# skalo: using SKA split k-mers with coloured de Brujin graphs to genotype indels

Romain Derelle[1], Kieran Madon[1], Nimalan Arinaminpathy[2], Ajit Lalvani[1], Simon R. Harris[3], John A. Lees[4], Leonid Chindelevitch[2]*

[1] NIHR Health Protection Research Unit in Respiratory Infections, National Heart and Lung Institute, Imperial College London, London, W21PG, UK
[2] MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, W12 0BZ, UK
[3] Bill and Melinda Gates Foundation, 62 Buckingham Gate, Westminster, London, SW1E 6AJ
[4] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK
*Corresponding author: l.chindelevitch@imperial.ac.uk

## Abstract
Insertions and deletions (indels) are important contributors to the genetic diversity and evolution of pathogens like *Mycobacterium tuberculosis*. However, accurately identifying them from genomic data remains challenging using current variant calling methods. We present skalo, a graph-based algorithm that complements the popular split k-mer approach implemented in the SKA software. skalo is designed for alignment-free inferences of indels between closely related haploid genomes, which are ignored by SKA. The graph traversal implemented in skalo enables rapid detection of indels and complex variants, while retaining the speed and alignment-free advantages of SKA. Through benchmarking on simulated and real *Mycobacterium tuberculosis* data, we demonstrated its ability to identify indels and complex variants with high precision, and explored their utility as phylogenetic markers to resolve isolates' relationships. By providing an efficient and easy-to-use method to extract additional variants from genomic data, skalo can enhance our understanding of pathogen evolution and transmission, with potential applications across diverse pathogen species. skalo is written in Rust and is freely available at https://github.com/rderelle/skalo.

## Introduction
The study of genomic variations allows researchers to decipher the evolutionary history of pathogens, understand their adaptive mechanisms, and trace the dynamics of infectious disease transmission (Armstrong et al. 2019). Among these genomic variations, insertions and deletions of nucleotides (indels) are found less frequently than single nucleotide polymorphisms (SNPs), occurring at a rate of approximately one indel for every ten SNPs (Sung et al. 2016), but are nonetheless important contributors to the genetic diversity and evolution of pathogens. For instance, indels in regulatory regions can lead to phase-variable expression of genes (van der Woude and Bäumler 2004; Vargas et al. 2023). These play a particularly important role in the evolution of some clonal organisms, such as *Mycobacterium tuberculosis* (Mtb), that are unable to acquire genes by horizontal gene transfer (Gagneux 2018). Large deletions are also common in Mtb genomes and have been used to classify Mtb strains (Jagielski et al. 2014). Comparative studies of Mtb genomes have recently revealed the important role of indels in the acquisition of antimicrobial resistance (Godfroid et al. 2020; Gomes et al. 2021; The, CRyPTIC Consortium 2022) and in reversible

gene silencing, which provides adaptive flexibility to this slowly evolving species (Gupta and Alland 2021). Finally, indels have also been proposed as phylogenetic characters to identify past tuberculosis transmission events (Godfroid et al. 2020; Sobkowiak and Colijn 2022).

The identification of genomic variants from whole-genome sequencing data is commonly done through the alignment of sequencing reads to a reference genome. The selection of programs for read-alignment and variant-calling, and the choice of reference genome are all pivotal elements in a variant-calling pipelines (Bush et al. 2020; Bush 2021; Li 2014; Valiente-Mullor et al. 2021). In order to minimise false positives, variants are filtered using various criteria and by ignoring variants from specific genomic regions, which commonly include repeats. In Mtb genomes excluded regions usually include genes belonging to the PE-PPE family. However, no filtering method is immune to errors, and downstream analyses using variants from such approaches can be strongly biased by the choice of pipeline (van der Putten et al. 2022; Walter et al. 2020). A particular challenge for many alignment algorithms are short indels, which often occur in low complexity regions prone to read misalignments, resulting in false positive indel calls (Fang et al. 2014; Narzisi and Schatz 2015). Consequently, the accuracy of short indel calling is generally lower than that of SNP detection (Poplin et al. 2018; Su et al. 2023), and usually requires read realignment and additional filtering steps to improve performance (Fang et al. 2016; DePristo et al. 2011; Gupta and Alland 2021).

Alignment-free approaches attempt to circumvent issues caused by misalignments by reducing the sequencing data of isolates into sets of k-mers, which are then compared directly without aligning to a reference genome (Vinga and Almeida 2003; Zielezinski et al. 2017). Among these methods, the split k-mer approach has been developed to infer SNPs from closely related haploid genomes (typically pathogen outbreak data) by further dividing each k-mer into left and right k-mers flanking a variable middle base. In this framework, the two flanking k-mers serve as pseudo-references for SNP calling by comparing the middle bases found in different isolates. Implemented in the programs kSNP (Gardner et al. 2015; Gardner and Hall 2013; Hall and Nisbet 2023) and SKA (Derelle et al. 2024; Harris 2018), the split k-mer approach demonstrates high accuracy in SNP calling and ultra-fast computational speeds. However, the fixed size of the middle base currently limits its ability to infer indels.

Here we introduce skalo, a graph-based algorithm designed to infer indels in a population. skalo uses the multi-samples split k-mer files generated by SKA as input, so inherits the computational efficiency of the SKA algorithm which generates these. skalo converts non-ubiquitous split k-mers (i.e., absent in at least one sample) into a local colored de Bruijn graph, as introduced by cortex (Iqbal et al. 2012), and leverages the sample distributions of k-mers to traverse the graph and identify variants. We demonstrate that skalo can rapidly and accurately identify indel variants from outbreak datasets, increasing the precision of genetic distance estimation, and thereby complementing the split k-mer approach. First, we benchmarked skalo's accuracy using simulated insertions, followed by testing it on a large Mtb dataset. Finally, we explored the utility of these indels in phylogenetic reconstruction of an Mtb outbreak.

## Methods

Algorithm description

skalo is a single core program written in Rust and is freely available at https://github.com/rderelle/skalo. Its name stands for 'SKA leftovers'. skalo takes as input a split k-mer file generated by SKA ('skf file') from at least two samples, which contains all split k-mers of size 'k'. These can be generated from assemblies or read data. All analyses were performed using skalo v0.3.1.

_ generating a coloured De Bruijn graph from skf file

The program first identifies non-ubiquitous split k-mers, as they represent signatures of variants. Focusing on these reduces the size of the De Bruijn graph, improving the program's speed and memory footprint. Non-ubiquitous split k-mers are converted into k-mers of size k by merging the middle-bases with their flanking sequences, which are then split into two overlapping k-mers of size k-1, considering both forward and reverse-complement sequences. The combination of all overlapping k-mers constitutes a coloured De Bruijn graph in which nodes are k-mers of size k-1 and edges contain the sample sets in which the two consecutive k-mers were found.

