Pan-Genome Analysis of Mycobacterium Africanum: Insights to Dynamics and Evolution.

Idowu Olawoye (olawoyei0303@run.edu.ng)
Redeemer's University https://orcid.org/0000-0002-6658-9917
Simon D.W. Frost
Microsoft Research
Christian T. Happi
Redeemer's University

Research article

Keywords: Mycobacterium tuberculosis, pan-genome, bioinformatics

Posted Date: June 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34142/v1

Abstract

Background: *Mycobacterium tuberculosis* complex (MTBC) consists of seven major lineages with three of them reported to circulate within West Africa: lineage 5 (West African 1) and lineage 6 (West African 2) which are geographically restricted to West Africa and lineage 4 (Euro-American lineage) which is found globally. It is unclear why the West African lineages are not found elsewhere; some hypotheses suggest that it could either be harboured by an animal reservoir which is restricted to West Africa, or strain preference for hosts of West African ethnicity, or inability to compete with other lineages in other locations.

We tested the hypothesis that *M. africanum* West African 2 (lineage 6) might have emigrated out of West Africa but was outcompeted by more virulent modern strains of *M. tuberculosis* (MTB).

Whole genome sequences of *M. tuberculosis* from Nigeria (n=21), South Africa (n=24) and *M. africanum* West African 2 from Mali (n=22) were retrieved, and a pan-genome analysis was performed after fully annotating these genomes.

Results: The outcome of this analysis shows that Lineages 2, 4 and 6 all have a close pan-genome. We also see a correlation in numbers of some multiple copy core genes and amino acid substitution with lineage specificity that may have contributed to geographical distribution of these lineages.

Conclusions: The findings in this study provides a perspective to one of the hypotheses that *M. africanum* West African 2 might find it difficult to compete against the more modern lineages outside West Africa hence its localization to the geographical region.

Background

Tuberculosis (TB) is the world leading cause of death from infectious diseases, with the latest TB report stating over 1.1 million deaths globally in 2018, with two thirds from low-to-middle income countries such as China, Indonesia, Nigeria, the Philippines, Pakistan, South Africa and Bangladesh (1). *Mycobacterium tuberculosis* complex (MTBC), the causative organism for tuberculosis has been studied widely and is known to be a monoclonal bacteria compared to other bacterial models acquiring variation predominantly through mutations (2). MTBC is grouped in two major ancestries, the ancient lineages and the modern lineages, both of which are further grouped into lineages and sub lineages based on the geographical regions that they are found. Lineage 1 (East Africa, Philippines, in the region of the Indian Ocean), lineage 2 (East Asia), lineage 3 (East Africa, Central Asia), lineage 4 (Europe, America and Africa), lineage 5 (West Africa 1), lineage 6 (West Africa 2) and lineage 7 in Ethiopia (3). The ancestry of these lineages was established by the study of 20 variable genomic regions that are caused by insertion or deletion events (4). For example, the absence or presence of an MTB specific deletion known as TbD1 and regions of difference (RD) helps to classify MTBC into modern (lineages 2,3 and 4) or ancient (lineages 1, 5. 6 and 7) strains.

Different lineages of MTBC have shown to present different symptoms and immunological responses as seen in several studies. Although there is limited knowledge of TB virulence, pathogenicity does not associate with classical virulence factors like toxins (5), but rather with other complex factors, such as other bacterial infections being harboured by the patient that can interact with the tubercle bacilli, the host immune system, or even environmental factors resulting in a complex cascade of events (6). Furthermore, reports have linked TB infectivity to human genetic susceptibility, with certain polymorphisms in the human genome relating to certain MTBC lineages in their respective geographic distribution (7–9). Interestingly, a study found that a particular variant in a gene responsible for autophagy in humans, IRGM-261T, influences the susceptibility of TB caused by MAF but not MTB (10). Additionally, findings from research on Ghanian populations showed that variants associated with 5-lipoxygenase (ALOX5) were associated with an increased TB risk (11). This suggests that MTBC might be adapted to certain populations just as they are geographically distributed (7).

In addition to host-pathogen compatibility of TB, environmental factors, lifestyle, living conditions and HIV co-infection play important roles in the outcome of the infection (12).

In Nigeria and West Africa at large, only three major lineages have been reported to cause tuberculosis: lineage 4, also known as the Euro-American lineage and lineages 5 and 6 also known as West Africa lineages 1 and 2. Lineage 4 comprises of 10 sub lineages: L4.1.1 (X); L4.1.2 (Haarlem); L4.1.3 (Ghana); L4.2; L4.3 (LAM); L4.4; L4.5; L4.6.1 (Uganda); L4.6.2 (Cameroon); and L4.10 (PGG3) (13). Lineage 4 is globally distributed, with the majority of TB infections in Nigeria being caused by the Cameroon sub lineage of lineage 4, followed by *M. africanum* (Lineages 5 and 6) (14).

With the advent of next generation sequencing and the continuous reduction in the cost to sequence the entire genome of an organism, studies have moved from the analysis of a single or few genomes to multiple or a collection of genomes. Pan-genome analysis is a product of the breakthrough of multi-genome study in molecular biology (15).

The pan-genome can be described as the collection of entire genes in a particular species. This comprises the core genes shared by all strains, dispensable genes shared by two or more strains, and unique genes also known as singletons that are peculiar to specific strains. This can be used in describing bacterial species as many species differ by their gene content to a large extent (16). The core genes are responsible for the major phenotype and basic biological processes of the bacteria, whilst the accessory and unique genes may be involved in other metabolic pathways such as adaptation to a particular host, virulence, antibiotic resistance and other functions that confer selective benefits over other species (15). Pan-genome analysis has been performed on more than 50 bacteria species in the past decade and this has revealed interesting information relating to pathogenesis, bacteria evolution, drug resistance, host specialization, horizontal gene transfer (HGT) (17, 18). Pan-genome analysis has also been adopted for viruses, fungi and plants (19, 20). Studies have also used pan-genome analysis for identifying potential vaccine candidates against bacterial infections (17).

