



HAL
open science

Evaluating impacts of syntenic block detection strategies on rearrangement phylogeny using *Mycobacterium tuberculosis* isolates

Afif Elghraoui, Siavash Mirarab, Krister M. Swenson, Faramarz Valafar

► To cite this version:

Afif Elghraoui, Siavash Mirarab, Krister M. Swenson, Faramarz Valafar. Evaluating impacts of syntenic block detection strategies on rearrangement phylogeny using *Mycobacterium tuberculosis* isolates. *Bioinformatics*, 2023, 39 (1), pp.#btad024. 10.1093/bioinformatics/btad024 . hal-03875983

HAL Id: hal-03875983

<https://hal.science/hal-03875983>

Submitted on 28 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Genome analysis

Evaluating impacts of syntenic block detection strategies on rearrangement phylogeny using *Mycobacterium tuberculosis* isolates

Afif Elghraoui^{1,2}, Siavash Mirarab², Krister M. Swenson³ and Faramarz Valafar^{4,*}

¹Department of Electrical and Computer Engineering, San Diego State University, San Diego, CA 92182, USA, ²Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093, USA, ³LIRMM, CNRS, Univ. Montpellier, Montpellier, France and ⁴School of Public Health, San Diego State University, San Diego, CA 92182, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on July 29, 2022; revised on December 21, 2022; editorial decision on December 27, 2022; accepted on January 11, 2023

Abstract

Motivation: The phylogenetic signal of structural variation informs a more comprehensive understanding of evolution. As (near-)complete genome assembly becomes more commonplace, the next methodological challenge for inferring genome rearrangement trees is the identification of syntenic blocks of orthologous sequences. In this article, we studied 94 reference quality genomes of primarily *Mycobacterium tuberculosis* (*Mtb*) isolates as a benchmark to evaluate these methods. The clonal nature of *Mtb* evolution, the manageable genome sizes, along with substantial levels of structural variation make this an ideal benchmarking dataset.

Results: We tested several methods for detecting homology and obtaining syntenic blocks and two methods for inferring phylogenies from them, then compared the resulting trees to the standard method's tree, inferred from nucleotide substitutions. We found that, not only the choice of methods, but also their parameters can impact results, and that the tree inference method had less impact than the block determination method. Interestingly, a rearrangement tree based on blocks from the Cactus whole-genome aligner was fully compatible with the highly supported branches of the substitution-based tree, enabling the combination of the two into a high-resolution supertree. Overall, our results indicate that accurate trees can be inferred using genome rearrangements, but the choice of the methods for inferring homology requires care.

Availability and implementation: Analysis scripts and code written for this study are available at <https://gitlab.com/LPCDRP/rearrangement-homology.pub> and <https://gitlab.com/LPCDRP/syntement>.

Contact: faramarz@sdsu.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Methods for phylogenetic inference based on genomic rearrangements have been developed and refined over the past several decades (Moret *et al.*, 2013), but the majority of biologists continue to rely primarily on phylogenetic inference methods based on nucleotide substitutions. As a major evolutionary mechanism, genomic structural variation cannot be neglected, yet the application of gene-order methods has been limited to specific cases of well-assembled eukaryotic genomes (e.g. Feng *et al.*, 2017; Pevzner and Tesler, 2003), as well as to small plastid genomes (Moret *et al.*, 2002). For example, the tool CREx is currently the standard software for analyzing plastid gene order evolution (Bernt

et al., 2007). The widespread application of these methods to larger genomes, however, has faced many challenges. Before the recent advent of accurate long-read sequencing technology, the proper *de novo* assembly of full genomes for use as input to a gene order analysis had proven difficult; indeed, the mechanisms of rearrangement are often inextricably linked to duplicated regions that are well known for confusing short-read reference-based sequencing methods (Liu *et al.*, 2012; Ranzet *et al.*, 2007).

Assembly of reference-quality bacterial genomes is now routine (Koren and Phillippy, 2015; Phillippy, 2017) and full assembly of the much larger and complex eukaryotic genomes is now possible (Liao *et al.*, 2021; Miga *et al.*, 2019; Nurk *et al.*, 2021; Rhie *et al.*,

2021). The improved assemblies have the potential to enable a wider use of rearrangements as a phylogenetic signal. However, methodological challenges associated with the use of such signal remain and call for better methods and better empirical understanding of the strengths and weaknesses of existing methods.

Most of the research in rearrangement phylogenetics is focused on the difficult problem of modeling complex scenarios that can arise from genomic structural variations mediated by many different mechanisms (Fertin, 2009). Methods that infer phylogenies based on rearrangements are of three varieties (Moret et al., 2013).

1. Model-free methods, which treat rearrangements as character evolution. For example, some methods encode adjacencies as simple characters (e.g. binary or copy-number) and use standard character evolution models together with maximum likelihood inference (e.g. Hu et al., 2014; Lin et al., 2012a; Wang et al., 2002).
2. Another family of methods uses distance-based tree reconstruction by finding the minimum number of events needed to transform one genome to another (e.g. Bohnenkämper et al., 2020; Feijao and Meidanis, 2011; Moret et al., 2002; Wang et al., 2006), a problem that remains challenging if there are duplicated genes (Fertin, 2009).
3. Finally, models of pairwise comparison can be generalized to the computation of rearrangements on a fixed phylogeny. This *small phylogeny problem* can be solved repeatedly, possibly combined with branch-and-bound, while searching for a tree topology. Instead of solving the small phylogeny problem directly, these approaches are usually based on the median of three genomes (Moret et al., 2013; Pe'er and Shamir, 1998; Sankoff and Blanchette, 1998), sometimes coupled with other heuristics (Bourque and Pevzner, 2002).

While the last approach is the most thorough, and a version of the median problem can now often be solved for duplication-free scenarios of reasonable size (Xu, 2010), the small phylogeny problem has proven very challenging in the presence of segmental duplication, and questions of scalability remain (Doerr and Chauve, 2021). Thus, unless compromises are made that resolve duplicated segments beforehand, the first two approaches are the only types of method that are practical for datasets of even moderate size. Several algorithms exist from both categories, and the relative accuracy of these methods has been the subject of study (e.g. Biller et al., 2016; Lin et al., 2011). Regardless of which one is used, a more prosaic question is that of preparing the input to these methods.

Although there are exceptions (Doerr et al., 2018), the input to rearrangement phylogeny reconstruction consists of a set of homologous *blocks* of nucleotide sequence, their homology assignments within and across genomes, and their relative positions and directions along the genomes. Clearly, there are many ways to define such blocks (Ghiurcuta and Moret, 2014), and detection of homology is far from trivial. The most obvious approach to homology detection is to annotate genes using gene models. This approach has to contend with difficulties of gene annotation (Salzberg, 2019), and outside prokaryotes, the more damaging issue is that only a small portion of the genome can be used. An alternative is to pairwise align genomes with respect to each other or a reference genome and use the alignments to define blocks of homologous sequence. Any definition of a block should allow some levels of heterogeneity within the block, often necessitating thresholds for defining how much variability is tolerated. Pairwise alignment has a fundamental limitation: there is no guarantee that the results are consistent (e.g. are transitive). Thus, a possibly better approach is to rely on multiple whole genome alignment (WGA). There has been much progress in recent years on scalable and accurate methods for multiple WGA (Armstrong et al., 2019; Earl et al., 2014) and new ways to compute syntenic blocks (Kolmogorov et al., 2018).

Defining the block-level input to rearrangement phylogeny algorithms remains a challenging problem (Lucas and Roest Crolius, 2017), as demonstrated by a thorough literature search which reveals roughly thirty tools developed for that purpose since 2004. Ghiurcuta

and Moret (2014) attempted to set basic standards for defining syntenic blocks but also acknowledged that the variety of criteria for selecting them corresponds to the variety of applications for which they are used, and so what is better for one application does not necessarily suit another. Despite these attempts, little is known about the relative accuracy of available options and the extent of their impact on the resulting phylogeny. One challenge when studying block inference methods is the lack of a sufficiently realistic genome simulator; to our knowledge existing simulators either do not allow for ancestral genomes to be specified as part of the input (Davín et al., 2020; Hindré et al., 2012) or are no longer maintained (i.e. Evolver). In the absence of such simulators, we have to rely on empirical data, which poses its own challenges. In particular, studying the impact of the block definition would be further complicated by large genome sizes, or very complex evolutionary histories including horizontal transfers, gene duplication events, or polyploidy. Thus, a relatively simple model organism is preferable.

As a way to minimize the issue of evolutionary model complexity, we consider *Mycobacterium tuberculosis* (*Mtb*) as a subject. Mycobacteria are unique in that they do not undergo horizontal gene transfer (HGT) in the traditional sense. While some mycobacteria have been observed to recombine via distributive conjugal transfer (DCT) (Gray and Derbyshire, 2018), the human-adapted pathogen *Mtb* in particular appears to have recently diverged yet contains appreciable diversity (Coscolla and Gagneux, 2014; Galagan, 2014). It does not show evidence of either DCT or traditional HGT and appears to have undergone strictly vertical evolution (Brosch et al., 2002; Gagneux, 2018). Nevertheless, structural variations do happen for these strains; even within the species, gene duplication and gene conversion (Uplekar et al., 2011), as well as inversions (Merrikkh and Merrikkh, 2018), have been observed. According to the classification of Koren et al. (2013), *Mtb* has a class II genome, characterized by many mid-scale repeats of ~1.5 kb insertion sequences. Focusing on *Mtb* simplifies the evolutionary models we must consider, and leaves us with the final difficulty of rearrangement phylogeny: defining suitable syntenic blocks. Despite its apparent clonal evolution, *Mtb* has diversified into several lineages (Fig. 1) distinguished by variations in repetitive regions (Kanduma et al., 2003), with the three ‘modern’ lineages further separated from four ‘ancestral’ lineages by the deletion of the TbD1 locus (Brosch et al., 2002; Gordon et al., 1999; Mostowy et al., 2002). Its evolution is driven in part by antibiotic pressure, though some lineages are more virulent and more successful than others (Merker et al., 2015), such as the globally prevalent Euro-American (L4) and East-Asian (L2) lineages (Coscolla and Gagneux, 2014).

