bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline

Gerry Tonkin-Hill[1,2], Neil MacAlasdair[1,3], Christopher Ruis[3,4,5], Aaron Weimann[3,4,5,6], Gal Horesh[1], John A. Lees[7], Rebecca A Gladstone[2], Stephanie Lo[1], Christopher Beaudoin[8], R Andrés Floto[4,9], Simon D.W. Frost[10,11], Jukka Corander[1,2,12*], Stephen D. Bentley[1*], and Julian Parkhill[3*]

[1]Parasites and Microbes, Wellcome Sanger Institute, Cambridge, United Kingdom
[2]Department of Biostatistics, University of Oslo, Blindern, 0317, Norway
[3]Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom
[4]Molecular Immunity Unit, Department of Medicine, University of Cambridge, Cambridge, United Kingdom
[5]Medical Research Council (MRC) – Laboratory of Molecular Biology, Cambridge, United Kingdom
[6]European Bioinformatics Institute, Cambridge, United Kingdom
[7]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, United Kingdom
[8]Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom
[9]Cambridge Centre for Lung Infection, Royal Papworth Hospital, Cambridge, CB23 3RE, United Kingdom
[10]Microsoft Research, Redmond, WA 98052
[11]London School of Hygiene & Tropical Medicine
[12]Helsinki Institute for Information Technology HIIT, Department of Mathematics and Statistics, University of Helsinki, 00014 Finland
[*]These authors contributed equally.

Population-level comparisons of prokaryotic genomes must take into account the substantial differences in gene content, resulting from frequent horizontal gene transfer, gene duplication and gene loss. However, the automated annotation of prokaryotic genomes is imperfect, and errors due to fragmented assemblies, contamination, diverse gene families and mis-assemblies accumulate over the population, leading to profound consequences when analysing the set of all genes found in a species. Here we introduce Panaroo, a graph based pangenome clustering tool that is able to account for many of the sources of error introduced during the annotation of prokaryotic genome assemblies. We verified our approach through extensive simulations of de novo assemblies using the infinitely many genes model and by analysing a number of publicly available large bacterial genome datasets. Using a highly clonal *Mycobacterium tuberculosis* dataset as a negative control case, we show that failing to account for annotation errors can lead to pangenome estimates that are dominated by error. We additionally demonstrate the utility of the improved graphical output provided by Panaroo by performing a pan-genome wide association study in *Neisseria gonorrhoeae* and by analysing gene gain and loss rates across 51 of the major global pneumococcal sequence clusters. Panaroo is freely available under an open source MIT licence at https://github.com/gtonkinhill/panaroo.

bacteria | pangenome | prokaryote | clustering | horizontal gene transfer
Correspondence: ***gt4@sanger.ac.uk***

Background

Prokaryotic genome evolution is driven by both the transfer of genetic material vertically from parent to offspring as well as by horizontal gene transfer between organisms (1). Large population sequencing studies of bacteria have confirmed that this results in large scale differences in intraspecies genome content (2). This has led to the description of the pangenome, the set of all genes that have been found in a species as a whole (3). Within the pangenome, genes are often then described as being part of the 'core' genome, the set of genes present in all members of a species, or the non-core ('accessory') genome. A common problem when inferring the pangenome of bacterial genomic datasets is the classification of homologous genes, usually defined by a percentage shared sequence identity, into either orthologous or paralogous clusters. Orthologs are homologous genes descended from the same ancestral sequence in the common ancestor, and not via gene duplication or acquisition. When analysing bacterial pangenomes we are often interested in not just the function of a gene or protein but also its location, as two nearly identical genes could be under differential regulation at different locations in the genome. Many programs for pangenome analysis therefore use location information to further identify paralogs, which occur when two genes descend from the same ancestral sequence due to gene duplication or when a homolog has been acquired horizontally.

Previous approaches for inferring the pangenome include Roary, OrthoMCL, PanOCT, PIRATE, PanX, PGAP, COGsoft, and MultiParanoid (4–10). The majority of methods for determining the pangenome tend to make use of one of two similar approaches (see Supplementary Figure 1). Most start by inferring similarity between predefined gene sequences using a homology search tool such as CD-HIT, BLAST or DIAMOND (11–13). Using this output a pairwise distance matrix is created and genes are then clustered into orthologous groups using either the popular Markov Clustering algorithm (MCL) or by looking at triangles of pairwise best hits (BeTs) (14, 15). A subset of these methods then continues by using the neighbourhood or genomic context of each gene to further split orthologous clusters into paralogs.

As bacterial genomic population studies have grown larger there has not been a corresponding increase in genome annotation accuracy or genome assembly contiguity. Thus as these databases have grown, so has the number of erroneous gene annotations. This can have profound implications for the resulting estimates of the pangenome whereby a higher number genomes leads to a higher number of errors (16, 17). Such errors can cause difficulties in any downstream mod-

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

eling of the pangenome, such as the modeling of negative frequency-dependent selection (NFDS) acting through the loci in the accessory genome (18, 19). Errors can be introduced into pangenome analyses by fragmented assemblies, mis-annotation, contamination and mis-assembly. Denton et al., have shown that fragmented assemblies were the major cause of inflated gene numbers in draft eukaryotic genomes (17). Whilst errors often lead to inflations in the estimates of the size of the accessory genome they can also lead to missing genes when the annotation software fails to identify a gene or where the gene is fragmented by a break in the assembly, which reduces the estimated size of the core genome. Many current pan-genome inference algorithms have not been subjected to rigorous verification using simulated data. Consequently, their ability to deal with the errors occurring in the initial genome annotations has received limited attention.

Here, we present an alternative approach to inferring the pangenome, Panaroo, which makes use of a graph based algorithm to share information between genomes, allowing us to correct for many of the sources of annotation error. Panaroo leverages the additional information provided by each genome in a dataset to improve annotation calls, and as a result, the clustering of orthologs and paralogs within the pangenome. We also provide a number of pre- and post-analysis scripts which further enrich the analysis package we provide, allowing integrated data quality control, gene association analysis, and to allow for the comparison of pangenomes between species. As Panaroo constructs a full graph representation of the pangenome, we are able to investigate structural variations within the resulting graph, allowing for associations between structural variations and phenotypes to be called. We demonstrate the success of the algorithm through extensive simulation using the Infinitely Many Genes model (20) and by analysing a diverse array of large bacterial genomic datasets including the major clades of the Global Pneumococcal Sequencing (GPS) project (21). We compare the output of Panaroo with the previous gold standard methods for analysing the pangenome and show that Panaroo produces superior ortholog clusters, often leading to both significant reductions in the size of the estimated accessory genome and increases in the size of the core genome.

Results

Overview. Panaroo builds a full graphical representation of the pangenome, where nodes are clusters of orthologous genes (COGs) and two nodes are connected by an edge if they are adjacent on a contig in any sample from the population. Using this graphical representation, Panaroo corrects for errors introduced during annotation by collapsing diverse gene families, filtering contamination, merging fragmented gene segments, and refinding missing genes (Figure 1). Panaroo generates the initial gene clusters using CD-HIT to cluster the collection of all gene sequences in all samples (11). Paralogs are then split by only allowing each genome to be present once in each cluster. Fragmented or mistranslated genes are identified and merged based on neighbourhood information of each node (22). Diverse gene families are identified using a relaxed alignment threshold along with neighbourhood information obtained from the graph. Finally, genes potentially missing from one or more samples are identified in the graph and the contig sequence near neighbouring nodes is searched to check for the presence of the gene.

Panaroo takes annotated assemblies in GFF3 format as input and generates a variety of output formats including a gene presence absence matrix (as in Roary) as well as a fully annotated graph in GML format for viewing in Cytoscape or other graph visualisation software (23). The Panaroo package includes a number of pre- and post-processing scripts that can be used for initial quality control as well as for determining pangenome size, gene gain and loss rates and to identify coincident genes. Panaroo interfaces easily with many other pangenome analysis packages including the latest version of pyseer allowing for associations between phenotypes and gene presence/absence as well as structural variation in the graph to be investigated (24). The package is written in python and is available under an open source MIT licence from https://github.com/gtonkinhill/panaroo/.

Corrected analysis of a *Mycobacterium tuberculosis* Outbreak in London. To assess the effectiveness of Panaroo and the impact of annotation errors on other pangenome inference methods we analysed a large outbreak of highly clonal, isoniazid resistant *Mycobacterium tuberculosis* (Mtb) in London (25). Mtb exhibits a very low mutation rate and is understood to have a 'closed' pangenome. Due to the short timescale of the outbreak, the maximum pairwise SNP distance within this dataset was 9. As we would expect to find no pangenome variation, this dataset provides a useful control to compare the different pangenome tools.

