Rapid Core-Genome Alignment and Visualization for Thousands of Microbial Genomes

Todd J. Treangen[1§*], Brian D. Ondov[1*], Sergey Koren[1], Adam M. Phillippy[1]

[1] National Biodefense Analysis and Countermeasures Center, 110 Thomas Johnson Drive, Frederick, MD 21702

[*] These authors contributed equally to this work

[§] Corresponding author

Email addresses:

TJT: treangent@nbacc.net

BDO: ondovb@nbacc.net

SK: korens@nbacc.net

AMP: phillippya@nbacc.net

Abstract

Though many microbial species or clades now have hundreds of sequenced genomes, existing whole-genome alignment methods do not efficiently handle comparisons on this scale. Here we present the Harvest suite of core-genome alignment and visualization tools for quickly analyzing thousands of intraspecific microbial strains. Harvest includes Parsnp, a fast core-genome multi-aligner, and Gingr, a dynamic visual platform. Combined they provide interactive core-genome alignments, variant calls, recombination detection, and phylogenetic trees. Using simulated and real data we demonstrate that our approach exhibits unrivaled speed while maintaining the accuracy of existing methods. The Harvest suite is open-source and freely available from: http://github.com/marbl/harvest.

Key Words

Genome alignment, phylogenomics, microbial genomics, SNPs, information visualization

Rationale

Microbial genomes represent over 93% of past sequencing projects, with the current total over ten thousand and growing exponentially. Multiple clades of draft and complete genomes comprising hundreds of closely related strains are now available from public databases [1], largely due to an increase in sequencing-based outbreak studies [2]. The quality of future genomes is also set to improve as short-read assemblers mature [3] and long-read sequencing enables finishing at greatly reduced costs [4, 5].

One direct benefit of high-quality genomes is that they empower comparative genomic studies based on multiple genome alignment. Multiple genome alignment is a fundamental tool in genomics essential for tracking genome evolution [6-8], accurate inference of recombination [9-14], identification of genomic islands [15, 16], analysis of mobile genetic elements [17, 18], comprehensive classification of homology [19, 20], ancestral genome reconstruction [21], and phylogenomic analyses [22-24]. The task of whole-genome alignment is to create a catalog of relationships between the sequences of each genome (ortholog, paralog, xenolog, etc. [25]) to reveal their evolutionary history [26, 27]. While several tools exist (LS-BSR [28], Magic[29], Mavid[30], Mauve [31-33], MGA [34], M-GCAT [35], Mugsy [36], TBA [37], multi-LAGAN [38], PECAN [39]), multiple genome alignment remains a challenging task due to the prevalence of horizontal gene transfer [40, 41], recombination, homoplasy, gene conversion, mobile genetic elements, pseudogenization, and convoluted orthology relationships [25]. In addition, the computational burden of multiple sequence alignment remains very high [42] despite recent progress [43].

The current influx of microbial sequencing data necessitates methods for large-scale comparative genomics and shifts the focus towards scalability. Current microbial genome alignment methods focus on all-versus-all progressive alignment [31, 36] to detect subset relationships (*i.e.* gene gain/loss), but these methods are bounded at various steps by quadratic time complexity. This exponential growth in compute time prohibits comparisons involving thousands of genomes. Chan and Ragan [44] reiterated this point, emphasizing that current phylogenomic methods, such as multiple alignment, will not scale with the increasing number of genomes, and that

“alignment-free” or exact alignment methods must be used to analyze such datasets. However, such approaches do not come without compromising phylogenetic resolution [45].

Core-genome alignment is a subset of whole-genome alignment, focused on identifying conserved orthologous regions contained in all aligned genomes. In contrast to the exponential complexity of multiple alignment, core-genome alignment is inherently more scalable because it ignores subset relationships. In addition, the core genome contains essential genes that are often vertically inherited and most likely to have the strongest signal-to-noise ratio for inferring phylogeny. The most reliable variants for building such phylogenies are single-nucleotide polymorphisms (SNPs). Thus, core-genome SNP typing is currently the standard method for reconstructing large phylogenies of closely related microbes [46]. Currently, there are three paradigms for core-genome SNP typing based on read mapping, k-mer analyses, and whole-genome alignment.

Read-based methods have dominated the bioinformatics methods landscape since the invention of high-fidelity, short-read sequencing (50-300bp) [47]. This has made it very affordable to sequence, yet extremely challenging to produce finished genomes [48, 49]. Thus, comparative genomics has turned to highly efficient and accurate read mapping algorithms to carry out assembly-free analyses, spawning many mapping tools [50-53] and variant callers [54-56] for detecting SNPs and short Indels. Read-based variant calling typically utilizes a finished reference genome and a sensitive read mapper (BWA [52], Smalt), variant caller (samtools/bcftools [56], GATK [54]), and variant filter (minimum mapping quality, core genomic regions). This method has

been shown to be effective in practice [57], effectively bypassing the need for *de novo* assembly in this context. However, mapping requires the read data, which is not always available and can be orders of magnitude larger than the genomes themselves. In addition, mapping can be sensitive to contaminant, overlook structural variation, misalign low-complexity and repetitive sequence, and introduce systematic bias in phylogenetic reconstruction [58-60].

Exact alignment methods, often formulated as k-mer matching, can produce high precision results in a fraction of the time required for gapped alignment methods [61-63]. Spectral k-mer approaches have been used to estimate genome similarity [64], and k-mer based methods are commonly used to identify or cluster homologous genomic sequence [65, 66]. Recently, k-mers have also been extended to SNP identification. kSNP [67] identifies odd-length k-mers between multiple samples that match at all but the central position. The matched k-mers are then mapped back to a reference genome to locate putative SNPs. Conveniently, this approach is suitable for both assembled genomes and read sets, but sensitivity is sacrificed for the improved efficiency of exact alignment [68].

Genome assembly [69-77], followed by whole-genome alignment [38, 78, 79], is the original method for variant detection between closely related bacterial genomes [80] and has been shown to perform well across multiple sequencing platforms [81]. In addition to SNPs, whole-genome alignment is able to reliably identify insertions and deletions (Indels) and other forms of structural variation. Thus, whole-genome alignment is the gold standard for comprehensive variant identification, but relies on highly accurate and continuous assemblies, which can be expensive to generate.

Lastly, and unlike reference mapping, whole-genome alignment is not easily parallelized or scaled to many genomes.

Specifically for the task of whole-genome SNP typing, the choice of read- or genome-based methods can often depend on data availability. For example, of the 24,000 bacterial genomes currently in NCBI RefSeq [82], only 55% have associated SRA read data and analysis of the remaining 45% requires genome-based methods. Thankfully, recent advances in both sequencing technology and assembly algorithms are making microbial genomes more complete than ever before. Modern *de Bruijn* assemblers like SPAdes [83]are able to generate high-quality assemblies from short reads [3], and long read technologies have enabled the automated finishing of microbial genomes for under $1,000 [84]. With the number of publically available genomes currently doubling every 18 months [1], and genome quality improving with the arrival of new technologies, we set out to solve the problem of aligning thousands of closely-related whole genomes.

Rapid Core-Genome Alignment and Visualization

Here we present Parsnp and Gingr for the construction and interactive visualization of massive core-genome alignments. For alignment, Parsnp combines the advantages of both whole-genome alignment and read mapping. Like whole-genome alignment, Parsnp accurately aligns assembled genomes to identify both structural and point variations, but like read mapping, Parsnp scales to thousands of closely related genomes. To achieve this scalability, Parsnp is based on a suffix graph data structure for the rapid identification of maximal unique matches (MUMs), which serve as a common foundation to many pairwise [78, 79, 85] and multiple genome alignment tools [31-36]. Parsnp uses MUMs to both recruit similar genomes and anchor the

multiple alignment. As input, Parsnp takes a directory of MultiFASTA files to be aligned; and as output, Parsnp produces a core-genome alignment, variant calls, and a SNP tree. These outputs can then be visually explored using Gingr. The details of Parsnp and Gingr are described below.

MUMi recruitment

Parsnp is designed for intraspecific alignments and requires input genomes to be highly similar (e.g. within the same subspecies group or >=97% average nucleotide identity). For novel genomes or an inaccurate taxonomy, which genomes meet this criterion is not always known. To automatically identify genomes suitable for alignment, Parsnp uses a recruitment strategy based on the MUMi distance [86]. Only genomes within a specified MUMi distance threshold are recruited into the full alignment.

Compressed suffix graph

Parsnp utilizes a Directed Acyclic Graph (DAG) data structure, called a Compressed Suffix Graph (CSG), to index the reference genome for efficient identification of multi-MUMs. CSGs have the unique property of representing an optimally compressed structure, in terms of number of nodes and edges, while maintaining all intrinsic properties of a Suffix Tree. CSGs were originally proposed as a more space-efficient alternative to Suffix Trees and first implemented in M-GCAT [35]. Node and edge compression of the Suffix Tree incurs a linear-time construction penalty, but facilitates faster traversal of the structure once built. Note that because multi-MUMs are necessarily present in all genomes, the choice of a reference genome has no effect on the resulting alignment.

Multi-MUM search

Once built for the reference genome, all additional genomes are streamed through the CSG, enabling rapid, linear-time identification of MUMs shared across all genomes. A divide-and-conquer algorithm, adapted from M-GCAT [35], recursively searches for smaller matches and iteratively refines the multi-MUMs. Next, locally collinear blocks (LCBs) of multi-MUMs are identified. These LCBs form the basis of the core-genome alignment.

Parallelized LCB alignment

The multi-MUMs within LCBs are used to anchor multiple alignments. Gaps between collinear multi-MUMs are aligned in parallel using MUSCLE [87]. To avoid the unnecessary overhead of reading and writing MultiFASTA alignment files, Parsnp makes direct library calls via a MUSCLE API. The MUSCLE library is packaged with Parsnp, but originally sourced from the Mauve code base [88]. For highly similar genomes, the gaps between collinear multi-MUMs are typically very short (e.g. a single SNP column in the degenerate case).

SNP filtering and trees

The final Parsnp multiple alignment contains all SNP, Indel, and structural variation within the core genome. However, given their ubiquity in microbial genome analyses, Parsnp performs additional processing of the core-genome SNPs. First, all polymorphic columns in the multiple alignment are flagged to identify: (i) repetitive sequence, (ii) small LCB size, (iii) poor alignment quality, (iv) poor base quality, and (v) possible recombination. Alignment quality is determined by a threshold of the number of SNPs and Indels contained within a given window size. Base quality is optionally determined using FreeBayes [55] to measure read support and mixed alleles. Bases likely to have undergone recent recombination are identified using PhiPack [89]. Only columns passing a set of filters based on these criteria are

considered reliable core-genome SNPs. The final set of core-genome SNPs is given to FastTree2 [90] for reconstruction of the whole-genome phylogeny.

