag,S., Reeh,M., Seifert,A.M., Mauermann,O., Izbicki,J.R., Pantel,K. and Riethdorf,S. (2013) Heterogeneity of epidermal growth factor receptor status and mutations of KRAS/PIK3CA in circulating tumor cells of patients with colorectal cancer.

Clin. Chem., **59**, 252–60.
429. Mcgranahan,N. and Swanton,C. (2017) Review Clonal Heterogeneity and Tumor Evolution : Past , Present , and the Future. *Cell*, **168**, 613–628.
430. De Bie,J., Demeyer,S., Alberti-Servera,L., Geerdens,E., Segers,H., Broux,M., De Keersmaecker,K., Michaux,L., Vandenberghe,P., Voet,T., *et al.* (2018) Single-cell sequencing reveals the origin and the order of mutation acquisition in T-cell acute lymphoblastic leukemia. *Leukemia*, 10.1038/s41375-018-0127-8.
431. Brenner,M.., Rill,D.., Krance,R.., Ihle,J.., Moen,R.., Mirro,J. and Anderson,W.. (1993) Gene-marking to trace origin of relapse after autologous bone-marrow transplantation. *Lancet*, **341**, 85–86.
432. Roesch,A., Vultur,A., Bogeski,I., Wang,H., Zimmermann,K.M., Speicher,D., Körbel,C., Laschke,M.W., Gimotty,P.A., Philipp,S.E., *et al.* (2013) Overcoming Intrinsic Multidrug Resistance in Melanoma by Blocking the Mitochondrial Respiratory Chain of Slow-Cycling JARID1Bhigh Cells. *Cancer Cell*, **23**, 811–825.
433. Luskin,M.R., Murakami,M.A., Manalis,S.R. and Weinstock,D.M. (2018) Targeting minimal residual disease: A path to cure? *Nat. Rev. Cancer*, **18**, 255–263.
434. Tirosh,I., Izar,B., Prakadan,S.M., Wadsworth Ii,M.H., Treacy,D., Trombetta,J.J., Rotem,A., Rodman,C., Lian,C., Murphy,G., *et al.* (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq HHS Public Access. *Sci. April*, **8**, 189–196.
435. Kim,C., Gao,R., Sei,E., Brandt,R., Hartman,J., Hatschek,T., Crosetto,N., Foukakis,T. and Navin,N.E. (2018) Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell*, **173**, 879–893.e13.
436. Brady,S.W., McQuerry,J.A., Qiao,Y., Piccolo,S.R., Shrestha,G., Jenkins,D.F., Layer,R.M., Pedersen,B.S., Miller,R.H., Esch,A., *et al.* (2017) Combating subclonal evolution of resistant cancer phenotypes. *Nat. Commun.*, **8**, 1231.
437. Guo,X., Zhang,Y., Zheng,L., Zheng,C., Song,J., Zhang,Q., Kang,B., Liu,Z., Jin,L., Xing,R., *et al.* (2018) Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.*, **24**, 978–985.
438. Tirosh,I., Izar,B., Prakadan,S.M., Wadsworth Ii,M.H., Treacy,D., Trombetta,J.J., Rotem,A., Rodman,C., Lian,C., Murphy,G., *et al.* (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq HHS Public Access. *Sci. April*, **8**, 189–196.
439. Savas,P., Virassamy,B., Ye,C., Salim,A., Mintoff,C.P., Caramia,F., Salgado,R., Byrne,D.J., Teo,Z.L., Dushyanthen,S., *et al.* (2018) Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med.*, **24**, 986–993.
440. Navin,N.E. (2015) The first five years of single-cell cancer genomics and beyond. *Genome Res.*, **25**, 1499–1507.
441. Lohr,J.G., Adalsteinsson,V.A., Cibulskis,K., Choudhury,A.D., Rosenberg,M., Cruz-Gordillo,P., Francis,J.M., Zhang,C.-Z., Shalek,A.K., Satija,R., *et al.* (2014) Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat. Biotechnol.*, **32**, 479–84.
442. Heitzer,E., Auer,M., Gasch,C., Pichler,M., Ulz,P., Hoffmann,E.M., Lax,S., Waldispuehl-Geigl,J., Mauermann,O., Lackner,C., *et al.* (2013) Complex Tumor Genomes Inferred from Single Circulating Tumor Cells by Array-CGH and Next-Generation Sequencing. *Cancer Res.*, **73**, 2965–2975.
443. Gao,Y., Ni,X., Guo,H., Su,Z., Ba,Y., Tong,Z., Guo,Z., Yao,X., Chen,X., Yin,J., *et al.* (2017) Single-cell sequencing deciphers a convergent evolution of copy number alterations from primary to circulating tumor cells. *Genome Res.*, **27**, 1312–1322.
444. Ni,X., Zhuo,M., Su,Z., Duan,J., Gao,Y., Wang,Z., Zong,C., Bai,H., Chapman,A.R., Zhao,J., *et al.* (2013) Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 21083–8.
445. Carter,L., Rothwell,D.G., Mesquita,B., Smowton,C., Leong,H.S., Fernandez-Gutierrez,F., Li,Y., Burt,D.J., Antonello,J., Morrow,C.J., *et al.* (2017) Molecular analysis of circulating tumor cells

identifies distinct copy-number profiles in patients with chemosensitive and chemorefractory small-cell lung cancer. *Nat. Med.*, **23**, 114–119.
446. Zhu,H., Lee,R.J., Desai,N., Trautwein,J., Arora,K.S., Shioda,T., Broderick,K.T., Zheng,Y., Fox,D.B., Toner,M., *et al.* (2015) RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science (80-.).*, **349**, 1351–1356.
447. Dago,A.E., Stepansky,A., Carlsson,A., Luttgen,M., Kendall,J., Baslan,T., Kolatkar,A., Wigler,M., Bethel,K., Gross,M.E., *et al.* (2014) Rapid phenotypic and genomic change in response to therapeutic pressure in prostate cancer inferred by high content analysis of single circulating tumor cells. *PLoS One*, **9**, e101777.
448. Meijerink,J.P.P. (2010) Genetic rearrangements in relation to immunophenotype and outcome in T-cell acute lymphoblastic leukaemia. *Best Pract. Res. Clin. Haematol.*, **23**, 307–318.
449. Rothenberg,E. V, Kueh,H.Y., Yui,M.A. and Zhang,J.A. (2016) Hematopoiesis and T-cell specification as a model developmental system. *Immunol. Rev.*, **271**, 72–97.
450. Rinn,J.L. and Chang,H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–166.
451. Pui,C.-H. and Jeha,S. (2007) New therapeutic strategies for the treatment of acute lymphoblastic leukaemia. *Nat. Rev. Drug Discov.*, **6**, 149–165.
452. Svensson,V. (2017) Moore ' s Law in Single Cell Transcriptomics.

2. Research objectives



Research objectives

Advances in NGS methods enabled to investigate whole genomes in a cost efficient and high-throughput manner. These methods are now widely used and contribute to our knowledge of molecular mechanisms involved in normal development and disease. Besides whole genomes, these methods also enable to unravel complete transcriptomes (RNA sequencing, RNA-seq). The first RNA-seq protocols only sequenced polyadenylated transcripts, whereby amongst others a large fraction of the long non-coding RNAs (lncRNAs) remained undetectable. Subsequently, total RNA-seq protocols were developed enabling to capture these non-polyadenylated as well as polyadenylated transcripts.

In this thesis, I combined polyA[+] and total RNA-seq of a large T-ALL cohort and an *in vitro TLX1* knockdown cell model system to identify lncRNAs involved in the development of TLX1/3 positive T-cell acute lymphoblastic leukemia (T-ALL) and more specifically lncRNAs regulated by TLX1 (**Aim 1**). As a bulk RNA-seq approach was followed, the average gene expression profile of cell populations was generated, possibly masking subtle differences among cells upon the perturbation. In contrast, single cell RNA-seq has the potential to detect, amongst others, transcriptional heterogeneity upon a perturbation. Therefore, I used a well-characterized cellular system upon chemical perturbation to develop a new single cell total RNA-seq protocol (**Aim 2**) and to evaluate three commercial single cell RNA-seq devices in terms of their ability to capture transcriptional heterogeneity and differentially expressed genes (**Aim 3**).

Aim 1: deciphering the TLX1 regulated lncRNAome in T-ALL

T-ALL is a highly aggressive haematological cancer associated with poor prognosis, however intensified therapeutic strategies have led to considerable improvements in patient survival in the past decade. Unfortunately, these treatments are associated with severe acute and long-term toxicities and a large fraction of the patients still relapse, underlying the need to better define the molecular basis of T-ALL. ’*T-cell leukemia homeobox 1’* (T*LX1*) is a major driver gene in T-ALL development, demarcating a molecular T-ALL subgroup with a specific gene expression profile for which downstream effects are already been thoroughly studied in terms of protein-coding genes. In my research project, I aimed to extend the TLX1 regulatory network in T-ALL towards lncRNAs, since it is now widely accepted that these lncRNAs can play an important role in the development of cancer and the role of lncRNAs in this disease remains largely unexplored. Therefore, I generated polyA[+] as well as total RNA-seq transcriptome data of ALL-SIL lymphoblasts upon *TLX1* knockdown and integrated ATAC-seq, H3K4me1, H3K4me3, H3K27ac and TLX1 ChIP-seq data to identify TLX1 regulated lncRNAs. I extended this dataset with polyA[+] and total RNA-seq of a large primary T-ALL cohort and aimed to identify TLX subgroup specific and possibly oncogenic lncRNAs (**Paper 1**). As this is a comprehensive and unique dataset in the T-ALL field that contains extensive unexplored information, I aimed to make the data publicly available and wrote a data descriptor with a detailed description of the methods used, enabling re-use of the dataset by the broader research community (**Paper 2**).

Aim 2: developing a single cell total RNA sequencing protocol

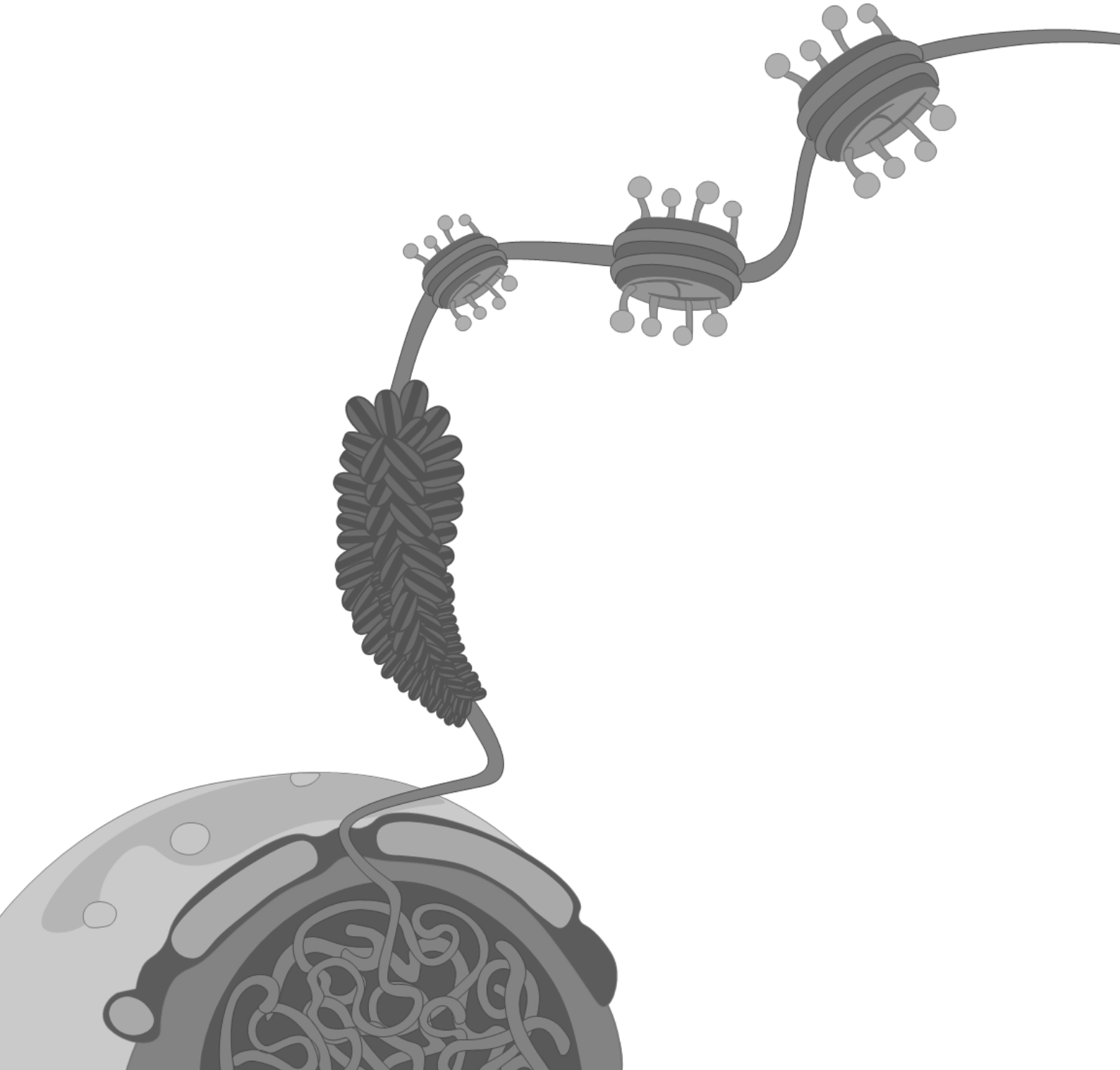
In 2012, Fluidigm released the C1, the first commercially available single cell RNA-seq device. Since then, the number of single cell RNA-seq methods raised rapidly thereby increasing the throughput and decreasing the cost per single cell. However, most methods could only capture polyadenylated transcripts, leaving the non-polyadenylated part of the transcriptome, including a large fraction of the lncRNAs and all circular RNAs (circRNAs), undetectable. Therefore, I aimed to develop a protocol that

enables capture of both polyadenylated and non-polyadenylated transcripts using the C1 instrument. Since the Fluidigm C1 is restricted to the capture of only 96 cells, I also aimed to validate the single cell total RNA-seq protocol for FACS sorted single cells, increasing the throughput and utility of the developed method (**Paper 3**).

Aim 3: evaluating transcriptional heterogeneity upon perturbation using three single cell RNA-seq devices

During my PhD mandate, the number of single cell sequencing devices increased rapidly, and consequently the number of single cells that can be captured in a single experiment increased drastically from a few to tens of thousands of single cells. Since each device has its own specifications, several studies compared these devices in terms of data quality and their ability to detect cellular subpopulations. However, none of these comparative studies focused on the detection of transcriptional heterogeneity upon a chemical perturbation. Therefore, I aimed to evaluate the C1 (Fluidigm), ddSeq (Bio-Rad, Illumina) and Chromium (10x Genomics) with respect to data quality, transcriptional heterogeneity and the ability to detect differential expressed genes (**Paper 4**).

3. Results



Paper 1

A comprehensive inventory of TLX1 controlled long non-coding RNAs in T-cell acute lymphoblastic leukemia through polyA+ and total RNA sequencing

Karen Verboom, Wouter Van Loocke, Pieter-Jan Volders, Bieke Decaesteker, Francisco Avila Cobos, Simon Bornschein, Charles E. de Bock, Zeynep Kalender Atak, Emmanuelle Clappier, Stein Aerts, Jan Cools, Jean Soulier, Tom Taghon, Pieter Van Vlierberghe, Jo Vandesompele, Frank Speleman and Kaat Durinck

Contribution: I performed H3K4me1 and H3K4me3 ChIP-seq and ATAC-seq, performed the data analysis of these experiments in R using ChIPSeeker and visualized the data in IGV. I contributed to the RNA-seq data analysis, starting from the count matrices by performing differential gene expression analysis, gene set enrichment analysis, cytoscape analysis and further downstream analyses in R. I made the figures and wrote the manuscript.

Published in Haematologica 103(12):e585-e589

Impact factor 2017: 9.09

A comprehensive inventory of TLX1 controlled long non-coding RNAs in T-cell acute lymphoblastic leukemia through polyA+ and total RNA sequencing

Karen Verboom[1,2], Wouter Van Loocke[1,2], Pieter-Jan Volders[1,2,3,4], Bieke Decaesteker[1,2], Francisco Avila Cobos[1,2,4], Simon Bornschein[5,6], Charles E. de Bock[5,6], Zeynep Kalender Atak[5,7], Emmanuelle Clappier[8], Stein Aerts[5,7], Jan Cools[5,6], Jean Soulier[8], Tom Taghon[2,9], Pieter Van Vlierberghe[1,2], Jo Vandesompele[1,2,4], Frank Speleman[1,2] and Kaat Durinck[1,2]

[1]*Center for Medical Genetics, Ghent University, Ghent, Belgium*

[2]*Cancer Research Institute Ghent, Ghent, Belgium*

[3]*Center for Medical Biotechnology, VIB-UGent, Ghent, Belgium*

[4]*Bioinformatics Institute Ghent from Nucleotides to Networks, BIG N2N, Ghent, Belgium*

[5]*KU Leuven Center for Human Genetics, Leuven, Belgium.*

[6]*VIB Center for Cancer Biology, Leuven, Belgium*

[7]*VIB Center for Brain & Disease Research, Laboratory of Computational Biology, Leuven, Belgium*

[8]*Hôpital Saint Louis, Institut Universitaire d'Hématologie, Paris, France*

[9]*Department of Clinical Chemistry, Microbiology and Immunology, Ghent University, Ghent, Belgium*

Corresponding author: Kaat Durinck, Kaat.Durinck@UGent.be

T-cell acute lymphoblastic leukemia (T-ALL) is an aggressive hematological malignancy arising from uncontrolled proliferation and arrested differentiation of precursor T-cells. T-ALL is a genetically heterogeneous disease and can be subdivided into different molecular cytogenetic subgroups associated with specific gene expression signatures.[1,2] The T-cell leukemia homeobox 1 (TLX1, HOX11) transcription factor is a key driver of the TLX subgroup in T-ALL with ectopic expression in developing thymocytes causing a maturation arrest at the early cortical stage of T-cell development. Aberrant TLX1 expression occurs in 5–10 % of pediatric and 30 % of adult T-ALL patients and predominantly results from t(7;10)(q34;q24) or t(10;14)(q24;q11) chromosomal translocations leading to juxtaposition of TLX1 to the T-cell receptor (TCR) δ or β promoter.[3] The TLX1 gene regulatory network has been extensively studied in terms of co-factors and downstream protein-coding gene targets.[4] Given that the protein-coding part of the genome only constitutes about 2 % while up to 70 % of the genome is transcribed (as non-coding ribonucleic acid (RNA)), a deeper exploration of the TLX1 driven non-coding transcriptome in T-ALL is warranted to support a more profound understanding of the molecular basis of this T-ALL subtype.[5] Long non-coding RNAs (lncRNAs) recently emerged as crucial transcriptional regulators in normal development and cancer, including normal and malignant hematopoiesis.[6,7] LncRNAs are arbitrarily defined as transcripts longer than 200 nucleotides and are poorly evolutionary conserved in terms of sequence.[8] Recently, our lab has identified a subset of lncRNAs that act in concert with NOTCH1 in both normal T-cell development and malignant T-cell transformation and a set of T-ALL subgroup-specific lncRNAs using microarray data.[9,10] In this study, we performed in vitro TLX1 knockdown in T-ALL cells as well as a deep exploration of the TLX subgroup-specific lncRNAome in primary T-ALLs. For the former, we applied an integrative

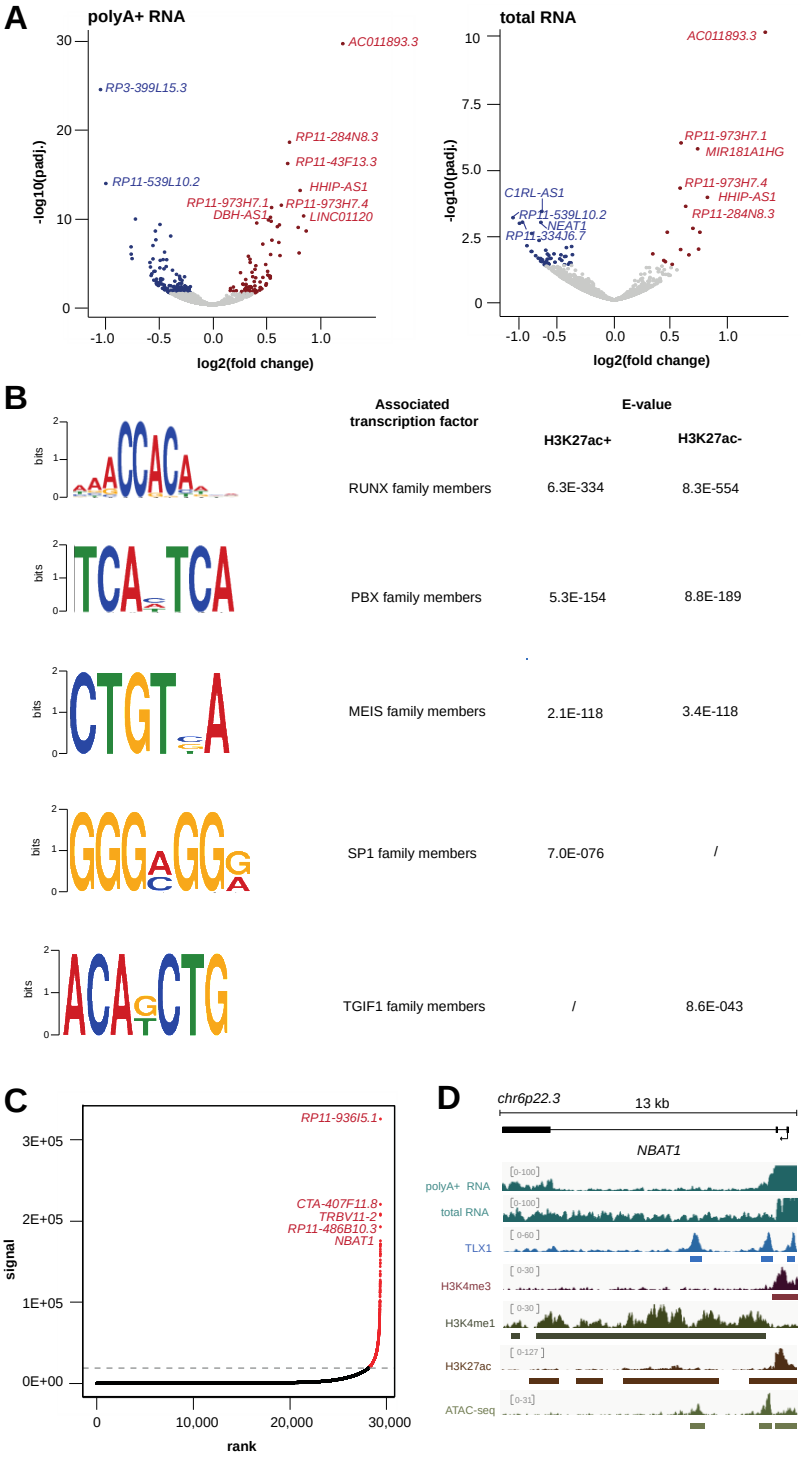


Figure 1: integrative TLX1 ChIP-seq and transcriptome analysis upon *TLX1* knockdown in ALL-SIL lymphoblasts for identification of a robust set of TLX1 directly regulated lncRNAs and super-enhancer associated lncRNAs. (A) Volcano plot representation of differentially expressed lncRNAs upon *TLX1* knockdown in ALL-SIL. Red (upregulated upon *TLX1* knockdown) and blue (downregulated upon *TLX1* knockdown) dots represent significantly differentially expressed lncRNAs detected with polyA+ RNA-seq (left panel) and total RNA-seq (right panel) (adjusted *P*-value <0.05). LncRNA names depicted in the plots are the top ten differentially regulated lncRNAs. Outliers with a -log10(padj) >30 are scaled to log10(padj)=30. (B) Motif enrichment analysis on the set of TLX1 bound regions with and without overlap of H3K27ac ChIP-seq peaks using MEME-ChIP suite identifies significant enrichment of the DNA binding motifs of the RUNX, PBX and MEIS family of transcription factors for both sets of peaks while the SP1 and TGIF1 families are only enriched in one set of peaks. (C) Hockey stick plot representing the normalized rank and cluster signal of clusters of H3K27ac ChIP-seq peaks at lncRNA transcripts. Red dots represent lncRNAs significantly associated with a super-enhancer (adjusted *P*-value <0.05). (D) IGV screen-shot of a super-enhancer associated lncRNA (*NBAT1*). PolyA+ and total RNA-seq tracks are depicted for control siRNA transfected samples. Bars represent the MACS2 peaks with FDR <0.05. RNA: ribonucleic acid. lncRNA: long non-coding RNA.

genomics approach combining quantitative data on the transcriptome and immunoprecipitated and open chromatin, using RNA-sequencing (RNA-seq), chromatin immunoprecipitation sequencing (ChIP-seq) and assay for transposase-accessible chromatin sequencing (ATAC-seq), respectively. Using this approach, we identified known and novel lncRNAs and gained insight into the super-enhancer marked lncRNA genetic landscape in TLX driven T-ALL, amongst others.

To elucidate the lncRNA repertoire under control of the TLX1 transcription factor, we performed transient TLX1 knockdown by electroporating two TLX1 targeting small interfering RNAs (siRNAs) in ALL-SIL lymphoblasts, displaying ectopic TLX1 expression as a result of a t(10;14)(q24;q11) translocation. From the resulting transcriptomes, both polyA+ and total RNA-seq libraries were generated in order to evaluate the expression changes of polyadenylated as well as non-polyadenylated lncRNA transcripts (Online Supplementary Figure S1A,B). By combining lncRNAs (biotype 'lincRNA' or 'antisense') detected with polyA+ (Figure 1A, left) and total RNA-seq (Figure 1A, right), more lncRNAs were significantly (adjusted P-value <0.05) downregulated (146 lncRNAs) than upregulated (80 lncRNAs) upon TLX1 knockdown (Online Supplementary Table S1). Up- or downregulation of nine of the top ten differentially TLX1-regulated lncRNAs detected by polyA+ and total RNA-seq could be validated by quantitative reverse transcription polymerase chain reaction (RT-qPCR) (Online Supplementary Figure S2). Of note, this significantly different ratio between up- and downregulated lncRNAs is contrary to the effect of TLX1 knockdown on protein-coding genes (mainly upregulated upon TLX1 knockdown) (Online Supplementary Figure S3A), in concordance with its previously described role as a transcriptional repressor.[4] Moreover, this opposite ratio remains intact upon integration of TLX1 ChIP-seq data (Online Supplementary Figure S3B). Using de novo motif analysis on transcriptionally active (H3K27Ac+) and inactive (H3K27Ac-) TLX1 bound regions, a significant enrichment for the RUNX, PBX and MEIS family of transcription factor motifs was observed for H3K27ac+ and H3K27ac- regions, as previously observed for TLX1-regulated protein-coding genes.[4] In contrast, some transcription factors such as SP1 and TGIF1 were only enriched in H3K27ac+ or H3K27ac- regions, suggesting that TLX1 activated genes can be transcriptionally regulated by different transcription factor families compared to TLX1 repressed genes (Figure 1B).

Among the 226 lncRNAs regulated by TLX1, 64 lncRNAs display a TLX1 chromatin binding peak in their immediate vicinity (max. 5 kb) (Online Supplementary Table S1), as illustrated for lncRNA RP11-539L10.2 (Online Supplementary Figure S4A). For 80 of the 226 differentially regulated lncRNAs upon TLX1 knockdown, the expression was significantly correlated with at least one neighboring protein-coding gene (|rho|(Rs) >0.5, P-value <0.05) located within a 100 kb window, irrespective of strand orientation (Online Supplementary Table S2). From the latter, 97.25 % are positively correlated with the expression of the differentially regulated lncRNAs, consistent with previous reports.[11]

Interestingly, three of the identified TLX1-regulated lncRNAs are in the vicinity (max 1 Mb) of a known differentially regulated T-ALL tumor suppressor gene[4] (Online Supplementary Figure S4B,C; Online Supplementary Table S1 and S3). To assign a possible function to the top five TLX1 up- and downregulated lncRNAs, a guilt-by-association approach was followed as described in the Online Supplementary Methods section of this paper (Online Supplementary Figure S5 and S6).

As it is known that some lncRNAs are located within super-enhancer regions, a hockey stick plot based on H3K27ac ChIP data for lncRNA loci was generated as described in Online Supplementary Methods (Figure 1C). Among the 2781 super-enhancer associated lncRNAs, 115 lncRNAs were significantly differentially expressed upon TLX1 knockdown with a

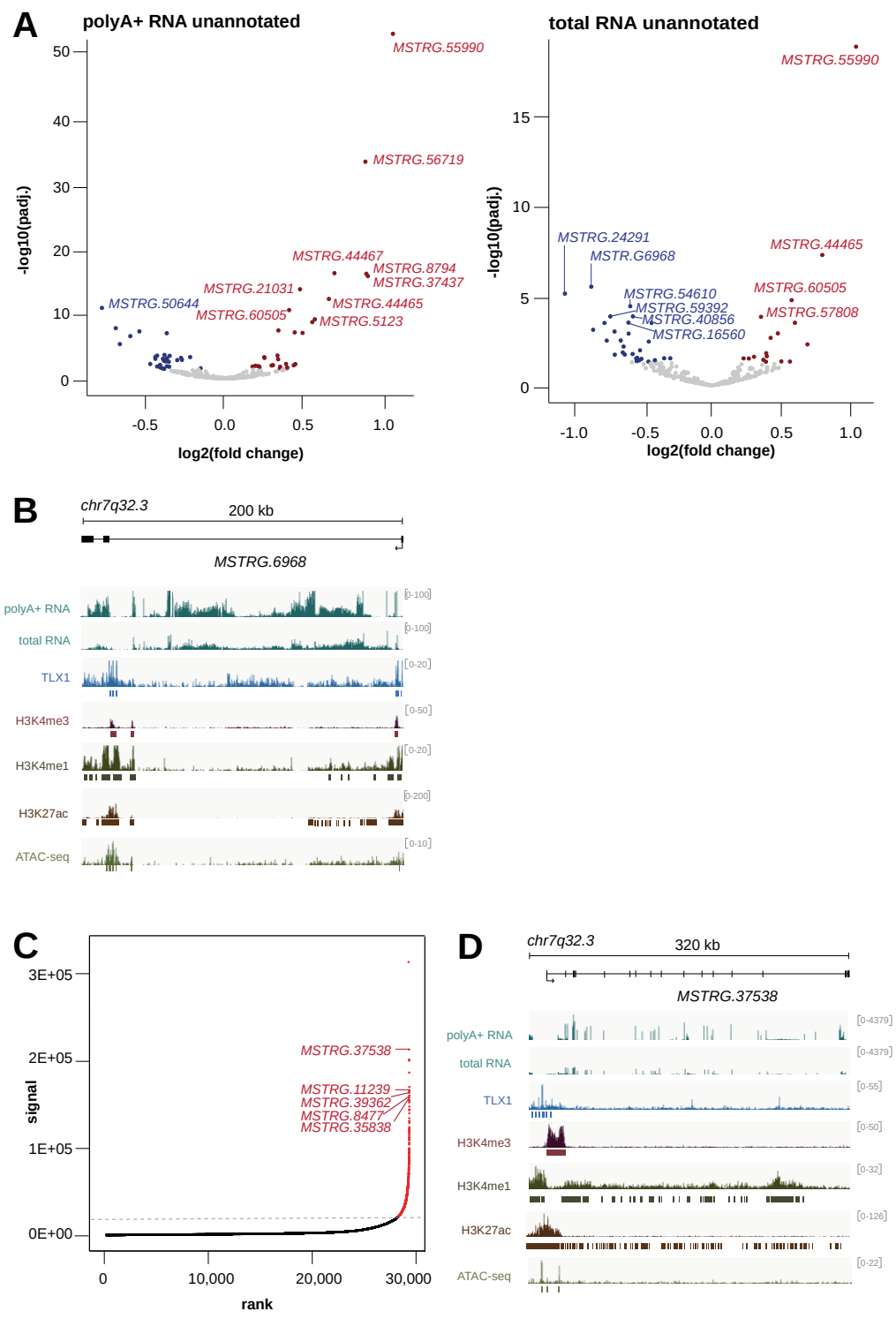


Figure 2: Identification of a set of previously unannotated TLX1 regulated lncRNAs in ALL-SIL lymphoblasts. (A) Volcano plot representation of unannotated differentially expressed lncRNAs upon *TLX1* knockdown in ALL-SIL. Red (upregulated upon TLX1 knockdown) and blue (downregulated upon *TLX1* knockdown) dots represent significantly differentially expressed lncRNAs detected with polyA+ RNA-seq (left panel) and total RNA-seq (right panel) (adjusted *P*-value <0.05). Gene names depicted in the plots are the top ten unannotated differentially regulated lncRNAs. (B) IGV screen-shot of an unannotated differentially expressed, TLX1 bound lncRNA (*MSTRG.6968*). PolyA+ and total RNA-seq tracks are depicted for control siRNA transfected samples. Bars represent the MACS2 peaks with FDR <0.05. (C) Hockey stick plot representing the normalized rank and cluster signal of clusters of H3K27ac ChIP-seq peaks. Red dots represent unannotated lncRNAs significantly associated with a super-enhancer. (D) IGV screenshot of an unannotated super-enhancer associated lncRNA (*MSTRG.37538*). PolyA+ and total RNA-seq tracks are depicted for control siRNA transfected samples. Bars represent the MACS2 peaks with FDR <0.05. RNA: ribonucleic acid. lncRNA: long non-coding RNA.

significant enrichment (42 lncRNAs) of TLX1 binding for these TLX1-regulated, super-enhancer associated lncRNAs, as exemplified for lncRNA *NBAT1* (Figure 1D; Online Supplementary Table S1). As super-enhancers are associated with regions of open chromatin, we also performed ATAC-seq and confirmed that 98.95 % of the super-enhancer regions overlap with regions of open chromatin. Moreover, we discovered that 66.4 % of the transcription start sites (TSSs) from highly expressed (top decile) genes had ATAC-seq peaks within +/− 5 kb (Online Supplementary Figure S7). To further explore the functional association of super-enhancers and expressed lncRNAs, the genome-wide transcriptional response of lncRNAs upon JQ1 treatment of ALL-SIL lymphoblasts was investigated, given that this bromodomain and extra-terminal motif (BET) inhibitor causes a decrease in the expression of super-enhancer associated genes (Online Supplementary Figure S8A).[12] Among 115 super-enhancer associated, TLX1-regulated lncRNAs, 41 lncRNAs were differentially expressed upon JQ1 inhibition (Online Supplementary Table S1). Moreover, 26 upregulated and 24 downregulated lncRNAs upon TLX1 knockdown were significantly overlapping with those lncRNAs downregulated upon JQ1 exposure (Online Supplementary Figure S8B).

In addition to previously annotated genes, 2788 lncRNAs that have not been previously annotated in Ensembl, Gencode, LNCipedia and RefSeq were also detected. Of these novel lncRNAs, 82 are differentially regulated upon TLX1 knockdown, of which 30 are directly bound by TLX1, as illustrated for MSTRG.6968 (Figure 2A,B; Online Supplementary Table S4). Of note, MSTRG.37538 is a lncRNA marked with one of the strongest genome-wide super-enhancer sites of all identified unannotated lncRNAs (Figure 2C, D).

In a complementary approach, TLX1 and TLX3 (further denoted as TLX) driven lncRNAs were retrieved from a primary T-ALL patient cohort as TLX1 and TLX3 induce T-ALL in a similar way and are associated with a similar gene expression profile.[13] By using polyA+ RNA-seq data of 60 T-ALL patients (including 17 TLX positive cases) as well as total RNA-seq of 25 T-ALL patients (including 10 TLX positive cases) 442 known and 158 novel TLX subgroup-specific lncRNAs were identified (Figure 3A; Online Supplementary Figure S9 and S10A; Online Supplementary Table S5 and S6). From these, 32 known and 14 novel lncRNAs overlapped significantly with the known and novel set of differentially expressed genes upon TLX1 knockdown, respectively (Figure 3B; Online Supplementary Figure S10B). Moreover, 22 known and three novel TLX subgroup-specific lncRNAs are in the vicinity (max 1 Mb) of a known differentially regulated T-ALL tumor suppressor gene (Online Supplementary Table S5 and S6).[4]

To identify possibly oncogenic TLX subtype-specific lncRNAs, this new data was integrated with our previously generated polyA+ RNA-seq data of OP9-DL1 cultured T-cells,[10] serving as reference material for lncRNA expression levels in untransformed T-cell progenitors. Therefore, lncRNAs that are significantly higher expressed in the TLX subgroup as compared to the other T-ALL subgroups of the primary T-ALL cohort (HOXA, immature, TAL) and significantly higher as compared to normal T-cells were selected. Those lncRNAs that were also differentially expressed among any of the other T-ALL subgroups and T-cells were excluded. In total, 144 TLX-specific, potentially oncogenic lncRNAs were identified (Figure 3C), as illustrated for lncRNA RP11-973H7.4 (Figure 3D, left), located in the immediate vicinity of the well-known T-ALL tumor suppressor gene PTPN2 (Figure 3D, right).

In this study, we present the first comprehensive analysis of the lncRNA transcriptome of TLX1+ ALL-SIL lymphoblasts and TLX subtype primary T-ALLs, uniquely integrating the polyadenylated and non-polyadenylated transcriptome and chromatin features. Our results reveal that TLX1 directly regulates a set of known and novel lncRNAs of which some are marked by super-enhancers. By integrating normal T-cell data and a primary T-ALL patient cohort, we also identified

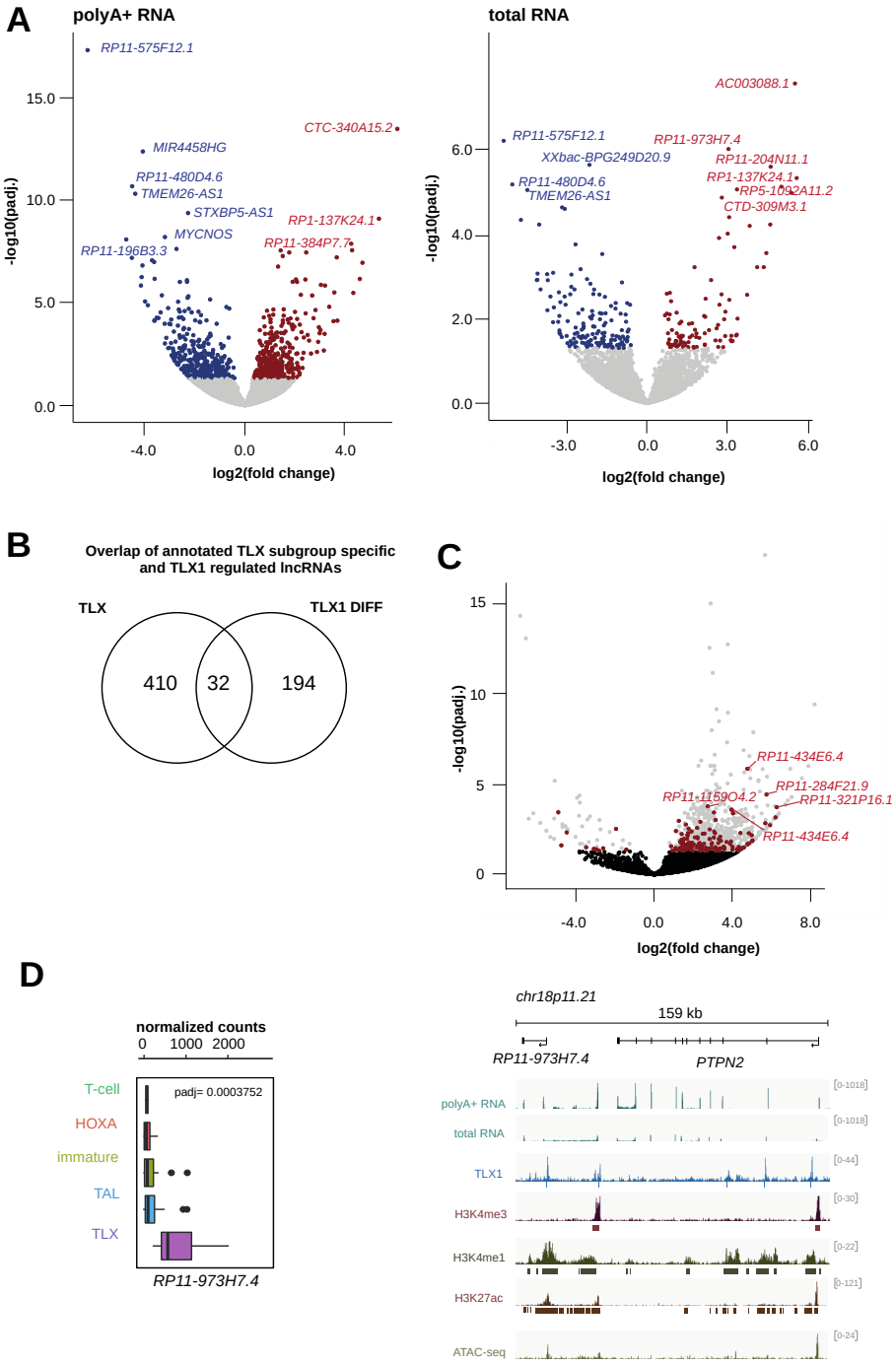


Figure 3: identification of TLX specific, possibly oncogenic lncRNAs in a primary T-ALL cohort. (A) Volcano plot representation of lncRNAs that are significantly higher or lower in the TLX group as compared with T-ALL patients belonging to other T-ALL subtypes (*TALR*, immature, HOXA). Red (upregulated in TLX subtype T-ALLs *versus* other subtypes) and blue (downregulated in TLX subtype T-ALLs *versus* other subtypes) dots represent significantly differentially expressed lncRNAs detected with polyA+ RNA-seq (left) and total RNA-seq (right) (adjusted *P*-value <0.05). LncRNA names depicted in the plots are the top ten differentially regulated lncRNAs. (B) Venn diagram depicting the overlap between significant differentially expressed lncRNAs upon *TLX1* knockdown and TLX subgroup specific lncRNAs (Fisher's exact test, adjusted *P*-value =1.601e-12). (C) Volcano plot representation of lncRNAs that are significantly higher expressed in the TLX group as compared with normal T-cell subsets. Gray and red dots represent significant differentially expressed lncRNAs (adjusted *P*-value <0.05). Red dots are TLX specific lncRNAs not differentially expressed between T-cells and other subgroups. LncRNA names depicted in the plots are the top five differentially regulated lncRNAs. (D) Boxplot and IGV screenshot for lncRNA *RP11-973H7.4*, that is significantly higher expressed in the TLX subgroup compared to the other subgroups and significantly higher expressed as in normal thymocytes. PolyA+ and total RNA-seq tracks are depicted for control siRNA transfected samples. Bars represent the MACS2 peaks with FDR <0.05. RNA: ribonucleic acid; lncRNA: long non-coding RNA.

144 putative TLX-specific oncogenic lncRNAs, which could be further tested for phenotypic effects upon knockdown and explored as new targets for RNA-based therapeutics. LncRNAs may serve as excellent therapeutic targets as these are often expressed in a cell-type-specific manner, offering potential advantages with respect to on-target toxicity as shown by our research group for lncRNA SAMMSON in melanoma.[14] In conclusion, our study delineates a TLX subgroup and TLX1-specific lncRNA network including a subset of super-enhancer associated lncRNAs. Our work, together with that of others, strongly suggest an important role of lncRNAs in T-ALL and warrant further functional investigation.[9,10,15]

REFERENCES

1. Soulier J, Clappier E, Cayuela JM, et al. HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). Blood. 2005;106(1):274–286.
2. Ferrando AA, Neuberg DS, Staunton J, et al. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. Cancer Cell. 2002;1(1), 75–87.
3. Hatano M, Robberts CW, Minden M, et al. Deregulation of a homeobox gene, HOX11, by the t(10;14) in T cell leukemia. Science. 1991;253(5015):79–82.
4. Durinck K, Van Loocke W, Van der Meulen J, et al. Characterization of the genome-wide TLX1 binding profile in T-cell acute lymphoblastic leukemia. Leukemia. 2015;29(12):2317–2327.
5. Djebali S, Davis CA, Merkel A, Gingeras TR. Landscape of transcription in human cells. Nature. 2012;489(7414):101–108.
6. Bhan A, Soleimani M, Mandal SS. Long Noncoding RNA and Cancer: A New Paradigm. Cancer Res. 2017;77(15):3965-3981.
7. Nobili L, Lionetti M, Neri A. Long non-coding RNAs in normal and malignant hematopoiesis. 7(5):.
8. Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458(7235):223–227.
9. Wallaert A, Durinck K, Van Loocke W, et al. Long noncoding RNA signatures define oncogenic subtypes in T-cell acute lymphoblastic leukemia. Leukemia. 2016;30(9):1927–1930.
10. Durinck K, Wallaert A, Van de Walle I, et al. The notch driven long non-coding RNA repertoire in T-cell acute lymphoblastic leukemia. Haematologica. 2014;99(12):1808–1816.
11. Casero D, Sandoval S, Seet CS, et al. LncRNA profiling of human lymphoid progenitors reveals transcriptional divergence of B and T lineages. Nat Immunol. 2016;16(12):1282–1291.
12. Lovén J, Hoke HA, Lin CY, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013;153(2):320–34.
13. Della Gatta G, Palomero T, Perez-Garcia A, et al. Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. Nat Med. 2012;18(3):436–40.
14. Leucci E, Vendramin R, Spinazzi M, et al. Melanoma addiction to the long non-coding RNA SAMMSON. Nature. 2016;531(7595):518–522.
15. Cao Thi Ngoc P, Hao Tan S, King Tan T, et al. Identification of novel lncRNAs regulated by the TAL1 complex in T- cell acute lymphoblastic leukemia. Leukemia 2018 Mar 26. [Epub ahead of print].

SUPPLEMENTARY METHODS

Cell lines

The TLX1 positive cell line ALL-SIL was obtained from the DSMZ cell line repository. Cells were maintained in RPMI-1640 medium (Life Technologies, 52400-025) supplemented with 20 % fetal bovine serum, 1 % of L-glutamine (Life Technologies, 15140-148) and 1 % penicillin/streptomycin (Life Technologies, 15160-047).

siRNA mediated knockdown in ALL-SIL lymphoblasts

The RNA isolated after *TLX1* knockdown that has been used for microarray based gene expression profiling (Agilent SurePrint G3, 8x60k) in the previous study has been used in this study for RNA-seq.[1] In short, ALL-SIL cells were electroporated (250 V, 1000 µF) using a Genepulser Xcell device (Bio-Rad) with 400 nM of Silencer Select Negative Control 1 siRNA (Ambion, #AM4635) or siRNAs targeting TLX1 (Silencer Select, Ambion, Carlsbad, CA, USA; #4392420, s6746 (siRNA1) and s6747 (siRNA2)). ALL-SIL cells were collected 24h post electroporation.

Compound treatment in ALL-SIL lymphoblasts

ALL-SIL cells were seeded at a density of $1x10^6$ cells/ml and treated with either DMSO or 1 µM of JQ1 compound (BPS Bioscience, 27401). Cells were harvested 12h post treatment for RNA isolation.

Clinical samples

Bone marrow lymphoblasts from 60 pediatric and adult T-ALL patients (13 immature, 23 *TALR*, 17 *TLX1/TLX3* and 7 *HOXA*) were collected with informed consent according to the declaration of Helsinki from Saint-Louis Hospital (Paris, France) and the study was approved by the Institut Universitaire d'Hématologie Institutional Review Board. This primary T-ALL cohort has previously been investigated[2] and these RNA samples were used for RNA-seq in this study.

RNA isolation

Total RNA was isolated using the miRNeasy mini kit (Qiagen) with DNA digestion on-column. By means of spectrophotometry, RNA concentrations were measured (Nanodrop 1000, Thermo Scientific).

PolyA+ RNA sequencing

Polyadenylated transcripts of the *TLX1* knockdown samples were sequenced using the TruSeq stranded mRNA sample preparation kit (Illumina). The libraries were quantified using the KAPA library quantification kit (Illumina) and samples were paired-end sequenced on the NextSeq 500 sequencer (Illumina) with a read length of 75 bp. The sequencing depth per sample is shown in **Supplementary Figure 1A (*left*)**. PolyA+ RNA-seq of the 60 primary T-ALL samples was performed using 100 ng of RNA as input material by Biogazelle (Belgium) with the TruSeq stranded mRNA sample preparation kit (Illumina). The samples were quantified using the KAPA library quantification kit and paired-end sequenced on a NextSeq 500 sequencer with a read length of 75 bp. The sequencing depth per sample is shown in **Supplementary Figure 9A (*left*)**. Dll1 data generated in the context of the paper of Durinck et al. were used as a reference dataset (GSE62006).[3]

Total RNA sequencing

Total RNA-seq of the knockdown samples and 25 primary T-ALL samples (5 immature, 5 *TALR*, 10 *TLX1/TLX3* and 5 *HOXA*) was performed using the TruSeq Stranded Total RNA (w/RiboZero Gold) sample prep kit (Illumina), involving depletion of ribosomal (rRNA) transcripts by Biogazelle (Belgium). Libraries were quantified using the Qubit 2.0 Fluorometer and paired-end sequenced on a NextSeq

500 sequencer (Illumina) with a read length of 75 bp. The sequencing depth per sample is shown in **Supplementary Figure 1A (*right*), 9A (*right*).** Total RNA-seq of the JQ1 treated ALL-SIL cells was performed using the TruSeq Stranded Total RNA sample prep kit (Illumina). The libraries were quantified using the KAPA library quantification kit and were single-end sequenced on a NextSeq 500 sequencer with a read length of 75 bp and an average sequencing depth of 38 million reads per sample.

Data processing RNA sequencing

FastQC was used for quality control of fastq files (available online at http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and RSeQC[4] was used for calculating read distribution over genomic features. All samples were aligned against GRCh38 with STAR_2.4.2a[5] using default settings and two pass methods as described in the STAR manual, using GENCODE v25 primary annotation as a guide during the first pass and the combined splice junction database of all samples generated in the first pass as a guide in the second step. Genes were quantified on GENCODE v25 during alignment with STAR. LncRNAs were defined as genes with biotype "lincRNA" or "antisense" (GENCODE v25). StringTie-1.3.3b was used for transcript assembly and transcriptome gtf files were subsequently merged with StringTie merge setting[6,7] to generate a transcriptome containing all detected transcripts over all samples. The merged file was subsequently used to quantify transcripts using HTSeq-0.6.1.[8] Differential gene expression was performed in R based on a negative binomial distribution using DESeq2.[9] Reported adjusted p-values were calculated using Wald statistic. IGV-tools[10] count was used on BAM files to generate a visualisation track for IGV.[11]

(Tumor suppressor) genes in the neighborhood of lncRNAs

BEDTools[12] was used to identify genes within a certain range from a selected lncRNA.

Protein-coding potential calculations

PhyloCSF[13] and the Coding-Potential Assessment Tool[14] (CPAT, version 1.2.2) were used to identify putative protein-coding transcripts in the unannotated genes obtained by RNA-seq. PhyloCSF employs codon substitution frequencies in whole-genome multi-species alignments to distinguish between coding and non-coding loci. Multiple alignments of 45 vertebrate genomes with human (hg19) are obtained from the UCSC website (multiz46way) and processed using the PHAST package (version 1.3) to obtain the required input format for PhyloCSF. To validate our workflow and obtain the optimal threshold for the PhyloCSF score, we benchmarked PhyloCSF with transcripts annotated in RefSeq.[15] Alignments in the BED format of 5859 RefSeq lncRNAs and 6051 RefSeq mRNAs were obtained from the UCSC genome browser website to serve as negative and positive sets, respectively. After ROC analysis, an optimal threshold for the PhyloCSF score of 60.7876 was found. This corresponds to a sensitivity and specificity of 92.75 %. For CPAT, the transcript sequences were provided in the FASTA format. The hexamer frequency table and logit model provided with the algorithm were used. We used the published cutoff of 0.364 as a threshold for the CPAT output. Bed files were converted to hg19 using LiftOver (UCSC).

Guilt-by-association analysis

Normalized counts were generated for the samples of the primary T-ALL cohort (polyA+, n=60). Spearman correlations were calculated for the top 5 TLX1 up and down regulated lncRNAs. The output was used as input for a GSEA pre-ranked analysis[16] using the c5.bp.v6.0 (GO biological processes) gene set. Next, the output was used as input for the Cytoscape plugin 'Enrichment map' to create networks of enriched gene set clusters. Functional gene set clusters that are correlated (blue nodes) and anti-correlated (red nodes) with the lncRNA of interest are depicted in these networks.

Chromatin immunoprecipitation sequencing

Our previously generated TLX1 and H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) data generated in the context of the paper of Durinck et al. were used (GSE62144).[1] H3K4me1 and H3K4me3 chromatin immunoprecipitation sequencing (ChIP-seq) were performed as previously described with minor changes.[17] In brief, $1x10^7$ cells were cross-linked with 1.1 % formaldehyde (Sigma-Aldrich, F1635) at room temperature for 10 min and the cross-linking reaction was quenched with glycine (125 mM final concentration, Sigma-Aldrich, G-8790). Nuclei were isolated and chromatin was purified by chemical lysis. Next, the purified chromatin was fragmented to 200-300 bp fragments by sonication (Covaris, M220, Focused-ultrasonicator). Chromatin immunoprecipitation was performed by incubation of the chromatin fraction overnight with 20 µl of protein-A coated beads (Thermo-Scientific, 53139) and 2 µg of H3K4me1 specific (Abcam, ab8895) or H3K4me3 specific (Abcam, ab8580) antibody. The next day, beads were washed to remove non-specific binding events and enriched chromatin fragments were eluted from the beads, followed by reverse cross-linking by incubation at 65 °C overnight. DNA was subsequently purified by phenol/chloroform extraction, assisted by phase lock gel tubes (5Prime). DNA obtained from the ChIP-assays was adaptor-ligated, amplified and quantified using the KAPA library quantification kit. The libraries were single-end sequenced on a NextSeq 500 sequencer with a read length of 75 bp and an average sequencing depth of 35 million reads.

Assay for transposase accessible chromatin sequencing

Assay for transposase accessible chromatin sequencing (ATAC-seq) was performed as previously described with minor changes.[18] In short, 50,000 cells were lysed and fragmented using Tn5 transposase (Illumina). Next, the samples were purified using the MinElute kit (Qiagen). The transposased DNA fragments were amplified and purified using the PCR Cleanup kit (Qiagen). Samples were quantified using the KAPA library quantification kit and single-end sequenced on a NextSeq 500 with a read length of 75 bp and an average sequencing read depth of 50 million reads.

Data processing ChIP and ATAC sequencing

All ChIP-seq files were aligned using STAR_2.4.2a[5] with --outFilterMultimapNmax 1 to exclude multimapping reads and --alignIntronMax 1 to turn off splice awareness. Peak calling was subsequently performed with MACS2[19] using input samples as control. Peak calling for histone marks was performed using --broad setting. BEDTools[12] was used to find overlaps between ChIP-seq tracks. All plots were generated in R using ggplot2. Any additional data manipulation was performed in R. IGV-tools[10] count was used on BAM files to generate a visualisation track for IGV.[11]

Visualization of distances between transcription start sites and closest ATAC-seq peaks

Based on the distribution of gene expression values from polyA+ RNA-seq performed on ALL-SIL lymphoblasts, we randomly selected 1000 genes belonging to the lowest decile, 1000 genes with expression levels between the 40 % and 60 % percentile and 1000 genes with expression values in the highest decile. Next, we retrieved the transcription start sites (TSSs) of those genes and built three Zipper plots showing the distribution of distances between the TSSs and the closest ATAC-seq peaks in a 5 kb window from the TSS.[20]

Motif analysis

MEME-ChIP[21] was used to find centrally enriched TF motifs in ChIP-seq peaks. To prepare peak files for input, peaks were adjusted to 500 bp centered around the peak summit.