In this article, we use a set of complete genomes for 92 largely drug-resistant *Mtb* clinical isolates and two reference strains, evaluating different methods of syntenic block determination with respect to how the block-order phylogeny built from them compares to standard trees inferred from substitutions. For each method, we use both an adjacency-based algorithm, maximum likelihood on whole-genome data (MLWD) (Lin et al., 2012a), and a recent distance-based method called DING (Bohnenkämper et al., 2020), to infer rearrangement-based phylogenies. We use two distinct approaches as input to these methods: (i) syntenic blocks determined using modern WGA methods [Cactus (Armstrong et al., 2020) with different parameters and SibeliaZ-LCB (Minkin and Medvedev, 2020)], and (ii) blocks determined by our in-house gene annotation pipeline. Phylogenies are built on these input blocks before and after applying a block aggregator called maf2synteny (Kolmogorov et al., 2018). We compare the resulting trees to each other, and to those inferred using substitutions alone, in order to quantify their levels of discordance, especially with respect to branches with high statistical support.

2 Materials and methods

2.1 Genome assemblies

Genome assemblies made available by Modlin et al. (2020) are used. These include 85 isolates resequenced from the set collected by the Global Consortium for Drug-resistant Tuberculosis Diagnostics

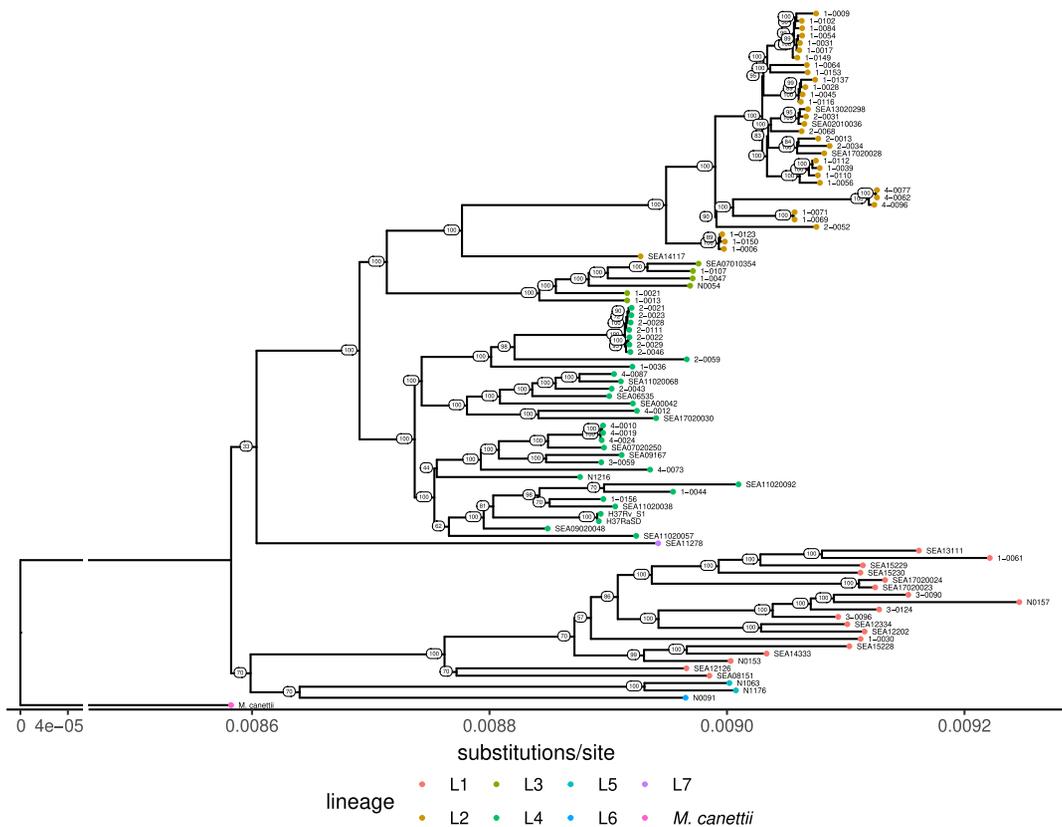


Fig. 1. Phylogeny of *Mycobacterium tuberculosis* (*Mtb*) clinical isolates used in this study. This maximum-likelihood phylogenetic tree inferred using the GTRCAT substitutions model from concatenated variable sites with respect to the reference strain H37Rv (NC_000962.3) shows the separation of our *Mtb* isolate set into 7 of the defined lineages and the level of diversity between them. This includes 3 *Mycobacterium africanum* isolates (lineages 5 and 6). *Mycobacterium canettii* is used as the outgroup

(GCDD) (Hillery *et al.*, 2014) (NCBI Bioproject PRJNA555636), 7 isolates from Berney *et al.* (2015) (NCBI Bioproject PRJEB8783), reference strains H37Rv (NCBI Bioproject PRJNA555636), H37Ra (NCBI accession NZ_CP016972.1) and *Mycobacterium canettii* (NCBI accession NC_019951.1). All sequencing data used for the assemblies, except the *M. canettii* assembly, are Pacific Biosciences SMRT sequencing reads. The assembly protocol of Modlin *et al.* (2020) was based on HGAP2 (Chin *et al.*, 2013) or, if that failed to produce a single contig, Canu (Koren *et al.*, 2017). The contigs were circularized using minimus2 (<http://amos.sourceforge.net/>) or circlator (Hunt *et al.*, 2015), followed by iterative assembly consensus polishing with Quiver (Chin *et al.*, 2013). To validate our assemblies, we used structural variant detection method PBHoney (English *et al.*, 2014), which detects irregularities such as soft-clippings in the alignment of reads to a reference. We applied PBHoney to the reads' alignment to the assembly that was generated from them, so any structural variant detected by PBHoney would in fact be a candidate misassembly. None of the genomes used here had any misassemblies detected.

Lineages were identified using TB-profiler (Phelan *et al.*, 2019) version 4.1.1 with database version 2022-01-25.

2.2 Block assignment

The synteny block assignment strategies we used here fall into the two categories of annotation-based and alignment-based methods and so differ in the type of markers they use for defining their blocks.

2.2.1 Annotation-based methods

Annotation and homology assignment: All genomes were simultaneously annotated using the Hybran pipeline (Elghraoui *et al.*, 2022). Hybran uses a combination of reference-based annotation transfer

implemented by RATT (Otto *et al.*, 2011) and Prokka, an *ab initio* method (Seemann, 2014). Annotations from Prokka are only retained in the places of gaps left where no suitable reference annotation could be transferred. The reference annotation used was that of *M. tuberculosis* H37Rv (NCBI accession NC_000962.3).

In the last stage of the annotation pipeline, orthologous genes across the genomes are identified using CD-HIT (Fu *et al.*, 2012; Li and Godzik, 2006) and MCL (Enright *et al.*, 2002) clustering. Genes that cluster together are assigned the same name if they have at least 95% protein sequence identity and 95% alignment coverage with the representative sequence of the cluster (this threshold is applied to both query/subject and subject/query coverage). These arbitrary thresholds clearly have the potential to impact results. Thus, we also ran the annotation pipeline with relaxed thresholds for the orthology mapping: 75% minimum identity and 66% minimum alignment coverage. These results are subsequently referred to as 'annotation-relaxed'. Note that due to the low divergence of sequences in our dataset, the use of CD-HIT is justified as it is designed for finding similarity at 60% or higher (Chen *et al.*, 2016).

2.2.2 Alignment-based methods

Cactus (Armstrong *et al.*, 2020) is a whole-genome aligner based on its namesake cactus graphs, which organize alignments hierarchically to reveal their substructures. The Cactus aligner is run with default parameters. Cactus requires soft-masking of repetitive sequences prior to alignment, so we applied nucmer (Kurtz *et al.*, 2004) to identify them and bedtools (Quinlan and Hall, 2010) to apply the masking. Cactus requires a guide tree as input, and the reference tree based on single nucleotide polymorphisms (SNP tree, described in Section 2.4) was provided by default. We tested the robustness to this choice by changing the guide tree. We constructed an alternative guide tree using FastME (Lefort *et al.*, 2015) on a

distance matrix computed using Mash (Ondov et al., 2016) using the maximum k-mer size (-k) of 32 due to the clonality of the dataset, and a sketch size (-s) of 1 billion, and phylogenetically corrected using the Jukes and Cantor (1969) (JC) model. We also used as a guide tree an adjacency-based tree (Section 2.4) inferred using synteny blocks computed from an alternative method, SibeliaZ-LCB, described in the next section. The guide tree used is denoted in parentheses if it is not the SNP tree. For example, Cactus(Mash) refers to the Cactus alignment based on the Mash guide tree.