Compressed alignment file

For simplicity and storage efficiency, the output of Parsnp includes a single binary file encoding the reference genome, annotations, alignment, variants, and tree. Thousand-fold compression of the alignment is achieved by storing only the columns that contain variants. The full multiple alignment can be faithfully reconstructed from this reference-compressed representation on demand. Since Parsnp focuses on aligning only core blocks of relatively similar genomes, the number of variant columns tends to increase at a sub-linear rate as the number of genomes increases, resulting in huge space savings versus alternative multiple alignment formats. Conversion utilities are provided for importing/exporting common formats to/from the binary archive file, including: BED, GenBank, FASTA, MAF, Newick, VCF, and XMFA.

Interactive Visualization

Developed in tandem with Parsnp, the visualization tool Gingr allows for interactive exploration of trees and alignments. In addition to the compressed alignment format, Gingr accepts standard alignment formats and can serve as a general-purpose multiple alignment viewer. Uniquely, Gingr is capable of providing dynamic exploration of alignments comprising thousands of genomes and millions of alignment columns. It is the first tool of its kind capable of dynamically visualizing multiple alignments of this scale. The alignment can be seamlessly zoomed from a display of variant density (at the genome level) to a full representation of the multiple alignment (at the nucleotide level). For exploration of phyletic patterns, the alignment is simultaneously presented along with the core-genome SNP tree, annotations, and dynamic variant highlighting. The tree can be zoomed by clade, or individual genomes selected to expand via a

fisheye zoom. Structural variation across the genome can also be displayed using Sybil coloring [91], where a color gradient represents the location and orientation of each LCB with respect to the reference. This is useful for identifying structurally variant regions of the core.

Evaluation of performance

We evaluated Parsnp on three simulated datasets (derived from *Escherichia coli* K-12 W3110) and three real datasets (*Streptococcus pneumoniae*, *Peptoclostridium difficile*, *Mycobacterium tuberculosis*). Parsnp is compared below versus two whole-genome alignment methods (Mugsy, Mauve), a k-mer based method (kSNP), and two commonly used mapping pipelines (based on Smalt and BWA). The Smalt pipeline replicates the methods of the landmark Harris *et al.* paper [92] that has been adopted in many subsequent studies. The BWA pipeline is similar to the Smalt pipeline, but uses BWA for read mapping (Methods).

Simulated *Escherichia coli* W3110 dataset

To precisely measure the accuracy of multiple tools across varying levels of divergence, we computationally evolved the genome of *E. coli* K-12 W3110 at three different mutation rates: 0.00001 (low), 0.0001 (medium), and 0.001 (high) SNPs per site, per branch. Each dataset comprises 32 simulated genomes, forming a perfect binary tree. Approximately ~65X coverage of Illumina MiSeq reads were simulated and assembled for each genome to create draft assemblies. For input, the whole-genome alignment programs were given the draft assemblies, and the mapping pipelines the raw reads. Table 1 gives the performance of each tool on each dataset. Figure 1 plots the performance of all tools averaged across all mutation rates.

The whole-genome alignment methods performed comparably across all three mutation rates (Figure 1, red squares), with Mauve exhibiting the highest sensitivity (97.42%) and Parsnp the highest precision (99.99%). Mugsy demonstrated slightly higher sensitivity than Parsnp but with lower precision. Mugsy's lower precision was traced to a single *fumA* paralog [93] misalignment that generated a high number of false-positive SNPs. All genome alignment methods were affected by misalignment of repeats and missing or low-quality bases in the assembly.

Performance of the individual methods was also measured in terms of branch SNP and length error with respect to the true phylogeny (Figure 2). These errors closely followed the false-negative and false-positive rates of each method, with no distinguishable pattern or branch biases. On draft genomes, precise methods such as Parsnp yielded underestimates of branch lengths while more aggressive methods like Mugsy resulted in more overestimates (outliers not shown). The aggressive methods also showed more variance in performance across branches. Again, genome alignment of finished genomes produced the most accurate results of all methods tested.

Comparison on closed genomes

Mugsy, Mauve, and Parsnp all performed similarly on finished genomes (Figures 1 and 2, green squares), offering a significant boost in sensitivity over both draft assemblies and reference mapping. Mugsy, Mauve, and Parsnp all exhibited near perfect false-discovery rates (FDR), with Parsnp being the only method to not report a single false positive across the three datasets. Both Mauve and Mugsy were similarly near-perfect in terms of true-positive rates (TPR). The drop in sensitivity (0.9%) for Parsnp on full genomes can be explained by a lack of an LCB extension method.

Mugsy was the most affected by draft genomes, going from best on closed genomes to demonstrating more false positives (Table 1) and LCB counts (Table 2) on draft genomes. Parsnp offered the overall best FDR of the genome alignment methods, and the fewest number of LCBs, averaged across both draft and closed genome datasets.

Comparison to read mapping methods

On average, mapping-based methods were as precise and 0.5–1% more sensitive than alignment of draft genomes (Figure 1, blue triangles). Smalt showed the highest sensitivity, while BWA was the most specific. The precision of the mapping approaches may be overestimated for this dataset due to the absence of non-core sequence that is known to confound mapping [94]. Parsnp was the only genome alignment method to match the precision of mapping, but with a slight reduction in sensitivity. However, when provided with finished genomes, the whole-genome alignment methods excel in both sensitivity and specificity compared to read mapping. Thus, the performance divide between whole-genome alignment and mapping is entirely due to assembly quality and completeness. Using short reads, both the mapping and assembly-based approaches suffer false-negatives due to ambiguous mappings or collapsed repeats, respectively. Exceeding 99% sensitivity for this test set requires either longer reads (for mapping) or complete genomes (for alignment) to accurately identify SNPs in the repetitive regions.

***Comparison on 31 Streptococcus pneumoniae* genomes**

Parsnp was compared to whole-genome alignment methods using the 31-genome *S. pneumoniae* dataset presented in the original Mugsy publication [36]. Angiuoli and Salzberg compared Mugsy, Mauve, and Nucmer+TBA to measure the number of LCBs and size of the core genome aligned. On this dataset, Parsnp is within 5–10% of

Mauve and Mugsy in terms of core genome size, while matching the NUCmer+TBA result but with fewer LCBs. In addition, Parsnp ran hundreds of times faster than the other methods (Table 3), finishing this 31-way alignment in less than 60 seconds.

Peptoclostridium difficile outbreak in the UK

Parsnp and Gingr are particularly suited for outbreak analyses of infectious diseases. To demonstrate this, we applied Parsnp to a recent *P. difficile* outbreak dataset [95]. To generate input suitable for Parsnp, we assembled all genomes using iMetAMOS [96]. It is important to note that this was a resequencing project not intended for assembly and represents a worst case for a core-genome alignment approach; reads ranged from 50 to 100bp in length and some genomes were sequenced without paired ends. The 826-way core genome alignment resulted in 1.4 Gbp being aligned in less than 5 hours. The core genome represented 40% of the *P. difficile* 630 reference genome, consistent with previous findings [97]. Figure 3 shows a Gingr visualization of the 826-way alignment and clade phylogeny. Related outbreak clusters are immediately visible from the phyletic patterns of the alignment, confirming the primary clades of the tree. In addition, the SNP heatmap highlights the phyletic signature of several subclades, in this case within the known hpdBCA operon [98] that is extremely well conserved across all 826 genomes.

Figure 4 shows a zoomed view of the 826 *P. difficile* genome alignment in Gingr, highlighting a single annotated gene. Although no metadata is publically available for this outbreak dataset, we identified that *bacA*, a gene conferring antibiotic resistance to bacitracin, is conserved in all 826 isolates. While alternative antibiotic treatments for *P. difficile* infections have been well-studied over the past 20-30 years [99], a recent study reported that 100% of 276 clinical isolates had high-level resistance to

bacitracin [100]. In concordance with this study, our results indicate there may be widespread bacitracin resistance across this outbreak dataset. Thus alternative antibiotics, such as vancomycin, could represent better treatment options.

Mycobacterium tuberculosis geographic spread

For a second case evaluation, we ran Parsnp on a *M. tuberculosis* global diversity dataset [101]. In this case, the raw SNP calls were kindly made available (Iñaki Comas, personal communication), facilitating a direct comparison to the published results. The variant pipeline of Comas *et al.* is similar to our BWA pipeline, but with all SNP calls intersected with MAQ SNPfilter, which discards any SNP with neighboring Indels ±3bp or surrounded by >3 SNPs within a 10bp window. To replicate this study using whole-genome alignment, we assembled all genomes from the raw reads using iMetAMOS and ran Parsnp on the resulting draft assemblies. Figure 5 summarizes the results of the comparison and Figure 6 shows a Gingr visualization of the resulting tree and alignment, with major clades confirmed by correlations in the SNP density display. The vast majority of Parsnp calls are shared with Comas *et al.* and the tree generated by Parsnp very closely matches the published lineages. Lower SNP counts on our end are likely due to fragmented assemblies that artificially shrink the core (low coverage, zero coverage). SNP columns unique to Parsnp (~10 per genome after Repeat, MGE, LCB, Recombination, Alignment, Assembly quality, and Indel filters) are likely contained in low coverage regions, misassemblies, or misalignments.

Runtime and storage analysis

To evaluate Parsnp's scalability, we profiled performance across six datasets ranging from 32 genomes to 10,000 genomes. Runtime was observed to increase linearly with additional genomes (Supplementary Figure 1), requiring a few minutes for the 32

genome *E. coli* dataset, 1.5 hours for the 826 genome *P. difficile* dataset, and a maximum of ~14 hours to align the 10,000 genome set on a 2.2 GHz, 32-core (64-thread) 1TB RAM server. In addition, for the 32 genome simulated *E. coli* datasets, Parsnp is 10–100 times faster than all other methods evaluated. Memory usage scales approximately by 15 bytes per total input dataset size (bps); maximum memory usage was 2GB of RAM for the 145Mbp *E. coli* dataset and 309GB for the 21Gbp *S. pneumoniae* dataset. Memory usage can be explicitly limited via a command-line parameter (--max-partition-size) but results in increased runtime.