SE calling

Super-enhancer analysis was performed using ROSE[22,23] with H3K27ac alignment file and input as control, excluding TSSs (+-2500 bp). BEDTools[12] was used to assign genes within 100 kb to the super-enhancer regions.

Statistical analyses

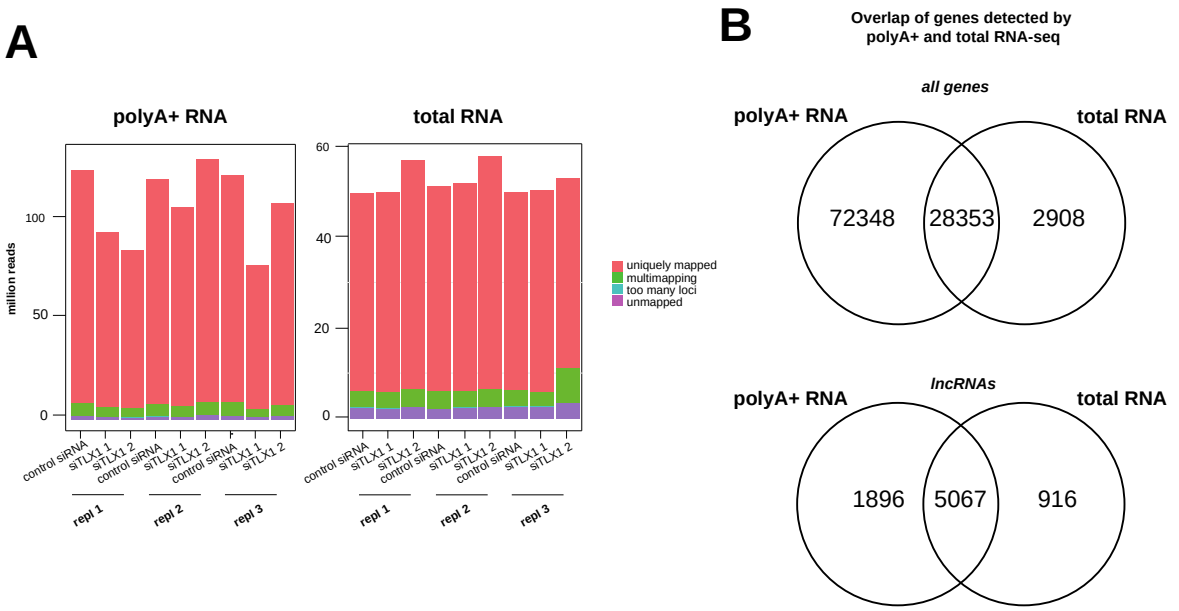
Statistical significance (pAdj.<0.05) of differences between conditions was determined by Fischer exact test using R package.

REFERENCES

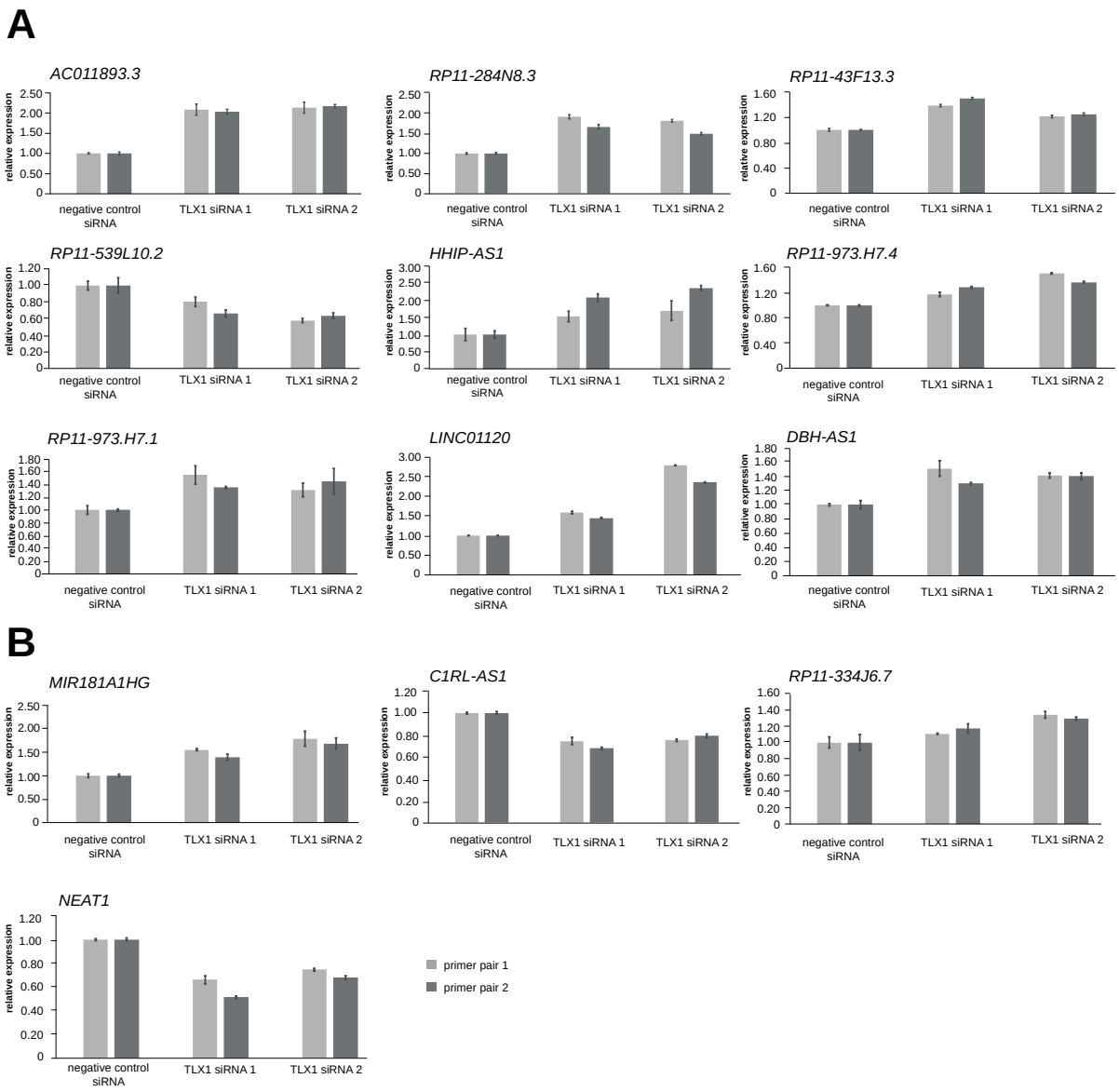
1. Durinck K, Van Loocke W, Van der Meulen J, et al. Characterization of the genome-wide TLX1 binding profile in T-cell acute lymphoblastic leukemia. Leukemia. 2015;29(12):2317–2327.
2. Wallaert A, Durinck K, Van Loocke W, et al. Long noncoding RNA signatures define oncogenic subtypes in T-cell acute lymphoblastic leukemia. Leukemia. 2016;30(9):1927–1930.
3. Durinck K, Wallaert A, Van de Walle I, et al. The notch driven long non-coding RNA repertoire in T-cell acute lymphoblastic leukemia. Haematologica. 2014;99(12):1808–1816.
4. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28(16):2184–2185.
5. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
6. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–295.
7. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11(9):1650–1667.
8. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
9. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
10. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.
11. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
12. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–842.
13. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011;27(13):i275–i282.
14. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41(6):e74.
15. Pruitt KD, Brown GR, Hiatt SM, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42(D1):D756–D763.
16. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
17. Lee TI, Johnstone SE, Young RA. Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc. 2006;1(2):729–48.
18. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current Protocols in Molecular Biology. 2015;119:21.29.1-21.29.9.

19. Zhang Y, Liu T, Meyer CA, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
20. Avila Cobos F, Anckaert J, Volders P-J, et al. Zipper plot: visualizing transcriptional activity of genomic regions. BMC Bioinformatics. 2017;18(1):231.
21. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202-8.
22. Whyte WA, Orlando DA, Hnisz D, et al. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell. 2013;153(2):307–319.
23. Lovén J, Hoke HA, Lin CY, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013;153(2):320–34.

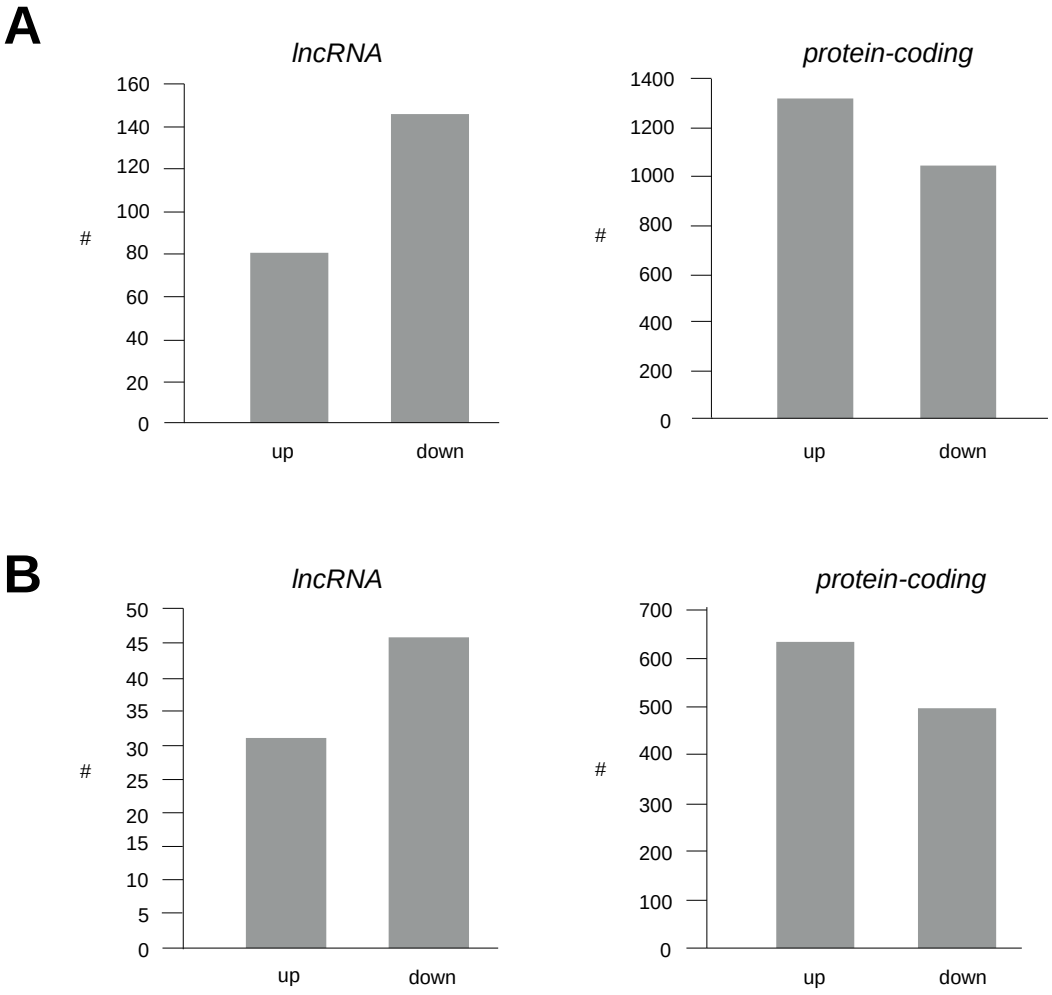
SUPPLEMENTARY FIGURES



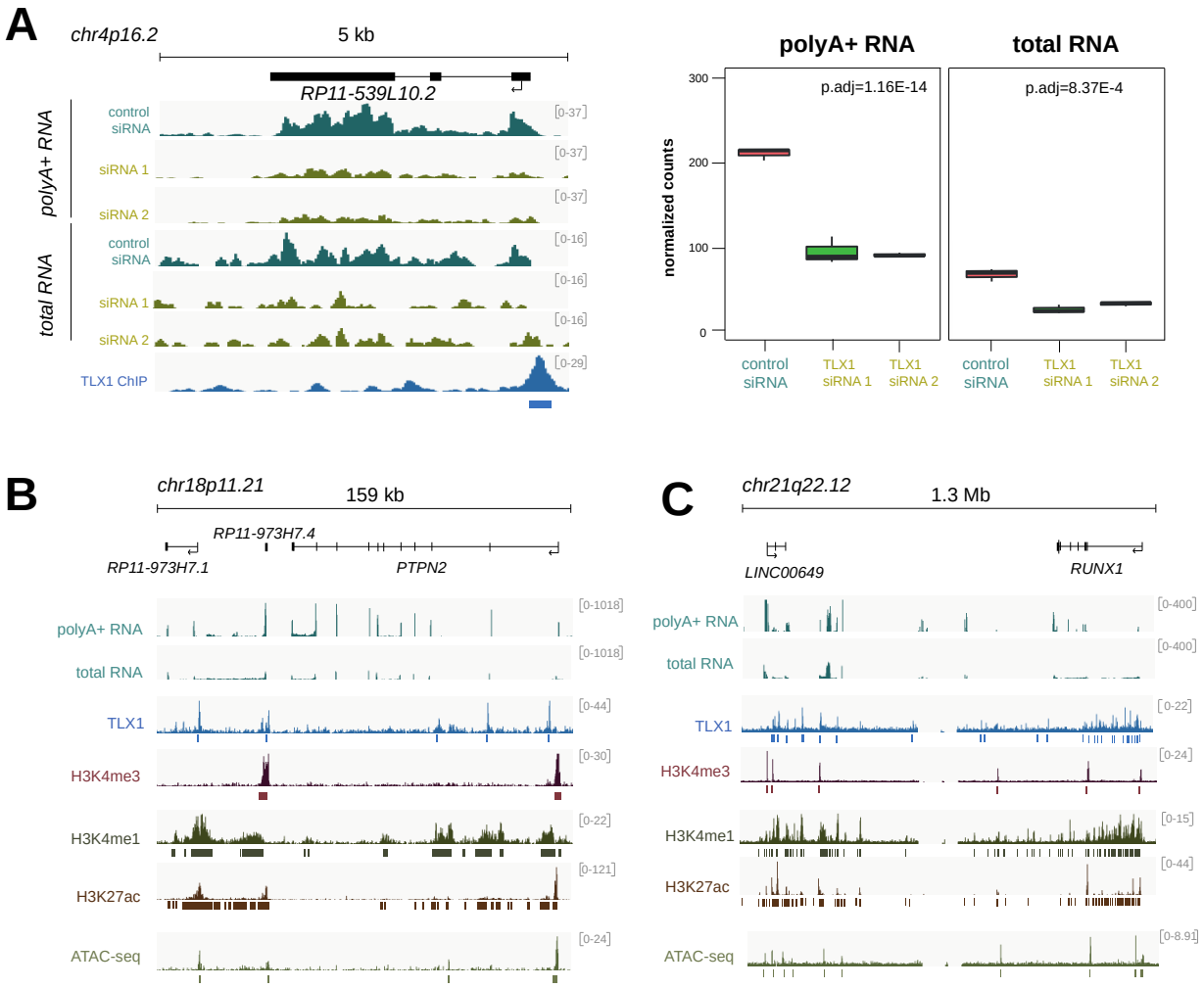
Supplementary Figure 1: qualitative and quantitative analysis of polyA+ and total RNA-seq data of ALL-SIL lymphoblasts with transient *TLX1* knockdown. (A) Number of counts for each sample of polyA+ (left) and total (right) RNA-seq libraries. (B) Venn diagram depicting all genes (upper panel, Fisher exact test, p-value < 2.200e-16) and lncRNAs (lower panel, Fisher exact test, p-value < 2.200e-16) detected by polyA+ and/or total RNA-seq upon *TLX1* knockdown in ALL-SIL cells.



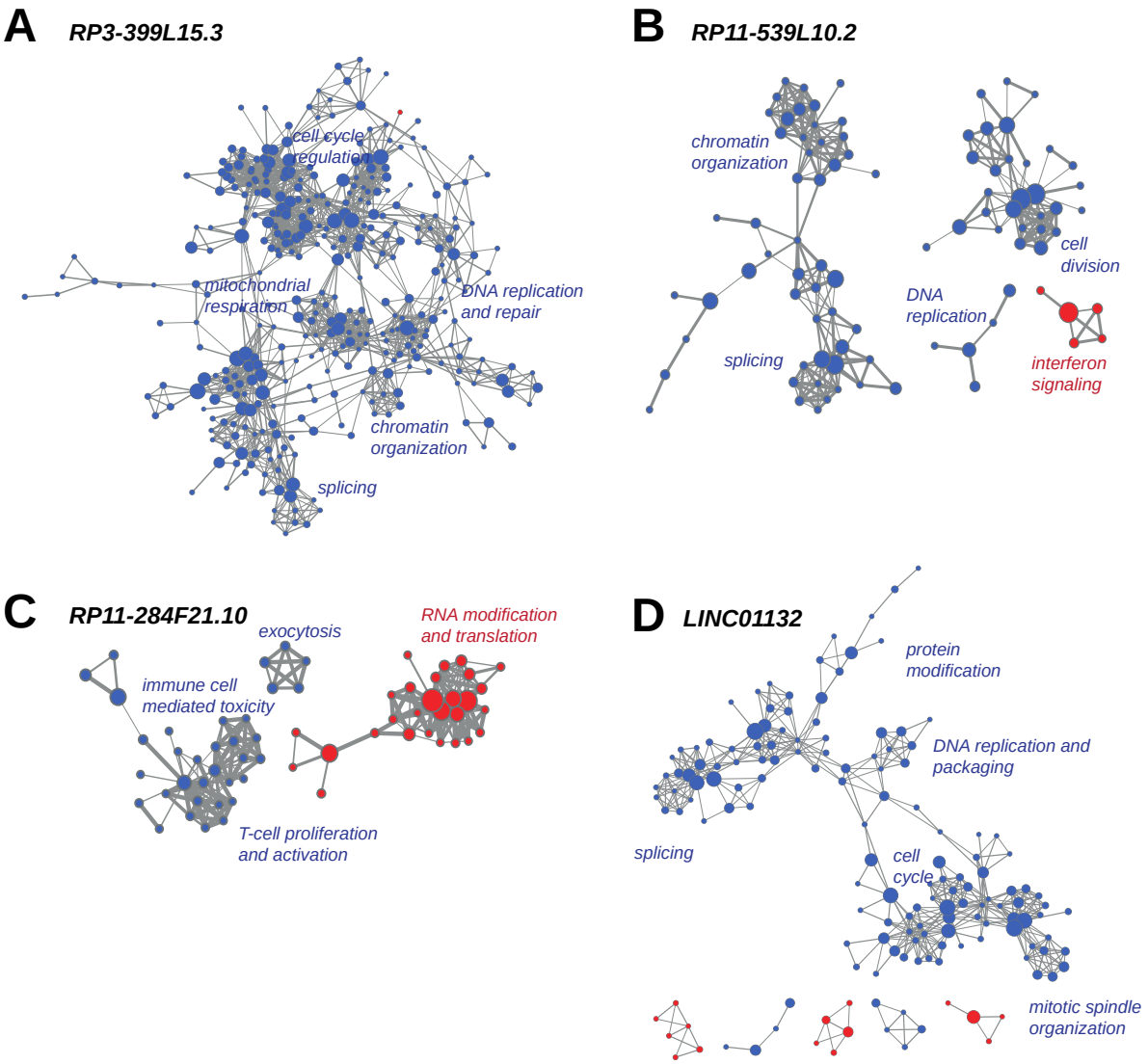
Supplementary Figure 2: RT-qPCR validation of the top 10 significantly differentially expressed genes. (A) Nine of the ten significantly differentially expressed lncRNAs detected by polyA+ could be validated by RT-qPCR on three replicates. No primers could be designed for lncRNA *RP3-399L15.3*. Barplots show a negative control siRNA and two independent TLX1 targeting siRNAs. (B) Nine of the ten significantly differentially expressed lncRNAs detected by total RNA-seq could be validated by RT-qPCR. Only lncRNAs that are not in the top ten differentially expressed lncRNAs detected by polyA+ RNA-seq are shown. Barplots show a negative control siRNA and two independent TLX1 targeting siRNAs. Error bars show the standard error. p-value < 0.05 for all shown lncRNAs (t-test).



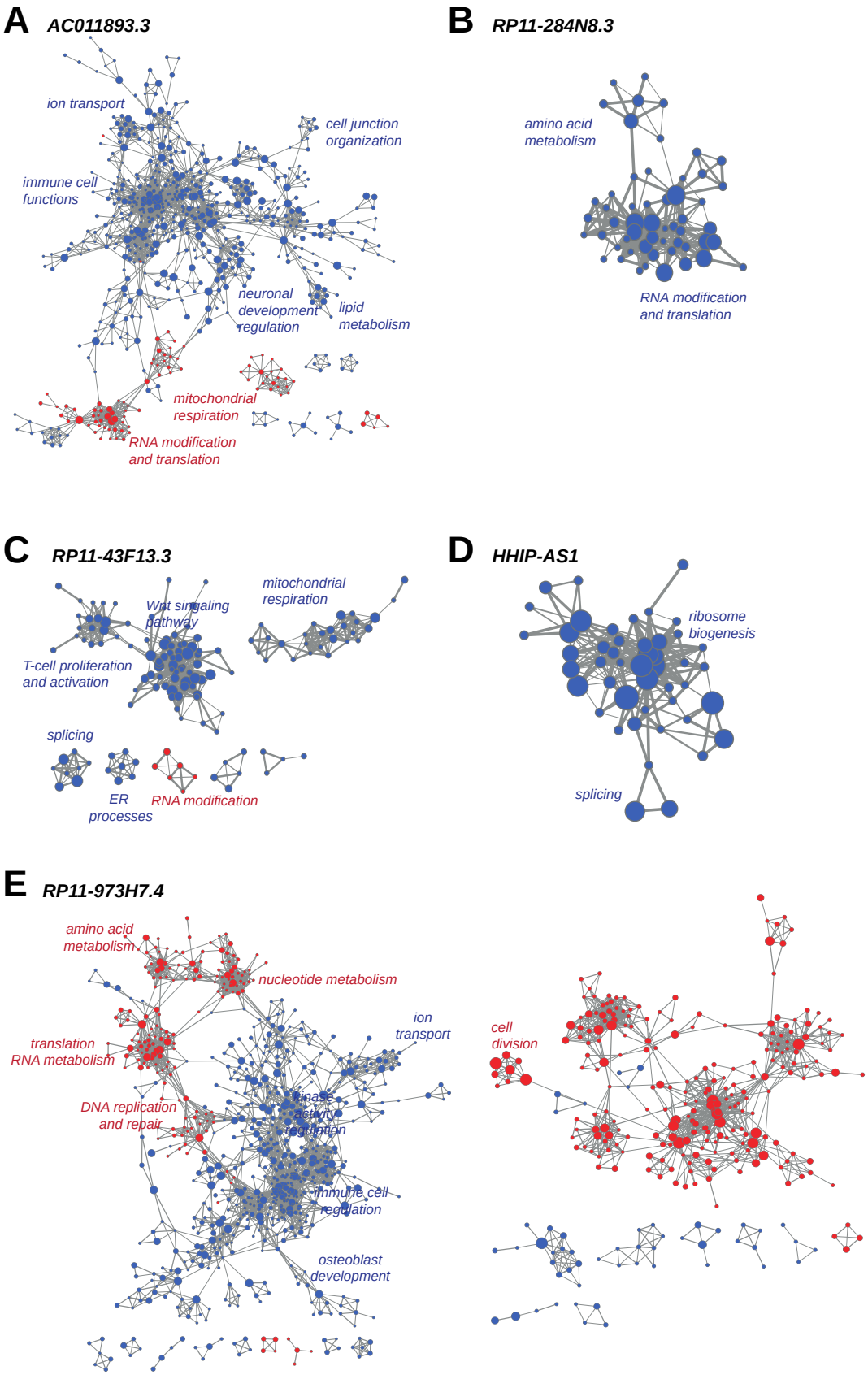
Supplementary Figure 3: TLX1 predominantly acts as an activator of lncRNA expression. (A) Barplot showing the total number of lncRNAs (left panel) and protein coding genes (right panel) differentially expressed upon *TLX1* knockdown without considering direct TLX1 binding as defined by ChIP-seq analysis. (B) Adapted barplot from (A) by only considering direct TLX1 binding events as defined by ChIP-seq analysis.



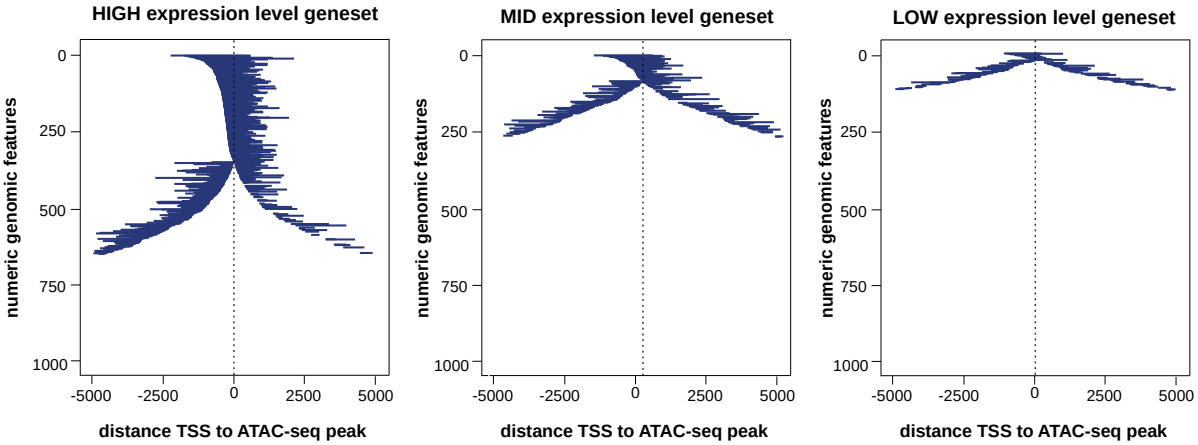
Supplementary Figure 4: TLX1 regulates lncRNAs in the vicinity (max. 1 Mb) of its well established T-ALL tumor suppressor genes. (A) Example of polyA+ RNA-seq, total RNA-seq and TLX1 ChIP-seq signals at the *RP11-539L10.2* locus. Boxplots show the effect of *TLX1* knockdown on the expression of the lncRNA. (B) TLX1 regulated lncRNAs *RP11-973H7.1* and *RP11-973H7.4* are located in the vicinity (max. 1 Mb) of *PTPN2*. (C) TLX1 regulated lncRNA *LINC00649* is located in the vicinity (max. 1 Mb) of *RUNX1*. PolyA+ and total RNA-seq tracks are depicted for control siRNA transfected samples. Bars represent the MACS2 peaks with FDR < 0.05.



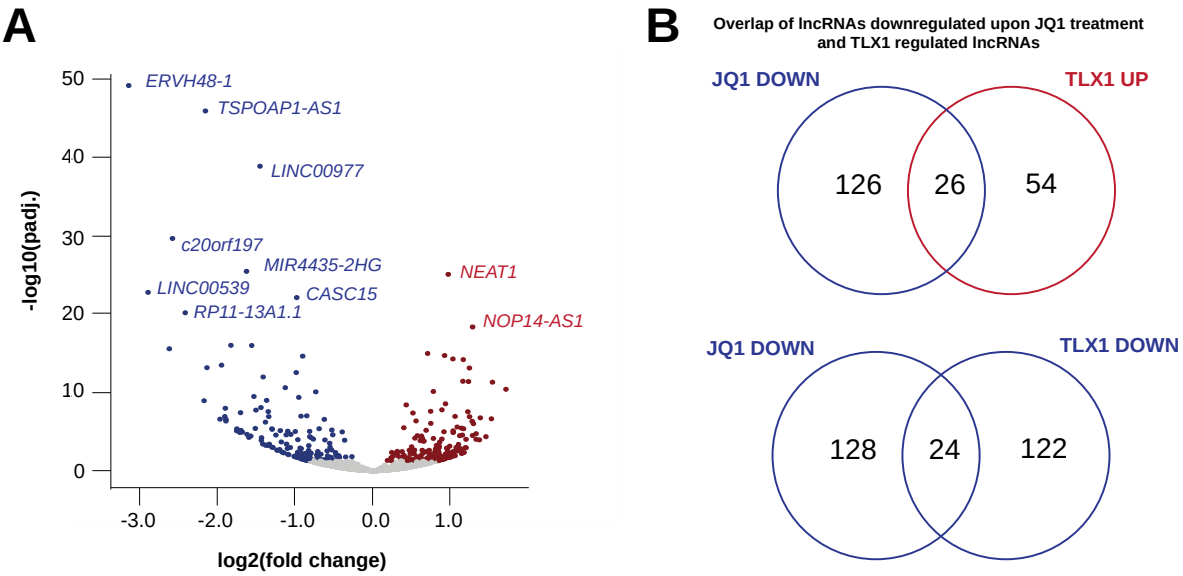
Supplementary Figure 5: functional annotation of the top-5 annotated lncRNA candidates downregulated upon *TLX1* knockdown in ALL-SIL cells through guilt-by-association analysis. (A) *RP3-399L15.3*, (B) *RP11-539L10.2*, (C) *RP11-284F21.10*, (D) *LINC01132*. For *RP11-18H21.1* no functional network could be built based on the primary patient cohort data since this lncRNA was not detected in this dataset. Red dots are positively enriched gene sets, blue dots are negatively enriched gene sets.



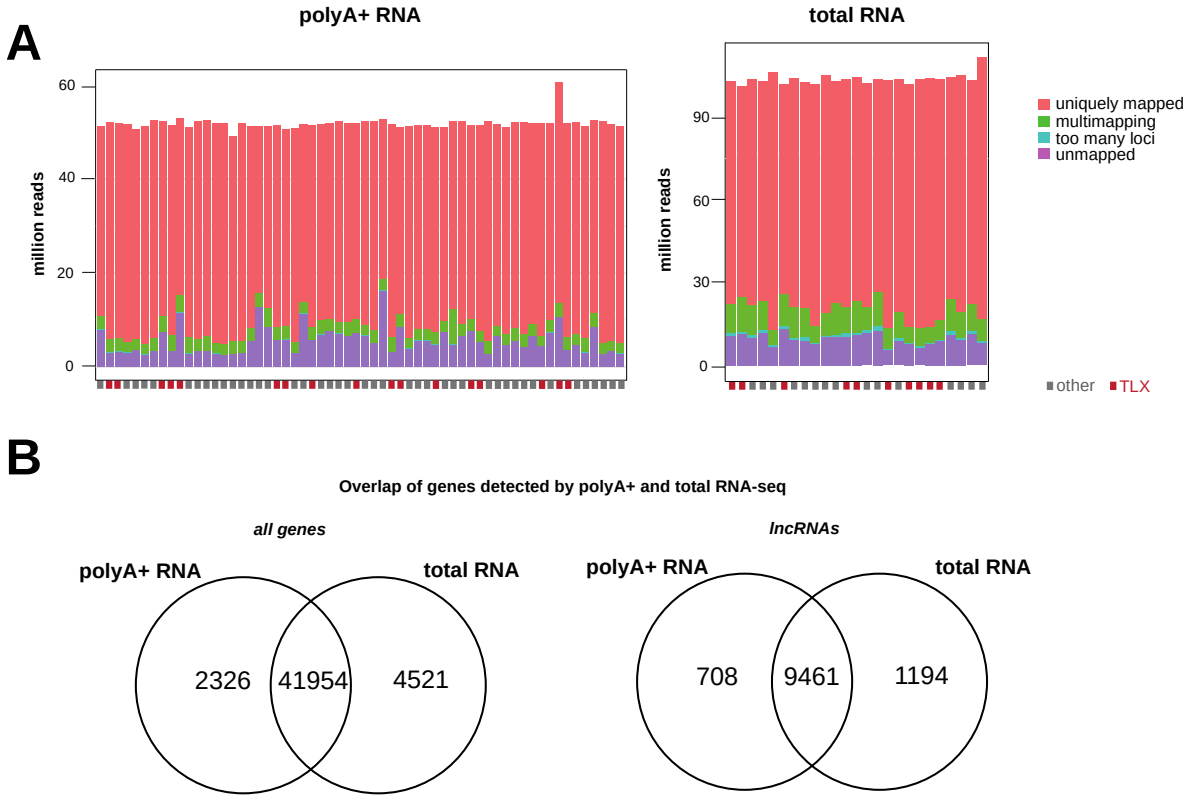
Supplementary Figure 6: functional annotation of the top-5 annotated lncRNA candidates upregulated upon *TLX1* knockdown in ALL-SIL cells through guilt-by-association analysis. (A) *AC011893.3*, (B) *RP11-284N8.3*, (C) *RP11-43F13.3*, (D) *HHIP-AS1*, (E) *RP11-973H7.4*. Red dots are positively enriched gene sets, blue dots are negatively enriched gene sets.



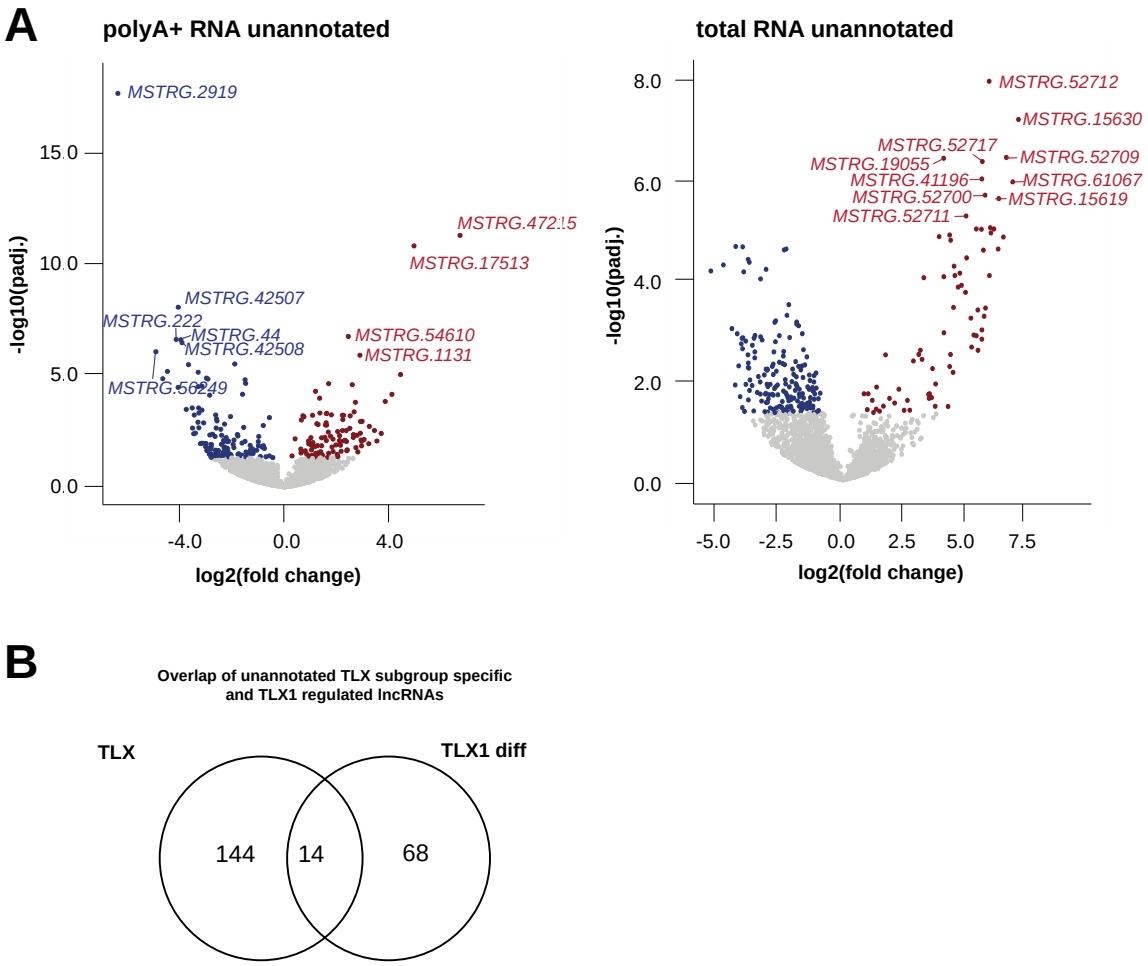
Supplementary Figure 7: open chromatin profiling by means of ATAC-seq. Zipper plots showing distances between TSSs and closest ATAC-seq peaks of 1000 TSSs of (from left to right) low, mid and high level expressed genes. Only ATAC-seq peaks within +/- 5 kb from the TSS are displayed.



Supplementary Figure 8: JQ1 treatment affects TLX1 regulated lncRNA expression. (A) Volcano plot showing significantly differentially expressed lncRNAs upon JQ1 inhibition in ALL-SIL cells. Red (upregulated upon JQ1 treatment) and blue (downregulated upon JQ1 treatment) dots represent significantly differentially expressed lncRNAs (adjusted P-value <0.05) detected with total RNA-seq. LncRNA names depicted in the plots are the top-10 differentially regulated lncRNAs. (B) Venn diagram depicting overlapping lncRNAs significantly downregulated upon JQ1 treatment of ALL-SIL cells with lncRNAs significantly upregulated (upper diagram, Fisher exact test, p-value < 2.200e-16) or downregulated (lower diagram, Fisher exact test, p-value < 2.200e-16) upon *TLX1* knockdown.



Supplementary Figure 9: qualitative and quantitative analysis of polyA+ and total RNA-seq data of a primary T-ALL cohort. (A) Number of counts for each sample of polyA+ (left) and total (right) RNA-seq libraries. (B) Venn diagram depicting all genes (left panel, Fisher exact test, p-value < 2.200e-16) and lncRNAs (right panel, Fisher exact test, p-value < 2.200e-16) detected by polyA+ and/or total RNA-seq in a primary T-ALL cohort.



Supplementary Figure 10: identification of a set of previously unannotated lncRNAs specific for TLX subtype T-ALL. (A) Volcano plot representation of unannotated differentially expressed lncRNAs in TLX subtype T-ALL patients versus those of other T-ALL subgroups. Red (higher in TLX subtype T-ALL versus other) and blue (lower in TLX subtype T-ALL versus other) dots represent significantly differentially expressed lncRNAs detected with polyA+ RNA-seq (left panel) and total RNA-seq (right panel) (adjusted P-value <0.05). Gene names depicted in the plots are the top 10 unannotated differentially regulated lncRNAs. (B) Venn diagram depicting the overlap between significant differentially expressed lncRNAs upon *TLX1* knockdown and TLX subgroup specific lncRNAs (Fisher exact test, p-value = 0.002977).

Paper 2

A comprehensive dataset of TLX1 positive ALL-SIL cells and primary T-cell acute lymphoblastic leukemias

Karen Verboom, Wouter Van Loocke, Emmanuelle Clappier, Jean Soulier, Jo Vandesompele, Frank Speleman and Kaat Durinck

Contribution: I performed the data analysis of all experiments in R starting from the count matrices. I made the figures and wrote the manuscript.

bioRxiv

A comprehensive dataset of TLX1 positive ALL-SIL lymphoblasts and primary T-cell acute lymphoblastic leukemias

Karen Verboom[1,2,3], Wouter Van Loocke[1,2,3], Emmanuelle Clappier[4], Jean Soulier[4], Jo Vandesompele[1,2,3], Frank Speleman[1,2,3] and Kaat Durinck[1,2,3]

[1]*Center for Medical Genetics, Ghent University, Ghent, Belgium*

[2]*Cancer Research Institute Ghent, Ghent, Belgium*

[3]*Department of Biomolecular Medicine, Ghent University, Ghent, Belgium*

[4]*Hôpital Saint Louis, Institut Universitaire d'Hématologie, Paris, France*

Corresponding author: Kaat Durinck, Kaat.Durinck@UGent.be

ABSTRACT

Most currently available transcriptome data of T-cell acute lymphoblastic leukemia (T-ALL) are based on polyA[+] RNA sequencing methods thus lacking non-polyadenylated transcripts. Here, we present the data of polyA[+] and total RNA sequencing in the context of *in vitro TLX1* knockdown in ALL-SIL cells and a primary T-ALL cohort. We extended this dataset with ATAC sequencing and H3K4me1 and H3K4me3 ChIP sequencing data to map putative gene regulatory regions. In this data descriptor, we present a detailed report of how the data were generated and which bioinformatics analyses were performed. Through several technical validations, we showed that our sequencing data are of high quality and that our *in vitro TLX1* knockdown was successful. We also validated the quality of the ATAC and ChIP sequencing data and showed that ATAC and H3K4me3 ChIP peaks are enriched at transcription start sites. We believe that this comprehensive set of sequencing data can be reused by others to further unravel the complex biology of T-ALL in general and TLX1 in particular.

SUMMARY

T-cell acute lymphoblastic leukemia (T-ALL) is a hematological cancer resulting from malignant transformation of normal precursor T-cells. T-ALL accounts for 15 % of the pediatric and 25 % of the adult ALL cases and can be subdivided into four molecular subgroups (TLX, HOXA, immature and TAL-R), each with a unique gene expression profile signature[1,2]. In addition, several other oncogenes (eg *NOTCH1*) and tumor suppressor genes (eg *CDKN2A* and *PTEN*), are involved in the multi-step process of leukemia formation across these subgroups[3,4]. Besides genetic alterations, also epigenetic mechanisms, such as deregulation of enhancers and histone modifications, are disturbed in T-ALL[5]. Several sequencing efforts, including whole genome, exome and transcriptome sequencing, have been performed to identify driver genes in T-ALL. The majority of these studies focused on the identification of mutations, fusion genes and transcriptional changes[6–8]. A major drawback of the current transcriptome studies is that polyA[+] RNA-seq only covers part of the transcriptome and provides no insights into the non-polyadenylated complement that includes biologically relevant transcripts such as circular RNAs (circRNAs) and long non-coding RNAs (lncRNAs)[9,10]. To resolve this issue, we generated polyA[+] as well as total RNA-seq data of 60 (17 TLX, 7 HOXA, 13 immature, 23 TAL-R) and 25 (10 TLX, 5 HOXA, 5 immature, 5 TAL-R) T-ALL patients, respectively. In this dataset, we identified TLX subgroup specific and possibly oncogenic lncRNAs as reviewed by Verboom et al.[11]. As we focused on the TLX subgroup and more specifically on TLX1, we also generated

polyA[+] and total RNA-seq data of *TLX1* knockdown samples generated using two independent siRNAs in triplicate. In addition, we performed assay for transposase accessible chromatin sequencing (ATAC-seq) as well as H3K4me3 and H3Kme1 chromatin immunoprecipitation sequencing (ChIP-seq) and used the publicly available TLX1 ChIP data (GSE62264) to identify regions directly regulated by TLX1 in TLX1 positive ALL-SIL lymphoblasts. This is a unique dataset in the T-ALL field as it combines for the first time polyA[+] and total RNA-seq of an *in vitro* model system as well as a large primary T-ALL cohort. Moreover, this dataset was extended with ChIP-seq and ATAC-seq to define gene regulatory regions. Here, we provide a detailed description of the methods and bioinformatics analyses used to generate the data to facilitate data-repurposing by other researchers. Moreover, we demonstrate that our data is of high quality. The T-ALL cohort has also been annotated according to their genetic subgroup. In our related publication, we used the dataset to identify TLX1 regulated lncRNAs and TLX subgroup specific lncRNAs[11]. Therefore, this dataset can be reused to characterize other subgroup specific or TLX1 regulated biotypes or

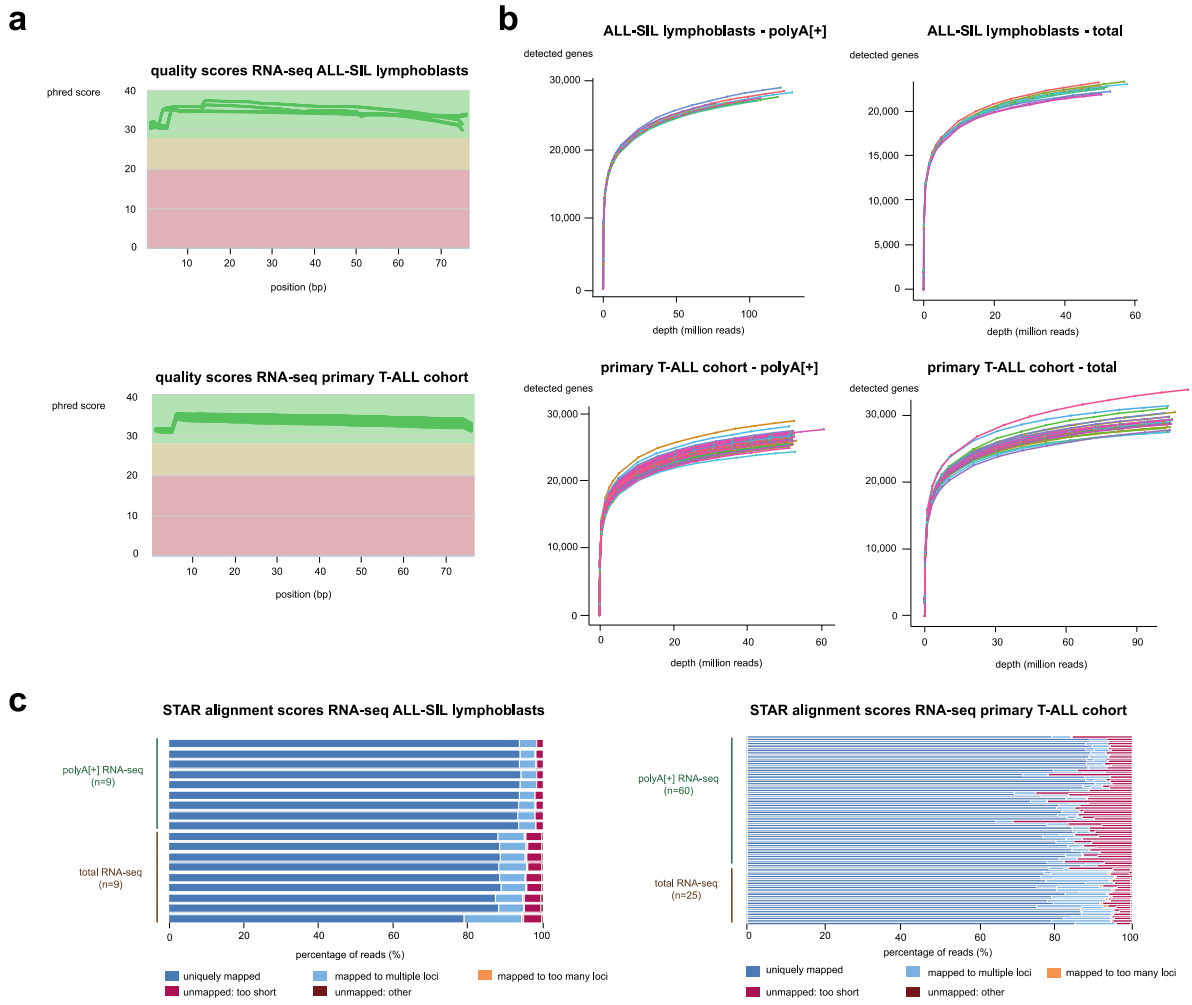


Figure 1: quality assessment of polyA[+] and total RNA-seq data of the *TLX1* knockdown and primary T-ALL cohort samples. (a) Average quality score per base position per sample generated by combining FastQC and MultiQC for the *TLX1* knockdown samples (upper panel) and the primary T-ALL cohort (lower panel). Each line represents a sample. Scores greater than 30 (green region) indicate a good quality. (b) Sequencing saturation plots showing the number of genes detected at a given sequencing depth. (c) Barplots showing the STAR alignment scores for the *TLX1* knockdown samples (left panel) and the primary T-ALL cohort (right panel) generated with MultiQC.

to study other subgroups in the primary T-ALL cohort as this has only been done using microarray data[12]. Besides gene expression analysis, this RNA-seq dataset can also be used

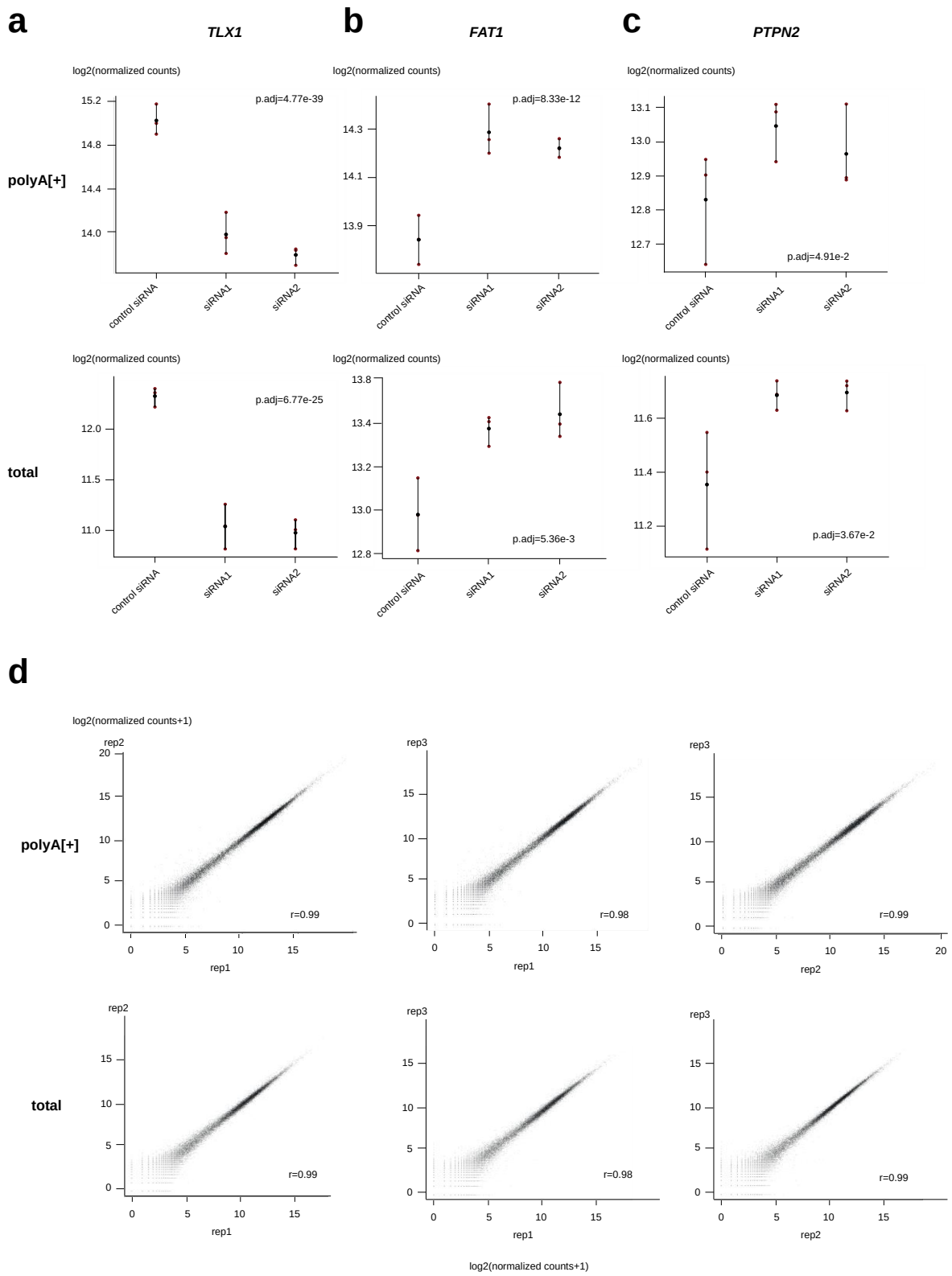


Figure 2: validation of *TLX1* knockdown in ALL-SIL lymphoblasts. (a) *TLX1* is significantly downregulated in the *TLX1* knockdown samples. Tumor suppressor genes *FAT1* (b) and *PTPN2* (c), known to be repressed by TLX1, are upregulated upon *TLX1* knockdown. (d) correlation plots for the replicates of siRNA 1. (d) Correlation plots for the three replicates of TLX1 siRNA1 for polyA[+] (upper panel) and total RNA-seq (lower panel). Pearson correlation coefficients are shown.

to identify gene mutations and gene fusions[6]. Furthermore, integrating the ChIP-seq and ATAC-seq data, together with GRO-seq data can be useful to study enhancer RNAs in TLX1 positive T-ALL. In summary, this data descriptor provides detailed information about the methods used to generate and analyze ChIP-seq and ATAC-seq data in the TLX1 positive ALL-SIL cell line as well as transcriptome data through polyA[+] and total RNA-seq of a *TLX1* knockdown *in vitro* model system and a primary T-ALL cohort. As this is a rich dataset including transcriptome as well as epigenome data, we believe that this dataset is of great value to the research community and will aid, amongst others, to further unravel the complex biology of T-ALL.

DATA DESCRIPTION

Validation of RNA sequencing data

To validate the quality of our RNA-seq data, FastQC analyses were performed on the fastq files. The mean quality scores for the polyA[+] and total RNA-seq data of the *TLX1* knockdown samples and the primary T-ALL cohort are shown in Figure 1a. These plots show the mean quality score across each base position per sample and indicate by a color scale if the quality is good (green), reasonable (orange) or bad (red). As the mean quality score per base position is located in the “good quality region” for each sample, we confirm that our sequencing data is of high quality. Of note, a small decrease in the quality towards the end of a read is typically observed, which is inherent to the Illumina sequencing by synthesis procedure [13]. As these samples have a high sequencing coverage, further increasing the number of reads would only give a small increase in the number of detectable genes (Fig. 1b). A high percentage of the reads were uniquely mapped to the Hg38 human reference genome, confirming a good mapping quality (Supplementary Table 1-2, Fig. 1c). After mapping, we detected an average of 26,412 genes over all samples.

Validation of *TLX1* knockdown in ALL-SIL lymphoblasts

Knockdown of *TLX1* has been validated as *TLX1* is significantly lower expressed in the knockdown samples compared to the control samples in the three replicates of polyA[+] (66 %, 58 % and 65 % knockdown) and total RNA-seq (56 %, 48 % and 62 % knockdown) data (Fig. 2a). In addition, we could demonstrate that the well-known tumor suppressor genes *FAT1* and *PTPN2*, shown to be repressed by TLX1 by microarray data[14], are significantly upregulated upon *TLX1* knockdown according to polyA[+] and total RNA-seq data (Fig. 2 b-c). Furthermore, correlation analysis shows that the replicates are concordant, as shown for a representative example (Fig. 2d).

Validation of ATAC sequencing data

Before sequencing, we validated the quality of the ATAC-seq library by inspecting Fragment Analyzer profiles. As expected for ATAC-seq libraries, the profile first showed a high peak around 180 nt containing the nucleosomal free DNA fragments, followed by two peaks of fragments containing 1 or 2 nucleosomes, respectively (Fig. 3a). After sequencing, the quality of the raw sequencing data was validated by performing FastQC. The average base quality score fell in the “good quality region”, confirming that we generated high quality ATAC-seq data (Fig. 3b). Raw sequencing reads were aligned against the Hg38 human reference genome resulting in 81.5 million mapped reads. 76.27 % of the reads mapped uniquely to the human reference genome and were used for subsequent analyses (Fig. 3c, Supplementary Table 2). In addition, we also confirmed that there is an enrichment of ATAC peaks around transcription start sites, known to contain open chromatin (Fig. 3d).