SibeliaZ (Minkin and Medvedev, 2020) is a whole-genome alignment method for closely-related genomes based on de Bruijn graph analysis. It was run with graph order (k-mer length) set to 15, the developer recommended value for bacterial datasets, and two different vertex frequency thresholds. To reduce the de Bruijn graph complexity, SibeliaZ removes k-mers that appear more frequently than the vertex frequency threshold. One of our runs used the default value of 150, and a second run (which we refer to as SibeliaZ-LCB-highVF) was performed using a documentation-guided setting of $8924 = 97$ (number of genomes) \times 23 [maximum multiplicity of any one gene from the annotation (Table 1)] \times 4 (scaling factor). Given that we required only the coordinates of the locally collinear blocks, rather than the full nucleotide alignments themselves, we ran SibeliaZ-LCB only, rather than the complete SibeliaZ alignment pipeline, (-n) following the developer recommendation.

2.2.3 Building synteny blocks

Cactus and SibeliaZ-LCB each produce a collection of sets of alignable genomic intervals that are inferred to be homologous, called *alignment blocks*. We used a custom script (<https://gitlab.com/LPCDRP/syntemem>) to formulate these alignment blocks as synteny block permutations for direct input into tree inference, as well as for formulating genes from the annotation output as such. The WGA blocks can be very short, which can lead to two problems: (i) the running time of tree inference methods increases rapidly with the number of blocks and especially the number of duplicates, and (ii) blocks that are very small may not represent true homology. A potential solution is maf2synteny (Kolmogorov et al., 2018), a program that can combine multiple adjacent small blocks into larger blocks if they are consistently syntenic across genomes. By creating these synteny blocks, maf2synteny may produce more accurate statements of homology and also will reduce the running time of tree inference. For consistency and to be able to run the more time-consuming distance inference methods, in this study, we also apply maf2synteny to annotations to combine syntenic genes into syntenic blocks.

In combining blocks, maf2synteny considers the lengths of the alignments and the length of the gaps between them. Specifically, maf2synteny takes two arguments: a set of simplification parameters

$S = \{(minBlock_1, maxGap_1), (minBlock_2, maxGap_2), \dots\}$ which govern when a block is expanded to incorporate adjacent aligned segments, and a synteny block scale (block_sizes), which we explore in Section 3. The method is based on an A-Bruijn graph, where there is a vertex for each alignment block (of length at least minBlock) and, for each genome, an edge between adjacent alignment blocks. Thus, a vertex can have maximum degree twice the number of genomes. A *collinear* path in the graph is one that includes only alignment blocks sharing the same set of genomes, and it indicates a set of collinear alignment blocks in those genomes. Maximal collinear paths, with the condition that no adjacent alignment blocks are more than maxGap apart, are aggregated. To permit heterogeneity within the syntenic blocks, maf2synteny also aggregates each set of alignment blocks that participate in a *bubble* (subject to the same maxGap parameter), which is a pair of collinear paths that share endpoints. The set of parameter pairs S is visited from smallest pairs to largest, and applied to the graph until the target syntenic block scale block_sizes is reached.

2.3 Evaluations

Because we analyze a real dataset, all trees in this study are inferred from the data. To use as a reference, we inferred trees using substitution models (Section 2.4). Clearly, there is no guarantee that these trees represent the true evolutionary history, and we invite readers to keep this point in mind when interpreting results. Nevertheless, when a method shows more similarity to the SNP tree, especially among highly supported branches, we can interpret this similarity as combined evidence for a branch in the true history. One caveat is the use of SNP tree as the guide tree for Cactus, which will be explored.

For each pair of fully resolved trees, we compare the trees using the normalized (Robinson and Foulds, 1981) (RF) distance and matching split distance (Bogdanowicz and Giaro, 2012) (MS) metrics, computed using TreeCmp (Bogdanowicz et al., 2012). Because we study the evolution of relatively closely related genomes, not every branch can be resolved with high confidence. Thus, in addition to comparing fully resolved trees, we also study highly supported branches. We considered each fully resolved tree as well as after contracting branches with bootstrap support (BS) below levels 33%, 50%, 75%, 95% and 100%. For contracted trees, RF and MS become difficult to interpret. Thus, instead, we report the total number of non-trivial branches in the tree (indicating its resolution), and the number of non-trivial branches that are compatible between two trees (recall that two bipartitions are compatible if they can both exist in the same tree).

Table 1. Characteristics of the blocks generated by each method

Method	Number of blocks	Average genome coverage	Number of duplicate blocks	Number of duplicate occurrences	Multiplicity	
					Mean	Max
annotation	6403	88.6%	35	3002	6.360	23
annotation+maf2synteny	2713	76.4%	20	1441	7.315	23
annotation-relaxed	5305	89.0%	50	3491	4.828	23
annotation-relaxed+maf2synteny	1658	78.6%	26	1777	4.689	23
Cactus	17 115	100%	3165	75 899	2.047	4
Cactus+maf2synteny	741	93.7%	41	821	2.037	3
Cactus-filtered	7527	98.9%	532	11 570	2.020	3
Cactus-filtered+maf2synteny	791	93.6%	41	821	2.037	3
SibeliaZ-LCB	3276	98.8%	229	4217	2.035	6
SibeliaZ-LCB+maf2synteny	682	93.5%	47	711	2.055	4
SibeliaZ-LCB-highVF	3479	99.7%	548	112 783	3.468	65
SibeliaZ-LCB-highVF+maf2synteny	454	91.8%	34	5159	2.467	7

Note: Synteny blocks produced by the methods in bold are the focus of subsequent analyses. They represent configurations that ultimately produced phylogenetic trees more congruent with the reference tree, as compared to alternative configurations of the same tools.

2.4 Phylogenetic tree inference

Reference SNP tree: SNPs were called with respect to the reference H37Rv (NCBI accession NC_000962.3) using show-snps from the MUMmer package (Kurtz *et al.*, 2004), then concatenated into a PHYLIP-formatted alignment. From this, RAxML (Stamatakis, 2014) was used to create a maximum-likelihood tree using the GTRCAT model with Felsenstein ascertainment bias correction using the count of sites in the reference genome where no SNP was observed, and bootstrapping with 100 replicates.

Rearrangement-based trees: For each of the block assignments specified in Section 2.2, we test both an adjacency-based and a distance-based method. Among adjacency-based methods, we used a tree inferred using MLGO (Hu *et al.*, 2014) phylogenetic tree reconstruction with bootstrapping on 100 replicates. This tool implements the MLWD algorithm (Lin *et al.*, 2012a), in which ordered markers are converted into a vector representing the copy-number of marker adjacencies. A maximum likelihood tree is then inferred, considering the transitions between these two states for each adjacency. Among distance-based methods, we used DING (Bohnenkämper *et al.*, 2020) to calculate a distance for every pair of samples based on the block assignments for each method (Section 2.2). DING computes distances under the double cut and join (DCJ) model that also accounts for duplications and segmental insertions/deletions. We reconstructed phylogenetic trees based on these distance matrices using FastME (Lefort *et al.*, 2015). The distance matrices could not be computed for the unaggregated annotation blocks, SibeliaZ-LCB-highVF, or the unfiltered and unaggregated Cactus blocks after running DING for one week on a machine with 256GB of RAM, but we were able to infer distance trees for the remaining configurations: annotation+maf2synteny, Cactus-filtered and SibeliaZ-LCB (both with and without maf2synteny), which we subsequently refer to with a +DING suffix.

Combining trees: After contracting low support branches, we are left with multifurcating trees. When two multifurcating trees are compatible, they can be easily combined by simply finding the union of their bipartitions, which implies a tree. We built such combined trees using Dendropy (Sukumaran and Holder, 2010) and a custom script.

3 Results

3.1 Composition of syntenic blocks

While the blocks produced by all the methods cover at least 75% of the genome, the Cactus alignment resulted in complete coverage. This full coverage is achieved through the creation of an order of magnitude more syntenic units than the other methods pre-aggregation. SibeliaZ-LCB-highVF comes close to Cactus, with an average genome coverage of 99.7% but using close to 5 times fewer blocks. In fact, SibeliaZ-LCB-highVF only has 203 more blocks than its sister run SibeliaZ-LCB, which has 98.8% coverage. Cactus' coverage becomes almost identical to SibeliaZ-LCB and slightly less than SibeliaZ-LCB-highVF once we filtered the nearly 10 000 alignments with fewer than 50 sites (Cactus-filtered) (Table 1) that accounted for 1.1% of the genome on average. The number 50 was chosen here as it was the minimum length of blocks produced by SibeliaZ-LCB. Even still, Cactus-filtered has over double the number of blocks as SibeliaZ-LCB (7527 versus 3276).

The annotation pipeline detects 6403 blocks (i.e. gene families). Gene orthology mapping using relaxed thresholds (annotation-relaxed) produced 17% fewer blocks compared to the default thresholds while maintaining the same genome coverage, reflecting the fact that more pairs of genes are identified as homologous as a result of the lower similarity thresholds. Most blocks appeared only once in each genome (i.e. were not duplicated), in contrast to pre-aggregation Cactus and SibeliaZ, both of which have at least hundreds of duplicates. Cactus, prior to filtering, had thousands of duplicate blocks with tens of thousands of occurrences. Most of these, however, were among the short aligned segments (often only a few bp long) excluded in Cactus-filtered, and so the number of duplicate blocks drops substantially—from 3165 to 532—after

filtering. Most alignment-based methods did not identify markers with a copy number over 6 in any single genome. Thus, they generally failed to capture into a single unit known high-duplicity markers such as the annotated transposases (of length 1.5 kb), which had a maximum multiplicity of 23 in the annotation pipeline. A notable exception is SibeliaZ-LCB-highVF, which delineated synteny blocks duplicated up to 65 times in a single genome. These duplicates, however, are mostly short (median length 129 bp) and dissipate upon processing with maf2synteny: the maximum multiplicity among the aggregated blocks falls to 7 for SibeliaZ-LCB-highVF+maf2synteny.