The graph is stored in two nested hash maps (dictionaries), with k-mers converted to 128-bits integers as dictionary entries. Using this format, skalo can handle all ranges of SKA split k-mer sizes (i.e., up to 63 nucleotides). Unique 32-bits integer identifiers are generated for each combination of samples (i.e. patterns), allowing a maximum of $4.3 \times 10^9$ sample patterns, and saved as values of the nested dictionary. skalo automatically stops if the maximum number of patterns is reached to avoid collision (the maximum number of patterns observed in this study was $1.6 \times 10^5$ for the analysis of 224 Mtb samples).

_ graph traversal and variant identification

As in other graph-based variant identification tools, the first step of the graph traversal consists of identifying entry nodes of the graph bubbles (Iqbal et al. 2012; Uricaru et al. 2015; Narzisi et al. 2014). These nodes are defined here as nodes leading to two or more paths whose edges contain non-identical sample sets. Since the graph contains both forward and reverse-complement k-mers, the set of nodes exiting the graph bubbles corresponds to the reverse-complement of the set of entry nodes.

The graph traversal takes advantage of the sample combinations stored in each edge to resolve graph bifurcations by selecting connected nodes having the most similar samples sets (Figure 1): for a given bubble path, at each node, the subsequent edge with the highest Jaccard similarity between its sample set and the sample set of the first edge is chosen. The traversal stops if no subsequent node is available, several subsequent nodes are available or the last visited node was an exit node.

A variant group is then identified when at least two paths start from the same entry node and end at the same exit node, with sample sets of variants defined as the majority-rule set of visited samples along the corresponding paths. This simple analytic process leads to classical graph-based variants mimicking the SKA split k-mer structure, in which the 'middle-base', represented by an inserted sequence ('indel') or more complex patterns (referred to as 'complex'), is surrounded by two k-mers of size k-1. In practice, the right k-mer is often found to have a length below k-1 depending on the sequences flanking the middle-base (examples provided in Supplementary figure S1). The term 'complex' refers here to any

sequence pattern that cannot be explained by a single SNP or indel (e.g., multiple SNPs, combination of SNP and indel, micro-inversion).

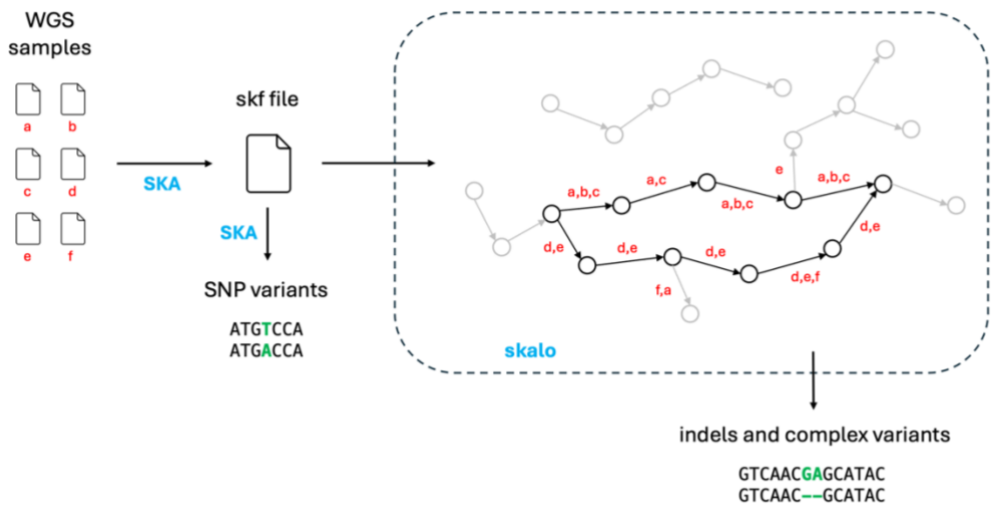


**Figure 1**: Schematic representation of skalo analysis of whole genome sequencing (WGS) data. Letters 'a', 'b', …, 'f' represent sample names. The two paths corresponding to the variant are highlighted in black, with sample sets indicated on each edge. In this dummy example, two alternative paths are not visited due to their divergent sample sets ('e' and 'f,a').

_ variant groups filtering and output

skalo first discards variant groups corresponding to SNPs. Then, two filters are applied to discard low-quality variant groups. First, the proportion of missing data of each variant group is calculated as the proportion of samples with either no variant or several variants (equivalent to a degenerate nucleotide in alignment-based SNP data). Variant groups having a proportion of missing data above a user-defined threshold 'm' are discarded (m=0.2 by default). Second, an optional parameter, turned off by default, discards indels inferred in homopolymers of size 'n' or more. These variants are known to be prone to homoplasy and therefore unsuitable for use as characters for phylogenetic analyses. The two output parameters 'm' and 'n' constitute the only additional parameters used by skalo. Finally, skalo attempts to realign the right k-mer among variants of a given variant group. If an alternative position for the right k-mer is found, the variant group is marked as 'unclear insert' in the output, as skalo cannot confidently infer the middle-base (example provided in Supplementary figure S2).

skalo generates three output files: the sequences of variant groups in FASTA format to be used for sequence visualisation or similarity searches, a binary alignment in FASTA format representing the distribution of minor and major variants among samples to be used for phylogenetic analyses, and a TSV file describing variant groups. More information about these output files is available on the skalo GitHub page.

_ limitations of this algorithm

The main limitation of skalo is similar, yet not fully identical, to that of SKA split-k-mer analyses: its sensitivity decreases with the density of variants (Harris 2018; Derelle et al. 2024). As opposed to SKA, if two consecutive variants appeared 'simultaneously' (i.e., on the same branch of the corresponding tree, both having identical sample sets), skalo will detect them in the form of a complex variant. However, similarly to SKA, in the case of two consecutive variants appearing 'successively' (i.e., on two different branches of the corresponding tree, having different sample sets) both variants and the combined complex variant will be detected by skalo but one, both or all three could be discarded depending on their level of missing samples (all three variants are never inferred together unless the proportion of missing data is set close to 1).

Second, while skalo's graph-based approach allows the detection of variants of in principle unlimited length, the detection of variants located in repeats will be limited by the k-mer size which must be larger than the repeat being mapped. The same limitation applies to tandem duplications and local inversions.