In addition to runtime efficiency, Parsnp requires much less storage than the other approaches due to its binary alignment format and the compressive effect of assembly. For the 32-genome *E. coli* dataset, Parsnp's output totals just 4.5MB, compared to 13GB required to store compressed FASTQ [102] and VCF [103] files and 149MB to store XMFA [38]. Storage reductions are amplified for larger datasets. For example, the raw read data for the *P. difficile* dataset requires 1.4TB of storage (0.6TB compressed). Assembling these data reduces the total to 3.3GB by removing the redundancy of the reads. The XMFA alignment of these assemblies is 1.4GB, and reference-compressed binary format occupies just 15MB. This equates to roughly a 100,000X (lossy) compression factor from raw reads to compressed archive, requiring only 0.08 bits per base to store the full core-genome alignment plus other related information, which is competitive with related techniques like CRAM [104]. As outbreak studies continue to expand in scale, whole-genome assembly and alignment presents a sustainable alternative to the current mapping-based strategies.

Discussion

Parsnp is orders of magnitude faster than current methods for whole-genome alignment and SNP typing, but it is not without limitations. Parsnp represents a compromise between whole-genome alignment and read mapping. Compared to whole-genome aligners, Parsnp is less flexible because it is designed to conservatively align the core genome and is less sensitive as a result. Compared to read mapping, Parsnp is less robust and requires high-quality assemblies to maximize sensitivity. Thus, the right tool depends on the data and task at hand.

Core-genome alignment and phylogeny reconstruction are critical to microbial forensics and modern epidemiology. When finished or high-quality genomes are available, Parsnp is both efficient and accurate for these tasks. In addition, even for fragmented draft assemblies, Parsnp exhibits a favorable compromise between sensitivity and specificity. Surprisingly, Parsnp matched the specificity of the mapping-based approaches on the simulated datasets. However, multiplexed short-read sequencing followed by mapping still remains the most economical approach for sensitive analysis of large strain collections. Thus, Parsnp is recommended for analyzing high-quality assemblies or when raw read data is not available.

Assembled genomes have a number of advantages over read data—primarily compression and convenience. Storing, sharing, and analyzing raw read datasets incurs significant overhead from the redundancy in sequencing (often 100 fold), and this burden nearly resulted in the closure of the NCBI SRA database [105]. Adding additional orders of magnitude to the already exponential growth of sequencing data is not sustainable. Instead, information in the reads not currently stored in common assembly formats (*e.g.* allelic variants) should be propagated to the assembled

representation, forming a compressed, but nearly lossless format. In this way, genomes could be shared in their native, assembled format, saving both space and time of analysis. Here, we have taken a small step in that direction by identifying low quality bases, as computed by FreeBayes [55]. This allows filtering of low quality and mixed alleles and improves the specificity of the assembly-based approaches. However, more comprehensive, graph-based formats are needed to capture the full population information contained in the raw reads.

Parsnp was also built around the observation that high-quality, finished genome sequences have become more common as sequencing technology and assembly algorithms continue to improve. New technologies, such as PacBio SMRT sequencing [106] are enabling the generation of reference-grade sequences at extremely reduced costs. This presents another opportunity for Parsnp—the construction and maintenance of core genomes and trees for clinically important species. With well defined reference cores, outbreaks could be accurately typed in real-time by mapping sequences directly to the tree using phylogenetically aware methods such as pplacer [107] or PAGAN [108]. Such a phylogenetic approach would be preferable to alternative typing schemes based on loosely defined notions of similarity, such as pulse-field electrophoresis (PFGE) [109] and multi-locus sequence typing (MLST) [110].

Conclusion

Parsnp offers a highly efficient method for aligning the core genome of thousands of closely related species, and Gingr provides a flexible, interactive visualization tool for the exploration of huge trees and alignments. Together, they enable analyses not previously possible with whole-genome aligners. We have demonstrated that Parsnp

provides highly specific variant calls, even for highly fragmented draft genomes, and can efficiently reconstruct recent outbreak analyses including hundreds of whole genomes. Future improvements in genome assembly quality and formats will enable comprehensive cataloging of microbial population variation, including both point and structural mutations, using genome alignment methods such as Parsnp.

Materials and Methods

Software and configurations

Mugsy [36] v1.23 and Mauve Aligner [31, 33] v2.3.1 were run using default parameters on assembled sequences. mauveAligner was selected instead of progressiveMauve due to improved performance on the simulated *E. coli* datasets, which do not contain subset relationships. kSNP v2.0 [67] was run with a k-mer size of 25 on both the raw read data and the assemblies; the assemblies were merged with Ns using the provided `merge_fasta_contigs.pl` utility. Raw MAF/XMFA/VCF output was parsed to recover SNPs and build MultiFASTA files using a conversion script [111].

Smalt version 0.7.5 was run with default parameters for paired reads, mirroring the pipeline used in several recent SNP typing studies [92, 112-114]. `Samtools view` was used to filter for alignments with mapping qualities greater than or equal to 30. Variants were called by piping `samtools mpileup` output into `bcftools view` with the -v (variants only), -g (genotype) and -I (skip Indels) flags. Variants were then filtered with `VCFUtils varFilter` with the -d (minimum read depth) parameter set to 3. Variants for all samples of each set were called concomitantly by providing `samtools mpileup` with all BAM files.

BWA [53] was run in its standard paired-end alignment mode with default parameters, using `aln` to align each set of ends and `sampe` to produce a combined SAM file. `Samtools view` was used to filter for alignments with mapping qualities greater than or equal to 30. Variants were called by piping `samtools mpileup` output into `bcftools view` with the -v (variants only), -g (genotype) and -I (skip Indels) flags. Variants were then filtered with `VCFUtils varFilter` with the -d (minimum read depth) parameter set to 3. As with Smalt, variants for all samples of each set were called concomitantly by providing `samtools mpileup` with all BAM files.

FastTree v2 [90] was used to reconstruct phylogenies using default parameters.

E. coli K-12 W3110 simulated dataset

The complete genome of *E. coli* K-12 W3110 [115], was downloaded from RefSeq (AC_000091). This genome was used as the ancestral genome and evolution was simulated along a balanced tree for three evolutionary rates using the Seq-Gen package [116] with parameters `mHKY -t4.0 -l4646332 -n1 -k1` and providing the corresponding binary tree evolved at three evolutionary rates: 0.00001, 0.0001, and 0.001 SNPs per site, per branch. This corresponds to a minimum percent identity of approximately 99%, 99.9%, and 99.99% between the two most divergent genomes, respectively, reflecting the variation seen in typical outbreak analyses. No small (< 5bp) or large Indels were introduced, but an average of ten 1Kbp rearrangements (inversions and translocations) were added, per genome, using a custom script (Supplement). Paired reads were simulated to model current MiSeq lengths (2x150bp) and error rates (1%). Moderate coverage, two million PE reads (64X coverage), was simulated for each of the 32 samples using wgsim (default parameters, no Indels), from samtools package version 0.1.17 [56].

Two of the simulated read sets were independently run through iMetAMOS [96] to automatically determine the best assembler. The consensus pick across both datasets was SPAdes version 3.0 [83], which was subsequently run on the remaining 30 simulated read sets using default parameters. The final contigs and scaffolds files were used as input to the genome alignment methods. For mapping methods, the raw simulated reads were used. For accuracy comparisons, Indels were ignored and called SNPs were required to be unambiguously aligned across all 32 genomes (i.e. not part of a subset relationship; SNPs present but part of a subset relationship were ignored).

S. pneumoniae dataset

A full listing of accession numbers for the 31-genome *S. pneumoniae* dataset is described in [36]. For scalability testing, *Streptococcus pneumoniae* TIGR4 (NC_003028.3) was used to create a pseudo-outbreak clade involving 10,000 genomes evolved along a star phylogeny with on average 10 SNPs per genome.

M. tuberculosis dataset

We downloaded and assembled sequencing data from a recently published study of *M. tuberculosis* [101]. 225 runs corresponding to project ERP001731 were downloaded from NCBI SRA and assembled using the iMetAMOS ensemble of SPAdes, MaSuRCA, and Velvet. The iMetAMOS assembly for each sample can be replicated with the following commands, which will automatically download the data for `RUN_ID` directly from SRA:

```
> initPipeline -d asmTB -W iMetAMOS -m RUN_ID -i 200:800
> runPipeline -d asmTB -a spades,masurca,velvet -p 16
```

The *M. tuberculosis* dataset included a mix of single and paired-end runs with a sequence length ranging from 51–108bp. The average k-mer size selected for

unpaired data was 26, resulting in an average of 660 contigs and an N50 size of 17Kbp. For paired-end data, the average selected k-mer was 35, resulting in an average of 333 contigs and an N50 size of 43Kbp. Assemblies containing more than 2,000 contigs, or 1.5X larger/smaller than the reference genome, were removed.

P. difficile dataset

Note, *Clostridium difficile* was recently renamed to *Peptoclostridium difficile [117]*. We downloaded and assembled sequencing data from a recently published study of *P. difficile* [95]. 825 runs corresponding to project ERP003850 were downloaded from NCBI SRA [88] and assembled within iMetAMOS this time only using SPAdes, which was identified as the best performer on the *M. tuberculosis* dataset. The iMetAMOS assembly for each sample can be replicated with the following commands, which will download the data for `RUN_ID` directly from SRA:

```
> initPipeline -d asmPD -W iMetAMOS -m RUN_ID -i 200:800
> runPipeline -d asmPD -a spades -p 16
```

The *P. difficile* dataset included paired-end runs with a sequence length ranging from 51–100bp. SPAdes was selected as the assembler and run with k-mer sizes of 21, 33, 55, and 77. The assemblies had an average of 660 contigs and an N50 size of 138Kbp. Assemblies containing more than 2,000 contigs, or 1.5X larger/smaller than the reference genome, were removed.

Abbreviations

Bp: base pair; Mbp: million base pairs; LCB: locally collinear block; MUM: maximal unique match; MUMi: similarity index based on maximal unique matches; SNP: single-nucleotide polymorphism; Indel: insertion or deletion; ERA: European Read Archive; SRA: Sequence Read Archive; PE: paired-end; NGS: Next generation sequencing; VCF: variant call format; XMFA: extendend multi-fasta format.

Data and software availability

All data, supplementary files, assemblies, packaged software binaries and scripts described in the manuscript are available from [111]. Source code of the described software, including Parsnp and Gingr, is available for download from: http://github.com/marbl/harvest

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TJT, BDO, and AMP conceived the method, designed the experiments, and drafted the manuscript. TJT implemented Parsnp and BDO implemented Gingr. TJT, BDO, and SK performed the experiments. All authors read and approved the final manuscript.