With regard to running time, the three methods were widely different, with SibeliaZ-LCB taking only minutes while annotations require hours and Cactus close to a day (Supplementary Table S1). Note that Cactus is the only method here that is producing a complete WGA, explaining its increased running time.

3.2 Impact of block aggregation with maf2synteny

Aggregating the markers with maf2synteny generally resulted in a slight drop in coverage (Table 1). Intriguingly, both alignment-based methods SibeliaZ-LCB and Cactus result in a similar numbers of syntenic blocks after aggregation (≈ 700). However, SibeliaZ-LCB-highVF's aggregation results in ~ 200 fewer blocks than SibeliaZ-LCB's despite starting with 203 more blocks. The coverage drop following maf2synteny for the annotation markers was more severe, from 88.6% to 76.4%, and the number of blocks was reduced by more than half (6403 to 2713). The number of duplicates also drops sharply as a result of maf2synteny, never exceeding 50 in any condition after aggregation whereas it could be as high as 3165 prior. In fact, the number of duplicate markers according to the aggregated alignment-based methods becomes similar to the raw annotation despite the order of magnitude difference in the number of blocks.

The block compositions, however, are sensitive to the parameters used with maf2synteny (Fig. 2). In particular, maf2synteny requires a block scale (`-block_sizes`), which is set by default to 5000 and which we have set to 500 to be approximately half the size of the average gene in the reference genome annotation. Exploring this parameter in one instance of each tool, we detected that it has a major impact on the number of blocks obtained, with the default resulting in 24–123 blocks. Reducing the block scale to 500 increased the number of syntenic blocks to around 700 for the alignment-based methods, with only minimal changes to the coverage. For all methods, average coverage generally decreases with increasing block scale. The highest coverage is attained at the lowest block scale tested, 50, and the next highest coverage comes with block scale 1000 rather than 500. This is likely due to the interaction of the block scale parameter and maf2synteny's simplification parameters (Section 2.2.3), where one more iteration of simplification takes place when using block scale 1000 versus block scale 500. The simplification parameters, however, are also tuneable to circumvent this lack of monotonicity between the block scale intervals we tested. The overall drop in coverage is more substantial for annotation and approaches 65% at larger block scales. The number of blocks and duplicate block occurrences decreases monotonically with increasing block scale for all methods. Thus, the impact of this parameter is mostly in how aggressively blocks are combined and not in how much of the genome is captured, except for annotation. Reassuringly, Cactus and SibeliaZ-LCB have comparable numbers of blocks at each block scale.

Because the default value, 5000, resulted in ~ 50 blocks, it provides very little phylogenetic signal. A tree computed for Cactus using this default setting had almost no resolution with a mean BS of 8% and only four branches with BS above 60%. The coverage difference between our setting of 500 and the slight improvement seen with 1000 did not strongly justify switching to it, as the data at this point were already tractable and further simplification would result in some further loss in signal, as seen in the extreme case of 5000.

By changing the composition of the blocks, method settings also impact the final tree in substantial ways. An increased number of blocks tends to result in trees with lower branch lengths. For example, the tree inferred from relaxed annotations has less than half

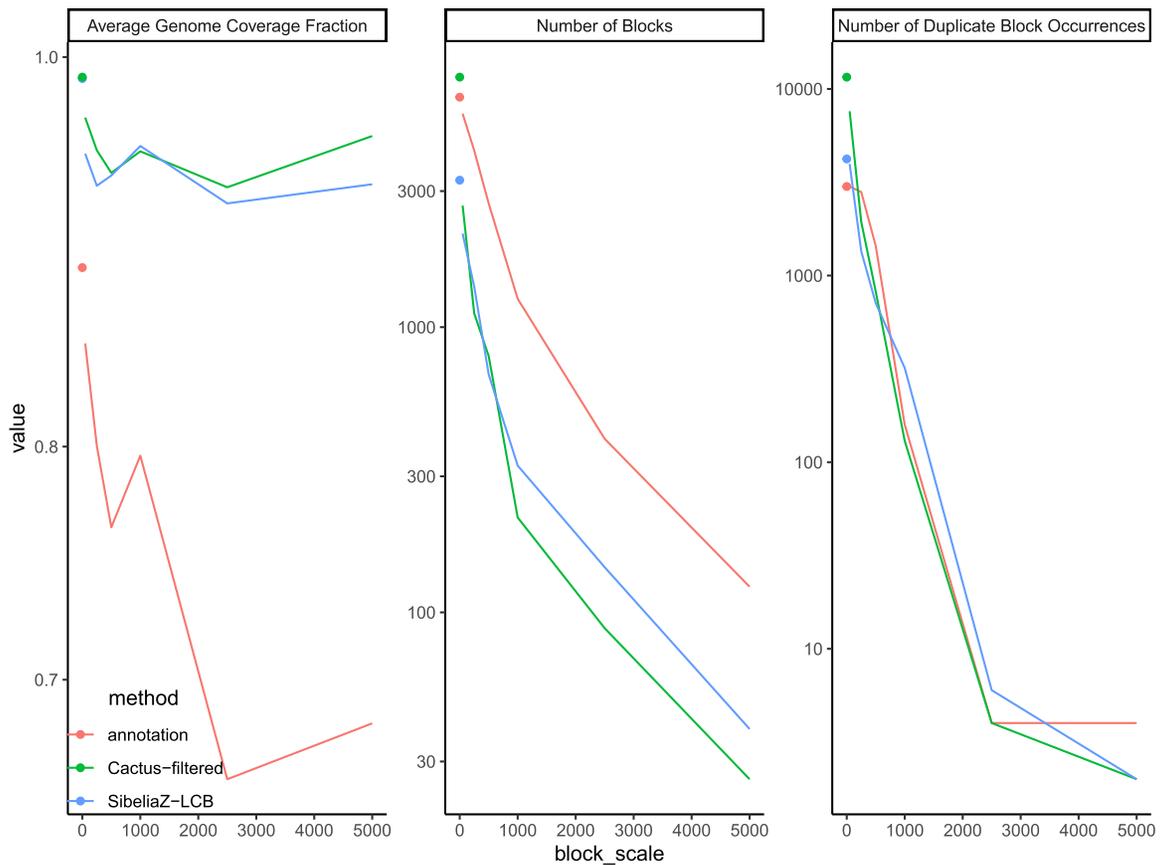


Fig. 2. maf2synteny parameter effects on synteny blocks. The effects of varying the—`block_sizes` parameter of maf2synteny (default 5000) on the average genome coverage fraction, resulting number of synteny blocks, and the number of duplicate occurrences. The isolated points represent the values for the raw markers. Only methods used in our subsequent phylogenetic analyses are shown here, representing a single run configuration for each tool used to identify the raw syntenic markers that maf2synteny acts upon

of the total branch length of the default annotations (0.053 versus 0.162), despite having only 20% fewer blocks. These changes are not necessarily surprising, because branch lengths are computed in the unit of mutations *per site*, which in this case can be interpreted as the number of changes *per adjacency*. When large blocks are divided into smaller blocks with little or no change in adjacencies, fewer changes will be observed per adjacency. Thus, the interpretation of branch lengths is very much tied to parameters.

3.3 Impact of block assignment method on trees

The choice of the block assignment methods had substantial impact on the resulting trees, especially among their less supported branches (Fig. 3). Tree resolution (i.e. the number of branches left after contracting low support branches) drops substantially at higher levels of support for rearrangement-based methods with the notable exception of Cactus-filtered without maf2synteny. The SNP tree has high resolution, with 93% mean BS and 63 (71) out of 91 branches having 100% ($\geq 95\%$) BS. Among the unaggregated blocks, the Cactus-filtered tree has the highest resolution, followed by SibeliaZ-LCB and annotation (mean BS: 98%, 84% and 79%, respectively). Moreover, block aggregation using maf2synteny substantially reduces resolution. Considering only the trees built from maf2synteny-aggregated blocks for each method, Cactus-filtered again has the highest, followed by annotation then SibeliaZ-LCB (mean BS: 82%, 79% and 77%, respectively).

Note that lower support should not be interpreted as less accuracy because if a method produces incorrect synteny blocks, the resulting tree can have high support for the wrong branches. A better measure of accuracy is the compatibility of trees with the reference trees. Taking the SNP tree as reference, we observe relatively high levels of compatibility with the reference tree among

adjacency-based methods and less so with the annotation-based tree (Fig. 3).

Compared to Cactus-raw, Cactus-filtered had similar compatibility with the SNP tree (Supplementary Fig. S1) and the added benefit of computational feasibility for running DING (the raw Cactus results in extremely large numbers of duplicate blocks that precluded running DING); so Cactus-filtered is exclusively used. The Cactus adjacency tree showed the greatest compatibility with the reference SNP tree, converging to perfect compatibility among branches with $\geq 75\%$ bootstrap support. Applying maf2synteny to it, however, substantially reduces compatibility. Perfect compatibility is not achieved except among branches with $\geq 95\%$ BS and this furthermore comes at a cost of much lower resolution.