Simulated pairs of Mtb samples

Pairs of samples were created using genome assemblies listed in Supplementary data. Each pair was composed of the original root genome and the same genome mutated with 100 insertions of random sequences (random lengths between 1 and 10 nucleotides or fixed lengths for long insertions) at random genomic positions, in 10 replicates. Please note that inserting or deleting sequences will produce similar results using skalo, as it operates in a reference-free mode. The following constraints were imposed on insertions: (i) to not be located within 100bp from another insertion so that successive indels are not merged into complex variants and (ii), for insertions of more than 1 nucleotide, the nucleotides corresponding to the extremities of the inserted sequences should be different to the nucleotides located on the opposite flanking insertion site. The latter constraint allows a unique alignment of these inserted sequences, which greatly facilitates their identification from read-alignment and skalo outputs, making precise benchmarking possible without significantly differing from true molecular processes of mutation. True positive indels were identified using the two following criteria: the insertion should have an exact match to one of the inserted sequences (forward or reverse-complement) and the original genome should display the alternative variant (i.e., no insertion at a position not treated as missing data).

Read-alignment pipeline

We choose the popular pipeline Snippy v4.6.0 (https://github.com/tseemann/snippy) for its speed and its ability to infer variants directly from genomic assemblies by generating pseudo high-quality reads. Snippy was run with its default parameters, considering as output the 'filtered' set of variants. The set of genomic positions for Mtb genome masking available on Snippy Github page did not have any noticeable impact in reducing the number of false positives indels. Instead, we ignored variants located at genomic positions corresponding to any PE-PPE genes and repeat regions as defined in Supplementary data (totalling 8.5% of the H37Rv genome).

Partitioning of Mtb real-world datasets

Fastq files of the two Mtb datasets were analysed using fastlin v0.2.1 (Derelle et al. 2023) with default parameters and subsequently partitioned as follows: mixed samples and

samples having a k-mer coverage below 30 or above 500 were discarded, and the remaining samples were partitioned into MTBC lineages having at least 10 samples. One lineage was found to be highly distant from the others, generating about 12 thousand SNPs, so was subsequently discarded, resulting in 37 final lineage partitions (lists of samples available in Supplementary data).

Variant inferences and phylogenetic analyses

Split k-mer files were generated using SKA v0.3.1 (Derelle et al. 2024), and SNP alignments were obtained using the command 'ska align --min-freq 0.8 –no-gap-only-site --filter-ambig-as-missing --filter no-ambig-or-const'. skalo was used with default parameters for the simulated benchmarks and the analysis of the Birmingham and Valencia datasets, and with n=6 and m=0.3 for the analysis of the large transmission cluster. Phylogenetic analyses were performed using IQtree2 v2.2.2.7 (Minh et al. 2020) under the GTR model with the option '—polytomy', which collapses branches with near-zero lengths into polytomies. For the phylogenetic analysis of SNPs combined with skalo variants, we define two joint partitions of characters: SNP alignment analysed under the GTR model and skalo binary alignment output analysed under the GTR2 model.

Homoplasy levels

The inference of homoplasy and homology of informative variants was then implemented using a custom Python script using the ete3 library (Huerta-Cepas et al. 2016). Briefly, the script roots the ML tree using the midpoint approach, identifies the node corresponding to the most recent common ancestor of samples sharing the minor variant and then counts the number of leaves within this clade. Homology is inferred if the number of leaves corresponds to the number of samples containing the minor variant. If homology is not found, the script repeats the analysis using the tree rooted on one of the samples having the major variant. These analyses were limited to informative characters for which the ancestral node in the ML tree was bifurcated, as no conclusions could be drawn in cases of ancestral polytomies.

Genomic location of variants

Variant groups were mapped to the H37Rv genome (accession ID: NC_000962.3) using Blastn v2.13.0 with the following parameters: '-evalue 0.00001 -dust no'. For each variant group, the genomic position was inferred as the position corresponding to the end of the left k-mer in the top blast hit. Permutation analyses were performed using a custom python script. Briefly, the script randomly assigned variants to genomic compartments in 10,000 replicates. The p-values were then calculated as the fractions of replicates in which the number of variants in genomic compartments were equal or superior to the observed values (all values are available in Supplementary data).

Efficiency measurements

Runtimes and maximum memory consumption were recorded on a MacBook Air 2022 (M2 processor; 8 GB of RAM memory).

Data availability

Sequencing data of the two large Mtb datasets were downloaded from NCBI (Bioproject accessions PRJEB25991 and PRJEB38719). Python scripts for the simulation of insertions and

to compute variant homoplasy levels, as well as all output files generated from the analysis of the two large Mtb datasets and the analysis of the transmission cluster, are available for download at Zenodo (https://zenodo.org/records/13731969).

## Results

Detection of simulated insertions

We first evaluated the performances of skalo compared to Snippy, using pairs of Mtb genomes differing by 100 random synthetic insertions. The pairs were generated from genome assemblies representing lineages with increasing phylogenetic distances from the reference genome H37Rv, spanning from lineage 4.9 (strain H37Rv) to lineage 1 (sequence divergence below 0.001 SNPs per site).

The detection of short insertions by skalo was found to be highly dependent on the SKA k-mer size: skalo missed on average 6 insertions at k=21, and reached a baseline of about 2 missed insertions at k=41, a sensitivity level similar to that observed with Snippy (Figure 2A). As for the identification of SNPs in split k-mer analyses, longer k-mers increase the chances to identify genomic regions that are unique. Importantly, skalo did not produce any false positive (FP) indels in any of these analyses, as opposed to Snippy that generated numbers of FPs increasing with the phylogenetic distance to the reference genome. Excluding variants located in PE-PPE genes and repeat regions ('masked' analyses) reduced the numbers of FPs by half, but also increased the numbers of missed insertions.

Accuracy to detect large indels, ranging from 100bp to 5kb, were only assessed for skalo as these variants cannot be identified by standard read-alignment pipelines for which variant detection is limited by the read length. skalo was able to detect the longest insertions of 5kb, and displayed similar performances to those observed with short insertions: the numbers of missed insertions decreased with the k-mer size (Figure 2B) and no FP was detected in any of these analyses (not shown).