We ran each of the pangenome inference methods on all 413 Mtb genome assemblies after first annotating them using Prokka (26). Panaroo identified both the highest number of core genes and the smallest accessory genome (Figure 2), consistent with the established biology of Mtb and a highly clonal dataset (27, 28). In contrast, PanX, PIRATE, COGsoft and Roary all reported inflated accessory genomes ranging in size from 2584 to 3670 genes representing a nearly ten-fold increase to that reported by Panaroo. The small number of accessory genes that Panaroo did predict mostly consisted of core genes where the algorithm was unable to refind the genes in a subset of the assemblies. The majority of the difference between the methods was driven by genes being fragmented during assembly (59%; see Supplementary Methods). A smaller subset of genes were only called in a small minority of the isolates despite the underlying sequence being nearly identical (10%). Whilst some of these differences could be due to frame shifts in the PE/PPE genes, 27.9% of the isolates were indistinguishable with only one isolate being more than 5 SNPs from this major clone. We found that the majority of the difference was due to the annotation algorithm optimising for each isolate individually, leading to inconsistent gene calls. However Panaroo's consensus approach helps to resolve these discrepancies. The magnitude of the difference observed in this dataset suggests that failing to account for

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

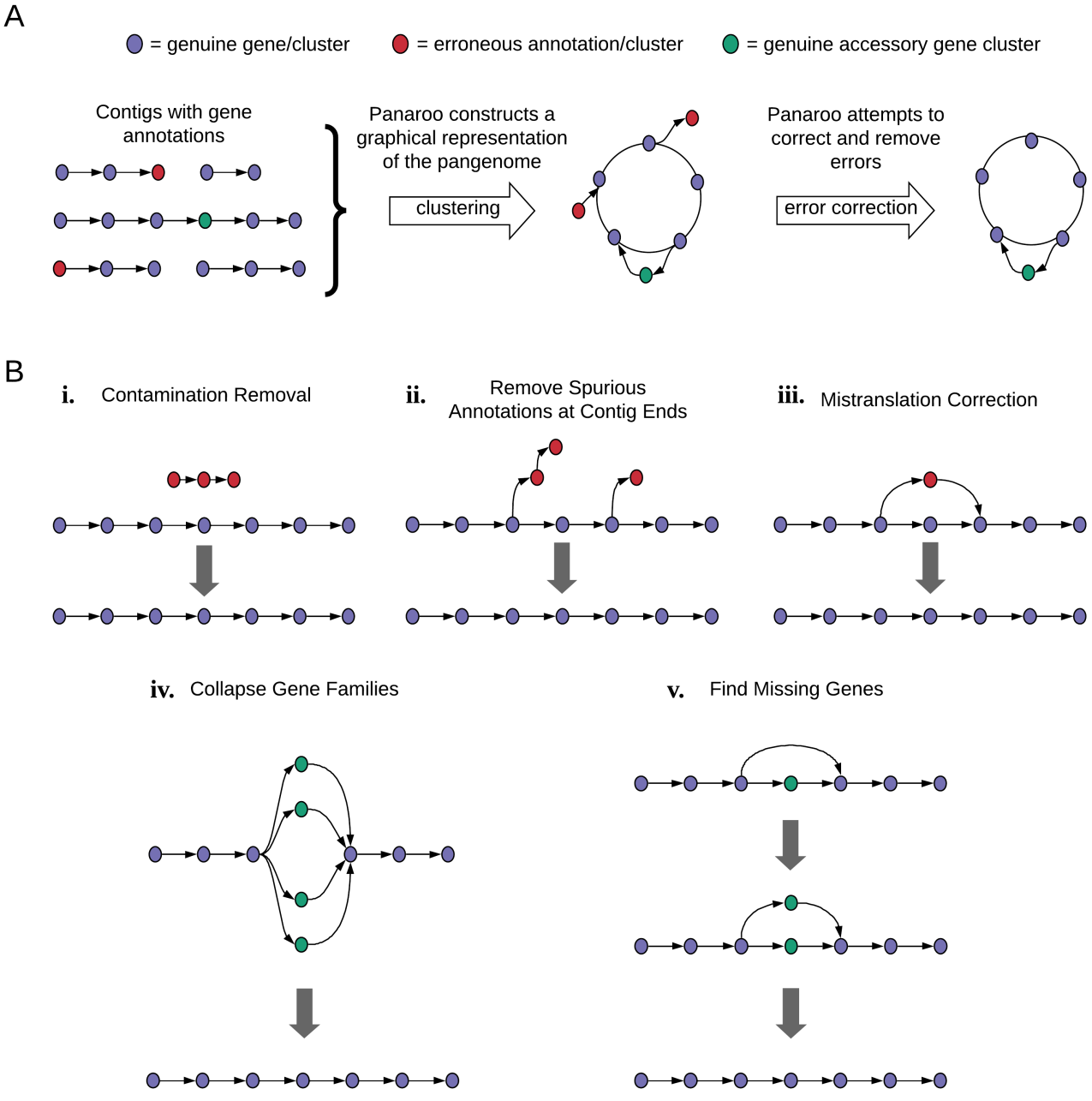


Fig. 1. A: An overview conceptualizing the problem with current gene annotation methods and the stages Panaroo uses to correct for annotation errors. **B:** Expanded specific stages in the process. **(i)** Contamination appears in the graph as poorly supported components. In the default mode, Panaroo removes contamination by recursively removing poorly supported nodes of degree 1. **(ii)** Genes are often mis-annotated near contig breaks (17). Panaroo corrects such mis-annotations by recursively removing poorly supported nodes of degree 1. **(iii)** Panaroo corrects cases where the same DNA sequence has been translated in multiple reading frames into a single gene by clustering concomitant genes at the DNA level. **(iv)** Panaroo uses context and a lower clustering threshold to combine diverse gene families into a single gene. **(v)** Annotation algorithms may predict a gene in some but not all samples, even when the samples share exactly the same DNA sequence. Panaroo finds missing genes by searching for the gene sequence in the surrounding DNA.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

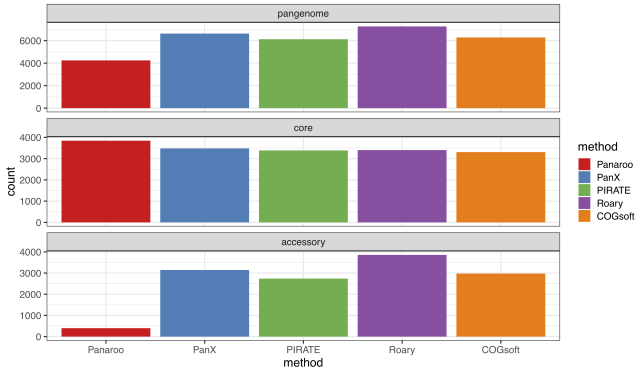


Fig. 2. Pangenome counts for 413 *Mycobacterium tuberculosis* genomes from an outbreak in London (25). The maximum pairwise SNP distance between these isolates was 9, suggesting extremely limited variation. Consequently, we would expect a very limited accessory genome and a core genome of approximately 4000 genes. All tools with the exception of Panaroo found in excess of 2500 accessory genes, which can be attributed to annotation errors.

annotation errors can have profound impacts on the resulting estimates of the pangenome.

Superior Performance on Simulated Populations. To further assess the ability of the different methods to identify the correct gene presence/absence matrix, we simulated pangenomes using the *Escherichia coli* reference genome ASM584v2 (accession number NC000913) and the Infinitely Many Genes model (20, 29). To more accurately simulate the kind of errors that typical annotation pipelines produce, we simulated short read assemblies from these pangenomes using Mason, ART and SPADES (30–32). A more detailed description is given in the methods. We conducted five simple and two more complicated simulations, each with three replicates (Supplementary Table 1). In the simple simulations, the gene gain/loss rate was varied with lower rates corresponding to a larger core genome and smaller accessory genome and higher rates corresponding to a larger accessory genome. The mutation rate of the accessory genome was also varied. In addition, we simulated two more complicated datasets, one of which had an increased level of fragmentation of the assembly by fragmenting the input genome prior to the NGS simulation. The second more complex simulation included contamination by randomly adding in short fragments of the *Staphylococcus epidermidis* reference genome, which is a common contaminant.

Figure 3a indicates the number of gene clusters which contained errors for each of the scenarios. Such errors included both genes that were incorrectly annotated as well as gene sequences that were incorrectly clustered together. Most methods performed fairly well when applied to the output from the simple simulation. All methods include some errors due to genes never being annotated except in the original reference. As each method relied on the same input files this was consistent between methods.

