Editorial

For reprint orders, please contact: reprints@futuremedicine.com

Future MICROBIOLOGY

Genome sequence-based curation of PubMLST data challenges interspecies recombination in the *Burkholderia cepacia* complex

Charlotte Peeters[1], Eliza Depoorter[1], Evelien De Canck[1] & Peter Vandamme*,[1]

[1]Laboratory of Microbiology, Department of Biochemistry & Microbiology, Faculty of Sciences, Ghent University, Ghent, Belgium

*Author for correspondence: peter.vandamme@ugent.be

"The pubMLST databases and underlying BIGSdb platform provide a well-documented, extensive tool set to not only query available sequence and provenance data, but also conveniently submit new data through an easy-to-use submission portal"

First draft submitted: 30 January 2020; Accepted for publication: 17 July 2020; Published online: 21 September 2020

Keywords: *Burkholderia cepacia* complex • epidemiology • genome sequence • identification • MLST • PubMLST • sequence type • sequencing error • typing

Multilocus sequence typing (MLST) has been the gold standard for typing and identification of a wide range of bacteria for two decades [1–3]. Public MLST databases allow researchers worldwide to analyze and deposit data, and enable the study of the global prevalence and epidemiology of a broad range of bacteria [3,4]. A *Burkholderia cepacia* complex (Bcc) MLST scheme based on partial *atpD*, *gltB*, *gyrB*, *recA*, *lepA*, *phaC* and *trpB* gene sequence analysis was developed for both species and strain level differentiation [5–7]. PCR primers were subsequently improved to reliably amplify the target loci from both Bcc and non-Bcc *Burkholderia* bacteria and to enable the use of a single primer set for amplification and sequencing [8]. The large number of recent publications that used MLST and the accompanying Bcc PubMLST database as a tool for epidemiological studies shows that MLST is a well-established method that enabled outbreak surveillance, shed light on the global distribution of strains and elucidated Bcc epidemiology and population structure [9]. While the first few hundred sequence types were primarily originating from European, American and Canadian isolates, there has been an increase in submissions from Australia and countries in Asia and South-America. Reproducibility and portability have been considered major advantages of MLST over earlier typing and identification methods [1,2]. In today's genomics era, traditional Sanger sequencing of MLST loci is gradually replaced by the extraction of MLST alleles from next-generation sequencing data, thus sustaining the continued use of the same MLST schemes [2,10]. As curators of the Bcc PubMLST database [3,11] we observed that genome sequence derived MLST data revealed several types of conflicts with earlier MLST data that were generated through Sanger sequencing. We generated genome sequences from high-coverage Illumina data for 113 Bcc isolates (method as previously described by Peeters *et al.* [12], [UNPUBLISHED DATA]) for which Sanger sequencing based MLST data were available; for 34 of these isolates (30%) there was a conflict between the genome sequence derived and earlier MLST data.

Generally, two types of conflict were found. In one type of conflict the genome sequence derived MLST alleles revealed one or a few single nucleotide polymorphisms compared with the earlier MLST data. An example of this type of error was found for *Burkholderia multivorans* outbreak strain C1576 [13], for which the initial MLST analysis yielded *lepA*-8 and *trpB*-6 [7], while more recent analyses uncovered *lepA*-224 [14] (1/397 nucleotide differences) and *trpB*-415 (GenBank/ENA accession number ERS784904) (3/301 nucleotide differences), changing the sequence type from ST-27 into ST-899. Because the coverage of Illumina data by far exceeds that of traditional Sanger sequencing, it is not unexpected to find a few false single nucleotide polymorphisms in the original MLST data [15,16].