Description of additional data files

- Supplementary Figure 1
- Portable Document Format (PDF)
- Timing performance on genomic datasets
- Parsnp core-genome alignment performance on the datasets presented in the manuscript, from 32 to 10k genomes.

Acknowledgments

We would like to thank Iñaki Comas for providing the raw data for the TB study; Xiangyu Deng for helpful comments on the manuscript and software; and Bradd Haley, Bill Klimke, Jason Sahl, Maliha Aziz, and Suzanne Bialek-Davenet for feedback on early versions of the software. This work was funded under Agreement No. HSHQDC-07-C-00020 awarded by the Department of Homeland Security Science and Technology Directorate (DHS/S&T) for the management and operation of the National Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. In no event shall the DHS, NBACC, or Battelle National Biodefense Institute (BNBI) have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. The Department of Homeland Security does not endorse any products or commercial services mentioned in this publication.

References

1. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC: **The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* 2012, **40:**D571-579.
2. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, et al: **Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany.** *N Engl J Med* 2011, **365:**709-717.
3. Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL: **GAGE-B: an evaluation of genome assemblers for bacterial organisms.** *Bioinformatics* 2013, **29:**1718-1725.
4. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM: **Hybrid error correction and de novo assembly of single-molecule sequencing reads.** *Nature Biotechnology* 2012.
5. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.** *Nat Methods* 2013, **10:**563-569.
6. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA: **Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing.** *Sci Transl Med* 2012, **4:**148ra116.
7. Gonzalez-Candelas F, Bracho MA, Wrobel B, Moya A: **Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source.** *BMC biology* 2013, **11:**76.
8. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, et al: **Helicobacter pylori genome evolution during human infection.** *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108:**5033-5038.
9. Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D: **Efficient inference of recombination hot regions in bacterial genomes.** *Mol Biol Evol* 2014.
10. Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EPC: **After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection.** *Genome research* 2012**:**gr.129544.129111--gr.129544.129111-.
11. Didelot X, Meric G, Falush D, Darling AE: **Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli.** *BMC Genomics* 2012, **13:**256.
12. Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D, Donnelly P: **Recombination and population structure in Salmonella enterica.** *PLoS Genetics* 2011, **7:**e1002191.

13. Didelot X, Lawson D, Darling A, Falush D: **Inference of homologous recombination in bacteria using whole-genome sequences.** *Genetics* 2010, **186:**1435-1449.
14. Didelot X, Darling A, Falush D: **Inferring genomic flux in bacteria.** *Genome Research* 2009, **19:**306-317.
15. Dhillon BK, Chiu TA, Laird MR, Langille MG, Brinkman FS: **IslandViewer update: Improved genomic island discovery and visualization.** *Nucleic Acids Res* 2013, **41:**W129-132.
16. Langille MG, Hsiao WW, Brinkman FS: **Evaluation of genomic island predictors using a comparative genomics approach.** *BMC Bioinformatics* 2008, **9:**329.
17. Touchon M, Rocha EPC: **Causes of Insertion Sequences Abundance in Prokaryotic Genomes.** *Mol Biol Evol* 2007, **24:**969-981.
18. Fuxelius H-HH, Darby A, Min C-KK, Cho N-HH, Andersson SG: **The genomic and metabolic diversity of Rickettsia.** *Research in microbiology* 2007, **158:**745-753.
19. Dessimoz C, Gabaldon T, Roos DS, Sonnhammer EL, Herrero J, Quest for Orthologs C: **Toward community standards in the quest for orthologs.** *Bioinformatics* 2012, **28:**900-904.
20. Gabaldon T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer EL, Lewis S: **Joining forces in the quest for orthologs.** *Genome Biol* 2009, **10:**403.
21. Alekseyev MA, Pevzner PA: **Breakpoint graphs and ancestral genome reconstructions.** *Genome Res* 2009, **19:**943-957.
22. Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Worheide G, Baurain D: **Resolving difficult phylogenetic questions: why more sequences are not enough.** *PLoS Biology* 2011, **9:**e1000602.
23. Philippe H, Delsuc F, Brinkmann H, Lartillot N: **PHYLOGENOMICS.** *Annual Review of Ecology, Evolution, and Systematics* 2005, **36:**541-562.
24. Hayward A, Grabherr M, Jern P: **Broad-scale phylogenomics provides insights into retrovirus-host evolution.** *Proc Natl Acad Sci U S A* 2013, **110:**20146-20151.
25. Gabaldon T, Koonin EV: **Functional and evolutionary implications of gene orthology.** *Nat Rev Genet* 2013, **14:**360-366.
26. Dagan T, Artzy-Randrup Y, Martin W: **Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution.** *Proc Natl Acad Sci U S A* 2008, **105:**10039-10044.
27. Dale J, Price EP, Hornstra H, Busch JD, Mayo M, Godoy D, Wuthiekanun V, Baker A, Foster JT, Wagner DM, et al: **Epidemiological tracking and population assignment of the non-clonal bacterium, Burkholderia pseudomallei.** *PLoS neglected tropical diseases* 2011, **5:**e1381.
28. Sahl JW, Caporaso JG, Rasko DA, Keim P: **The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes.** *PeerJ* 2014, **2:**e332.

29. Swidan F, Rocha EPCP, Shmoish M, Pinter RYY: **An Integrative Method for Accurate Comparative Genome Mapping.** *PLoS Comput Biol* 2006, **2**.
30. Bray N, Pachter L: **MAVID: Constrained Ancestral Alignment of Multiple Sequences.** *Genome Research* 2004, **14:**693-699.
31. Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.** *PLoS One* 2010, **5:**e11147.
32. Darling AE, Treangen TJ, Messeguer X, Perna NT: **Analyzing patterns of microbial evolution using the mauve genome alignment system.** *Methods in molecular biology (Clifton, NJ)* 2007, **396:**135-152.
33. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements.** *Genome Res* 2004, **14:**1394-1403.
34. Hohl M, Kurtz S, Ohlebusch E: **Efficient multiple genome alignment.** *Bioinformatics* 2002, **18:**S312-320.
35. Treangen T, Messeguer X: **M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species.** *BMC Bioinformatics* 2006, **7**.
36. Angiuoli SV, Salzberg SL: **Mugsy: fast multiple alignment of closely related whole genomes.** *Bioinformatics* 2011, **27:**334-342.
37. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome Res* 2004, **14:**708-715.
38. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.** *Genome Res* 2003, **13:**721-731.
39. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E: **Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs.** *Genome Research* 2008, **18:**1814-1828.
40. Dagan T, Artzy-Randrup Y, Martin W: **Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution.** *Proceedings of the National Academy of Sciences* 2008, **105:**10039-10044.
41. Ochman H, Lerat E, Daubin V: **Examining bacterial species under the specter of gene transfer and exchange.** *Proc Natl Acad Sci U S A* 2005, **102 Suppl:**6595-6599.
42. Wang L, Jiang T: **On the complexity of multiple sequence alignment.** *Journal of computational biology : a journal of computational molecular cell biology* 1994, **1:**337-348.
43. Liu K, Warnow T: **Large-scale multiple sequence alignment and tree estimation using SATe.** *Methods Mol Biol* 2014, **1079:**219-244.
44. Chan CX, Ragan MA: **Next-generation phylogenomics.** *Biol Direct* 2013, **8:**3.
45. Posada D: **Phylogenetic models of molecular evolution: next-generation data, fit, and performance.** *J Mol Evol* 2013, **76:**351-352.

46. Robinson ER, Walker TM, Pallen MJ: **Genomics and outbreak investigation: from sequence to consequence.** *Genome medicine* 2013, **5:**36.
47. Bennett S: **Solexa Ltd.** *Pharmacogenomics* 2004, **5:**433-438.
48. Schatz MC, Delcher AL, Salzberg SL: **Assembly of large genomes using second-generation sequencing.** *Genome Res* 2010, **20:**1165-1173.
49. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, et al: **Assemblathon 1: a competitive assessment of de novo short read assembly methods.** *Genome Res* 2011, **21:**2224-2241.
50. Langmead B: **Aligning short sequencing reads with Bowtie.** *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2010, **Chapter 11:**Unit 11 17.
51. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nature Methods* 2012, **9:**357-359.
52. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26:**589-595.
53. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
54. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.
55. Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** *arXiv:12073907* 2012.
56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
57. Harris SR, Torok ME, Cartwright EJ, Quail MA, Peacock SJ, Parkhill J: **Read and assembly metrics inconsequential for clinical utility of whole-genome sequencing in mapping outbreaks.** *Nat Biotechnol* 2013, **31:**592-594.
58. Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E: **Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads.** *Mol Biol Evol* 2014.
59. Li H: **Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples.** *Bioinformatics* 2014.
60. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nature Reviews Genetics* 2011.
61. Vinga S, Almeida J: **Alignment-free sequence comparison-a review.** *Bioinformatics* 2003, **19:**513-523.
62. Patro R, Mount SM, Kingsford C: **Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.** *Nat Biotechnol* 2014, **32:**462-464.
63. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biol* 2014, **15:**R46.

64. Chor B, Horn D, Goldman N, Levy Y, Massingham T: **Genomic DNA k-mer spectra: models and modalities.** *Genome Biol* 2009, **10:**R108.
65. Hauser M, Mayer CE, Soding J: **kClust: fast and sensitive clustering of large protein sequence databases.** *BMC Bioinformatics* 2013, **14:**248.
66. Ghodsi M, Liu B, Pop M: **DNACLUST: accurate and efficient clustering of phylogenetic marker genes.** *BMC Bioinformatics* 2011, **12:**271.
67. Gardner SN, Hall BG: **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* 2013, **8:**e81760.
68. Hohl M, Ragan MA: **Is multiple-sequence alignment required for accurate inference of phylogeny?** *Syst Biol* 2007, **56:**206-221.
69. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data.** *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108:**1513-1518.
70. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Research* 2009, **19:**1117-1123.
71. Simpson JT, Durbin R: **Efficient de novo assembly of large genomes using compressed data structures.** *Genome Research* 2012, **22:**549-556.
72. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G: **Aggressive assembly of pyrosequencing reads with mates.** *Bioinformatics* 2008, **24:**2818-2824.
73. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98:**9748-9753.
74. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Adam MP: **Hybrid error correction and de novo assembly of single-molecule sequencing reads.** *Nature Biotechnology* 2012, **30:**693-700.
75. Zerbino DR: **Using the Velvet de novo assembler for short-read sequencing technologies.** *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2010, **Chapter 11:**Unit 11 15.
76. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
77. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al: **The sequence and de novo assembly of the giant panda genome.** *Nature* 2010, **463:**311-317.
78. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome biology* 2004, **5:**R12-R12.

79. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL: **Alignment of whole genomes MUMMER.** *Nucl Acids Res* 1999, **27:**2369-2369.
80. Perna NT, Plunkett G, 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, et al: **Genome sequence of enterohaemorrhagic Escherichia coli O157:H7.** *Nature* 2001, **409:**529-533.
81. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: **Performance comparison of benchtop high-throughput sequencing platforms.** *Nat Biotechnol* 2012, **30:**434-439.
82. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I: **RefSeq microbial genomes database: new representation and annotation strategy.** *Nucleic Acids Res* 2014, **42:**D553-559.
83. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.** *J Comput Biol* 2012, **19:**455-477.
84. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH, Phillippy AM: **Reducing assembly complexity of microbial genomes with single-molecule sequencing.** *Genome Biol* 2013, **14:**R101.
85. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic acids research* 2002, **30:**2478-2483.
86. Deloger M, El Karoui M, Petit MA: **A genomic distance based on MUM indicates discontinuity between most bacterial species and genera.** *J Bacteriol* 2009, **191:**91-99.
87. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Research* 2004, **32:**1792-1797.
88. [http://www.ncbi.nlm.nih.gov/sra]
89. Bruen TC, Philippe H, Bryant D: **A simple and robust statistical test for detecting the presence of recombination.** *Genetics* 2006, **172:**2665-2681.
90. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for large alignments.** *PLoS One* 2010, **5:**e9490.
91. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15:**589-594.
92. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, et al: **Evolution of MRSA during hospital transmission and intercontinental spread.** *Science* 2010, **327:**469-474.
93. van Vugt-Lussenburg BM, van der Weel L, Hagen WR, Hagedoorn PL: **Biochemical similarities and differences between the catalytic [4Fe-4S] cluster containing fumarases FumA and FumB from Escherichia coli.** *PLoS One* 2013, **8:**e55549.
94. Li H: **Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples.** *arxivorg* 2014.

95. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CL, Golubchik T, Batty EM, Finney JM, et al: **Diverse sources of C. difficile infection identified on whole-genome sequencing.** *N Engl J Med* 2013, **369:**1195-1205.
96. Sergey Koren TJT, Christopher M Hill, Mihai Pop and Adam M Phillippy: **Automated ensemble assembly and validation of microbial genomes.** *BMC Bioinformatics* 2014, **15**.
97. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeno-Tarraga AM, Wang H, et al: **The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome.** *Nat Genet* 2006, **38:**779-786.
98. Dawson LF, Donahue EH, Cartman ST, Barton RH, Bundy J, McNerney R, Minton NP, Wren BW: **The analysis of para-cresol production and tolerance in Clostridium difficile 027 and 012 strains.** *BMC Microbiol* 2011, **11:**86.
99. Young GP, Ward PB, Bayley N, Gordon D, Higgins G, Trapani JA, McDonald MI, Labrooy J, Hecker R: **Antibiotic-associated colitis due to Clostridium difficile: double-blind comparison of vancomycin with bacitracin.** *Gastroenterology* 1985, **89:**1038-1045.
100. Venugopal AA, Johnson S: **Current state of Clostridium difficile treatment options.** *Clin Infect Dis* 2012, **55 Suppl 2:**S71-76.
101. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, et al: **Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans.** *Nat Genet* 2013, **45:**1176-1182.
102. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.** *Nucleic Acids Res* 2010, **38:**1767-1771.
103. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27:**2156-2158.
104. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E: **Efficient storage of high throughput DNA sequencing data using reference-based compression.** *Genome Res* 2011, **21:**734-740.
105. Lipman D, Flicek P, Salzberg S, Gerstein M, Knight R: **Closure of the NCBI SRA and implications for the long-term future of genomics data storage.** *Genome Biol* 2011, **12:**402.
106. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW: **Real-time DNA sequencing from single polymerase molecules.** *Methods Enzymol* 2010, **472:**431-455.
107. Matsen FA, Kodner RB, Armbrust EV: **pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree.** *BMC Bioinformatics* 2010, **11:**538.
108. Loytynoja A, Vilella AJ, Goldman N: **Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm.** *Bioinformatics* 2012, **28:**1684-1691.
109. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B: **Interpreting chromosomal DNA**

restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* 1995, **33:**2233-2239.

110. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, et al: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci U S A* 1998, **95:**3140-3145.
111. **Data, Supplementary Files, Assemblies, and Code.** [http://www.cbcb.umd.edu/software/harvest]
112. Castillo-Ramirez S, Corander J, Marttinen P, Aldeljawi M, Hanage WP, Westh H, Boye K, Gulay Z, Bentley SD, Parkhill J, et al: **Phylogeographic variation in recombination rates within a global clone of Methicillin-Resistant Staphylococcus aureus (MRSA).** *Genome Biology* 2012, **13:**R126.
113. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M: **Population genomics of post-vaccine changes in pneumococcal epidemiology.** *Nat Genet* 2013, **45:**656-663.
114. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, et al: **Rapid pneumococcal evolution in response to clinical interventions.** *Science* 2011, **331:**430-434.
115. Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner BL, Mori H, Horiuchi T: **Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110.** *Molecular systems biology* 2006, **2:**2006 0007.
116. Rambaut A, Grassly NC: **Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Comput Appl Biosci* 1997, **13:**235-238.
117. Yutin N, Galperin MY: **A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia.** *Environ Microbiol* 2013, **15:**2631-2641.

Figure Legends

Figure 1. Core-genome SNP accuracy for simulated *E. coli* datasets

Results are averaged across low, medium, and high mutation rates. Red squares denote alignment-based SNP calls on draft assemblies, green squares alignment-based SNP calls on closed genomes, and blue triangles for read mapping. Full results for each dataset are given in Table 1.

Figure 2. Branch errors for simulated *E. coli* datasets. Simulated *E. coli* trees are shown for medium mutation rate (0.0001 per base per branch). Panel A shows branch length errors as bars, with overestimates of branch length above each branch and underestimates below each branch. Maximum overestimate of branch length was 2.15% (bars above each branch) and maximum underestimate was 4.73% (bars below each branch). Panel B shows branch SNP errors as bars, with false-positive errors above each branch and false-negative errors below each branch. The maximum FP SNP value is 6 (bars above each branch) and maximum FN SNP value is 23 (bars below each branch). Note that the bar heights have been normalized by the maximum value for each tree and are not comparable across trees. Outlier results from Mugsy were excluded from the branch length plot, and kSNP results are not shown. All genome alignment methods performed similarly on closed genomes, with Mauve and Mugsy exhibiting the best sensitivity (Table 1).

Figure 3. Gingr visualization of 826 *P. difficile* genomes aligned with Parsnp. The leaves of the reconstructed phylogenetic tree (left) are paired with their corresponding rows in the multi-alignment. A genome has been selected (rectangular aqua highlight), resulting in a fisheye zoom of several leaves and their rows. A SNP density

plot (center) reveals the phylogenetic signature of several clades, in this case within the fully-aligned hpd operon (hpdB, hpdC, hpdA). The light gray regions flanking the operon indicate unaligned sequence. When fully zoomed (right), individual bases and SNPs can be inspected.

Figure 4. Conserved presence of *bacA* antiobiotic resistance gene in *P. difficile* outbreak. Gingr visualization of conserved bacitracin resistance gene within the Parsnp alignment of 826 *P. difficile* genomes. Vertical lines indicate SNPs, providing visual support of subclades within this outbreak dataset.

Figure 5. Comparison of Parsnp and Comas *et al.* result on *M. tuberculosis* dataset. On top, SNPs unique to Comas *et al.* analysis [101] at left (red), SNPs unique to Parsnp at right (blue), and shared SNPs in the middle (purple). On bottom, comparison of unrooted phylogenies as reconstructed by Comas *et al.* (left) and Parsnp (right). The reconstructions are largely concordant, with all major clades shared and well supported by both trees.

Figure 6. Gingr visualization of 224 *M. tuberculosis* genomes aligned with Parsnp. The visual layout is the same as Figure 3, but unlike Figure 3, a SNP density plot across the entire genome is displayed. Major clades are visible as correlated SNP densities across the length of the genome.

Tables

Table 1. Core-genome SNP accuracy for simulated *E. coli* datasets

Data shown indicates performance metrics of the evaluated methods on the three simulated *E. coli* datasets (low, medium and high). Descr: Paradigm employed by each method, *WGA* indicates whole-genome alignment, *CGA* indicates core genome alignment, *KMER* indicates k-mer based SNP calls, and *MAP* indicates read mapping. False positive (FP) and false negative (FN) counts for the three mutation rates (low, med, and high). TP: number of SNP calls that agreed with the truth. FP: number of SNP calls that are not in truth set. FN: number of truth SNP calls not detected. TPR: TP/(TP+FN). FDR: FP/(TP+FP). A total of 1,299,178 SNPs were introduced into the 32-genome dataset, across all three mutational rates.