Upon further examination of the Cactus adjacency tree (Fig. 4A), several patterns emerge. The Cactus adjacency tree is consistent with all assignments to standard lineages. Branches separating lineages are relatively long, especially for the East-Asian lineage (L2), separated by a branch of length 0.002 (i.e. 0.2% of block adjacencies have shifted in this clade). There is high BS (100%) for Indo-Oceanic (L1) to be the first to diverge from the rest, consistent with its classification as an ancestral lineage (Brosch et al., 2002; Gagneux, 2018). There is also strong support (100%) for uniting the East-African-Indian lineage (L3) with East-Asian. The diameter of the tree is 0.01, showing that 1% of adjacencies are different between the most divergent pair of isolates. The mean distance between any pair of sequences is 0.0058.

SibeliaZ-LCB's adjacency tree was less compatible than Cactus with the SNP tree before aggregation. It had four (five) incompatible branches at 100% BS ($\geq 95\%$) with the SNP tree. maf2synteny slightly improved its compatibility, leaving one (three) incompatible branch(es) at 100% BS ($\geq 95\%$), but at a cost of further reducing resolution of the tree at 100% BS (32% versus 59% resolution for

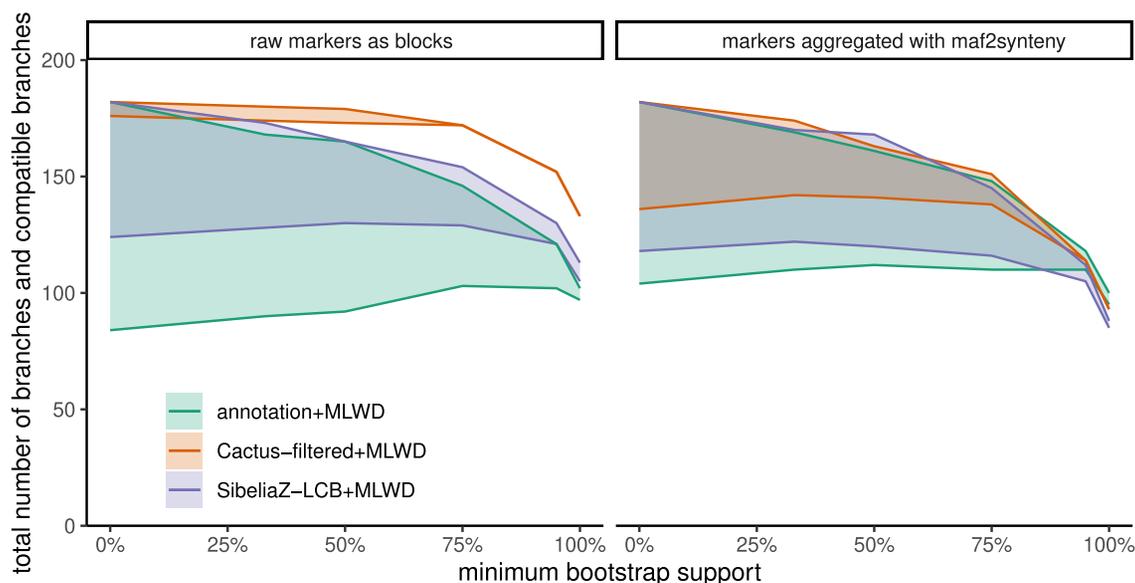


Fig. 3. Resolution and branch compatibility versus minimum bootstrap support in MLWD rearrangement trees. Two lines are plotted for each method. The upper lines represent the total number of branches in each tree and the reference SNP tree after contracting branches with support below a threshold (x-axis). The lower lines represent the total number of compatible branches in the two trees. Convergence of the two lines indicates perfect compatibility. The left panel uses trees based on the raw markers produced by the method, while the right uses those based on markers aggregated using maf2synteny

after and before maf2synteny, respectively). The deterioration in resolution of the alignment-based markers' adjacency trees following maf2synteny is suggestive of reduced signal as a result of merging of orthologous alignment blocks across the taxa. As for SibeliaZ-LCB-highVF (+maf2synteny), the result is strictly worse than SibeliaZ-LCB (+maf2synteny) in that both resolution and compatibility are lower (Supplementary Fig. S2). Furthermore, the high duplicity of blocks in SibeliaZ-LCB-highVF makes computing a distance matrix for them with DING infeasible. Because the use of a higher VF value resulted in no obvious improvement, we use the default version in the rest of these analyses.

Annotation also had substantial incompatibilities with the SNP tree (nine branches at 95% BS and two branches at 100% BS), which improved with maf2synteny to four incompatible branches at 95% BS and 3 three at 100% BS, with little loss in resolution compared to the unaggregated annotation. Comparing relaxed annotation and the default annotation, we did not observe a reduction in incompatibilities with the SNP tree (Supplementary Fig. S3), leading us to focus on the default annotation for the rest of the analyses. These high levels of fully supported incompatibilities in the raw annotation tree are consistent with it including strong but incorrect signal, perhaps as a result of incomplete orthology assignments resulting in missing adjacencies. The improvement in its compatibility with the SNP tree following maf2synteny may indicate that some of the incomplete and incorrect orthology assignments are masked by aggregation, making the final result a net improvement.

Combining complementary signals: Because the Cactus tree is fully compatible with the SNP tree at 75% BS, the two trees can easily be combined into a supertree (Fig. 4B) where every branch has $\geq 75\%$ BS in at least one of the two source trees. This supertree makes it clear that the SNP tree and the Cactus adjacency tree include complementary signals; while the two base trees have 85 and 87 branches with $\geq 75\%$ BS, the supertree has 90 out of 91 such branches (i.e. is 99% resolved). This supertree has all main lineages as monophyletic and the overall topology matches current understanding of Mtb's evolution (Gagneux, 2018): separating an East-African-Indian (L3) + East-Asian (L2) clade, first from Euro-American (L4), and then from the remaining lineages. The only remaining polytomy in the combined tree is between isolates 1-0156, SEA11020038, and the pair SEA11020092 and 1-0044. Even here, however, both the SNP and Cactus trees resolve it the same way, the former tree with BS 29% and the latter with BS 55%.

3.4 Impacts of guide tree

While the Cactus adjacency tree was highly compatible with the SNP tree, a caveat is that the guide tree used to infer the Cactus WGA is the SNP tree. Thus, to examine the impact of the Cactus guide tree on these results, we reran Cactus with two alternative guide trees: An independent one based on Mash distances (see Section 2.2.2) and the SibeliaZ-LCB+maf2synteny adjacency tree. We chose SibeliaZ-LCB+maf2synteny as one of the guide trees in order to test if Cactus results are biased toward the input guide tree. The results show that Cactus is indeed sensitive to the guide tree. Judging by matching split (MS) distance (Fig. 5), we see that the most similar trees to each other, by a large margin of at least 50 points, are those between a Cactus adjacency/distance tree and its guide tree or between the Cactus adjacency and distance trees based on the same alignment (explored more generally in Section 3.5). While the choice of the guide tree does not impact the amount of the resolution of the Cactus adjacency tree, it dramatically impacts the topology, especially for branches with lower resolution. While Cactus(SNP)-filtered shares 97% of the SNP tree branches, Cactus(Mash)-filtered shares 81%, and Cactus(SibeliaZ-LCB)-filtered only shares 64%. Similar patterns are observed at higher bootstrap support. Among branches with $\geq 75\%$ BS, while Cactus(SNP)-filtered had no incompatibility with the SNP tree, Cactus(SibeliaZ-LCB)-filtered has 31 such incompatible branches at $\geq 75\%$ BS level and even 12 incompatible branches at 100% BS. Similarly, while Cactus(SibeliaZ-LCB)-filtered has only one incompatibility with the SibeliaZ-LCB guide tree at $\geq 75\%$ BS, Cactus(SNP)-filtered has 16 incompatible branches with SibeliaZ-LCB at that level, 5 at $\geq 95\%$ BS, and 2 at 100% BS (Supplementary Fig. S4). Thus, each Cactus tree is most congruent with its own guide tree and quite incongruent with other trees, even at high support. In other words, the impact of the guide tree on the adjacency-based tree resulted in a bias toward the guide tree (as opposed to noisy variations). The fact that the impact of the guide tree extends to highly supported branches shows that the choice of the guide tree is critical.