The purpose of this study was to compare the entire gene set of *M. africanum* and the most commonly found TB sub lineage in Nigeria L4.6.2 (also known as the Cameroon sub lineage) and L2.2.1 (also known as the Beijing lineage) as an out-group to understand the evolution, genome dynamics, metabolic pathways and also genes involved in biological processes.

Methods

Sample collection and filtering:

We first retrieved fully assembled genomes of *M. africanum* West African 2 also known as lineage 6 (n = 22) out of 30 available genomes on the NCBI Reference Sequence Database (RefSeq), selected sequences were those that were not derived from surveillance projects or contain anomalies, but from a sequencing project in Mali between 2006 and 2010 (21). Also, we retrieved an additional 30 raw sequence datasets from the Senghore et al., (14) study, and selected 21 genomes after filtering and profiling them according to respective lineages using TB-Profiler version 2.3 (22). Thus, we selected only Euro-American lineage 4 sequences which were collected between 2011 and 2014. An out-group set of lineages was used for comparison by obtaining 334 genome assemblies from a project that was sequenced from South Africa between 2011 and 2014 (23). These selected genomes were classified into subtype lineages using Biohansel (24), and only genomes that belongs to the Beijing lineage 2 (n = 24) were selected.

All genome assemblies were generated from Illumina Miseq or Hiseq sequencing platforms and quality control on the assemblies was performed using Quast version 5.0.2 (25) to evaluate genome coverage, GC content, contig sizes and duplication ratio against the H37Rv (NC_000962.3) reference genome.

Bioinformatics analysis:

The genomes of the Euro-American lineages were assembled using a de novo approach with SPAdes version 3.11.1 (26). The scaffolds of the lineage 6, lineages 4 and 2 were annotated using Prokka version 1.12 (27). The annotated genomes were analysed with the Bacterial Pan Genome Analysis Tool (BPGA) version 1.3 (15), using 90% similarity for orthologous clustering and both pan- and core genomes were calculated over 30 permutations to prevent bias. Additionally, the pan-genome functional analysis was carried out utilizing KEGG, COG metabolic and functional pathways, which were all visualised with LibreOffice Calc plot functions. Panaroo (28) and PopPUNK (29) were also employed for pan-genome investigation of gene profiles, phylogenetics, clustering core and accessory genomes.

Results

Assessing completeness of assembled genomes was done with Quast by mapping them against the H37Rv reference genome. Lineage 6 genomes had ≥ 98.069% alignment, lineage 4 had ≥ 96.145% and lineage 2 had ≥ 96.856% alignment with the reference genome. A complete report on genome assembly statistics is shown in Supplementary table 1. In order to avoid sub-population bias, a transmission

analysis was performed by constructing a phylogenetic tree using core genomes rooted against the H37Rv reference genome in each lineages in this study as seen in Fig. 1.

Mycobacterium africanum (Lineage 6):

The identification of core genes (genes shared between all strains), accessory genes (genes shared by two or more strains but not all) and unique genes (genes peculiar to individual strains) were clustered respectively. The pan-genome analysis across 22 genomes of *M. africanum* showed that they have 3974 ± 13 genes (mean ± standard deviation). The number of accessory genes ranged from 115 to 159 genes and the unique set of genes varied from 1 to 54 genes (Supplementary table 2A). Empirical power law equations and exponential equations were used to generate pan- and core genome curves, which showed that the pan-genome curve has almost reached a plateau as the exponent parameter calculated is 0.03 (Fig. 2). This argues that the 22 genomes analysed are sufficient to obtain an accurate estimate of pan- and core genome size, with additional samples yielding diminishing returns.

Pan-genome functional analysis was done using KEGG pathway database and KEGG IDs assigned to orthologous protein clusters in the core, accessory and singleton genes and matched against the database, as shown in Fig. 3. The highest gene contents in the core genomes of *M. africanum* were responsible for biological processes such as metabolism, whilst a remarkable amount of unique genes account for environmental information processing and organismal systems. A more detailed KEGG classification showed that majority of unique gene sets were responsible for signalling molecules and interaction, signal transduction, infectious diseases, digestive system and cellular community (Fig. 4).

Mycobacterium tuberculosis **lineage 4 (Euro-American lineage)**:

The classification of core, accessory and unique genes in numbers after clustering into orthologous groups was performed. The 21 genomes from the Euro-American lineage had 4067 ± 18 genes (mean ± standard deviation). Accessory gene numbers ranged from 242 to 289 and singleton gene numbers were from 19 to 55 genes (Supplementary table 2B). Power law equations and exponential equations were used to generate pan and core genome curves which reflects the curve has almost reached a plateau as the exponent parameter calculated is 0.06 (Fig. 2).

A high percentage of genes in the core genome were linked to metabolic processes after functional classification of the pan-genome (Fig. 3). A more detailed KEGG classification of the pan genome showed that majority of the accessory genes are related to infectious diseases and cellular community (Fig. 4).

Mycobacterium tuberculosis **lineage 2 (Beijing lineage)**:

Orthologous clustering of the pan-genome showed that the analysis of 24 genomes have 4074 ± 29 genes (mean ± standard deviation). The number of accessory genes ranged from 216 to 262 genes, with 3825 genes shared between all strains (core genes) and unique gene sets varied from 1 to 121 genes per strain (Supplementary table 2C). Exponential regression was used to generate the number of core genes

whilst power law regression was used to extrapolate the pan-genome curve which reflects a slightly closed pan-genome with an exponent parameter 0.03 (Fig. 2).

KEGG classification of core, accessory and unique genes into functional roles by clustering genes into orthologous groups argued that the majority of core, accessory and unique genes are responsible for metabolic functions (Fig. 3) and a detailed classification of functional genes showed that a relative quantity of the core genes are associated with carbohydrate metabolism, whilst the accessory genes are linked with cellular community, amino acid metabolism and infectious diseases (Fig. 4).

Pan-genome comparative analysis

Following the clustering of respective TB datasets into core, accessory and unique genes, the core genes of lineage 6 were higher in number than core genes in lineages 4 and 2 but had fewer accessory genes than these lineages (Supplementary table 2). The red, orange and cyan lines in Fig. 2 (a, b and c) represent the power-fit curve derived from the equation ($f(x) = a.x^b$), where the exponent $b > 0$ implying that the genome is open, however the parameter b values are 0.03, 0.06 and 0.03 (Lineage 6, Lineage 4 and Lineage 2 respectively) meaning the pan genome is almost closed and addition of new genomes may not lead to the discovery of novel functions (30).