For the simple simulations, PanX and Panaroo produced the fewest errors, followed by PIRATE, Roary and COGsoft. Roary was the most sensitive method to the substitution rate, with higher rates leading to more errors. This can be attributed to its reliance on a strict BLAST e-value threshold. COGsoft gave variable results, performing poorly on pangenomes with larger accessories suggesting it may over collapse genes. This interpretation was further supported in our analysis of a diverse *Klebsiella pneumoniae* dataset (see below).

Whilst most methods were able to perform adequately on relatively error-free simulated data, the introduction of more realistic significant sources of annotation error had a large impact. Figure 3b indicates the resulting error counts after simulating both contamination and highly fragmented assemblies. Here, the importance of Panaroo's multiple annotation error correction approaches becomes apparent. As expected, when small amounts of contaminating *S. epidermidis* DNA were added to the simulated NGS data all methods except Panaroo and COGsoft incorrectly called a larger accessory genome. This is due to their inability to account for and remove contaminating contigs. In contrast, Panaroo achieved error rates similar to that found for the clean assemblies. COGsoft had a similar number of total errors to the other programs but rather than calling a larger accessory genome tended to incorrectly merge the contamination with other genes.

The highly fragmented assemblies led to the largest error rates in each pangenome analysis tool. Fragmentation can lead to gene annotation software such as Prokka miscalling genes near the ends of contigs. It can also impact on the consistency of the training step in some annotation algorithms. This resulted in a large increase in the estimated accessory genome size for all methods except Panaroo. Similarly, miscalling can lead to genes being left unannotated resulting in smaller estimates of the core genome. In both cases Panaroo's error correction and refinding steps were able to accurately recover the true pangenome, while PanX, COGsoft, PIRATE and Roary all produced nearly an order of magnitude higher error rates. This result mirrors that observed in the analysis of the highly clonal *M. tuberculosis* outbreak, helping to confirm the impact that such errors can have on pangenome estimates.

Greater Internal Consistency in a Diverse *Klebsiella pneumoniae* Collection. We then went on to compare each method on a more complex real dataset – 328 globally sourced *Klebsiella pneumoniae* genomes from both human and animal hosts (2). *K. pneumoniae* is a highly diverse gram-negative bacterium that can colonize both plants and animals and has previously been found to have a large pangenome (2). The high recombination rate and often multiple plasmids per bacterium complicates analysis of the *K. pneumoniae* pangenome. Nine of the 328 isolates were identified as outliers by the Panaroo quality control script due to the number of contigs or number of genes they contained (see Supplementary Figures 3-5). These isolates were removed before running each algorithm. Figure 4a indicates the resulting total, core and accessory gene counts inferred by each method, using the 99% presence threshold for core genes as used in Roary (4).

As species such as *K. pneumoniae* are known to have many

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

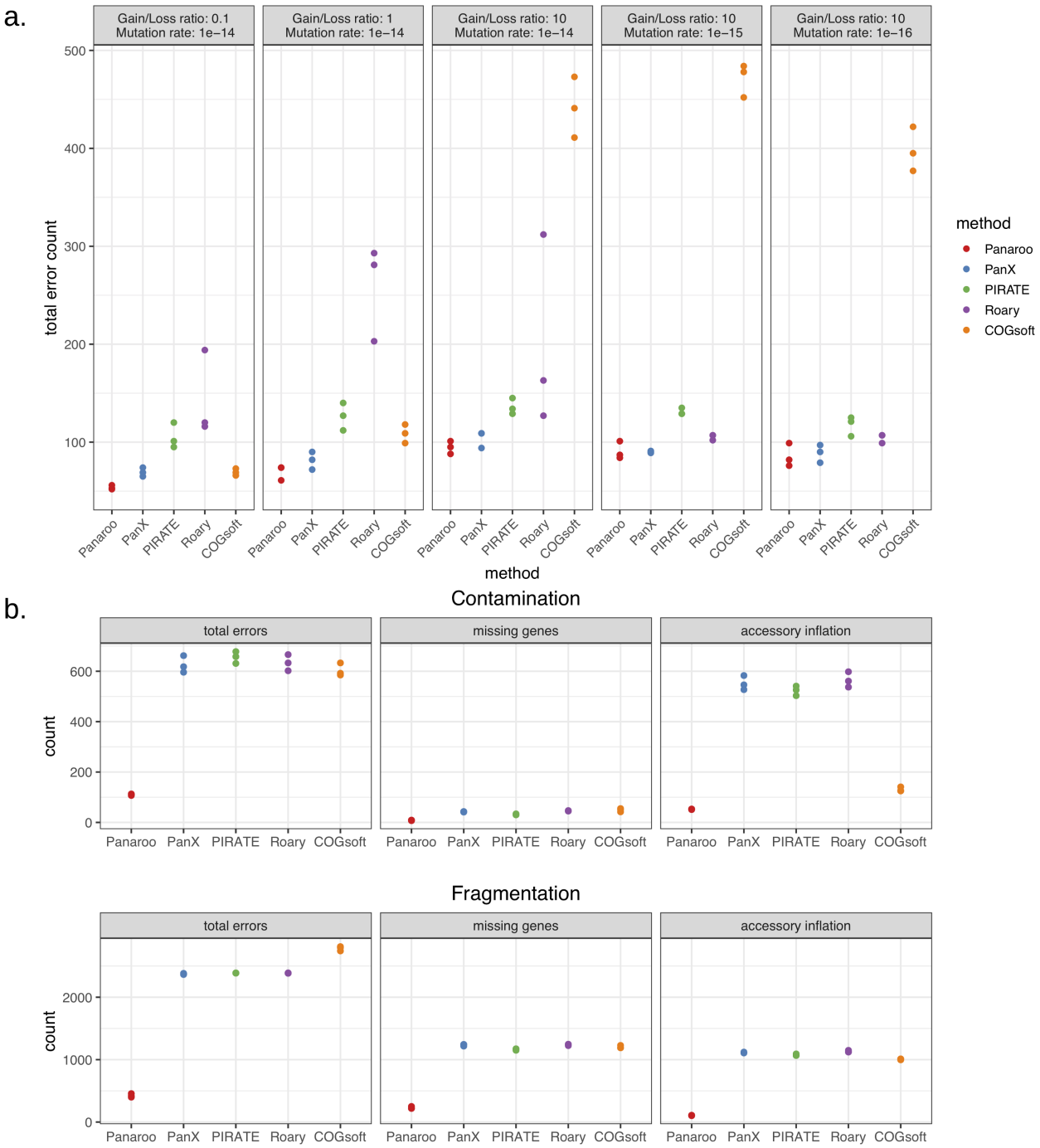


Fig. 3. Error counts for the different algorithms after comparing with simulated data on different scenarios. Accessory genome inflation refers to the number of erroneous clusters that do not correspond to any simulated gene cluster. Even in simulations of pangenome variation from a single *E. coli* reference with only relatively simple sources of error, a), panaroo outperforms other methods across a variety of gene gain/loss rates and mutation rates. In more realistic simulations of sequencing data, b), the only method with reasonable control of the error rate is panaroo.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

rare plasmids which are difficult to distinguish from contamination, we developed a 'sensitive' mode for when the default 'strict' mode of Panaroo contamination filter can be overly stringent. Panaroo identified the highest number of core genes in both its default and sensitive modes, 3372 and 3376 respectively. Hence for these genomes there was only a minor difference in the estimated core between the two pipeline options. Roary identified the smallest core genome of 1800 genes. Given the result of the simulations, this is likely due to gene clusters being incorrectly split into multiple smaller clusters, as the default Roary pairwise identity threshold of 95% is too stringent for such a diverse dataset. PIRATE relaxes the strict threshold required in Roary and it identified a similar number of core genes to Panaroo (3318) but a smaller number of accessory genes than both the Panaroo (sensitive) and PanX methods which agreed more closely with the original estimates in Holt *et al.* (2).

Whilst there is no gold standard with which to compare these results, we can look at the gene annotations within clusters to identify cases where a gene cluster contains multiple different annotations, which would suggest separate gene clusters have been incorrectly merged. Figure 4b indicates the number of conflicting annotations in the clusters of each method. As gene fragments and genes annotated as "hypothetical" are often the result of errors and thus can have erroneous annotations, we did not consider conflicts that involved these. Panaroo in both its default and sensitive modes had the lowest number of conflicting annotations. PanX had the second lowest number whilst COGsoft recorded the highest number of conflicts which is consistent with the tendency of its method to over-cluster genes. Overall, Panaroo identified a larger core genome and fewer conflicting annotations than any other method showing that its error correction approach is also suitable for diverse datasets of highly recombinogenic bacteria.