Validation of ChIP-seq data

The quality of the raw sequencing data was determined using FastQC analyses and we could show that the average base quality score fell in the “good quality region” for each ChIP-seq sample (Fig. 4a). On average, 32.9 million reads

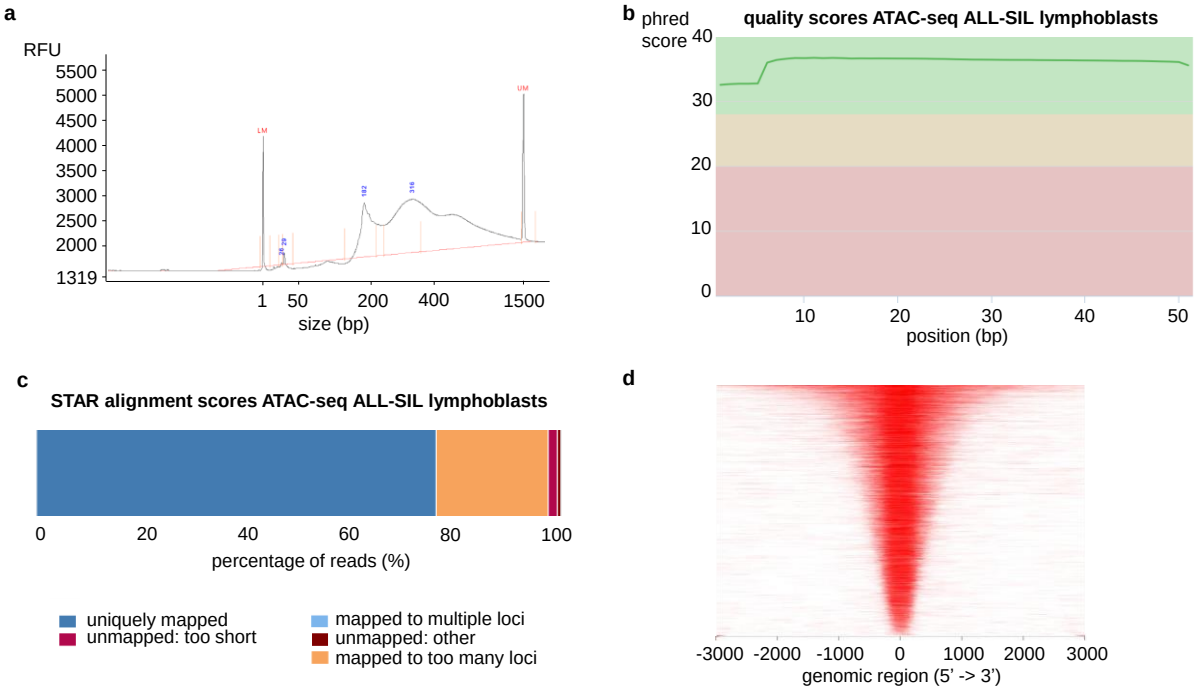


Figure 3: quality assessment of ATAC-seq in ALL-SIL lymphoblasts. (a) Fragment analyzer profile showing a high nucleosome-free peak followed by a mono-nucleosome and di-nucleosome peak. (b) Average quality score per base position generated by combining FastQC and MultiQC. Scores greater than 30 (green region) indicate a good quality. (c) Barplot showing the STAR alignment scores. (d) Enrichment of ATAC peaks around transcription start sites.

per sample were aligned to the Hg38 human reference genome. 69.79 %, 89.49 % and 89.85 % of the reads were uniquely aligned to the Hg38 human reference genome for the input, H3K4me1 and H3K4me3 sample, respectively (Fig. 4b, Supplementary Table 2). As H3K4me3 is a marker of promotor regions, we showed that H3K4me3 is enriched on the promotor of *EEF2*, one of the highest expressed genes in parental ALL-SIL lymphoblasts (Fig. 4c).

In conclusion, this data descriptor shows that we have generated high quality RNA-seq, ChIP-seq and ATAC-seq data that can be reused by other users to further unravel the complex biology of T-ALL.

METHODS

Cell lines

The TLX1 positive cell line ALL-SIL was obtained from the DSMZ cell line repository (ACC 511). Cells were maintained in RPMI-1640 medium (Life Technologies, 52400-025) supplemented with 20 % fetal bovine serum (PAN Biotech, P30-3306), 1 % of L-glutamine (Life Technologies, 15140-148) and 1 % penicillin/streptomycin (Life Technologies, 15160-047) at 37 °C in a 5 % CO_2 atmosphere. Short tandem repeat genotyping was used to validate cell line authenticity prior to performing the described experiments and mycoplasma testing is done on a monthly basis in our laboratory using the MycoAlert Mycoplasma Detection Kit (Lonza, T07-318), according to manufacturer's instructions.

siRNA mediated knockdown in ALL-SIL lymphoblasts

The RNA isolated upon *TLX1* knockdown that has been used for microarray based gene expression profiling in a previous study has been used in this study for RNA-seq[14]. In short, 400 nM of Silencer Select Negative Control 1 siRNA (Ambion, #AM4635) or siRNAs targeting TLX1 (Ambion, Carlsbad, #4392420, s6746 (siRNA1) and s6747 (siRNA2)) was added to 8 million ALL-SIL cells in a total volume of 400 µl. The samples were electroporated (250 V, 1000 µF) in electroporation cuvettes (Bio-Rad, 1652086)

using a Genepulser Xcell device (Bio-Rad). ALL-SIL cells were collected 24h post electroporation. RNA concentrations were measured (Nanodrop 1000, Thermo Scientific). RNA quality scores

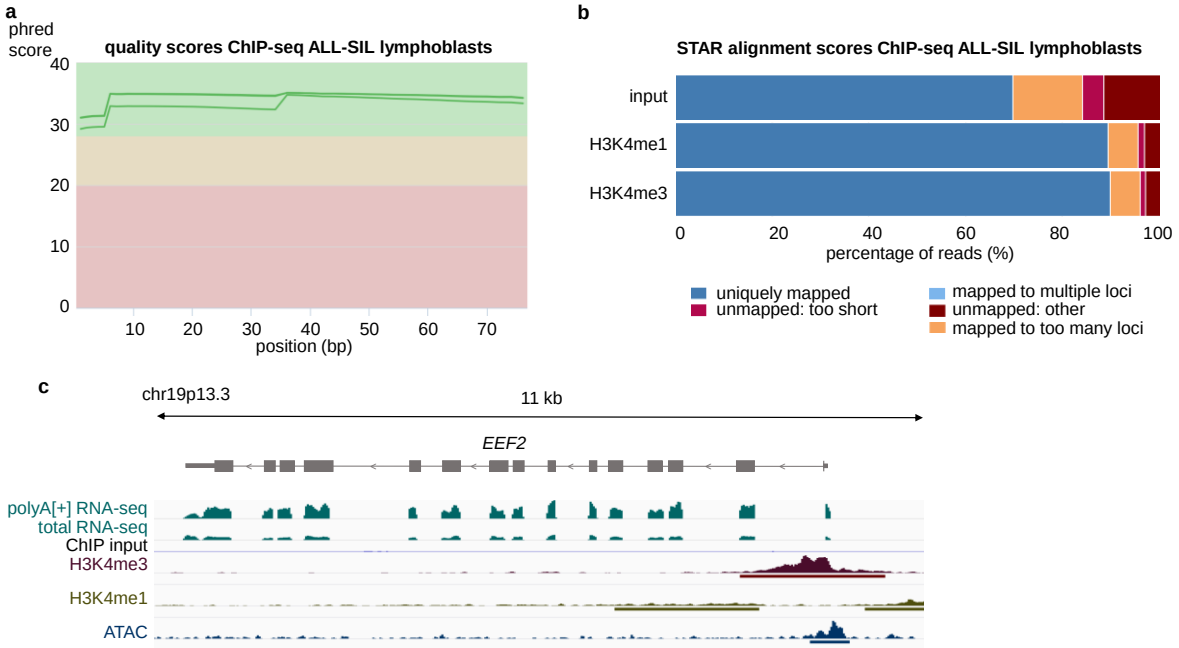


Figure 4: quality assessment of H3K4me1 and H3K4me3 ChIP-seq in ALL-SIL lymphoblasts. (a) Average quality score per base position generated by combining FastQC and multiQC. Scores greater than 30 (green region) indicate a good quality. (b) Barplots showing the STAR alignment scores per sample. (c) IGV screenshot showing H3K4me3 peaks on the promotor of *EEF2,* one of the highest expressed genes in ALL-SIL lymphoblasts. PolyA[+] and total RNA-seq tracks are depicted for control siRNA transfected samples. H3K4me3 bar represent the MACS2 peak with FDR < 0.05.

Clinical samples

Bone marrow lymphoblast and blood samples from 50 pediatric and 10 adult (> 18 year) T-ALL patients (13 immature, 23 TAL-R, 17 TLX1/TLX3 and 7 HOXA) were collected with informed consent according to the declaration of Helsinki from Saint-Louis Hospital (Paris, France) and the study was approved by the Institut Universitaire d'Hématologie Institutional Review Board and the Ghent University Hospital (approval number B670201627319). Subgroup annotation is given in Supplementary Table 1. White blood cells from the patients samples were isolated by Ficoll centrifugation and cryopreserved using standard procedures. After thawing, cells were spun down and RNA was extracted[15].

RNA isolation

Total RNA was isolated using the miRNeasy mini kit (Qiagen, 217084) with DNA digestion on-column according to the manufacturer's instructions. By means of spectrophotometry, (RNA integrity number (RIN)) of the RNA of the cell line experiments were high (RIN: 9-10). The RIN score of the RNA of the primary samples were determined using a bioanalyzer and are shown in Supplementary Table 1. All bioanalyzer profiles suggest RNA of good quality.

PolyA+ RNA sequencing

Library prep of the polyA[+] transcripts of the *TLX1* knockdown samples was performed using the TruSeq stranded mRNA sample preparation kit (Illumina, RS-122-2101) according to manufacturer's instructions. The libraries were quantified using the KAPA library quantification kit (Roche, KK4854) and samples were paired-end sequenced on the NextSeq 500 sequencer (Illumina) with a read length of 2 x 75 bp. PolyA[+] RNA-seq of the 60 primary T-ALL samples was performed using 100 ng of RNA as input material by Biogazelle (Belgium) with the TruSeq stranded mRNA sample preparation kit according to manufacturer's instructions. The

samples were quantified using the KAPA library quantification kit and paired-end sequenced on a NextSeq 500 sequencer with a read length of 2 x 75 bp. The number of uniquely mapped reads are shown in Supplementary Table 1-2.

Total RNA sequencing

Total RNA-seq of the *TLX1* knockdown samples and 25 primary T-ALL samples (5 immature, 5 *TAL-R*, 10 *TLX1/TLX3* and 5 *HOXA*) was performed using the TruSeq Stranded Total RNA (w/RiboZero Gold) sample prep kit (Illumina, RS-122-2301), involving depletion of ribosomal (rRNA) transcripts by Biogazelle according to manufacturer's instructions. Libraries were quantified using the Qubit 2.0 Fluorometer (Thermo Fischer Scientific) and paired-end sequenced on a NextSeq 500 sequencer with a read length of 2 x 75 bp. The number of uniquely mapped reads are shown in Supplementary Supplementary 1-2.

Data processing RNA sequencing

FastQC v0.11.3 was used for quality control of fastq files (available online at http://www.bioinformatics.babraham.ac.uk/projects/fastqc). All samples were aligned against the human reference genome (GRCh38) with STAR_2.4.2a[16] using default settings and two pass methods as described in the STAR manual. GENCODE v25 primary annotation was used as a guide during the first pass and the combined splice junction database of all samples generated in the first pass as a guide in the second step. Genes were quantified on GENCODE v25 during alignment with STAR. Raw data files have been deposited in the NCBI Gene Expression Omnibus (GEO) database. MultiQC v0.9 was used to aggregate the FastQC and STAR results of all samples. Sample IDs and subgroups are shown in Supplementary Table 1. TDF files were loaded in IGV v2.3.98 to visualize the data.

Assay for transposase accessible chromatin sequencing

Assay for transposase accessible chromatin sequencing (ATAC-seq) was performed as previously described with minor changes [17]. In short, 50,000 cells were lysed and fragmented using Tn5 transposase (Illumina, FC-121-1030). Next, the samples were purified using the MinElute kit (Qiagen, 28204). Ten µl transposased DNA fragments were amplified by adding 10 µl H_2O, 2.5 µl of each forward (5'AATGATACGGCGACCACCGAGATCTACACTCGT CGGCAGCGTCAGATGTG3') and reverse (5'CAAGCAGAAGACGGCATACGAGATAAAATGGT CTCGTGGGCTCGGAGATGT3') primer (25 µM) and 25 µl NEBNext High-Fidelity 2x PCR master mix (Bioké, M0541) using the following PCR program: 5 min at 72 °C, 30 sec at 98 °C and 5 cycles of 10 sec at 98 °C, 30 sec at 63 °C and 1 min at 72 °C. To reduce GC-content and size bias, qPCR was performed to determine the exact number of PCR cycles to prevent saturation. 5 µl DNA, 3 µl H_2O, 1 µl of each forward and reverse primer (5 uM) and 10 µl SYBR green (Roche, 4707516001) were used for qPCR (30 sec at 98 °C and 19 cycles of 10 sec at 98 °C, 30 sec at 63 °C and 1 min at 72 °C). The number of extra PCR cycles was calculated as the number of qPCR cycles that correspond to ¼ of the maximum fluorescent intensity. Ten extra PCR cycles were performed on the 45 µl transposed DNA fragments. Samples were purified using the PCR Cleanup kit (Qiagen, 28104). Sample quality was determined using Fragment Analyzer (Advanced Analytical) and samples were quantified using the KAPA library quantification kit. Samples were single-end sequenced on a NextSeq 500 with a read length of 75 bp and an average sequencing read depth of 81.5 million reads.

Chromatin immunoprecipitation sequencing

H3K4me1 and H3K4me3 chromatin immunoprecipitation sequencing (ChIP-seq) were performed as previously described with minor changes [18]. In brief, $1x10^7$ cells were cross-linked with 1.1 % formaldehyde (Sigma-Aldrich, F1635) at room temperature for 10 min and the cross-linking reaction was quenched with an access of glycine (125 mM final concentration, Sigma-Aldrich, G-8790). Nuclei were isolated and chromatin was purified by chemical lysis. Next, the purified chromatin was fragmented to 200-300 bp fragments by

sonication (Covaris, M220, Focused-ultrasonicator). Chromatin immunoprecipitation was performed by incubation of the chromatin fraction overnight with 20 µl of protein-A coated beads (Thermo-Scientific, 53139) and 2 µg of H3K4me1 specific (Abcam, ab8895) or 2 µg of H3K4me3 specific (Abcam, ab8580) antibody. The next day, beads were washed to remove non-specific binding events and enriched chromatin fragments were eluted from the beads, followed by reverse cross-linking by incubation at 65 °C overnight. DNA was subsequently purified by phenol/chloroform extraction, assisted by phase lock gel tubes (5Prime, 733-2478). DNA obtained from the ChIP-assays was adaptor-ligated and amplified using the NEBNext multiplex oligo kit (New England BioLabs, E7335S) according to manufacturer's instructions. Libraries were quantified using the KAPA library quantification kit. The input, H3K4me1 and H3K4me3 libraries were single-end sequenced on a NextSeq 500 sequencer with a read length of 75 bp and an average sequencing depth of 32.9 million reads.

Data processing ChIP-seq and ATAC-seq

FastQC was used for quality control of fastq files. All ChIP-seq and ATAC-seq files were aligned against the human reference genome (GRCh38) using STAR_2.4.2a[16] with --outFilterMultimapNmax 1 to exclude multimapping reads and --alignIntronMax 1 to turn off splice awareness. MultiQC was used to aggregate the FastQC and STAR results of all samples. Raw data files have been deposited in the NCBI Gene Expression Omnibus (GEO) database . To detect enriched regions (ChIP-seq) and regions of open chromatin (ATAC-seq), peak calling has been performed with MACS2 v2.1.20150731 using input samples as control for ChIP-seq [19]. Peak calling for histone marks was performed using --broad setting. Generated files show peak location, enrichment of the peak and score of the identified peaks. TDF and bed files were loaded in IGV v2.3.98 to visualize the data. ChIPseeker v1.16.1 was used to identify enrichment of ATAC-seq peaks around the transcription start sites [20].

USER NOTES

The raw data and gene count tables of the polyA[+] and total RNA-seq data can be downloaded from the GEO database via accession numbers GSE110632 (polyA[+]) and GSE110635 (total) for the *TLX1* knockdown samples and via GSE110633 (polyA[+]) and GSE110636 (total) for the primary T-ALL cohort. The sample IDs and subgroup annotation can be found in Supplementary Table 1. The raw data and peak tables of the ATAC-seq and ChIP-seq can also be downloaded from the GEO database via accession numbers GSE110631 (ATAC-seq) and GSE110630 (ChIP-seq). The analyses can be repeated using the specifications described in the method section starting from the raw sequencing data. By providing the raw sequencing data, users can also use their own pipeline to analyse the data. Other pipelines, such as the bcbio-nextgen pipeline (https://github.com/bcbio/bcbio-nextgen), can also be used to analyze the ChIP-seq and ATAC-seq data. Matching TLX1 and H3K27ac ChIP-seq data in the ALL-SIL cell line, generated by Durinck et al., can be integrated with the datasets described in this paper and are available through GEO (GSE70734, GSE62264)[14]. Downstream differential analysis can be easily performed using the provided count tables and sample annotation (Supplementary Table 1-2) as input for Deseq2 [21] (as described by Verboom et al. [11]) or other packages as Limma [22] or EdgeR [23].

DATASETS

1. polyA+ RNA-seq in ALL-SIL upon TLX1 knockdown. *Gene Expression Omnibus,* http://identifiers.org/geo:GSE110632 (2018).
2. total RNA-seq in ALL-SIL upon TLX1 knockdown. *Gene Expression Omnibus,* http://identifiers.org/geo:GSE110635 (2018).

3. polyA+ RNA-seq in a primary T-ALL cohort. *Gene Expression Omnibus,* http://identifiers.org/geo:GSE110633 (2018).
4. total RNA-seq in a primary T-ALL cohort. *Gene Expression Omnibus,* http://identifiers.org/geo:GSE110636 (2018).
5. ATAC-seq of ALL-SIL cells. *Gene Expression Omnibus,* http://identifiers.org/geo:GSE110630 (2018).
6. H3K4me1 CHIP-seq in ALL-SIL. *Gene Expression Omnibus,* http://identifiers.org/geo:GSE110631 (2018).

AUTHOR CONTRIBUTIONS

Karen Verboom performed laboratory experiments and data analysis, was involved in the design of the experiments, coordination of the research, generation of the figures and writing of the manuscript. Wouter Van Loocke assisted with the data analysis of RNA-seq, ChIP-seq and ATAC-seq. Jean Soulier and Emmanuelle Clappier collected the primary T-ALL samples. Jo Vandesompele coordinated the research, designed experiments and helped writing the paper. Frank Speleman coordinated the research, designed experiments and helped writing the paper. Kaat Durinck coordinated the laboratory experiments, was involved in the design of the experiments and coordination of the research and helped writing the paper.

All authors approved the final version of the manuscript.

FUNDING

The authors want to thank the following funding agencies: Kinderkankerfonds, the Fund for Scientific Research Flanders (FWO Vlaanderen, research project G051416N; postdoctoral grant to KD), BOF (PhD grant to KV), Ghent University (GOA grant BOF16/GOA/023) and ANR 10-IBHU-0002 to JS. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI.

REFERENCES

1 Pui C-H, Robison LL, Look AT: Acute lymphoblastic leukaemia. Lancet 2008;371:1030–1043.

2 Soulier J, Clappier E, Cayuela JM, Regnault A, García-Peydró M, Dombret H, et al.: HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). Blood 2005;106:274–286.

3 De Keersmaecker K, Marynen P, Cools J: Genetic insights in the pathogenesis of T-cell acute lymphoblastic leukemia. Haematologica 2005 [cited 2018 Jul 26];90:1116–27.

4 Weng AP, Ferrando AA, Lee W, Morris JP, Silverman LB, Sanchez-Irizarry C, et al.: Activating Mutations of NOTCH1 in Human T Cell Acute Lymphoblastic Leukemia. Science (80-) 2004;306:269–271.

5 Peirs S, Van der Meulen J, Van de Walle I, Taghon T, Speleman F, Poppe B, et al.: Epigenetics in T-cell acute lymphoblastic leukemia. Immunol Rev 2015;263:50–67.

6 Atak ZK, Gianfelici V, Hulselmans G, De Keersmaecker K, Devasia AG, Geerdens E, et al.: Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. PLoS Genet 2013;9:e1003997.

7 Liu Y, Easton J, Shao Y, Maciaszek J, Wang Z, Wilkinson MR, et al.: The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. Nat Genet 2017; DOI: 10.1038/ng.3909

8 Gianfelici V, Chiaretti S, Demeyer S, Di Giacomo F, Messina M, La Starza R, et al.: RNA sequencing unravels the genetics of refractory/relapsed T-cell acute lymphoblastic leukemia. Prognostic and

therapeutic implications. Haematologica 2016;101:941–50.

9 Yang L, Duff MO, Graveley BR, Carmichael GG, Chen L-L: Genomewide characterization of non-polyadenylated RNAs. Genome Biol 2011;12:R16.

10 Jeck WR, Sharpless NE: Detecting and characterizing circular RNAs. Nat Biotechnol 2014;32:453–61.

11 Verboom K, Van Loocke W, Volders P-J, Decaesteker B, Avila Cobos F, Bornschein S, et al.: A comprehensive inventory of TLX1 controlled long non-coding RNAs in T-cell acute lymphoblastic leukemia through polyA+ and total RNA sequencing. Haematologica 2018;haematol.2018.190587.

12 Wallaert A, Durinck K, Van Loocke W, Van de Walle I, Matthijssens F, Volders PJ, et al.: Long noncoding RNA signatures define oncogenic subtypes in T-cell acute lymphoblastic leukemia. Nat Publ Gr 2016;30:1927–1930.

13 Minoche AE, Dohm JC, Himmelbauer H: Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol 2011;12:R112.

14 Durinck K, Van Loocke W, Van der Meulen J, Van de Walle I, Ongenaert M, Rondou P, et al.: Characterization of the genome-wide TLX1 binding profile in T-cell acute lymphoblastic leukemia. Leukemia 2015;29:2317–2327.

15 Touzri F, Clappier E, Ballerini P, Sigaux F, Gerby B, Baruchel A, et al.: Clonal selection in xenografted human T cell acute lymphoblastic leukemia recapitulates gain of malignancy at relapse. J Exp Med 2011;208:653–661.

16 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al.: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.

17 Buenrostro JD, Wu B, Chang HY, Greenleaf WJ: ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide; in : Current Protocols in Molecular Biology. Hoboken, NJ, USA, John Wiley & Sons, Inc., 2015, p 21.29.1-21.29.9.

18 Lee TI, Johnstone SE, Young RA: Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc 2006;1:729–48.

19 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al.: Model-based Analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137.

20 Yu G, Wang L-G, He Q-Y: ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015;31:2382–2383.

Supplementary Table 1: sample information of the primary T-ALL cohort

sample	subgroup	sequencing	uniquely mapped (%)	total reads	RNA quality scores (RIN)	GEO
primary T-ALL cohort	TAL	polyA[+] RNA-seq	79.27	51134041	9.6	GSM3004545
primary T-ALL cohort	IMM	polyA[+] RNA-seq	89.11	52131482	8.7	GSM3004546
primary T-ALL cohort	TLX	polyA[+] RNA-seq	89.9	50866673	9.6	GSM3004547
primary T-ALL cohort	IMM	polyA[+] RNA-seq	88.12	51643482	8.8	GSM3004548
primary T-ALL cohort	IMM	polyA[+] RNA-seq	89.85	52284445	8.6	GSM3004549
primary T-ALL cohort	HOXA	polyA[+] RNA-seq	88.18	51255315	9.5	GSM3004550
primary T-ALL cohort	IMM	polyA[+] RNA-seq	90.34	52379578	8.7	GSM3004551
primary T-ALL cohort	TAL	polyA[+] RNA-seq	89.38	51693949	8.9	GSM3004552
primary T-ALL cohort	TLX	polyA[+] RNA-seq	90.32	51305639	7.1	GSM3004553
primary T-ALL cohort	TLX	polyA[+] RNA-seq	88.86	52038015	9.7	GSM3004554
primary T-ALL cohort	HOXA	polyA[+] RNA-seq	88.35	51955313	7.1	GSM3004555
primary T-ALL cohort	TAL	polyA[+] RNA-seq	90.14	51681626	9.6	GSM3004556
primary T-ALL cohort	TAL	polyA[+] RNA-seq	88.27	50654317	8.6	GSM3004557
primary T-ALL cohort	HOXA	polyA[+] RNA-seq	90.78	51352215	9.1	GSM3004558
primary T-ALL cohort	TAL	polyA[+] RNA-seq	88.82	52453197	7.5	GSM3004559
primary T-ALL cohort	TAL	polyA[+] RNA-seq	79.66	52174896	7.7	GSM3004560
primary T-ALL cohort	IMM	polyA[+] RNA-seq	87.36	51381227	8.7	GSM3004561
primary T-ALL cohort	IMM	polyA[+] RNA-seq	71.69	53058858	9.9	GSM3004562
primary T-ALL cohort	TLX	polyA[+] RNA-seq	87.75	51038574	9.4	GSM3004563
primary T-ALL cohort	IMM	polyA[+] RNA-seq	88.02	52502134	9.8	GSM3004564
primary T-ALL cohort	TAL	polyA[+] RNA-seq	90.28	51909933	9.7	GSM3004565
primary T-ALL cohort	IMM	polyA[+] RNA-seq	90.62	51911212	9.4	GSM3004566
primary T-ALL cohort	HOXA	polyA[+] RNA-seq	87.4	59081633	9.2	GSM3004567
primary T-ALL cohort	TAL	polyA[+] RNA-seq	89.43	51919497	8.1	GSM3004568
primary T-ALL cohort	TAL	polyA[+] RNA-seq	84.16	51255849	9.2	GSM3004569
primary T-ALL cohort	TLX	polyA[+] RNA-seq	69.59	51305951	8.7	GSM3004570
primary T-ALL cohort	IMM	polyA[+] RNA-seq	76.36	51338014	8.9	GSM3004571

primary T-ALL cohort	TLX	polyA[+] RNA-seq	83.74	51368719	7.8	GSM3004572
primary T-ALL cohort	IMM	polyA[+] RNA-seq	83.05	50643895	9.6	GSM3004573
primary T-ALL cohort	IMM	polyA[+] RNA-seq	73.78	51591573	9,00	GSM3004574
primary T-ALL cohort	TLX	polyA[+] RNA-seq	83.88	51475202	8.9	GSM3004575
primary T-ALL cohort	TAL	polyA[+] RNA-seq	81.29	51657232	9.1	GSM3004576
primary T-ALL cohort	TAL	polyA[+] RNA-seq	80.47	51916311	8.9	GSM3004577
primary T-ALL cohort	HOXA	polyA[+] RNA-seq	81.92	52274790	8.9	GSM3004578
primary T-ALL cohort	TLX	polyA[+] RNA-seq	81.94	51773369	3.8	GSM3004579
primary T-ALL cohort	TLX	polyA[+] RNA-seq	80.82	51875328	9.4	GSM3004580
primary T-ALL cohort	HOXA	polyA[+] RNA-seq	83.16	52264800	8.6	GSM3004581
primary T-ALL cohort	TAL	polyA[+] RNA-seq	85.51	52235267	8,00	GSM3004582
primary T-ALL cohort	TAL	polyA[+] RNA-seq	64.75	52826763	7.5	GSM3004583
primary T-ALL cohort	TAL	polyA[+] RNA-seq	78.01	51091717	8.1	GSM3004584
primary T-ALL cohort	TLX	polyA[+] RNA-seq	88.18	51199311	8.2	GSM3004585
primary T-ALL cohort	IMM	polyA[+] RNA-seq	84.61	51344603	9.4	GSM3004586
primary T-ALL cohort	TAL	polyA[+] RNA-seq	84.81	51362219	7.8	GSM3004587
primary T-ALL cohort	TAL	polyA[+] RNA-seq	85.7	51100095	1.7	GSM3004588
primary T-ALL cohort	TAL	polyA[+] RNA-seq	81.17	50899540	7.5	GSM3004589
primary T-ALL cohort	TAL	polyA[+] RNA-seq	77.04	52229414	NA	GSM3004590
primary T-ALL cohort	TAL	polyA[+] RNA-seq	83.0	52319669	9.9	GSM3004591
primary T-ALL cohort	TAL	polyA[+] RNA-seq	80.41	51322739	6.2	GSM3004592
primary T-ALL cohort	TAL	polyA[+] RNA-seq	85.31	51531625	9.2	GSM3004593
primary T-ALL cohort	TAL	polyA[+] RNA-seq	83.45	51764239	9.4	GSM3004594
primary T-ALL cohort	TLX	polyA[+] RNA-seq	86.55	50986744	8.3	GSM3004595
primary T-ALL cohort	TLX	polyA[+] RNA-seq	84.49	52010225	8.4	GSM3004596
primary T-ALL cohort	TLX	polyA[+] RNA-seq	87.03	52000045	8.9	GSM3004597
primary T-ALL cohort	TLX	polyA[+] RNA-seq	82.79	51774667	8,00	GSM3004598
primary T-ALL cohort	TLX	polyA[+] RNA-seq	87.26	51916137	8.6	GSM3004599
primary T-ALL cohort	TLX	polyA[+] RNA-seq	81.15	52018296	10,00	GSM3004600
primary T-ALL cohort	HOXA	polyA[+] RNA-seq	77.87	60514863	10,00	GSM3004601
primary T-ALL cohort	TLX	polyA[+] RNA-seq	88.0	51956038	9.4	GSM3004602

primary T-ALL cohort	IMM	polyA[+] RNA-seq	86.93	52084376	9.6	GSM3004603

Supplementary Table 2: sample information of ALL-SIL lymphoblasts

sample	subgroup	sequencing	uniquely mapped (%)	total reads	GEO
***TLX1* knockdown**	siRNA 1 - repl 1	polyA[+] RNA-seq	93.8	124147993	GSM3004536
***TLX1* knockdown**	siRNA 2 - repl 1	polyA[+] RNA-seq	94.01	93035379	GSM3004537
***TLX1* knockdown**	scrambled siRNA - repl 1	polyA[+] RNA-seq	93.71	84583398	GSM3004538
***TLX1* knockdown**	siRNA 1 - repl 2	polyA[+] RNA-seq	94.25	119636318	GSM3004539
***TLX1* knockdown**	siRNA 2 - repl 2	polyA[+] RNA-seq	94.12	105500445	GSM3004540
***TLX1* knockdown**	scrambled siRNA - repl 2	polyA[+] RNA-seq	93.74	129759760	GSM3004541
***TLX1* knockdown**	siRNA 1 - repl 3	polyA[+] RNA-seq	93.55	121660448	GSM3004542
***TLX1* knockdown**	siRNA 2 - repl 3	polyA[+] RNA-seq	93.32	77081841	GSM3004543
***TLX1* knockdown**	scrambled siRNA - repl 3	polyA[+] RNA-seq	93.71	107915058	GSM3004544
***TLX1* knockdown**	scrambled siRNA - repl 3	total RNA-seq	88.05	49625332	GSM3004611
***TLX1* knockdown**	siRNA 1 - repl 2	total RNA-seq	88.48	49943664	GSM3004612
***TLX1* knockdown**	siRNA 2 - repl 2	total RNA-seq	88.77	56881159	GSM3004613
***TLX1* knockdown**	siRNA 2 - repl 3	total RNA-seq	88.4	51132644	GSM3004614
***TLX1* knockdown**	scrambled siRNA - repl 2	total RNA-seq	88.51	51878120	GSM3004615
***TLX1* knockdown**	siRNA 1 - repl 3	total RNA-seq	88.9	57781287	GSM3004616
***TLX1* knockdown**	siRNA 2 - repl 1	total RNA-seq	87.37	49905219	GSM3004617
***TLX1* knockdown**	siRNA 1 - repl 1	total RNA-seq	88.37	50441264	GSM3004618
***TLX1* knockdown**	scrambled siRNA - repl 1	total RNA-seq	78.93	52942076	GSM3004619
ALL-SIL lymphoblasts	-	ATAC-seq	76.72	84064905	GSM3004532
ALL-SIL lymphoblasts	H3K4me1	ChIP-seq	89.49	41543910	GSM3004533
ALL-SIL lymphoblasts	input	ChIP-seq	69.79	28590025	GSM3004534
ALL-SIL lymphoblasts	H3K4me3	ChIP-seq	89.85	36408072	GSM3004535

Paper 3

SMARTer single cell total RNA sequencing

Verboom Karen*, Everaert Celine*, Bolduc Nathalie, Livak J. Kenneth, Yigit Nurten, Rombaut Dries, Anckaert Jasper, Simon Lee, Venø T Morten, Kjems Jørgen, Speleman Frank, Mestdagh Pieter and Vandesompele Jo

*contributed equally

Contribution: I adapted the bulk library method to a single cell library prep method on the C1. I performed all serum starvation experiments and subsequently performed cell cycle analysis. I performed the C1 experiments with the help of a lab technician. The co-first author and I equally contributed to the figures and manuscript.

Accepted in Nucleic Acids Research

Impact factor 2017: 11.561

SMARTer single cell total RNA sequencing

Verboom Karen[1,2a], Everaert Celine[1,2a], Bolduc Nathalie[3], Livak J. Kenneth[4], Yigit Nurten[1,2], Rombaut Dries[1,2], Anckaert Jasper[1,2], Simon Lee[3], Venø T Morten[5], Kjems Jørgen[5b], Speleman Frank[1,2], Mestdagh Pieter[1,2] and Vandesompele Jo[1,2]

[1] *Center for Medical Genetics, Ghent University, Ghent, Belgium*

[2] *Cancer Research Institute Ghent, Ghent, Belgium*

[3] *Takara Bio USA, Mountain View, California, 94043, USA*

[4] *Fluidigm Corporation, South San Francisco, California, 94080, USA*

[5] *Department of Molecular Biology and Genetics and Interdisciplinary Nanoscience Center, Aarhus University, Aarhus, DK-8000, Denmark*

[a] *These authors contributed equally to this work.*

[b] *current address: Omiics, Aarhus, DK-8200, Denmark*

Corresponding author: Jo Vandesompele, Jo.Vandesompele@UGent.be

ABSTRACT

Single cell RNA sequencing methods have been increasingly used to understand cellular heterogeneity. Nevertheless, most of these methods suffer from one or more limitations, such as focusing only on polyadenylated RNA, sequencing of only the 3' end of the transcript, an exuberant fraction of reads mapping to ribosomal RNA, and the unstranded nature of the sequencing data. Here, we developed a novel single cell strand-specific total RNA library preparation method addressing all the aforementioned shortcomings. Our method was validated on a microfluidics system using three different cancer cell lines undergoing a chemical or genetic perturbation and on two other cancer cell lines sorted in microplates. We demonstrate that our total RNA-seq method detects an equal or higher number of genes compared to classic polyA[+] RNA-seq, including novel and non-polyadenylated genes. The obtained RNA expression patterns also recapitulate the expected biological signal. Inherent to total RNA-seq, our method is also able to detect circular RNAs. Taken together, SMARTer single cell total RNA sequencing is very well suited for any single cell sequencing experiment in which transcript level information is needed beyond polyadenylated genes.

INTRODUCTION

To understand the complexity of life, knowledge of cells as fundamental units is key. Recently, technological advances have emerged to enable single cell RNA sequencing (RNA-seq). In 2009, Tang et al. published the first single cell RNA-seq protocol in which cells were picked manually and transcripts reverse transcribed using a polydT primer (1). As the throughput was low, new methods using early multiplexing, such as STRT-seq and SCRB-seq, were introduced in which cells were pooled at an early step in the workflow, enabling processing of many cells in parallel (2–4). In contrast to these methods that have inherent 3' end or 5' end bias, Smart-seq2 generates read coverage across the whole transcript expanding the spectrum of applications as this method can be used for fusion detection, single nucleotide variants

(SNV) analysis, and splicing, beyond typical gene expression profiling applications (5, 6). To reduce the PCR bias generated in the aforementioned methods, CEL-seq and MARS-seq were introduced using linear in vitro transcription (IVT) instead of PCR to obtain enough cDNA for sequencing (7–9). Most recently, droplet and split-pool ligation based methods capturing thousands of single cells were developed, providing new insights in cellular heterogeneity and rare cell types (10–14). The main drawback of these methods is that analyses are typically confined to gene expression of only (3' ends of) polyadenylated transcripts (Table 1). More complex analyses with respect to alternative splicing, allele specific expression, mutation analysis, assembly of (novel) transcripts, circular RNA (circRNA) quantification and post-transcriptional regulation, require full-length and full-transcriptome methods. Moreover, sequencing a large number of cells is often compromising sequencing depth, resulting in low coverage per cell and detection of only the most abundant transcripts (24). In contrast to these droplet based methods, microfluidic chip and flowcytometry based platforms typically capture fewer cells, but are able to sequence entire transcripts and detect a substantially higher number of genes per cell providing a more complete view of the complexity and richness of single cells' transcriptomes (6, 25). Of note, most single cell RNA-seq studies assess only 3' end polyadenylated (polyA[+]) transcripts, ignoring non-polyadenylated (polyA[-]) transcripts (Table 1) (6, 12, 14). Since a substantial part of the human transcriptome is non-polyadenylated, various RNA types including circRNAs, enhancer RNAs, histone RNAs, and a sizable fraction of long non-coding RNAs (lncRNAs) are not quantified using these classic methods (26–28). In order to study polyA[-] transcripts at the single cell level, total RNA-seq workflows were developed (22, 29, 30). While in principle both polyA[+] and polyA[-] transcripts are converted into a sequencing-ready library using random primer mediated reverse transcription, these methods suffer from one or more of the following limitations: the strand-orientation information is lost and a high percentage of reads map to ribosomal RNA (rRNA) (Table 1). Therefore, new methods circumventing these limitations are warranted. A rRNA depletion step is essential as up to 95 % of the total RNA content in a mammalian cell consists of rRNA. Moreover, to discriminate sense and antisense overlapping transcripts, stranded sequencing is crucial; at least 38 % of the annotated transcripts in cancer cells have antisense expression (31). Here, we developed a novel easy to use and efficient single cell total RNA-seq workflow based on the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian combining for the first time strandeness and effective removal of ribosomal cDNA (Table 1). We ported the method to Fluidigm's C1 single cell microfluidics instrument, and demonstrated that the method works equally well on FACS sorted cells in microplates. In total, 458 cells from 5 different human cancer cell lines in 4 experiments were sequenced with a total sequencing depth of 1528 million reads. Using our novel method, we consistently observe less than 3 % of ribosomal reads and we detect more than 5360 genes by at least four reads, including novel genes, polyA[-] genes and circular RNAs.

METHODS

Cell lines

The neuroblastoma cell line NGP, used for the C1 experiments, is a kind gift of prof. R. Versteeg (Amsterdam, the Netherlands). Cells were maintained in RPMI-1640 medium (Life Technologies, 52400-025) supplemented with 10 % fetal bovine serum (PAN Biotech, P30-3306), 1 % of L-glutamine (Life Technologies, 15140-148) and 1 % penicillin/streptomycin (Life Technologies, 15160-047) (referred to as complete medium) at 37 °C in a 5 % CO_2 atmosphere. Short tandem repeat genotyping was used to validate cell line authenticity prior to performing the described experiments and mycoplasma testing was done on a monthly basis using the MycoAlert Mycoplasma

Detection Kit (Lonza, T07-318), according to manufacturer's instructions. The A375 (ATCC CRL-1619) and Jurkat (clone E6.-1; ATCC TIB-152) cells, used for the FACS experiments, were grown in Dulbecco's modified Eagle's medium (DMEM; Millipore-Sigma, D5796) supplemented with 10 % Tet system approved fetal bovine serum (FBS) (Takara, 631106) and RPMI-1640 medium (RPMI; Millipore-Sigma, R0883) supplemented with 10 % Tet system approved FBS, respectively. Cell lines were sub-cultured every two days or when they reached > 80% confluence (A375) or >1x10^6 cells/ml (Jurkat).

Cell cycle synchronization and nutlin-3 treatment of NGP cells

NGP cells were synchronized using serum starvation prior to nutlin-3 treatment. First, cells were seeded at low density for 48 hours in complete medium. Then, cells were refreshed with serum-free medium for 24 hours. Finally, the cells were treated with either 8 µM of nutlin-3 (Cayman Chemicals, 10004372, dissolved in ethanol) or vehicle. Cells were trypsinized (Gibco, 25300054) 24 hours post treatment and harvested for single cell analysis, bulk RNA isolation and cell cycle analysis.

Cell cycle analysis

Four million cells were washed with PBS (Gibco, 14190094) and the pellet was resuspended in 300 µl PBS. Next, 700 µl of 70 % ice-cold ethanol was added dropwise while vortexing to fix the cells. Cells were stored at -20 °C for at least 1 hour. After incubation, cells were washed with PBS and the pellet was resuspended in 1 ml PBS containing RNAse A (Qiagen, 19101) at a final concentration of 0.2 mg/ml. After 1 hour incubation at 37 °C, propidium iodide (BD biosciences, 556463) was added to a final concentration of 40 µg/ml. Samples were loaded on a S3 cell sorter (Bio-Rad) and analyzed using the FlowJo v.10 software.

RNA isolation and cDNA synthesis

Total RNA was isolated using the miRNeasy mini kit (Qiagen, 217084) with DNA digestion on-column according to the manufacturer's instructions. RNA concentration was measured using spectrophotometry (Nanodrop 1000, Thermo Fisher Scientific). cDNA was synthesized using the iScript Advanced cDNA synthesis kit (Bio-Rad, 1708897) using 500 ng RNA as input in a 20 µl reaction. cDNA was diluted to 2.5 ng/µl with nuclease-free water prior to RT-qPCR measurements.

Reverse transcription quantitative PCR

PCR mixes containing 2.5 µl 2x SsoAdvanced SYBR qPCR supermix (Bio-Rad, 04887352001), 0.25 µl each forward and reverse primer (5 µM, IDT), and 2 µl diluted cDNA (5 ng total RNA equivalents) were analyzed on the LightCycler480 instrument (Roche) using two replicates. Expression levels were normalized using expression data of four stable reference genes (*SDHA, YWHAZ, TBP, HPRT1*). These reference genes were selected based on geNorm analysis with the qbase+ software v3.0 (Biogazelle), identifying the most stable references genes for normalization. RT-qPCR data was analyzed using the qbase+ software v3.0 (Biogazelle). Primer sequences are available in Supplementary Table 1.

FACS sorting of A375 and Jurkat cells in microplates

Before sorting, cells were washed twice in 1X PBS buffer (DPBS without calcium chloride and magnesium chloride; Sigma Aldrich, D8537) and labelled with 7-AAD (BD Pharmingen, 51-68981E) for live/dead differentiation and FITC-conjugated antibody [anti-CD47 (BD Pharmingen, 556045) for A375 and anti-CD81 (BD Pharmingen, 551108) for Jurkat]. After washing off the unbound antibodies in 1X PBS, cells were resuspended in BD FACS Pre-Sort Buffer (BD, 563503). Single cell sorting in 8-tube PCR strips was done using a BD FACSJazz Cell Sorter. A375 cells were sorted in 7 µl 1X PBS buffer and Jurkat cells in 8 µl lysis solution [100 µl 10X Lysis buffer (Takara, 635013), 5 µl RNase Inhibitor (Takara, 635013) and 700 µl water]. Following sorting, tubes were sealed and subjected to a quick spin and immediately frozen on dry ice and finally stored at –80 °C until use.

All sorting experiments included negative controls (no cell in a well).

Single cell total RNA library preparation of nutlin-3 treated NGP cells

Cells were washed with PBS and pellets of vehicle treated cells were resuspended and incubated in 1 ml pre-warmed (37 °C) cell tracker (CellTracker Green BODIPY Dye, Thermo fisher Scientific, C2102) for 20 minutes at room temperature. After incubation, cells were washed in PBS and resuspended in 1 ml wash buffer (Fluidigm, 100-6201). An equal number of stained (vehicle treated) and non-stained (nutlin-3 treated) cells were mixed and diluted to 300,000 cells per ml. Suspension buffer (Fluidigm) was added to the cells in a 3:2 ratio and 6 µl of this mix of was loaded on a primed C1 Single-Cell Open App IFC (Fluidigm, 100-8134) designed for medium-sized cells (10-17 µm). Cells were captured using the 'SMARTer single cell total RNA-seq' script deposited in Script Hub (Fluidigm). Upon capture, cells were visualized using the Axio Observer Z1 (Zeiss) and a median multiplet rate of 34.54 % was observed over all experiments. These cells were excluded from further analyses. Sequencing libraries were generated using the C1 running the 'SMARTer single cell total RNA-seq' script deposited on Script Hub. In short, the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Pico v2, total RNA, Takara, 634413) was used to synthesize cDNA with following modifications. Cells were fragmented and lysed by loading 7 µl of 10x reaction mix [2.3 µl SMART Pico Oligo Mix v2, 6 µl 5x first-strand buffer, 1 µl 20x C1 loading reagent (Fluidigm), 3 µl lysis mix (19 µl 10x lysis buffer, 1 µl RNAse inhibitor (40 U/µl)), 1 µl 1/1000 diluted ERCC spikes (Ambion, 4456740), 6.7 µl water] and incubating the cells at 85 °C for 6 minutes (to lyse cells and fragment RNA) followed by 2 minutes at 10 °C. Next, 8 µl first strand master mix [1 µl C1 loading reagent, 4 µl 5x first-strand buffer, 0.9 µl RNAse inhibitor (40 U/µl), 3.5 µl SMARTScribe reverse transcriptase (100 U/µl), 7.9 µl SMART TSO Mix v2 (from Takara kit, 634413), 2.7 µl water] was loaded and incubated at 42 °C for 90 minutes followed by 70 °C for 10 minutes. Finally, a PCR master mix for each well was made [1 µl 20x loading reagent, 2 µl 2.4 µM forward primer (Takara, 634413), 2 µl 2.4 µM reverse primer, 13.1 µl 1.5x PCR mix (1050 µl 2x SeqAmp CB buffer, 42 µl SeqAmp DNA polymerase, 308 µl water)] and 5 µl of each of these mixes was loaded in the harvest wells of the IFC. The samples were incubated for 1 minute at 94 °C followed by 11 PCR cycles (30 s at 98 °C, 15 s at 55 °C, 30 s at 68 °C) and 2 minutes at 68 °C. Following this initial cDNA amplification, 12 wells were pooled per tube using 8 µl of cDNA per cell. Next steps of the library prep were performed according to manufacturer's instructions with minor modifications. 13 PCR cycles were used for PCR2 and a 1:1 ratio was used for beads cleanup after PCR2. Next, the samples were resuspended in 22 µl 5 mM tris buffer (from kit) and 20 µl was used to perform a second beads cleanup using a 0.9:1 ratio. Finally, the samples were resuspended in 12 µl tris buffer and the quality was determined on the Fragment Analyzer (Advanced Analytical). Of note, the protocol can also be executed using the single cell specific version of the kit, released by Takara (SMART-Seq Stranded Kit, 634442) after we had completed our C1 experiments.

Single cell polyA[+] RNA library preparation of nutlin-3 treated NGP cells

Vehicle treated cells were stained with cell tracker as described above. An equal number of stained (vehicle treated) and non-stained (nutlin-3 treated) cells were mixed and diluted to 300,000 cells per ml. Suspension buffer was added to the cells in a 3:2 ratio and 6 µl of this mix of was loaded on a primed C1 Single-Cell Auto Prep Array for mRNA Seq (Fluidigm, 100-6041) designed for medium-sized cells (10-17 µm). Single cell polyA[+] RNA sequencing on the C1 was performed using the SMART-Seq v4 Ultra Low Input RNA Kit for the Fluidigm C1 System (SMART-Seq v4, polyA[+] RNA, Takara, 635026) according to manufacturer's instructions. One microliter of the ERCC spike-in mix was diluted in 999 µl loading buffer to get a 1/1000 dilution of the ERCC spikes. One microliter of this dilution was added to the 20 µl lysis mix. The quality of

the cDNA was checked for 11 random single cells on the Fragment Analyzer. The concentration of the cells was measured using the quantifluor dsDNA kit (Promega, E2670) and glomax (Promega) according to manufacturer's instructions. The samples were 1/5 diluted in C1 harvest reagent (Fluidigm). Next, library prep was performed using the Nextera XT library prep kit (Illumina, FC-131-1096) according to manufacturer's instructions, followed by quality control on the Fragment Analyzer.

Single cell total RNA library preparation of FACS sorted A375 and Jurkat cells

Cells were processed using the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian (Takara, 634413) or the SMART-Seq Stranded Kit (Takara, 634444) reagents according to the manufacturer's instructions with some modifications that were also implemented in the C1 protocol. For the SMART-Seq Stranded Kit, the Ultra Low Input workflow described in the user manual was followed by pooling of 8 samples according to Appendix A of the user manual. For the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian, the cells were also processed as described for the SMART-Seq Stranded Kit, but using the reagents specific to the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian, which were also used for the C1 protocol. For both kits, cells sorted in a lysis solution instead of 1X PBS were processed without addition of lysis buffer. Identital to the C1 protocol, the initial RNA shearing step was performed at 85 °C for 6 min and 10 and 13 PCR cycles were carried out for PCR1 and PCR2, respectively.

Library sequencing

All libraries were quantified using the KAPA library quantification kit (Roche) and libraries were diluted to 4 nM. For NGP, the polyA[+] RNA library and total RNA library were pooled in a 1/4 ratio and 1.5 pM of the pooled library was single-end sequenced on a NextSeq 500 (Illumina) with a read length of 75 bp and a total sequencing read depth of 274 million reads, combining single cell polyA[+] and total RNA libraries to prevent inter-run bias. A median sequencing read depth of 0.81 and 3.67 million reads per cell was reached for the single cell polyA[+] and total RNA libraries, respectively. In addition, 1.3 pM of the total RNA library was also sequenced in 2x75 paired-end sequencing run mode on the NextSeq 500, yielding 327 million reads and a median sequencing read depth of and 4.04 million per cell. The fastq data is deposited in GEO (GSE119984). A375 and Jurkat total RNA libraries were pooled and 1.2 pM of the pooled library was sequenced in 2x75 paired-end run mode on the NextSeq 500, yielding 41 million reads. FASTQ data is deposited in GEO (GSE130578).

Sequencing data quality control

While single-end sequencing libraries do not require pre-trimming, the paired-end libraries were trimmed using cutadapt (v.1.16) (32) to remove 3 nucleotides of the 5' end of read 2. To assess the quality of the data, the reads were mapped using STAR (v.2.5.3) (33) on the hg38 genome including the full ribosomal DNA (45S, 5.8S and 5S) and mitochondrial DNA sequences. The parameters of STAR were set to retain only primary mapping reads, meaning that for multi-mapping reads only the best scoring location is retained. Using SAMtools (v1.6) (34), reads mapping to the different nuclear chromosomes, mitochondrial DNA and rRNA were extracted and annotated as exonic, intronic or intergenic. In contrast to the unstranded nature of polyA[+] Smart-seq v4 data, the total RNA SMARTer-seq data is stranded and processed accordingly (unless explicitely mentioned). Gene body coverage was calculated using the full Ensembl (v91) (35) transcriptome. The coverage per percentile was calculated, followed by a loess regression fit.

Quantification of Ensembl and LNCipedia genes

Genes were quantified by Kallisto (v.0.43.1) (36) using both Ensembl (v.91) (35) extended with the ERCC spike sequences and LNCipedia (v.5.0) (37). The strandedness of the total RNA-seq reads was considered by running the –rf-

stranded mode and omitted for unstranded analysis of the data. Subsampling 1 million reads (polyA[+] RNA libraries) or 1, 4 or 8 million reads (total RNA libraries) was performed by seqTK (v.1.2) followed by Kallisto quantification. Further processing was done with R (v.3.5.1) making use of tidyverse (v.1.2.1). To measure the biological signal we first performed differential expression analysis between the treatment groups using DESeq2 (v.1.20.0) (38) in combination with Zinger (v.0.1.0) (39). To identify enriched gene sets a fsgea (v.1.6.0) analysis was performed, calculating enrichment for the hallmark gene sets retrieved from MSigDB (v.6.2).

Circular RNA detection

CircRNAs were detected using the deeper sequenced paired-end sequencing data. Trim_galore (v.0.4.1) was used to trim adaptor sequences, perform quality filtering and remove 3 nucleotides from the 5' end of read 2. Subsequently, reads from all samples were combined, adding originating sample names to read names for later splitting of data. The combined data was used for circRNA detection using find_circ (v.1) (40) using the reads2sample (find_circ.py -r) option to allow circRNA detection on the combined dataset while dividing out the contribution from each sample in the output. Only circRNAs with unique mapping on both anchors were accepted. Human genome hg19 was used for circRNA analysis. CircRNAs were annotated with host gene names from RefSeq (release 75) and circBase IDs from circbase.org. The Database for Annotation, Visualisation and Integrated Discovery (DAVID, v.6.8) (41, 42) was used for Gene Ontology (GO) analysis for the circRNA host genes using biological processes (BP) and molecular function (MF). P-value < 0.05 was used for statistical significance.

Single cell transcriptome assembly

A transcriptome per cell was created by combining STAR (v.2.5.3) and Stringtie (v.1.3.0) (43), using the deeper sequenced paired-end sequencing data. The parameters of Stringtie were set to require a coverage of 1. These single cell transcriptomes were merged with the Ensembl (v.91) transcriptome as a reference. From the merged multi-cell transcriptome, only multi-exonic genes with a minimum length of 200 nt were retained. To define the set of novel genes, genes annotated in Ensembl (35) or LNCipedia (v.5.0) (37) were filtered out. All genes in this novel multi-cell transcriptome were quantified using Kallisto on single-end subsample data (1, 4 or 8 million reads per cell).