3.5 Impact of tree inference method

Having established that the choice of methods of block assignment do indeed matter, we next ask whether the results are robust to the choice of the tree estimation method. As bootstrapping is not commonplace for

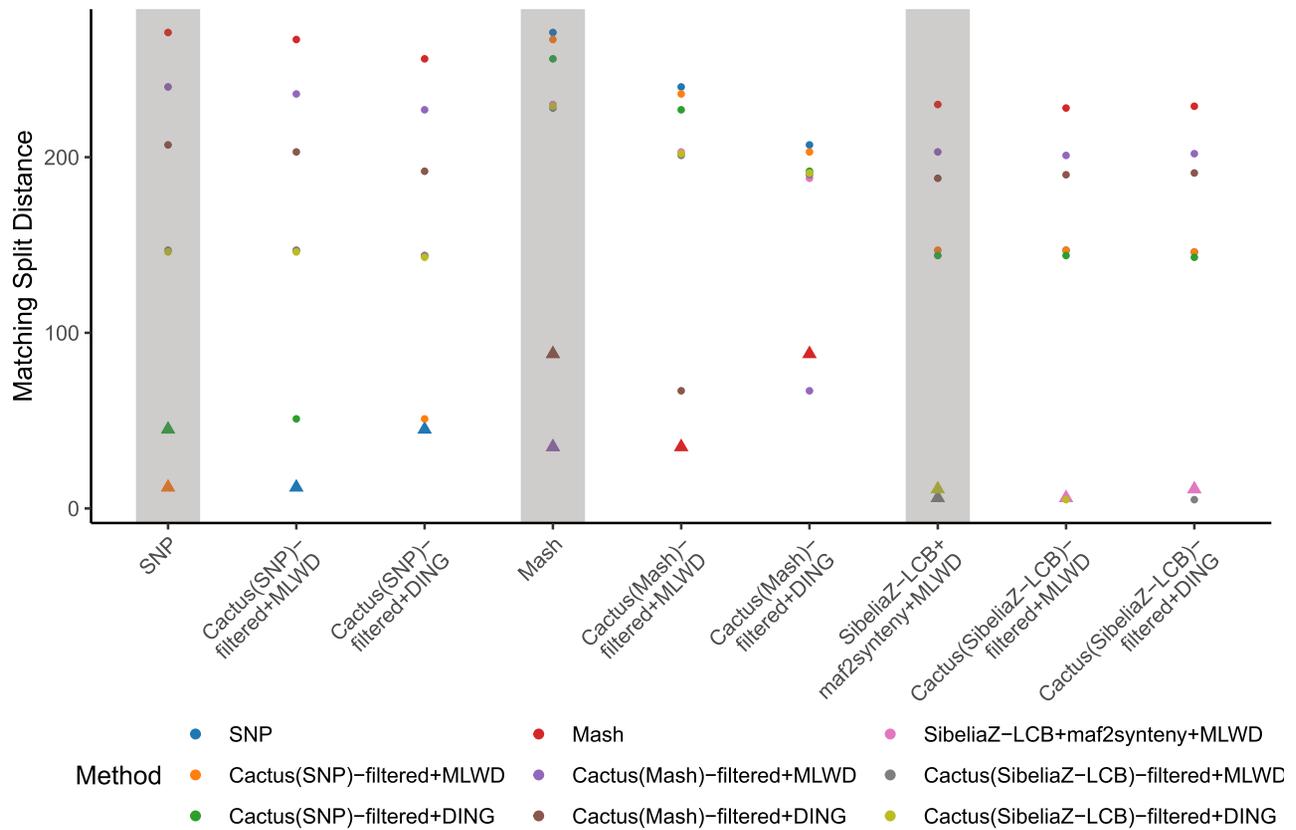


Fig. 5. Cactus whole-genome alignment is sensitive to the input guide tree. The plot shows matching split (MS) distances between rearrangement trees inferred from Cactus alignments based on three different guide trees: SNP, Mash and SibeliaZ-LCB. Each point corresponds to the MS distance between the method indicated on the x-axis and that denoted by the point's color. The MS distances between MLWD/DING trees and the corresponding guide tree are shown as triangular points. The most similar tree to a given MLWD/DING tree is invariably either the guide tree used for the underlying Cactus alignment itself, or the tree built using the same synteny blocks but the other inference method (i.e. *methodA*+MLWD and *methodA*+DING). Furthermore, there is a large difference between these and the remaining independent Cactus results. The use of SibeliaZ-LCB+maf2synteny+MLWD as one of the guide trees further shows how Cactus can, to an extent, mimic another method's alignment and thereby produce a similar rearrangement tree

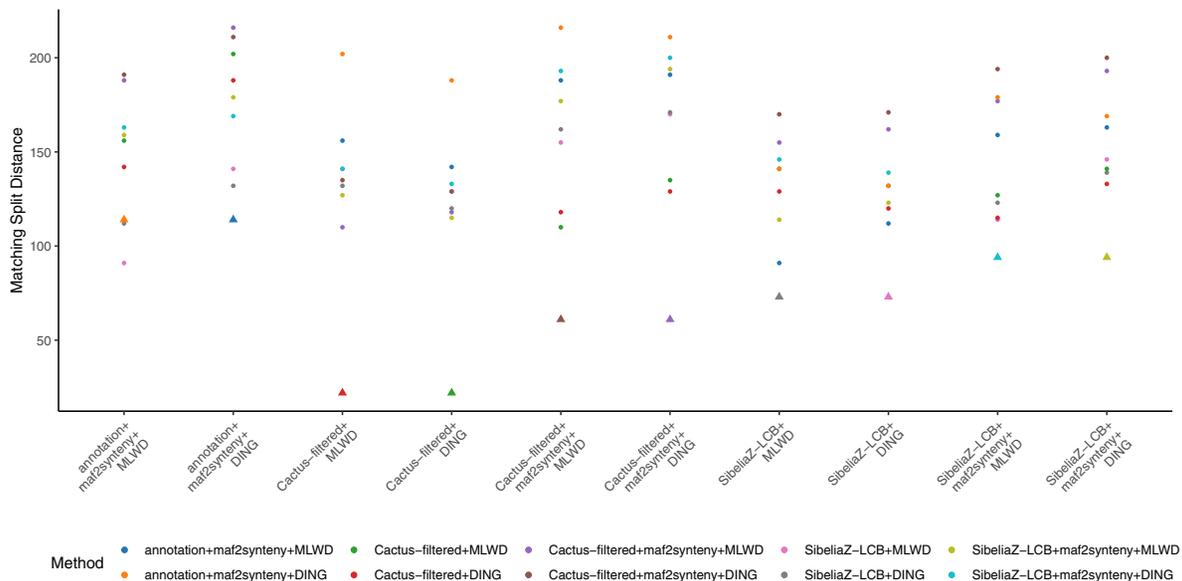


Fig. 6. Comparison of fully resolved MLWD and DING trees using Matching Split distance. As in Fig. 5, each point corresponds to the matching split distance between the method indicated at the x position and that denoted by the point's color. The MS distances between trees built from the same underlying synteny blocks but using a different tree inference method are shown as triangular points. In the majority of cases, the trees closest to each other are in fact these pairs, with the exception of annotation+maf2synteny+MLWD having greater similarity to both trees from SibeliaZ-LCB+maf2synteny than to annotation+maf2synteny+DING. A comparison of annotation+MLWD to annotation+DING (without maf2synteny) cannot be shown due to the computational infeasibility of computing the distance matrix for the latter

using the same inference method applied to the same synteny blocks differed from as little as 4% RF, 22 MS (Cactus-filtered) to 29% RF, 94 MS (SibeliaZ+maf2synteny). The only case where the tree inference method mattered more than the block determination method was annotation+maf2synteny+MLWD, where this tree was slightly closer to both SibeliaZ-LCB+MLWD (31% RF, 91 MS) and SibeliaZ-LCB+DING (29% RF, 112 MS) than to its partner tree annotation+maf2synteny+DING (27% RF, 114 MS).

Interestingly, the MS distances of DING trees (annotation+maf2synteny+DING, SibeliaZ-LCB+maf2synteny+DING and Cactus+DING, both with and without maf2synteny) to the SNP tree did tend to be higher than those of the corresponding trees with MLWD, indicating a potential advantage of MLWD as a tree-inference method. For SibeliaZ-LCB (unaggregated), however, the MLWD tree had a slightly higher MS distance than the corresponding DING tree (161 versus 158) (Supplementary Fig. S6).

4 Discussion

As inferring phylogenies based on large-scale mutations becomes increasingly more feasible in terms of data availability, many questions about the best practices remain unanswered. Before these reconstruction methods are more broadly adopted, we need more empirical analyses that guide the practitioners in building robust and reliable analysis pipelines. In this article, we took a step in that direction. Using a dataset of high-quality *Mtb* genomes, we interrogated the robustness of methods used for preparing the syntenic blocks, which form the input to methods that infer phylogenies using rearrangements.

The space of possible methods for obtaining input to rearrangement phylogenies is wide and this study is necessarily incomplete. While we have made an effort to represent the most modern WGA-based strategies for block assignment, many alternatives exist and future work should explore them.

While we considered two sets of parameters for orthology assignment of the annotation blocks, we did not consider alternative assignment methods (Linard et al., 2021), and the full phylogenetic impact of the different methods for annotation, orthology detection and post-processing of orthology, remains a daunting task. Additionally, we limited our study to the use of maf2synteny for the agglomeration of basic homology statements into syntenic blocks. Other methods such as i-ADHoRe3.0 (Proost et al., 2012) and Cyntenator (Rödelsperger and Dieterich, 2010) exist for this purpose. Unfortunately, we were unsuccessful in running each of these on our dataset. i-ADHoRe3.0 has an array of parameters and, using the recommended values, we received empty output. Running Cyntenator with default parameters produced only 2 synteny blocks. After adjusting parameters and rerunning, Cyntenator did not complete after a week of computation.

Cactus' requirement of a guide tree to determine the order of pairwise alignment became an important variable for the subsequent inference of the tree from the alignment. The yet-unpublished Cactus Pangenome Pipeline is presented as a solution to the problem of guide trees, though rather than a guide tree, it requires specifying a reference genome. This option should be explored as an alternative solution.