Using BPGA to cluster the core, accessory and unique genes and assigning KEGG functional pathways to them, about 92% of the core genes of the three lineages are responsible for metabolism related pathways. However, accessory gene assignment of metabolic, organismal system and environmental information related pathways in the lineages are: Lineage 6 (33%, 59% and 58%), Lineage 4 (23%, 28% and 33%) and Lineage 2 (47%, 18% and 21%) as shown in Fig. 4.

Evolution of M. africanum

Core genes are responsible for survival and majority of biological processes in the bacteria. Those that exist in copies, also known as “multi copy core genes” (MCG) were studied as these genes, especially rRNAs in bacteria, have been seen to influence adaptation to environmental pressures and the structure of microbiomes (31). We examined conserved genes of these lineages that exhibit copy number variation (CNV) and constructed a phylogeny based on their core genomes (Fig. 5). Eleven (11) PE/PPE family proteins and four (4) *fadD* genes fall under the category of MCG. PPE family proteins and *fadD* genes were selected as they have been reported to be crucial factors for mycobacterial virulence and *in vivo* pathogenicity (32–34).

PPE family proteins 15 and 26 all had similar gene copy number across all strains, which shows that these proteins have remained unchanged over time, whilst family PPE proteins 20, 47/48, 57, 51 and 42 showed little evolution in the CNV and lastly, PPE proteins 40, 32, 29 and 33 displayed lineages distinct CNV across the genomes (Fig. 5).

On the other set of gene families studied, *fadD13*, *fadD15*, *fadD25* and *fadD32* genes which are responsible for fatty acid CoA ligase, all showed CNV in the TB strains during the course of evolution except the *fadD32* gene (Fig. 5). Additionally, multiple sequence alignment of *fadD13* genes in MTBC genomes showed substitution mutations A8G in *fadD13_2*, S108F and A123V in *fadD13_3* (Fig. 6).

Discussion

Supporting the clonal nature of MTBC, our findings shows that the addition of new genomes to the analysis will not lead to the discovery of new phenotypes in *M. africanum* due to its close pan-genome. In addition, we also observed copy number variations of some core genes that may be related to the geographically restricted specialist (lineage 6) and globally distributed generalists (lineage 4 and lineage 6).

The reduction in number of core genes in lineages 4 and 2 compared to lineage 6 can be linked to increased virulence as previous work have shown that modern lineages (lineage 2 and 4) are more pathogenic than ancient lineages such as lineage 6. Genome reduction causes bacteria to have increased virulence compared to their counterparts with larger genomes (35, 36). Varying numbers of accessory and unique genes which could also be responsible for drug resistance, virulence or preference to a particular host which plays a key role in lineage classification, geographical distribution and distinct lifestyles of some TB strains as reported in earlier studies (7, 15). Furthermore, investigation into some MCG show copy number variation and amino acid substitution during the evolution of MTBC, this could also attribute to host preference and the geographical peculiarity of *M. africanum* West African 2.

In the detailed KEGG classification shown in Fig. 4, a large number of orthologous genes belonging to cellular community, digestive system, immune system, infectious diseases, signal transduction and signalling molecule interactions pathways were grouped under unique and accessory genes in lineage 6, whilst a significant reduction of orthologous genes in these same pathways were seen in lineages 2 and 4. We speculate that genetic loss in the dispensable genome of the modern lineages has a selective advantage for virulence and global distribution that allows the TB strain to persist in a wide range of host as reported in previous works ((13, 37). We also saw a reduction in copy number of some core genes in lineages 2 and 4, which was higher in lineage 6 such as PPE family protein 40, 51 and 29 (Fig. 5). This may also be related to the geographic distribution and host specificity as housekeeping genes are responsible for survival of pathogens (38).

One of the multiple copy core genes, *fadD13*, which codes for long chain fatty acid COA ligase in MTB and maintains appropriate mycolic acid composition and permeability of the envelope on its exposure to acidic pH (39, 40), was investigated for evolution in the lineages. The *fadD13* gene is encoded by operon *MymA* and is essential for virulence and *in vivo* progression of MTBC (41). Multiple alignments of gene copies of *fadD13* (Fig. 6) in the TB genomes suggests that these lineage-specific mutations could have also shaped the distinctive features of the modern and ancient lineages in this study, as numerous

fadD13 protein variants have been seen to influence ATP binding (41). This protein also had more copies in lineage 2 and 4 than found in lineage 6 (Fig. 6).

Conclusion

Pan-genome analysis of *M. africanum* West African 2 in this work showed that a higher number of orthologous genes in the dispensable genome may have contributed to the restriction of global distribution and reduced virulence compared to modern lineages 2 and 4. Also, substitutions in some multiple copy core genes and copy number variations might have influenced evolution of *M. africanum* West African 2 and its geographic distribution.

Inasmuch as it is hard to say if lineage 6 migrated out of West Africa and got outcompeted by modern lineages, it is almost certain that it might soon be outcompeted by modern virulent strains in West Africa due to its close pan genome.

Declarations

Availability of data and materials

The datasets used and/or analysed during the current study are available on NCBI under the following BioProjects: PRJNA211726, PRJEB15857 and PRJNA476470.

Acknowledgements

We would like to show appreciation to our colleagues at the African Centre of Excellence for Genomics of Infectious Diseases and the facility made available at the research centre which continues to provide an enabling environment for ground breaking research.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by grants from the National Institute of Allergy and Infectious Diseases. NIH-H3Africa (U01HG007480 and U54HG007480 to Redeemer's University [Dr. Happi]), and a grant from the World Bank grant (project ACE019 to Redeemer's University [Dr. Happi]).

References

1. World Health Organisation. Global Tuberculosis Report 2019 [Internet]. 2019 [cited 2019 Oct 18]. Available from: https://www.who.int/tb/data/en/.