Pyseer Association Analysis with Panaroo Finds Antibiotic Resistance Mechanisms. Panaroo provides a number of outputs as well as post processing scripts for analysing the cleaned pangenome graph. Panaroo outputs both a gene presence/absence matrix as well as structural variation presence/absence matrix that can be used as input to pyseer or Scoary for association analyses (24, 33). Panaroo generates structural variation calls by identifying distinct consecutive triplets of gene families in the graph that describe different paths through a node (see Figure 5a). As larger insertion and deletion events will only be represented once in the structural presence/absence matrix rather than repeatedly for each gene, this approach increases the power of such association analyses. The approach also identifies associations with large structural re-arrangements although these are often more difficult to interpret. Once a significant association between a gene triplet and a phenotype of interest have been identified, the context of the structural rearrangement can be investigated manually by interrogating the pangenome graph in Cytoscape (23).

To validate the pan-genome wide association study (pan-GWAS) and pan-genome structural variant association study

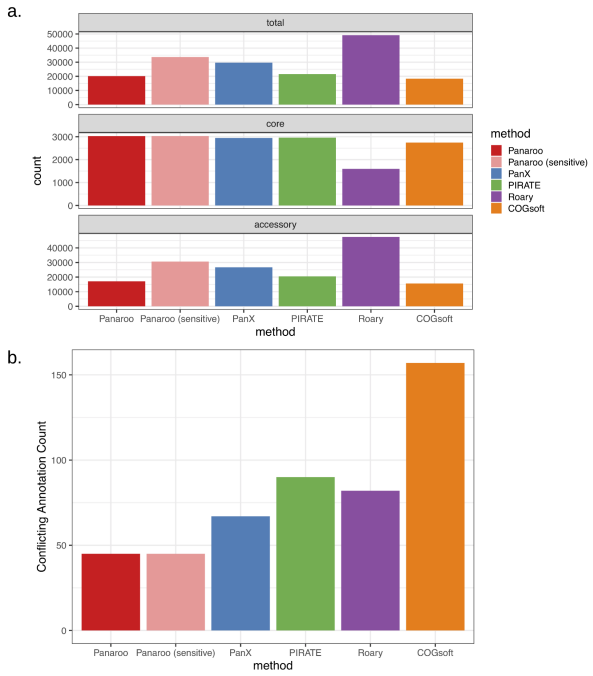


Fig. 4. a.) the estimated pangenome, core and accessory sizes from the different algorithms in the global *K. pneumoniae* dataset. (b) The number of conflicting gene annotations in the inferred clusters of the different algorithms.

(sv-pan-GWAS) pipelines, we ran panaroo on the Euro-GASP collection of 1054 *Neisseria gonorrhoeae* isolates collected from 20 countries across Europe from September to November 2013 (34). We combined the Panaroo output with antimicrobial MIC testing results for seven different antibiotics performed in the original study and carried out association studies on the gene presence-absence patterns and structural variants using pyseer (24).

The gene presence-absence pan-GWAS approach returned 67 genes (Supplementary Table 2) associated with various antibiotics (adjusted p-value ≤ 0.05). This included many probable candidates for genes causing resistance, including an uncharacterised ABC transporter (group_464), for penicillin resistance. ABC transporters are a common resistance mechanism against ribosome-targeting antimicrobials, as they can function as efflux pumps (35).

The structural variant pan-GWAS returned 138 triplets (Supplementary Table 3) associated with antibiotic resistance (adjusted p-value ≤ 0.05). These included many triplets containing phage-associated, transposase, or pilin genes, all of which are known to be mobile within the genome.

Among these, the sv-pan-GWAS identified a number of insertions and deletions of whole genes which were associated with antibiotic resistance. One of these, group_1138, is a transmembrane protein which, when inserted, is associated with ceftriaxone resistance. All four possible gene triplets bypassing or going through the insertion were significantly associated with either susceptibility or resistance depending on if they included group_1138. The mechanisms of ceftriaxone resistance in *N. gonorrhoeae* are not yet fully understood, but it has been suggested that efflux and permeability

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

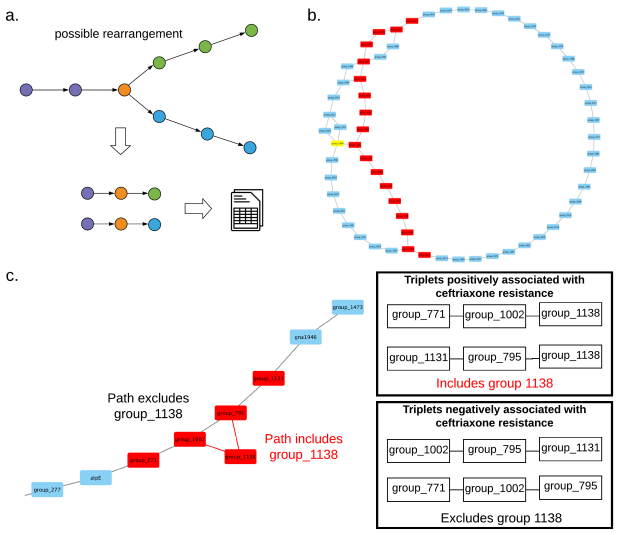


Fig. 5. a) A diagram indicating how gene triplets are called in the graph. A single genome can only pass through a node once; thus, variations in the arrangement of genes in different genomes can be called using triplets. These triplets are summarised as a binary presence absence matrix. **b)** A family of related plasmids present in the *N. gonorrhoeae* pangenome gene network. The path highlighted in red contained 4 structural variant gene triplets significantly negatively associated with tetracycline resistance, or associated with tetracycline susceptibility by a structural variant pan-GWAS (all adjusted p-value < 0.05). The gene highlighted in yellow, group_1999, was found to be a tetM resistance gene. **c)** A subsection of the *N. gonorrhoeae* pangenome gene network of the region surrounding gene group_1138. The presence of gene triplets (group_771-group_1002-group_1138) and (group_1131-group_795-group_1138) is positively associated with tetracycline resistance while the triplets (group_1002-group_795-group_1131) and (group_771-group_1002-group_795) are negatively associated with tetracycline resistance (all adjusted p-value < 0.05).

must play a role (36). Group_1138, as it is a transmembrane protein, could have either of these functions.

The sv-pan-GWAS approach allows for closely related genetic architectures to be disentangled, including highly related plasmids and phages. For example, analysis of the pangenome graph showed that a common *N. gonorrhoeae* plasmid present 430 times in this dataset is actually a family of several closely related plasmids. These highly similar plasmids share the majority of their genes, but there are several differences in gene content, which appear as bubbles in the pangenome graph (Figure 5B). One of the plasmid versions (highlighted in red in Figure 5B) is negatively associated with tetracycline resistance, with four gene triplets significantly negatively associated with this phenotype in the sv-pan-GWAS. The other plasmid variants each contain group_1999, a tetM tetracycline resistance gene, providing a mechanism to explain the differential resistance profiles. Together, these analyses demonstrate that multiple members of the same plasmid family with different resistance profiles are circulating in the European *N. gonorrhoeae* population, and illustrate the value of an the sv-pan-GWAS approach.

Improved Methods for Analysing Pangenome Evolutionary Dynamics. The higher accuracy obtained by Panaroo allows for the comparison of gene gain and loss rates between lineages and species as well as the more accurate inference of pangenome size. Whilst it is common practice to plot gene accumulation curves in the analysis of pangenomes, these are not robust to errors and fail to account for sampling biases and population structure. Thus, accumulation curves should not be used to compare pangenome characteristics of different lineages or species. Recently, a number of phylogenetically informed methods for investigating pangenome dynamics have been published, including the Infinitely Many Genes (IMG) model and the Finitely Many Genes (FMG) model (8, 20, 37). Both of these approaches account for the diversity of the sample and have been implemented as post-processing scripts in Panaroo.

To demonstrate the utility of using the corrected pangenome graph to infer gene gain/loss rates and pangenome size, we used the FMG model to investigate 51 of the major Global Pneumococcal Sequence Clusters (GPSCs) for which reliable dated phylogenies could be constructed (21). The major clades of the pneumococcus have distinct accessory gene profiles (38). We ran Panaroo on each GPSC separately and used the resulting gene presence/absence matrix with the corresponding dated phylogeny to infer gene gain and loss rates for each cluster. We compared the inferred parameters with other variables of interest calculated by Gladstone *et al.* (21), including the inferred recombination rate (r/m), odds ratio of invasive disease and the number of distinct serotypes for each cluster. The parameters along with these variables are plotted in Supplementary Figure 2. We found that the estimated effective pangenome size correlated positively with the recombination rate of a cluster (Spearman correlation coefficient 0.53, p<0.001) and the number of serotypes present in the cluster (Spearman coefficient 0.51, p=0.001). This is consistent with biological understanding of the genome diversification and gives confidence to our results, as a higher recombination rate would allow for a clade to more easily gain and lose genes, including serotype-defining gene clusters, resulting in a larger pangenome. Interestingly, GPSCs that have lower gene gain rates were more likely to have a significant odds ratio for invasive disease (p=0.04)(see Figure 6). The association with gene loss rate was weaker, although the effect was in the same direction (p=0.08). Genome reduction has previously been associated with increasingly obligate interactions with the host in multiple unrelated bacterial pathogens (39).