To combine methods inferred from two types of data, we relied on the fact that they were fully compatible for highly supported branches. Such simple supertree methods facilitate interpretation, as one knows each branch has support from at least one data source and no strong conflict from the other. Nevertheless, full compatibility will not always be achieved, necessitating more complex procedures for obtaining supertrees (Bininda-Emonds, 2004), which may reduce interpretability. An arguably more principled approach for combining the phylogenetic signals is to combine the data and infer one tree from the entirety of the data. This combination, however, is not trivial. We can perhaps concatenate data and use data partitioning, but proper weighting of the data sources is not obvious. Such approaches need to be further explored.

In terms of tree inference methods, the aim of this study was not to exhaustively test and compare methods. Rather, we chose two

methods to demonstrate the impact of synteny block formation on different families of methods. Many more methods for phylogenetic inference from syntenic blocks exist (e.g. Drillon et al., 2020; Zhao et al., 2021). While comparing these methods is beyond the scope of this work, it seems reasonable to expect that similarly to the two methods that we did test, they would also show sensitivity to block formation strategies.

It should also be noted that our dataset consists of primarily *Mtb* isolates, chosen partially because the vertical evolution of *Mtb* is advantageous for our purposes. Thus, our conclusions may be specific to closely-related or microbial genomes without many horizontal events. Alignment difficulty is clearly a function of divergence, and as the divergence levels increase, various aligners may be impacted unevenly. In fact, it is possible that, for very divergent species, annotation works better than alignment. Moreover, the choice of best parameters for aligners and annotators will likely depend on the divergence levels. Thus, additional future studies on more datasets are needed to test whether the patterns observed here are replicated across longer evolutionary times and more complex evolutionary scenarios.

Finally, method evaluation on real data, while free of concern about realism of the data, is complicated by a lack of access to the ground truth. Realistic genome simulation remains challenging; options that do exist tend to have important shortcomings such as lack of user support, bugs, and limited features. We require a method that supports genome-scale rearrangement as well as gene-scale substitutions, a model for intergenic nucleotide evolution, and the specification of a real genome at the root of the simulated species tree. Most existing methods lack some of these features. For example, the program ALF (Dalquen et al., 2011) supports genome-scale rearrangement, substitutions, and specification of a root genome, but does not model intergenic nucleotide evolution. As far as we know, Evolver (<https://www.drive5.com/evolver>) is the only option that has the complete feature set, but the software is complex to set up, was designed specifically for eukaryotic genomes, and is no longer supported by its authors. Once improved methods for genome simulation are available, repeating our analyses will allow a more direct measurement of accuracy, at every step of the pipeline.

5 Conclusions

Our results both create cause for caution and room for optimism. It would be ideal if the inferred phylogenies were robust to the block inference and reconstruction methods. Instead, we saw that the choice of the block assignment method matters, but also parameter settings of the methods can impact results. The block composition seems sensitive to the parameters of alignment, on one hand, and parameters of the methods used to group alignment segments into larger groups (e.g. maf2synteny), on the other. The block properties, in turn, impact the phylogeny. Furthermore, some parameters, such as the guide tree used for WGA, do not impact block composition in obvious ways but impact the final tree in substantial ways. Thus, practitioners are encouraged to remain cautious about these choices, and our results call for more extensive empirical analyses.

The apparent importance of the choice of synteny block construction method on the downstream phylogenetic inference should not surprise us. Traditional phylogenetics has long wrestled with impacts of incorrect homology detection and alignment on tree inference (Li-San Wang et al., 2011; Liu et al., 2009; Lunter et al., 2008; Ogden and Rosenberg, 2006; Philippe et al., 2017; Springer and Gatesy, 2018), and it would be naive to expect rearrangement phylogenies would be spared that concern. In fact, some of our findings are analogous to similar observations for multiple sequence alignment. For example, the impact of guide trees on final adjacency trees reminds one of the impact of guide trees on multiple sequence alignments and resulting trees (Boyce et al., 2014; Lake, 1991; Nelesen et al., 2008).

Beyond parameters, our results provide reasons to prefer some methods versus others. Compared to using gene annotations, the alignment-based methods showed superior compatibility with the reference SNP tree and with each other. In contrast, the annotation-

based tree has numerous incompatibilities, even among branches with 100% bootstrap support, with the SNP tree. The large number of blocks in the annotation input is cause for concern, and not greatly improved by further aggregation. It appears that annotation pipelines, even run with relaxed settings, fail to make complete orthology calls in our dataset. Such failures to find orthology can easily lead to inconsistent adjacencies across genomes, and perhaps high support for wrong branches. Thus, our results support the use of WGA.

Despite the variability that we observed, some encouraging patterns emerged. The Cactus adjacency tree did have high levels of compatibility and a *complementary* signal to the SNP tree, allowing us to combine their highly supported branches into a supertree with more resolution than either tree. It is true that the compatibility is helped by the choice of the SNP tree as the guide, complicating the interpretation of compatibility. Nevertheless, if we are not using the SNP tree for benchmarking, this compatibility shows a path forward. Practitioners can infer preliminary substitution-based trees, using reference genomes and infer a WGA using those as guide tree. The WGA can then be used both with adjacency-based and substitution-based models to infer alternative trees, which when compatible, can easily and unambiguously be combined. When this process is successful, as it was here, the long-standing goal of combining signal from two types of events is achieved.

Acknowledgements

The authors thank Jijun Tang for providing us with the source code for MLGO to allow us to run it locally.

Author contributions

A.E. wrote the analysis code and collected the data. S.M., A.E. and K.M.S. designed the study. A.E., S.M., K.M.S. and F.V. analyzed the data, wrote the article and approved the final version.

Funding

This work was funded through a grant (R01AI105185) by the National Institute for Allergy and Infectious Diseases (NIAID) awarded to F.V. and National Science Foundation (NSF) [III-1845967 to S.M.]. The funding body had no role in the design of the study or in collection, analysis and interpretation of data or in writing the article.

Competing Interests

The authors declare that they have no competing interests.

Data availability

Sequences used in this study are available from NCBI Bioprojects PRJNA555636, PRJEB8783 and PRNJA555636. Analysis scripts and code written for this study are available at <https://gitlab.com/LPCDRP/rearrangement-homology.pub> and <https://gitlab.com/LPCDRP/syntement>.

References

Armstrong, J. *et al.* (2019) Whole-genome alignment and comparative annotation. *Annu. Rev. Anim. Biosci.*, **7**, 41–64.
 Armstrong, J. *et al.* (2020) Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, **587**, 246–251.
 Berney, M. *et al.* (2015) Essential roles of methionine and S-adenosylmethionine in the autarkic lifestyle of *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA*, **112**, 10008–10013.
 Bernt, M. *et al.* (2007) CREx: inferring genomic rearrangements based on common intervals. *Bioinformatics*, **23**, 2957–2958.

Billar, P. *et al.* (2016) Comparative genomics on artificial life. In: *Conference on Computability in Europe*. Springer, Paris, France. pp. 35–44. <https://link.springer.com/book/10.1007/978-3-319-40189-8>.
 Bininda-Emonds, O.R.P. (ed.) (2004) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, Vol. 4. Kluwer Academic Publishers, Dordrecht, The Netherlands. ISBN 140203294.
 Bogdanowicz, D. and Giaro, K. (2012) Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **9**, 150–160.
 Bogdanowicz, D. *et al.* (2012) TreeCmp: comparison of trees in polynomial time. *Evol. Bioinform. Online*, **8**, EBO.S9657.
 Bohnenkämper, L. *et al.* (2020) Computing the rearrangement distance of natural genomes. In: Schwartz, R. (ed.) *Research in Computational Molecular Biology, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 3–18. ISBN 978-3-030-45257-5. doi: [10.1007/978-3-030-45257-5_1](https://doi.org/10.1007/978-3-030-45257-5_1).
 Bourque, G. and Pevzner, P.A. (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.*, **12**, 26–36.
 Boyce, K. *et al.* (2014) Simple chained guide trees give high-quality protein multiple sequence alignments. *Proc. Natl. Acad. Sci. USA*, **111**, 10556–10561.
 Brosch, R. *et al.* (2002). A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl. Acad. Sci. USA*, **99**, 3684–3689.
 Chen, Q. *et al.* (2016) Evaluation of CD-HIT for constructing non-redundant databases. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 703–706. doi: [10.1109/BIBM.2016.7822604](https://doi.org/10.1109/BIBM.2016.7822604).
 Chin, C.-S. *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, **10**, 563–569.
 Coscolla, M. and Gagneux, S. (2014) Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin. Immunol.*, **26**, 431–444.
 Dalquen, D.A. *et al.* (2012) ALF—a simulation framework for genome evolution. *Mol. Biol. Evol.*, **29**, 1115–1123.
 Davin, A.A. *et al.* (2020) Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead lineages. *Bioinformatics*, **36**, 1286–1288.
 Doerr, D. and Chauve, C. (2021) Small parsimony for natural genomes in the DCJ-indel model. *J. Bioinform. Comput. Biol.*, **19**, 2140009.
 Doerr, D. *et al.* (2018) Family-free genome comparison. In: Setubal, J.C., Stoye, J., and Stadler, P.F. (ed.) *Comparative Genomics: Methods and Protocols, Methods in Molecular Biology*. Springer, New York, NY, pp. 331–342. ISBN 978-1-4939-7463-4. doi: [10.1007/978-1-4939-7463-4_12](https://doi.org/10.1007/978-1-4939-7463-4_12).
 Drillon, G. *et al.* (2020) Phylogenetic reconstruction based on synteny block and gene adjacencies. *Mol. Biol. Evol.*, **37**, 2747–2762.
 Earl, D. *et al.* (2014) Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res.*, **24**, 2077–2089.
 Elghraoui, A. *et al.* (2022) Hybran: Hybrid reference transfer and ab initio prokaryotic genome annotation. *bioRxiv*. doi: [10.1101/2022.11.09.515824](https://doi.org/10.1101/2022.11.09.515824).
 English, A.C. *et al.* (2014) PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*, **15**, 180.
 Enright, A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
 Feijao, P. and Meidanis, J. (2011) SCJ: a breakpoint-like distance that simplifies several rearrangement problems. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 1318–1329.
 Feng, B. *et al.* (2017) Reconstructing yeasts phylogenies and ancestors from whole genome data. *Sci. Rep.*, **7**, 15209.
 Fertin, G. (ed.) (2009) *Combinatorics of Genome Rearrangements. Computational Molecular Biology*. MIT Press, Cambridge, MA. ISBN 978-0-262-06282-4.
 Fu, L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
 Gagneux, S. (2018) Ecology and evolution of *Mycobacterium tuberculosis*. *Nat. Rev. Microbiol.*, **16**, 202–213.
 Galagan, J.E. (2014) Genomic insights into tuberculosis. *Nat. Rev. Genet.*, **15**, 307–320.
 Ghiurcata, C.G. and Moret, B.M.E. (2014) Evaluating synteny for improved comparative studies. *Bioinformatics*, **30**, i9–i18.
 Gordon, S.V. *et al.* (1999) Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol. Microbiol.*, **32**, 643–655.
 Gray, T.A. and Derbyshire, K.M. (2018) Blending genomes: distributive conjugal transfer in mycobacteria, a sexier form of HGT. *Mol. Microbiol.*, **108**, 601–613.