2. Cristina M, Brisse G, Brosch S, Fabre R, Omaïs M, Marmiesse B. M, et al. Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. PLoS Pathog. 2005;1(1):0055–61.
3. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, et al. Mycobacterial lineages causing pulmonary and extrapulmonary Tuberculosis, Ethiopia. Emerg Infect Dis. 2013 Mar;19(3):460–3.
4. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A. 2002 Mar;19(6):3684–9. 99(.
5. Smith I. Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. Clin Microbiol Rev. 2003;Vol. 16:463–96.
6. Comas I, Gagneux S. The past and future of tuberculosis research. PLoS Pathog [Internet]. 2009 Oct [cited 2019 Sep 13];5(10):e1000600. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19855821.
7. S A [Internet]
Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A [Internet]. 2006 Feb 21 [cited 2019 Apr 25];103(8):2869–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16477032.
8. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS Biol [Internet]. 2008 Dec 16 [cited 2019 Sep 13];6(12):e311. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19090620.
9. Reed MB, Pichler VK, McIntosh F, Mattia A, Fallow A, Masala S, et al. Major Mycobacterium tuberculosis lineages associate with patient country of origin. J Clin Microbiol [Internet]. 2009 Apr [cited 2019 Sep 13];47(4):1119–28. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19213699.
10. Intemann CD, Thye T, Niemann S, Browne ENL, Amanua Chinbuah M, Enimil A, et al. Autophagy gene variant IRGM – 261T contributes to protection from tuberculosis caused by Mycobacterium tuberculosis but not by M. africanum strains. PLoS Pathog [Internet]. 2009 Sep [cited 2019 Nov 5];5(9):e1000577. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19750224.
11. Herb F, Thye T, Niemann S, Browne ENL, Chinbuah MA, Gyapong J, et al. ALOX5 variants associated with susceptibility to human pulmonary tuberculosis. Hum Mol Genet. 2008 Apr 1;17(7):1052–60.
12. Lönnroth K, Jaramillo E, Williams BG, Dye C, Raviglione M. Drivers of tuberculosis epidemics: The role of risk factors and social determinants. Soc Sci Med. 2009 Jun;68(12):2240–6.
13. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet. 2016 Dec 1;48(12):1535–43.
14. Senghore M, Otu J, Witney A, Gehre F, Doughty EL, Kay GL, et al. Whole-genome sequencing illuminates the evolution and spread of multidrug-resistant tuberculosis in Southwest Nigeria. PLoS

One. 2017.

15. Chaudhari NM, Gupta VK, Dutta C. BPGA-an ultra-fast pan-genome analysis pipeline. Sci Rep. 2016;6.
16. S A [Internet]
    Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A [Internet]. 2005 Sep 27 [cited 2019 Sep 14];102(39):13950–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16172379.
17. Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, et al. Identification of a universal Group B streptococcus vaccine by multiple genome screen. Science [Internet]. 2005 Jul 1 [cited 2019 Sep 14];309(5731):148–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15994562.
18. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol [Internet]. 2015 Feb [cited 2019 Apr 11];23:148–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25483351.
19. Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G. Analysis of the Saccharomyces cerevisiae pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. Genome Res [Internet]. 2012 May [cited 2019 Sep 14];22(5):908–24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22369888.
20. Muzzi A, Masignani V, Rappuoli R. The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. Drug Discovery Today. 2007;Vol. 12:429–39.
21. Winglee K, Manson McGuire A, Maiga M, Abeel T, Shea T, Desjardins CA, et al. Whole Genome Sequencing of Mycobacterium africanum Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction. PLoS Negl Trop Dis. 2016 Jan 11;10(1).
22. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015 Dec 14;7(1).
23. Shah NS, Auld SC, Brust JCM, Mathema B, Ismail N, Moodley P, et al. Transmission of extensively drug-resistant tuberculosis in South Africa. N Engl J Med. 2017 Jan;19(3):243–53. 376(.
24. Labbé G, Kruczkiewicz P, Mabon P, Robertson J, Schonfeld J, Kein D, et al. Rapid and accurate SNP genotyping of clonal bacterial pathogens with BioHansel. bioRxiv. 2020 Jan 11;2020.01.10.902056.
25. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142–50.
26. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May 1;19(5):455–77.
27. 10.1093/bioinformatics/btu153
    Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics [Internet]. 2014 Jul 15 [cited 2019 Apr 15];30(14):2068–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu153.

28. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline. bioRxiv. 2020 Jan 28;2020.01.28.922989.
29. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019 Feb;29(2)(1):304–16.
30. Bosi E, Fani R, Fondi M. Defining orthologs and pangenome size metrics. Methods Mol Biol [Internet]. 2015 [cited 2019 Sep 19];1231:191–202. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25343867.
31. Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol. 2000 Apr;66(4):1328–33.
32. Ates LS. New insights into the mycobacterial PE and PPE proteins provide a framework for future research. Mol Microbiol [Internet]. 2019 Nov 24 [cited 2020 Jan 10];mmi.14409. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/mmi.14409.
33. 10.1086/651452
Dunphy KY, Senaratne RH, Masuzawa M, Kendall LV, Riley LW. Attenuation of Mycobacterium tuberculosis Functionally Disrupted in a Fatty Acyl–Coenzyme A Synthetase Gene fadD5. J Infect Dis [Internet]. 2010 Apr 15 [cited 2020 Jan 10];201(8):1232–9. Available from: https://academic.oup.com/jid/article-lookup/doi/10.1086/651452.
34. Rindi L, Fattorini L, Bonanni D, Iona E, Freer G, Tan D, et al. Involvement of the fadD33 gene in the growth of Mycobacterium tuberculosis in the liver of BALB/c mice. Vol. 148, Microbiology. Society for General Microbiology; 2002. p. 3873–80.
35. Ribeiro SCM, Gomes LL, Amaral EP, Andrade MRM, Almeida FM, Rezende AL, et al. Mycobacterium tuberculosis strains of the modern sublineage of the beijing family are more likely to display increased virulence than strains of the ancient sublineage. J Clin Microbiol. 2014;52(7):2615–24.
36. Weinert LA, Welch JJ. Why Might Bacterial Pathogens Have Small Genomes? Vol. 32, Trends in Ecology and Evolution. Elsevier Ltd; 2017. p. 936–47.
37. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nat Genet. 2015 Mar 1;47(3):242–9.
38. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Vol. 15, Current Opinion in Genetics and Development. 2005. p. 589–94.
39. Cheruvu M, Plikaytis BB, Shinnick TM. The acid-induced operon Rv3083-Rv3089 is required for growth of Mycobacterium tuberculosis in macrophages. Tuberculosis (Edinb) [Internet]. 2007 Jan [cited 2019 Sep 23];87(1):12–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16893682.
40. Khare G, Gupta V, Gupta RK, Gupta R, Bhat R, Tyagi AK. Dissecting the role of critical residues and substrate preference of a Fatty Acyl-CoA Synthetase (FadD13) of Mycobacterium tuberculosis. PLoS One [Internet]. 2009 Dec 21 [cited 2019 Sep 23];4(12):e8387. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20027301.