Computational Performance. Panaroo uses a similar level of computational resources to competing methods. Figure 7 indicates the memory and cpu time required for the analysis of 10, 100 and 1000 *N. gonorrhoeae* isolates subsampled from the Euro-GASP collection. PanX and COGsoft used the most resources with COGsoft not completing the largest dataset in under a week. Roary, PIRATE and Panaroo all performed similarly.

PREPRINT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

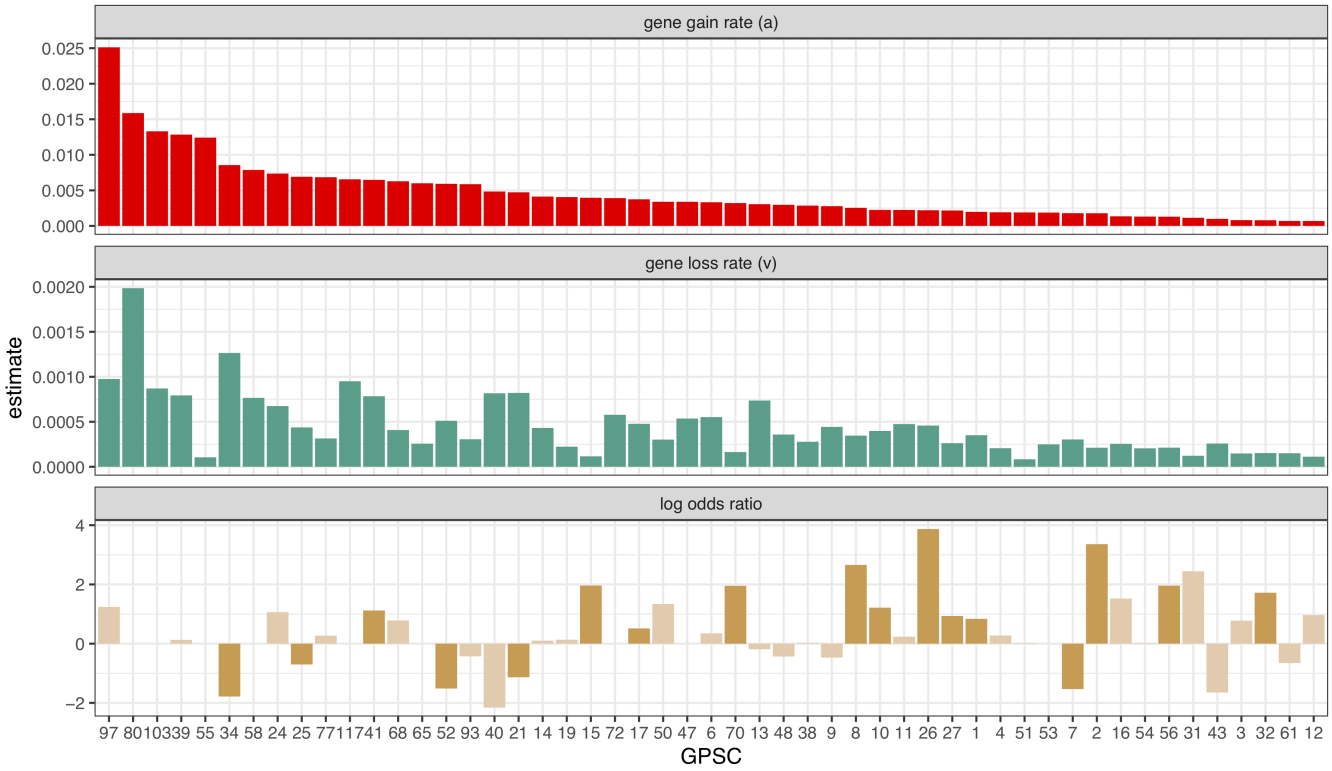


Fig. 6. The inferred gene gain and loss rates of each of the 51 major clades of the Global Pneumococcal Sequencing project plotted above the respective log odds ratio of invasive disease in that clade. Clades which had significant odds ratios in Gladstone *et al.* (21) are represented in dark yellow.

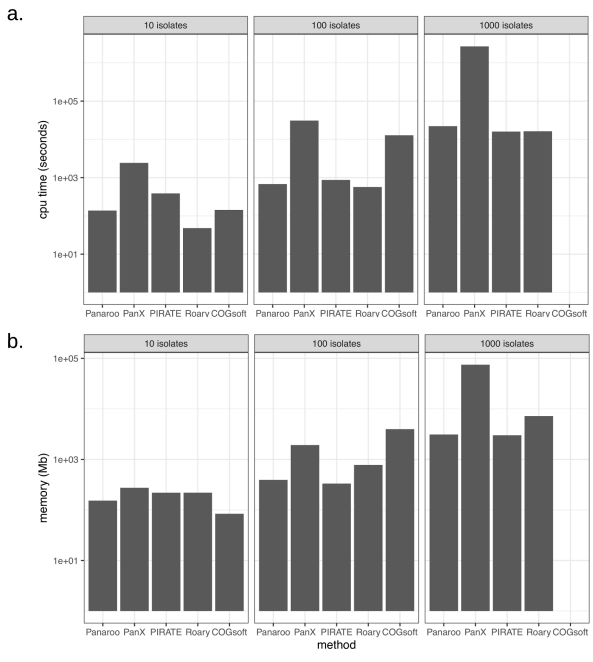


Fig. 7. the cpu time and memeory required for each of the algorithms for 10, 100 and 1000 *N. gonorrhoeae* isolates. Each tool was run with 5 cpus

Discussion

Annotation errors, fragmented assemblies and contamination represent a major challenge for pangenome analysis. We have designed Panaroo to tackle these challenges using a sophisticated framework for error-correction that leverages information across strains through a population graph-based pangenome representation. Using both simulations and well-characterised real world datasets, we demonstrated that many commonly used methods greatly inflated the size of the accessory genome while reducing the estimated size of the core genome. In contrast, Panaroo exhibited far lower error rates and reconstructed highly accurate core and accessory genomes for simulated datasets that included contamination and genome fragmentation. Analysis of both a highly conserved *M. tuberculosis* dataset and a highly diverse *K. pneumoniae* dataset indicated that Panaroo provides superior solutions in challenging real world population genomics applications.

Panaroo also includes a number of pre- and post-processing scripts for the analysis of bacterial pangenomes that assist in quality control of the input data and facilitate down-stream processing of the pangenome. We used the Panaroo pre-processing QC scripts to identify nine *K. pneumoniae* samples that were outliers based on the number of contigs or genes and excluded these samples from our analysis. We recommend that such pre-processing QC be carried out on all datasets to identify potentially erroneous samples.

We used the output from Panaroo as input to pyseer to run pan-GWAS and sv-pan-GWAS analyses on *N. gonor-*

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

rhoeae. Through this approach, we identified a deletion in the genome of *N. gonorrhoeae* in a large European collection that confers resistance to tetracycline. We demonstrated the utility of Panaroo to disentangle highly similar genetic structures through identification of a plasmid family in *N. gonorrhoeae* (Figure 5C). By combining this high resolution picture with structural variant pan-GWAS we identified that some members of this plasmid family carry tetracycline resistance and were able to accurately determine the tetM gene as the cause of resistance.

As part of the Panaroo package, we include implementations of recently proposed pangenome evolution models, which are more appropriate than the more frequently used gene accumulation curves. We demonstrated the effectiveness of such methods through the analysis of the 51 major GPSCs where we observed an association between recombination rate and pangenome size (Supplementary Figure 2). We also identified an association between pneumococcal clade invasiveness and gene gain rate.

Panaroo is written in python (versions 3.6+) and is available under the open source MIT licence from https://github.com/gtonkinhill/panaroo. The code used to produce the analyses described above along with summary data is available from https://github.com/gtonkinhill/panaroo_manuscript. The raw GFF3, FASTA and all intermediate post-processing files are available from https://doi.org/10.5281/zenodo.3599800.

Taking gene annotation errors into account is vital to recover an accurate pangenome, something previous methods have struggled to do in a systematic manner. Panaroo uses gene adjacency in a population-graph to provide a fast method for pangenome analysis, which is robust to a wide range of error sources. In the future, we plan to further improve the computational performance of Panaroo to allow it to scale to datasets involving hundreds of thousands or millions of genomes. We will also extend the post-processing tools available to analyse the resulting pangenome graph.