Method	Descr.	FP low	FN low	FP med	FN med	FP high	FN high	TPR	FDR
Mauve	*WGA*	148	318	198	2877	100	30378	0.974	0.0004
Mauve_full	*WGA*	0	0	2	38	6	649	0.999	0
Mugsy	*WGA*	*1261	395	1928	3371	1335	34923	0.970	0.0036
Mugsy_full	*WGA*	2	0	2	0	1	81	0.999	0
Parsnp	*CGA*	23	423	45	3494	7	35466	0.970	0.0001
Parsnp_full	*CGA*	0	24	0	603	0	10989	0.992	0
kSNP	*KMER*	259	600	908	19730	1968	916127	0.280	0.0086
Smalt	*MAP*	33	110	0	1307	55	22957	0.981	0.0001
BWA	*MAP*	0	168	16	1947	27	27091	0.9775	0.0000

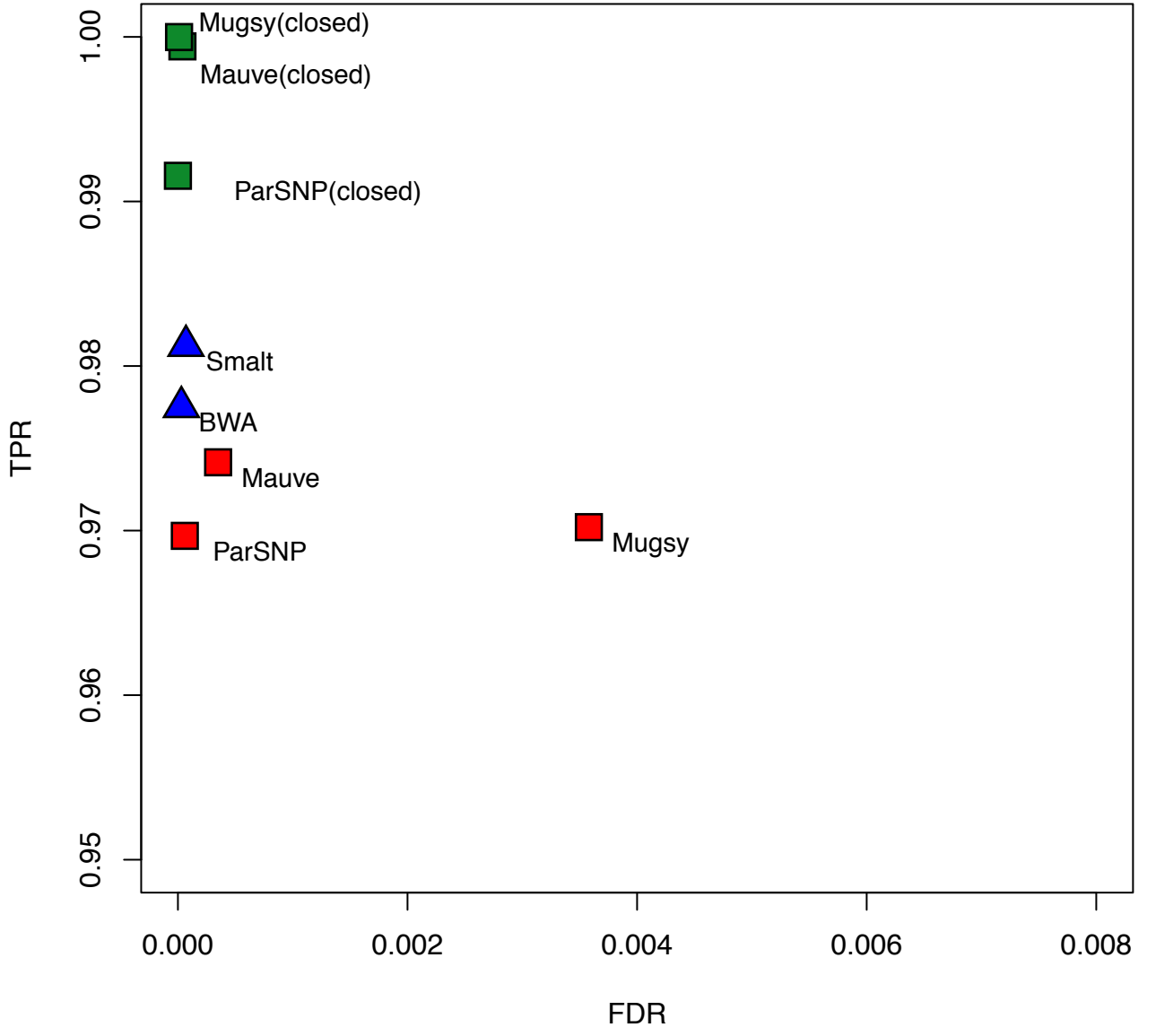
Table 2. Comparison of locally collinear alignment block (LCB) count for simulated *E. coli* datasets, on assembled and finished genomes. Low: >99.99% similarity, Medium: >99.9% similarity, High: >99% similarity. *_full indicates genome aligners ran on closed genomes.

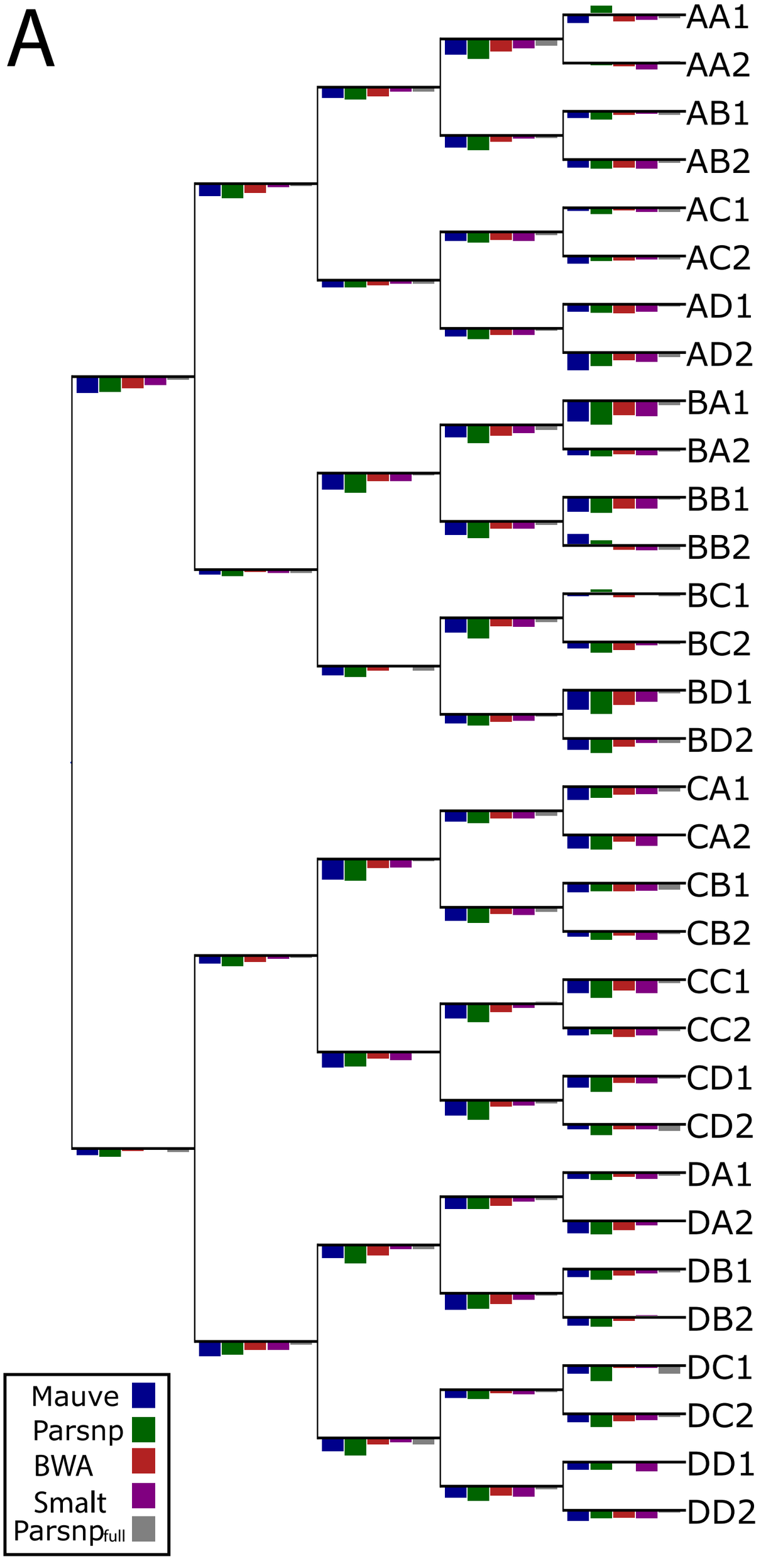
Method	Low	Medium	High
Mauve	325	363	519
Mauve_full	150	174	333
Mugsy	10 977	11 194	16 632
Mugsy_full	237	247	351
Parsnp	205	271	344
Parsnp_full	139	190	506

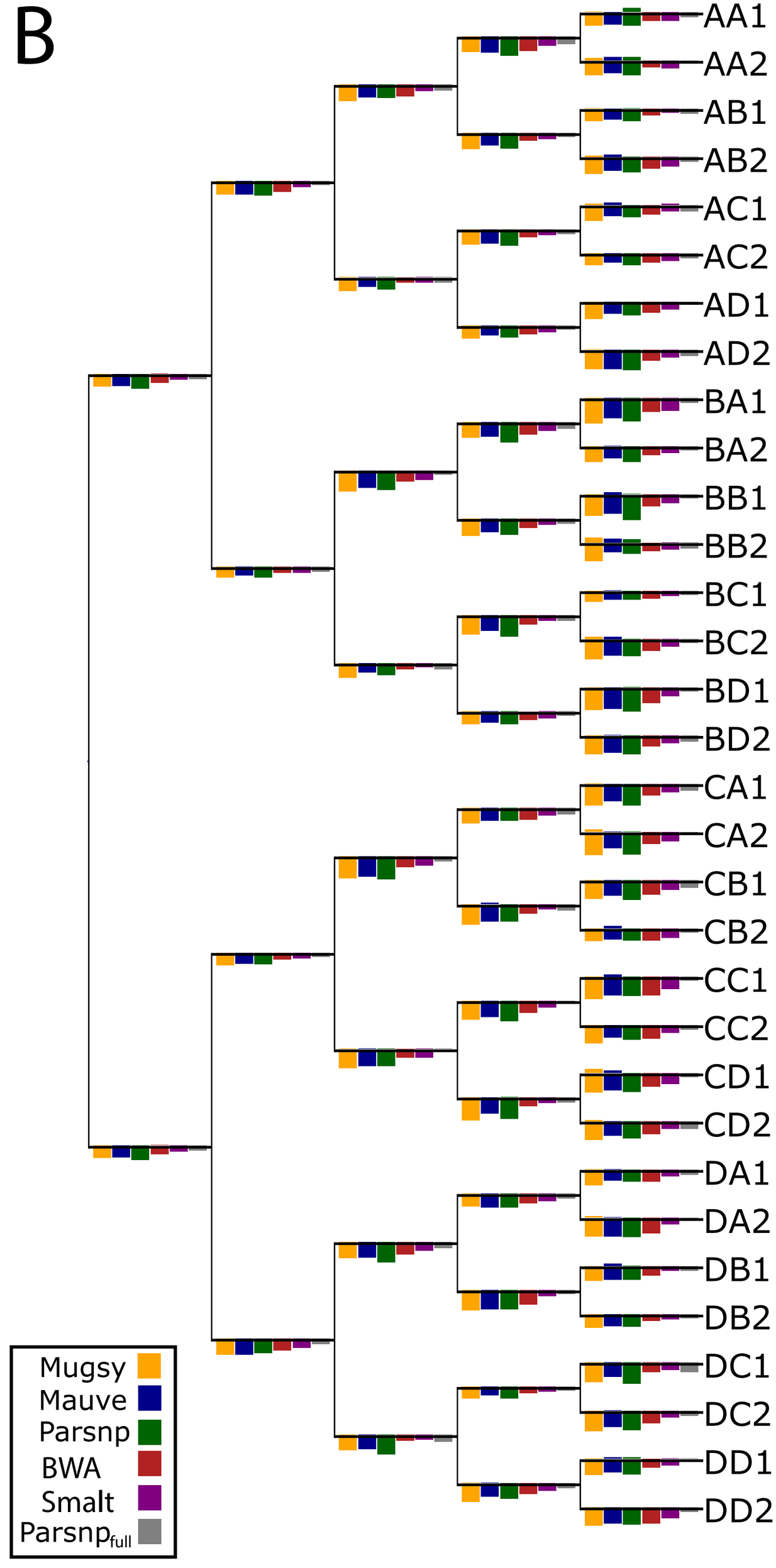
Table 3. Comparison to the 31 *S. pneumoniae* Mugsy benchmark