41. Andersson CS, Lundgren CAK, Magnúsdóttir A, Ge C, Wieslander Å, Molina DM, et al. The Mycobacterium tuberculosis very-long-chain fatty acyl-CoA synthetase: Structural basis for housing lipid substrates longer than the enzyme. Structure. 2012 Jun;6(6):1062–70. 20(.

Figures

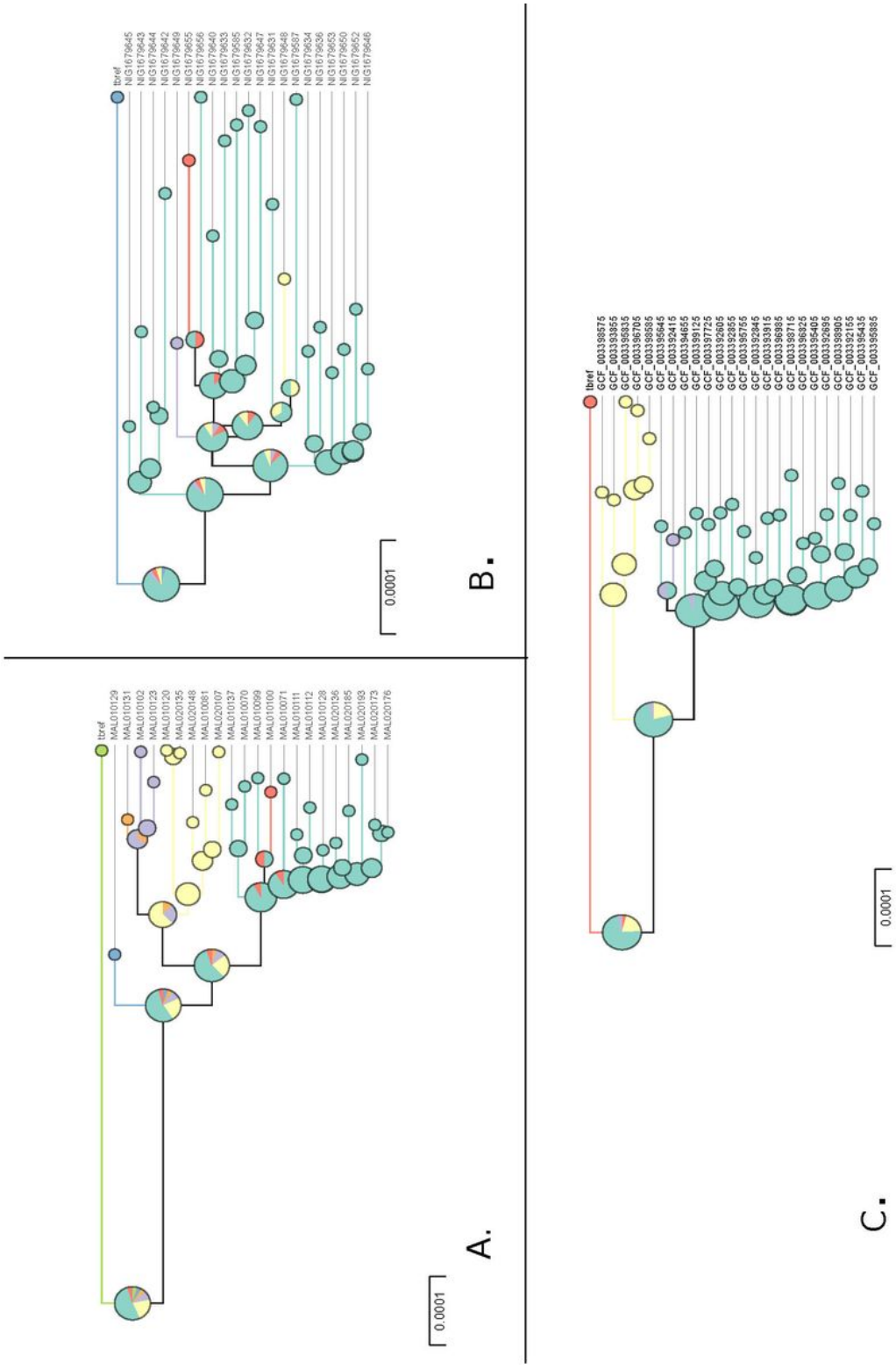

Figure 1

Transmission analysis of Mycobacterium tuberculosis genomes represented in this study. (A) Lineage 6: M. africanum West African 2 (B) Lineage 4: Euro-American lineage (C) Lineage 2: Beijing lineage.