- Hillery, N. et al. (2014) The Global Consortium for Drug-Resistant Tuberculosis Diagnostics (GCDD): design of a multi-site, head-to-head study of three rapid tests to detect extensively drug-resistant tuberculosis. *Trials*, **15**, 434.
- Hindré, T. et al. (2012) New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nat. Rev. Microbiol.*, **10**, 352–365.
- Hu, F. et al. (2014) MLGO: phylogeny reconstruction and ancestral inference from gene-order data. *BMC Bioinformatics*, **15**, 354.
- Hunt, M. et al. (2015) Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.*, **16**, 294.
- Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In: Munro, H.N. (ed) *Mammalian protein Metabolism*, Vol. III. Academic Press, Inc., New York, New York. pp. 21–132.
- Kanduma, E. et al. (2003) Molecular methods for *Mycobacterium tuberculosis* strain typing: a users guide. *J. Appl. Microbiol.*, **94**, 781–791.
- Kolmogorov, M. et al. (2018) Chromosome assembly of large and complex genomes using multiple references. *Genome Res.*, **28**, 1720–1732.
- Koren, S. and Phillippy, A.M. (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.*, **23**, 110–120.
- Koren, S. et al. (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.*, **14**, R101.
- Koren, S. et al. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.
- Kurtz, S. et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
- Lake, J.A. (1991) The order of sequence alignment can bias the selection of tree topology. *Mol. Biol. Evol.*, **8**, 378–385.
- Lefort, V. et al. (2015) FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.*, **32**, 2798–2800.
- Li, W. and Godzik, A. (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
- Liao, Y. et al. (2021) Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*. *Genome Res.*, **31**, 397–410.
- Lin, Y. et al. (2011) Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator. *J. Comput. Biol.*, **18**, 1131–1139.
- Lin, Y. et al. (2012a) Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. In: *Biocomputing 2013*. WORLD SCIENTIFIC, pp. 285–296. ISBN 978-981-4596-36-7. doi: [10.1142/9789814447973_0028](https://doi.org/10.1142/9789814447973_0028).
- Lin, Y. et al. (2012b) Bootstrapping phylogenies inferred from rearrangement data. *Algorithms Mol. Biol.*, **7**, 21.
- Linard, B. et al.; QFO Consortium. (2021) Ten years of collaborative progress in the quest for orthologs. *Mol. Biol. Evol.*, **38**, 3033–3045.
- Li-San Wang, L.-S. et al. (2011) The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 1108–1119.
- Liu, K. et al. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, **324**, 1561–1564.
- Liu, P. et al. (2012) Mechanisms for recurrent and complex human genomic rearrangements. *Curr. Opin. Genet. Dev.*, **22**, 211–220.
- Lucas, J.M. and Roest Crolius, H. (2017) High precision detection of conserved segments from synteny blocks. *PLoS One.*, **12**, e0180198.
- Lunter, G. et al. (2008) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, **18**, 298–309.
- Merkler, M. et al. (2015) Evolutionary history and global spread of the *mycobacterium tuberculosis* Beijing lineage. *Nat Genet.*, **47**, 242–249.
- Merrick, C.N. and Merrick, H. (2018) Gene inversion potentiates bacterial evolvability and virulence. *Nat. Commun.*, **9**, 4662.
- Miga, K.H. et al. (2019) Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, **585**(7823), 79–84. doi: [10.1011/735928](https://doi.org/10.1011/735928).
- Minkin, I. and Medvedev, P. (2020) Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nat. Commun.*, **11**, 6327.
- Modlin, S.J. et al. (2020) Drivers and sites of diversity in the DNA adenine methylomes of 93 *Mycobacterium tuberculosis* complex clinical isolates. *eLife*, **9**, e58542.
- Moret, B.M.E. et al. (2002) Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, **65**, 508–525.
- Moret, B.M.E. et al. (2013) Rearrangements in phylogenetic inference: compare, model, or encode? In: Chauve, C., El-Mabrouk, N., and Tannier, E. (eds.) *Models and Algorithms for Genome Evolution, Computational Biology*. Springer London, London, pp. 147–171. ISBN 978-1-4471-5298-9. doi: [10.1007/978-1-4471-5298-9_7](https://doi.org/10.1007/978-1-4471-5298-9_7).
- Mostowy, S. et al. (2002) Genomic deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. *J. Infect. Dis.*, **186**, 74–80.
- Nelesen, S. et al. (2008) The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. In: *Pacific Symposium on Biocomputing*, pp. 25–36. ISSN 2335-6928 (Print). doi: [10.1142/9789812776136{_}0004](https://doi.org/10.1142/9789812776136{_}0004).
- Nurk, S. et al. (2021) The complete sequence of a human genome. *Science*, **376**(6588), 44–53. doi: [10.1101/2021.05.26.445798](https://doi.org/10.1101/2021.05.26.445798).
- Ogden, T.H. and Rosenberg, M.S. (2006) Multiple sequence alignment accuracy and phylogenetic inference. *Syst. Biol.*, **55**, 314–328.
- Ondov, B.D. et al. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
- Otto, T.D. et al. (2011) RATT: rapid annotation transfer tool. *Nucleic Acids Res.*, **39**, e57.
- Pe'er, I. and Shamir, R. (1998) The median problems for breakpoints are NP-complete. *Elec. Colloq. Comput. Complexity*, **71**, 1–16.
- Pevzner, P. and Tesler, G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.*, **13**, 37–45.
- Phelan, J.E. et al. (2019) Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.*, **11**, 41.
- Philippe, H. et al. (2017) Pitfalls in supermatrix phylogenomics. *EJT*. **283**, 1–25. ISSN 2118-9773. doi: [10.5852/ejt.2017.283](https://doi.org/10.5852/ejt.2017.283).
- Phillippy, A.M. (2017) New advances in sequence assembly. *Genome Res.*, **27**, xi.
- Proost, S. et al. (2012) i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.*, **40**, e11.
- Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
- Ranz, J.M. et al. (2007) Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.*, **5**, e152.
- Rhie, A. et al. (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737–746.
- Robinson, D. and Foulds, L. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.
- Rödelsperger, C. and Dieterich, C. (2010) CYTENATOR: progressive gene order alignment of 17 vertebrate genomes. *PLoS One*, **5**, e8861.
- Salzberg, S.L. (2019) Next-generation genome annotation: we still struggle to get it right. *Genome Biol.*, **20**, 92.
- Sankoff, D. and Blanchette, M. (1998) Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, **5**, 555–570.
- Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
- Springer, M.S. and Gatesy, J. (2018) On the importance of homology in the age of phylogenomics. *Syst. Biodivers.*, **16**, 210–228.
- Stamatakis, A. (May 2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313. ISSN 1367-4803. doi: [10.1093/bioinformatics/btu033](https://doi.org/10.1093/bioinformatics/btu033).
- Sukumaran, J. and Holder, M.T. (2010) DendroPy: a python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.
- Uplekar, S. et al. (2011) Comparative genomics of Esx genes from clinical isolates of *Mycobacterium tuberculosis* provides evidence for gene conversion and epitope variation. *Infect. Immun.*, **79**, 4042–4049.
- Wang, L.-S. et al. (2002) Fast phylogenetic methods for the analysis of genome rearrangement data: an empirical study. In: *Pacific Symposium on Biocomputing*, pp. 524–535. ISSN 2335-6928 (Print).
- Wang, L.-S. et al. (2006) Distance-based genome rearrangement phylogeny. *J. Mol. Evol.*, **63**, 473–483.
- Xu, A.W. (2010) The median problems on linear multichromosomal genomes: graph representation and fast exact solutions. *J. Comput. Biol.*, **17**, 1195–1211.
- Yu, G. et al. (2017) ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*, **8**, 28–36.
- Zhao, T. et al. (2021) Whole-genome microsynteny-based phylogeny of angiosperms. *Nat. Commun.*, **12**, 3498.