Future Medicine

In a second type of conflict, strongly different alleles were observed. An example of this type of error was found for *B. multivorans* LMG 18824, for which the initial MLST analysis yielded *gyrB*-307, while the genome sequence uncovered *gyrB*-445 (36/454 nucleotide differences), changing the sequence type from ST-523 into ST-1530. This kind of discrepancy was most likely introduced by human error, for example, a mix-up of amplicons during PCR or a typographical error during the submission of the MLST data to the PubMLST database. This type of error not only causes a change in sequence type but may also alter the strain's phylogenetic position. Indeed, the *gyrB*-307 allele is detected only in *Burkholderia cenocepacia* IIIB while *gyrB*-445 thus far occurs only in *B. multivorans,* so this mix-up incorrectly revealed shared MLST loci between Bcc species and suggested interspecies recombination [5]. So far, we found that 92% (i.e., 48/52) of those alleles that were originally thought to be shared among several Bcc species and that were verified through new sequence data could in fact be explained by human error, challenging the concept of interspecies recombination within the Bcc. Of note, we confirmed three alleles that were shared between *B. cenocepacia* lineages IIIA and IIIB (*atpD*-16, *phaC*-6 and *phaC*-121), and detected only a single case of genuine interspecies recombination, more specifically *atpD*-123 that was found both in *B. multivorans* and *B. pseudomultivorans*.

The errors described above not only affected the allele profile and sequence type of taxonomic reference strains such as the *Burkholderia anthina*, *Burkholderia dolosa*, '*Burkholderia paludis*', *Burkholderia plantarii*, *Burkholderia pyrrocinia* and *Burkholderia vietnamiensis* type strains, but also well-documented Bcc strains such as *B. cenocepacia* K56-2, a representative of the ET12 epidemic lineage [17], and *B. cenocepacia* H111 [18].

While the occurrence of sequencing errors that lead to a few false single nucleotide polymorphisms can be expected to gradually disappear with increasing use of genome sequence derived MLST data, researchers, but also referees and editors ought to be more alert for human error and strain mix-ups. For example, a recent study [19] presented both Sanger sequencing based MLST data and a genome sequence for strain MSh1^T that was presented as the type strain of the novel Bcc species '*B. paludis*'. Its Sanger sequencing based sequence type ST-1043 was marked as suspicious during our November 2016 curation of the database because it formed an unusual long branch in the phylogenetic tree of concatenated allele sequences of all Bcc STs (data not shown). Closer inspection of the data revealed that ST-1043 comprised alleles from seven different *Burkholderia* species. The sequence type extracted from the genome sequence generated in the Ong *et al.* [19] study yielded ST-1347, and comprised new alleles for all seven *loci*. Furthermore, when we accessioned the deposited '*B. paludis*' type strain from the BCCM/LMG bacteria collection (i.e., LMG 30113^T [20]), its resequenced genome sequence yielded ST-1381 which differed in its *gltB* (2/400 nucleotide differences), *gyrB* (10/454 nucleotide differences), *recA* (1/393 nucleotide difference) and *phaC* (2/385 nucleotide differences) alleles from ST-1347, suggesting that MSh1^T and LMG 30113^T do not represent the same strain, or possibly another cause of large experimental error.

We are also cocurator of the *Achromobacter* PubMLST database [21] and this database too suffers from the same conflicts between earlier MLST data and genome sequence derived MLST data. Thus far, genome sequences of 84 *Achromobacter* isolates (method as previously described by Peeters *et al.* [12], [UNPUBLISHED DATA]) for which earlier MLST data were in the database, revealed similar conflicts for 26 of these isolates (31%).

Clearly the Sanger sequencing methodology and human errors introduced an unexpectedly high number of errors in MLST databases that are considered of superior reproducibility and portability [1,2]. The pubMLST databases and underlying BIGSdb platform provide a well-documented, extensive tool set to not only query available sequence and provenance data, but also conveniently submit new data through an easy-to-use submission portal [3]. Together with the increasing numbers of high-quality genome sequences that can be expected in the near future, this will lead to a gradual disappearance of false single nucleotide polymorphisms and mistaken cases of interspecies recombination. The true extent of interspecies recombination within the Bcc will become clear and it will be most interesting to study the alleles – and larger sets of genes – that are genuinely shared among species. However, the '*B. paludis*' example also shows that vigilance of researchers, database curators, referees and editors is mandatory to safeguard the quality of the data in these public databases and ensure a correct understanding of distribution, population structure and epidemiology of the bacteria involved.

Author contributions

C Peeters, E Depoorter and P Vandamme conceived the study. C Peeters and P Vandamme wrote the manuscript. E Depoorter and E De Canck critically revised the manuscript. E De Canck performed the wet-lab procedures and acquired the data. C Peeters and E Depoorter analyzed and interpreted the data. P Vandamme generated the required funding. All authors read and approved the final manuscript.