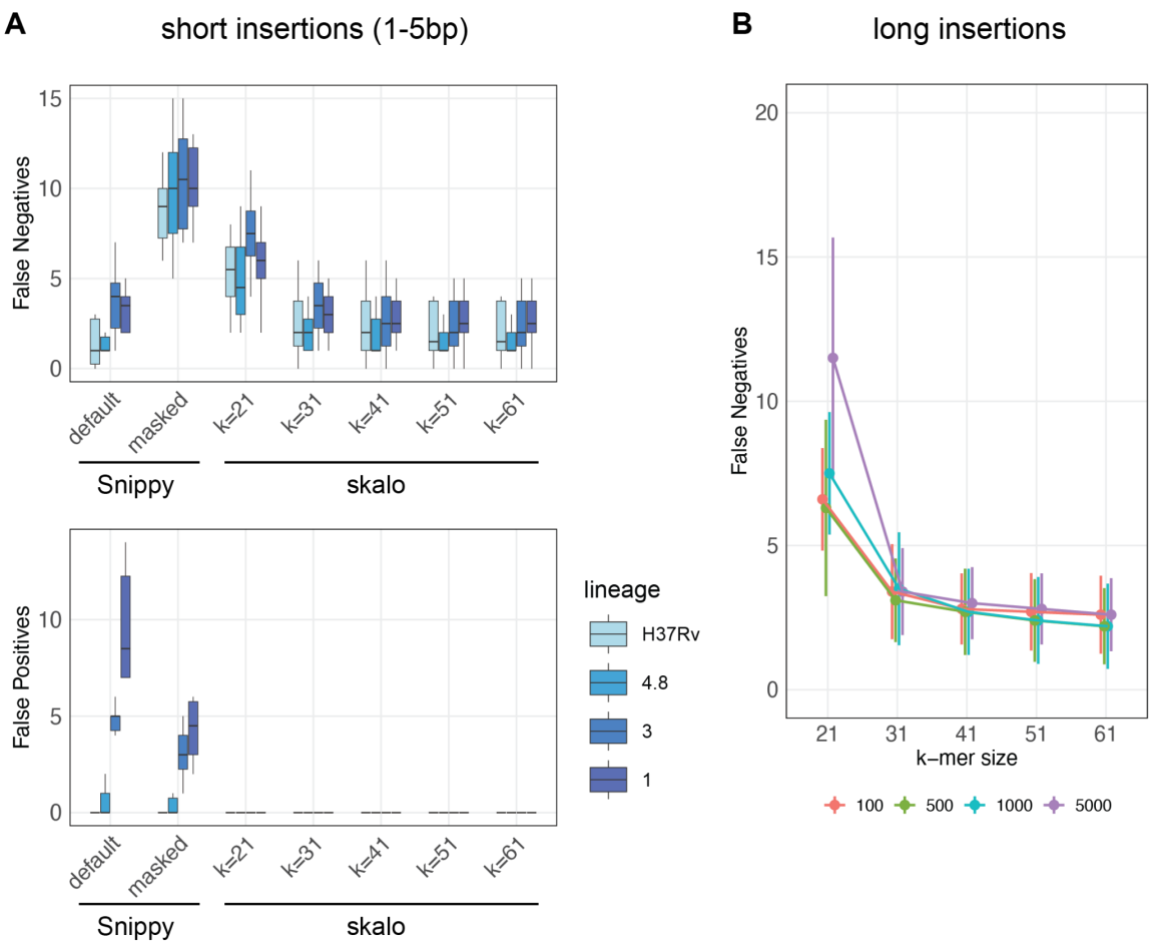


**Figure 2**: **A**: Detection of short simulated insertions (n=100). The parameters are indicated in the x-axis: with or without masking for Snippy, and the SKA k-mer size (k) for skalo. **B**: Detection of large simulated insertions by skalo (n=100). The different colours indicate different insert sizes, ranging from 100 to 5000 nucleotides. Error bars represent the standard deviations over 10 replicates.

Variant diversity in two large real-world Mtb datasets

The simulated benchmarks presented in the previous section revealed a high specificity of skalo with no false positive variant calls. However, a more realistic outbreak datasets benchmark should contain more than two samples, and indels are not randomly distributed in the genome but instead occur more often in low-complexity regions. To address these issues, we reanalysed in this section two large Mtb sequencing datasets obtained from tuberculosis surveillance programs in Birmingham (UK) (Walker et al. 2022) and the Valencia region (Spain) (Cancino-Munoz et al. 2022). Since split k-mer analyses perform better between closely related isolates (Harris 2018; Derelle et al. 2024), these two datasets were partitioned into 37 MTBC lineages, totalling 1,439 samples, that were independently analysed using SKA (k-mer size of 41) and skalo.

The numbers of indels and complex variants strongly correlated with the number of SNPs ($r^2$= 0.97 in both cases; Figure 3A; values available in Supplementary data). Using the slopes of the linear regression analyses, we determined that indels occurred at a rate of about one in every 11 SNPs, while complex variants occurred at a rate of about one in every 83 SNPs.

This indel rate is in the upper range of previous estimates (Godfroid et al. 2020; Sung et al. 2016; Gomes et al. 2021; Stritt et al. 2024). These analyses identified 199 indels of more than 1kb, with a maximum length of 26 kb, consistent with published results (Gomes et al. 2021; Baena et al. 2023). We then modified the code source of skalo to bypass the path selection based on sample set similarities, meaning that skalo now stops the graph traversal every time a multifurcation is encountered. Using this altered version, we observed a reduction in the numbers of inferred variants across all partitions, up to a reduction of 22% for the partition with the highest numbers of samples (Supplementary data). The number of indels of more than 1kb was particularly impacted as only 29 were identified (maximum length of 7 kb). The complexity of the De Bruijn graph is expected to increase with the number of samples, thereby hindering variant identification due to the increased number of multifurcations. However, this analysis shows that the simple graph traversal based on sample distributions, as implemented in skalo, mitigates this negative impact, making this tool suitable for analysing of large datasets.

Out of the 8,317 variants inferred by skalo, only 49 (0.6%) did not have a Blastn hit on the H37Rv genome assembly. We observed significant enrichments of both indels and complex variants in intergenic regions, repeat elements, and genes belonging to the PE-PPE family (p-value < 0.0001 in all 6 cases; Figure 3B and Supplementary data). Here again, these results are in line with previous observations (Stritt et al. 2024; Godfroid et al. 2020; Gupta and Alland 2021). Importantly, these three genomic compartments are known to be enriched in low-complexity sequences, indicating that skalo is able to infer indels and complex variants among repeated sequences. Only 106 variants were identified as having an unclear insert (i.e., multiple possible insert sizes), two-thirds of which were located within repeat elements (Supplementary data).

skalo was found to be relatively fast in these analyses, with runtimes increasing linearly with the number of analysed samples (0.4 second per sample on average; Figure 3C). The maximum memory consumption never exceeded a few gigabytes (Supplementary data). For instance, the analysis of the largest partition composed of 224 samples took 1.5 minutes and a maximum memory consumption of 2.7 GB. Please note however that these values do not include the creation of split k-mer files by SKA, which are prerequisite to skalo analyses. For instance, when analysing lineage 4.8.2 from the Birmingham dataset (55 samples), the creation of the skf file by SKA and its analysis by skalo took 10.7 minutes and 20 seconds respectively.