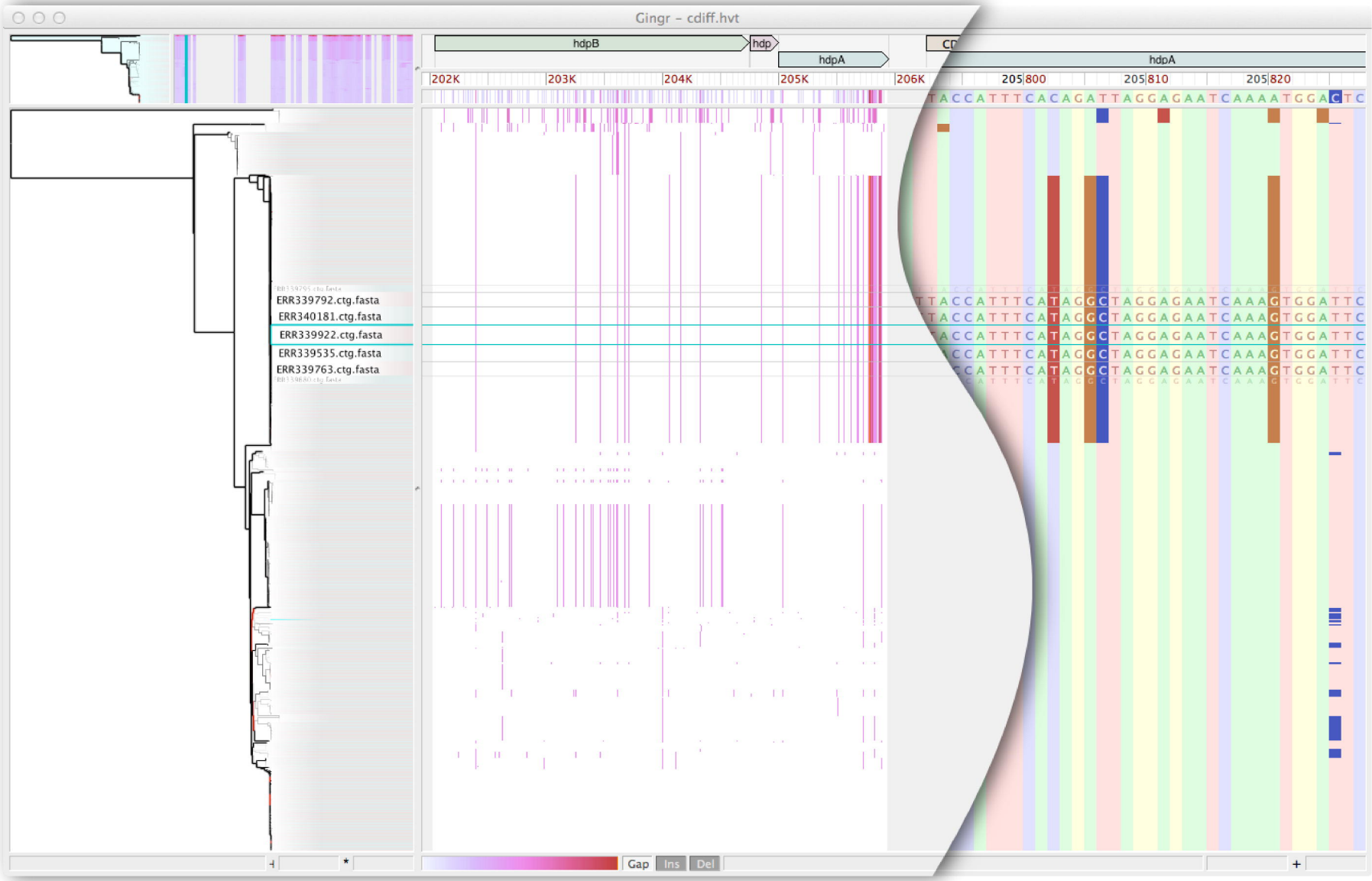
Time: method runtime from input to output. Core: size of the aligned core genome measured in base pairs. LCBs: number of locally colinear blocks in the alignment.

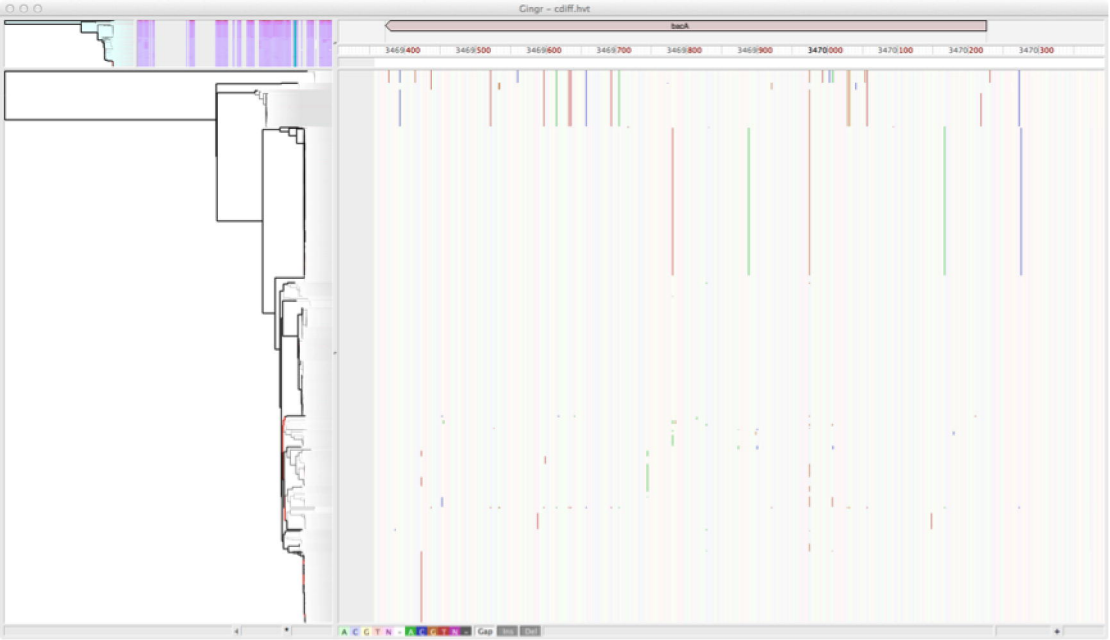
Method	Time	Core bp	LCBs
Parsnp	**0.3 min**	**1 428 407**	**1171**
Mugsy	100 min	1 590 820	2394
Mauve	377 min	1 568 715	1366
NUCmer+TBA	80 min	1 457 575	27075

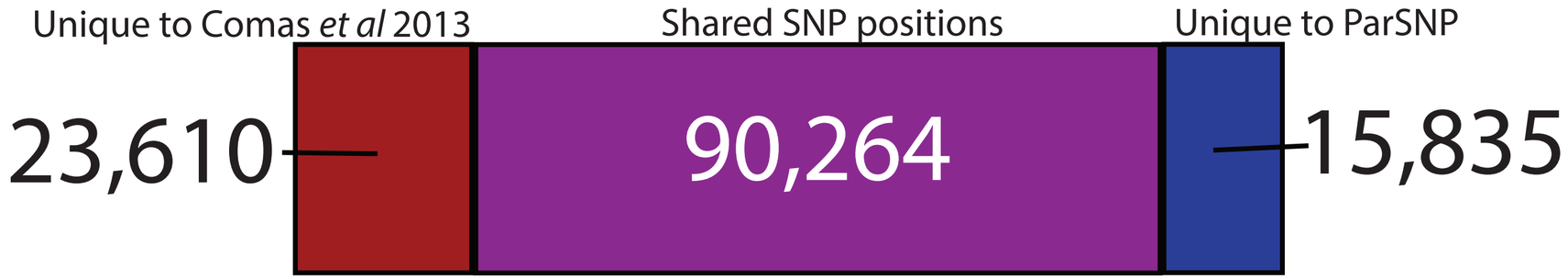

Mugsy(closed)
Mauve(closed)
ParSNP(closed)
Smalt
BWA
Mauve
ParSNP
Mugsy
TPR
1.00
0.99
0.98
0.97
0.96
0.95
0.000
0.002
0.004
0.006
0.008
FDR