Table 1: Characteristics of the top ten cited single cell polyA[+] RNA-seq in Web of Science and four available single cell total RNA-seq methods (including our SMARTer method).

	total RNA-seq	full length	rRNA < 5 %	stranded	reference
Drop-seq	-	-	+	-	(15)
Tang et al.	-	+	+	-	(16)
InDrop	-	-	+	-	(17)
MARS-seq	-	-	+	-	(9)
Smart-seq2	-	+	+	-	(5,18)
CEL-seq	-	-	+	+	(7)
STRT-seq	-	-	+	+	(3)
Quartz-seq	-	+	+	-	(19)
CEL-seq2	-	-	+	+	(8)
cytoSeq	-	-	+	-	(20)
SuPeR-seq	+	+	+	-	(21)
RamDA-seq	+	+	-	-	(22)
MATQ-seq	+	+	NA	-	(23)
SMARTer	+	+	+	+	

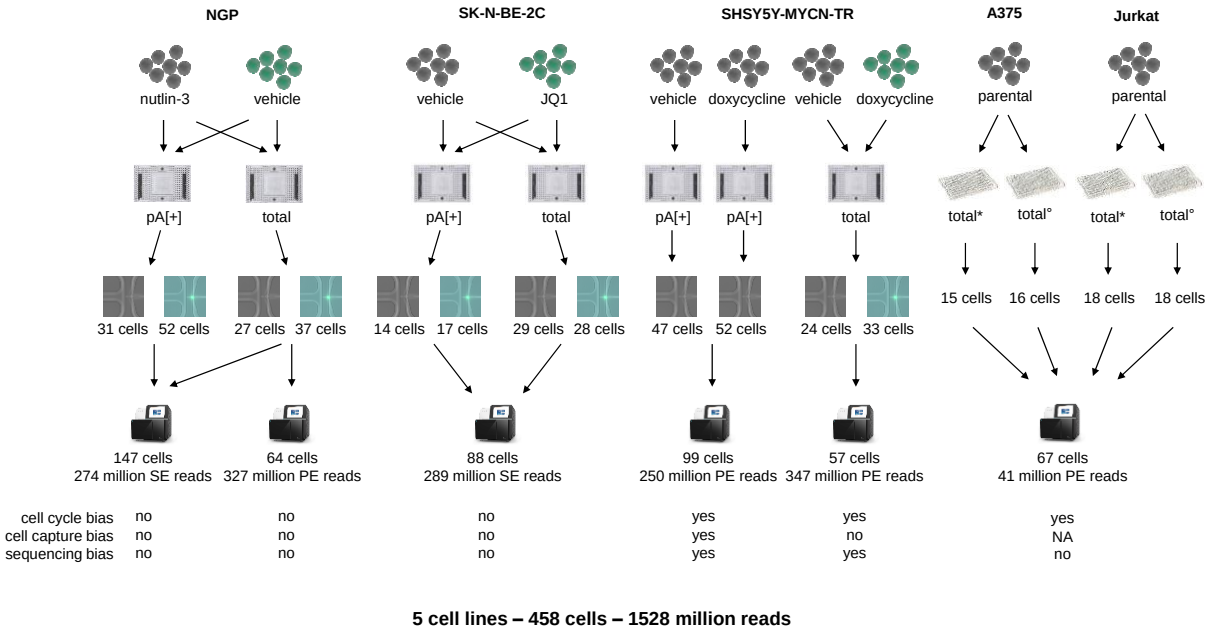


Figure 1: overview of experimental set-up. Single cell total RNA libraries of the FACS sorted cells were generated using 2 different reagent kits (#634413, denoted with * and #634444, denoted with °).

Genes with an estimated count higher than 1 were retained.

RESULTS

Principle of SMARTer single cell total RNA sequencing

We developed a single cell total RNA-seq protocol for unbiased, full transcript and strand-specific analysis of both polyadenylated and non-polyadenylated transcripts from mammalian cells. The method uses reagents from the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Pico v2, total RNA), a kit that is meant for low input bulk total RNA-seq, whereby we optimized reaction volumes, number of PCR cycles, and duration and temperature of the RNA fragmentation. The library preparation method employs random primers and a template switching mechanism to capture full transcript fragments of both polyadenylated (polyA[+]) and non-polyadenylated (polyA[-]) transcripts. Unwanted ribosomal cDNA is removed using probes, complementary to mammalian rRNA. After successfully porting the bulk library prep protocol to Fluidigm's C1 single cell instrument, we assessed the performance of the single cell total RNA-seq protocol through three distinct experiments in which nutlin-3, JQ1 or doxycycline was used to treat NGP, SK-N-BE-2C, and SHSY5Y-MYCN-TR neuroblastoma cell lines, respectively (with vehicle treated cells as control) (Figure 1). In addition, we performed matched single cell polyA[+] RNA-seq as a reference using cells from the same pool. While all experiments were successful, we focus our analyses and performance assessment on the NGP data. In this experiment, the treated and control cells were processed in the same microfluidic chip (preventing possible chip bias), the highest number of cells were captured, and the highest sequencing depth was reached.

SMARTer single cell total RNA sequencing yields high-quality data

In single cell sequencing experiments, it is important to prevent or limit potential biases that mask true biological differences. In

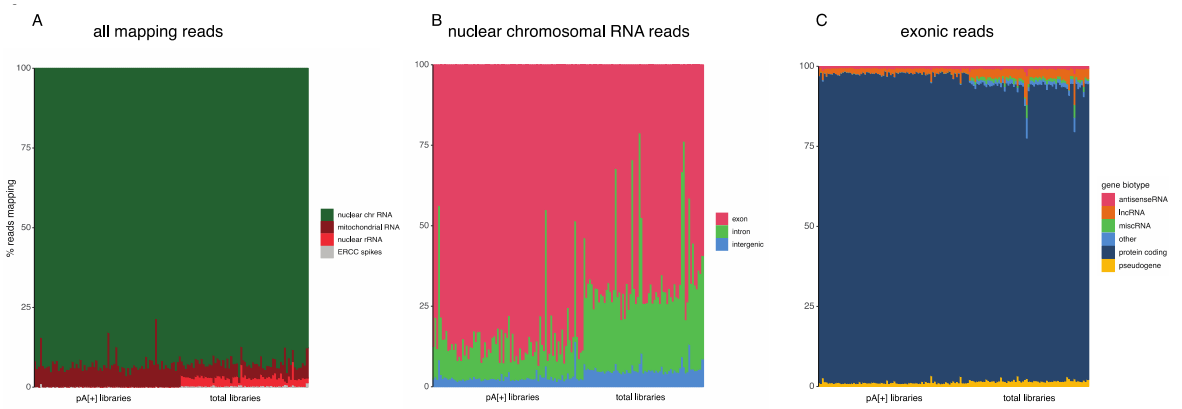


Figure 2: read distribution differs between polyA[+] and total RNA libraries. A) Percentage of reads derived from nuclear RNA, mitochondrial RNA and ribosomal RNA per cell quantified with STAR. B) Percentage of the reads originating from nuclear chromosomes derived from exonic, intronic and intergenic regions per cell quantified with STAR. C) Percentage of exonic reads attributed to the different biotypes per cell quantified with Kallisto.

particular, the cell cycle state is a known confounder (44). Therefore, we synchronized the cells through serum starvation for 24 hours. Upon synchronization, 80.3 % of the NGP cells showed an arrest at the G0/G1 stage compared to only 53.3 % for non-synchronized NGP cells (Supplementary Figure 1 A-B). Subsequently, the synchronized NGP cells were treated for 24 hours with vehicle or nutlin-3, the latter known to release TP53 from its negative regulator MDM2. As expected, nutlin-3 treatment resulted in cell cycle arrest (Supplementary Figure 1 C-D). To prevent possible C1 batch effects (45), vehicle treated NGP cells were stained and loaded together with the non-stained nutlin-3 treated cells on the same C1 chip. Based on the fluorescent label and the transparency of the C1 system, vehicle and nutlin-3 treated cells were discriminated by fluorescence microscopy. By loading two C1 chips, one for polyA[+] RNA and one for total RNA library preparation, we captured 31 and 27 nutlin-3 treated versus 52 and 37 vehicle treated single cells, respectively. High-quality cDNA libraries of polyA[+] and total RNA were generated using the SMART-Seq v4 Ultra Low Input RNA Kit for the Fluidigm C1 System (SMART-Seq v4, polyA[+]) and our novel SMARTer single cell total RNA-seq protocol, respectively (Supplementary Figure 1 E-F). ERCC spike-in molecules were added for external quality control in the lysis mix (Supplementary Figure 2). For the recovered spikes (with a concentration in the original mix of at least 10 attomoles/µl), linear models were calculated (Supplementary Figure 3), retrieving similar R^2 values for the polyA[+] RNA and total RNA library preparation protocol (Supplementary Figure 4). The transcripts detected in the polyA[+] libraries were somewhat shorter compared to the total RNA libraries (Supplementary Figure 5). In addition, the total RNA-seq libraries show a more uniform transcript coverage (Supplementary Figure 6).

As expected, a higher fraction of reads mapped to nuclear rRNA in the total RNA-seq libraries compared to the polyA[+] RNA libraries (average of 2.739 % [2.488, 2.990; 95 % confidence interval (CI)] vs. 0.031 % [0.026, 0.035; 95 % CI], respectively). Nevertheless, the fraction of nuclear rRNA is very low in the total RNA libraries considering the use of random priming data (Figure 2A), and substantially lower compared to the RAMDA-seq method (9.667 % rRNA [9.615, 9.719; 95 % CI], Supplementary Figure 7). Furthermore, the single cell total RNA libraries contain more intronic (27.99 % [25.06, 30.91; 95 % CI] vs. 11.87 % [10.14, 13.60; 95 % CI]) and intergenic (5.38 % [5.00, 5.76; 95 % CI] vs. 2.90 % [2.54, 3.26; 95 % CI]) reads originating from nuclear chromosomes compared to polyA[+] RNA libraries (Figure 2B). Non-polyadenylated histone genes are highly abundant in the total RNA libraries, while low or absent in the polyA[+] libraries, confirming the validity of our single cell total RNA-seq workflow (Supplementary Figure

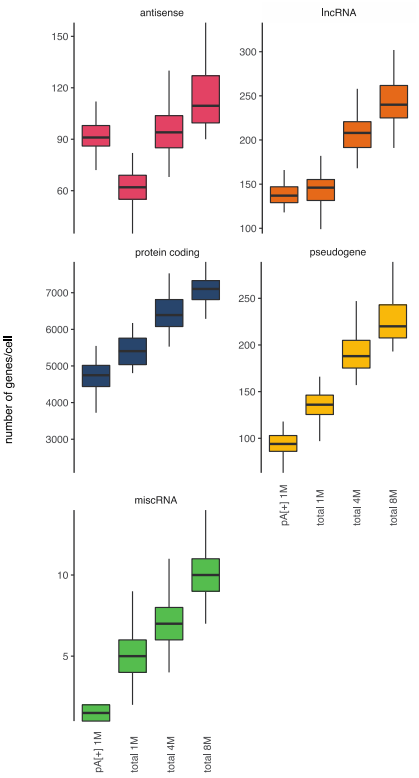


Figure 3: total RNA libraries comprise more genes per RNA biotype. All genes in Ensembl v.91 were quantified on subsampled data (1, 4 or 8 million reads per cell). Only genes with at least 10 reads were included.

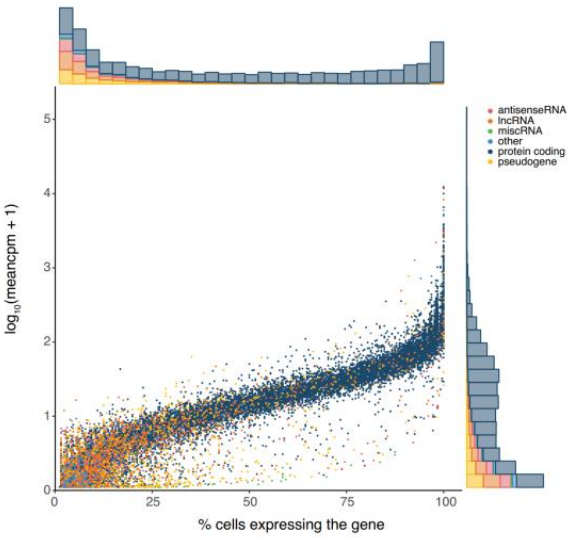


Figure 4: gene biotype and abundance are correlated to fraction of expressed cells. In general, the fraction of cells in which a gene is expressed is related to the mean expression level of that gene; exceptionally, some low abundant genes are present in a large fraction of cells. RNA biotypes that are known to be more cell-type specifically expressed, such as lincRNAs, are expressed in fewer cells.

8). Equal results were obtained for the SK-N-BE-2C, and SHSY5Y-MYCN-TR cell lines (Supplementary Figure 9).

SMARTer single cell total RNA sequencing reveals a unique set of genes

More reads map to long intergenic RNAs (lincRNAs) using the single cell total RNA-seq protocol (2.64 % [2.523, 2.756; 95 % CI]) compared to polyA[+] RNA sequencing (1.67 % [1.489, 1.849; 95 % CI]). In addition, the single cell total RNA-seq protocol detects an equal or higher number of genes (subsampled to 1 million reads/cell and detected by more than 10 reads) covering the different biotypes, including lincRNAs (144 [139, 148; 95 % CI]), protein coding (5124 [4874, 5372; 95 % CI]) genes, and pseudogenes (132 [127, 137; 95 % CI]) (Figure 2C, 3). Of note, antisense genes are the only biotype for which the total RNA protocol detects fewer genes (62 [59-64; 95 % CI]), likely because of the unstranded nature of the polyA[+] RNA libraries, which results in erroneous quantification of sense/antisense overlapping genes (Supplementary Figure 10). Considering both polyA[+] RNA-seq and total RNA-seq data, 3978 different antisense-sense relationships with an overlap of more than 200 nucleotides were detected with expression of the sense or antisense gene in at least one cell. These loci are prone to erroneous quantification. Quantification of the stranded SMARTer data in an unstranded way shows that 42.1 % (median of 180 of the 428 detected antisense genes per cell) of the detected antisense genes (in 6 random cells) are receiving counts, while they have zero counts when properly treated as stranded data; further, 10.1 % of the antisense genes detected in both analyses display fold change differences larger than 2 (Supplemental Figure 11). Most of these genes with fold change differences (87.0 %) are more abundant in the unstranded analysis compared to the stranded analysis, explained by the fact that these antisense genes are consuming counts from the sense gene. LincRNAs, antisense genes and pseudogenes are clearly expressed in fewer cells compared to protein coding genes. We hypothesize that low abundant genes might be

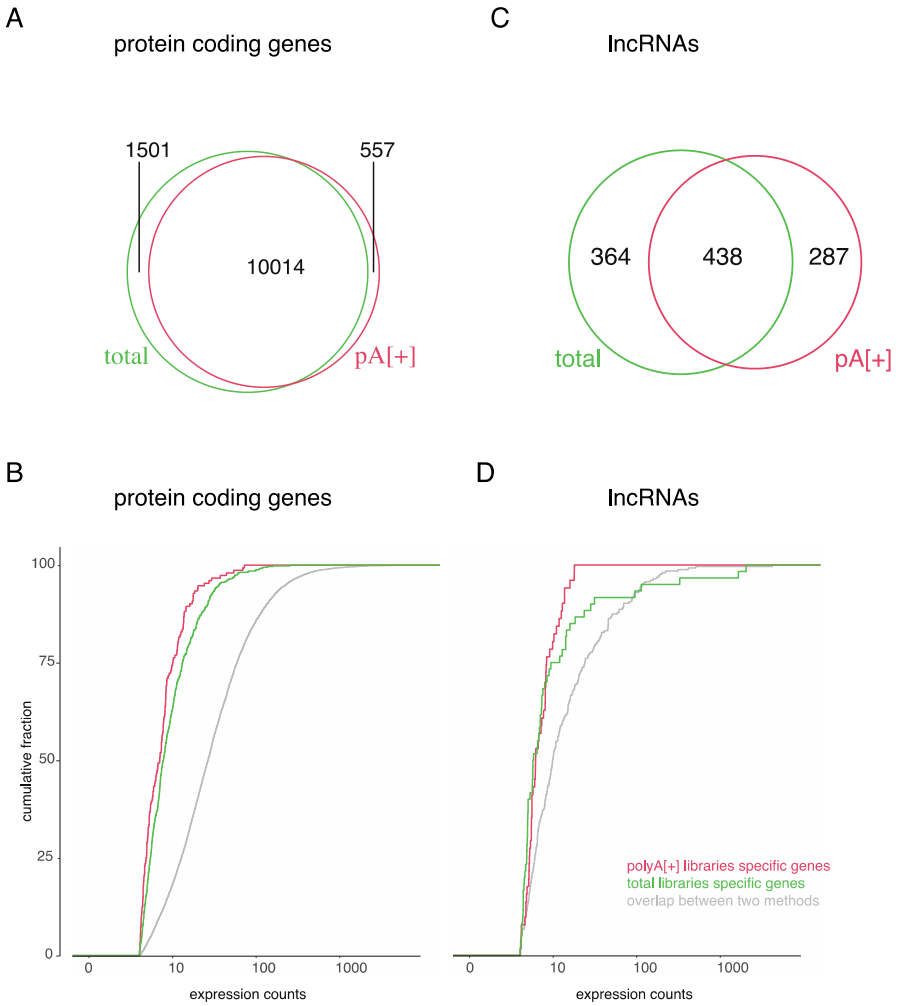


Figure 5: while most protein coding genes are commonly detected, lincRNAs appear more method specific. (A) Overlap between protein coding genes detected in polyA[+] (1 million reads) and total RNA (1 million reads) libraries. (B) Expression counts for protein coding genes detected in only polyA[+] libraries (red), only total RNA libraries (green) or both (gray). (C) Overlap between lncRNAs detected in polyA[+] (1 million reads) and total RNA (1 million reads) libraries. (D) Expression counts for lncRNAs detected in only polyA[+] libraries (red), only total RNA libraries (green) or both (gray).

missed because of sampling bias during the sequencing workflow or that lincRNAs, often low abundant in nature, are expressed under specific conditions or stimuli (Figure 4) (46). As expected, increasing the number of reads (up to 4 or 8 million) in the total RNA library protocol results in the detection of a higher number of genes. We observed no saturation when generating 8 million reads per cell, suggesting that deeper sequencing could yield even more detected genes (Figure 3). The overlap between protein coding genes detected in the polyA[+] and total RNA libraries (subsampled for 1 million reads/cell and mean expression of at least 1 read over all cells) (Figure 5A) is high. Genes detected in only one of the library types are generally lower abundant compared to genes detected with both methods (Figure 5B). In contrast to protein coding genes, the overlap for lincRNAs between the methods is much smaller (Figure 5C). Importantly, a significant fraction of the total RNA-seq specific lncRNAs display a high expression, thus possibly representing functionally important RNAs (Figure 5D). LincRNA *RMRP* is one of the most abundant lincRNAs that is solely detected by our novel single cell total RNA-seq workflow. This gene is known to be 3' non-adenylated and is the first known RNA encoded by a single-copy nuclear gene imported into mitochondria (47, 48). As only a subset of the lincRNAs and antisense genes are currently annotated in Ensembl, we also quantified our libraries with the LNCipedia transcriptome (the most comprehensive human

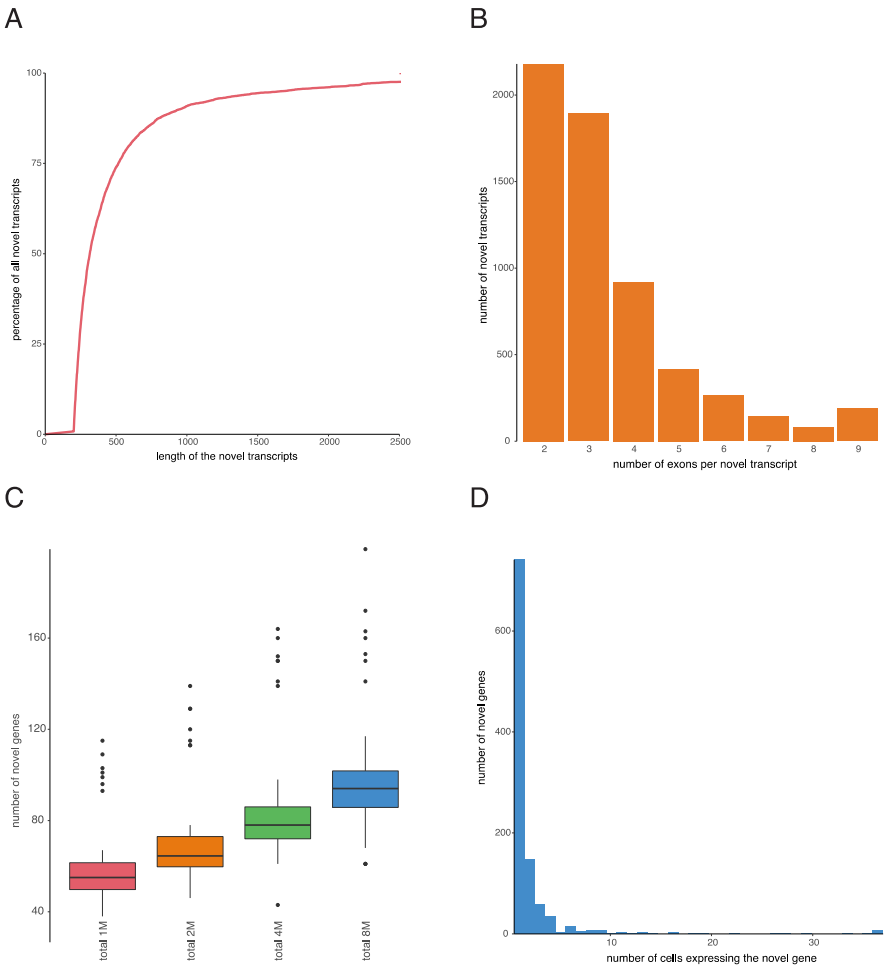


Figure 6: total RNA libraries enable assembly of single cell transcriptomes. A) Transcripts were filtered at a length of 200 nt. The remaining transcripts have a mean length of 537 nt. B) Transcripts were required to have at least two exons. The remaining transcripts are on average 3.4 exons long. C) All novel genes were quantified on subsampled data (1, 4 or 8 million single-end reads per cell). Genes with at least 1 count were retained. D) While some novel genes are expressed in all cells, most novel genes are detected in only 1 cell.

resource of both antisense and lincRNA genes, further referred to as lncRNAs). While the number of detected lncRNAs is slightly lower in the total RNA-seq libraries if an equal number of reads (1 million) is used, each library type contains a certain proportion of unique lncRNAs (Supplementary Figure 12). LNCipedia is likely biased towards medium-to-high abundant polyadenylated lncRNAs.

SMARTer single cell total RNA sequencing detects circular RNAs and novel genes

In addition to linear RNA biotypes, we tested whether the single cell total RNA-seq protocol is able to quantify circRNAs as this class of non-coding RNAs lacks a polyA-tail and in principle can only be detected using unbiased total RNA-seq. With a requirement of at least two unique back-spliced junction reads, 537 circRNAs were identified derived from 460 host genes (Supplementary Table 2, online available). The majority of the circRNAs were found in fewer than 3 out of 64 cells, with only 14 circRNAs detected in at least 4 cells. Gene Ontology analysis for molecular functions and biological processes was performed on the circRNA host genes from both treated and untreated cells. A significant enrichment of TP53 binding, TP53 pathway, cell cycle, and chromosome organization suggests that the identified circRNAs may play a role in these biological functions.

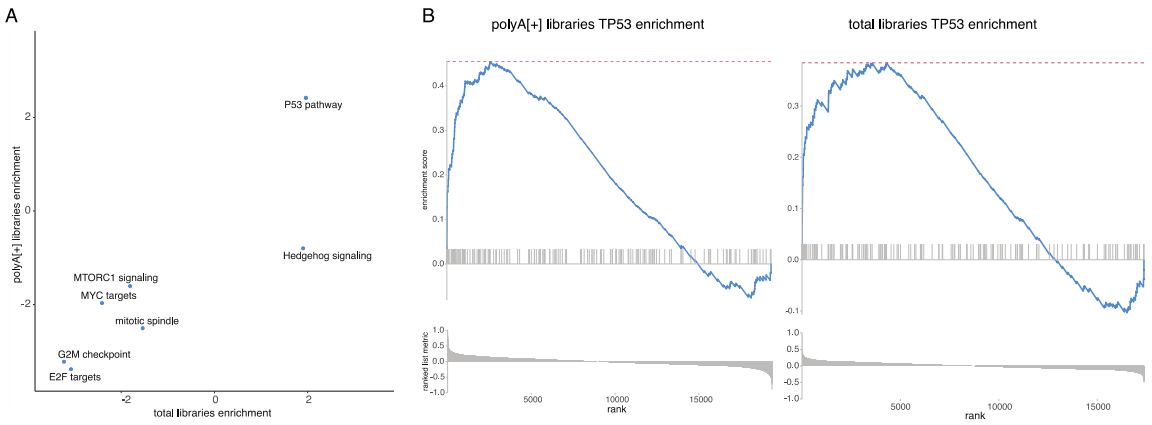


Figure 7: pathway analysis for polyA[+] RNA and total RNA libraries is similar. A) Gene set enrichment analysis for all hallmark pathways resulted in the same significant (p_{adj} < 0.05) pathway predictions. B) The TP53 pathway is, as expected, enriched in both library prep methods.

In the single cell total RNA libraries, the fraction of intergenic reads (relative to existing Ensembl and LNCipedia annotation) is high, suggesting that these reads originate from novel unannotated transcripts. To validate this hypothesis, we generated genome and transcriptome guided transcriptome assembly of the paired-end single cell total RNA-seq data resulting in 5360 novel, multi-exonic genes. The novel transcripts have a median length of 317 nucleotides (Figure 6A) and consist on average of more than 3 exons (Figure 6B). Quantification of this novel transcriptome using the single-end data subsampled at 1 million reads per cell resulted in a median number of 59 novel genes per cell [55 - 63; 95 % CI] (Figure 6C). Of note, most novel genes are expressed in only one cell (Figure 6D).

SMARTer single cell total RNA profiles reflect the biological signal

To assess whether the single cell total RNA-seq protocol is also able to reveal known biological signal, we performed differential expression analysis using DESeq2 combined with the Zinger method coping with zero inflated data. Based on the ranking obtained by the DESeq2 test statistic, gene set enrichment analysis using the hallmark gene sets was performed. Firstly, the same gene sets are significantly enriched in both library preparation protocols (Figure 7A); secondly, TP53 target genes are - as expected - the most significantly enriched gene set (Figure 7B), confirming that the biological signal is recapitulated through single cell total RNA-seq analyses.

SMARTer single cell total RNA sequencing of FACS sorted cells in microplates

To demonstrate that our novel single cell total RNA seq method also efficiently works on FACS sorted cells in microplates, we processed A375 and Jurkat sorted cells. In parallel, Takara's single cell purposed SMART-Seq Stranded Kit (used in all our previous experiments) was also tested on these cells (Figure 1). Equally low amounts of ribosomal cDNA were sequenced using both reagent kits, i.e. 1.46 % [0.77, 2,15; 95 % CI] and 0.66 % [0.48, 0.85; 95 % CI] for the A375 cells and 1.17 % [1.05, 1.29; 95 % CI] and 0.94 % [0.80, 1.09; 95 % CI] for the Jurkat cells (Figure 8, Supplemental Figure 13). Similar to the total RNA seq libraries generated on the C1 system, we analysed the number of reads assigned to intron, exon and intergenic regions and the read fraction for all RNA biotypes. The microplate sorted single cell data was very comparable to the C1 data (Figure 8, Supplemental Figure 13).

DISCUSSION

In this study, we developed a single cell total RNA-seq method to sequence full transcripts from single cells in an essentially unbiased manner. To demonstrate the performance of the method, we applied single cell total RNA-seq in four experiments on five different cancer cell

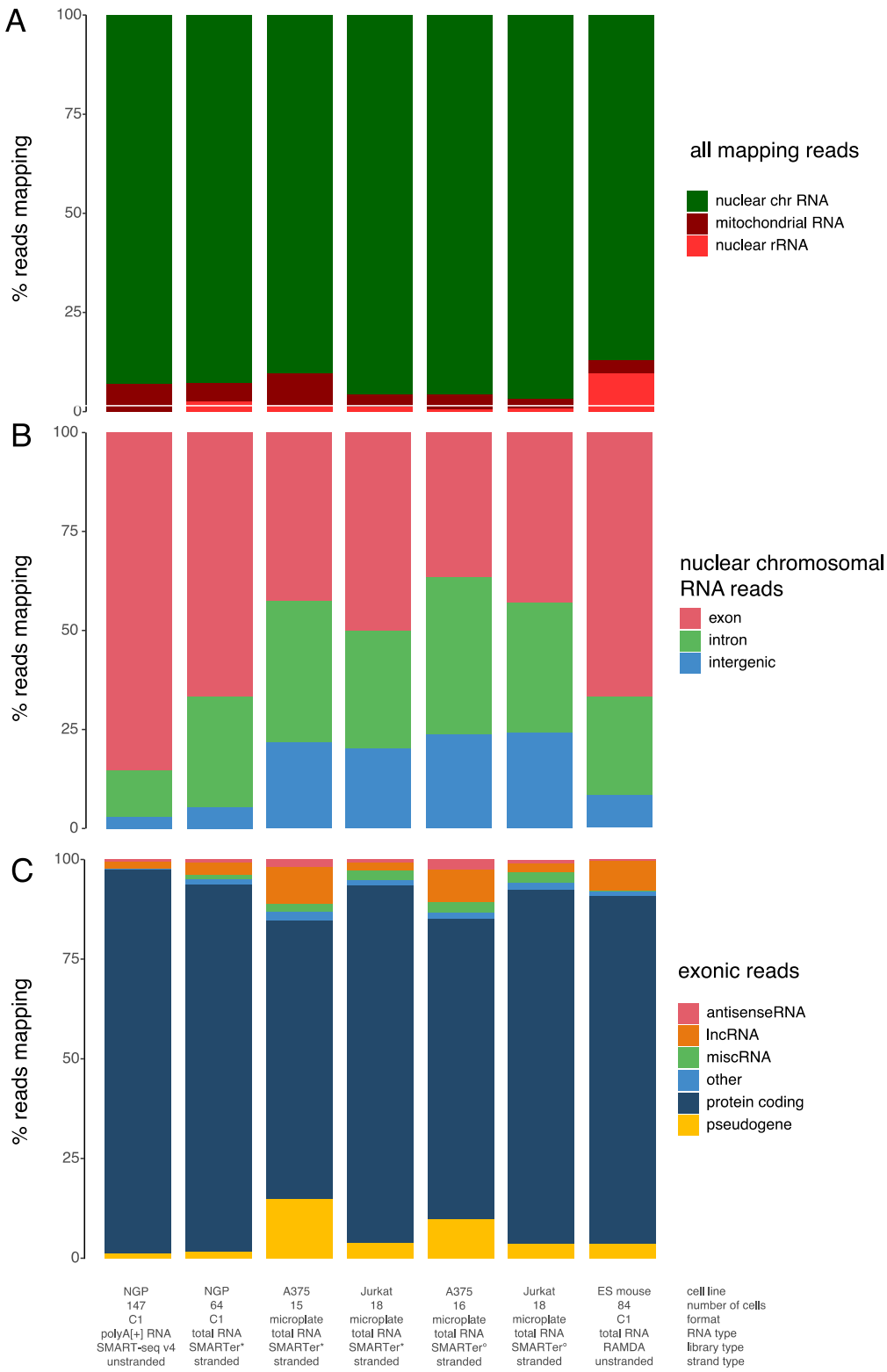


Figure 8: mean read distributions are similar for total RNA sequencing libraries generated on C1 or in microplates. A) Mean percentage of reads derived from nuclear RNA, mitochondrial RNA and ribosomal RNA quantified with STAR. Single cell total RNA libraries of the FACS sorted cells were generated using 2 different reagent kits (#634413, denoted with * and #634444, denoted with °). B) Mean percentage of reads originating from nuclear chromosomes derived from exonic, intronic and intergenic regions quantified with STAR. C) Mean percentage of exonic reads attributed to the different RNA biotypes quantified with Kallisto.

lines, of which three undergoing a specific perturbation. In parallel, we also performed single cell polyA[+] RNA-seq on three cell lines using the well-established Smart-seq v4 method

(6, 18). As in any genomics study, the experimental set-up may suffer from confounding factors, such as variations in cell cycle states of the cells and batch effects of single cell capture and sequencing, masking real biological differences. In two of the four experiments, we carefully controlled all these experimental biases. The cell cycle bias was minimized by cell cycle synchronization using serum starvation. We also avoided potential cell selection bias by capturing differentially labeled treated and untreated cells on the same chip (44, 45). Finally, sequencing bias was minimized by sequencing both polyA[+] and total RNA libraries on the same Illumina flow cells.

The single cell total RNA-seq method has some distinctive advantages compared to other methods. First, in any total RNA-seq library, depletion of rRNA is essential as this makes up the bulk of the total RNA mass. Depletion of rRNA from single cells prior to cDNA synthesis is technically very difficult. Here, we used ribosomal cDNA specific removal probes, resulting in less than 3 % of ribosomal reads per single cell library. This highly efficient rRNA depletion step is a major improvement compared to RAMDA-seq, where 10-35 % of the reads map to rRNA (22). Second, given the stranded nature of the single cell total RNA sequencing data, quantification of antisense genes is accurate, which is not possible when using unstranded data. In contrast to the three existing single cell total RNA-seq methods, our method uniquely combines these two features that are highly desirable for total RNA-sequencing (22, 29, 30). Third, as expected, our single cell total RNA libraries contain substantially more intronic reads compared to polyA[+] RNA libraries (49, 50). Such intronic reads can be used to detect changes in nascent transcription, whereby the difference in exonic and intronic reads provides insights in post-transcriptional regulation (51). As such, we believe that our method may be particularly well suited for “RNA velocity analysis” of single cells (52). Fourth, the single cell total RNA-seq workflow presented in this paper detects relatively more protein coding genes, pseudogenes, lincRNAs and miscellaneous RNA (miscRNA) compared to single cell polyA[+] RNA libraries, when corrected for equal sequencing depth. While the number of detected genes increases with sequencing depth, there seems to be no plateau yet at 8 million reads, suggesting that further increasing the sequencing depth, could enable low abundant gene detection. Fifth, our method also detects non-polyadenylated RNA molecules, such as histone genes, lncRNAs and circRNAs. In the NGP dataset, 537 circRNAs were detected using reads with evidence for back splicing. In order to detect more circRNAs in an individual cell, a higher sequencing depth is required or libraries should be enriched for circRNAs by selectively removing linear RNA by exonuclease treatment prior to library prep and sequencing (28, 29). Sixth, the data enables reference guided transcriptome assembly, resulting in the detection of 5360 novel genes. Finally, differential gene expression analysis and gene set enrichment of NGP cells treated with nutlin-3 confirmed activation of the TP53 pathway at the transcriptional level.

One limitation of the implementation of the single cell total RNA library preparation method on the C1 instrument is the relatively low throughput, as maximally 96 cells are simultaneously captured. In contrast, current droplet based single cell methods capture thousands of individual cells, but these systems are limited to 3’ end sequencing of polyadenylated RNA, preventing quantification of splice variants and non-polyadenylated transcripts. To enable the analysis of higher cell numbers, we demonstrated that the method works equally well on FACS sorted cells in microplates. By using FACS sorted cells the throughput can be increased and no specialized devices, such as the C1, are required. Finally, an advantage of our total RNA-seq protocol on both C1 and in microplates is that single-end sequencing is sufficient while more expensive paired-end sequencing is required for most droplet-based methods. We advice to use the single cell total RNA-seq method rather than polyA[+] methods if it is desired to study non-

polyadenylated RNA molecules such as lncRNAs or circRNAs, if stranded-specific data is a must and if full transcript sequencing is priority (e.g. analysis of alternative splicing, RNA editing or somatic mutations).

AVAILABILITY

The SMARTer single cell total RNA sequencing script is deposited in Script Hub (Fluidigm).

CONFLICT OF INTEREST

N.B. and S.L. are employees of Takara Bio USA whose reagent kits are used in this study. K.L. is employee of Fluidigm whose C1 instrument was used for single cell RNA seq library preparation.

REFERENCES

1. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A., *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. **6**.
2. Junker,J.P. and van Oudenaarden,A. (2014) Every Cell Is Special: Genome-wide Studies Add a New Dimension to Single-Cell Biology. *Cell*, **157**, 8–11.
3. Islam,S., Kjällquist,U., Moliner,A., Zajac,P., Fan,J.-B., Lönnerberg,P. and Linnarsson,S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, **21**, 1160–7.
4. Soumillon,M., Cacchiarelli,D., Semrau,S., van Oudenaarden,A. and Mikkelsen,T.S. (2014) Characterization of directed differentiation by high-throughput single-cell RNA-Seq - SI. *bioRxiv*, 10.1101/003236.
5. Picelli,S., Faridani,O.R., Björklund,Å.K., Winberg,G., Sagasser,S. and Sandberg,R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.
6. Picelli,S., Björklund,Å.K., Faridani,O.R., Sagasser,S., Winberg,G. and Sandberg,R. smart-seq2 for sensitive full-length transcriptome profiling in single cells. 10.1038/nMeth.2639.
7. Hashimshony,T., Wagner,F., Sher,N. and Yanai,I. (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, **2**, 666–73.
8. Hashimshony,T., Senderovich,N., Avital,G., Klochendler,A., de Leeuw,Y., Anavy,L., Gennert,D., Li,S., Livak,K.J., Rozenblatt-Rosen,O., *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, **17**, 77.
9. Jaitin,D.A., Kenigsberg,E., Keren-Shaul,H., Elefant,N., Paul,F., Zaretsky,I., Mildner,A., Cohen,N., Jung,S., Tanay,A., *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, **343**, 776–9.
10. Zheng,S., Papalexi,E., Butler,A., Stephenson,W. and Satija,R. (2018) Molecular transitions in early progenitors during human cord blood hematopoiesis. *Mol. Syst. Biol.*, **14**, e8041.
11. Rosenberg,A.B., Roco,C.M., Muscat,R.A., Kuchina,A., Sample,P., Yao,Z., Graybuck,L.T., Peeler,D.J., Mukherjee,S., Chen,W., *et al.* (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, **360**, 176–182.
12. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M., *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.
13. Zilionis,R., Nainys,J., Veres,A., Savova,V., Zemmour,D., Klein,A.M. and Mazutis,L. (2016) Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.*, **12**, 44–73.
14. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J., *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
15. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M.,

et al. (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.

16. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A., *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.
17. Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.
18. Fish,R.N., Bostick,M., Lehman,A. and Farmer,A. (2016) Transcriptome Analysis at the Single-Cell Level Using SMART Technology. In *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc., Hoboken, NJ, USA, Vol. 116, p. 4.26.1-4.26.24.
19. Sasagawa,Y., Nikaido,I., Hayashi,T., Danno,H., Uno,K.D., Imai,T. and Ueda,H.R. (2013) Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.*, **14**, 3097.
20. Fan,H.C., Fu,G.K. and Fodor,S.P.A. (2015) Combinatorial labeling of single cells for gene expression cytometry. *Science (80-.).*, **347**.
21. Fan,X., Zhang,X., Wu,X., Guo,H., Hu,Y., Tang,F. and Huang,Y. (2011) Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. 10.1186/s13059-015-0706-1.
22. Hayashi,T., Ozaki,H., Sasagawa,Y., Umeda,M., Danno,H. and Nikaido,I. (2018) Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.*, **9**, 619.
23. Sheng,K., Cao,W., Niu,Y., Deng,Q. and Zong,C. (2017) Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods*, **14**.
24. Kashima,Y., Suzuki,A., Liu,Y., Hosokawa,M., Matsunaga,H., Shirai,M., Arikawa,K., Sugano,S., Kohno,T., Takeyama,H., *et al.* (2018) Combinatory use of distinct single-cell RNA-seq analytical platforms reveals the heterogeneous transcriptome response. *Sci. Rep.*, **8**, 3482.
25. Hashimshony,T., Senderovich,N., Avital,G., Klochendler,A., de Leeuw,Y., Anavy,L., Gennert,D., Li,S., Livak,K.J., Rozenblatt-Rosen,O., *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, **17**, 77.
26. Yang,L., Duff,M.O., Graveley,B.R., Carmichael,G.G. and Chen,L.-L. Genomewide characterization of non-polyadenylated RNAs. 10.1186/gb-2011-12-2-r16.
27. Lai,F., Gardini,A., Zhang,A. and Shiekhattar,R. (2015) Integrator mediates the biogenesis of enhancer RNAs. *Nature*, **525**, 399–403.
28. Jeck,W.R. and Sharpless,N.E. (2014) Detecting and characterizing circular RNAs. *Nat. Biotechnol.*, **32**, 453–61.
29. Fan,X., Zhang,X., Wu,X., Guo,H., Hu,Y., Tang,F. and Huang,Y. (2015) Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.*, **16**, 148.
30. Sheng,K., Cao,W., Niu,Y., Deng,Q. and Zong,C. (2017) Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods*, **14**, 267–270.
31. Balbin,O.A., Malik,R., Dhanasekaran,S.M., Prensner,J.R., Cao,X., Wu,Y.-M., Robinson,D., Wang,R., Chen,G., Beer,D.G., *et al.* (2015) The landscape of antisense gene expression in human cancers. *Genome Res.*, **25**, 1068–1079.
32. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
33. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
34. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
35. Zerbino,D.R., Achuthan,P., Akanni,W.,

Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.

36. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
37. Volders,P.J., Verheggen,K., Menschaert,G., Vandepoele,K., Martens,L., Vandesompele,J. and Mestdagh,P. (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.*, **43**, 4363–4364.
38. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
39. Van den Berge,K., Perraudeau,F., Soneson,C., Love,M.I., Risso,D., Vert,J.-P., Robinson,M.D., Dudoit,S. and Clement,L. (2018) Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.*, **19**, 24.
40. Memczak,S., Jens,M., Elefsinioti,A., Torti,F., Krueger,J., Rybak,A., Maier,L., Mackowiak,S.D., Gregersen,L.H., Munschauer,M., *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**, 333–338.
41. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
42. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
43. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
44. Buettner,F., Natarajan,K.N., Casale,F.P., Proserpio,V., Scialdone,A., Theis,F.J., Teichmann,S.A., Marioni,J.C. and Stegle,O. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.
45. Tung,P.-Y., Blischak,J.D., Hsiao,C.J., Knowles,D.A., Burnett,J.E., Pritchard,J.K. and Gilad,Y. (2017) Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.*, **7**, 39921.
46. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
47. Livyatan,I., Harikumar,A., Nissim-Rafinia,M., Duttagupta,R., Gingeras,T.R. and Meshorer,E. (2013) Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation. *Nucleic Acids Res.*, **41**, 6300–15.
48. Hsieh,C.L., Donlon,T.A., Darras,B.T., Chang,D.D., Topper,J.N., Clayton,D.A. and Francke,U. (1990) The gene for the RNA component of the mitochondrial RNA-processing endoribonuclease is located on human chromosome 9p and on mouse chromosome 4. *Genomics*, **6**, 540–4.
49. Ameur,A., Zaghlool,A., Halvardson,J., Wetterbom,A., Gyllensten,U., Cavelier,L. and Feuk,L. (2011) Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.*, **18**, 1435–1440.
50. Zhao,S., Zhang,Y., Gamini,R., Zhang,B. and von Schack,D. (2018) Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.*, **8**, 4781.
51. Gaidatzis,D., Burger,L., Florescu,M. and Stadler,M.B. (2015) Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.*, **33**, 722–729.
52. La Manno,G., Soldatov,R., Zeisel,A., Braun,E., Hochgerner,H., Petukhov,V., Lidschreiber,K., Kastriti,M.E., Lönnerberg,P., Furlan,A., *et al.* (2018) RNA velocity of single cells. *Nat*, **560**, 494–498.

SUPPLEMENTARY METHODS

Cell lines

The MYCN shRNA doxycycline inducible cell line SHSY5Y-MYCN-TR is a kind gift of prof. R. Versteeg (Amsterdam, the Netherlands). The neuroblastoma cell line SK-N-BE-2C is a kind gift of prof. John Lunec (Newcastle, United Kingdom). Cells were maintained in RPMI-1640 medium (Life Technologies, 52400-025) supplemented with 10 % fetal bovine serum, 1 % of L-glutamine (Life Technologies, 15140-148) and 1 % penicillin/streptomycin (Life Technologies, 15160-047) (referred to as complete medium) at 37 °C in a 5 % CO_2 atmosphere. Short tandem repeat (STR) genotyping was used to validate cell line authenticity prior to performing the described experiments and mycoplasma testing was done on a monthly basis using the MycoAlert Mycoplasma Detection Kit (Lonza, T07-318), according to manufacturer's instructions.

Cell cycle synchronization and chemical or genetic perturbation of SK-N-BE-2C and SHSY5Y-MYCN-TR cells

SHSY5Y-MYCN-TR cells were seeded in a T75 culture flask in complete medium. After 24 hour, cells were refreshed with complete medium with either 1 µg/ml doxycycline (sigma Aldrich, D9891-1G, dissolved in ethanol) or vehicle. SK-N-BE-2C cells were synchronized as described for NGP cells. 24 hour after serum starvation, SK-N-BE-2C cells were treated with either 1 µM of JQ1 (PBS Bioscience, 27402, dissolved in DMSO) or vehicle. Cells were trypsinized 24 hour post treatment and harvested for single cell analysis and bulk RNA isolation.

Single cell total RNA sequencing of SHSY5Y-MYCN-TR and SK-N-BE-2C cells

Doxycycline treated SHSY5Y-MYCN-TR cells were stained with 4 µM cell tracker as described for NGP cells. An equal number of stained (doxycycline treated) and non-stained (vehicle treated) cells were mixed and diluted to 300,000 cells per ml. Suspension buffer (Fluidigm, 100-6201) was added to the cells in a 7:3 ratio and 6 µl was loaded on a primed C1 Single-Cell Open App IFC (Fluidigm, 1**00-8134**) designed for medium-sized cells (10-17 µm). Cells were captured and cDNA synthesized as described for the NGP cells with minor modifications. The reagents of the SMARTer Stranded Total RNA-Seq Kit v1 - Pico Input Mammalian (Pico v1, Takara, 635007) were used. The lysis and fragmentation were performed by incubating the cells for 3 minutes at 94 °C and 2 minutes at 10 °C and by using 9 instead of 11 PCR cycles in PCR1. Following the initial cDNA amplification, all cells were pooled in a tube using 4 µl of cDNA per cell. Next steps of the library prep were performed according to manufacturer's instructions with minor modifications. 500 µl of 80 % ethanol was used to wash the beads. 15 PCR cycles were used for PCR2. Library quality was determined on the Bioanalyzer (Agilent).

JQ1 treated cells were stained with cell tracker as described for NGP cells. An equal number of stained (JQ1 treated) and non-stained (vehicle treated) cells were mixed and diluted to 300,000 cells per ml. Suspension buffer was added to the cells in a 8:2 ratio and 6 µl of this mix was loaded on a primed C1 Single-Cell Open App IFC designed for medium-sized cells (10-17 µm). Cells were captured and cDNA synthesized as described for the SHSY5Y-MYCN-TR cells by using the reagents of the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Takara, 634413). One microliter of the ERCC spike-in mix (Ambion, 4456740) was diluted in 999 µl loading buffer to get a 1/1000 dilution of the ERCC spikes. One microliter of this dilution was added to the 20 µl lysis mix. Following the initial cDNA amplification, all cells were pooled in a tube using 5 µl of cDNA per cell. Library prep was performed as described for SHSY5Y-MYCN-TR cells.

Single cell polyA[+] RNA sequencing of treated SHSY5Y-MYCN-TR and SK-N-BE-2C cells

Doxycycline treated SHSY5Y-MYCN-TR cells were diluted to 300,000 cells per ml. Suspension buffer was added to the cells in a 7:3 ratio and 6 µl of this mix of was loaded on a primed C1 Single-Cell Auto Prep Array for mRNA Seq (Fluidigm, 100-6041) designed for medium-sized cells (10-17 µm) (separate IFC for treated and untreated cells). Single cell polyA[+] RNA-sequencing was performed on the C1 using the SMART-Seq v1 Ultra Low Input RNA Kit for the Fluidigm C1 System (Takara, 634833) according to manufacturer's instructions. ArrayControl RNA spikes (Ambion, AM1780) were added as described in the manual. The concentration was measured using the quantifluor dsDNA kit (Promega, E2670) and glomax (Promega) according to manufacturer's instructions. The samples were 1/3 diluted in C1 harvest reagent (Fluidigm). Next, library preparation was performed using the Nextera XT library prep kit (Illumina, FC-131-1096) according to manufacturer's instructions, followed by quality control on the Bioanalyzer.

JQ1 treated cells were stained with cell tracker as described for NGP cells. An equal number of stained (JQ1 treated) and non-stained (DMSO treated) cells were mixed and diluted to 300,000 cells per ml. Suspension buffer was added to the cells in a 7:2 ratio and 6 µl of this mix of was loaded on a primed C1 Single-Cell Auto Prep Array for mRNA Seq designed for medium-sized cells (10-17 µm). Single cell polyA[+] RNA-sequencing on the C1 was performed using the SMART-Seq v1 Ultra Low Input RNA Kit for the Fluidigm C1 System (Takara) according to manufacturer's instructions. One microliter of the ERCC spike-in mix was diluted in 999 µl loading buffer to get a 1/1000 dilution of the ERCC spikes. One microliter of this dilution was added to the 20 µl lysis mix. The quality of the cDNA was checked for 11 random single cells on the Bioanalyzer. The concentration was measured using the qubit dsDNA HS kit (Invitrogen) according to manufacturer's instructions. The samples were 1/4 diluted in C1 harvest reagent. Next, library preparation was performed using the Nextera XT library prep kit according to manufacturer's instructions, followed by quality control on the Bioanalyzer.

Library sequencing

For SHSY5Y-MYCN-TR, the polyA[+] and total RNA libraries were quantified using the KAPA library quantification kit (Roche). 1.5 pM of the total RNA library was paired-end sequenced on a NextSeq 500 (Illumina) with a read length of 36 bp and a total sequencing read depth of 347 million reads.

For the polyA[+] library, 1.2 pM of the library was paired-end sequenced on a NextSeq 500 with a read length of 75 bp and a total sequencing read depth of 250 million reads.

For SK-N-BE-2C, the polyA[+] and total RNA libraries were quantified using the KAPA library quantification kit (Roche) and libraries were diluted to 4 nM. The polyA[+] RNA library and total RNA library were pooled in a 1/2 ratio. 1.3 pM of the pooled library was single-end sequenced on a NextSeq 500 (Illumina) with a read length of 75 bp and a total sequencing read depth of 289 million reads, combining single cell polyA[+] and total RNA libraries to prevent inter-run bias.

Sequencing data quality control

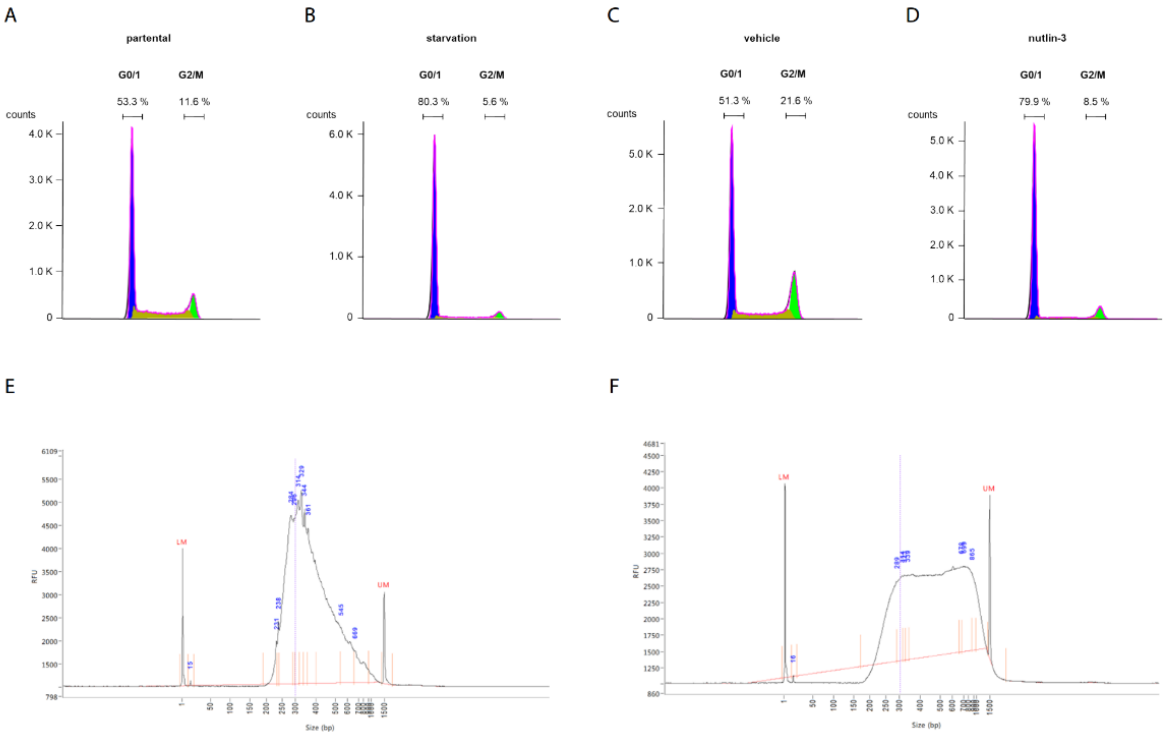
The paired-end sequencing output of the SHSY5Y-MYCN-TR cells were trimmed using cutadapt (v.1.16) (1) to remove 3 nucleotides of the 5' end of read 1. Remark that this changed to read 2 for the NGP version of the protocol. The SK-N-BE-2C libraries were single-end sequenced, so no trimming was needed. For both experiments, the quality of the data was assessed by mapping the reads using STAR (v.2.5.3) (2) on the hg38 genome including the full ribosomal DNA (45S, 5.8S and 5S) and mitochondrial DNA sequences. The parameters of STAR were set to retain only primary mapping reads. Using SAMtools (v.1.6) (3), reads mapping to the different nuclear chromosomes, mitochondrial DNA and rRNA were extracted and annotated as exonic, intronic or intergenic. Genes were quantified by Kallisto (v.0.43.1) (4) using both Ensembl (v.91) (5) extended with the ERCC spike sequences and LNCipedia

(v.5.0) (6). The strandedness of the total RNA-seq reads was taken into account by running the –rf-stranded mode.

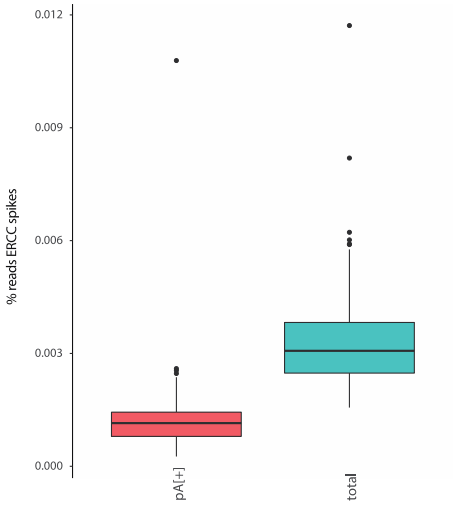
References

1. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
2. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
3. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
4. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
5. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
6. Volders,P.J., Verheggen,K., Menschaert,G., Vandepoele,K., Martens,L., Vandesompele,J. and Mestdagh,P. (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.*, **43**, 4363–4364.