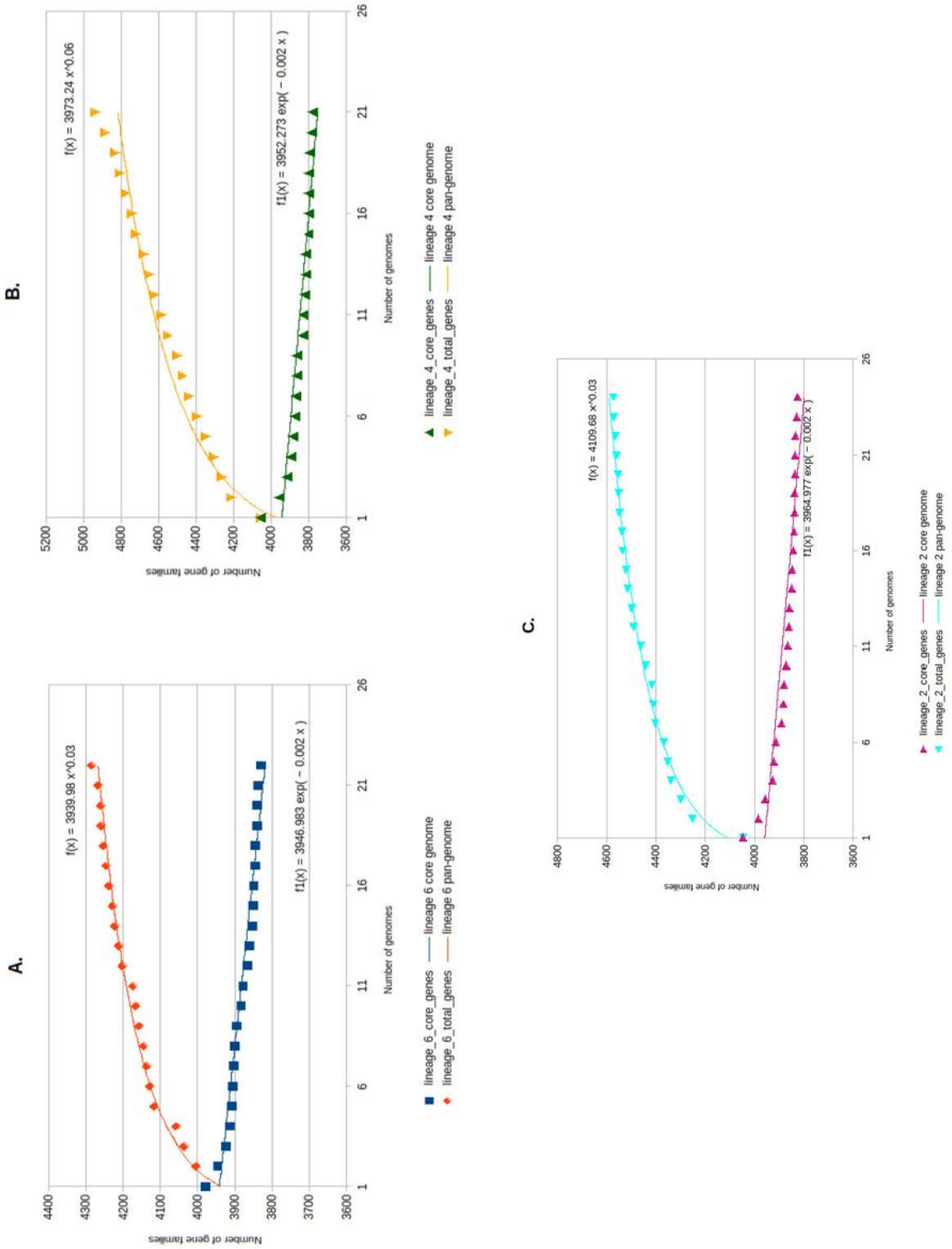

Figure 2

The pan- and core genome of (A) Lineage 6: M. africanum West Africa 2, (B) lineage 4: Euro-American lineage and (C) lineage 2: Beijing lineage profile plot with power-fit curve equation shown as f and exponential curve equation as f1.