Methods

Panaroo algorithm. The Panaroo algorithm builds a graphical representation of the pangenome where nodes are genes and edges connect nodes if two genes appear adjacent to one another on at least one contig. The algorithm then uses this initial graph structure to perform a number of cleaning steps which correct for many of the problems encountered in genome annotation. Panaroo accepts annotated assemblies in GFF3 format as output by the popular annotation pipeline Prokka (26). Unlike similar pangenome software, Panaroo attempts to preserve the full global context of each gene in the graph. This is in contrast to other programs such as Roary (4, 7, 10) which uses only the local context surrounding genes to build the graph.

Initial graph creation. To first build the graph, Panaroo runs CD-HIT (v4.8.1) at a high sequence identity threshold (98%) (11). The resulting clusters are then either classified as non-paralogous gene clusters, if they contain at most one instance of each genome, or paralogous clusters if they contain more than one gene from any single genome. Initially, non-paralogous gene clusters are represented by a single node in the graph whilst paralogous clusters are split into a single node for every occurrence of that cluster in the dataset. For instance, if a paralogous gene appears twice in two genomes and once in another, the initial graph will contain five nodes representing that paralog. The graph is then built by connecting cluster nodes with edges between them if the two clusters appear adjacent to one another on any contig. Paralogous nodes are collapsed back into the maximum number of nodes in which those genes appear in a single genome using the global context of the graph. In the above example, this would result in the final graph having two instances of the paralog node.

Contig Ends. Fragmented assemblies can cause issues for gene annotation software, whereby genes are often misannotated near contig breaks (17). These spurious annotations appear as short paths of low support edges and nodes that end in a node of degree one that splits off from the main graph. To deal with this, Panaroo recursively removes nodes of degree one that are below a given support threshold as indicated in Figure 1.

Contamination. Contigs originating from sample contamination are generally significantly diverged from the target species pangenome. Thus, contaminating contigs tend to appear as disconnected components from the main graph with low support. To remove these, Panaroo uses the same approach as described for contig ends to recursively delete low supported nodes with less than or equal to one degree (see Figure 1). This approach has the advantage of retaining rare genes which are present in the main graph whilst removing likely contaminants. Whilst this has in general been found to be very successful it can occasionally lead to rare plasmids being removed. We have found that the benefits of removing unwanted noise far exceed the small loss in sensitivity that this approach provides. However, we also provide three settings for the algorithm with the most sensitive retaining such rare calls which can be useful when one is interested in rare plasmids.

Mistranslation Correction. Many annotation algorithms rely on an initial training phase where their parameters are adapted to the dataset at hand (40–42). Often this training is performed separately on each genome. This is the case in the Prokka pipeline, which makes use of Prodigal to perform the initial gene annotation (26, 40). This can result in an identical sequence being annotated differently in different genomes. To correct for this Panaroo checks genes that are within close proximity in the pangenome graph to determine if any are likely to be mistranslations, frame shift mutations, or pseudogenised gene copies by comparing their sequence at the nucleotide level. If two gene sequence matches at a high coverage and identity, typically 95% and 99% respectively a mistranslation is called and the gene node with the lower support is collapsed into the node with higher support.

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Collapse Gene Families. Gene families diversify at different rates due to the influence of positive and purifying selection. This makes choosing a strict sequence identity threshold for defining orthologous clusters difficult. Most pangenome analysis software relies on either a pairwise sequence identity or BLAST e-value threshold. This reliance can lead to both over clustering, where separate gene families are incorrectly merged, and over splitting where a single gene family is incorrectly split into several smaller clusters. Many approaches attempt to deal with the former of these problems by utilising contextual information to split apart clusters that have different gene neighbourhoods (4, 6). More recently, alternatives that make use of clustering at lower thresholds followed by more involved splitting techniques have been proposed (7, 8). As an alternative to these approaches we extend the idea of using gene context to the over splitting problem. Panaroo utilises gene contextual information to collapse diverse gene families that have been incorrectly split into multiple clusters during the initial pangenome graph creation. Initial gene clusters that share a common neighbour in the graph are compared at a lower pairwise sequence threshold (default 70%). If they fall within this threshold the two nodes are collapsed and the resulting node is annotated to indicate it consists of a more diverse family. We have found that utilising this additional contextual information leads to more robust clusters.

Identifying Missing Genes. Previous pangenome clustering software tools are unable to identify missing annotations. Gene annotations can be lost due to variability in model training, fragmented assemblies and mis-assemblies. Panaroo remedies this issue by identifying pairs of nodes in the pangenome graph where one node is present in a genome and its neighbour is not. The potentially missing node is then searched for in the sequence surrounding the neighbouring node. If a match of sufficient coverage and identity is found, the graph is corrected to include an annotation for this missing gene in that genome. The alignment tool edlib (v1.3.4) is used to perform these searches which enables millions of checks in a reasonable time frame (22).

Output. To allow for simple integration with existing bioinformatics pipelines Panaroo outputs many of the same file formats as Roary. This includes the same gene presence/absence file format as well as core and accessory genome alignments created using either MAFFT, Prank or Clustal Omega (43–45). In addition, Panaroo outputs a fully annotated pangenome graph in GML format for easy viewing in Cytoscape (23). Each gene node and edge is annotated with the genomes to which it belongs as well as the gene annotations given by Prokka, gene sequence and whether or not the node has been classified as being a paralog. This graph format provides a valuable tool for visually inspecting the results of Panaroo. As Panaroo attempts to build the full pangenome graph rather than only using local context, this graph is able to provide insights hidden in many of the outputs of similar tools such as Roary (4).

Structural Variation. As Panaroo constructs the full pangenome graph, it is possible to go beyond gene presence/absence and look at the underlying structure of the graph. To facilitate the analysis of this structure, Panaroo generates a gene triplet presence/absence matrix, indicating when three genes are present in a path along a genome. This is demonstrated in Figure 5a and the resulting presence/absence matrix can be used in association studies to investigate differences in rearrangements between genomes in a species. The context of each triplet can then be analysed by looking at the full graph in Cytoscape.

Pre- and Post-Processing. The Panaroo pipeline comes packaged with a number of pre and post processing scripts for analysing pangenomes. We have included a wrapper for the popular Mash and Mash screen algorithms which generates diagnostic plots for quality control prior to running the Panaroo pipeline (46, 47). These plots include a Multidimensional scaling (MDS) projection of pairwise Mash distances, interactive bar charts to investigate contamination, as well as gene and contig counts.

In addition we have included post processing scripts for estimating gene gain and loss rates using both the infinitely many genes (IMG) model (37, 48) and the finitely many genes (FMG) model of (8, 48). These are preferable to the common practice of plotting accumulation curves to indicate pangenome size as they account for the diversity and timescale of the sampled isolates. This allows for a clearer comparison between the pangenomes of different species or clades. Panaroo also includes an implementation of the Spydrpick algorithm which allows for the identification of gene presence/absence patterns that are either highly correlated or anti-correlated whilst accounting for population structure (49). Such correlations can indicate that the genes involved have epistatic effects on fitness or that their presence or absence is a result of similar selective pressures. Finally, the output of Panaroo seamlessly interfaces with pyseer (v1.3.0), a bacterial GWAS package (24, 50). pyseer includes a wide range of methods for performing association studies allowing for phenotypic associations to be found with gene or structural presence/absence patterns.

Simulation and Comparison with Previous Methods. Using the *E. coli* reference genome ASM584v2 as a starting point, we simulated variation in the accessory genome by varying the rates of gene gain and loss using a phylogeny simulated with the Kingman coalescent in dendropy (v4.4.0) (51). In addition, we simulated various degrees of sequence variation by varying the within gene codon substitution rate. Three replicate datasets of 100 sampled genomes were created for each set of model parameters outlined in Supplementary Table 1. Realistic sequence assemblies were generated by first simulating NGS sequencing reads using either Mason (v2.0.9) or ART (v2.5.8) (30, 32). These were assembled using SPAdes (v3.13.0) (31). The resulting assemblies were annotated using Prokka (v1.13.3) with a custom BLAST database containing the correctly assigned proteins from the simulation prior to assembly. This extensive simula-

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tion pipeline provided more realistic data and included many of the sources of error encountered in pangenome analyses. To simulate the problems that fragmentation can bring to the analysis of pangenomes we also simulated a fragmented assembly by breaking the simulated whole genomes into fragments prior to simulating the NGS reads. This resulted in highly fragmented final assemblies. Contamination was also simulated by randomly adding 10kb segments of the *S. epidermidis* reference genome ASM764v1 a common lab contaminant to the simulated genomes prior to NGS simulation. These segments were added by sampling from a Poisson distribution with mean 1. The gene presence/absence matrix was then generated for PanX (v1.5.1), Roary (v1.007002), PIRATE (v1.0), COGsoft (v201204) and Panaroo (v1.0.0). These were compared with the simulated matrix and the number of inferred orthologous clusters that contained an error was counted and is shown in Figure 3.