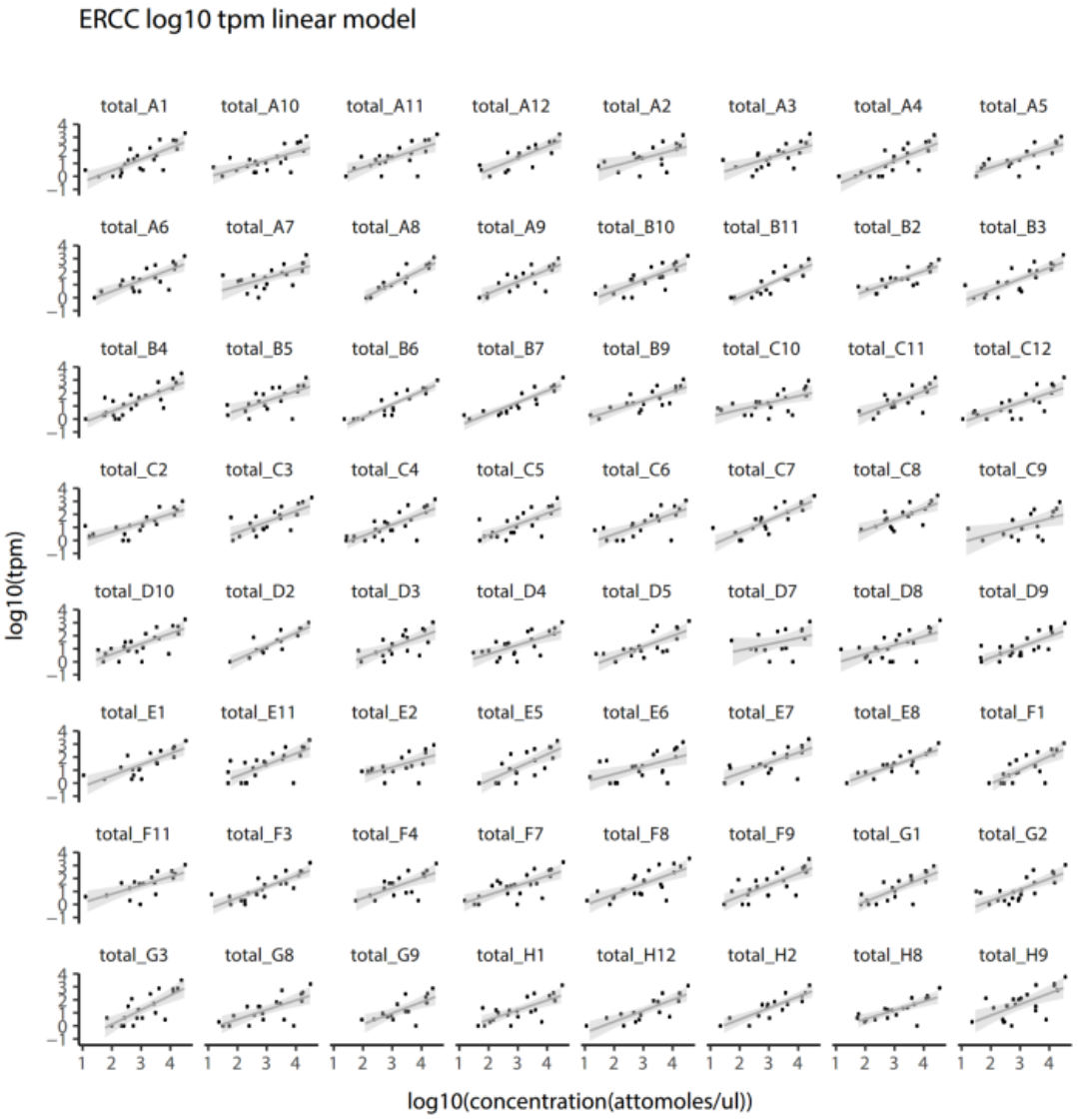
SUPPLEMENTARY FIGURES



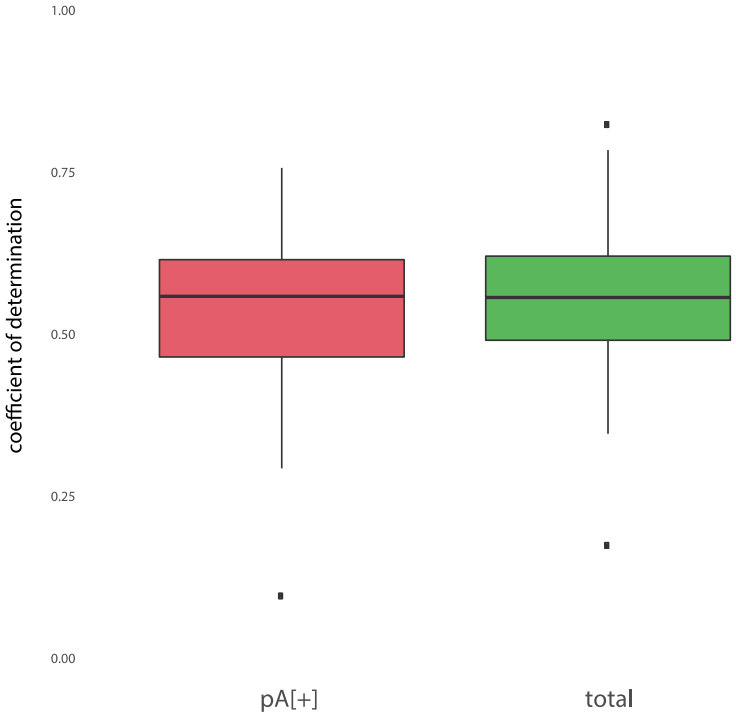
Supplementary Figure 1: cell cycle profiles for NGP cells maintained in RPMI-1640 medium supplemented with 10 % serum (A) or 0 % serum (B) for 24 hours. Cell cycle profiles for NGP cells maintained in 0 % serum for 24 hours and treated with vehicle (C) or nutlin-3 (D) for 24 hours. Fragment analyzer profiles of a NGP total RNA library (E) and polyA[+] RNA library (F).



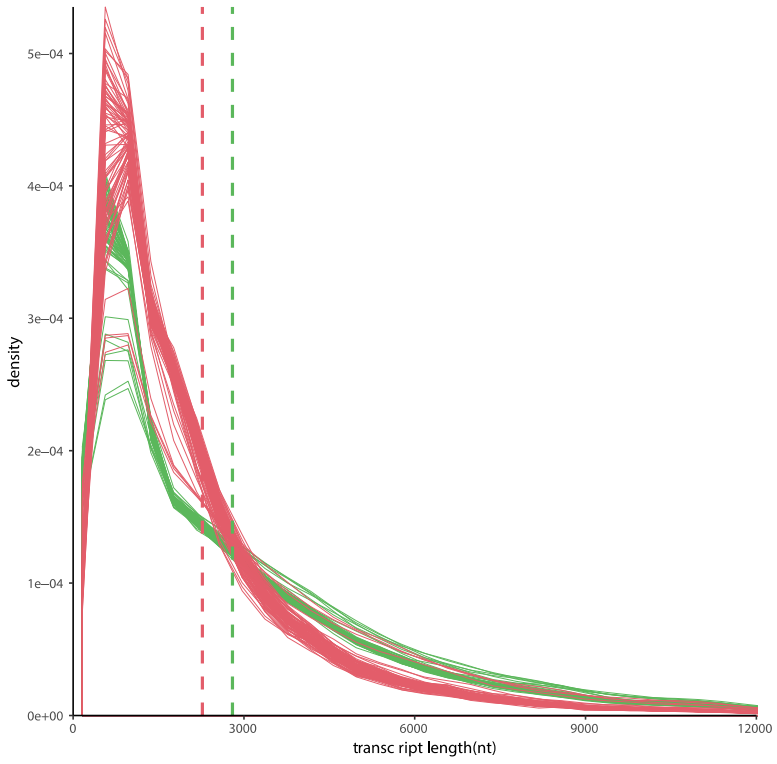
Supplementary Figure 2: the percentage of reads mapped on ERCC spikes is higher in the total RNA libraries compared to polyA[+] RNA libraries.



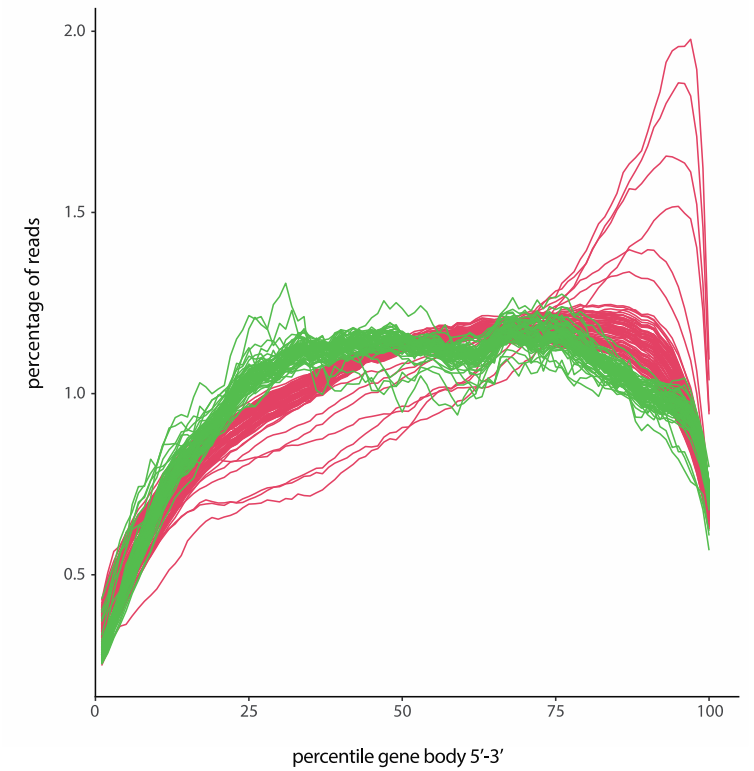
Supplementary Figure 3: linear modeling of ERCC spike abundance demonstrates quantitative performance.



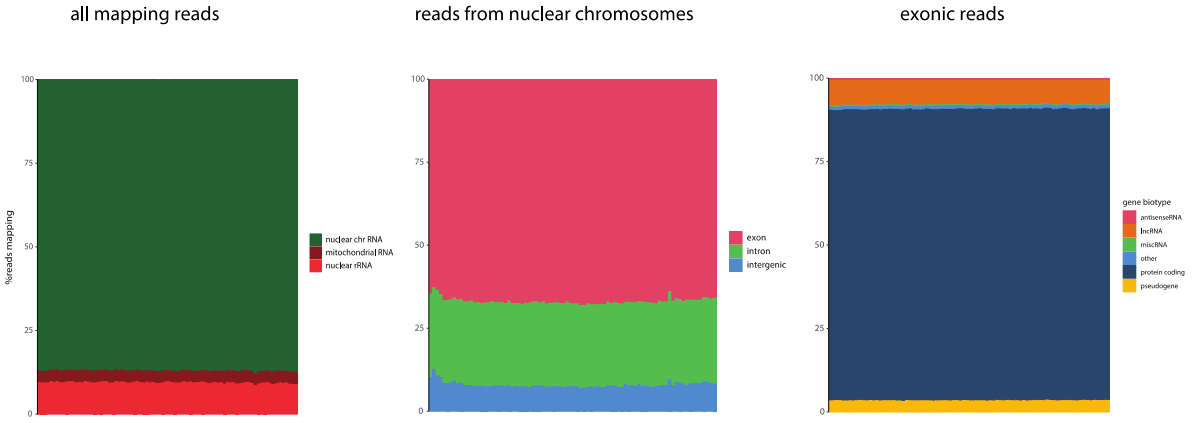
Supplementary Figure 4: the coefficients of determination obtained through linear regression of ERCC spikes are equal for polyA[+] and total RNA libraries.



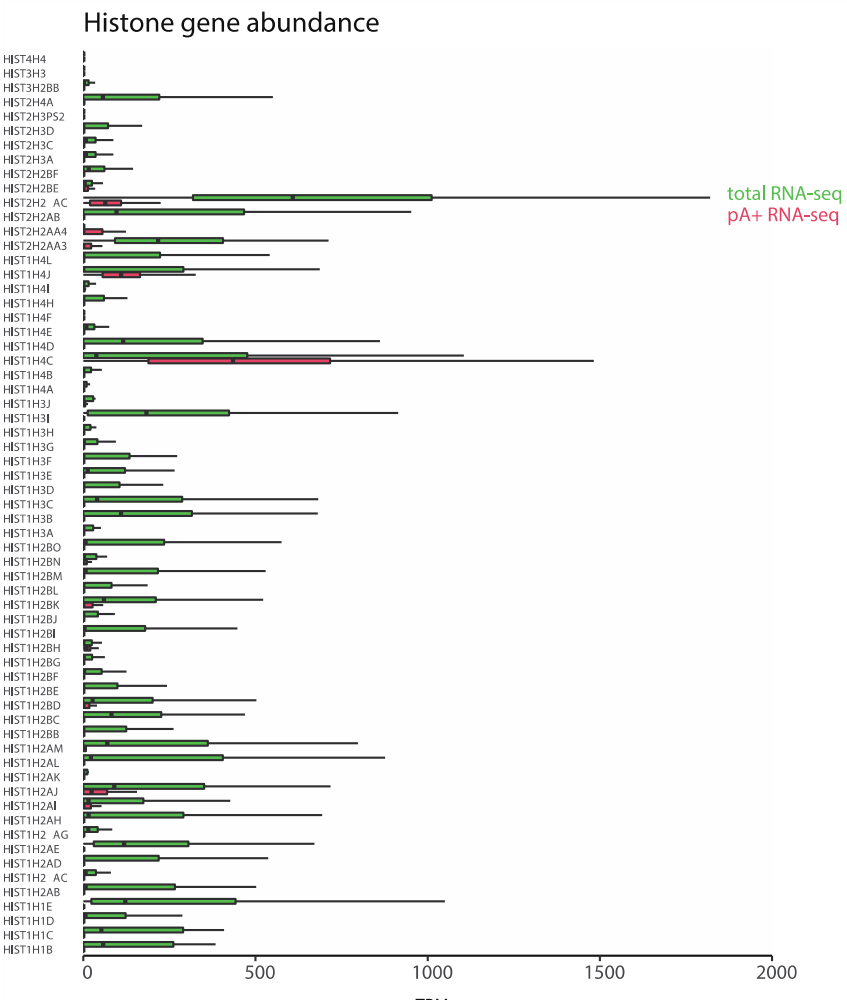
Supplementary Figure 5: the length distributions differ between polyA[+] RNA transcripts (red) and total RNA transcripts (green) per cell. PolyA[+] RNA libraries typically result in the detection of shorter transcripts.



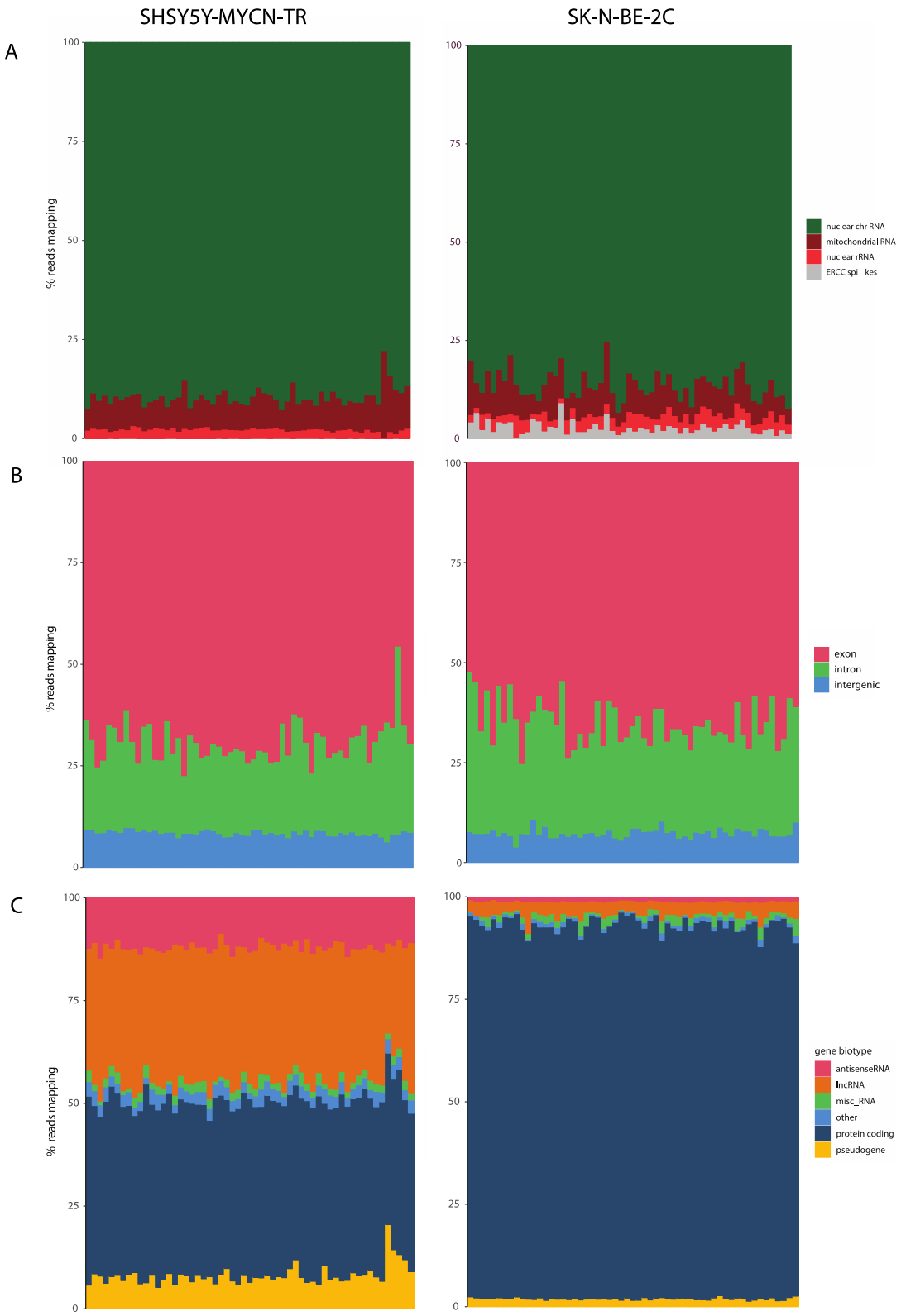
Supplementary Figure 6: the gene body coverage differs slightly between polyA[+] RNA libraries (red) and total RNA libraries (green). Equal gene body coverage would result in 1 % read fraction along the entire gene body. Both library types show some bias towards the 3' and 5' end.



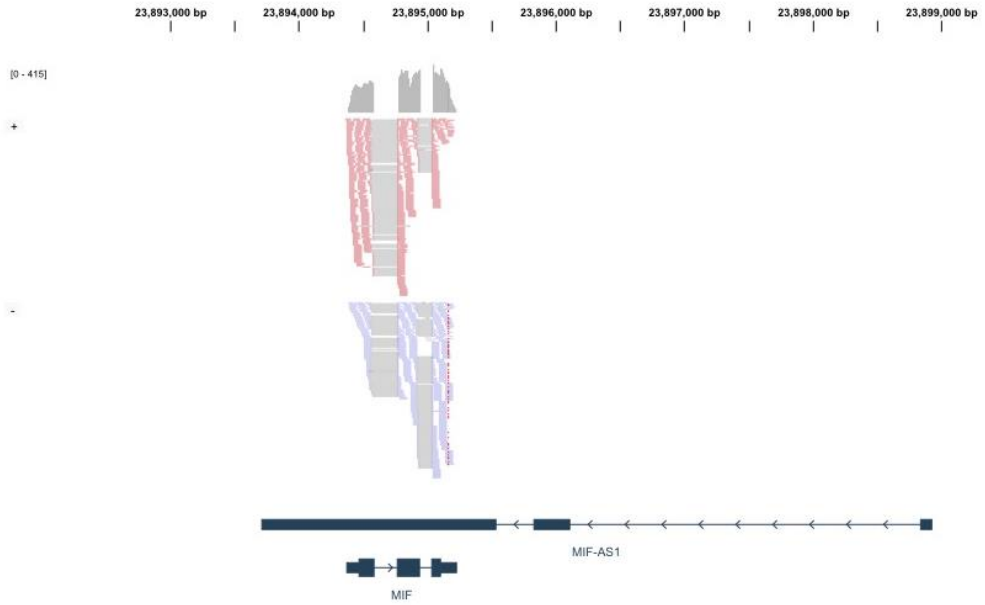
S**upplementary Figure 7: the QC results for RAMDA-seq libraries show higher rRNA percentages.** A) Percentage of reads derived from nuclear RNA, mitochondrial RNA and ribosomal RNA per cell quantified with STAR. B) Percentage of reads originating from nuclear chromosomes derived from exonic, intronic and intergenic regions per cell quantified with STAR. C) Percentage of exonic reads attributed to the different biotypes per cell quantified with Kallisto.



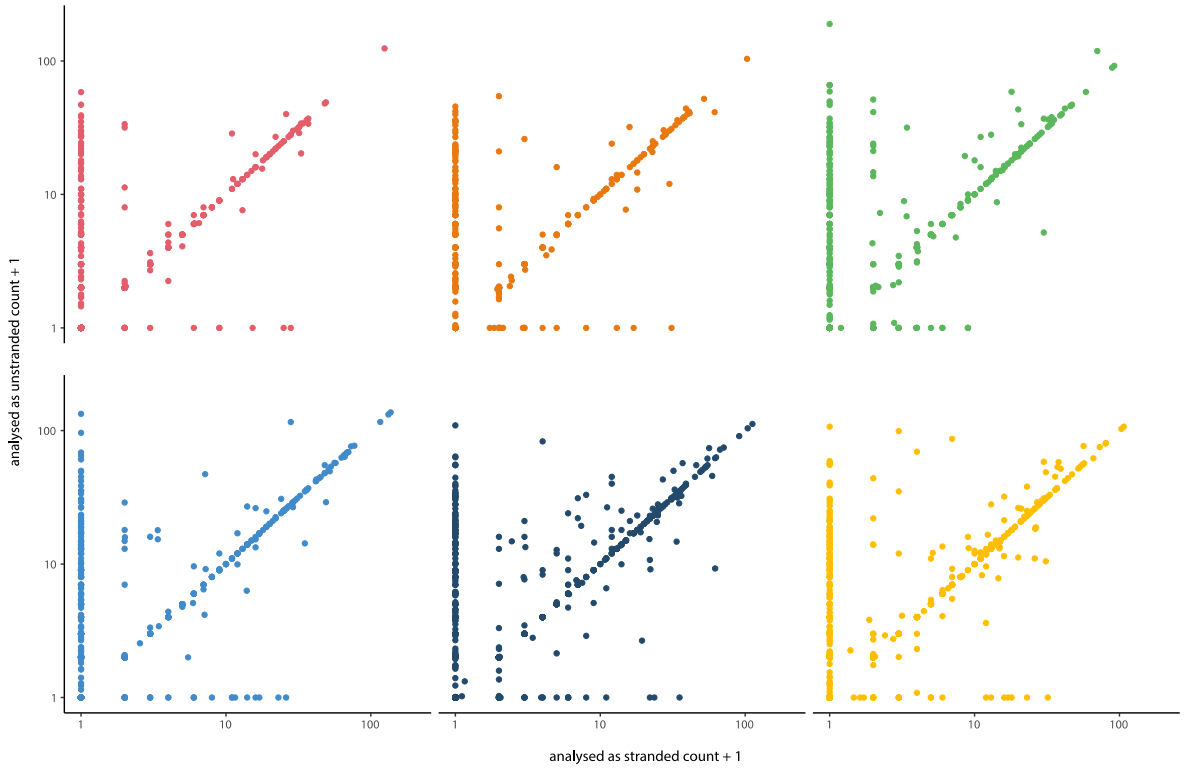
Supplementary Figure 8: TPM expression of histone genes, typically non-polyadenylated, shows that total RNA libraries (green) efficiently capture non-polyadenylated transcripts.



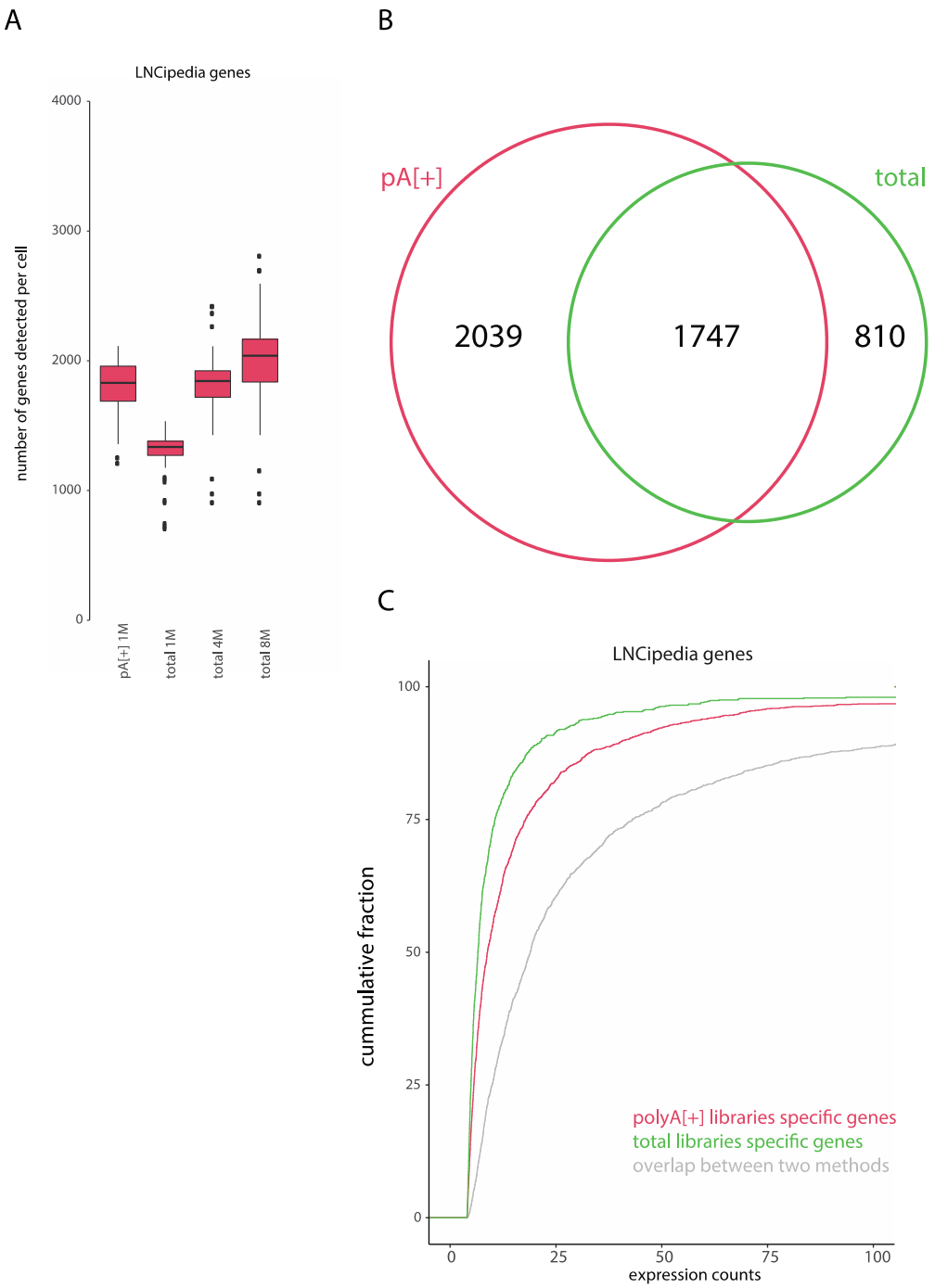
Supplementary Figure 9: the QC results for total RNA libraries of SHSY5Y-MYCN-TR and SK-N-BE-2C cells are similar to NGP cells. A) Percentage of reads derived from nuclear RNA, mitochondrial RNA and ribosomal RNA per cell quantified with STAR. B) Percentage of reads originating from nuclear chromosomes derived from exonic, intronic and intergenic regions per cell quantified with STAR. C) Percentage of exonic reads attributed to the different biotypes per cell quantified with Kallisto.



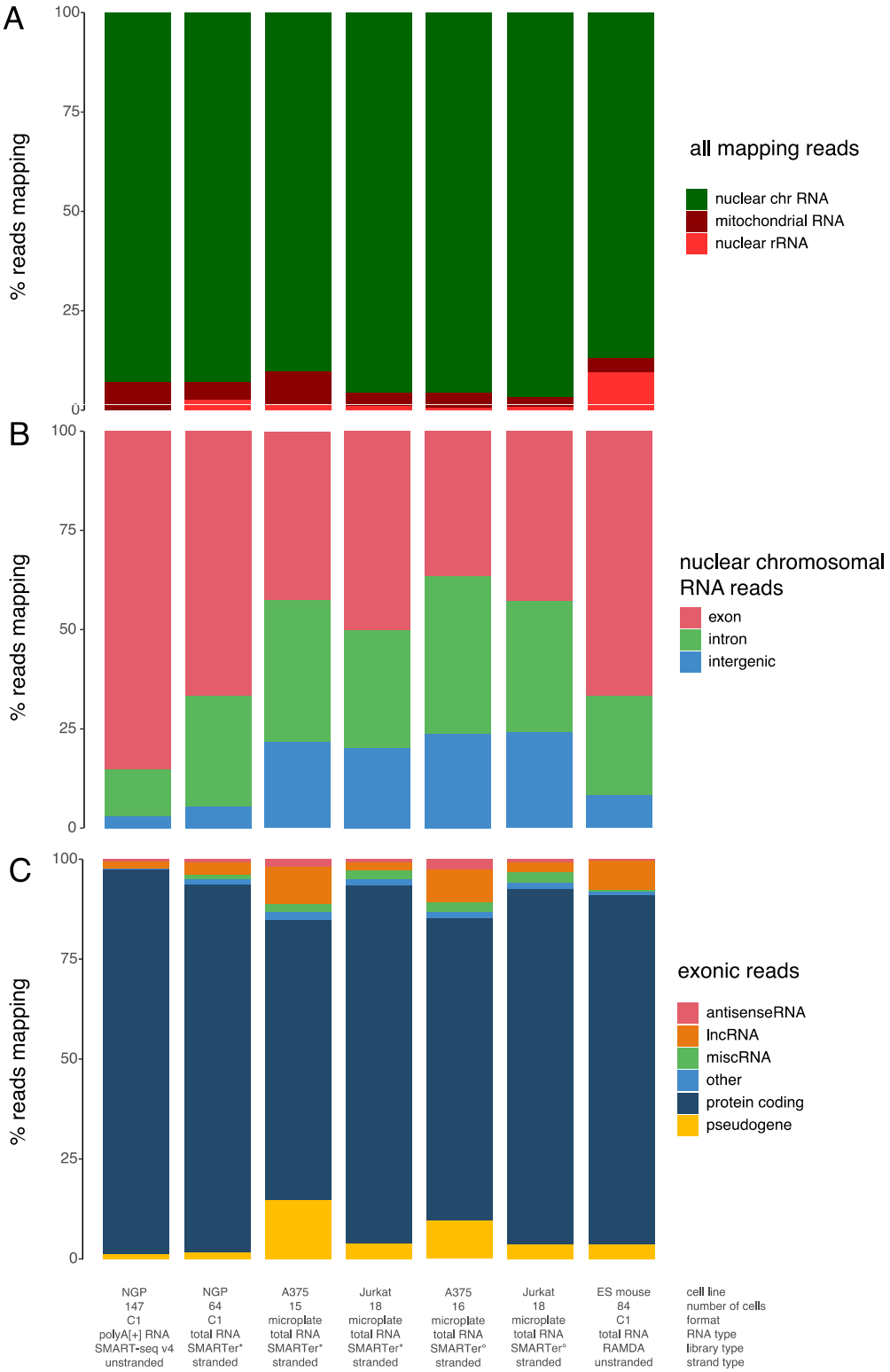
Supplementary Figure 10: IGV visualisation of sense-antisense gene pairs for polyA[+] RNA. The reads mapping on the sense (red) and antisense (blue) strand can not be unambigously assigned to the MIF gene as the data is unstranded. While the counts will be partially mis-assigned to the MIF-AS1 gene, the reads clearly have the splice pattern of only the MIF gene.



Supplementary Figure 11: Quantification in stranded and unstranded analysis mode of SMARTer total RNA seq data on antisense genes in six randomly selected cells demonstrates substantial misquantification of antisense genes in unstranded mode.



Supplementary Figure 12: by using the LNCipedia transcriptome for quantification, a higher number of lncRNAs was discovered. (A) Number of LNCipedia genes detected in subsampled data (1, 4 and 8 million reads per cell). The proportions are equal compared to Ensembl lncRNAs overlap. (B) Overlap between lncRNAs detected in polyA[+] and total RNA libraries. (C) Expression counts for lncRNAs detected in only polyA[+] RNA libraries (red), only total RNA libraries (green) or both (gray).



Supplementary Figure 13: SMARTer single cell total RNA sequencing works well on FACS sorted cells in microplates. Both reagents kits have similar performance (#634413, denoted with * and #634444, denoted with °; see Materials and Methods). A) Percentage of reads derived from nuclear RNA, mitochondrial RNA and ribosomal RNA per cell quantified with STAR. B) Percentage of reads originating from nuclear chromosomes derived from exonic, intronic and intergenic regions per cell quantified with STAR. C) Percentage of exonic reads attributed to the different biotypes per cell quantified with Kallisto.

SUPPLEMENTARY TABLES

Supplementary Table 1: RT-qPCR primer sequences.

CDKN1A_F	CCTCATCCCGTGTTCTCCTTT
CDKN1A_R	GTACCACCCAGCGGACAAGT
BAX_F	GATGCGTCCACCAAGAAGCT
BAX_R	CGGCCCCAGTTGAAGTTG
BBC3_F	CCTGGAGGGTCCTGTACAATCT
BBC3_R	GCACCTAATTGGGCTCCATCT
SDHA_F	TGGGAACAAGAGGGCATCTG
SDHA_R	CCACCACTGCATCAAATTCATG
TBP_F	CACGAACCACGGCACTGATT
TBP_R	TTTTCTTGCTGCCAGTCTGGAC
YWHAZ_F	ACTTTTGGTACATTGTGGCTTCAA
YWHAZ_R	CCGCCAGGACAAACCAGTAT
HPRT1_F	TGACACTGGCAAAACAATGCA
HPRT1_R	GGTCCTTTTCACCAGCAAGCT

Paper 4

Comprehensive benchmarking of single cell RNA sequencing technologies for characterizing cellular perturbation systems

Karen Verboom, Alemu T Assefa, Nurten Yigit, Jasper Anckaert, Niels Vandamme, Dries Rombaut, Yvan Saeys, Olivier Thas, Kaat Durinck, Frank Speleman and Jo Vandesompele

Contribution: I performed all serum starvation experiments and subsequent cell cycle analysis. I performed the C1 and ddSeq experiments with the help of a lab technician. I performed the data analysis in R to determine the quality of the data and performed gene set enrichment analysis. I made the figures and wrote the manuscript.

In preparation

Comprehensive benchmarking of single cell RNA sequencing technologies for characterizing cellular perturbation systems

Karen Verboom[1,2], Alemu T Assefa[3], Nurten Yigit[1,2], Jasper Anckaert[1,2], Niels Vandamme[2,4,5], Dries Rombaut[1,2], Yvan Saeys[2,4,5], Olivier Thas[2,3], Kaat Durinck[1,2], Frank Speleman[1,2] and Jo Vandesompele[1,2].

[1] *Center for Medical Genetics, Ghent University, Ghent, Belgium*

[2] *Cancer Research Institute Ghent, Ghent, Belgium*

[3] *Department of Data Analysis and Mathematical Modeling, Ghent University, Ghent, Belgium*

[4] *Data Mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium*

[5] *Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium*

Corresponding author: Jo Vandesompele, Jo.Vandesompele@UGent.be

ABSTRACT

Technological advances in transcriptome sequencing of single cells has provided an unprecedented view on tissue composition and cellular heterogeneity. While several studies have compared different single cell RNA sequencing methods with respect to data quality and their ability to distinguish cellular subpopulations, none of these comparative studies investigated the heterogeneity of the cellular transcriptional response upon a chemical perturbation. In this study, we evaluated the transcriptional response of NGP neuroblastoma cells upon nutlin-3 treatment using the C1, ddSeq and Chromium single cell systems. These systems and library preparation methods are representative for the wide variety of platforms, ranging from microfluids chips to droplet-based systems and from full transcript sequencing to 3' end sequencing. In parallel, we used bulk RNA-seq for molecular characterization of the transcriptional response. Two complementary metrics to evaluate performance were applied: the first is the number and identification of differentially expressed genes as robustly assessed by two statistical models, and the second is enrichment analysis of biological signals, which is independent of sample size or number of cells evaluated. Where relevant, we downsampled sequencing library size, selected cell subpopulations based on specific RNA abundance features, or created pseudobulk samples to make the data more comparable. While the C1 detects the highest number of genes per cell and better resembles bulk RNA-seq, the Chromium identifies most differentially expressed genes, albeit still substantially fewer than bulk RNA-seq. Gene set enrichment analyses reveals that detection of the most abundant genes in single cell RNA-seq experiments is sufficient for molecular phenotyping. Finally, single cell RNA-seq reveals a heterogeneous response of NGP cells upon nutlin-3 treatment, pinpointing putative late-responders or resistant cells, hidden in bulk RNA-seq experiments.

INTRODUCTION

Almost a decade ago, the first single cell RNA-seq study was published, in which cells were manually isolated and polyadenylated transcripts were captured using oligo(dT) reverse transcription primers (1). Since then, various single cell RNA-seq methods and devices have emerged, unveiling an unanticipated cellular heterogeneity underestimated or masked through bulk cell population gene expression profiles. As such, single cell RNA-seq enabled the identification of subtle differences among cells and the detection of rare (novel) subpopulations. This has led to revolutionary discoveries in several research fields, including cancer (2, 3)

and development (4–6). The first automated single cell isolation devices used flow cytometry or microfluidic chips and could only capture a hundred cells. Most RNA library preparation protocols for these systems provide full gene body read coverage, enabling mutation and splice isoform analysis on top of standard gene abundance profiling (7–10). Using these methods, single cells can be visualized to remove cell doublets and select cells of interest. Later, commercially available and custom made droplet-based methods, such as Chromium, ddsSeq and InDrop, were developed, increasing the throughput to thousands of cells and reducing the cost per cell considerably (11–15). One disadvantage is that these methods typically sequence the 3' or 5' end of a transcript, limiting the analyses to gene expression profiling. Further, these droplet-based methods only quantify the most abundant genes, excluding for instance the detection of medium to low abundant mRNAs and the majority of long non-coding RNAs (lncRNAs). Consequently, lower complexity RNA libraries are generated using these droplet-based methods, resulting in more PCR bias. Fortunately, this bias can largely be reduced through unique molecular indices (UMI), incorporated in the to-be-sequenced molecules in droplet-based systems (14–17). Also, virtually all these initial single cell RNA-seq methods only capture polyadenylated transcripts, ignoring the vast non-polyadenylated part of the transcriptome, representing roughly two third of the entire transcriptome. Since flow cytometry and microfluidic chip based methods are mostly open systems, single cell total RNA-seq protocols were recently custom developed enabling the sequencing of both polyadenylated as well as non-polyadenylated transcripts (18–20). The extensive advances in the single cell RNA-seq technologies raise the question which method to use for a given application. While several studies compared single cell RNA-seq methods in terms of data quality, costs, reproducibility, and the ability to discriminate subpopulations, our study focuses on the added value of three single-cell RNA-seq technologies for differential gene expression analysis and revealing putative transcriptional heterogeneity (21–25). Therefore, cell cycle synchronized NGP neuroblastoma cells were treated with the TP53 activator nutlin-3, whose transcriptional effects are well characterized in bulk, resulting in activation of the *TP53* pathway and consequently in cell cycle arrest and apoptosis (26, 27). Single cell RNA-seq of this well-characterized model system has been performed using three commercially available single cell devices, i.e. the microfluidic chip-based C1 (Fluidigm) and the droplet-based ddSeq (Bio-Rad, Illumina) and Chromium (10X Genomics), single cell RNA-seq platforms, representing a range of throughputs from 96, 300 or more than 10,000 cells per condition, respectively. As a reference, the same experiment was also performed using bulk RNA-seq of ten replicates.

We revealed that despite the lower number of differentially expressed genes detected in single cell RNA-seq experiments compared to bulk population analysis, the biological signal can be faithfully recognized through gene set enrichment analysis for all single cell devices. Furthermore, we show that single cell transcriptome analyses reveal a certain degree of cellular heterogeneity in response to nutlin-3 treatment, possible pinpointing late-responders or resistant cells, hidden in bulk RNA-seq experiments.

METHODS

Cell lines

The neuroblastoma cell line NGP is a kind gift of prof. R. Versteeg (Amsterdam, the Netherlands). Cells were maintained in RPMI-1640 medium (Life Technologies, 52400-025) supplemented with 10 % fetal bovine serum (PAN Biotech, P30-3306), 1 % of L-glutamine (Life Technologies, 15140-148) and 1 % penicillin/streptomycin (Life Technologies, 15160-047) (referred to as complete medium) at 37 °C in a 5 % CO_2 atmosphere. Short tandem repeat genotyping was used to validate cell line authenticity prior to

performing the described experiments and verification of absence of mycoplasma was done on a monthly basis using the MycoAlert Mycoplasma Detection Kit (Lonza, T07-318), according to manufacturer's instructions.

Cell cycle synchronization and nutlin-3 treatment of NGP cells

NGP cell cycle synchronization, nutlin-3 treatment and cell cycle analysis were performed as previously described by Verboom et al (20).

RNA isolation, cDNA synthesis and reverse transcription quantitative PCR

RNA was isolated, cDNA synthesized and RT qPCR performed as described by Verboom et al (20).

Bulk RNA library preparation of NGP cells

The RNA of ten biological replicates of NGP cells treated with either nutlin-3 or vehicle, without serum starvation was extracted using the RNeasy mini kit. The RNA concentration was measured using spectrophotometry (Nanodrop 1000) and quality ascertained using the fragment analyzer (Advanced Analytical). 100 ng of total RNA was used as input for the TruSeq stranded mRNA library prep kit (Illumina, 20020594), according to manufacturer's instructions.

Single cell RNA library preparation of C1 isolated NGP cells

Sequencing data of single cells isolated with the C1 were previously generated and used here (20) [GEO: GSE119984].

Single cell RNA library preparation of ddSeq isolated NGP cells

Single cell RNA-seq on the ddSeq system (Bio-Rad) was performed using the SureCell WTA 3' library prep kit (Illumina, 20014279) according to manufacturer's instructions with minor modifications. Four samples were prepared: (1) nutlin-3 treated cells with external RNA controls consortium (ERCC) spikes diluted to 1/1000 (N704 index), (2) nutlin-3 treated cells with ERCC spikes diluted to 1/10,000 (N705 index), (3) vehicle treated cells with ERCC spikes diluted to 1/1000 (N706 index) and (4) vehicle treated cells with ERCC spikes diluted to 1/10,000 (N707 index). Cells were diluted to 5000 cells/µl and ERCC spikes were diluted to 1/500 and 1/5000. Cells and ERCC spikes were mixed 1:1 resulting in a final concentration of 2500 cells/µl and a dilution of 1/1000 and 1/10,000 for the ERCC spikes, respectively. After library preparation, the quality of the RNA libraries was confirmed on the Bioanalyzer (Agilent).

Single cell RNA library preparation of Chromium isolated NGP cells

Single cell RNA-seq on the Chromium system (10X Genomics) was performed for nutlin-3 (SI-GA-8E index) and vehicle (SI-GA-8D index) treated NGP cells using the GemCode Single Cell 3' Gel Bead and Library Kit (V2 chemistry, 10X Genomics, PN-120237, PN-120236, PN-120262) according to manufacturer's instructions with minor modifications. Cells were centrifuged at 4 °C at 400 g and resuspended in PBS + 0.04 % BSA to yield an estimated concentration of 1000 cells/µl. 3.5 µl of the cell suspension was used to obtain a cell recovery of about 2000 cells per sample. Per sample, 2.5 µl of an 1/10 dilution of ERCC spikes was added to the mastermix. After library preparation, the quality of the RNA libraries was confirmed on the Bioanalyzer.

Library sequencing

Bulk RNA libraries were quantified using KAPA library quantification kit (Roche) and diluted to 4 nM. 1.2 pM of the RNA library was paired-end sequenced on a NextSeq 500 (Illumina) with a read length of 75 bp. The C1 RNA libraries were quantified using the KAPA library quantification kit and libraries were diluted to 4 nM. 1.5 pM of the library was single-end sequenced on a NextSeq 500 (Illumina) with a read length of 75 bp. The ddSeq RNA libraries were quantified using the Qubit dsDNA HS kit (Thermo Fischer Scientific, Q32854) and libraries were diluted to

2 nM. 3 pM of the library was paired-end sequenced on a NextSeq 500 with a read length of 68 and 75 bp and a custom sequencing primer included in the SureCell WTA 3' library prep kit. The Chromium RNA libraries were quantified using the KAPA library quantification kit and libraries were diluted to 4 nM. 1.2 pM of the library was paired-end sequenced twice on a NextSeq 500 with a read length of 26 and 98 bp.

Data analysis of the bulk RNA sequencing data

Raw fastq files were processed with Kallisto (v.0.43.1) (28) using Ensembl (v.91) annotation (29) .

Data analysis of the C1 RNA sequencing data

To assess the quality of the data, the reads were mapped using STAR (v.2.5.3) (30) on the hg38 genome including the full ribosomal DNA (45S, 5.8S and 5S) and mitochondrial DNA sequences. The STAR parameters were set to retain only primary mapping reads, meaning that for multi-mapping reads only the best scoring location is retained. Genes were quantified by Kallisto (v.0.43.1) (28) using Ensembl (v.91) (29) annotation supplemented with the ERCC spike-in RNA sequences.

Data analysis of the ddSeq RNA sequencing data

To analyze the ddSeq data, ddSeeker, a custom pipeline based on the Drop-seq Core Computational Protocol (version 2.0.0 -9/28/18), was used (31). ddSeeker.py was run on paired-end gzipped fastq files with default parameters using Python (v.3.6.4), pysam (v.0.14) and Biopython (v.1.71). First, fastq files were converted to unaligned BAM files using Picard FastqToSam. These BAM files were subsequently tagged with both cell (XC) and molecular (XM) barcodes using TagBamWithReadSequenceExtended. Next, these tagged BAM files were filtered to remove reads below the base quality threshold (XQ) and to remove erroneous barcodes (XE). The SMART adapter that can occur at the 5' end of the read was trimmed using TrimStartingSequence and polyA tails were trimmed using PolyATrimmer. Next, the trimmed and filtered BAM files were converted to fastq files and were used for subsequent alignment. Reads were aligned using STAR (v.2.6.0) (30) and Ensembl (v.91) (29) annotation and the BAM file was sorted by query name using SortSam (Picard). The sorted alignment files and the unaligned (tagged) BAM files were then merged to recover BAM tags, lost during alignment (MergeBamAlignment from Picard). TagReadWithGeneFunction provides three tags for each read (gene name, gene strand and gene function) required to create a digital expression matrix. This cell matrix contains two subpopulations of cells, one cell population with many genes and reads and one with few genes and reads per cell. As the cell population with few genes and reads does not recapitulate biological signal, these needed to be removed. The average number of genes per cell (5045) clearly separated the two subpopulations, therefore, only cells with more than 5045 genes were retained (MIN_NUM_GENES_PER_CELL=5045).
Furthermore, only genes with at least 2 read counts were retained. The matrices for 1/1000 and 1/10,000 diluted ERCC spikes were merged.

Data analysis of the Chromium RNA sequencing data

Demultiplexing of the raw sequencing data was done by 10x Cell Ranger (v.2.0.2) software 'cellranger mkfastq' which wraps Illumina's bcl2fastq. The fastq files obtained after demultiplexing were used as input for 'cellranger count', which aligns the reads to the hg38 human reference genome using STAR (30) using Ensembl (v.91) (29) annotation and collapses to UMI counts. This was extended with mapping to ERCC spike-in RNA sequences, generating two separate matrices. Aggregation of samples to one dataset was done using 'cellranger aggr'. The gene and ERCC count matrices were merged and only cells containing ERCC spikes were retained.

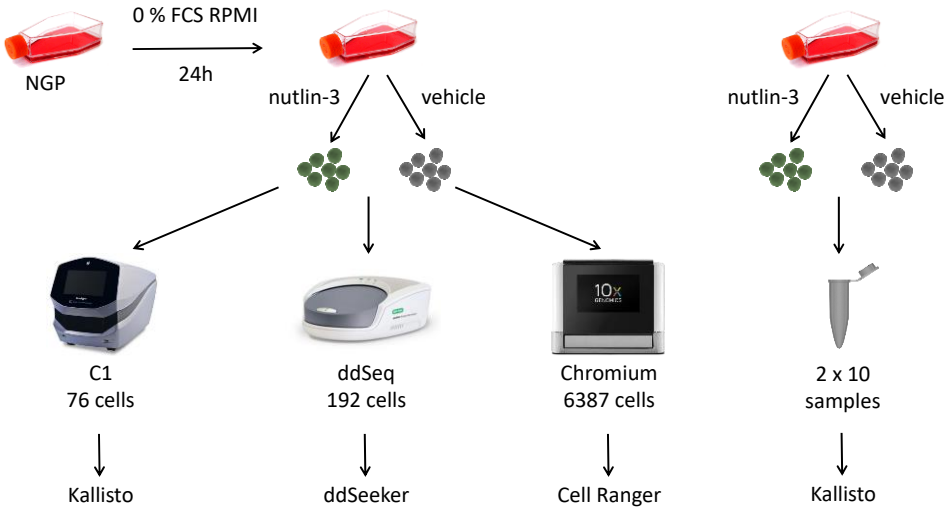


Figure 1: overview of the experimental set-up. Synchronized NGP cells were treated with either nutlin-3 or vehicle and single cell RNA-seq was performed using the C1, ddSeq and Chromium device. In parallel, bulk RNA-seq of 10 replicates of NGP cells treated with nutlin-3 and vehicle was carried out. Each dataset was analyzed with the appropriate pipeline.

Quality control and filtering of the single cell sequencing data

Quality assessment and further filtering were done in R (v.3.5.0) using Seurat (v.2.3.4) (32) and Scater (v.1.8.0) (33) as described by Lun et al. (34). For the C1 dataset, only genes with at least 5 counts were retained, as described previously (35). To retain a similar fraction of genes for the other two single cell devices, genes in at least 17 and 20 cells were retained for ddSeq and Chromium, respectively. The cyclone function of the scran (v.1.8.4) package was used to determine the cell cycle stage of the cells.

Differential analysis of the single cell sequencing data using PIM and EdgeR-Zinger

For testing differential gene expression (DGE) between the nutlin-3 and vehicle treated cells, edgeR in combination with Zinger for the single cell experiments (36, 37) and probabilistic index models (PIM) (38) were used. Zinger calculates weights from zero-inflated negative binomial models, which is used by edgeR to fit a weighted generalized linear model (GLM) with negative binomial distribution. The PIM is a distribution-free regression model that models the probabilistic index (PI) as a function of the treatment factor (38). If there is a strong evidence that a gene is DE, then the estimated PI becomes close to 1 (if the gene expression is higher in the nutlin-3 group) or 0 (if lower in the nutlin-3 group). Under the null hypothesis (no DE), the estimated PI is expected to be 0.5, indicating that there is a 50% chance that the expression of the gene in a randomly selected cell from the nutlin-3 group is lower than that of a randomly selected cell from the vehicle group (and vice versa).

Ranking of cells based on TP53 pathway

Cells were ranked based the total count for TP53 pathway genes (39). In particular, cells were ranked according to the sum of log-CPM for 116 TP53 pathway genes (39). Ranks were then compared between the treatment and control group and significance was determined using the Wilcoxon rank sum test.

Gene set enrichment analysis

Genes were ranked according to their log fold change in decreasing order and used as input for a preranked gene set enrichment analysis (GSEA) (40). The C2 (curated gene sets) gene sets were

used to identify significantly enriched gene sets (q<0.05) in the datasets.

RESULTS

Experimental design

To compare single cell polyA[+] RNA-seq data generated with the C1 (Fluidigm), ddSeq (Bio-Rad/Illumina) and Chromium (10x Genomics), the same cellular perturbation experiment was evaluated on all three devices. Additionally, the same experiment was also performed in bulk for ten replicates to contrast with the single cell RNA-seq results (Figure 1). Since cell cycle status may be a confounder in single cell experiments, cell cycle synchronization by serum starvation of NGP neuroblastoma cells was carried out for all single cell experiments, but not for the bulk experiment, prior treatment, resulting in an arrest in the G0/G1 phase (Supplementary Figure 1A). Next, NGP cells were treated with nutlin-3 or vehicle (ethanol). Nutlin-3 is a TP53 activator by inhibiting the interaction between TP53 and its negative regulator MDM2, resulting in an activation of the TP53 pathway and consequently in cell cycle arrest and apoptosis (1). The effect of nutlin-3 treatment was confirmed using bulk RT-qPCR by a 28-fold upregulation of *CDKN1A*, a known TP53 target gene (Supplementary Figure 1B). ERCC spike-in RNA was added in all single cell experiments, but not in the bulk RNA-seq experiment.

Quality control and filtering of sequencing data

All three single cell methods generated high quality libraries as confirmed by Bioanalyzer or Fragment Analyzer (Supplementary Figure 1C). Single cell RNA-seq data differ amongst others in the generated read structure, as ddSeq and Chromium reads for instance contain UMIs, while this is not the case for C1 reads. Therefore, each device has its own pipeline to analyze the data, although all reads, including those generated with the bulk RNA-seq protocol, were mapped against Ensembl v91, making the data comparable (Figure 1). For C1, the number of single cells was determined visually and 83 of the 96 capture sites contained single cells without visible debris. In contrast, single cells isolated with ddSeq and Chromium cannot be visualized and the number of single cells is determined by the computational pipeline, resulting in 260 and 7514 single cells for ddSeq and Chromium, respectively. No ERCC spikes were detected in 7 out of the 7514 Chromium isolated cells and these cells were removed from further analysis. To filter out low quality cell data, all cells with a log-transformed number of reads or genes more than three times the median absolute deviation (MAD) below the log-transformed median were removed from further analysis, since transcripts are likely not efficiently captured in these cells (2). Similarly, cells above this cutoff were also removed, as these data may be derived from cell doublets. Since we added ERCC spike-in molecules in all three single cell experiments, the same MAD cutoff was used to remove low quality cells and cell doublets based on the percentage of ERCC spike-in reads per cell. Finally, 76, 192 and 6387 single cells were retained for the C1, ddSeq and Chromium, respectively (Table 1, Figure 1). Besides low-quality cells, also genes that are only expressed in a few cells were removed. Due to the differences in throughput, the selected cutoff differs depending on the device and ~58 % of the genes were maintained by retaining only genes expressed in at least 5 (16,921 genes), 17 (12,753 genes) and 20 (15,307 genes) cells for the C1, ddSeq and Chromium, respectively. For the bulk experiment, genes expressed in fewer than three

Table 1: overview of the number of cells removed based on library size, number of genes and percentage ERCC spikes per cell.

	library size	number of genes	ERCC spikes (%)	remaining cells
C1	0	0	7	76
ddSeq	28	36	6	192
Chromium	360	211	822	6387

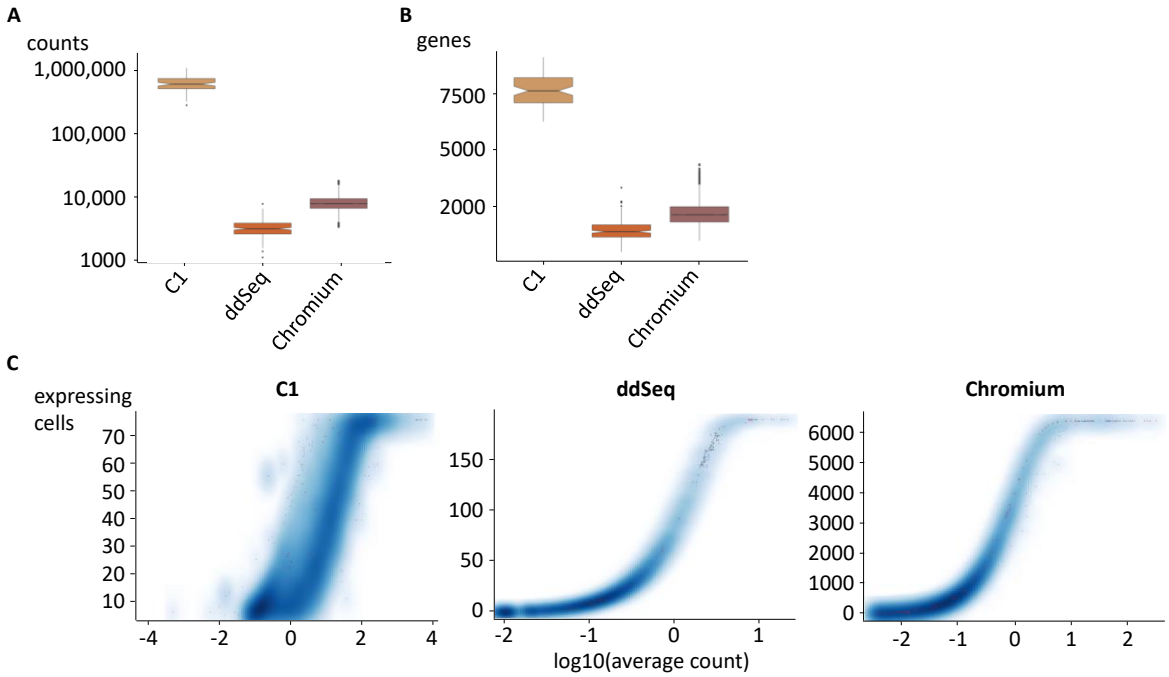


Figure 2: number of counts and genes detected per device. Boxplots depicting the number of counts (A) and genes (B) detected for the C1, ddSeq and Chromium after filtering. (C) Smoothscatter plot shows the correlation between the gene expression level and the number of cells that express the gene. Red dots show the ERCC spikes.

samples were removed, retaining 33,700 genes. In general, the average gene expression correlation among the platforms was high. As expected, the correlation between ddSeq and Chromium was slightly higher (r=0.84), compared to each of these methods with the C1 (ddSeq: r=0.77, Chromium: r=0.78) as ddSeq and Chromium generate sequencing libraries in a similar way (supplementary Figure 2A). Furthermore, the average gene expression over all cells in the C1 dataset correlates best with bulk (r=0.83) (supplementary Figure 2B), with both methods sequencing full transcripts.