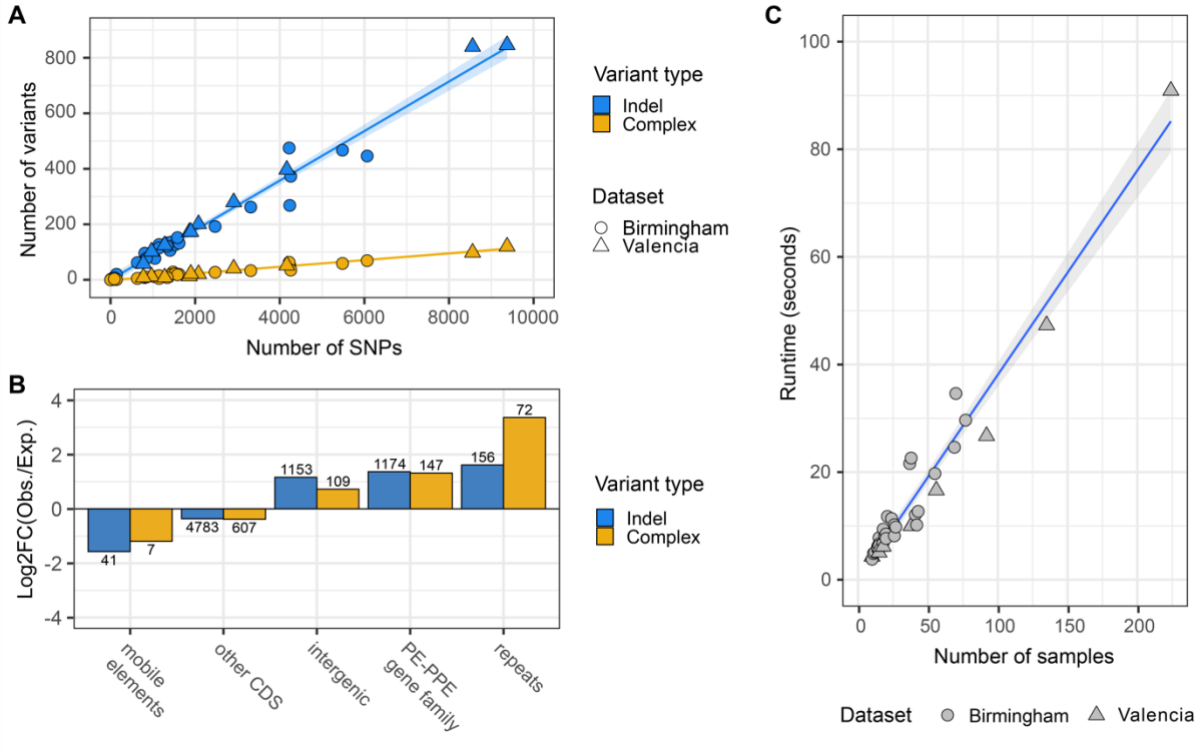


**Figure 3**: **A**: Number of indels and complex variants as a function of the number of SNPs observed in the 37 lineage groups. **B**: Log2 fold change enrichment of the observed ('Obs.') number of indels and multis in the different genomic compartments with respect to the expected ('Exp.') numbers obtained by chance. 'other CDS' includes all protein-coding genes except those of the PE-PPE family. **C**: skalo runtimes for each partition as a function of the number of samples.

Indels and complex variants as potential phylogenetic markers

We then tested variants inferred by skalo for two important metrics of phylogenetic markers: the percentage of phylogenetically informative variants (i.e., percentage of minor variants shared by at least two isolates; non-singletons) and the level of homoplasy. We used for this all variants inferred by SKA and skalo from the 37 Mtb lineage partitions described above.

As expected, the fraction of informative variants was found to be highly similar between SNPs, indels and complex variants, with median values of 28.7%, 30.8%, and 30.5% respectively (Figure 4A). Levels of homoplasy among informative variants were then estimated using phylogenies inferred from SNPs (see Methods). We observed similar levels of homoplasy between SNPs (12.6%; 2,630 out of 20,832), indels (11.8%; 214 out of 1,817) and complex variants (11.9%; 29 out of 243; Supplementary data). These two results indicate that variants identified by skalo have similar informative properties to SNPs, and could therefore be used as phylogenetic markers.

We then repeated all skalo analyses to assess the relevance of the two filtering parameters n (maximum homopolymer length) and m (maximum proportion of missing data) implemented in skalo. As shown in Figure 4B and 4C, we observed high levels of homoplasy for indels inferred within long homopolymers and variants presenting high levels of missing

data. The high degree of homoplasy of indels found in long homopolymers is likely attributed to the extensive polymorphism of these genomic tracks, as noted in previous studies (Godfroid et al. 2020; Vargas et al. 2023). As for variants with high levels of missing data, we observed that the higher the level of missing data, the higher the proportion of these characters having multiple hits on the H37Rv genome (Supplementary figure S3). This suggests that skalo generates some background noise for variants located in duplicated genomic regions, probably due to stochastic variations in coverages within and between samples leading to erroneous sample distributions of variants. This noise appeared however to be concentrated in variants with relatively high proportions of missing data.

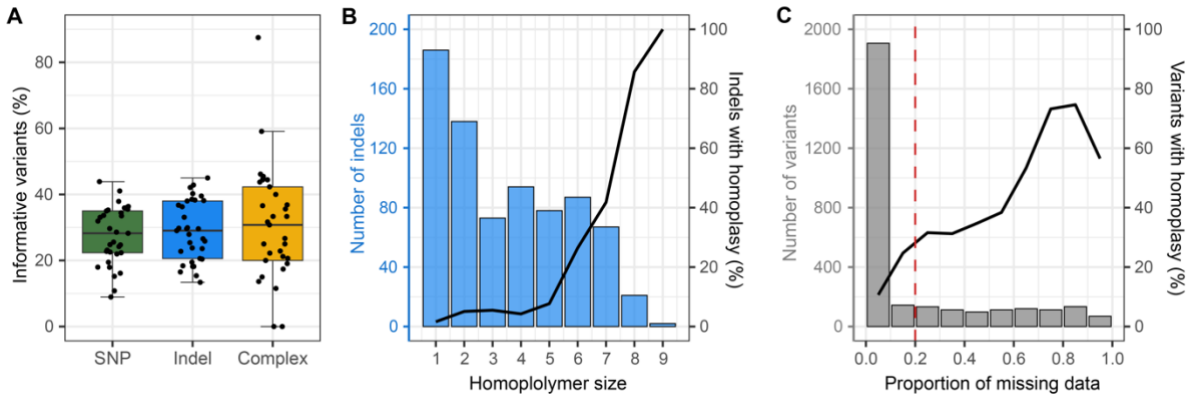


**Figure 4**: **A**: Percentage of informative SNPs, indels and complex variants. Data points represent each of the 33 partitions displaying at least 100 SNPs. **B**: Indels as a function of homopolymer sizes. skalo did not infer any indel in homopolymers of size 10 or more. **C**: Variants as a function of the proportion of missing data. The plot was based on an analysis using skalo parameter '-m 2', which effectively turns off the filter based on the proportion of missing data. The red dotted line indicates skalo default value.