SOFTWARE AVAILABILITY

Source code available from:
https://github.com/gtonkinhill/panaroo
Code for reproducing figures from:
https://github.com/gtonkinhill/panaroo_manuscript
Archived data for replication at time of publication to bioRxiv:
https://doi.org/10.5281/zenodo.3599800

ACKNOWLEDGEMENTS

Many thanks to Lauren Bell for designing the Panaroo logo and to the members of teams 81 and 284 at the Wellcome Sanger Institute for helpful comments and testing.

FUNDING

Wellcome Trust [206194 to S.D.B]; Wellcome Trust PhD Scholarship Grant [204016 to G.T.H]; ERC [742158 to J.C.]. J.A.L. is funded by MR/R015600/1. This award is jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2 programme supported by the European Union.

Bibliography

1. D S Guttman and D E Dykhuizen. Clonal divergence in escherichia coli as a result of recombination, not mutation. *Science*, 266(5189):1380–1383, November 1994. ISSN 0036-8075. doi: 10.1126/science.7973728.
2. Kathryn E Holt, Heiman Wertheim, Ruth N Zadoks, Stephen Baker, Chris A Whitehouse, David Dance, Adam Jenney, Thomas R Connor, Li Yang Hsu, Juliëtte Severin, Sylvain Brisse, Hanwei Cao, Jonathan Wilksch, Claire Gorrie, Mark B Schultz, David J Edwards, Kinh Van Nguyen, Trung Vu Nguyen, Trinh Tuyet Dao, Martijn Mensink, Vien Le Minh, Nguyen Thi Khanh Nhu, Constance Schultsz, Kuntaman Kuntaman, Paul N Newton, Catrin E Moore, Richard A Strugnell, and Nicholas R Thomson. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci. U. S. A.*, 112(27):E3574–81, July 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1501049112.
3. Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome. *Curr. Opin. Genet. Dev.*, 15(6):589–594, December 2005. ISSN 0959-437X. doi: 10.1016/j.gde.2005.09.006.
4. Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew T G Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, November 2015. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btv421.
5. Li Li, Christian J Stoeckert, Jr, and David S Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13(9):2178–2189, September 2003. ISSN 1088-9051. doi: 10.1101/gr.1224503.
6. Derrick E Fouts, Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res.*, 40(22):e172–e172, December 2012. ISSN 0305-1048. doi: 10.1093/nar/gks757.
7. Sion C Bayliss, Harry A Thorpe, Nicola M Coyle, Samuel K Sheppard, and Edward J Feil. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. April 2019.
8. Wei Ding, Franz Baumdicker, and Richard A Neher. panx: pan-genome analysis and exploration. *Nucleic Acids Res.*, 46(1):e5, January 2018. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkx977.
9. Yongbing Zhao, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. PGAP: pan-genomes analysis pipeline. *Bioinformatics*, 28(3):416–418, February 2012. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btr655.
10. David M Kristensen, Lavanya Kannan, Michael K Coleman, Yuri I Wolf, Alexander Sorokin, Eugene V Koonin, and Arcady Mushegian. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, 26 (12):1481–1487, June 2010. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btq229.
11. Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, July 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158.
12. Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, September 1997. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/25.17.3389.
13. Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12(1):59–60, January 2015. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3176.
14. A J Enright, S Van Dongen, and C A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584, April 2002. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/30.7.1575.
15. R L Tatusov, E V Koonin, and D J Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, October 1997. ISSN 0036-8075. doi: 10.1126/science.278.5338.631.
16. Steven L Salzberg. Next-generation genome annotation: we still struggle to get it right. *Genome Biol.*, 20(1):92, May 2019. ISSN 1465-6906, 1474-760X. doi: 10.1186/s13059-019-1715-2.
17. James F Denton, Jose Lugo-Martinez, Abraham E Tucker, Daniel R Schrider, Wesley C Warren, and Matthew W Hahn. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput. Biol.*, 10(12):e1003998, December 2014. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1003998.
18. Jukka Corander, Christophe Fraser, Michael U Gutmann, Brian Arnold, William P Hanage, Stephen D Bentley, Marc Lipsitch, and Nicholas J Croucher. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nature Ecology & Evolution*, page 1, October 2017. ISSN 2397-334X, 2397-334X. doi: 10.1038/s41559-017-0337-x.
19. Alan McNally, Teemu Kallonen, Christopher Connor, Khalil Abudahab, David M Aanensen, Carolyne Horner, Sharon J Peacock, Julian Parkhill, Nicholas J Croucher, and Jukka Corander. Diversification of colonization factors in a Multidrug-Resistant escherichia coli lineage evolving under negative Frequency-Dependent selection. *MBio*, 10(2), April 2019. ISSN 2150-7511. doi: 10.1128/mBio.00644-19.
20. Franz Baumdicker, Wolfgang R Hess, and Peter Pfaffelhuber. The infinitely many genes model for the distributed genome of bacteria. *Genome Biol. Evol.*, 4(4):443–456, February 2012. ISSN 1759-6653. doi: 10.1093/gbe/evs016.
21. Rebecca A Gladstone, Stephanie W Lo, John A Lees, Nicholas J Croucher, Andries J van Tonder, Jukka Corander, Andrew J Page, Pekka Marttinen, Leon J Bentley, Theresa J Ochoa, Pak Leung Ho, Mignon du Plessis, Jennifer E Cornick, Brenda Kwambana-Adams, Rachel Benisty, Susan A Nzenze, Shabir A Madhi, Paulina A Hawkins, Dean B Everett, Martin Antonio, Ron Dagan, Keith P Klugman, Anne von Gottberg, Lesley McGee, Robert F Breiman, and Stephen D Bentley. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine*, 43:338–346, May 2019. ISSN 2352-3964. doi: 10.1016/j.ebiom.2019.04.021.
22. Martin Šošić and Mile Šikić. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, 2017. ISSN 1367-4803.
23. Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, November 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303.
24. John A Lees, Marco Galardini, Stephen D Bentley, Jeffrey N Weiser, Jukka Corander, and Oliver Stegle. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, July 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty539.
25. Nicola Casali, Agnieszka Broda, Simon R Harris, Julian Parkhill, Timothy Brown, and Francis Drobniewski. Whole genome sequence analysis of a large Isoniazid-Resistant tuberculosis outbreak in london: A retrospective observational study. *PLoS Med.*, 13(10):e1002137, October 2016. ISSN 1549-1277, 1549-1676. doi: 10.1371/journal.pmed.1002137.
26. Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14): 2068–2069, July 2014. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btu153.
27. Maxime Godfroid, Tal Dagan, and Anne Kupczok. Recombination signal in mycobacterium tuberculosis stems from reference-guided assemblies and alignment artefacts. *Genome Biol. Evol.*, 10(8):1920–1926, August 2018. ISSN 1759-6653. doi: 10.1093/gbe/evy143.
28. S T Cole, R Brosch, J Parkhill, T Garnier, C Churcher, D Harris, S V Gordon, K Eiglmeier, S Gas, C E Barry, 3rd, F Tekaia, K Badcock, D Basham, D Brown, T Chillingworth, R Connor, R Davies, K Devlin, T Feltwell, S Gentles, N Hamlin, S Holroyd, T Hornsby, K Jagels, A Krogh, J McLean, S Moule, L Murphy, K Oliver, J Osborne, M A Quail, M A Rajandream, J Rogers, S Rutter, K Seeger, J Skelton, R Squares, S Squares, J E Sulston, K Taylor, S Whitehead, and B G Barrell. Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544, June 1998. ISSN 0028-0836. doi: 10.1038/31159.
29. R Eric Collins and Paul G Higgs. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol. Biol. Evol.*, 29(11):3413–3425, November 2012. ISSN 0737-4038. doi: 10.1093/molbev/mss163.
30. Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, February 2012. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btr708.
31. Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, Alexey V Pyshkin, Alexander V Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A Alekseyev, and Pavel A Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, 19(5):455–477, May 2012. ISSN 1066-5277, 1557-8666. doi: 10.1089/cmb.2012.0021.