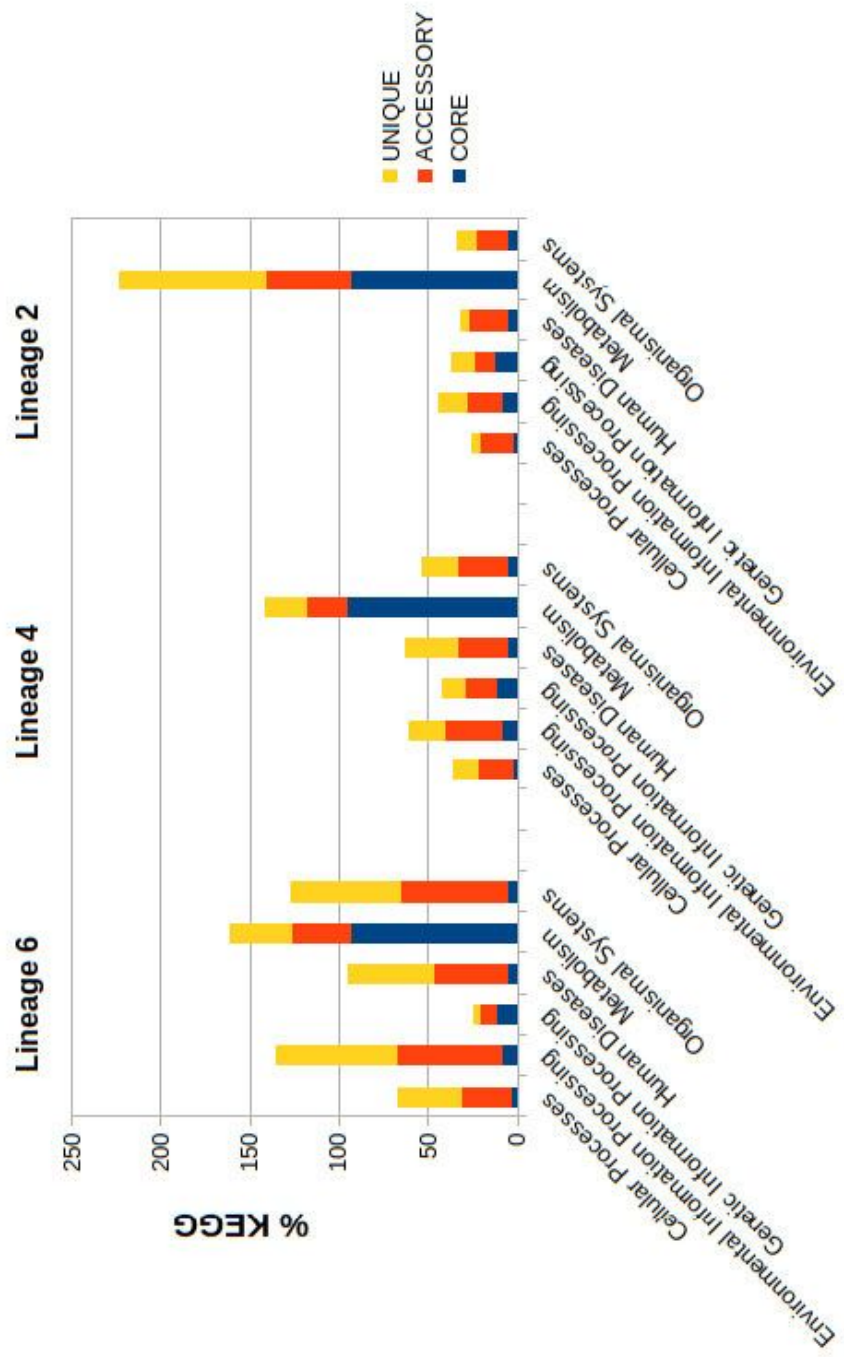

Figure 3

KEGG distribution of core, accessory and unique genes for the MTBC genomes

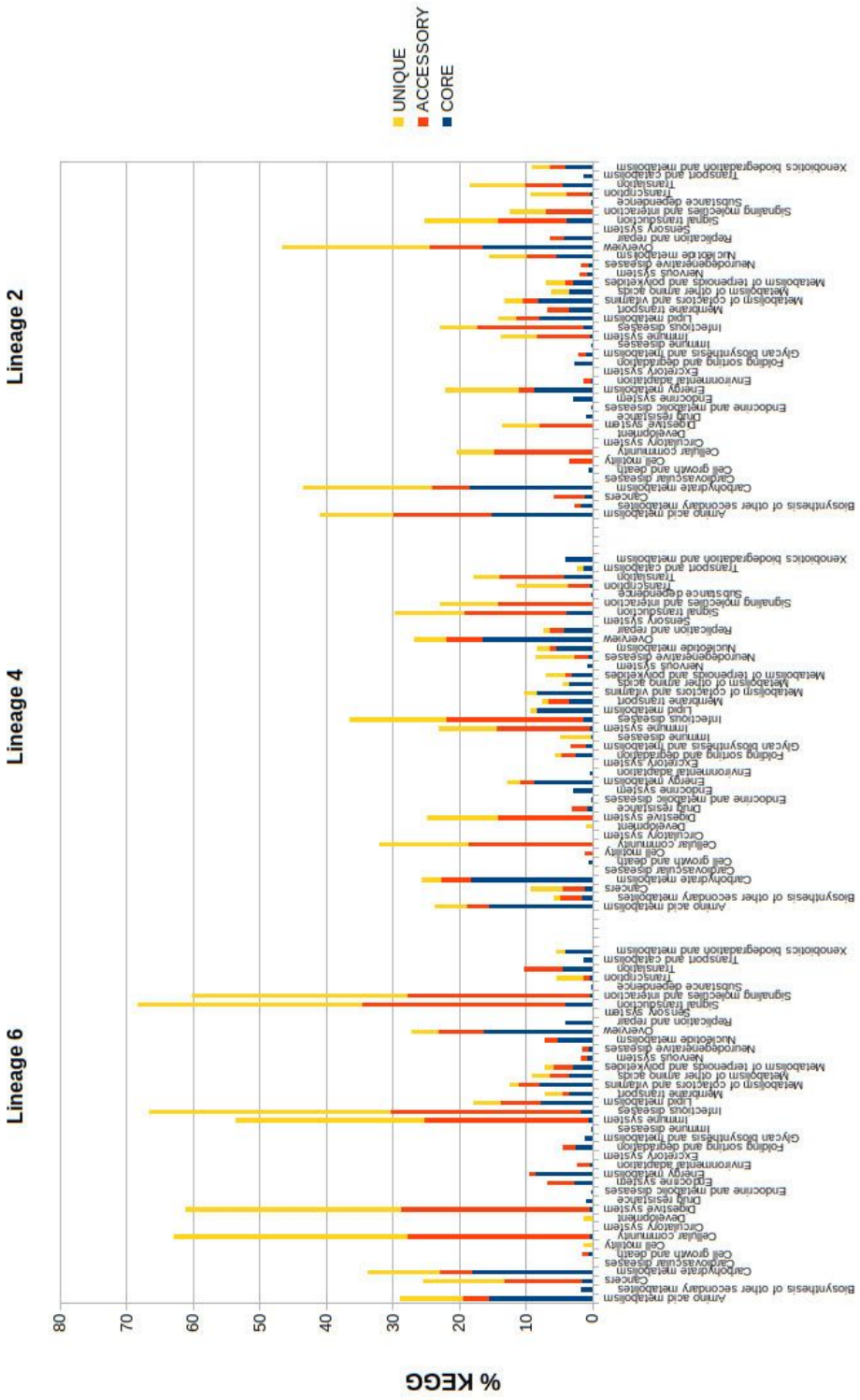

Figure 4

Detailed KEGG classification of core, accessory and unique genes for the MTBC genomes