Phylogenetic analysis of a Mtb outbreak

We finally assessed the impact of skalo variants on phylogenetic inferences by reanalysing the largest transmission cluster present in the two Mtb datasets previously analysed: 55 samples belonging to lineage 4.8.2 of the Birmingham dataset (12-SNPs cluster number 15 from (Walker et al. 2022)).

skalo inferred nine variants at k=41 for this Mtb outbreak, including one phylogenetically informative variant. At k=51, ten variants were inferred, which included two informative variants (minor variants shared by 7 and 17 samples; proportion of missing data equal to 0 in both cases). Considering variants inferred at k=51, phylogenetic analyses based on SNPs combined with skalo variants generated an ML tree having two more partitions than the tree obtained from the 89 SNPs alone, partially resolving a large polytomy (Figure 5A; trees with leaf names available in Supplementary figures S4 and S5). Other differences between the two ML trees correspond to variations of terminal branch lengths generated by non-phylogenetically informative skalo variants. However, the largest polytomy observed in the ML trees remained unresolved.

Interestingly, the two informative variants correspond to two in-frame complex variants located in the coding sequence of the genes pks12 and PGRS28 (Figure 5B), whose proteins have respectively known and hypothesised interactions with the human immune-system (Matsunaga et al. 2004; Guo et al. 2023). The highly complex patterns displayed by these two variants might correspond to genomic 'scars' identified by the pipeline ScarTrek and involved in reverse silencing (Gupta and Alland 2021). However, the multiple SNPs/indels generating these patterns must have all appeared in a very short timeframe, almost simultaneously. Therefore, these scars might have been selected due to the genomic diversity they created in these virulence genes, without any association with reverse silencing.

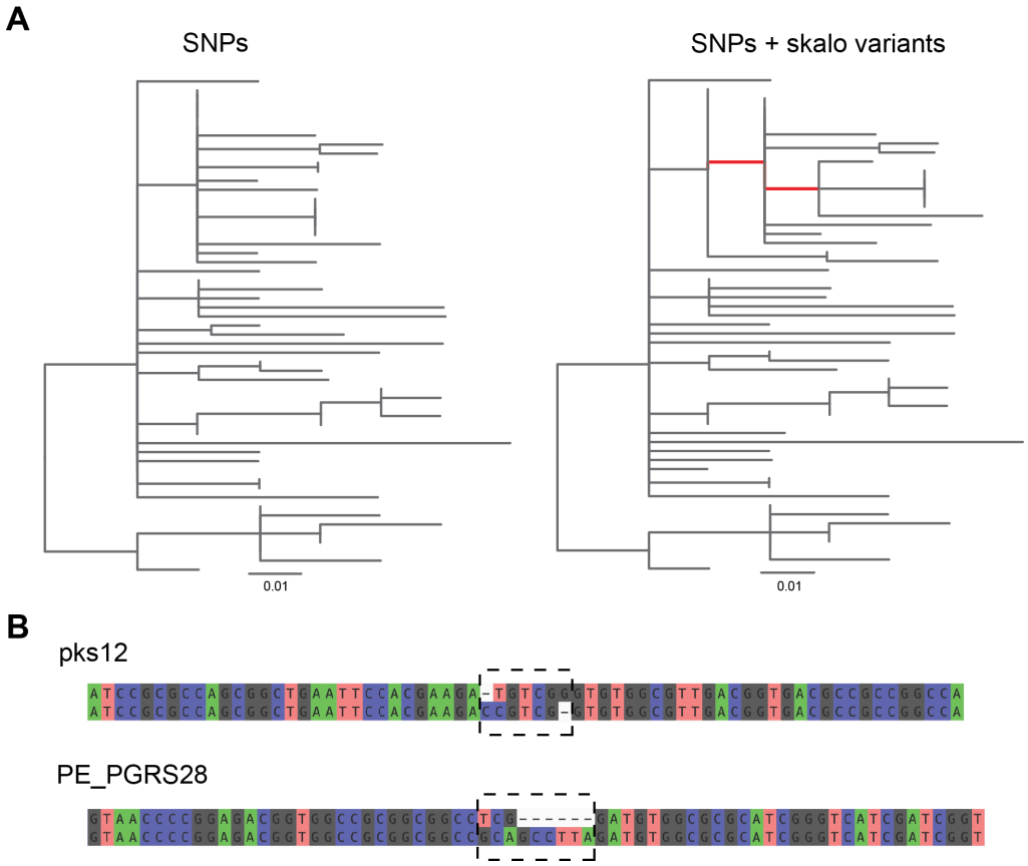


**Figure 5**: **A**: ML phylogenetic trees obtained from SNPs alone and SNPs combined with skalo variants, using arbitrary rooting. The two new bifurcations observed in the tree generated from SNPs combined with skalo variants are highlighted in red. **B**: Sequences of the two informative skalo variants, cropped to 20bp on both sides and with middle-bases manually aligned for illustrative purposes (skalo does not give full coordinates of the alignment). Gene names are indicated on top of each variant group and the middle-bases are highlighted by a dotted box.

## Discussion

Resolving indels and other complex variants is of particular importance for genomic resolution of outbreaks. SNPs alone cannot always be used to distinguish between pairs of isolates, and therefore resolve the direction of transmission. This is of particular importance for slowly evolving pathogens such as Mtb. Any additional information on differences in genomic content between samples is therefore of particular value in an outbreak setting.

Our tool skalo aims to make such analyses more accessible to researchers. skalo is very simple to use with no *ad hoc* criteria or parameter optimisation other than the split k-mer size used by SKA. As shown in this study, skalo can be used both to investigate the evolutionary ramifications of indels and complex variants found in outbreak genomic data, or for the purpose of phylogenetic analyses.

While only tested here on Mtb whole-genome sequencing data, skalo could be employed in the analysis of datasets from other species. Indeed, a lack of phylogenetic signal is often found in sequencing data of pathogen outbreaks, hampering inferences of transmission events (Campbell et al. 2018). Given its high speed, skalo would only marginally increase the computational times of split k-mer analyses, and potentially provide much needed additional phylogenetic markers. We note however that in species where recombination is frequent, skalo does not yet support positioning variants along a reference genome for the detection and removal of recombination events (Croucher et al. 2015; Didelot and Wilson 2015).