The C1 has the highest gene detection sensitivity

After filtering, an average of 0.71 million, 3780 and 9466 reads were retained per cell, resulting in the detection of on average 7621, 1487 and 2220 genes per cell for the C1, ddSeq and

Chromium, respectively, demonstrating that the C1 has the highest sensitivity (Figure 2A-B). Of note, 0.1 %, 1.5 % and 16.8 % of the reads were respectively attributed to ERCC spikes. Single cell RNA-seq experiments suffer from a lot of missing data points (dropouts) that can be either biological or technical. For C1, 54.96 % of the values are dropouts, while this is much higher for ddSeq (88.34 %) and Chromium (85.50 %). PCA plots show a separation between nutlin-3 and vehicle treated cells for all single cell devices. While the distinction is clear for ddSeq, there is more overlap between treated and untreated cells for the C1 and Chromium (Supplementary Figure 3A). In general, genes that are low abundant are detected in a few cells, while more abundant genes are expressed in a higher fraction of cells (Figure 2C). Both ddSeq and Chromium display a tighter curve compared to the C1, probably due to the higher number of cells and removal of amplification bias by UMIs. Furthermore, ddSeq and Chromium data contain more genes that are expressed in only a few cells compared to C1, where genes are generally detected in a higher fraction of cells (Figure 2C). Comparing the genes detected with bulk RNA-seq to these detected using single cell RNA-seq revealed a large overlap, although, some of the genes are only detected by one of the devices (Figure 3A). In general, genes detected by all platforms have a higher expression compared to genes detected by only one device (Figure 3 B-D).

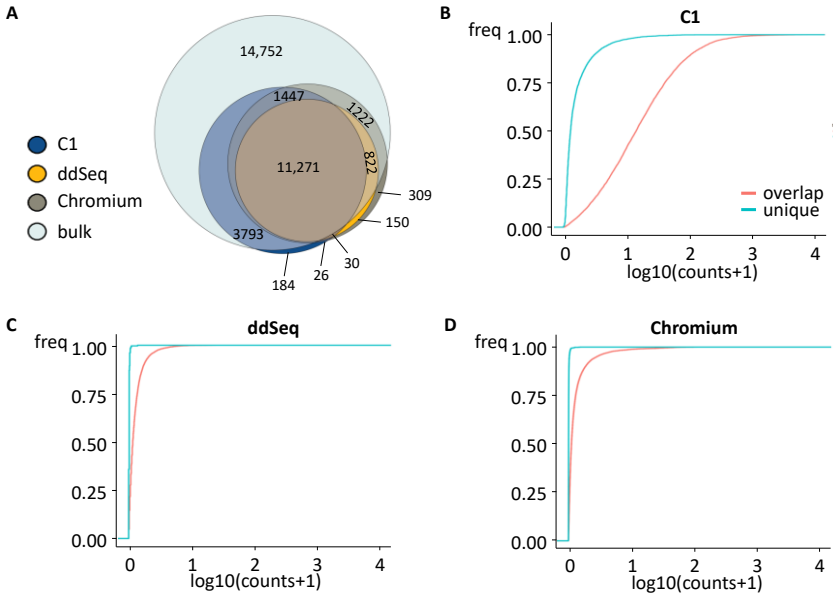


Figure 3: each platform detects a unique set of genes. (A) Overlap between detected genes using bulk RNA-seq, C1, ddSeq and Chromium. Cumulative expression plots of genes detected with all single cell devices or with only C1 (B), ddSeq (C) or Chromium (D).

16.3 % and 7.2 % of all reads map on the top 25 expressed genes for the C1 and ddSeq, respectively, while this number is higher for Chromium (29.6 %), highlighting the lower library complexity of Chromium libraries. The top 25 abundant genes contain many ribosomal and mitochondrial genes (Supplementary Figure 3B). Overlap shows that the top 25 genes expressed genes differ per platform (Supplementary Figure 3C).

Bulk experiments detect most differential expressed genes, while Chromium most enriched gene sets

As the number of differentially expressed genes in part depends on the statistical tool, we

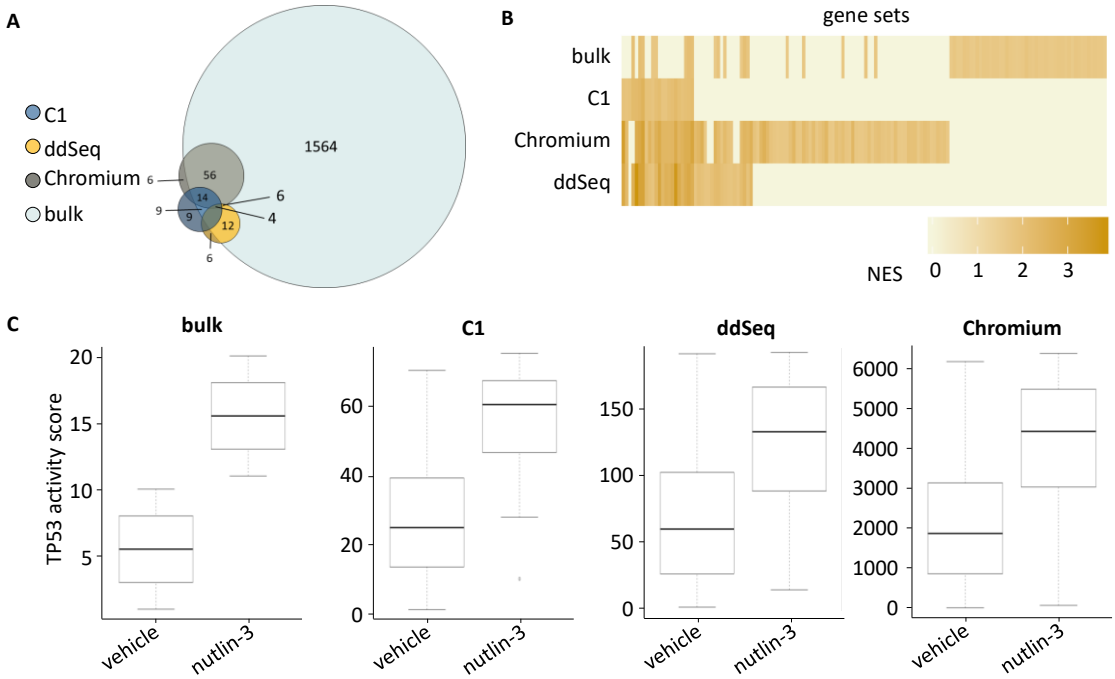


Figure 4: single cell RNA-seq results recapitulate the biological signal. (A) Overlap of genes detected by all platforms and significantly differentially expressed with both EdgeR in combination with Zinger as well as PIM. (B) Heatmap of significantly positively (q-value <0.05) enriched gene sets after GSEA for the C2-curated gene sets for each method. Gene sets are color-coded according to their normalized enrichment score (NES). (C) Boxplots depicting the TP53 activity score per cell, whereby ranking was based on the expression of 116 TP53 target genes.

performed both EdgeR in combination with Zinger as well as probabilistic index model (PIM) analysis and retained high-confident genes that are called significantly differentially expressed between nutlin-3 and vehicle treated NGP cells with both tools (3–5). In a comparison study, EdgeR was shown to be one of the better tools for single cell differential gene expression analysis and PIM is a new tool, developed specifically for differential gene expression analysis of single cells (Assafa et al., manuscript in preparation) (3). Genes were called significantly differentially expressed by EdgeR if FDR < 0.05 and absolute log fold change > 1, while genes are significantly differentially expressed with PIM if adjusted p-value < 0.05 and PI < 0.4 (downregulated) or PI > 0.6 (upregulated). For bulk, C1, ddSeq and Chromium, 7010, 40, 28 and 88 significantly differentially expressed genes were identified, respectively (Supplementary Table 1). By only including genes that are detected by all four platforms, the number of differentially expressed genes in the bulk dataset drastically dropped to 1665, while only little differences were noticed for C1 (36 genes), ddSeq (28 genes) and Chromium (86 genes), in line with the fact that many genes are only detected in the bulk experiment. Most differentially expressed genes in the single cell datasets overlap with those detected in the bulk dataset, however, some genes are uniquely differentially expressed in only one of the datasets (Figure 4A). Genes that are differentially expressed in only one of the single cell datasets are mostly borderline in significance and effect sizes (Supplementary Figure 4). Interestingly, although more genes are significantly differentially expressed in the bulk dataset compared to the single cell datasets, GSEA analysis shows that Chromium identifies more significantly (q-value <0.05) positively enriched gene sets, demonstrating that biological signal can be effectively captured with only the most abundant genes (Supplementary Table 1, Figure 4B). Of note, several TP53 gene sets pop up in all positively enriched gene sets, while cell cycle gene sets are common in the negatively enriched gene sets, validating the effect of nutlin-3 on the TP53 pathway and the cell cycle arrest in nutlin-3 treated cells for all datasets. Furthermore, ranking cells according to the expression of 116 TP53 activated genes shows that these genes are significantly higher expressed (p-value < 0.01) in nutlin-3 treated cells compared to vehicle treated cells for all devices, showing that these all recapitulate biological signal (Figure 4C). Of note, the bulk experiment has the clearest separation between treated and untreated cells, but this may be in part due to the fact that TP53 target genes were defined based on bulk gene expression profiles (6).

Single cell RNA sequencing reveals a heterogeneous response upon nutlin-3 treatment and uncovers hidden biological signals

To get a first view on the heterogeneity of the response of NGP cells on nutlin-3 treatment, the expression of *CDKN1A,* a known TP53 target, was determined for the three single cell and he bulk RNA-seq experiments. While *CDKN1A* is significantly upregulated in all datasets upon nutlin-3 treatment, there is a remarkable heterogeneity of *CDKN1A* expression in the single cell datasets (Figure 5A). To understand the differences between cells with a low and high expression of *CDKN1A*, nutlin-3 treated cells with *CDKN1A* expression in the lowest quartile were compared to cells with expression in the highest quartile. To have a sufficiently large number of cells in each group, this analysis was only done for the Chromium dataset. A total of 83 genes were significantly differentially expressed, of which 76 overlapped with the set of genes significantly differentially expressed between nutlin-3 and vehicle treated cells in the full Chromium dataset (Supplementary Table 1). In addition, 93 of the 103 significantly positively enriched gene sets overlap with those of the full Chromium dataset (Figure 5B, Supplementary Table 1). These results demonstrate that the same signals can be detected between vehicle and nutlin-3 treated cells and between nutlin-3 treated cells with low and high *CDKN1A* expression.

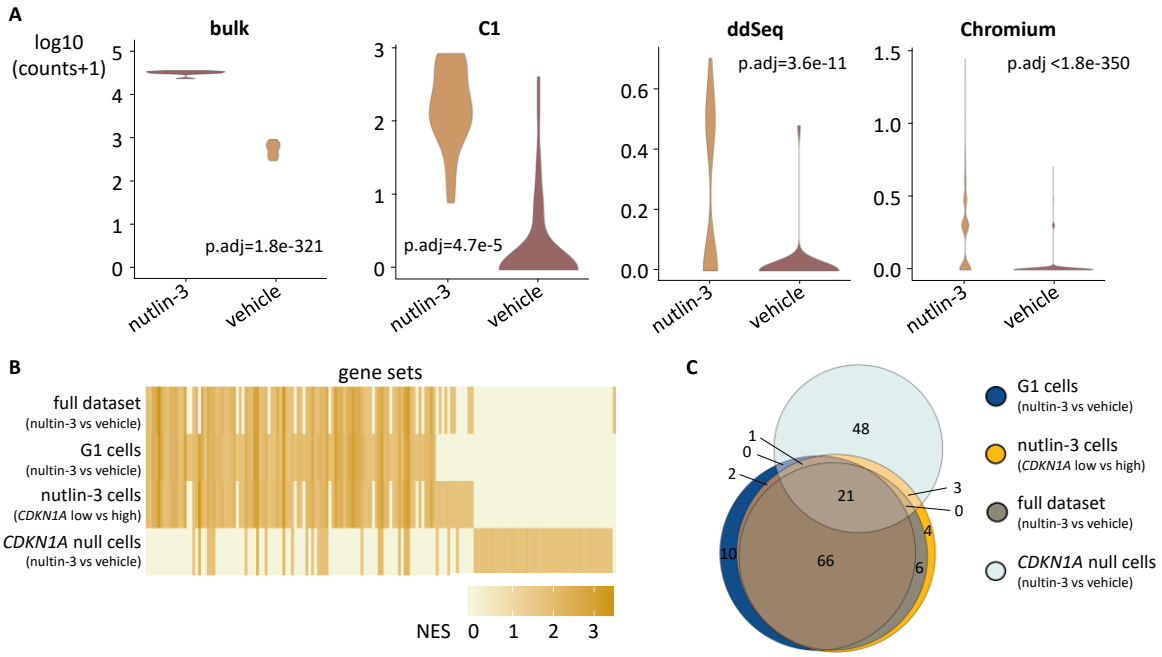


Figure 5: differences between cells with varying *CDKN1A* expression. (A) *CDKN1A* expression in the bulk and single cell RNA-seq datasets. (B) Heatmap of significantly positively (q-value <0.05) enriched gene sets after GSEA for the C2-curated gene sets for each platform. Gene sets are color-coded according to their normalized enrichment score (NES). (C) Overlap between significantly positively enriched gene sets for the three cellular subgroups and full Chromium dataset.

As the cell cycle can be a major confounder in single cell experiments masking putative biological effects, cells in the G1 phase of the cell cycle were selected based on the expression of a G1 cell cycle signature in the Chromium data set. Doing so, 105 genes were significantly differentially expressed, of which 80 overlapped with the differentially expressed genes of the full Chromium dataset (Supplementary Table 1). As we detected slightly more (105 instead of 88) differentially expressed genes between nutlin-3 and vehicle treated cells in the G1 phase compared to the full dataset, these differentially expressed genes might have been masked by cell cycle effects in the full dataset. Several genes that are downregulated in the G1 cells, but not in the full dataset, including *UBE2C* and *PCLAF,* are known to be repressed by TP53 and also downregulated in the bulk RNA-seq dataset (7, 8). Likewise, several genes that are upregulated in the G1 cells only, including *DDIT4* and *KRT17,* are known to be induced by TP53 and also upregulated in the bulk RNA-seq dataset, showing that biologically interesting targets are identified in RNA-seq data from single cells in the same cell cycle phase (9, 10). Interestingly, *PTTG1* is significantly differentially expressed in G1 cells, but not in the full Chromium, nor the bulk dataset, and known to be repressed by TP53 (11). Additionally, 87 of the 100 significantly positively enriched gene sets overlap with those of the full dataset (Figure 5B). One gene set (CONCANNON_APOPTOSIS_BY_EPOXOMICIN_UP, NES= 1.80, FDR = 0.03) containing genes upregulated because of apoptosis was only enriched in the G1 cells, showing the relevance of signals that are only detected in the G1 cells and not in the full dataset.

Finally, as mentioned above, single cell experiments are characterized by a high dropout rate. To determine the differences between nutlin-3 and vehicle treated cells without detectable expression of *CDKN1A,* such (so-called *CDKN1A* null) cells were selected for each treatment arm. A total of 71 genes were significantly differentially expressed, of which 68 overlapped with the full set (Supplementary Table 1. In contrast, only 21 of the 73 significantly positively enriched gene sets overlap with those of the full dataset, depicting that nutlin-3 and vehicle treated cells without detectable expression of *CDKN1A* behave differently

compared to all nutlin-3 and vehicle treated cells in the full dataset.

Overlap of the positively enriched gene sets of the three subsets and the full dataset confirms that nutlin-3 treated cells with low or high expression of *CKDN1A* and vehicle and nutlin-3 treated cells in the G1 phase resemble the vehicle and nutlin-3 treated cells in the full dataset, while cells without expression of *CDKN1A* in both groups are different (Figure 5C).

Pseudobulk data resembles real bulk data better than single cell data

To understand the differences between bulk RNA-seq and single cell RNA-seq patterns better, pseudobulk data from the single cell data were created by pooling and averaging subsets of single cells. Chromium data were pooled in ten pseudosamples per treatment arm, resulting in the same sample size as the bulk data. Chromium was taken as an example as this dataset contains the highest number of cells. To make the data even more comparable, the bulk library size was downsampled to obtain the same number of reads for the single cell and bulk dataset summarized over all (pseudo)samples. Originally, the bulk library size was 4.8 times larger compared to the single cell library size. After downsampling, the total number of reads in each experiment was 70.2 million, with a mean of 3.5 million reads per (pseudo)sample (Figure 6A). Only genes expressed in at least 3 samples were retained in both datasets. The correlation between the average gene expression in the downsampled bulk and pseudobulk dataset was higher (r=0.83) compared to the correlation in the original bulk and single cell dataset (r=0.69) (Supplementary Figures 2 and 5A). As expected, fewer genes (26,845 instead of 33,700) and fewer significantly differentially expressed genes (5277 instead of 7010) were detected in the downsampled bulk dataset compared to the original bulk dataset, due to the lower sequencing depth (Supplementary Table 1). Of these differentially expressed genes, the large majority (5105, 96.74 %) overlapped with the differentially expressed genes of the original bulk dataset. For the pseudobulk dataset, almost 10-fold more genes (810 instead of 88) were significantly differentially expressed compared to the original single cell dataset. Of note, this

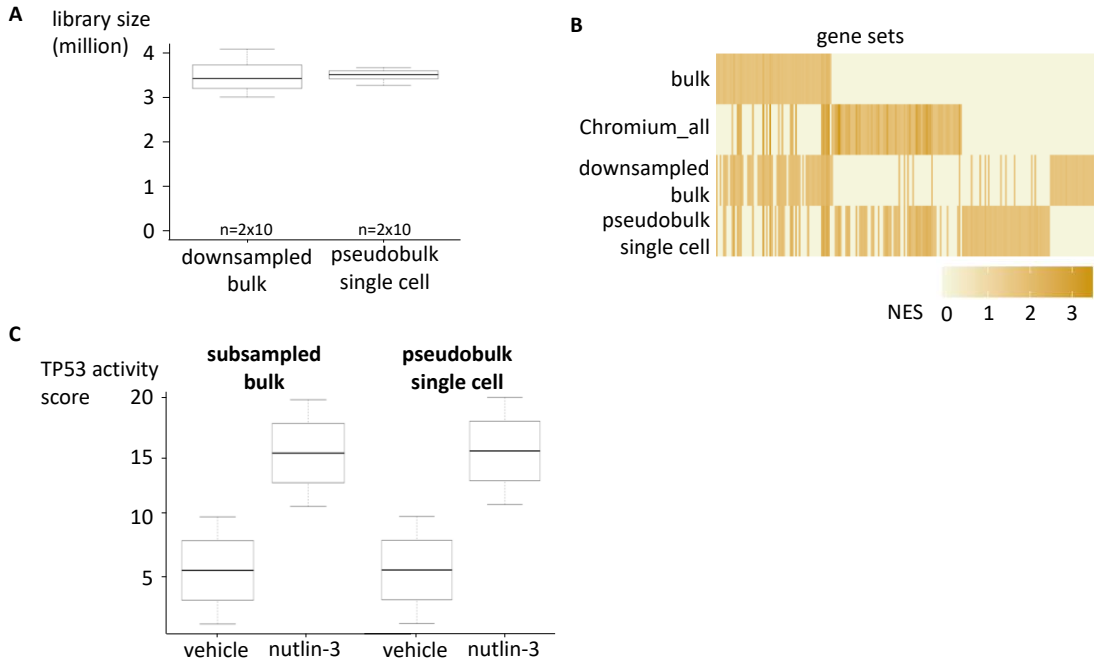


Figure 6: downsampling of bulk and pseudobulkification of single cell data. (A) After downsampling of the bulk data and generating pseudobulk data from the Chromium single cell RNA-seq data, the mean number of reads per (pseudo)sample is 3.5 million. (B) Heatmap of significantly positively (q-value <0.05) enriched gene sets after GSEA for the C2 curated gene sets for the original and downsampled bulk and pseudobulk Chromium datasets. (C) Boxplots depicting the TP53 activity score per cell, whereby ranking was based on the expression of 116 TP53 target genes.

number is still considerably lower compared to bulk at equal sequencing depth. The higher number of differentially expressed genes in the pseudobulk dataset compared to the original single cell dataset is probably owing to the reduction in noise after pooling. 519 of the 810 significantly differentially expressed genes in the pseudobulk dataset overlap with the differentially expressed genes of the downsampled bulk dataset. Furthermore, 125 and 97 significantly positively enriched datasets were identified for the pseudobulk and downsampled bulk dataset, respectively. With 42 of the 97 positively enriched gene sets in the pseudobulk dataset overlapping with those of the downsampled bulk, and only 18 of the 94 positively enriched gene sets overlapping in the original single cell data and the bulk data, it is clear that the pseudobulk data better resembles the bulk data (Figure 6B, supplementary Figure 5B). Furthermore, ranking cells according to the expression of 116 activated TP53 genes shows that these genes are significantly higher expressed (p-value < 0.01) in nutlin-3 treated cells compared to vehicle treated cells for both the downsampled bulk and pseudobulk dataset, showing that these continue to recapitulate biological signal. Interestingly, there is a clearer separation for the pseudobulk dataset compared to the original single cell dataset (Figure 4C, 6C).

To determine the effect of the sequencing depth on single cell experiments, the total library size of the Chromium dataset was downsampled to the library size of the C1 dataset, since the library size over all cells was only 1.17 times higher for Chromium compared to the C1, downsampling gave similar results as the original experiment (Supplementary Figure 5C-D).

DISCUSSION

Over the last years, several single cell RNA-seq methods emerged, whereby the number of single cells analyzed in a single experiment drastically increased from a few up to tens of thousands of single cells. While several studies attempted to compare these single cell RNA-seq methods, most studies focused on the quality of the generated data and their ability to distinguish cellular subpopulations (1–5). Furthermore, the more recent ddSeq instrument was included in only one comparative study (2). Here, we evaluated for the first time three commercially available single cell devices, i.e. C1, ddSeq and Chromium, to study transcriptional heterogeneity upon a chemical perturbation and to contrast it with a bulk cell population response. To this purpose, NGP neuroblastoma cells were treated with the TP53 activator nutlin-3 or vehicle as negative control followed by single cell RNA-seq using the C1, ddSeq and Chromium. Since the cell cycle state is a known confounder of single cell experiments, this effect was minimized by synchronizing cells prior to treatment. To further characterize the results of the single cell experiments, bulk RNA-seq was performed in parallel on the same model system in ten biological replicates. We showed that the highest gene detection rate and lowest number of droupouts were obtained by the C1 device, confirming that this platform has the highest detection sensitivity, which may partially be explained by the higher sequencing depth (5). Downsampling read depth to an equal number of reads per cells for all three devices should be carried out to effectively confirm that the C1 displays the highest sensitivity, independent of sequencing depth. In addition, the overlap between the detected genes in the bulk and single cell datasets was the highest for the C1 with an overlap of more than 50 %, which is slightly higher than reported previously (5). Possible explanations for this difference are sequencing depth and the applied bulk library prep method. The C1 average gene expression levels correlated better with bulk gene expression data compared to ddSeq and Chromium, owing to the higher sequencing depth and higher transcriptome complexity of the C1 cDNA libraries. In contrast, correlation of average gene expression among the single cell devices revealed a slightly better correlation between the ddSeq and Chromium, in line with their similarity in terms of RNA-seq library preparation. The correlation between C1 and

Chromium was the lowest, as previously reported (5, 6). The gene expression profiles of the ddSeq and Chromium seem less noisy compared to the C1, owing to the higher number of isolated single cells and the use of UMIs. It has been reported that technical noise can be reduced by 50 % using the UMI enabled counting of cDNA molecules (7, 8). Although a large overlap in the genes quantified with the three single cell devices was seen, each device also detected some unique genes. It has been previously reported that unique C1 genes do not have 3' ends that are difficult to capture, preventing their detection by 3' end sequencing technologies such as the ddSeq and Chromium. Hence, the large set of unique C1 genes results from higher C1 mRNA capture efficiency (1, 5). In our attempt to make the devices somewhat comparable, ERCC spike-in RNA molecules were added to all three single cell experiments. Of note, ERCC spikes are generally not added to droplet-based experiments, since these spikes-in molecules are added to every droplet and consequently also amplified and sequenced in droplets without cells, increasing the sequencing costs considerably (1). Due to the lack of guidelines for droplet-based experiments, too many reads (17 %) in our Chromium dataset mapped to ERCC spikes, consequently losing endogenous reads and indicating that lower amounts of ERCC spikes should be added in future experiments. Apart from being used as workflow control, ERCC spike-in molecules can also be used for normalization, although that use is still under debate (9–11).

To test the ability to identify differentially expressed genes upon nutlin-3 treatment in single cell RNA-seq datasets, two different statistical methods were used, i.e. EdgeR in combination with Zinger, and PIM. As differential gene expression analysis tools typically vary in the number of genes called as differentially expressed, we here continued with the intersection of both tools to conservatively identify truly differentially expressed genes (12). The largest number of differentially expressed genes was detected in the Chromium dataset, in line with the observation that more genes are called differentially expressed with increasing number of single cells (5, 12). Although many more genes were differentially expressed in the bulk dataset, the biological signal is faithfully recapitulated in the tested single cell datasets as strong enrichment of several TP53 gene sets was present in all datasets. This result suggests that detecting the most abundant genes (through single cell RNA-seq data) is sufficient for pathway activity analysis. Of note, single cell datasets also reveal some unique enrichment signals, of which the relevance should be determined by further investigation.

To characterize the effect of nutlin-3 treatment at the single cell level, three cell subpopulations from the full Chromium dataset were selected based on their cell cycle stage and TP53 transcriptional target gene *CDKN1A* expression levels, and compared with the entire cell population. This subpopulation analyses were only performed for the large Chromium dataset in order to have a sufficient number of cells per subset. In order to avoid cell cycle effects as much as possible, nutlin-3 and vehicle treated cells in the G1 phase were selected in the first subset. Although a large fraction of the differentially expressed genes in the G1 cells overlapped with the full dataset, more significantly differentially expressed genes were detected in the G1 cells, possibly hidden by cell cycle effects in the full dataset. Many of these genes are known to be regulated by TP53, showing the utility of subpopulation analysis by and the relevance of the genes differentially expressed in cells in the G1 phase. This type of subpopulation analysis could in principle be extended to the other cell cycles stages. In a second subset, differential gene expression analysis and gene set enrichment analysis on nutlin-3 treated cells with low or high expression of *CDKN1A* revealed that these subsets resemble vehicle and nutlin-3 treated cells from the full dataset. This indicates that treated cells with low expression of *CDKN1A* are similar to vehicle treated cells and may thus represent cells that react in a later stage to nutlin-3 or show primary

resistance. To further investigate this intriguing observation, time-course experiments should be set up to reveal if *CDKN1A* is upregulated in a larger fraction of cells at later timepoints, in line with a delayed nutlin-3 response. In the third subset, nutlin-3 and vehicle treated cells without *CDKN1A* expression were compared. Few enriched gene sets overlapped between this subpopulation analysis and the full dataset, indicating that the cells that do not express *CDKN1A* in nutlin-3 and vehicle treated cells are not representative for the whole population of nutlin-3 and vehicle treated cells. Again, it might be that some of the nutlin-3 treated cells without *CDKN1A* expression react at a later timepoint or are resistant, whereby these cells resemble vehicle treated cells. Consequently, fewer differences are noticeable between the nutlin-3 treated cells and vehicle treated cells without *CDKN1A* expression compared to the full dataset. On the other hand, gene set enrichment analysis revealed that TP53 pathways are positively enriched in this subset, indicating that some of these cells do react on nutlin-3. In these cells, *CDKN1A* may not be detected due to the high dropout rates, typically seen in single cell experiments. Although these analyses gave first insights in the heterogeneous response of NGP cells on nutlin-3 treatment, more in depth analyses are required to better understand the observations. Amongst others, our results should be confirmed by performing similar analyses with other bona fide TP53 targets, such as *PUMA* and *BAX*. Of note, due to the relatively low detection sensitivity, *BAX* cannot be used for such a confirmatory analysis as it is completely missing in the data. These analyses should also be repeated using the C1 and ddSeq datasets, although this may not be robust as only a few cells per subset will be retained.

Since single cell and bulk RNA-seq experiments differ at several points, such as the library prep method, the sequencing depth, and the 'sample' size, we set up an additional analysis in which we attempted to cancel out these differences. To account for the sample size, several Chromium cell data were pooled to create 10 pseudobulk samples for each condition. In addition, the bulk dataset was downsampled to obtain the same number of reads as the single cell RNA seq dataset. Correlation analysis between the gene expression profiles of the pseudobulk and bulk samples depicted a higher correlation compared to the correlation between the bulk and the Chromium single cell dataset, validating that the pseudobulk data resembles the bulk data. More genes were differentially expressed upon pooling, likely because of a reduction in measurement noise, which is typically high in single cell experiments (7, 13). Still, the number of differentially expressed genes is lower compared to the bulk dataset, owing to the marked higher detection sensitivity of bulk RNA-seq methods. In addition, to make the single cell datasets more comparable, the library size of the Chromium dataset was subsampled to the library size of the C1 dataset to obtain equal library sizes over all cells. As there was only a small difference in library size, the subsampled data gave similar results. Due to the higher number of cells in the Chromium dataset, the average number of reads per cell is much lower for Chromium compared to C1. Therefore, future subsampling of reads to obtain an equal number of reads per cell should be performed. This will also reveal if the C1 is still the most sensitive method, despite the reduction in sequencing depth per cell.

In conclusion, we evaluated for the first time three commercial single cell RNA-seq devices in terms of their ability to characterize a cellular perturbation system. We revealed that despite the lower number of differentially expressed genes in single cell RNA-seq experiments compared to bulk RNA-seq experiments, the biological signal can faithfully detected through gene set enrichment analysis for all single cell devices. We also showed that single cell RNA-seq analyses reveal transcriptional heterogeneity in response to nutlin-3 treatment and may help to identify potentially late-responders or resistant cells that are hidden in bulk RNA-seq experiments.

REFERENCES

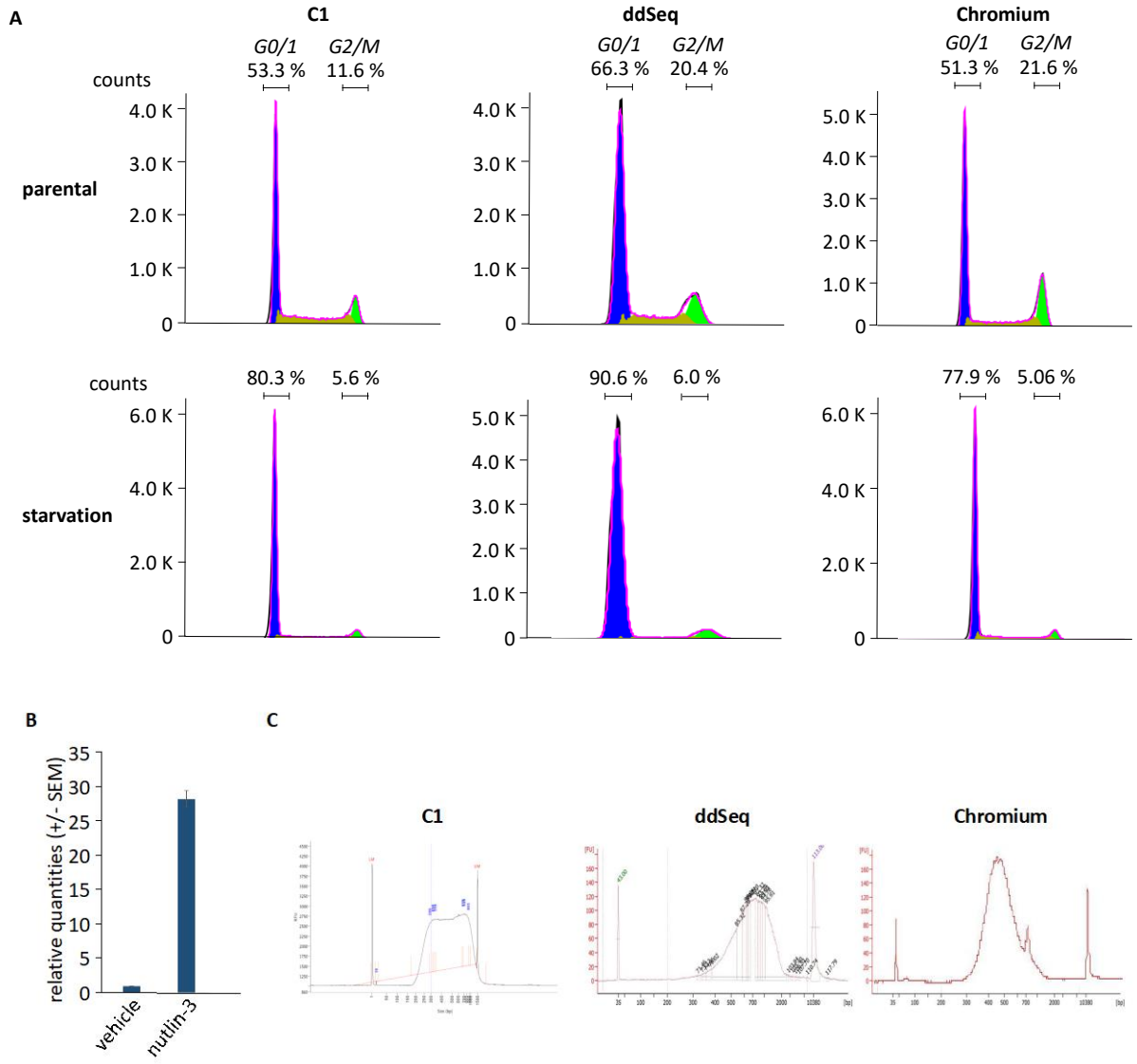
1. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A., *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. **6**.
2. Venteicher,A.S., Tirosh,I., Hebert,C., Yizhak,K., Neftel,C., Filbin,M.G., Hovestadt,V., Escalante,L.E., Shaw,M.L., Rodman,C., *et al.* (2017) Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science*, **355**, eaai8478.
3. Rambow,F., Rogiers,A., Marin-Bejar,O., Aibar,S., Femel,J., Dewaele,M., Karras,P., Brown,D., Chang,Y.H., Debiec-Rychter,M., *et al.* (2018) Toward Minimal Residual Disease-Directed Therapy in Melanoma. *Cell*, **174**, 843–855.e19.
4. Tang,F., Barbacioru,C., Bao,S., Lee,C., Nordman,E., Wang,X., Lao,K. and Surani,M.A. (2010) Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis. *Cell Stem Cell*, **6**, 468–478.
5. Semrau,S., Goldmann,J.E., Soumillon,M., Mikkelsen,T.S., Jaenisch,R. and van Oudenaarden,A. (2017) Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat. Commun.*, **8**, 1096.
6. Yan,L., Yang,M., Guo,H., Yang,L., Wu,J., Li,R., Liu,P., Lian,Y., Zheng,X., Yan,J., *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, **20**, 1131–1139.
7. Hashimshony,T., Senderovich,N., Avital,G., Klochendler,A., de Leeuw,Y., Anavy,L., Gennert,D., Li,S., Livak,K.J., Rozenblatt-Rosen,O., *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, **17**, 77.
8. Picelli,S., Björklund,Å.K., Faridani,O.R., Sagasser,S., Winberg,G. and Sandberg,R. smart-seq2 for sensitive full-length transcriptome profiling in single cells. 10.1038/nMeth.2639.
9. Schroth,G.P., Gertz,J., Myers,R.M., Williams,B.A., McCue,K., Marinov,G.K. and Wold,B.J. (2013) From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.*, **24**, 496–510.
10. Ramsköld,D., Luo,S., Wang,Y.-C., Li,R., Deng,Q., Faridani,O.R., Daniels,G.A., Khrebtukova,I., Loring,J.F., Laurent,L.C., *et al.* (2012) Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells HHS Public Access. *Nat Biotechnol*, **30**, 777–782.
11. Zheng,S., Papalexi,E., Butler,A., Stephenson,W. and Satija,R. (2018) Molecular transitions in early progenitors during human cord blood hematopoiesis. *Mol. Syst. Biol.*, **14**, e8041.
12. Soumillon,M., Cacchiarelli,D., Semrau,S., van Oudenaarden,A. and Mikkelsen,T.S. (2014) Characterization of directed differentiation by high-throughput single-cell RNA-Seq - SI. *bioRxiv*, 10.1101/003236.
13. Islam,S., Kjällquist,U., Moliner,A., Zajac,P., Fan,J.-B., Lönnerberg,P. and Linnarsson,S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, **21**, 1160–7.
14. Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.
15. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M., *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.
16. Kashima,Y., Suzuki,A., Liu,Y., Hosokawa,M., Matsunaga,H., Shirai,M., Arikawa,K., Sugano,S., Kohno,T., Takeyama,H., *et al.* (2018) Combinatory use of distinct single-cell RNA-seq analytical platforms reveals the heterogeneous transcriptome response. *Sci. Rep.*, **8**, 3482.
17. Zhang,X., Li,T., Liu,F., Chen,Y., Yao,J., Li,Z., Huang,Y. and Wang,J. (2018) Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol. Cell*, 10.1016/J.MOLCEL.2018.10.020.
18. Fan,X., Zhang,X., Wu,X., Guo,H., Hu,Y., Tang,F. and Huang,Y. (2011) Single-cell

RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. 10.1186/s13059-015-0706-1.

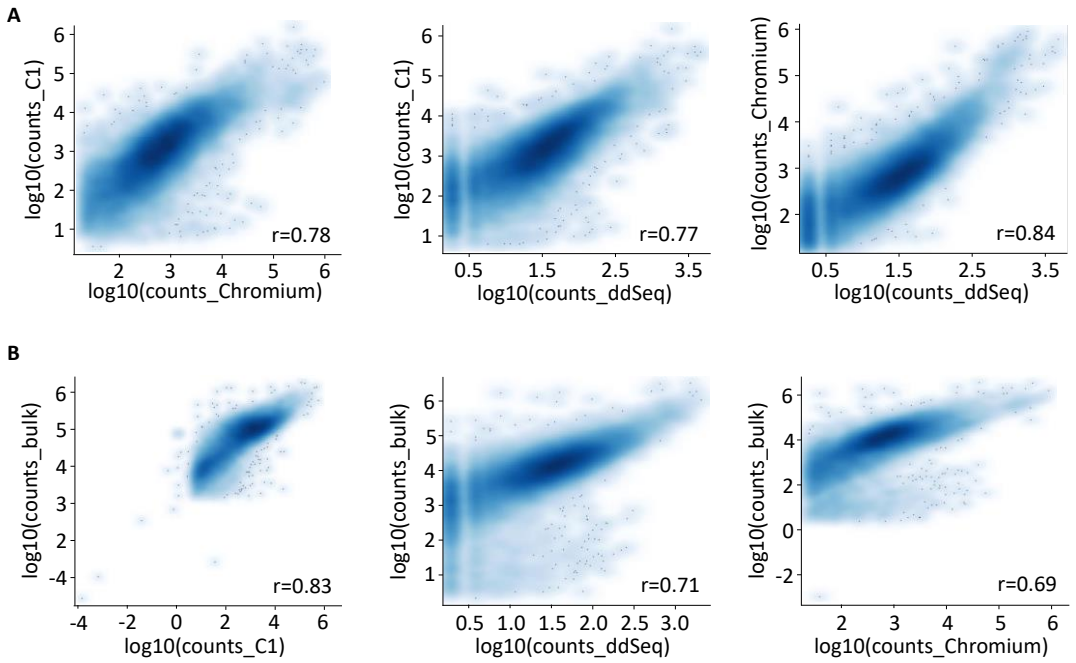
19. Hayashi,T., Ozaki,H., Sasagawa,Y., Umeda,M., Danno,H. and Nikaido,I. (2018) Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.*, **9**, 619.
20. Verboom,K., Everaert,C., Bolduc,N., Livak,K.J., Yigit,N., Rombaut,D., Anckaert,J., Lee,S., Venø,M.T., Kjems,J., *et al.* (2019) SMARTer single cell total RNA sequencing. *Nucleic Acids Res.*, 10.1093/nar/gkz535.
21. Svensson,V., Natarajan,K.N., Ly,L., Miragaia,R.J., Labalette,C., Macaulay,I.C., Cvejic,A. and Teichmann,S.A. (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat. Publ. Gr.*, **14**, 381–387.
22. Wu,A.R., Neff,N.F., Kalisky,T., Dalerba,P., Treutlein,B., Rothenberg,M.E., Mburu,F.M., Mantalas,G.L., Sim,S., Clarke,M.F., *et al.* (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, **11**, 41–46.
23. Ziegenhain,C., Vieth,B., Parekh,S., Reinius,B., Guillaumet-Adkins,A., Smets,M., Leonhardt,H., Heyn,H., Hellmann,I. and Enard,W. (2017) Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell*, **65**, 631–643.e4.
24. Zhang,X., Li,T., Liu,F., Chen,Y., Li,Z., Huang,Y. and Wang,J. (2018) Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. *bioRxiv*, 10.1101/313130.
25. Mereu,E., Lafzi,A., Moutinho,C., Ziegenhain,C., Maccarthy,D.J., Alvarez,A., Batlle,E., Grün,D., Lau,J.K. and Boutet,S.C. (2019) Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects.
26. Van Maerken,T., Speleman,F., Vermeulen,J., Lambertz,I., De Clercq,S., De Smet,E., Yigit,N., Coppens,V., Philippé,J., De Paepe,A., *et al.* (2006) Small-Molecule MDM2 Antagonists as a New Therapy Concept for Neuroblastoma. *Cancer Res.*, **66**, 9646–9655.
27. Barbieri,E., De Preter,K., Capasso,M., Johansson,P., Man,T.-K., Chen,Z., Stowers,P., Tonini,G.P., Speleman,F. and Shohet,J.M. (2013) A p53 drug response signature identifies prognostic genes in high-risk neuroblastoma. *PLoS One*, **8**, e79843.
28. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
29. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
30. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
31. Romagnoli,D., Boccalini,G., Bonechi,M., Biagioni,C., Fassan,P., Bertorelli,R., Sanctis,V. De, Leo,A. Di, Migliaccio,I., Malorni,L., *et al.* (2018) ddSeeker: a tool for processing Bio-Rad ddSEQ single cell RNA-seq data. *BMC Genomics*, **19**, 960.
32. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
33. McCarthy,D.J., Campbell,K.R., Lun,A.T.L. and Wills,Q.F. (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, **33**, 1179.
34. Lun,A.T.L., McCarthy,D.J. and Marioni,J.C. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data. *F1000Research*, **5**, 2122.
35. Van den Berge,K., Perraudeau,F., Soneson,C., Love,M.I., Risso,D., Vert,J.-P., Robinson,M.D., Dudoit,S. and Clement,L. (2018) Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.*, **19**, 24.
36. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–40.
37. Van den Berge,K., Perraudeau,F., Soneson,C., Love,M.I., Risso,D., Vert,J.-P., Robinson,M.D., Dudoit,S. and Clement,L. (2018) Observation weights unlock bulk

RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.*, **19**, 24.
38. Assefa,A.T., Vandesompele,J. and Thas,O. (2019) Probabilistic index models for testing differential expression in single cell RNA sequencing data. *bioRxiv*, 10.1101/718668.
39. Fischer,M. (2017) Census and evaluation of p53 target genes. *Oncogene*, **36**, 3943–3956.
40. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–50.

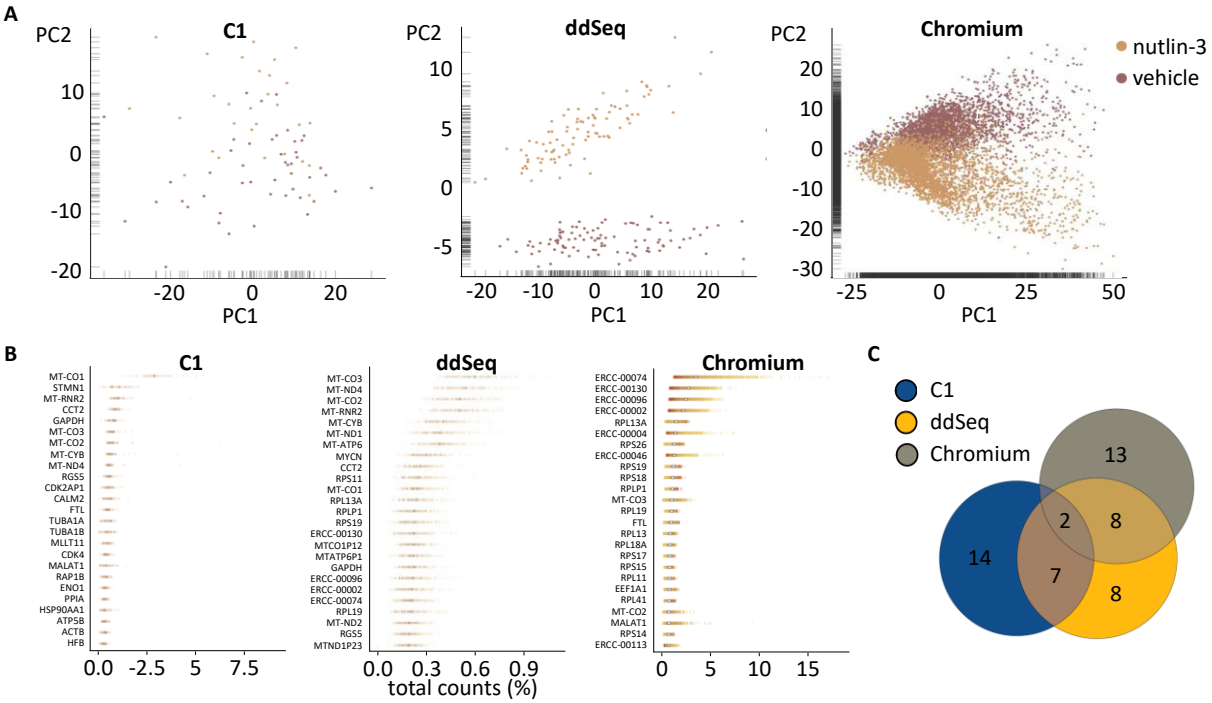
SUPPLEMENTARY FIGURES



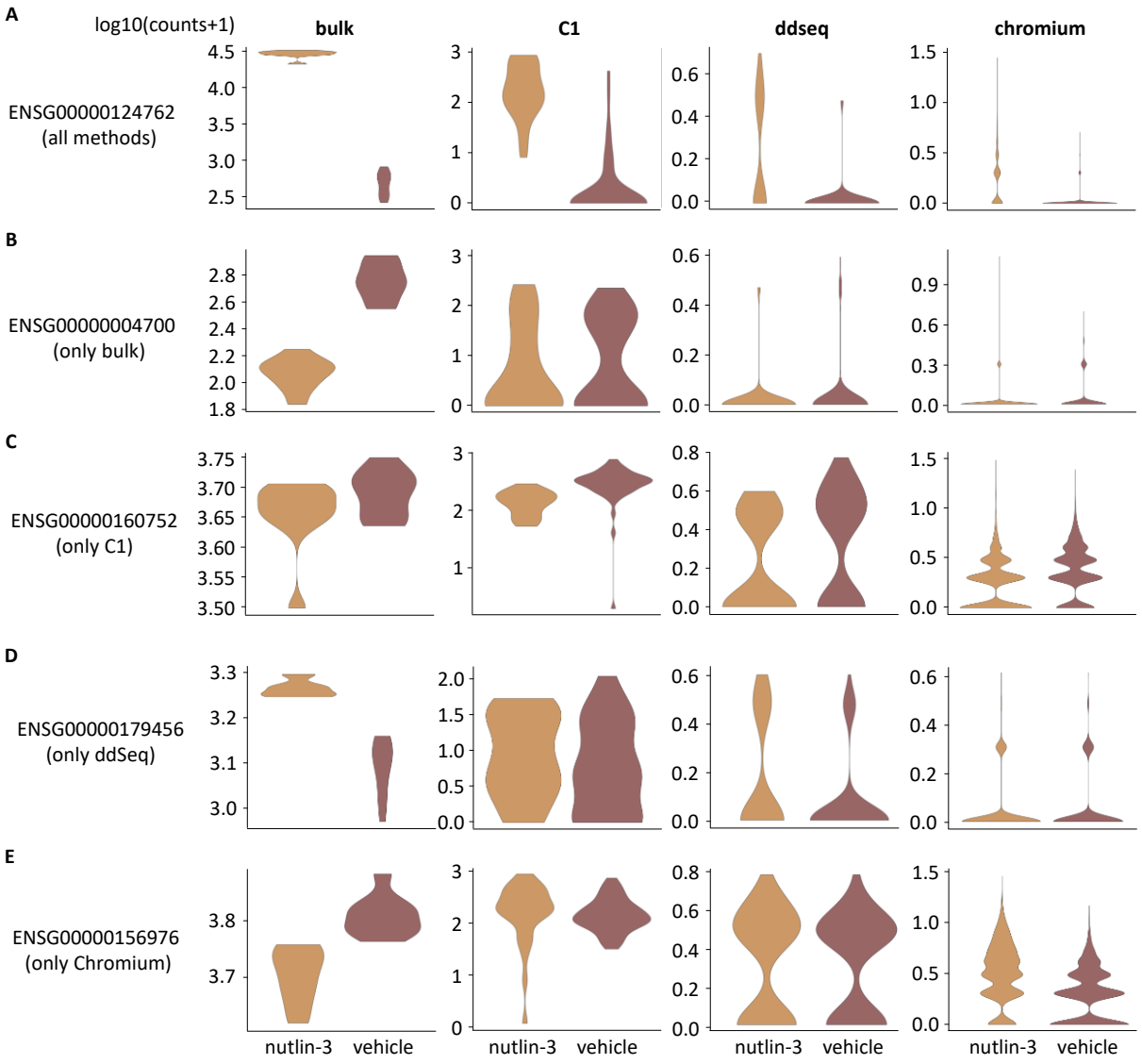
Supplementary Figure 1: cell cycles profiles and technical validation of the single cell RNA libraries. (A) Cell cycle profiles of parental cells (up) and nutlin-3 treated cells (down). (B) RT-qPCR validation of TP53 target gene *CDKN1A* expression levels. Barplot shows the mean expression for the three single cell experiments. Bars represent the standard error of measurement (SEM). (C) Fragment Analyzer profile for C1 RNA libraries (left), Bioanalyzer profiles for ddSeq (middle) and Chromium (right).



Supplementary Figure 2: gene expression correlation plots show high correlation between the devices. (A) Smoothscatter plots showing the correlation in gene expression between C1, ddSeq and Chromium. (B) Smoothscatter plots showing the correlation in gene expression between bulk and each of the single cell devices.

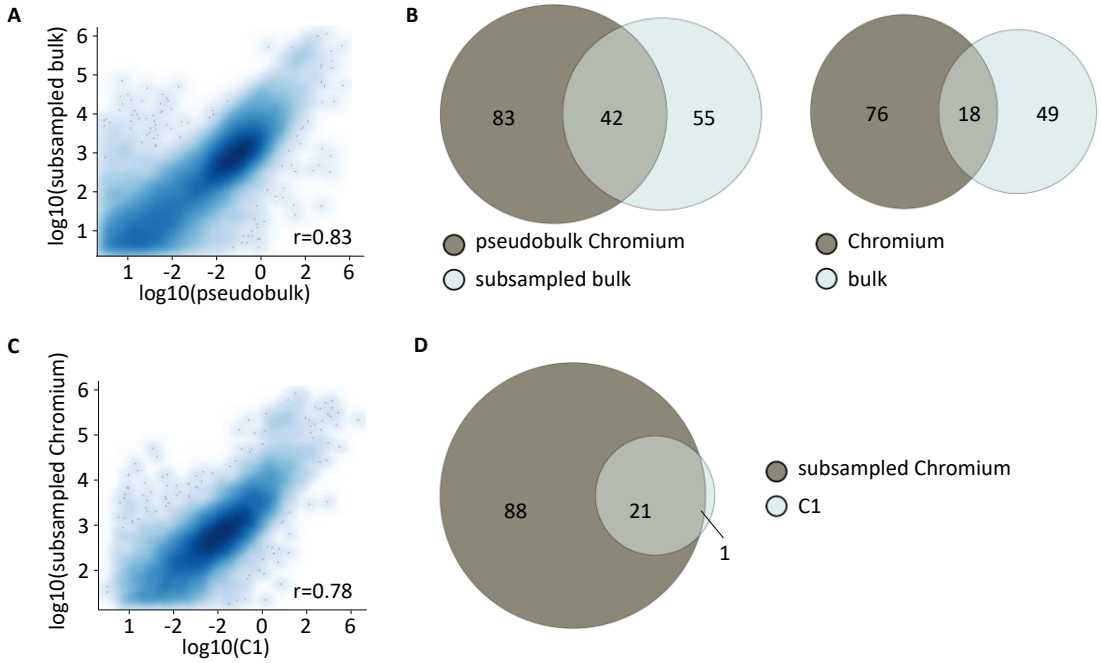


Supplementary Figure 3: quality assessment of the single cell RNA-seq data. (A) PCA plots show a separation between nutlin-3 and vehicle treated cells for all devices. (B) Percentage of counts mapping on the top 25 highest expressed genes for the three single cell devices. (C) Overlap of the top 25 expressed genes for the three single cell devices.



	bulk				C1				ddSeq				Chromium			
	PI	p.adj	LFC	FDR	PI	p.adj	LFC	FDR	PI	p.adj	LFC	FDR	PI	p.adj	LFC	FDR
ENSG00000124762	1.00	1.19e-221	5.71	1.81e-321	0.97	4.73e-5	3.93	2.00e-3	0.73	6.64e-11	3.14	3.37e-18	0.79	0.00	2.33	0.00
ENSG00000004700	1.37e-10	0.00	-1.12	1.54e-60	0.43	0.77	-0.10	0.99	0.49	0.87	-0.42	0.91	0.45	5.73e-10	-0.50	5.83e-15
ENSG00000160752	0.57	0.78	-0.12	0.07	0.09	7.93e-6	-1.10	4.18e-5	0.42	0.58	-0.42	0.33	0.40	1.41e-43	-0.52	1.69e-43
ENSG00000179456	1.00	1.94e-167	0.57	5.30e-35	0.57	0.874	0.99	0.53	0.64	3.60e-3	1.27	3.86e-3	0.50	1.00	-0.02	0.80
ENSG00000156976	1.00	8.15e-167	0.34	2.70e-17	0.59	0.70	0.43	0.62	0.56	0.65	0.15	0.75	0.74	2.48e-244	1,21	9.51e-301

Supplementary Figure 4: violin plots for significantly differentially expressed genes in all or only one of the datasets. (A) Expression of ENSG00000124762 (*CDKN1A*), significantly differentially expressed in all datasets. (B) Expression of ENSG00000004700 (*RECQL*), uniquely significantly differentially expressed in the bulk data. (C) Expression of ENSG00000160752 (*FDPS*), uniquely significantly differentially expressed in the C1 data. (D) Expression of ENSG00000179456 (*ZBTB18*), uniquely significantly differentially expressed in the ddSeq dataset. (E) Expression of ENSG00000156976 (*EIF4A2*), uniquely significantly differentially expressed in the Chromium dataset.