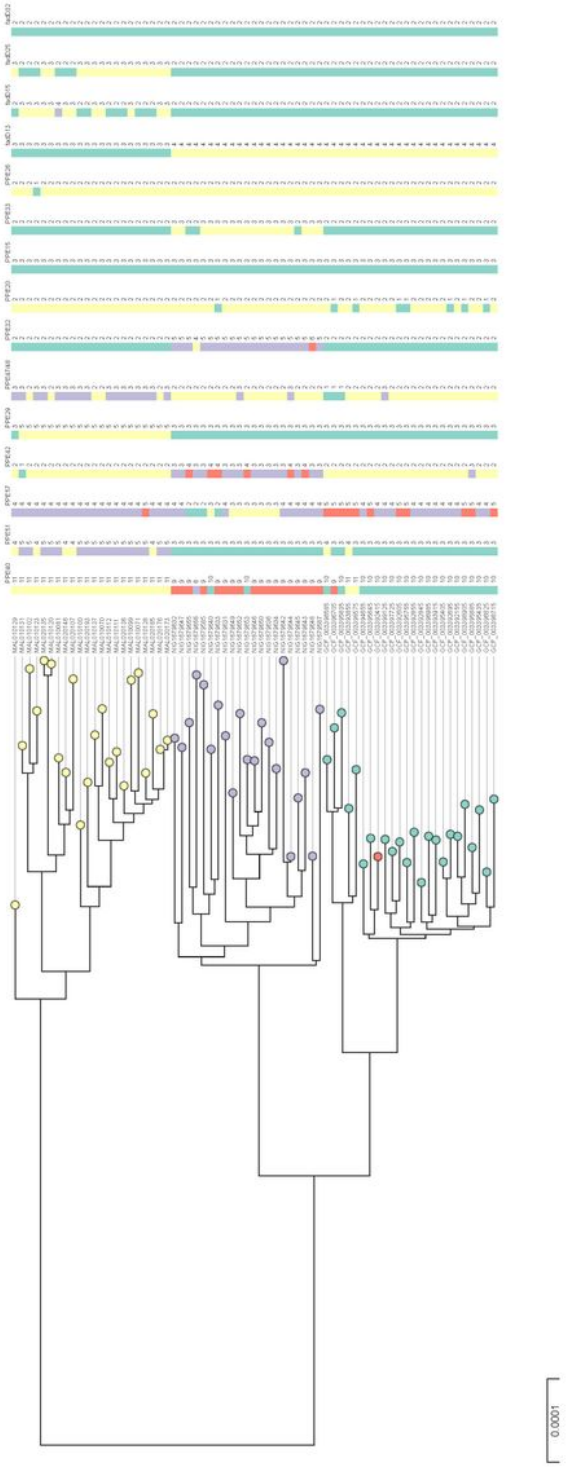

Figure 5

Core genome phylogeny with tree nodes coloured according to clusters generated by PopPUNK. Yellow (lineage 6), purple (lineage 4), blue and red (lineage 2). Columns display multi-copy core PPE and fadD genes. Genes are coloured according to copy numbers.

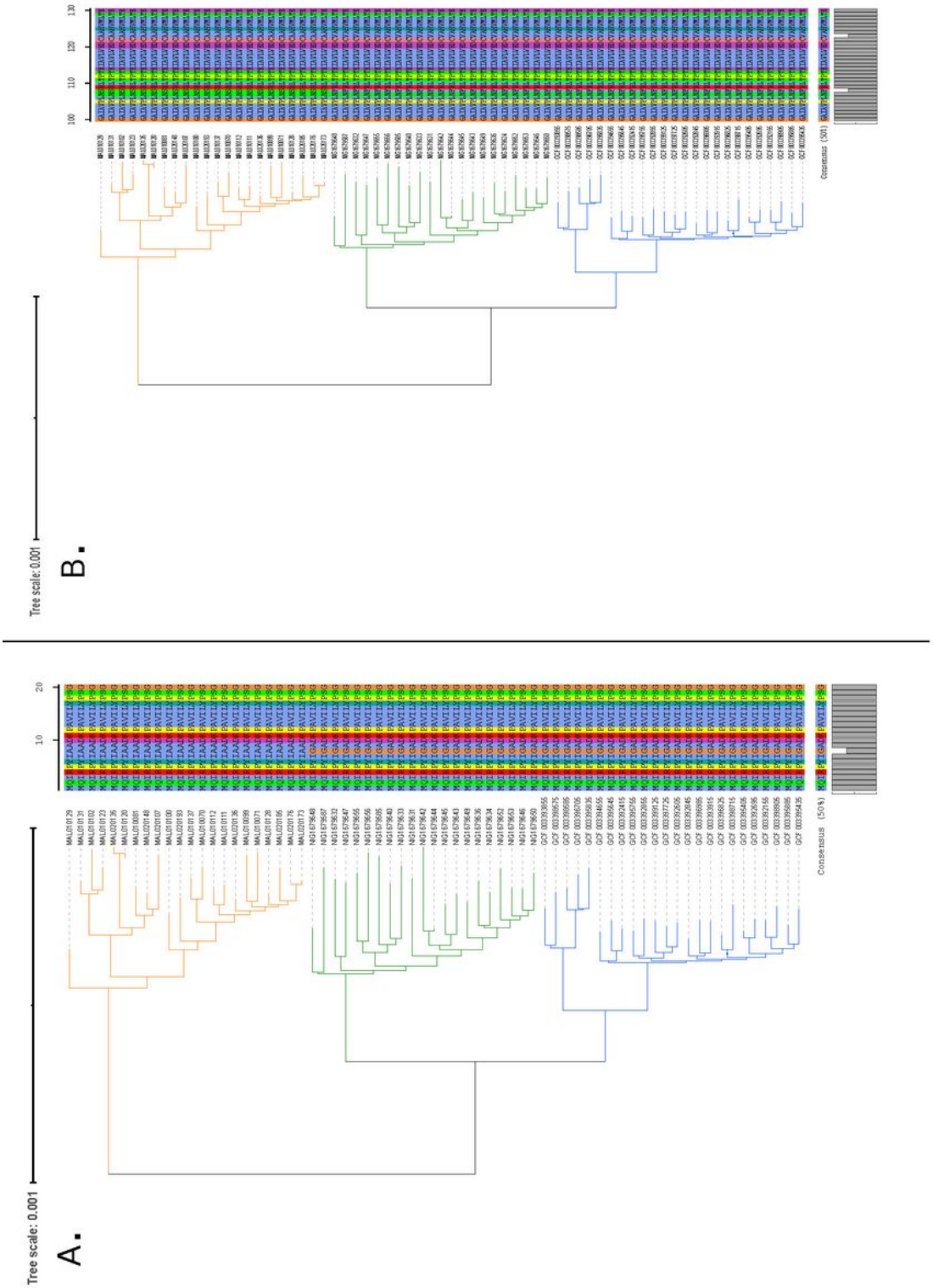

Figure 6

Core genome phylogeny and multiple sequence alignment of genes (A) fadD13_2 showing amino acid substitutions A8G, (B) S108F and A123V in fadD13_3 within lineage 6 (orange), lineage 4 (green) and lineage 2 (blue).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytable2.xlsx
- Supplementarytable1.xlsx