Several improvements of skalo are still possible. First, skalo will be fully integrated into SKA so that all variants (SNPs, indels and complex variants) can be inferred using a single program. Second, a full local-alignment algorithm could be implemented to automatically position variants on a reference genome. This would eliminate the need to rely on a third-party program such as blast for similarity searches to infer the genomic positions from similarity hits. Finally, a recent analysis has revealed the significant contribution of chromosomal rearrangements to the genomic diversity observed among closely related Mtb isolates (Stritt et al. 2024). The detection of such structural variants, represented by crossovers in the de Bruijn graph, could also be implemented in this program.

While indels and complex variants could already be inferred using read-alignment by seasoned bioinformaticians, quality filtering, multiple tools and internal benchmarking would typically be required to produce a high-confidence set of calls. The development of skalo is part of an ongoing wave of simple, fast, and accurate k-mer based programs created to facilitate and democratize bioinformatic analyses (Bray et al. 2016; Lu et al. 2022; Lees et al. 2019).

## Conflict of Interest

The authors declare that there are no conflicts of interest.

## Acknowledgements

RD, KM and AL were supported by the NIHR Health Protection Research Unit in Respiratory Infections, in partnership with the UK Health Security Agency [NIHR200927]. JAL was

supported by the European Molecular Biology Laboratory. NA. and LC acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and are also part of the EDCTP2 programme supported by the European Union.

## References

Armstrong, G. L., D. R. MacCannell, J. Taylor, H. A. Carleton, E. B. Neuhaus, R. S. Bradbury, J. E. Posey, and M. Gwinn. 2019. 'Pathogen Genomics in Public Health', *N Engl J Med*, 381: 2569-80.

Baena, A., F. Cabarcas, J. C. Ocampo, L. F. Barrera, and J. F. Alzate. 2023. 'Large genomic deletions delineate Mycobacterium tuberculosis L4 sublineages in South American countries', *PLoS One*, 18: e0285417.

Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter. 2016. 'Near-optimal probabilistic RNA-seq quantification', *Nat Biotechnol*, 34: 525-7.

Bush, S. J. 2021. 'Generalizable characteristics of false-positive bacterial variant calls', *Microb Genom*, 7.

Bush, S. J., D. Foster, D. W. Eyre, E. L. Clark, N. De Maio, L. P. Shaw, N. Stoesser, T. E. A. Peto, D. W. Crook, and A. S. Walker. 2020. 'Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines', *Gigascience*, 9.

Campbell, F., C. Strang, N. Ferguson, A. Cori, and T. Jombart. 2018. 'When are pathogen genome sequences informative of transmission events?', *PLoS Pathog*, 14: e1006885.

Cancino-Munoz, I., M. G. Lopez, M. Torres-Puente, L. M. Villamayor, R. Borras, M. Borras-Manez, M. Bosque, J. J. Camarena, C. Colijn, E. Colomer-Roig, J. Colomina, I. Escribano, O. Esparcia-Rodriguez, F. Garcia-Garcia, A. Gil-Brusola, C. Gimeno, A. Gimeno-Gascon, B. Gomila-Sard, D. Gonzales-Granda, N. Gonzalo-Jimenez, M. R. Guna-Serrano, J. L. Lopez-Hontangas, C. Martin-Gonzalez, R. Moreno-Munoz, D. Navarro, M. Navarro, N. Orta, E. Perez, J. Prat, J. C. Rodriguez, M. M. Ruiz-Garcia, H. Vanaclocha, Group Valencia Region Tuberculosis Working, and I. Comas. 2022. 'Population-based sequencing of Mycobacterium tuberculosis reveals how current population dynamics are shaped by past epidemics', *Elife*, 11.

Croucher, N. J., A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, and S. R. Harris. 2015. 'Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins', *Nucleic Acids Res*, 43: e15.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nat Genet*, 43: 491-8.

Derelle, R., J. Lees, J. Phelan, A. Lalvani, N. Arinaminpathy, and L. Chindelevitch. 2023. 'fastlin: an ultra-fast program for Mycobacterium tuberculosis complex lineage typing', *Bioinformatics*.

Derelle, Romain, Johanna von Wachsmann, Tommi Mäklin, Joel Hellewell, Timothy Russell, Ajit Lalvani, Leonid Chindelevitch, Nicholas J. Croucher, Simon R. Harris, and John A.

Lees. 2024. 'Seamless, rapid and accurate analyses of outbreak genomic data using Split K-mer Analysis (SKA)', *bioRxiv*: 2024.03.25.586631.
Didelot, X., and D. J. Wilson. 2015. 'ClonalFrameML: efficient inference of recombination in whole bacterial genomes', *PLoS Comput Biol*, 11: e1004041.
Fang, H., E. A. Bergmann, K. Arora, V. Vacic, M. C. Zody, I. Iossifov, J. A. O'Rawe, Y. Wu, L. T. Jimenez Barron, J. Rosenbaum, M. Ronemus, Y. H. Lee, Z. Wang, E. Dikoglu, V. Jobanputra, G. J. Lyon, M. Wigler, M. C. Schatz, and G. Narzisi. 2016. 'Indel variant analysis of short-read sequencing data with Scalpel', *Nat Protoc*, 11: 2529-48.
Fang, H., Y. Wu, G. Narzisi, J. A. O'Rawe, L. T. Barron, J. Rosenbaum, M. Ronemus, I. Iossifov, M. C. Schatz, and G. J. Lyon. 2014. 'Reducing INDEL calling errors in whole genome and exome sequencing data', *Genome Med*, 6: 89.
Gagneux, S. 2018. 'Ecology and evolution of Mycobacterium tuberculosis', *Nat Rev Microbiol*, 16: 202-13.
Gardner, S. N., and B. G. Hall. 2013. 'When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes', *PLoS One*, 8: e81760.
Gardner, S. N., T. Slezak, and B. G. Hall. 2015. 'kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome', *Bioinformatics*, 31: 2877-8.
Godfroid, M., T. Dagan, M. Merker, T. A. Kohl, R. Diel, F. P. Maurer, S. Niemann, and A. Kupczok. 2020. 'Insertion and deletion evolution reflects antibiotics selection pressure in a Mycobacterium tuberculosis outbreak', *PLoS Pathog*, 16: e1008357.
Gomes, L. C., S. Campino, C. R. F. Marinho, T. G. Clark, and J. E. Phelan. 2021. 'Whole genome sequencing reveals large deletions and other loss of function mutations in Mycobacterium tuberculosis drug resistance genes', *Microb Genom*, 7.
Guo, F., J. Wei, Y. Song, B. Li, Z. Qian, X. Wang, H. Wang, and T. Xu. 2023. 'Immunological effects of the PE/PPE family proteins of Mycobacterium tuberculosis and related vaccines', *Front Immunol*, 14: 1255920.
Gupta, A., and D. Alland. 2021. 'Reversible gene silencing through frameshift indels and frameshift scars provide adaptive plasticity for Mycobacterium tuberculosis', *Nat Commun*, 12: 4702.
Hall, B. G., and J. Nisbet. 2023. 'Building Phylogenetic Trees From Genome Sequences With kSNP4', *Mol Biol Evol*, 40.
Harris, S. R. 2018. 'SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology', *bioRxiv*.
Huerta-Cepas, J., F. Serra, and P. Bork. 2016. 'ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data', *Mol Biol Evol*, 33: 1635-8.
Iqbal, Z., M. Caccamo, I. Turner, P. Flicek, and G. McVean. 2012. 'De novo assembly and genotyping of variants using colored de Bruijn graphs', *Nat Genet*, 44: 226-32.
Jagielski, T., J. van Ingen, N. Rastogi, J. Dziadek, P. K. Mazur, and J. Bielecki. 2014. 'Current methods in the molecular typing of Mycobacterium tuberculosis and other mycobacteria', *Biomed Res Int*, 2014: 645802.
Lees, J. A., S. R. Harris, G. Tonkin-Hill, R. A. Gladstone, S. W. Lo, J. N. Weiser, J. Corander, S. D. Bentley, and N. J. Croucher. 2019. 'Fast and flexible bacterial genomic epidemiology with PopPUNK', *Genome Res*, 29: 304-16.
Li, H. 2014. 'Toward better understanding of artifacts in variant calling from high-coverage samples', *Bioinformatics*, 30: 2843-51.