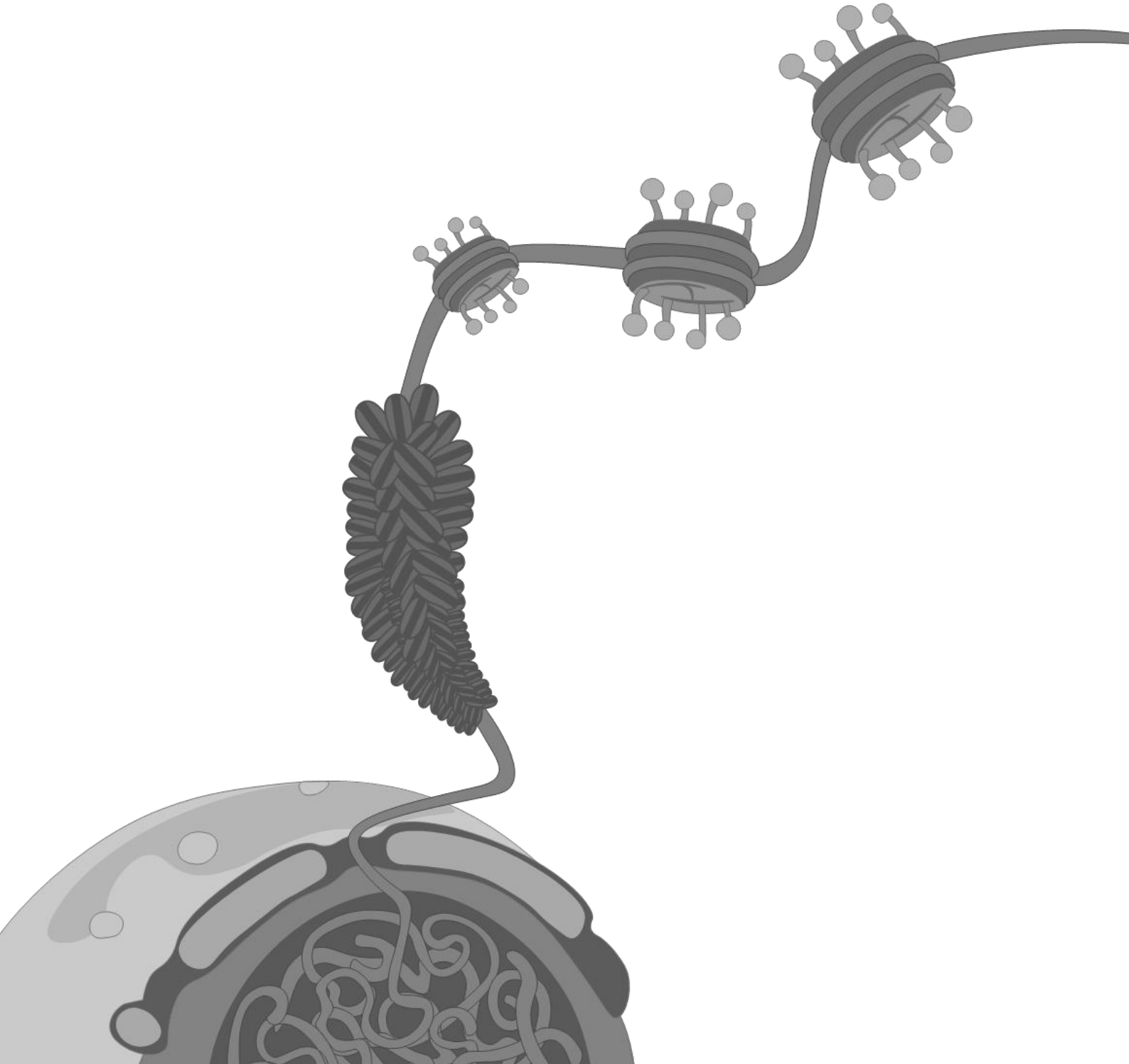
Supplementary Figure 5: pseudobulk data resembles true bulk data. (A) Smoothscatter plot showing the correlation in gene expression between the pseudobulk and downsampled bulk dataset. (B) Overlap of significantly positively (q-value < 0.05) gene sets in the pseudobulk and downsampled bulk dataset and the original bulk and Chromium dataset. (C) Smoothscatter plot showing the correlation in gene expression between the C1 and downsampled Chromium dataset. (D) Overlap of significantly positively (q-value < 0.05) gene sets in the C1 and downsampled Chromium dataset.

Supplementary Table 1: number of significantly differentially expressed genes determined by EdgeR and PIM. EdgeR is used in combination with Zinger for the single cell RNA sequencing datasets. FDR: false discovery rate; LFC: log fold change; PIM: probabilistic index model; p.adj: p.adjusted; PI: probabilistic index.

.

	subsampled_bulk	*CDKN1A* null cells nutlin-3 vs vehicle	Nutlin-3 cells *CDKN1A* low vs high	G1 cells nutlin-3 vs vehicle	Chromium nutlin-3 vs vehicle	ddSeq nutlin-3 vs vehicle	C1 nutlin-3 vs vehicle	bulk nutlin-3 vs vehicle
EdgeR FDR < 0.05 abs(LFC) > 1	6358	88	112	104	97	107	52	8303
PIM p.adj < 0.05 PI > 0.6 or PI < 0.4	1361	442	504	492	502	193	301	17327
EdgeR+PIM	810	71	105	83	88	28	40	7010
positively enriched gene sets (q-value < 0.05)	125	73	100	103	94	39	22	67
	16487	5554	5277	97				

4. Discussion and future perspectives



T-ALL is a haematological cancer resulting from the malignant transformation of immature T-cells. Advances in NGS technologies broadened our understanding of the genetic basis of T-ALL and enabled further genetic dissection of T-ALL through detection of novel key driver genes and delineation of distinct molecular T-ALL subgroups (1, 2). Despite the advances in our knowledge about the genetic basis and the increasing survival rates of T-ALL cases, the prognosis of adult patients remains poor, with 20 % of the pediatric and 40 % of the adult T-ALL patients still suffering from relapse, hinting to the need for further refinement of the molecular basis of T-ALL and exploit this knowledge to develop more effective and targeted therapies (3, 4). Over the past years, it became clear that genetic alterations in the non-coding part of the genome, which comprises the largest fraction of the genome and has been seen as 'junk DNA' for a long time, plays major roles in the development of diseases and is just starting to be explored (5). Extensive advances in NGS methods revealed that CNVs and base pair variants in cancer cells often occur in regulatory elements or non-coding genes (6, 7). One class of these non-coding genes are the lncRNAs, for which it is now widely accepted that they can exert important functions and when perturbed, can be implicated in cancer development. A few studies investigated lncRNA expression in the context of T-ALL, but the number of functionally annotated lncRNAs in T-ALL development remains very limited with only two studies reporting on *LUNAR1* and *NALT1* (8–10). By microarray profiling of a large T-ALL patient cohort, Wallaert et al. defined a subgroup specific lncRNA expression profile, demonstrating that lncRNAs can be used to delineate cancer subtypes, as shown for other cancer entities such as breast cancer (11, 12). Cao Thi Ngoc et al. focused on the TAL-R subgroup and revealed 57 lncRNAs that are directly regulated by TAL1, of which some are absent during T-cell development, hinting to a potential role as ectopically expressed oncogenic lncRNAs (13). During my PhD, I generated a unique and comprehensive dataset in the T-ALL field by combining for the first time polyA[+] and total RNA-seq of an *in vitro TLX1* knockdown system and a large primary T-ALL cohort. This dataset was extended with ATAC-seq as well as H3K4me3, H3K4me1, H3K27ac and TLX1 ChIP-seq, enabling to detect TLX1 regulated/TLX subgroup specific and super-enhancer associated lncRNAs. Some of these lncRNAs are potentially oncogenic, marking them as highly interesting targets for further in-depth characterization (**Paper 1, Figure 15**). Since this is a comprehensive dataset containing unexplored features, I wrote a data descriptor with detailed information about the data quality, the specifications of the methods and the bioinformatics pipelines applied in order to make the data available and re-usable for the research community (**Paper 2, Figure 15**).

For this first part, I used bulk transcriptome profiles resulting in average expression levels across a cell population. However, at that time, the first single cell devices were emerging, and single cell methods revealed that bulk average expression profiles can hamper the detection of true biological effects and hide cellular heterogeneity. Since the utility and richness of single cell data arouse my interest, I focused on the optimization of single cell RNA-seq technologies in the second part of my PhD. At that time, the methods that existed only captured polyadenylated transcripts. Therefore, I developed a new single cell total RNA-seq protocol enabling to capture both polyadenylated and non-polyadenylated transcripts at the single cell level by combining for the first time strandedness and effective removal of ribosomal cDNA (**Paper 3, Figure 15**). During my PhD mandate, the number of single cell sequencing methods and devices expanded quickly, whereby the number of single cells that can be isolated in a single experiment increased from a few to more than 10,000 single cells (14–16). Therefore, I eventually evaluated three commercial single cell devices (C1, ddSeq and Chromium) with respect to data quality and the ability to detect differentially expressed genes and revealed that single cell data can detect biological signal faithfully through gene set enrichment analysis and may help to identify potentially late-responders or resistant cells upon compound treatment (**Paper 4, Figure 15**).

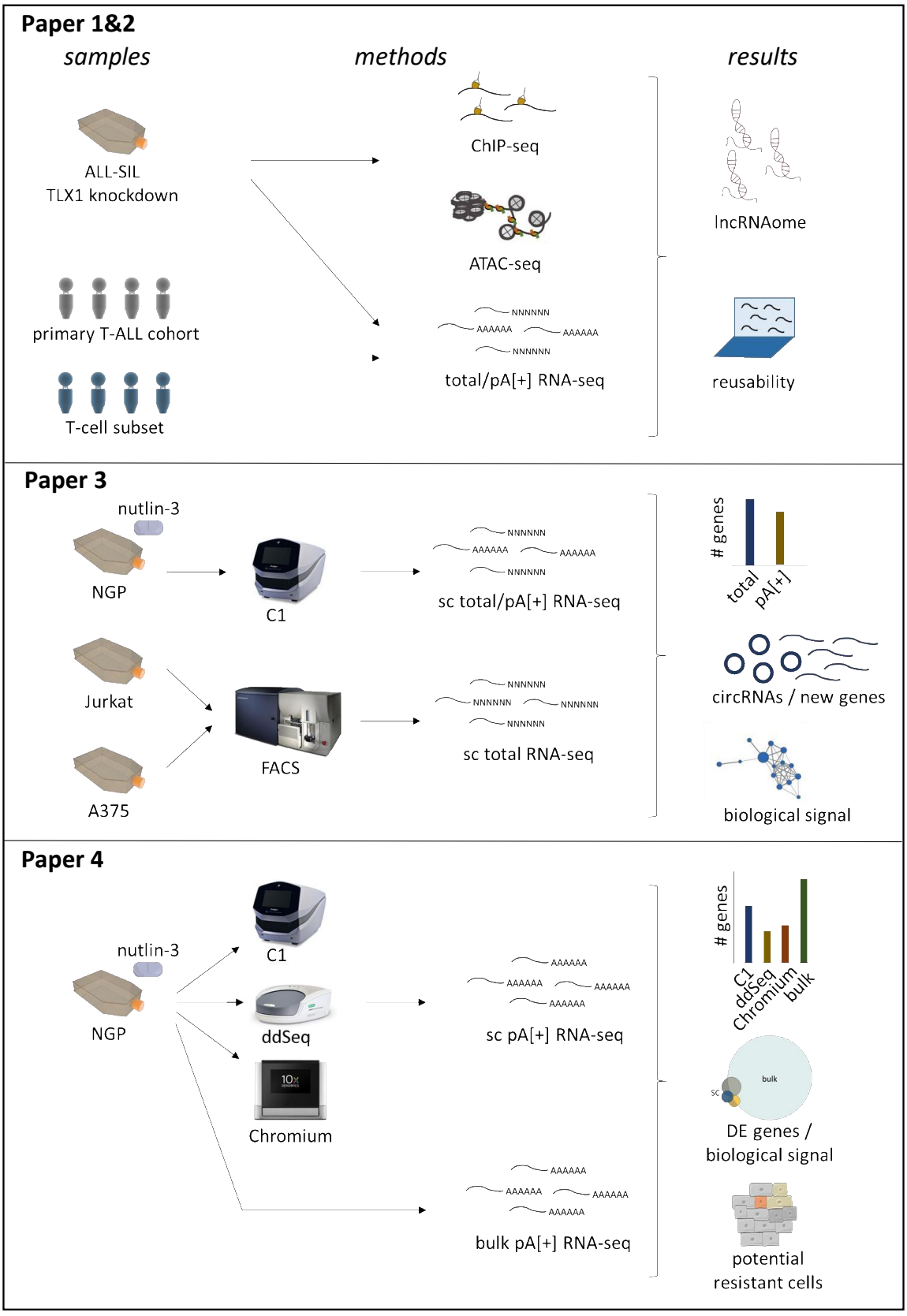


Figure 15: overview of the results of this PhD thesis.

4.1. **Investigating the TLX1 lncRNAome in T-ALL**

TLX1 is a driver oncogene in T-ALL development and is ectopically expressed in 5-10 % of pediatric and 30 % of adult T-ALL patients, resulting in an arrest at the early cortical stage of T-cell development (17–19). The long latency of TLX1 positive T-ALL development in mice, indicated that *TLX1* overexpression is not sufficient for T-cell transformation and that cooperating genetic alterations are required to fully transform progenitor T-cells to leukemic blasts (20, 21). To identify these cooperating events, TLX1 positive T-ALL has already been studied extensively in terms of protein-coding genes and revealed secondary mutations and/or deletions in *WT1*, *PHF6*, *PTEN*, *PTPN2* and *BCL11B* and *NUP214-ABL1* and a high frequency of *NOTCH1* mutations. However, the role of lncRNAs in this TLX1 positive T-ALL remained unexplored (22–27). Therefore, I investigated the role of lncRNAs in this T-ALL subgroup by uniquely combining polyA[+] and total RNA-seq of a primary T-ALL cohort and the TLX1 positive ALL-SIL cell line upon *TLX1* knockdown.

To obtain a knockdown of *TLX1*, electroporation of two *TLX1* targeting siRNAs and one control siRNA was used. In the course of my PhD research, I optimized new transfection methods to further increase the transfection efficiency and reduce the cell death. Therefore, we now typically use a nucleofector in the lab, where the voltage, number of pulses and duration of the pulses can be adapted depending on the cell type to obtain an efficient knockdown with low cell death. Furthermore, we are currently exploring photoporation as this method allows to specifically select cells that have taken up the siRNA. This method uses transient permeabilisation of the cell membrane by laser irradiation of gold particles that adsorb to the membrane. After irradiation, the photothermal effects of these gold nanoparticles transiently make pores in the cell membrane, enabling entrance of macromolecules such as siRNAs (28). Transcriptome profiling upon *TLX1* knockdown revealed both polyadenylated and non-polyadenylated lncRNAs regulated by TLX1. In a next step, I further interrogated these candidates using in-house generated TLX1 ChIP-seq data to identify those lncRNAs directly bound by TLX1. Amongst those, I identified *NEAT1* and *MALAT1*, two well-known lncRNAs involved in cancer, and in my dataset shown to be downregulated upon *TLX1* knockdown. To determine if these two lncRNAs also have important roles in the development of TLX1 positive T-ALL, knockdown experiments of these lncRNAs in the TLX1 positive T-ALL cell line ALL-SIL should be carried out.

Unexpectedly, I revealed an opposite regulation of protein-coding genes and lncRNAs by TLX1, which had not been described previously. The majority of the identified TLX1 downstream lncRNAs are shown to be activated by TLX1, while most protein-coding genes are repressed. Intrigued by this novel finding, I performed motif analysis to identify co-factors that may explain the opposite regulation of lncRNAs and protein-coding genes by TLX1. However, since the same motifs were enriched, this could not explain the difference. Another approach that could be followed to further explore potential differential regulation and is currently explored by the host lab, is the use of BioID in combination with dCas9. Here, dCas9 is ligated to a biotin ligase and guided to the TSS of the gene of interest. Subsequently, co-factors in close proximity are biotinylated, after which these can be isolated and identified by mass spectrometry (29). Using this approach for the top downregulated lncRNAs and top upregulated protein-coding genes could reveal different co-factors involved in the regulation of these genes (30). This opposite regulation of lncRNAs and protein-coding genes made us hypothesize that TLX1 can activate a subset of lncRNAs, which are possibly involved in the negative regulation of TLX1 regulated protein-coding genes, a hypothesis that requires further investigation. This can be studied by transcriptome profiling after knockdown of a selected lncRNA to identify if the expression of some

protein-coding genes increases. If this is the case, close proximity between the lncRNA and protein-coding gene can be confirmed using 3C sequencing.

By integrating H3K27ac ChIP-seq data, super-enhancer associated lncRNAs were identified. Some of those super-enhancer associated lncRNAs probably act *in cis* as their expression is significantly correlated with the expression of neighboring genes. To reveal if this is the case, knockdown experiments of the lncRNA should have an effect on the expression of the neighboring genes, while upregulation of the gene should have no effect. In contrast, to reveal if some of these lncRNAs also work *in trans*, the effect of random integration of these candidates should be investigated (31). These interactions can then further be explored using 3C or 4C sequencing experiments. Besides *TLX1* knockdown in ALL-SIL lymphoblasts, I also performed treatment with JQ1 and evaluated transcriptional response upon drug exposure. JQ1 is a bromodomain and extra-terminal motif (BET) inhibitor causing depletion of amongst others BRD4, known to be enriched on enhancers and to recruit RNAPII for eRNA production (32, 33). Since 50 TLX1 regulated lncRNAs are also significantly downregulated upon JQ1 treatment, the latter known to downregulate eRNA transcription, I hypothesize that some of these lncRNAs may act as eRNAs (34, 35). These eRNAs can be determined using my dataset as these are characterized by open chromatin (ATAC-seq peak), H3K27ac peaks and a high H3K4me1/H3K4me3 ratio (36, 37). Since most eRNAs lack a polyA tail, the epigenetic marks can be integrated with the total RNA-seq dataset to identify these eRNAs. However, eRNAs are often low abundant and unstable, making it difficult to identify them using classic total RNA-seq protocols (37, 38). Therefore, nascent RNA-seq methods such as global run-on sequencing (GRO-seq), precision run-on sequencing (PRO-seq) and BruUV-seq, should be performed as an additional layer, enabling to detect nascent RNAs before these are degraded (39–43).

4.2. **Identification of subgroup specific and possibly oncogenic lncRNAs**

Besides TLX1 regulated lncRNAs, I aimed to verify TLX1/3 linked lncRNA signatures in primary T-ALLs as *TLX1* and *TLX3* induce T-ALL in a similar way and are associated with a similar gene expression profile (44). Similar as to the *in vitro* knockdown system, I identified TLX subgroup specific lncRNAs and lncRNAs associated with super-enhancers. Of interest, I integrated CD34+ T-cell data to identify potential oncogenic lncRNAs and identified a set of 144 lncRNAs with low expression in T-cells and high expression in T-ALL, more specifically high in the TLX subgroup versus the other subtypes. As these lncRNAs are potentially oncogenic, they can serve as new potential targets for T-ALL therapy development. However, I used CD34+ cells cultured on OP9 stromal cells expressing the NOTCH1 ligand DLL1, whereby only information about the lncRNA expression at that stage was obtained. Since T-ALL subgroups are characterized by an arrest at a specific stage of the T-cell development, further studies should compare the lncRNA expression profile of a specific subgroup with T-cells of the corresponding development stage. In my study, a lncRNA can be identified as non-oncogenic when it is expressed in CD34+ T-cells, while the lncRNA expression may drop at a later stage, whereby overexpression at that specific stage can cause transformation of T-cells and marks the lncRNA as potential oncogenic. This is for instance also the case for *NOTCH1*, which is required at the initial stage for T-cell specification, but drops during further development to be able to develop in the αβ lineage, whereas overexpression causes T-cell transformation (45). As lncRNAs can be erroneous classified as potential oncogenic based on expression in CD34+ T-cells, a larger T-cell subset is warranted for this analysis. Knockdown studies are eventually required to further validate that these lncRNAs are truly oncogenic.

To get a first insight into the function of the top-regulated lncRNAs, I used a guilt-by-association approach, in which lncRNA functions are predicted based on correlations with the expression of protein-coding genes. These functional predictions give a first hint into the possible cellular roles of the selected lncRNAs and mark them as interesting targets for further functional analysis and eventually as potential targets for new therapies. However, before further functional validation, it should be validated that the lncRNAs are truly independent transcriptional units as some lncRNAs have been shown to be transcriptional read-through from neighboring genes (46). To validate this, the H3K4me3 and H3K27ac ChIP-seq data should be extended with CAGE-seq data as this identifies transcription start sites (47, 48). In addition, lncRNAs are characterized by a lack of protein-coding potential, however studies have shown that lncRNAs can contain a small ORF and can be occupied with ribosomes. However, occupancy with ribosomes does not ensure translation as some lncRNAs, such as the well-known *XIST*, contain an ORF and are occupied with ribosomes, but are however not translated, underscoring that ribosome occupancy is not sufficient for translation (49). On contrast, some other lncRNAs that contain an ORF and are occupied by ribosomes, generate short peptides. These were previously not detected as most prediction algorithms discard ORFs with less than 100 amino acids and these peptides are often low abundant and lost during sample preparation for mass spectrometry (50–52). A large study using tandem mass spectrometry revealed that 8 % of the lncRNAs are translated in small peptides and consequently mis-assigned as non-coding (52, 53). Therefore, it may be interesting to validate with mass spectrometry that the lncRNAs that I identified are really non-coding and not translated in small peptides.

4.3. **Functional characterization of the identified TLX1 regulated and TLX subgroup specific lncRNAs requires further investigation**

From the large dataset I generated, multiple lncRNAs were marked as interesting targets for further in-depth characterization. Therefore, I prioritized five lncRNAs: *lnc-DAD1-2* was selected as this lncRNA is located in the TRCα locus and might consequently be involved in TCR rearrangements and T-cell development; *lnc-THADA-1* and *lnc-PTPN2* as these are respectively located nearby *ZFP36L2* and *PTPN*2, two known T-ALL tumor suppressor genes and *RP11-973H7.4* and *FOXP4-AS* as these were identified as potentially oncogenic and TLX subgroup specific. Unfortunately, knockdown of these lncRNAs by LNAs seemed to be challenging in our hands as by testing ten LNAs per lncRNA, I only obtained sufficient knockdown for *lnc-DAD1-2* with 3 LNAs, while the expression of the other lncRNAs was barely affected upon transfection. Further optimization of the electroporation conditions did not result in better knockdown efficiencies. Although knockdown of *lncDAD1-2* had no effect on proliferation and apoptosis in T-ALL cell lines, knockdown of the lncRNA in T-cells resulted in an increase in double positive T-cells, hinting to a role in T-cell development. This was further confirmed by 4C-seq, since I was able to show that *lncDAD1-2* interacts with the TCRδ locus. Unfortunately, I was not able to validate the phenotype in T-cells in a second replicate.

Although I was not able to obtain knockdown for the other prioritized lncRNAs at that time, these remain highly interesting targets that should be further characterized. New and more efficient methods have now been developed and can be used in a second attempt to obtain knockdown of these lncRNAs. Recent studies show promising results for the knockdown of lncRNAs using the CRISPR technology. Using the original CRISPR technology for lncRNAs is challenging, since lncRNAs often overlap with enhancer regions or protein-coding genes whereby the effect can be due to a deletion in the enhancer or protein-coding gene or due to deletion of the lncRNA (54). To circumvent this, CRISPRi can be used, where an inactive CAS9 (dCas9) is bound to a repressor domain, such as KRAB, and is

guided to the promotor of a lncRNA by the guide RNA (gRNA) to silence the lncRNA (55, 56). Since this results in variable knockdown efficiencies, a recent study has shown that combining KRAB with the repressor domain of MeCP2 (dCas9-KRAB-MeCP2) results in considerably increased gene repression (57). Likewise, attaching an activator, such as VP64, to dCas9 can be used for the overexpression of lncRNAs (58, 59). Besides targeting specific lncRNAs, CRISPRi and CRISPRa screens can be setup to obtain a knockdown or overexpression of thousands of lncRNAs in parallel and to identify potential functions (31, 60). More recently, also methods combining CRISPR screens with single cell RNA-seq, such as CRISPR droplet sequencing (crop-seq) and Perturb-seq were developed, allowing to detect the effect of the perturbation at the single cell level. By performing bulk read-out, 50 % reduction of expression of a gene can mean that the gene has halve of its expression in all cells or that there is a 100 % reduction in halve of the cells. This information is lost by bulk sequencing, underscoring the need to investigate the effect of perturbations at the single cell level (61–63).

Besides the CRISPR technology, siPools have also been developed, in which 60 siRNAs are combined at low concentrations, resulting in undetectable off-targets effects (64). Since siRNAs mainly work in the cytoplasm, knowing the localization of the lncRNAs is helpful to decide if it is useful to use these siPools. However, it has been shown that the RISC machinery of the siRNA is also present in the nucleus and able to efficiently degrade nuclear RNAs (65, 66). RNA-FISH or RNAscope, based on binding of fluorescently labeled probes, or cell fractionation experiments can be carried out to determine the localization of the lncRNA (67). In addition, defining the cellular localization of a lncRNA can also give a first hint towards possible functions as lncRNAs involved in the regulation of gene expression are more likely to be expressed in the nucleus, while lncRNAs involved in the regulation of translation or miRNA sequestration are more commonly expressed in the cytoplasm (68). Further insights in the function of a lncRNA can be acquired by determining the interaction partners of the lncRNA. As discussed in section 1.2.3, lncRNAs often bind with chromatin to remodel their structure and binding of the lncRNA with chromatin can be revealed using chromatin isolation by RNA purification (ChIRP), capture hybridization analysis of RNA targets (CHART) or RNA antisense purification (RAP), all using biotinylated antisense DNA oligonucleotides to capture and isolate lncRNA-chromatin interactions (69–71). ChIRP and CHART can also be used to identify binding of proteins by western blot analysis or by combining it with mass spectrometry as lncRNAs often serve as a scaffold, decoy or guide for proteins (69, 71). Domain specific ChIRP (dChIRP) enables to identify functional domains required for chromatin binding and enables to identify interactions with DNA, RNA and proteins by means of RT-qPCR, high-throughput sequencing and Western blot or mass spectrometry, respectively (72).

Validation of lncRNAs *in vivo* has been challenging as lncRNAs are generally less conserved compared to protein-coding genes, which limits the use of animal models to study these lncRNAs (73, 74). It has been shown that 81 % of the lncRNAs are primate-specific and that conserved lncRNAs often only display a short conserved region (75). For zebrafish for instance, it has been shown that only 29 lincRNAs show detectable sequence conservation with human lincRNAs (76). To solve this problem, the human transcript can be overexpressed in a model system or xenografts can be used by implanting human cell lines after knockdown or overexpression of the lncRNA in the model system (77, 78).

4.4. **Sharing data accelerates scientific breakthroughs**

Previously, researchers were reluctant to share data as experiments are time-consuming and expensive, thus researchers wanted to take full advantage of the generated data by publishing multiple

papers using the same dataset. Over the last years, the scientific community is moving to more openness as this accelerates research and new breakthroughs, allows to investigate larger sample cohorts by combining public datasets, and paves a way for new collaborations. Besides sharing, data should also be clearly annotated and methods described in detail to allow other researchers to re-use the data. I generated a unique and comprehensive dataset in the T-ALL field by combining polyA[+] as well as total RNA-seq of a large T-ALL cohort and *TLX1* knockdown system, extended with epigenetic layers through ATAC-seq and ChIP-seq. As I only investigated the role of lncRNAs, with a focus on the TLX subgroup, my dataset contains extensive unexplored information. Hence, I wrote a data descriptor providing detailed information about the methods and analyses performed, enabling other researchers to further explore this comprehensive dataset to further unravel the complex biology of T-ALL in general and TLX1 in particular.

In this PhD thesis, I only investigated TLX1 specific lncRNAs and differences in lncRNA expression between the TLX subgroup and the other subgroups, whereby my dataset remains to be explored for plenty of other analyses. First, this dataset can be further used to reveal the lncRNA expression profile of the other subgroups, although it should be noted that the HOXA subgroup is underrepresented (13 immature, 17 TLX, 23 TAL-R and 7 HOXA patients). Of interest, this analysis has previously been scrutinized in the lab by Wallaert et al. using microarray profiling of the same T-ALL cohort. Comparing this dataset with my RNA-seq dataset is complicated as only a few of the probes on the microarray are also defined as lncRNA in Ensembl (11). In contrast to microarray based profiling, RNA-seq has the advantage to detect all lncRNAs in an unbiased way and enables to detect new lncRNAs as no probes are required. Second, I only investigated lncRNAs, leaving the opportunity to study other biotypes. Third, I have generated total RNA-seq data containing reads mapping to immature/unspliced RNA (introns), allowing to further exploit this type of data to differentiate transcriptional vs post-transcriptional regulation (79). Finally, the total RNA-seq data can be used to detect circRNAs as these circRNAs lack a polyA tail and have been shown to play a role in cancer development. circRNA *PVT1* for example sponges *miRNA-497,* which normally represses the anti-apoptotic protein BCL2, resulting in the inhibition of apoptosis and induction of proliferation in lung cancer (80). As a follow-up on the paper, I tried to investigate subgroup specific or TLX1 regulated circRNAs in my dataset using CircExplorer, but unfortunately identified only a few differentially expressed circRNAs. Since these circRNAs are mostly low abundant, deeper sequencing will probably be needed or circRNAs should be enriched by selectively removing linear RNA by exonuclease treatment prior to library prep and sequencing (81, 82).

In conclusion, by providing a detailed description of the methods and analyses performed in the data descriptor, I believe that other researchers can re-use this dataset to further unravel the complexity of T-ALL. This dataset can be further explored in term of other biotypes, other subgroups and post-transcriptional regulatory analyses can be carried out. Although I believe that re-using this dataset will be beneficial to further unravel the complexity of T-ALL, it should be noted that the data is generated using bulk experiments and will consequently hide heterogeneity and some biological effects by generating average expression profiles. Mutation analysis of diagnosis and relapse samples of T-ALL patients revealed that some of the mutations identified at relapse were already present in a minor clone at diagnosis, while other mutations were only detected at relapse. Since relapse can result from a subclone that is resistant to therapy and further expands, the detection of these subclones is important (26, 83, 84). It has recently been shown that some of these subclones only consist of 1 % of all cells and these are consequently missed by bulk sequencing, underscoring the need to perform single cell sequencing. To further unravel the clonal evolution of T-ALL and to obtain an in-depth view

on the heterogeneity of T-ALL at diagnosis and during treatment, large single cell studies are warranted (85). In addition, single cell RNA sequencing studies are required to investigate the heterogeneous response of single cells on a treatment to identify potential resistan cells. As single cell sequencing methods were starting to emerge at the start of my PhD, I optimized the single cell technology in the lab as a second part of my PhD, to be able to use this technology in future studies in the lab.

4.5. **Deciphering the non-polyadenylated fraction of the transcriptome at the single cell level**

In 2012, the first commercial single cell sequencing device, the C1, had just been released by Fluidigm. Before, individual cells were picked manually or FACS sorted resulting in labor intensive protocols requiring expertise and technical noise due to multiple pipetting steps. The release of the Fluidigm C1 single cell autoprep system enabled to automatically isolate up to 96 single cells in a microfluidic chip and synthesize cDNA in the same chip. This reduced the technical noise owing to human handling and increased the sensitivity as it has been shown that reaction efficiencies increase in lower reaction volumes (86–91). Since then, the number of single cell sequencing methods and devices expanded quickly and new methods now enable to capture tens of thousands of single cells in an experiment and to significantly reduce the cost per single cell (14–16). However, almost all of the currently existing methods focus on sequencing of the ends of polyadenylated transcripts. Since these represent only 1-5 % of the total RNA present in each cell, the non-polyadenylated fraction of the transcriptome, including eRNAs, a considerable fraction of the lncRNAs and all circRNAs, remains unexplored (14, 82, 92–95). To detect non-polyadenylated RNAs, three workflows –SUPeR-seq, RamDA-seq and MATQ-seq– for single cell total RNA-seq have recently been developed. Unfortunately, these methods suffer from either an unstranded nature of the protocol or results in a high fraction of ribosomal reads (96–98). Retaining strand information is warranted to assign reads to the correct gene as a considerable fraction of the genes overlap on opposite strands (99). In addition, a rRNA depletion step is essential as up to 95 % of the total RNA content in a mammalian cell consists of rRNA. Since none of the single cell total RNA-seq methods combined these desirable features at the start of my thesis, I developed a new single cell total RNA-seq method that meets these requirements. I have introduced the protocol for the C1 and for FACS sorted cells using Fluidigm script builder as the C1 and FACS are, in contrast to the frequently used droplet-based methods, open and flexible systems for which users can easily adapt and develop protocols. I showed that the method generates an average of only 3 % ribosomal reads and retains strand information. Furthermore, my method permits to detect relatively more protein-coding genes, pseudogenes, lincRNAs and miscellaneous RNA (miscRNA) compared to single cell polyA[+] RNA libraries, when corrected for equal sequencing depth. By further increasing the sequencing depth up to 8 million reads, no plateau is reached, showing the high transcriptome complexity of the libraries. Besides these known gene classes, also novel genes (not annotated in Ensembl and LNCipedia) were identified. As expected for total RNA libraries, more intronic reads were present compared to polyA[+] libraries. The exon-intron ratio can be used for the analysis of post-transcriptional regulation or RNA velocity analysis (100–102). The latter allows to predict a cell's future state on a timescale of hours based on the balance between spliced and non-spliced transcripts. This is possible as an increase of transcription first results in a concomitant increase of non-spliced pre-mRNA expression followed by an increase of mature spliced mRNA expression and the other way around for a transcriptional drop. Therefore, the balance between non-spliced and spliced transcripts provides an indication for the future state of mature mRNA in a cell (103). Moreover, these introns can be used to perform intron based expression analysis to identify the variance in pre-mRNA expression, which reflects the effect of transcriptional bursting (98). Besides single cell total RNA-seq, single nucleus RNA-seq also detects a higher proportion of intronic reads compared to whole cell preparation methods owing to the unprocessed RNA in the nucleus. Therefore, snRNA-seq can also be used for the abovementioned analyses (104). Finally, I also detected 537 circRNAs of which 14 were detected in at

least 4 out of 64 cells. To detect more circRNAs in multiple single cells, total RNA libraries should be sequenced deeper or enriched for circRNAs by selective removal of linear RNA by exonuclease treatment prior to library prep and sequencing (81, 82).

Since single cell experiments are confounded by several types of biases, including cell cycle, chip and sequencing bias, I eliminated as many as possible of these biases. Cell cycle is the major confounder of single cell experiments and can hide true biological effects. Therefore, I reduced cell cycle bias by performing cell cycle synchronization by serum starvation. Bulk cell cycle analysis after serum starvation demonstrated that 80.3 % of the cells were arrested in the G0/G1 phase. As this analysis was done in bulk, I was not able to define the cell cycle stage of each cell separately. In order to do so, fluorescence ubiquitination-based cell cycle indicator (FUCCI) can be used whereby cells are stained according to their cell cycle stage. This is based on the fact that CTD1 is highly expressed in the G1 phase and subsequently ubiquitinated, while geminin is highly expressed during the S, M and G2 phase. By transducing cells with fluorescently labeled CTD1 and geminin, each cell's individual cell cycle phase can be determined (105). After sequencing, the cell cycle phase per cell can also be determined bioinformatically based on cell cycle stage specific gene expression patterns. To further eliminate the effect of cell cycle, bioinformatics tools, such as the single-cell latent variable model (scLVM), can be applied during data analysis (106). In addition, chip bias can be introduced by analyzing single cells of treated and untreated cells on different chips. Therefore, I stained the treated cells to be able to process treated and untreated cells on one chip and to visually distinguish them after cell capture. Finally, sequencing bias can be introduced by sequencing samples on different runs. As I wanted to compare my single cell total RNA-seq protocol with the standard single cell polyA[+] protocol, I sequenced both libraries in one sequencing run.

Despite many advantages, my method also has some limitations. First, the throughput is low as only up to 96 cells can be captured on a C1 microfluidic chip. Moreover, the cell capture rate of the C1 depends on the cell type as some cell types (e.g. leukemic cells) are very motile and can move through the channels resulting in multiple cells per capture site. To partially solve the throughput problem, I showed that the single cell total RNA-seq C1 protocol also works for FACS sorted cells, enabling to sort cells in 384 well plates. Although this increases the throughput, it is still low compared to the droplet-based systems where tens of thousands of cells can be captured, but which are less flexible to adapt. Second, the C1 suffers from a high cell doublet rate, further reducing the number of truly single cells per chip (107). For the three experiments that were performed, I observed at least one cell in 609 of the 672 capture sites (7 chips) with a mean multiplet rate of 34.54 %. However, the C1 chip has the important advantage that cells can be visualized using a microscope to exclude these multiplets from further analysis, which is not possible for droplet-based systems. However, stacked cells (i.e. cells on top of each other) cannot be identified as two cells since they seem to be one cell (108). To get an accurate estimate of the doublet rate of specific devices, mixed mouse-human experiments are typically carried out by counting the number of cells that contain a considerable fraction of human as well as mouse transcripts or by using two cell types that can be genetically distinguished (107). Also bioinformatics tools have been developed to remove these doublets *in silico* (109, 110). Finally, three chips for the C1 exist, depending on the size of the cells: small (5-10 µM), medium (10-17 µm), and large (17-25 µm) chips. Therefore, heterogeneous cell populations consisting of cells with different cell sizes cannot be fully captured using the C1 (111). Nevertheless, this problem can be solved using FACS sorted cells as I demonstrated that my novel total RNA single cell method works equally well on FACS sorted cells. Furthermore, this has the advantage that cells of interest can be sorted based on known surface markers and that cell doublets and debris are removed. In addition, rare cells can be enriched using FACS sorting, for which it has been shown to have only minor effect on gene expression profiles (112).

In contrast to the typically used single cell polyA[+] methods, the developed single cell total RNA-seq method in my PhD thesis enables to quantify non-polyadenylated genes, although small RNAs, including microRNAs and tRNAs will remain undetected. To detect these small RNAs, other methods for single cell small RNA-seq have recently been developed (113). As single cell small, polyA[+] and total RNA-seq methods each detect a unique set of the transcriptome, these layers should ideally be combined to get a more complete view of a cell's transcriptome (114). Combining small and mRNA-seq is possible by first isolating the polyadenylated transcripts with an oligo dT primer followed by small RNA-seq on the supernatants fraction, in which the small RNAs are still present. This combination can be used to validate microRNA targets as changes of microRNA expression can have influences on the expression of hundreds of mRNAs (115).

4.6. **The hurdles to get a complete view of a single cell's transcriptome are device dependent**

Since the number of single cell sequencing devices and library prep methods has increased substantially over the last years, I compared three -C1, ddSeq and Chromium- commercial available single cell devices, representative for the wide variety of platforms, ranging from microfluidic chips to droplet-based systems and from full transcript sequencing to 3' end sequencing. The C1 uses microfluidic channels and pressure-controlled valves, in which 96 cells can be isolated and processed in nanoliter volume reaction chambers. In contrast, ddSeq and Chromium are droplet-based devices that capture ~300 and tens of thousands of cells per well, respectively. To compare these three devices, I treated NGP neuroblastoma cells with nutlin-3 and performed single cell polyA[+] RNA-seq on these samples with each of the three devices. Cell synchronization was carried out to reduce the cell cycle bias and ERCC spike-in RNA molecules were added in all three experiments. These ERCC spikes are 92 synthetic RNA molecules that have a bacterial sequence composition in order not to interfere with the human sequences. These spikes differ in length, GC content and abundance and are used in bulk experiments to measure accuracy and sensitivity (116). In 2012, these ERCC spikes were for the first time used in a single cell experiment and can be used for normalization (117). Of note, ERCC spikes are mostly not added in droplet-based experiments since these are added to every droplet and consequently also sequenced in droplets without cells, increasing the sequencing costs considerably (118). However, we have added ERCC spikes to all three experiments to make them comparable. In order to remove batch effects, nutlin-3 treated NGP cells were stained with a cell tracker dye in the C1 experiment, enabling to process treated and untreated cells on one chip. In contrast, for the ddSeq and Chromium, treated and untreated cells were isolated in different wells of the chip, resulting in possible small batch effects. In future experiments, these batch effects could be reduced by 'cell hashing' where oligo coupled antibodies against ubiquitously expressed surface markers are used, enabling to distinguish cell types that are processed together and to identify multiplets as each cell type has another oligo (119). Furthermore, cell hashing can be used to distinguish low-quality cells from ambient (extracellular) RNA as the antibodies will only bind in droplets containing real single cells. Of note, cell hashing requires a deeper sequencing coverage to be able to sequence and detect the oligos bound to the antibodies (119, 120).

As a first step in the comparison of these devices, low-quality cell data needed to be removed. Therefore, I removed all cells that have a log number of genes or reads more than three times the median absolute deviation (MAD) below the median value as the mRNA of these cells has not efficiently been captured. Also, cells with a number of reads or genes above this cutoff were removed as these may be cell doublets. In addition, cells with a number of ERCC spikes above or below three times the MAD are also removed as this indicates a too low or high endogenous mRNA content, respectively. In such low-quality cells, the endogenous mRNA is less efficiently captured, thus ERCC spikes will be preferentially reverse transcribed, amplified and sequenced. The opposite is true for cell

doublets, where relatively speaking too much endogenous mRNA is captured (121–123). However, it should be noted that these filters may remove specific subpopulations in heterogeneous cell populations as some cells may effectively contain less or more mRNA (121–123). To further remove low-quality data, cells with a high mitochondrial RNA content are typically removed, as dying cells with impaired cellular membranes are known to result in more mitochondrial reads (107, 124). Using this filter should be done carefully as some cell types such as heart cells have a high mitochondrial content and should not be removed (124). Moreover, particular treatments, such as the TP53 activator nutlin-3, result in apoptosis, which can lead to a high fraction of mitochondrial reads.

All single cell devices, including also the C1, ddSeq and Chromium that I evaluated, have a certain number of challenges in common, some more than others. The first challenge is that all the developed technologies suffer from technical noise due to the low input volumes. This low input requires several PCR cycles resulting in amplification bias, which can only partially be removed using UMIs, since these UMIs are random sequences that tag each unique mRNA molecule. My data revealed a noisier gene expression pattern for the C1 compared to ddSeq and Chromium. This can partially be explained by the lower throughput and by the fact that C1 library prep methods do not integrate UMIs, resulting in a higher amplification bias compared to ddSeq and Chromium that include UMIs (125–128). A second challenge is that only 10 – 40 % of transcripts are typically captured per single cell (107, 114, 129, 130). This subsampling of transcripts has a major effect on the interpretation of single cell transcriptome profiles as single cell RNA-seq does not comprise a full picture of a cell's transcriptome. The most abundant genes will be detected in almost all cells, while medium expressed genes will only be detected in some cells implying that these are rare, while true low abundant transcripts will probably not be detected. Therefore, care should be taken to draw conclusions and results should be validated using single molecule FISH (smFISH). smFISH uses a set of fluorescent labeled probes to bind on target RNA, enabling quantification and localization of single RNA molecules in individual cells (131, 132). Nevertheless, based on the gene expression profiles of multiple cells, we can obtain insights in cellular heterogeneity and subpopulations (125). In line with the literature, I showed that using the C1, more transcripts per cell can be captured compared to the ddSeq and Chromium and that C1 gene expression data better correlates with bulk gene expression data, due to the higher sequencing depth and higher transcriptome complexity of the C1 cDNA libraries. Of note, it has been shown that even after subsampling, the sensitivity of SMART-seq protocol that I used on the C1 is still higher compared to for instance Chromium, showing that the SMART-seq method used on the C1 has a higher mRNA capture efficiency (107). By subsampling my single cell datasets, I will validate if the C1 still has the highest sensitivity. Furthermore, I revealed that correlation of average gene expression patterns among the single cell devices was slightly better between ddSeq and Chromium, in line with their similarity in terms of RNA-seq library preparation. The third challenge is that many dropouts are generated due to technical limitations associated with the low input volumes whereby some transcripts are not captured (107, 133, 134). These zero-values for a gene can be both biological or technical in origin. Biological dropouts are zero-values for genes that are simply not expressed. These genes can be expressed in some cells, but not in other cells due to heterogeneity or transcriptional bursts (108, 127, 134). In contrast, technical dropouts are zero-values for genes that are expressed, but not detected due to the low capture efficiency of low abundant genes (108, 127, 134). In line with the higher sensitivity of the C1, less dropouts (55 %) were detected in the C1 dataset compared to the ddSeq (88 %) and Chromium (86 %) dataset, due to the higher RT efficiency and higher number of reads that are typically generated for this low-throughput method. Several tools have been developed to take into account these dropout events (130, 135, 136).

As discussed above, each single cell RNA-seq method has its own strengths and weaknesses, which should be taken into account to select the optimal method for a specific application. The device one

should use depends on the available equipment, the number of required single cells, the sequencing coverage, the sample availability and the research question. To investigate embryo development, only a hundred of cells are sufficient to identify the critical steps during development, while more cells are needed to study transcriptional heterogeneity (120, 137). To unravel heterogeneity, one can first do a shallow sequencing experiment on the Chromium with many cells. If the results are of interest, a more in-depth analysis can be done on fewer cells with a deeper and more complete transcriptome coverage (120). Tools exist to calculate the required number of cells based on the number of subpopulations one expects, the fraction of cells that belong to the rarest subpopulation and the number of cells wanted per subpopulation (120). Since the throughput of single cell devices differs, the number of required cells will have a direct influence on the choice of the device. A low coverage is sufficient to capture heterogeneity since the expression of high abundant genes is sufficient to discriminate cell types, while deeper sequencing is required to study transcriptional heterogeneity of low abundant genes (125, 138). Moreover, deeper sequencing is required for full length methods to perform splicing and mutation analysis (139, 140). Dr. Hadfield and colleagues have generated a web portal to collect single cell quality data of the scientific community with information about the cell type, sequencing depth and number of genes detected, providing some evidence for the number of reads that are warranted for your specific question and cell type (http://10xqc.com)(120). The device of choice also depends on the material used. For precious clinical samples with few cells, devices with a high cell capture efficiency are warranted in order not to lose too much material (125). Furthermore, the price can play a role in choosing a device. The C1 is expensive due to the expensive chips and commercial kits. However, the price of this type of experiments can be reduced using nanoliter dispensing robots decreasing the volumes needed and consequently the price (139, 140).

4.7. **Single cell RNA sequencing reveals transcriptional heterogeneity and hidden biological signals**

Besides data quality, I also evaluated the three single cell devices with respect to transcriptional heterogeneity and their ability to detect differentially expressed genes, which is unique since other comparative studies compared single cell methods with a focus on data quality, costs, reproducibility, and the ability to discriminate subpopulations (86, 104, 117, 118, 128). One recent study also evaluated the ability of the C1 and Chromium to detect differentially expressed genes, however did not use a model system upon chemical perturbation (107). Furthermore, this study only used Limma Voom to detect differentially expressed genes, while I combined PIM and EdgeR with Zinger to identify truly differentially expressed genes, since it has been shown that the number of genes called as differentially expressed varies among tools (141). The largest number of genes was identified as differentially expressed using Chromium, in line with the other study and the observation that more genes are called differentially expressed with increasing numbers of single cells (107, 142). Although the number of differentially expressed genes in the single cell datasets is much lower compared to the bulk dataset, a strong enrichment of TP53 gene sets was identified in all datasets, showing that capturing only the most abundant genes is sufficient to recapitulate the biological signal. Of note, only a small overlap in enriched gene sets between the bulk and single cell datasets was identified, in line with a previous study (107). This discrepancy might be partially explained by the fact that in contrast to the single cell experiments, no cell cycle synchronization was performed for the bulk experiment. Further in-depth investigation should reveal the relevance of the unique enriched signals in the single cell datasets. Removing the cell cycle effect by only investigating cells in the G1 cell cycle phase revealed that more genes are differentially expressed compared to the full dataset, showing that these might be hidden by cell cycle effects in the full dataset. Furthermore, enrichment of TP53 regulated genes underlines the relevance of the genes only differentially expressed in G1 cells. To validate these results, the analysis should be repeated for cells in other cell cycle phases. In addition, this analysis has only been

carried out for the Chromium dataset as this is the largest dataset and repeating these analyses for the C1 and ddSeq dataset might not be robust as only a few cells per cell cycle phase will be retained. However, this issue can be partially solved by pre-selecting cells in a specific cell cycle stage by FACS sorting prior to single cell isolation. Furthermore, for the C1 the high-throughput chip, in which 800 single cells can be captured, can be used in future experiments, increasing throughput considerably. Of interest, comparing nutlin-3 treated cells with low and high expression of the TP53 target *CDKN1A* and nutlin-3 and vehicle treated cells without expression of *CDKN1A* enabled to detect potential late-responders and resistant cells, an intriguing finding that requires further in-depth investigation. This analysis should also be validated using other TP53 target genes. Selecting these target genes is complicated by the fact that only the most abundant genes are detected in single cell experiments, excluding for instance *BAX* as candidate. Finally, I showed that pseudobulk data, generated by merging single cell profiles, reconstitute the bulk data.

4.8. **Limitations of the current methods drive the development of new single cell sequencing methods**

Over the past decade, multiple single cell RNA-seq methods have been developed with increasing numbers of cells that can be processed, from a few cells to tens of thousands of cells in a single experiment (129, 143, 144). Although this also resulted in a drop of the costs of a single cell sequencing experiment, sequencing thousands of single cells remains costly (119). Furthermore, only 10 – 40 % of transcripts are typically captured per single cell, missing information to obtain a complete view of a cellular transcriptome (107, 114, 129, 130). Therefore, further evolution in the single cell RNA-seq methods is warranted to capture more transcripts and to further reduce the costs. As most high-throughput single cell RNA-seq methods currently require paired-end sequencing, costs could be reduced by developing methods that only need single-end sequencing (145). To increase the transcript capture rate, cell lysis and RT steps need to be further optimized (87, 146).

A major drawback of current single cell RNA-seq methods is the lack of spatial information that is embedded in the tissue of origin (89). One of the first efforts to retain spatial information was the use of smFISH, which combines spatial organization with copy number variations by using fluorescent probes. This method does not need pre-amplification, but is limited by the number of colors that can be visualized and can consequently only be used for the quantification of a handful of targets (132, 147, 148). Sequential FISH (seqFISH) and multiplexed error robust FISH (MERFISH) circumvent this limitation by including sequential rounds of labeling and imaging, enabling an unlimited number of transcripts to be visualized. However, this is only possible for known markers (147–149). To obtain the whole transcriptome and retain the spatial information in a single cell, fluorescent in situ sequencing (FISSEQ) has been developed. In FISSEQ, RNA is reverse transcribed and amplified in fixed cells, followed by cDNA sequencing by the 'sequencing by ligation' approach, providing detailed spatial information of the transcripts (150–152). Further improvements will be needed to get better coverage and longer reads and to increase the sensitivity (151, 153). A limitation of the current single cell whole transcript methods, which can be used for mutation and splicing analysis, is that these are limited to a few hundred cells. Therefore, single-cell isoform RNA-Seq (ScISOr-Seq) has been developed enabling to sequence full length transcripts of thousands of cells using long-read sequencing. Furthermore, 3' end RNA-seq of the same cells can be carried out in parallel. First, single cells are captured and labeled using Chromium followed by dividing the pool of cells in two populations, one part for the 3' end counting for gene expression profiling and one part for the long read sequencing, used for isoform identification using Pacific Biosciences (PacBio) or Oxford Nanopore sequencing instrument (154).

Numerous single cell sequencing methods have been developed and optimized over the last decade, enabling to sequence tens of thousands of single cells. However, all currently existing single cell

sequencing methods are end-point experiments, eliminating the possibility to further follow-up these cells over time. Furthermore, due to the increase of single cell analysis research articles, numerous single cell sequencing datasets are publicly available. To make the data also easily available and interpretable for biologists, often lacking skills to analyze these complex datasets, several tools with a user-friendly interface have been developed. The single-cell analysis pipeline (ASAP) allows to load single cell sequencing data and to perform subsequent analysis steps such as filtering and normalization. Afterwards, results can be visualized making them easily interpretable (155). Panglao DB is another tool that contains hundreds of mouse and human single cell data analyzed with the same pipeline allowing analysis, visualization and interpretation of the data (156).

4.9. **Unraveling tumor heterogeneity by single cell RNA sequencing can have major clinical implications**

Over the last decades, cancer research has been focusing on the detection of genetic alterations in cancer specific genes and their therapeutic targeting. While this enabled the development and application of precision oncology treatments for some tumor types, a large fraction of the patients acquires therapy resistance by circumventing the mechanisms of action (157, 158). This resistance can be due to a specific subpopulation that has primary resistance or develops resistance to the therapy (158). Despite advances in NGS and computational methods on bulk tumor tissue, with the ability to detect subpopulations, single cell RNA-seq methods are required to get a full view of the complexity of the tumor, i.e. its clonal subpopulations, stromal cells and infiltrating immune cells (158–160).

Single cell sequencing offers great benefits for cancer research. First, identifying the cells of origin can significantly contribute to early tumor detection (158, 161). Second, identification of a pre-malignant disease state and consequently providing early treatment can improve patient survival (158, 162). Identifying such a pre-malignant stage is difficult as the CNVs and SNVs are only present in a few cells at the initial stage and missed performing bulk RNA-seq underscoring the need to perform single cell sequencing (158). Identifying this pre-malignant state contributes to the prediction of tumor progression and decision of the treatment (158, 163). Third, despite the fact that the primary tumor is mostly investigated, more research is needed to characterize intra-tumor heterogeneity as characterization of the different subpopulations is important to give the right combination of drugs that can target all these populations. Finally, comparing primary and metastatic cells enables phylogenetic analysis to determine how the tumor evolved in a metastatic tumor. Single cell RNA-seq can unravel the transcriptional changes a cell undergoes to emerge from a cancer cell to a migrating cancer cell and can contribute to the identification of new therapeutic targets (158).