Acknowledgments

This publication made use of the PubMLST website (https://pubmlst.org/) developed by K Jolley [3] and sited at the University of Oxford. The development of that website was funded by the Wellcome Trust. We thank M Bull, S Cardona, K Dyet, H Davies, V Gautam, C Hall, B Jovcic, T Kidd, JJ LiPuma, E Mahenthiralingam, T Spilker and J Zlosnik for their help with the correction of MLST data.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

1. Maiden MCJ, Bygraves JA, Feil E *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* 95(6), 3140–3145 (1998).
2. Jolley KA, Maiden MC. Using MLST to study bacterial variation: prospects in the genomic era. *Future Microbiol.* 9(5), 623–630 (2014).
3. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* 3(124), 1–20 (2018).
4. Maiden MCJ. Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* 60, 561–588 (2006).
5. Baldwin A, Mahenthiralingam E, Drevinek P *et al.* Elucidating global epidemiology of Burkholderia multivorans in cases of cystic fibrosis by multilocus sequence typing. *J. Clin. Microbiol.* 46(1), 290–295 (2008).
6. Baldwin A, Mahenthiralingam E, Drevinek P *et al.* Environmental Burkholderia cepacia complex isolates in human infections. *Emerg. Infect. Dis.* 13(3), 458–461 (2007).
7. Baldwin A, Mahenthiralingam E, Thickett KM *et al.* Multilocus sequence typing scheme that provides both species and strain differentiation for the Burkholderia cepacia complex. *J. Clin. Microbiol.* 43(9), 4665–4673 (2005).
8. Spilker T, Baldwin A, Bumford A, Dowson CG, Mahenthiralingam E, LiPuma JJ. Expanded multilocus sequence typing for Burkholderia species. *J. Clin. Microbiol.* 47(8), 2607–2610 (2009).
9. Bcc pubMLST database. Recent publications using MLST in Burkholderia research. https://pubmlst.org/bcc/references.shtml
10. Pérez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infect. Genet. Evol.* 16, 38–53 (2013).
11. Bcc pubMLST database. https://pubmlst.org/bcc/
12. Peeters C, De Canck E, Cnockaert M *et al.* Comparative genomics of Pandoraea, a genus enriched in xenobiotic biodegradation and metabolism. *Front. Microbiol.* 10(2556), 1–21 (2019).
13. Whiteford ML, Wilkinson JD, McColl JH *et al.* Outcome of Burkholderia (Pseudomonas) cepacia colonisation in children with cystic fibrosis following a hospital outbreak. *Thorax* 50(11), 1194–1198 (1995).
14. Denman CC, Brown AR. Mannitol promotes adherence of an outbreak strain of Burkholderia multivorans via an exopolysaccharide-independent mechanism that is associated with upregulation of newly identified fimbrial and afimbrial adhesins. *Microbiology* 159, 771–781 (2013).
15. Liu L, Li Y, Li S *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012, 1–12 (2012). https://www.hindawi.com/journals/bmri/2012/251364/
16. Ferres I, Iraola G. MLSTar: automatic multilocus sequence typing of bacterial genomes in R. *PeerJ.* 6, e5098 (2018).
17. Mahenthiralingam E, Coenye T, Chung JW *et al.* Diagnostically and experimentally useful panel of strains from the Burkholderia cepacia complex. *J. Clin. Microbiol.* 38(2), 910–913 (2000).
18. Schwager S, Agnoli K, Köthe M *et al.* Identification of Burkholderia cenocepacia strain H111 virulence factors using nonmammalian infection hosts. *Infect. Immun.* 81(1), 143–153 (2013).
19. Ong KS, Aw YK, Lee LH, Yule CM, Cheow YL, Lee SM. Burkholderia paludis sp. nov., an antibiotic-siderophore producing novel Burkholderia cepacia complex species, isolated from Malaysian tropical peat swamp soil. *Front. Microbiol.* 7(2046), 1–14 (2016).
20. BCCM/LMG bacteria collection. http://bccm.belspo.be/about-us/bccm-lmg
21. Achromobacter pubMLST database. https://pubmlst.org/achromobacter/