Lu, J., N. Rincon, D. E. Wood, F. P. Breitwieser, C. Pockrandt, B. Langmead, S. L. Salzberg, and M. Steinegger. 2022. 'Metagenome analysis using the Kraken software suite', *Nat Protoc*, 17: 2815-39.

Matsunaga, I., A. Bhatt, D. C. Young, T. Y. Cheng, S. J. Eyles, G. S. Besra, V. Briken, S. A. Porcelli, C. E. Costello, W. R. Jacobs, Jr., and D. B. Moody. 2004. 'Mycobacterium tuberculosis pks12 produces a novel polyketide presented by CD1c to T cells', *J Exp Med*, 200: 1559-69.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Mol Biol Evol*, 37: 1530-34.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer. 2011. 'Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems', *Genome Biol*, 12: R112.

Napier, G., S. Campino, Y. Merid, M. Abebe, Y. Woldeamanuel, A. Aseffa, M. L. Hibberd, J. Phelan, and T. G. Clark. 2020. 'Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies', *Genome Med*, 12: 114.

Narzisi, G., J. A. O'Rawe, I. Iossifov, H. Fang, Y. H. Lee, Z. Wang, Y. Wu, G. J. Lyon, M. Wigler, and M. C. Schatz. 2014. 'Accurate de novo and transmitted indel detection in exome-capture data using microassembly', *Nat Methods*, 11: 1033-6.

Narzisi, G., and M. C. Schatz. 2015. 'The challenge of small-scale repeats for indel discovery', *Front Bioeng Biotechnol*, 3: 8.

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. 2018. 'Scaling accurate genetic variant discovery to tens of thousands of samples', *bioRxiv*: 201178.

Sobkowiak, Benjamin, and Caroline Colijn. 2022. 'Characterising indel diversity in a large Mycobacterium tuberculosis outbreak – implications for transmission reconstruction', *bioRxiv*: 2022.10.26.513840.

Stoler, N., and A. Nekrutenko. 2021. 'Sequencing error profiles of Illumina sequencing instruments', *NAR Genom Bioinform*, 3: lqab019.

Stritt, Christoph, Michelle Reitsma, Galo Goig, Anna Dötsch, Sonia Borrell, Christian Beisel, Daniela Brites, and Sebastien Gagneux. 2024. 'Large contribution of repeats to genetic variation in a transmission cluster of Mycobacterium tuberculosis', *bioRxiv*: 2024.03.08.584093.

Su, J., W. W. Lui, Y. Lee, Z. Zheng, G. K. Siu, T. T. Ng, T. Zhang, T. T. Lam, H. Y. Lao, W. C. Yam, K. K. Tam, K. S. Leung, T. W. Lam, A. W. Leung, and R. Luo. 2023. 'Evaluation of Mycobacterium tuberculosis enrichment in metagenomic samples using ONT adaptive sequencing and amplicon sequencing for identification and variant calling', *Sci Rep*, 13: 5237.

Sung, W., M. S. Ackerman, M. M. Dillon, T. G. Platt, C. Fuqua, V. S. Cooper, and M. Lynch. 2016. 'Evolution of the Insertion-Deletion Mutation Rate Across the Tree of Life', *G3 (Bethesda)*, 6: 2583-91.

The, CRyPTIC Consortium. 2022. 'Genome-wide association studies of global Mycobacterium tuberculosis resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms', *PLoS Biol*, 20: e3001755.

Uricaru, R., G. Rizk, V. Lacroix, E. Quillery, O. Plantard, R. Chikhi, C. Lemaitre, and P. Peterlongo. 2015. 'Reference-free detection of isolated SNPs', *Nucleic Acids Res*, 43: e11.

Valiente-Mullor, C., B. Beamud, I. Ansari, C. Frances-Cuesta, N. Garcia-Gonzalez, L. Mejia, P. Ruiz-Hueso, and F. Gonzalez-Candelas. 2021. 'One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads', *PLoS Comput Biol*, 17: e1008678.

van der Putten, B. C. L., N. A. H. Huijsmans, D. R. Mende, and C. Schultsz. 2022. 'Benchmarking the topological accuracy of bacterial phylogenomic workflows using in silico evolution', *Microb Genom*, 8.

Vinga, S., and J. Almeida. 2003. 'Alignment-free sequence comparison-a review', *Bioinformatics*, 19: 513-23.

Walker, T. M., M. Choisy, M. Dedicoat, P. G. Drennan, D. Wyllie, F. Yang-Turner, D. W. Crook, E. R. Robinson, A. S. Walker, E. G. Smith, and T. E. A. Peto. 2022. 'Mycobacterium tuberculosis transmission in Birmingham, UK, 2009-19: An observational study', *Lancet Reg Health Eur*, 17: 100361.

Walter, K. S., C. Colijn, T. Cohen, B. Mathema, Q. Liu, J. Bowers, D. M. Engelthaler, A. Narechania, D. Lemmer, J. Croda, and J. R. Andrews. 2020. 'Genomic variant-identification methods may alter Mycobacterium tuberculosis transmission inferences', *Microb Genom*, 6.

Zielezinski, A., S. Vinga, J. Almeida, and W. M. Karlowski. 2017. 'Alignment-free sequence comparison: benefits, applications, and tools', *Genome Biol*, 18: 186.