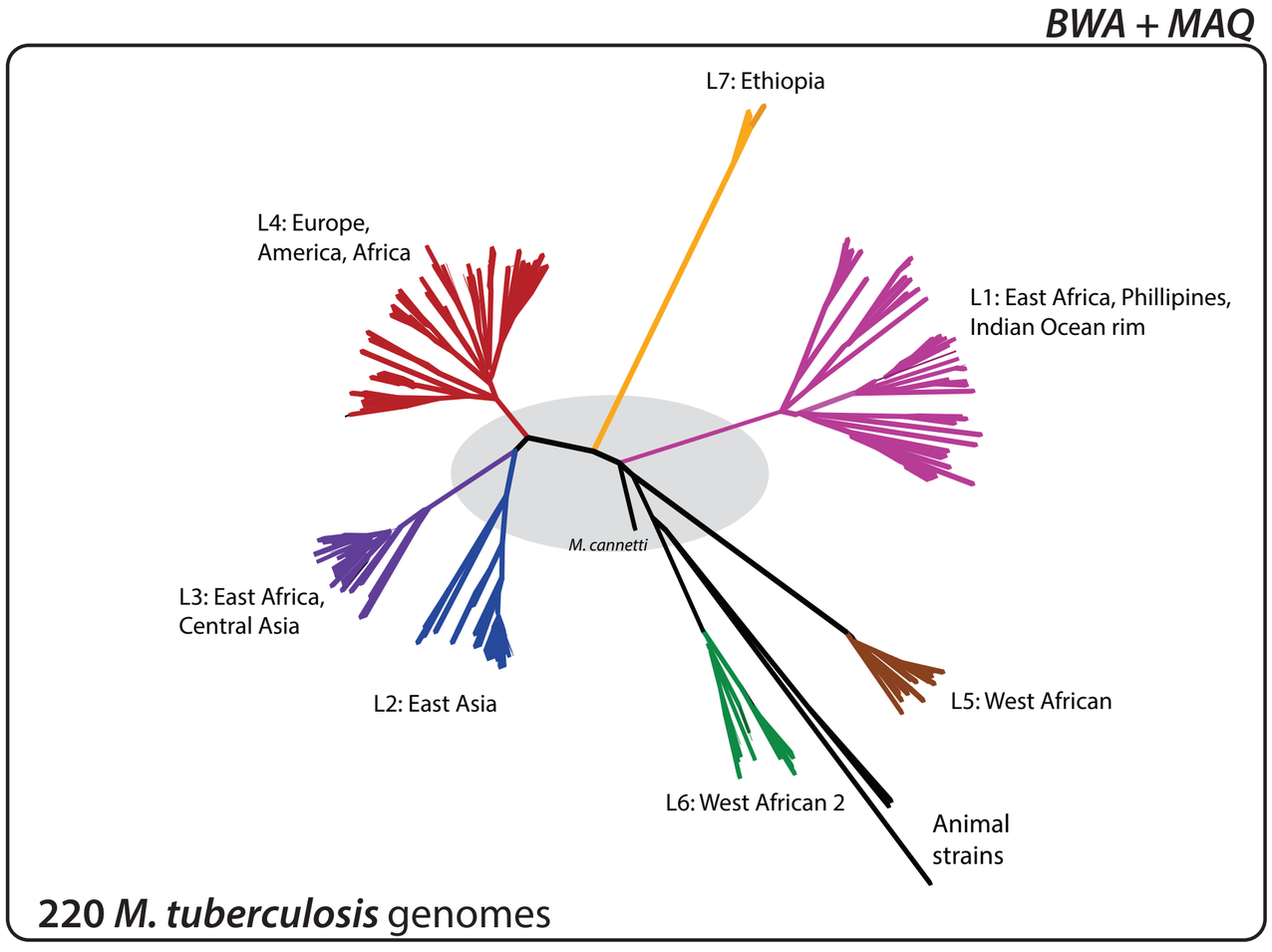

A
AA1
AA2
AB1
AB2
AC1
AC2
AD1
AD2
BA1
BA2
BB1
BB2
BC1
BC2
BD1
BD2
CA1
CA2
CB1
CB2
CC1
CC2
CD1
CD2
DA1
DA2
DB1
DB2
DC1
DC2
DD1
DD2
Mauve
Parsnp
BWA
Smalt
Parsnp$_{full}$

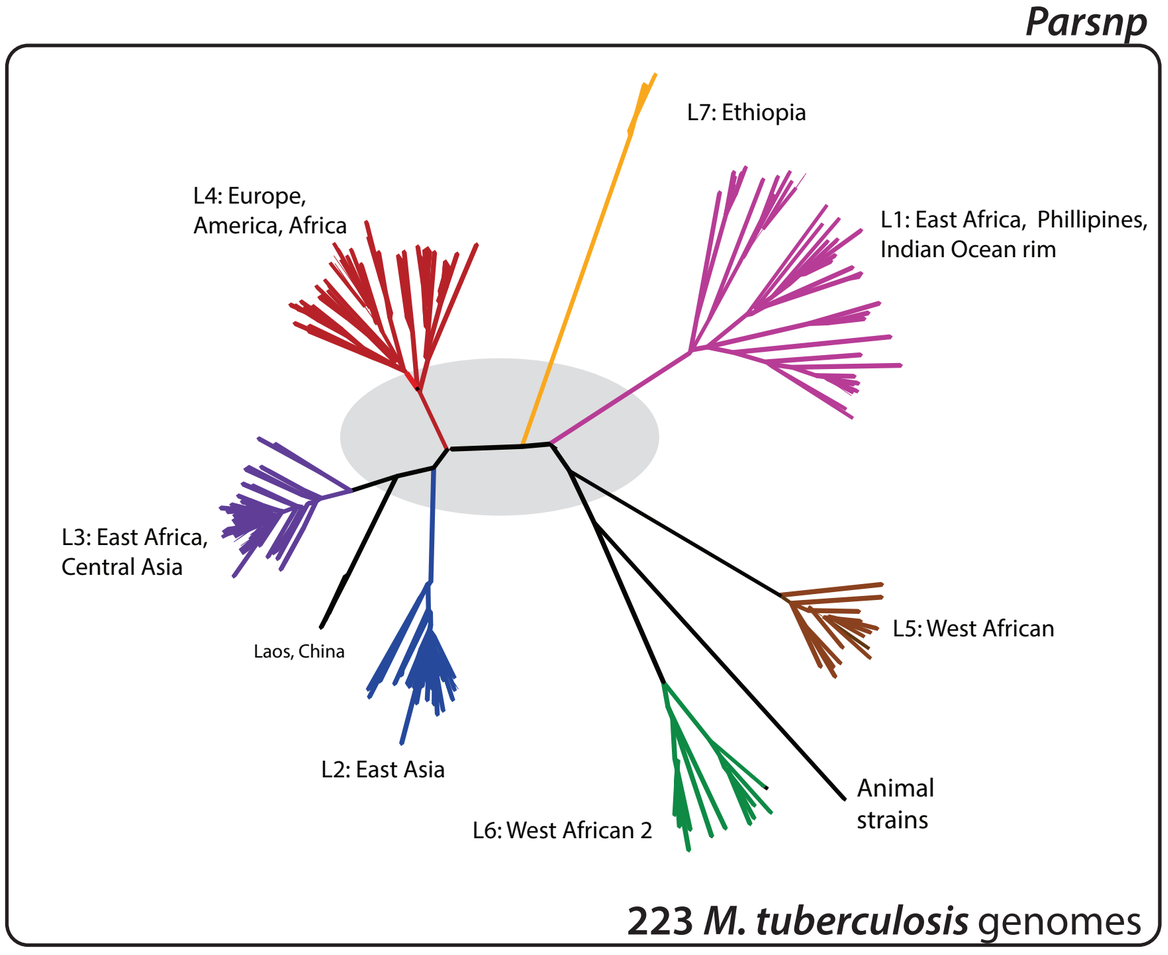

B
AA1
AA2
AB1
AB2
AC1
AC2
AD1
AD2
BA1
BA2
BB1
BB2
BC1
BC2
BD1
BD2
CA1
CA2
CB1
CB2
CC1
CC2
CD1
CD2
DA1
DA2
DB1
DB2
DC1
DC2
DD1
DD2
Mugsy
Mauve
Parsnp
BWA
Smalt
Parsnp$_{full}$

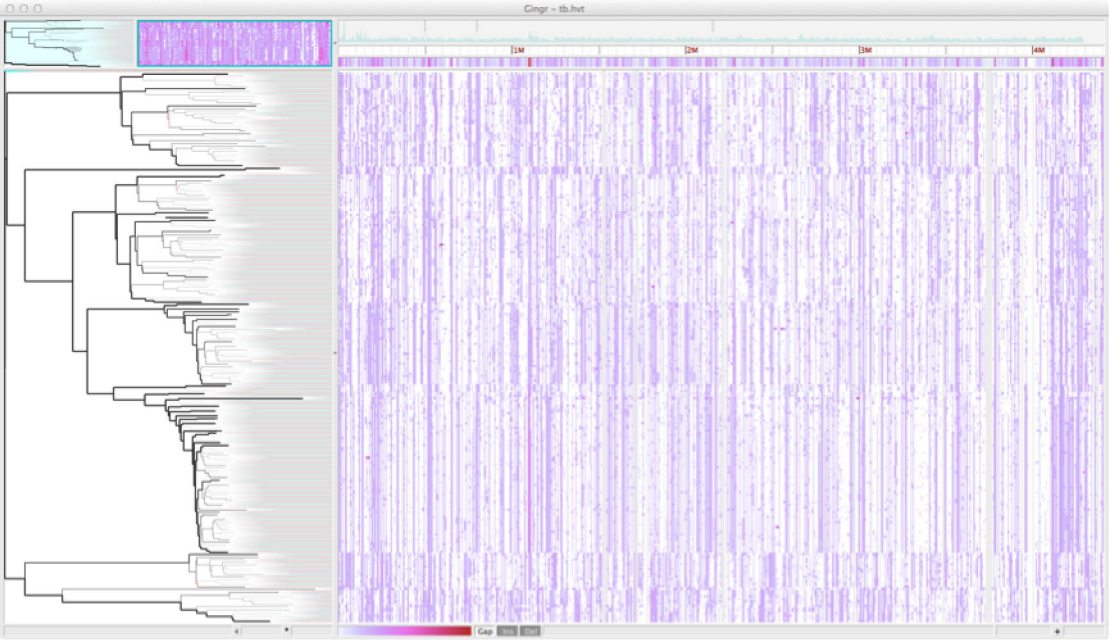

Gingr – cdiff.hvt
hdpB
hdp
hdpA
hdpA
202K
203K
204K
205K
206K
205 800
205 810
205 820
ERR339792.ctg.fasta
ERR340181.ctg.fasta
ERR339922.ctg.fasta
ERR339535.ctg.fasta
ERR339763.ctg.fasta
Gap
Ins
Del


Gingr - cdiff.hvt
tacA
3469400
3469500
3469600
3469700
3469800
3469900
3470000
3470100
3470200
3470300
A C G T N - A C G T N -
Gap
Ins
Del


Unique to Comas *et al* 2013
Shared SNP positions
Unique to ParSNP
23,610
90,264
15,835


BWA + MAQ
L7: Ethiopia
L4: Europe,
America, Africa
L1: East Africa, Phillipines,
Indian Ocean rim
M. cannetti
L3: East Africa,
Central Asia
L2: East Asia
L5: West African
L6: West African 2
Animal
strains
220 *M. tuberculosis* genomes


Parsnp
L7: Ethiopia
L4: Europe,
America, Africa
L1: East Africa, Phillipines,
Indian Ocean rim
L3: East Africa,
Central Asia
Laos, China
L5: West African
L2: East Asia
L6: West African 2
Animal
strains
223 *M. tuberculosis* genomes


Gingr – tb.hvt
1M
2M
3M
4M
Gap
Ins
Del