Nowadays, most single cell sequencing methods still start from fresh material, limiting the applications as tumors are often flash-frozen or preserved formalin-fixed paraffin-embedded (FFPE). This often results in membrane rupture, although nuclear membranes are shown to retain intact (164, 165). Therefore, new methods that can use FFPE material or cells that have been frozen as input have been developed (166). These methods isolate nuclei instead of whole cells as nuclei are more resistant to the stress during freeze-thawing and FFPE preserving (165, 167, 168). For FFPE material, intact nuclei can be isolated and DNA that has been disrupted can be repaired by adding DNA repair enzymes which can subsequently be used for CNV analyses, resulting in similar results as for fresh material (166). Also methods to perform snRNA-seq of frozen cells show a high concordance between nuclei and whole cell derived transcriptomes and detect more intronic reads as compared to whole cell derived methods (104, 165, 167, 168). A drawback of snRNA-seq is that sorting based on surface markers is not possible (168). In order to be able to store cells without influencing their gene expression profile, it has recently been shown that cells can be fixed using aldehyde or alcohol and enables to store cells for several weeks. The sensitivity, doublet rate and gene expression profile obtained are similar as to fresh cells

enabling to store cells and process them at later timepoints (158, 169, 170). This is preferable for clinical samples that often need to be shipped because sample collection and sample processing often occur at different locations (145). Since invasive biopsies are needed to isolate cells from tumors, single CTC sequencing has gained interest over the last years. CTCs are cells from the primary tumor, shed in the blood stream which indicates the presence of metastasis. These CTCs can be isolated in a non-invasive way and can be used for follow-up of tumor progression and treatment. As only a few CTCs per 10 ml blood are present, methods to enrich for these CTCs are warranted. These can be isolated from blood using the epithelial cell adhesion molecule (EpCam) marker, which is a tumor-specific surface maker expressed on epithelial tumor cells while absent on most blood cells. In contrast, CD45 is only expressed on most blood cells and can be used for negative selection. Isolation of CTCs based on these markers has the disadvantage that EpCam negative CTCs are missed (164, 171). In contrast, the DepArray system uses fluorescent markers where CTCs can be isolated based on specific markers of the tumor type, whereby also EpCam negative CTCs can be isolated (172). These methods are biased as they are based on known markers or EpCam expression. To capture CTCs in an unbiased way, nanofabricated filters were developed as CTCs are in general larger than normal blood cells (164, 173). New enrichment methods are still warranted, to more efficiently enrich for these rare, valuable CTCs.

4.10. **Conclusions**

In conclusion, I identified a set of TLX1 regulated and TLX subgroup specific lncRNAs, of which some are potentially oncogenic, marking them as highly interesting targets for further in-depth characterization. Since the T-ALL dataset I generated contains unexplored features, I wrote a data descriptor with detailed information about the methods and bioinformatics pipelines applied to make the data re-usable for the research community, enabling to further unravel the complex biology of T-ALL. Furthermore, I developed a single cell total RNA-seq protocol that for the first time combines strandedness and effective removal of ribosomal cDNA and enables the detection of both polyadenylated and non-polyadenylated transcripts, including lncRNAs, circRNAs and novel genes. Finally, I evaluated three commercial single cell devices (C1, ddSeq and Chromium) with respect to data quality and the ability to detect differentially expressed genes and revealed that single cell data can detect biological signal faithfully through gene set enrichment analysis and may help to identify potentially late-responders or resistant cells upon compound treatment.

4.11. **References**

1. Ferrando,A.A., Neuberg,D.S., Staunton,J., Loh,M.L., Huard,C., Raimondi,S.C., Behm,F.G., Pui,C.-H., Downing,J.R., Gilliland,D.G., *et al.* (2002) Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell*, **1**, 75–87.
2. Soulier,J., Clappier,E., Cayuela,J.-M., Regnault,A., García-Peydró,M., Dombret,H., Baruchel,A., Toribio,M.-L. and Sigaux,F. (2005) HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood*, **106**, 274–86.
3. T.B.,H., R.B.,M. and G.H.,R. (2009) Late effects in long-term survivors after treatment for childhood acute leukemia. *Clin. Pediatr. (Phila).*, **48**, 601–608.
4. Pui,C.-H., Robison,L.L. and Look,A.T. (2008) Acute lymphoblastic leukaemia. *Lancet*, **371**, 1030–1043.
5. Girardi,T., Vicente,C., Cools,J. and De Keersmaecker,K. (2017) The genetics and molecular biology of T-ALL. *Blood*, **129**, 1113–1123.
6. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J., *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–5.
7. Nik-Zainal,S., Davies,H., Staaf,J., Ramakrishna,M., Glodzik,D., Zou,X., Martincorena,I., Alexandrov,L.B., Martin,S., Wedge,D.C., *et al.* (2016) Landscape of somatic mutations in 560

breast cancer whole-genome sequences. *Nature*, **534**, 47–54.
8. Wang,Y., Wu,P., Lin,R., Rong,L., Xue,Y. and Fang,Y. (2015) LncRNA NALT interaction with NOTCH1 promoted cell proliferation in pediatric T cell acute lymphoblastic leukemia. *Sci. Rep.*, **5**, 1–10.
9. Trimarchi,T., Bilal,E., Ntziachristos,P., Fabbri,G., Dalla-Favera,R., Tsirigos,A. and Aifantis,I. (2014) Genome-wide Mapping and Characterization of Notch-Regulated Long Noncoding RNAs in Acute Leukemia. *Cell*, **158**, 593–606.
10. Durinck,K., Wallaert,A., Van de Walle,I., Van Loocke,W., Volders,P.J., Vanhauwaert,S., Geerdens,E., Benoit,Y., Van Roy,N., Poppe,B., *et al.* (2014) The notch driven long non-coding RNA repertoire in T-cell acute lymphoblastic leukemia. *Haematologica*, **99**, 1808–1816.
11. Wallaert,A., Durinck,K., Van Loocke,W., Van de Walle,I., Matthijssens,F., Volders,P.J., Avila Cobos,F., Rombaut,D., Rondou,P., Mestdagh,P., *et al.* (2016) Long noncoding RNA signatures define oncogenic subtypes in T-cell acute lymphoblastic leukemia. *Leukemia*, **30**, 1927–1930.
12. Yan,X., Hu,Z., Feng,Y., Hu,X., Yuan,J., Zhao,S.D., Zhang,Y., Yang,L., Shan,W., He,Q., *et al.* (2015) Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell*, **28**, 529–540.
13. Cao Thi Ngoc,P., Hao Tan,S., King Tan,T., Min Chan,M., Li,Z., J Yeoh,A.E., Tenen,D.G. and Sanda,T. Identification of novel lncRNAs regulated by the TAL1 complex in T- cell acute lymphoblastic leukemia. *Leukemia*, 10.1038/s41375-018-0110-4.
14. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M., *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.
15. Zilionis,R., Nainys,J., Veres,A., Savova,V., Zemmour,D., Klein,A.M. and Mazutis,L. (2017) Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.*, **12**, 44–73.
16. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J., *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
17. Ferrando,A.A., Neuberg,D.S., Dodge,R.K., Paietta,E., Larson,R.A., Wiernik,P.H., Rowe,J.M., Caligiuri,M.A., Bloomfield,C.D. and Look,A.T. (2004) Prognostic importance of TLX1 (HOX11) oncogene expression in adults with T-cell acute lymphoblastic leukaemia. *Lancet*, **363**, 535–536.
18. Kees,U.R., Heerema,N.A., Kumar,R., Watt,P.M., Baker,D.L., La,M.K., Uckun,F.M. and Sather,H.N. (2003) Expression of HOX11 in childhood T-lineage acute lymphoblastic leukaemia can occur in the absence of cytogenetic aberration at 10q24: a study from the Children's Cancer Group (CCG). *Leukemia*, **17**, 887–893.
19. Berger,R., Dastugue,N., Busson,M., van den Akker,J., Pérot,C., Ballerini,P., Hagemeijer,A., Michaux,L., Charrin,C., Pages,M.P., *et al.* (2003) t(5;14)/HOX11L2-positive T-cell acute lymphoblastic leukemia. A collaborative study of the Groupe Français de Cytogénétique Hématologique (GFCH). *Leukemia*, **17**, 1851–1857.
20. De Keersmaecker,K., Real,P.J., Gatta,G. Della, Palomero,T., Sulis,M.L., Tosello,V., Van Vlierberghe,P., Barnes,K., Castillo,M., Sole,X., *et al.* (2010) The TLX1 oncogene drives aneuploidy in T cell transformation. *Nat. Med.*, **16**, 1321–1327.
21. Rakowski,L.A., Lehotzky,E.A. and Chiang,M.Y. (2011) Transient responses to NOTCH and TLX1/HOX11 inhibition in T-cell acute lymphoblastic leukemia/lymphoma. *PLoS One*, **6**.
22. Kleppe,M., Soulier,J., Asnafi,V., Mentens,N., Hornakova,T., Knoops,L., Sigaux,F., Meijerink,J.P., Vandenberghe,P., Tartaglia,M., *et al.* (2013) lymphoblastic leukemia PTPN2 negatively regulates oncogenic JAK1 in T-cell acute lymphoblastic leukemia. **117**, 7090–7098.
23. Van Vlierberghe,P., Palomero,T., Khiabanian,H., Van Der Meulen,J., Castillo,M., Van Roy,N., De Moerloose,B., Philippé,J., González-García,S., Toribio,M.L., *et al.* (2010) PHF6 mutations in T-cell acute lymphoblastic leukemia. *Nat. Genet.*, **42**, 338–342.
24. De Keersmaecker,K. and Ferrando,A.A. (2011) TLX1-Induced T-cell Acute Lymphoblastic Leukemia. *Clin. Cancer Res.*, **17**, 6381–6386.
25. Graux,C., Cools,J., Melotte,C., Quentmeier,H., Ferrando,A., Levine,R., Vermeesch,J.R., Stul,M., Dutta,B., Boeckx,N., *et al.* (2004) Fusion of NUP214 to ABL1 on amplified episomes in T-cell acute

lymphoblastic leukemia. *Nat. Genet.*, **36**, 1084–1089.
26. Tosello,V., Mansour,M.R., Barnes,K., Paganin,M., Sulis,M.L., Jenkinson,S., Allen,C.G., Gale,R.E., Linch,D.C., Palomero,T., *et al.* (2009) WT1 mutations in T-ALL. *Blood*, **114**, 1038–1045.
27. Durinck,K., Van Loocke,W., Van der Meulen,J., Van de Walle,I., Ongenaert,M., Rondou,P., Wallaert,A., de Bock,C.E., Van Roy,N., Poppe,B., *et al.* (2015) Characterization of the genome-wide TLX1 binding profile in T-cell acute lymphoblastic leukemia. *Leukemia*, **29**, 2317–2327.
28. Xiong,R., Joris,F., Liang,S., De Rycke,R., Lippens,S., Demeester,J., Skirtach,A., Raemdonck,K., Himmelreich,U., De Smedt,S.C., *et al.* (2016) Cytosolic Delivery of Nanolabels Prevents Their Asymmetric Inheritance and Enables Extended Quantitative in Vivo Cell Imaging. *Nano Lett.*, **16**, 5975–5986.
29. Varnaitė,R. and MacNeill,S.A. (2016) Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID. *Proteomics*, **16**, 2503–2518.
30. Roux,K.J., Kim,D.I., Burke,B. and May,D.G. (2018) BioID: A Screen for Protein-Protein Interactions. *Curr. Protoc. protein Sci.*, **91**, 19.23.1-19.23.15.
31. Joung,J., Engreitz,J.M., Konermann,S., Abudayyeh,O.O., Verdine,V.K., Aguet,F., Gootenberg,J.S., Sanjana,N.E., Wright,J.B., Fulco,C.P., *et al.* (2017) Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature*, 10.1038/nature23451.
32. Roe,J.-S., Mercan,F., Rivera,K., Pappin,D.J. and Vakoc,C.R. (2015) BET Bromodomain Inhibition Suppresses the Function of Hematopoietic Transcription Factors in Acute Myeloid Leukemia. *Mol. Cell*, **58**, 1028–39.
33. Baud,M.G.J., Lin-Shiao,E., Cardote,T., Tallant,C., Pschibul,A., Chan,K.-H., Zengerle,M., Garcia,J.R., Kwan,T.T.-L., Ferguson,F.M., *et al.* (2014) Chemical biology. A bump-and-hole approach to engineer controlled selectivity of BET bromodomain chemical probes. *Science*, **346**, 638–641.
34. Lee,J.-E., Park,Y.-K., Park,S., Jang,Y., Waring,N., Dey,A., Ozato,K., Lai,B., Peng,W. and Ge,K. (2017) Brd4 binds to active enhancers to control cell identity gene induction in adipogenesis and myogenesis. *Nat. Commun.*, **8**, 2217.
35. Kanno,T., Kanno,Y., LeRoy,G., Campos,E., Sun,H.-W., Brooks,S.R., Vahedi,G., Heightman,T.D., Garcia,B.A., Reinberg,D., *et al.* (2014) BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. *Nat. Struct. Mol. Biol.*, **21**, 1047–57.
36. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W., *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–12.
37. Kim,T.-K., Hemberg,M., Gray,J.M., Costa,A.M., Bear,D.M., Wu,J., Harmin,D.A., Laptewicz,M., Barbara-Haley,K., Kuersten,S., *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–7.
38. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T., *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–61.
39. Gardini,A. (2017) Global Run-On Sequencing (GRO-Seq). *Methods Mol. Biol.*, **1468**, 111–20.
40. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science (80-.).*, **322**, 1845–1848.
41. Kwak,H., Fuda,N.J., Core,L.J. and Lis,J.T. (2013) Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science (80-.).*, **339**, 950–953.
42. Magnuson,B., Veloso,A., Kirkconnell,K.S., Lima,L.C. de A., Paulsen,M.T., Ljungman,E.A., Bedi,K., Prasad,J., Wilson,T.E. and Ljungman,M. (2016) Identifying transcription start sites and active enhancer elements using BruUV-seq. *Sci. Rep.*, **5**, 17978.
43. Li,W., Notani,D. and Rosenfeld,M.G. (2016) Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.*, **17**, 207–223.
44. Della Gatta,G., Palomero,T., Perez-Garcia,A., Ambesi-Impiombato,A., Bansal,M., Carpenter,Z.W., De Keersmaecker,K., Sole,X., Xu,L., Paietta,E., *et al.* (2012) Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. *Nat. Med.*, **18**, 436–40.
45. Van Walle,I. De, De Smet,G., De Smedt,M., Vandekerckhove,B., Leclercq,G., Plum,J. and Taghon,T.

(2009) An early decrease in Notch activation is required for human TCR-αβ lineage differentiation at the expense of TCR-γδ T cells. *Blood*, **113**, 2988–2998.
46. Ebisuya,M., Yamamoto,T., Nakajima,M. and Nishida,E. (2008) Ripples from neighbouring transcription. *Nat. Cell Biol.*, **10**, 1106–1113.
47. Avila Cobos,F., Anckaert,J., Volders,P.-J., Everaert,C., Rombaut,D., Vandesompele,J., De Preter,K. and Mestdagh,P. (2017) Zipper plot: visualizing transcriptional activity of genomic regions. *BMC Bioinformatics*, **18**, 231.
48. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,T., *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 15776–81.
49. Guttman,M., Russell,P., Ingolia,N.T., Weissman,J.S. and Lander,E.S. (2013) Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell*, **154**, 240–251.
50. Basrai,M.A., Hieter,P. and Boeke,J.D. (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res.*, **7**, 768–71.
51. Andrews,S.J. and Rothnagel,J.A. (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.*, **15**, 193–204.
52. Ruiz-Orera,J., Messeguer,X., Subirana,J.A. and Alba,M.M. (2014) Long non-coding RNAs as a source of new peptides. *Elife*, **3**.
53. Bánfai,B., Jia,H., Khatun,J., Wood,E., Risk,B., Gundling,W.E., Kundaje,A., Gunawardena,H.P., Yu,Y., Xie,L., *et al.* (2012) Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.*, **22**, 1646–57.
54. Groff,A.F., Sanchez-Gomez,D.B., Soruco,M.M.L., Gerhardinger,C., Barutcu,A.R., Li,E., Elcavage,L., Plana,O., Sanchez,L.V., Lee,J.C., *et al.* (2016) In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements. *Cell Rep.*, **16**, 2178–2186.
55. Thakore,P.I., D'Ippolito,A.M., Song,L., Safi,A., Shivakumar,N.K., Kabadi,A.M., Reddy,T.E., Crawford,G.E. and Gersbach,C.A. (2015) Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods*, **12**, 1143–9.
56. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell*, **152**, 1173–1183.
57. Yeo,N.C., Chavez,A., Lance-Byrne,A., Chan,Y., Menn,D., Milanova,D., Kuo,C.-C., Guo,X., Sharma,S., Tung,A., *et al.* (2018) An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nat. Methods*, **15**, 611–616.
58. Rankin,C.R., Treger,J., Faure-Kumar,E., Benhammou,J., Anisman-Posner,D., Bollinger,A.E., Pothoulakis,C. and Padua,D.M. (2019) Overexpressing Long Noncoding RNAs Using Gene-activating CRISPR. *J. Vis. Exp.*, 10.3791/59233.
59. Maeder,M.L., Linder,S.J., Cascio,V.M., Fu,Y., Ho,Q.H. and Joung,J.K. (2013) CRISPR RNA–guided activation of endogenous human genes. *Nat. Methods*, **10**, 977–979.
60. Liu,S.J., Horlbeck,M.A., Cho,S.W., Birk,H.S., Malatesta,M., He,D., Attenello,F.J., Villalta,J.E., Cho,M.Y., Chen,Y., *et al.* (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, **355**.
61. Dixit,A., Parnas,O., Li,B., Chen,J., Fulco,C.P., Jerby-Arnon,L., Marjanovic,N.D., Dionne,D., Burks,T., Raychndhury,R., *et al.* (2016) Perturb-seq: Dissecting molecular circuits with scalable single cell RNA profiling of pooled genetic screens HHS Public Access. **16711**.
62. Datlinger,P., Rendeiro,A.F., Schmidl,C., Krausgruber,T., Traxler,P., Klughammer,J., Schuster,L.C., Kuchler,A., Alpar,D. and Bock,C. (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods*, **14**, 297–301.
63. Xie,S., Duan,J., Li,B., Zhou,P. and Hon,G.C. (2017) Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol. Cell*, **66**, 285–299.e5.
64. Hannus,M., Beitzinger,M., Engelmann,J.C., Weickert,M.T., Spang,R., Hannus,S. and Meister,G. (2014) SiPools: Highly complex but accurately defined siRNA pools eliminate off-target effects.

Nucleic Acids Res., **42**, 8049–8061.
65. Robb,G.B., Brown,K.M., Khurana,J. and Rana,T.M. (2005) Specific and potent RNAi in the nucleus of human cells. *Nat. Struct. Mol. Biol.*, **12**, 133–137.
66. Gagnon,K.T., Li,L., Chu,Y., Janowski,B.A. and Corey,D.R. (2014) RNAi Factors Are Present and Active in Human Cell Nuclei. *Cell Rep.*, **6**, 211–221.
67. Wang,F., Flanagan,J., Su,N., Wang,L.C., Bui,S., Nielson,A., Wu,X., Vo,H.T., Ma,X.J. and Luo,Y. (2012) RNAscope: A novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Diagnostics*, **14**, 22–29.
68. Chiu,H.S., Somvanshi,S., Patel,E., Chen,T.W., Singh,V.P., Zorman,B., Patil,S.L., Pan,Y., Chatterjee,S.S., Caesar-Johnson,S.J., *et al.* (2018) Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. *Cell Rep.*, **23**, 297–312.e12.
69. Ci Chu1, Kun Qu1, Franklin Zhong2, Steven E. Artandi2, and H.Y.C. (2012) Genomic maps of lincRNA occupancy reveal principles of RNA- chromatin interactions. **44**, 667–678.
70. Engreitz,J.M., Pandya-Jones,A., McDonel,P., Shishkin,A., Sirokman,K., Surka,C., Kadri,S., Xing,J., Goren,A., Lander,E.S., *et al.* (2013) The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science (80-.).*, **341**, 1237973–1237973.
71. Simon,M.D., Wang,C.I., Kharchenko,P. V., West,J.A., Chapman,B.A., Alekseyenko,A.A., Borowsky,M.L., Kuroda,M.I. and Kingston,R.E. (2011) The genomic binding sites of a noncoding RNA. *Proc. Natl. Acad. Sci.*, **108**, 20497–20502.
72. Quinn,J.J., Ilik,I.A., Qu,K., Georgiev,P., Chu,C., Akhtar,A. and Chang,H.Y. (2014) Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nat. Biotechnol.*, **32**, 933–940.
73. Johnsson,P., Lipovich,L., Grandér,D. and Morris,K. V (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta*, **1840**, 1063–71.
74. Prabhakar,B., Zhong,X.-B. and Rasmussen,T.P. (2017) Exploiting Long Noncoding RNAs as Pharmacological Targets to Modulate Epigenetic Diseases. *Yale J. Biol. Med.*, **90**, 73–86.
75. Noviello,T.M.R., Di Liddo,A., Ventola,G.M., Spagnuolo,A., D'Aniello,S., Ceccarelli,M. and Cerulo,L. (2018) Detection of long non–coding RNA homology, a comparative study on alignment and alignment–free metrics. *BMC Bioinformatics*, **19**, 407.
76. Ulitsky,I., Shkumatava,A., Jan,C.H., Sive,H. and Bartel,D.P. (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, **147**, 1537–50.
77. Trimarchi,T., Bilal,E., Ntziachristos,P., Fabbri,G., Dalla-Favera,R., Tsirigos,A. and Aifantis,I. (2014) Genome-wide Mapping and Characterization of Notch-Regulated Long Noncoding RNAs in Acute Leukemia. *Cell*, **158**, 593–606.
78. Wallaert,A., Durinck,K., Taghon,T., Van Vlierberghe,P. and Speleman,F. (2017) T-ALL and thymocytes: a message of noncoding RNAs. *J. Hematol. Oncol.*, **10**, 66.
79. Gaidatzis,D., Burger,L., Florescu,M. and Stadler,M.B. (2015) Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.*, **33**, 722–729.
80. Qin,S., Zhao,Y., Lim,G., Lin,H., Zhang,X. and Zhang,X. (2019) Circular RNA PVT1 acts as a competing endogenous RNA for miR-497 in promoting non-small cell lung cancer progression. *Biomed. Pharmacother.*, **111**, 244–250.
81. Fan,X., Zhang,X., Wu,X., Guo,H., Hu,Y., Tang,F. and Huang,Y. (2015) Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.*, **16**, 148.
82. Jeck,W.R. and Sharpless,N.E. (2014) Detecting and characterizing circular RNAs. *Nat. Biotechnol.*, **32**, 453–61.
83. Ferrando,A. (2018) Can one target T-cell ALL? *Best Pract. Res. Clin. Haematol.*, **31**, 361–366.
84. Oshima,K., Khiabanian,H., da Silva-Almeida,A.C., Tzoneva,G., Abate,F., Ambesi-Impiombato,A., Sanchez-Martin,M., Carpenter,Z., Penson,A., Perez-Garcia,A., *et al.* (2016) Mutational landscape, clonal evolution patterns, and role of RAS mutations in relapsed acute lymphoblastic leukemia. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 11306–11311.

85. De Bie,J., Demeyer,S., Alberti-Servera,L., Geerdens,E., Segers,H., Broux,M., De Keersmaecker,K., Michaux,L., Vandenberghe,P., Voet,T., *et al.* (2018) Single-cell sequencing reveals the origin and the order of mutation acquisition in T-cell acute lymphoblastic leukemia. *Leukemia*, 10.1038/s41375-018-0127-8.
86. Wu,A.R., Neff,N.F., Kalisky,T., Dalerba,P., Treutlein,B., Rothenberg,M.E., Mburu,F.M., Mantalas,G.L., Sim,S., Clarke,M.F., *et al.* (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, **11**, 41–46.
87. Hashimshony,T., Senderovich,N., Avital,G., Klochendler,A., de Leeuw,Y., Anavy,L., Gennert,D., Li,S., Livak,K.J., Rozenblatt-Rosen,O., *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, **17**, 77.
88. Liu,Y., Chen,X., Zhang,Y. and Liu,J. (2019) Advancing single-cell proteomics and metabolomics with microfluidic technologies. *Analyst*, **144**, 846–858.
89. Kolodziejczyk,A.A. and Lönnberg,T. (2018) Global and targeted approaches to single-cell transcriptome characterization. *Brief. Funct. Genomics*, **17**, 209–219.
90. Baran-Gale,J., Chandra,T. and Kirschner,K. (2018) Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics*, **17**, 233–239.
91. Streets,A.M., Zhang,X., Cao,C., Pang,Y., Wu,X., Xiong,L., Yang,L., Fu,Y., Zhao,L., Tang,F., *et al.* (2014) Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 7048–53.
92. Picelli,S., Björklund,Å.K., Faridani,O.R., Sagasser,S., Winberg,G. and Sandberg,R. smart-seq2 for sensitive full-length transcriptome profiling in single cells. 10.1038/nMeth.2639.
93. Yang,L., Duff,M.O., Graveley,B.R., Carmichael,G.G. and Chen,L.-L. (2011) Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.*, **12**, R16.
94. Lai,F., Gardini,A., Zhang,A. and Shiekhattar,R. (2015) Integrator mediates the biogenesis of enhancer RNAs. *Nature*, **525**, 399–403.
95. Suzuki,Y. Single Molecule and Single Cell Sequencing.
96. Fan,X., Zhang,X., Wu,X., Guo,H., Hu,Y., Tang,F. and Huang,Y. (2011) Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. 10.1186/s13059-015-0706-1.
97. Hayashi,T., Ozaki,H., Sasagawa,Y., Umeda,M., Danno,H. and Nikaido,I. (2018) Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.*, **9**, 619.
98. Sheng,K., Cao,W., Niu,Y., Deng,Q. and Zong,C. (2017) Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods*, **14**.
99. Islam,S., Kjällquist,U., Moliner,A., Zajac,P., Fan,J.-B., Lönnerberg,P. and Linnarsson,S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, **21**, 1160–7.
100. Ameur,A., Zaghlool,A., Halvardson,J., Wetterbom,A., Gyllensten,U., Cavelier,L. and Feuk,L. (2011) Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.*, **18**, 1435–1440.
101. Zhao,S., Zhang,Y., Gamini,R., Zhang,B. and Von Schack,D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. 10.1038/s41598-018-23226-4.
102. Gaidatzis,D., Burger,L., Florescu,M. and Stadler,M.B. (2015) Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.*, **33**, 722–729.
103. La Manno,G., Soldatov,R., Zeisel,A., Braun,E., Hochgerner,H., Petukhov,V., Lidschreiber,K., Kastriti,M.E., Lönnerberg,P., Furlan,A., *et al.* (2018) RNA velocity of single cells. *Nature*, **560**, 494–498.
104. Mereu,E., Lafzi,A., Moutinho,C., Ziegenhain,C., Maccarthy,D.J., Alvarez,A., Batlle,E., Grün,D., Lau,J.K. and Boutet,S.C. (2019) Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects.

105. Sakaue-Sawano,A., Kurokawa,H., Morimura,T., Hanyu,A., Hama,H., Osawa,H., Kashiwagi,S., Fukami,K., Miyata,T., Miyoshi,H., *et al.* (2008) Visualizing Spatiotemporal Dynamics of Multicellular Cell-Cycle Progression. *Cell*, **132**, 487–498.
106. Buettner,F., Natarajan,K.N., Casale,F.P., Proserpio,V., Scialdone,A., Theis,F.J., Teichmann,S.A., Marioni,J.C. and Stegle,O. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.
107. Wang,Y.J., Schug,J., Lin,J., Wang,Z. and Kossenkov,A. (2019) Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues.
108. L. Lun,A.T., Bach,K. and Marioni,J.C. (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 75.
109. Wolock,S.L., Lopez,R. and Klein,A.M. (2019) Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.*, 10.1016/j.cels.2018.11.005.
110. McGinnis,C.S., Murrow,L.M. and Gartner,Z.J. (2019) DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.*, 10.1016/j.cels.2019.03.003.
111. Fan,H.C., Fu,G.K. and Fodor,S.P.A. (2015) Combinatorial labeling of single cells for gene expression cytometry. *Science (80-.).*, **347**.
112. Richardson,G.M., Lannigan,J. and Macara,I.G. (2015) Does FACS perturb gene expression? *Cytom. Part A*, **87**, 166–175.
113. Faridani,O.R., Abdullayev,I., Hagemann-Jensen,M., Schell,J.P., Lanner,F. and Sandberg,R. (2016) Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.*, **34**, 1264–1266.
114. Macaulay,I.C., Ponting,C.P. and Voet,T. (2017) Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends Genet.*, **33**, 155–168.
115. Xiao,Z., Cheng,G., Jiao,Y., Pan,C., Li,R., Jia,D., Zhu,J., Wu,C., Zheng,M. and Jia,J. (2018) Holo-Seq: Single-cell sequencing of holo-transcriptome. *Genome Biol.*, **19**, 1–22.
116. Jiang,L., Schlesinger,F., Davis,C.A., Zhang,Y., Li,R., Salit,M., Gingeras,T.R. and Oliver,B. Synthetic spike-in standards for RNA-seq experiments. 10.1101/gr.121095.111.Freely.
117. Svensson,V., Natarajan,K.N., Ly,L., Miragaia,R.J., Labalette,C., Macaulay,I.C., Cvejic,A. and Teichmann,S.A. (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat. Publ. Gr.*, **14**, 381–387.
118. Ziegenhain,C., Vieth,B., Parekh,S., Reinius,B., Guillaumet-Adkins,A., Smets,M., Leonhardt,H., Heyn,H., Hellmann,I. and Enard,W. (2017) Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell*, **65**, 631–643.e4.
119. Hao,S., Mauck,W.M., Satija,R., Houck-Loomis,B., Zheng,S., Stoeckius,M., Yeung,B.Z. and Smibert,P. (2018) Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.*, **19**.
120. Baran-Gale,J., Chandra,T. and Kirschner,K. (2018) Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics*, **17**, 233–239.
121. AlJanahi,A.A., Danielsen,M. and Dunbar,C.E. (2018) An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Mol. Ther. Methods Clin. Dev.*, **10**, 189–196.
122. Lun,A.T.L., McCarthy,D.J. and Marioni,J.C. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data. *F1000Research*, **5**, 2122.
123. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
124. Ilicic,T., Kim,J.K., Kolodziejczyk,A.A., Bagger,F.O., McCarthy,D.J., Marioni,J.C. and Teichmann,S.A. (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.*, **17**, 29.
125. Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.
126. Grün,D., Kester,L. and van Oudenaarden,A. (2014) Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**, 637–640.

127. Islam,S., Zeisel,A., Joost,S., La Manno,G., Zajac,P., Kasper,M., Lönnerberg,P. and Linnarsson,S. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, **11**, 163–166.
128. Zhang,X., Li,T., Liu,F., Chen,Y., Li,Z., Huang,Y. and Wang,J. (2018) Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. *bioRxiv*, 10.1101/313130.
129. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M., *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.
130. Ren,X., Kang,B. and Zhang,Z. (2018) Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biol.*, **19**.
131. Bonasio,R., Shaffer,S., Torre,E., Gupte,R., Murray,J., Kim,J., Raj,A., Gospocic,J. and Dueck,H. (2018) Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. *Cell Syst.*, **6**, 171–179.e5.
132. Raj,A., van den Bogaard,P., Rifkin,S.A., van Oudenaarden,A. and Tyagi,S. (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, **5**, 877–879.
133. Hicks,S.C., Townes,F.W., Teng,M. and Irizarry,R.A. (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, **19**, 562–578.
134. Kharchenko,P. V, Silberstein,L. and Scadden,D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.
135. Gong,W., Kwak,I.-Y., Pota,P., Koyano-Nakagawa,N. and Garry,D.J. (2018) DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, **19**, 220.
136. Li,W.V. and Li,J.J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, **9**, 997.
137. Yan,L., Yang,M., Guo,H., Yang,L., Wu,J., Li,R., Liu,P., Lian,Y., Zheng,X., Yan,J., *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, **20**, 1131–1139.
138. Pollen,A.A., Nowakowski,T.J., Shuga,J., Wang,X., Leyrat,A.A., Lui,J.H., Li,N., Szpankowski,L., Fowler,B., Chen,P., *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, **32**, 1053–1058.
139. Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann and Wolfgang Enard7 (2017) Comparative Analysis of Single-Cell RNA Sequencing Methods: Molecular Cell. 10.1016/j.molcel.2017.01.023.
140. Schroth,G.P., Gertz,J., Myers,R.M., Williams,B.A., McCue,K., Marinov,G.K. and Wold,B.J. (2013) From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.*, **24**, 496–510.
141. Soneson,C. and Robinson,M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, **15**, 255–261.
142. Soneson,C. and Robinson,M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, **15**, 255–261.
143. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A., *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods 2009 65*, **6**, 377.
144. Zhang,X., Li,T., Liu,F., Chen,Y., Yao,J., Li,Z., Huang,Y. and Wang,J. (2018) Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol. Cell*, 10.1016/J.MOLCEL.2018.10.020.
145. Hochgerner,H., Lönnerberg,P., Hodge,R., Mikes,J., Heskol,A., Hubschle,H., Lin,P., Picelli,S., La Manno,G., Ratz,M., *et al.* (2017) STRT-seq-2i: dual-index 5′ single cell and nucleus RNA-seq on an addressable microwell array. *Sci. Rep.*, **7**, 16327.
146. Picelli,S., Faridani,O.R., Björklund,Å.K., Winberg,G., Sagasser,S. and Sandberg,R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.

147. Chen,K.H., Boettiger,A.N., Moffitt,J.R., Wang,S. and Zhuang,X. (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-.).*, **348**, 1360–1363.
148. Kulkarni,A., Anderson,A.G., Merullo,D.P. and Konopka,G. (2019) Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr. Opin. Biotechnol.*, **58**, 129–136.
149. Lubeck,E., Coskun,A.F., Zhiyentayev,T., Ahmad,M. and Cai,L. (2014) Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods*, **11**, 360–361.
150. Paper,W., Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., *et al.* (2017) The human cell atlas. *Elife*, **6**, 1–30.
151. Lee,J.H., Daugharthy,E.R., Scheiman,J., Kalhor,R., Amamoto,R., Peters,D.T., Turczyk,B.M. and Marblestone,A.H. (2014) Sequencing in Situ. 10.1126/science.1250212.
152. Ke,R., Mignardi,M., Pacureanu,A., Svedlund,J., Botling,J., Wählby,C. and Nilsson,M. (2013) In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods*, **10**, 857–860.
153. Kolodziejczyk,A.A., Kim,J.K., Svensson,V., Marioni,J.C. and Teichmann,S.A. (2015) The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell*, **58**, 610–620.
154. Gupta,I., Collier,P.G., Haase,B., Mahfouz,A., Joglekar,A., Floyd,T., Koopmans,F., Barres,B., Smit,A.B., Sloan,S.A., *et al.* (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.*, **36**, 1197–1202.
155. Gardeux,V., David,F.P.A., Shajkofci,A., Schwalie,P.C. and Deplancke,B. (2017) ASAP: A web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics*, **33**, 3123–3125.
156. Franzén,O., Gan,L.-M. and Björkegren,J.L.M. (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford).*, **2019**.
157. Dumas,P.-Y., Naudin,C., Martin-Lannerée,S., Izac,B., Casetti,L., Mansier,O., Rousseau,B., Artus,A., Dufossée,M., Giese,A., *et al.* (2019) Hematopoietic niche drives FLT3-ITD acute myeloid leukemia resistance to quizartinib via STAT5- and hypoxia- dependent up-regulation of AXL. *Haematologica*, 10.3324/haematol.2018.205385.
158. Baslan,T. and Hicks,J. (2017) Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat. Rev. Cancer*, **17**, 557–569.
159. Landau,D.A., Carter,S.L., Stojanov,P., McKenna,A., Stevenson,K., Lawrence,M.S., Sougnez,C., Stewart,C., Sivachenko,A., Wang,L., *et al.* (2013) Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell*, **152**, 714–726.
160. Carter,S.L., Cibulskis,K., Helman,E., Mckenna,A., Shen,H., Beroukhim,R., Pellman,D., Levine,D.A. and Lander,E.S. (2015) HHS Public Access. **30**, 413–421.
161. Visvader,J.E. (2011) Cells of origin in cancer. *Nature*, **469**, 314–322.
162. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A., Kinzler,K.W. and Kinzler,K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–58.
163. Stachler,M.D., Taylor-Weiner,A., Peng,S., McKenna,A., Agoston,A.T., Odze,R.D., Davison,J.M., Nason,K.S., Loda,M., Leshchiner,I., *et al.* (2015) Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat. Genet.*, **47**, 1047–55.
164. Navin,N.E. (2014) Cancer genomics: one cell at a time. *Genome Biol.*, **15**, 452.
165. Gao,R., Kim,C., Sei,E., Foukakis,T., Crosetto,N., Chan,L.K., Srinivasan,M., Zhang,H., Meric-Bernstam,F. and Navin,N. (2017) Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nat. Commun.*, **8**.
166. Martelotto,L.G., Baslan,T., Kendall,J., Geyer,F.C., Burke,K.A., Spraggon,L., Piscuoglio,S., Chadalavada,K., Nanjangud,G., Ng,C.K.Y., *et al.* (2017) Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. *Nat. Med.*, **23**, 376–385.
167. Bakken,T.E., Hodge,R.D., Miller,J.A., Yao,Z., Nguyen,T.N., Aevermann,B., Barkan,E., Bertagnolli,D., Casper,T., Dee,N., *et al.* (2018) Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One*, **13**, e0209648.
168. Leung,M.L., Wang,Y., Waters,J. and Navin,N.E. (2015) SNES: Single nucleus exome sequencing. *Genome Biol.*, **16**.
169. Thomsen,E.R., Mich,J.K., Yao,Z., Hodge,R.D., Doyle,A.M., Jang,S., Shehata,S.I., Nelson,A.M.,

Shapovalova,N. V, Levi,B.P., *et al.* (2016) Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat. Methods*, **13**, 87–93.
170. Alles,J., Karaiskos,N., Praktiknjo,S.D., Grosswendt,S., Wahle,P., Ruffault,P.-L., Ayoub,S., Schreyer,L., Boltengagen,A., Birchmeier,C., *et al.* (2017) Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.*, **15**, 44.
171. Lampignano,R., Yang,L., Neumann,M.H.D., Franken,A., Fehm,T., Niederacher,D. and Neubauer,H. (2017) A Novel Workflow to Enrich and Isolate Patient-Matched EpCAMhigh and EpCAMlow/negative CTCs Enables the Comparative Characterization of the PIK3CA Status in Metastatic Breast Cancer. *Int. J. Mol. Sci.*, **18**.
172. Di Trapani,M., Manaresi,N. and Medoro,G. (2018) DEPArray™ system: An automatic image-based sorter for isolation of pure circulating tumor cells. *Cytom. Part A*, **93**, 1260–1266.
173. Adams,D.L., Martin,S.S., Alpaugh,R.K., Charpentier,M., Tsai,S., Bergan,R.C., Ogden,I.M., Catalona,W., Chumsri,S., Tang,C.-M., *et al.* (2014) Circulating giant macrophages as a potential biomarker of solid tumors. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 3514–9.

DANKWOORD

Met het schrijven van dit dankwoord ben ik bijna aan het einde gekomen van mijn doctoraat. Het waren vier fantastische jaren, met zoals in elk doctoraat af en toe een dipje, maar vooral heel veel goede momenten. Ik kwam elke dag met veel plezier werken, enerzijds omdat ik elke dag opnieuw uitgedaagd werd op wetenschappelijk vlak en anderzijds door de fantastische collega's en de goede sfeer op het werk. Het was leuk om in zo een team te werken, waar de wetenschap centraal stond, maar waar er ook plaats was voor ontspannende activiteiten na het werk, zoals de vele lab weekends, FSP lunches en team activiteiten, maar even goed de zomerse avonden waarop we impulsief besloten om nog even van het zonnetje te gaan genieten.

Eerst en vooral wil ik mijn promotoren, Frank, Jo en Kaat, bedanken om mij de kans te geven om dit doctoraat te doen. Het was uitdagend, maar ook super interessant om twee uiteenlopende projecten te hebben. Bedankt voor de interessante discussies, waarbij jullie input vanuit een andere invalshoek ongelooflijk waardevol was voor de vooruitgang van mijn onderzoek. Bedankt Frank, om mij steeds te motiveren met jouw enthousiasme voor het onderzoek en veel interessante nieuwe hypotheses. Bedankt Jo, voor jouw optimisme, om mij te blijven uitdagen om net iets verder te gaan en toch nog dat extra experiment te doen. Bedankt Kaat, voor de opvolging van dichtbij, om mee te denken over de experimenten en interpretatie, voor de vele input die je mij gegeven hebt. Daarnaast wil ik ook graag Annelynn bedanken, om tijdens mijn masterproef mijn interesse voor het onderzoek te wekken en altijd klaar te staan om te helpen. De euforie die ik soms zag bij jou voor een goed resultaat tijdens mijn masterthesis, en ik toen niet helemaal begreep, heb ik tijdens mijn doctoraat ook verschillende keren mogen ondervinden, waarna ik met een grote glimlach en al huppelend door de gang liep. Siebe, blij dat wij 'team leukemie' vormden. Bedankt voor de wetenschappelijke babbels, maar even goed voor de drinks en ontspannende momenten tussendoor. Lieve Nurten, bedankt voor de aangename samenwerking. Met jou heb ik het allereerste en ook allerlaatste experiment van mijn PhD gedaan. Als een 'tandem' hebben we samen veel *single cell* experimenten gedaan; bedankt om enthousiast te blijven, ondanks de vele tegenslagen en mij te motiveren als ik een dipje had! Bedankt Laura, om last minute nog mijn voorblad te maken. Ook mijn paranimfen, Eva en Lisa, wil ik bedanken voor al de praktische zaken, maar zeker ook voor de talloze ontspannende momenten en motiverende woorden. Elke reden was goed om te klinken met een goed glas cava!

Lieve eetclub-vriendinnen, bedankt voor de maandelijkse gezellige etentjes, om mij even te laten ontspannen tijdens drukke periodes. Na de drukke verbouwingen in combinatie met mijn doctoraat heb ik vanaf nu weer meer tijd om samen op pad te gaan, beginnend met ons vriendinnenweekend. Graag wil ik ook Lisa bedanken, om er steeds te zijn, om klaar te staan met de juiste woorden of snel een dessertje binnen te steken om mij te motiveren! Een grote dankjewel ook aan mijn ouders en zus. Bedankt om 'even' wat huishoudelijke taken over te nemen, om te blijven helpen verbouwen, ondanks dat ik boven aan het schrijven was, en om mij te steunen tijdens mijn doctoraat.

Bedankt aan al de CMGG collega's voor de vier fantastische jaren, jullie zorgden ervoor dat ik steeds met heel veel plezier kwam werken en met een grote glimlach zal terug denken aan deze periode!

CURRICULUM VITAE

Karen Verboom

Personalia

Address:	Jan Dhondtstraat 20
	9050 Gentbrugge
	Belgium
Phone:	+32497061391
E-mail:	Karen.Verboom@UGent.be
	karen_verboom@hotmail.com
Date of birth:	June 7, 1992
Nationality:	Belgian

Experience

Function:	PhD researcher
Period:	September 2015 – September 2019
Institute:	Center for Medical Genetics, Ghent University, Ghent, Belgium
Thesis:	Deciphering complete cancer transcriptomes from bulk to single cell level
Promoter:	Prof. dr. Frank Speleman
Co-promoters:	Prof. dr. Jo Vandesompele and Kaat Durinck
Techniques:	RNA sequencing, ChIP sequencing, ATAC sequencing, single cell RNA sequencing, RT-qPCR, tissue culture, Western Blot and FACS sorting

Education

Master of Science in the Biomedical Sciences – Medical Genetics

Period:	2010-2015
Institute:	Ghent University, Ghent, Belgium
Thesis:	Study of long noncoding RNAs in the development of T-cell acute lymphoblastic leukemia
Promoter:	Prof. dr. Frank Speleman
Techniques:	tissue culture, RNA isolation, RT-qPCR, PCR, transformation of bacteria, plasmid purification, electroporation, FACS sorting and Western blotting

Doctoral training program

- Career management – Ugent (1, 2, 15 October 2018, Gent, Belgium)
- Scientific writing – Ugent (October-December 2017, Gent, Belgium)
- Basic statisctics theory – VIB (December 16, 2016, Leuven, Belgium)
- Introductory statistics in GraphPad Prism – VIB (December 9, 2016, Leuven, Belgium)
- Effective scientific communication – UGent (November 9, 29, 30, 2016, Gent, Belgium)
- RNA-seq analysis for differential expression in Genepattern – VIB (October 17, 2016, Leuven, Belgium)
- Course and master classes on molecular aspects of hematological disorders–ErasmusMC (June 7-8, 2016, Rotterdam, Netherlands)
- Non-coding genome – VIB (May 12, 2016, Leuven, Belgium)
- Introduction to HPC @Ugent– Ugent (April 26, 2016, Gent, Belgium)
- Conference skills– Ugent (January-May, 2016, Gent, Belgium)
- RNA-seq analysis for differential expression – VIB (September 25 and 28, 2015, Leuven, Belgium)

Scientific achievements

Publications

Verboom K, Van Loocke W, Clappier E, Soulier J, Vandesompele J, Speleman F and Durinck K (2019). **A comprehensive dataset of TLX1 positive ALL-SIL lymphoblasts and primary T-cell acute lymphoblastic leukemias**. Submitted to Data.

Verboom K*, Everaert C*, Bolduc N, Livak JK, Yigit N, Rombaut D, Anckaert J, Venø TM, Kjems J, Speleman F, Mestdagh P and Vandesompele J (2019). **SMARTer single cell total RNA-sequencing.** Nucleic Acids Research. Accepted. *contributed equally.

Lorenzi L, Avila Cobos F, Decock A, Everaert C, Helsmoortel H, Lefever S, Verboom K, Volders PJ, Speleman F,Vandesompele J, Mestdagh P (2019). **Long noncoding RNA expression profiling in cancer: challenges and opportunities.** Genes Chromosomes Cancer 58(4):191-199.

Loontiens S, Depestel L, Vanhauwaert S, Dewyn G, Gistelinck C, Verboom K, Van Loocke W, Matthijssens F, Willaert A, Vandesompele J, Speleman F, Durinck K (2019). **RNA isolation method for single embryo transcriptome analysis in zebrafish.** BMC Genomics 20(1):228.

D'haene E, Bar-Yaacov R, Bariah I, Vantomme L, Van Loo S, Cobos FA, Verboom K, Eshel R, Alatawna R, Menten B, Birnbaum RY, Vergult S (2018). **A neural enhancer network upstream of MEF2C is compromised in patients with rett-like characteristics.** Hum. Mol. Genet. 28(5):818-827.

Verboom K, Van Loocke W, Volders PJ, Decaesteker B, Avila Cobos F, Bornschein S, de Bock CE, Kalender Atak Z, Clappier E, Aerts S, Cools J, Soulier J, Taghon T, Van Vlierberghe P, Vandesompele J, Speleman F, Durinck K(2018). **A comprehensive inventory of TLX1 controlled long non-coding RNAs in T-cell acute lymphoblastic leukemia through polyA+ and total RNA sequencing.** Haematologica 103(12):E585-589.

Decaesteker B, Denecker G, Van Neste C, Dolman EM, Van Loocke W, Gartlgruber M, De Vloed F, Depuydt P, Verboom K, Rombaut D, Loontiens S, De Wyn Jolien, Kholosy WM, Koopmans B, Essing AHW, Herrmann C, Dreidax D, Durinck K, Deforce D, van Nieuwerburgh F, henssen A,

Versteeg R, Boeva V, Schleiermacher G, van Nes J, Mesdagh P, Vanhauwaert S, Schulte JH,Westermann F, Molenaar JJ, De Preter K, Speleman F (2018). **TBX2 is a neuroblastoma core regulatory circuitry component enhancing MYCN/FOXM1 reactivation of DREAM targets**. Nature communications 9(1):4866.

Grants

BOF grant: Epigenetic re-activation of T-ALL tumor suppressors under control of TLX1 driven enhancer RNAs. October 1, 2017, Belgium.

BOF grant: Epigenetic re-activation of T-ALL tumor suppressors under control of TLX1 driven enhancer RNAs. October 1, 2015, Belgium.

Oral presentations

Verboom K. **The TLX1 oncogene modulates the enhancer RNA landscape in T-ALL**. Course and master classes on molecular aspects of hematological disorders. June 20-21, 2017, Rotterdam, Netherlands.

Verboom K. **The TLX1 oncogene modulates the enhancer RNA landscape in T-ALL**. fTALES: Cancer an old dog with new tricks. March 6-7, 2017, Leuven, Belgium.

Verboom K. **The T-ALL oncogene TLX1 controls enhancer lncRNA expression.** f-TALES: Light on the dark side of the genome . September 15, 2016, Ghent, Belgium.

Verboom K. Dissecting the NOTCH1 driven transcriptional landscape in T-cell acute lymphoblastic leukemia at single cell level. **Oncopoint. March 2, 2016, Ghent, Belgium.**

Verboom K. **Single cell RNA sequencing at CRIG**. Single cell workshop. January 19, 2016, Ghent, Belgium.

Poster presentations

Verboom K, Everaert C, Bolduc N, Livak JK, Yigit N, Rombaut D, Anckaert J, Venø MT, Kjems J, Speleman F, Mestdagh P and Vandesompele J. **SMARTer single cell total RNA sequencing.** Keystone Symposia: Single cell biology. January 13-17, 2019, Breckenridge, USA.

Verboom K, Everaert C, Bolduc N, Livak JK, yigit N, Rombaut D, Anckaert J, Venø MT, Kjems J, Speleman F, Mestdagh P and Vandesompele J. **SMARTer single cell total RNA sequencing**. CRIG's single cell mini-symposium. December 5, 2018, Gent, Belgium.

Verboom K, Van Loocke W, Vandesompele J, Vandamme N, Berx G, Saeys Y, Martnes L, Van Vlierberghe P, Speleman F and Durinck K. **The TLX1 oncogene modulates the enhancer RNA landscape in T-ALL.** Chromatin architecture and chromosome organization. March 24-27, 2018, Whistler, Canada.

Verboom K, Van Loocke W, Vandesompele J, Van Vlierberghe P, Speleman F and Durinck K. **A comprehensive inventory of TLX1 controlled long non-coding RNAs in T-cell acute**

lymphoblastic leukemia. Keystone Symposia: The epigenome in development and disease. February 16, 2018, Gent, Belgium.

Verboom K, Durinck K, De Decker M, Van Loocke W, Matthijssens F, Soulier J-J, De Laat W, Taghon T, Van Vlierberge P and Speleman F. **The TLX1 oncogene modulates the enhancer RNA landscape in T-ALL.** T-ALL workshop. March 17-19, 2017, Leuven, Belgium.

Verboom K, Durinck K, De Decker M,Van Loocke W, Matthijssens F, Soulier J-J, De Laat W, Taghon T, Van Vlierberge P and Speleman F. **The TLX1 oncogene modulates the enhancer RNA landscape in T-ALL.** fTALES: Cancer an old dog with new tricks. March 6-7, 2017, Leuven, Belgium.

Verboom K, Durinck K, Van Loocke W, Matthijssens F, Van de Walle I, Wallaert A, Volders P-J, Van Roy N, Benoit Y, Poppe B, Rondou P, Mestdagh P, Vandesompele J, De laat W, Soulier J-J, Taghon T, Van Vlierberge P and Speleman F. **The T-ALL oncogene TLX1 controls enhancer lncRNA expression.** BeSHG: The epigenome in development and disease. February 17, 2017, Louvain-La-Neuve, Belgium.

Verboom K, Durinck K, Van Loocke W, Matthijssens F, Van de Walle I, Wallaert A, Volders P-J, Van Roy N, Benoit Y, Poppe B, Rondou P, Mestdagh P, Vandesompele J, De laat W, Soulier J-J, Taghon T, Van Vlierberge P and Speleman F. **The T-ALL oncogene TLX1 controls enhancer lncRNA expression.** Keystone Symposia: Noncoding RNAs: from disease to targeted therapeutics. February 5-9, 2017, Banff, Canada.

Verboom K, Durinck K, Yigit N, Everaert C, Cannoodt R, Van Vlierberghe P, Vandesompele J and Speleman F. **Dissecting the NOTCH1 driven transcriptional landscape in T-cell acute lymphoblastic leukemia at single cell level.** Single cell biology 2016. March 8-10, 2016, Hinxton, UK.

Conferences

- BeSHG: Precision medicine, March 15, 2019, Luik, Belgium
- Keystone Symposia: Single cell biology. January 13-19, 2019, Breckenridge, USA.
- CRIG's single cell mini-symposium. December 5, 2018, Ghent, Belgium.
- Keystone Symposia: Chromatin architecture and chromosome organization. March 24-27, 2018, Whistler, Canada.
- BeSHG: The epigenome in development and disease. February 16, 2018, Gent, Belgium.
- Course and master classes on molecular aspects of hematological disorders. June 20-21, 2017, Rotterdam, Netherlands.
- Enhancer structure and function. April 19-21, 2017, Bordeaux, France.
- Zebrafish cancer modelling: state of the art and novel tools. March 20, 2017, Gent, Belgium.
- T-ALL workshop. March 17-19, 2017, Leuven, Belgium.
- Oncopoint. March 15, 2017, Ghent, Belgium
- f-TALES: Cancer an old dog with new tricks. March 6-7, 2017, Leuven, Belgium.
- BeSHG. February 17, 2017, Louvain-La-Neuve, Belgium.
- Keystone Symposia: Noncoding RNAs: from disease to targeted therapeutics. February 5-9, 2017, Banff, Canada.
- Hallmarks of cancer. December 11-13, 2016, Ghent, Belgium.

- f-TALES: Light on the dark side of the genome. September 15-16, 2016, Ghent, Belgium.
- Course and master classes on molecular aspects of hematological disorders. June 7-8, 2016, Rotterdam, Netherlands.
- Single cell biology. March 8-10, 2016, Hinxton, UK.
- Oncopoint. March 2, 2016, Ghent, Belgium
- Keystone Symposia: Noncoding RNAs in health and disease/ Enhancer malfunctions in cancer, February 21-24, 2016, Santa Fe, USA.
- Genome engineering and synthetic biology: second edition, January 28-29, 2016, Ghent, Belgium.

Student Guidance

Baptiste Oosterlinck (September 2017 – June 2018). Master thesis, master in science in biomedicine engineering (Ghent University, Faculty of bioscience engineering). **Studie van enhancer RNA's in de ontwikkeling van T-cel acute lymfoblastische leukemie.** Promotor: Speleman F, Durinck K, De Vos W, mentor: Verboom K.

Tom Coucke (2017). Z-line paper, bachelor of Medicine (Ghent University, Faculty of Medicine and health sciences). **Gamma secretase inhibitor combinatietherapieën in de behandeling van T-cel acute lymfoblastische leukemie.** Promotor: Speleman F, mentor: Verboom K.

Celine Haegeman (2017). Z-line paper, bachelor of Medicine (Ghent University, Faculty of Medicine and health sciences). **Gamma secretase inhibitor combinatietherapieën in de behandeling van T-cel acute lymfoblastische leukemie.** Promotor: Speleman F, mentor: Verboom K.

Astrid Rycx (2016). Z-line paper, bachelor of Medicine (Ghent University, Faculty of Medicine and health sciences). **CRISPR: een nieuwe revolutionaire techniek voor gentherapie: het voorbeeld van β-thalassemie**. Promotor: Speleman F, mentor: Verboom K.

James Schelfaut (2016). Z-line paper, bachelor of Medicine (Ghent University, Faculty of Medicine and health sciences). **CRISPR: een nieuwe revolutionaire techniek voor gentherapie: het voorbeeld van β-thalassemie.** Promotor: Speleman F, mentor: Verboom K.