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

32. Manuel Holtgrewe. Mason: a read simulator for second generation sequencing data. Technical Report TR-B-10-06, Institut für Mathematik und Informatik, Freie Universität Berlin, 2010.
33. Ola Brynildsrud, Jon Bohlin, Lonneke Scheffer, and Vegard Eldholm. Rapid scoring of genes in microbial pan-genome-wide association studies with scoary. *Genome Biol.*, 17(1): 238, November 2016. ISSN 1465-6906. doi: 10.1186/s13059-016-1108-8.
34. Simon R Harris, Michelle J Cole, Gianfranco Spiteri, Leonor Sánchez-Busó, Daniel Golparian, Susanne Jacobsson, Richard Goater, Khalil Abudahab, Corin A Yeats, Beatrice Bercot, et al. Public health surveillance of multidrug-resistant clones of neisseria gonorrhoeae in europe: a genomic survey. *The Lancet Infectious diseases*, 18(7):758–768, 2018.
35. Daniel N Wilson. The abc of ribosome-related antibiotic resistance. *MBio*, 7(3):e00598–16, 2016.
36. Zijian Gong, Wei Lai, Min Liu, Zhengshuang Hua, Yayin Sun, Qingfang Xu, Yue Xia, Yue Zhao, and Xiaoyuan Xie. Novel genes related to ceftriaxone resistance found among ceftriaxone-resistant neisseria gonorrhoeae strains selected in vitro. *Antimicrobial agents and chemotherapy*, 60(4):2043–2051, 2016.
37. Franz Baumdicker and Peter Pfaffelhuber. The infinitely many genes model with horizontal gene transfer. *Electron. J. Probab.*, 19, 2014. ISSN 1083-6489. doi: 10.1214/EJP.v19-2642.
38. Nicholas J Croucher, Paul G Coupland, Abbie E Stevenson, Alanna Callendrello, Stephen D Bentley, and William P Hanage. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat. Commun.*, 5:5471, November 2014. ISSN 2041-1723. doi: 10.1038/ncomms6471.
39. Nancy A Moran. Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, 108 (5):583–586, March 2002. ISSN 0092-8674. doi: 10.1016/s0092-8674(02)00665-7.
40. Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, March 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119.
41. A L Delcher, D Harmon, S Kasif, O White, and S L Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, 27(23):4636–4641, December 1999. ISSN 0305-1048. doi: 10.1093/nar/27.23.4636.
42. Arthur L Delcher, Kirsten A Bratke, Edwin C Powers, and Steven L Salzberg. Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics*, 23(6):673–679, March 2007. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btm009.
43. Kazutaka Katoh, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.*, 30(14):3059–3066, July 2002. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkf436.
44. Ari Löytynoja. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.*, 1079:155–170, 2014. ISSN 1064-3745, 1940-6029. doi: 10.1007/978-1-62703-646-7_10.
45. Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, and Others. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, 7(1), 2011.
46. Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 17(1):132, June 2016. ISSN 1465-6906. doi: 10.1186/s13059-016-0997-x.
47. Brian D Ondov, Gabriel J Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B Buck, and Adam M Phillippy. Mash screen: High-throughput sequence containment estimation for genome discovery.
48. Seyed Alireza Zamani-Dahaj, Mohamed Okasha, Jakub Kosakowski, and Paul G Higgs. Estimating the frequency of horizontal gene transfer using phylogenetic models of gene gain and loss. *Mol. Biol. Evol.*, 33(7):1843–1857, July 2016. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msw062.
49. Johan Pensar, Santeri Puranen, Brian Arnold, Neil MacAlasdair, Juri Kuronen, Gerry Tonkin-Hill, Maiju Pesonen, Yingying Xu, Aleksi Sipola, Leonor Sánchez-Busó, John A Lees, Claire Chewapreecha, Stephen D Bentley, Simon R Harris, Julian Parkhill, Nicholas J Croucher, and Jukka Corander. Genome-wide epistasis and co-selection study using mutual information. *Nucleic Acids Res.*, July 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkz656.
50. John A Lees, Minna Vehkala, Niko Välimäki, Simon R Harris, Claire Chewapreecha, Nicholas J Croucher, Pekka Marttinen, Mark R Davies, Andrew C Steer, Steven Y C Tong, Antti Honkela, Julian Parkhill, Stephen D Bentley, and Jukka Corander. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.*, 7:12797, September 2016. ISSN 2041-1723. doi: 10.1038/ncomms12797.
51. Jeet Sukumaran and Mark T Holder. DendroPy: a python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, June 2010. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btq228.

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figures:

Gain/Loss Ratio	Mutation Rate	Source of Error
0.1	1.00E-14	NA
1	1.00E-14	NA
10	1.00E-14	NA
10	1.00E-15	NA
10	1.00E-16	NA
10	1.00E-14	Fragmentation
10	1.00E-14	Contamination with *S. epidermidis*

Supplementary Table 1: Parameters used to simulate pangenomes using the Infinitely Many Genes model.

Gene	lrt-pvalue	beta	beta-std-err	Antibiotic	Annotation
gatC	9.38E-06	0.402	0.0903	PEN	aspartyl/glutamyl-tRNA amidotransferase subunit C
group_111	4.82E-07	0.319	0.063	PEN	ATP synthase F0F1 subunit delta
group_1126	3.60E-08	0.18	0.0324	PEN	putative thiosulfate sulfurtransferase
group_1140	9.18E-06	-0.108	0.0242	AZM	ArsR family transcriptional regulator
group_1168	2.95E-10	0.317	0.0498	PEN	integral membrane protein
group_1237	1.92E-20	0.165	0.0174	TET	putative phage associated protein;phage associated protein
group_1333	4.44E-07	0.263	0.0517	CFM	two-component system transcriptional response regulator
group_1333	5.93E-18	0.499	0.0568	PEN	two-component system transcriptional response regulator
group_1333	4.44E-07	0.263	0.0517	CRO	two-component system transcriptional response regulator
group_134	4.48E-06	0.211	0.0458	CRO	TonB-dependent receptor protein
group_134	2.30E-19	0.439	0.0478	PEN	TonB-dependent receptor protein
group_134	4.48E-06	0.211	0.0458	CFM	TonB-dependent receptor protein
group_1387	5.01E-09	0.12	0.0203	TET	putative maltose phosphorylase
group_144	9.10E-06	0.149	0.0334	PEN	elongation factor G
group_1491	3.64E-08	0.261	0.047	PEN	phage protein
group_1496	2.09E-08	0.332	0.0587	PEN	membrane protein
group_1511	2.17E-07	0.377	0.0722	PEN	phage protein
group_1611	9.18E-06	-0.108	0.0242	AZM	phage associated protein;hypothetical protein
group_1623	2.65E-07	0.256	0.0494	CRO	phage associated protein
group_1623	1.34E-17	0.486	0.0559	PEN	phage associated protein
group_1623	2.65E-07	0.256	0.0494	CFM	phage associated protein
group_1693	8.09E-15	0.122	0.0155	TET	putative cytochrome C
group_1708	2.65E-07	0.256	0.0494	CFM	VapD-like protein
group_1708	1.34E-17	0.486	0.0559	PEN	VapD-like protein
group_1708	2.65E-07	0.256	0.0494	CRO	VapD-like protein
group_172	7.24E-06	0.202	0.0447	PEN	IS1016 transposase
group_1753	2.42E-08	0.419	0.0745	CRO	YegA
group_1753	4.11E-09	0.472	0.0796	PEN	YegA
group_1753	2.42E-08	0.419	0.0745	CFM	YegA
group_238	1.78E-08	-0.0668	0.0118	TET	amidophosphoribosyltransferase
group_299	4.82E-07	0.319	0.063	PEN	NADH:ubiquinone dehydrogenase L subunit
group_380	9.10E-06	0.149	0.0334	PEN	arsenate reductase
group_438	9.18E-06	-0.108	0.0242	AZM	phosphoribosylaminoimidazole carboxylase ATPase subunit
group_451	2.42E-08	0.419	0.0745	CRO	Protein rnfH
group_451	4.11E-09	0.472	0.0796	PEN	Protein rnfH
group_451	2.42E-08	0.419	0.0745	CFM	Protein rnfH
group_464	2.22E-08	0.178	0.0316	PEN	ABC transporter ATP-binding protein
group_914	9.62E-07	0.0566	0.0115	CIP	phage repressor phage associated protein
group_945	4.82E-07	0.319	0.063	PEN	ABC transporter permease amino acid
porB	1.63E-06	-0.152	0.0316	PEN	major outer membrane protein porin P.IB; P.I
rplW	1.53E-08	-0.327	0.0574	AZM	50S ribosomal protein L23

Supplementary Table 2: Significant pan-GWAS results for antibiotic resistance in the EuroGASP collection

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

variant	af	filter-pvalue	lrt-pvalue	beta	beta-std-err	variant_h2	antibiotic
metG-group_1858-pilE_1_pilE_pilE_3_pilS5_pilS2_1_pilE_2	0.0398	1.79E-09	4.85E-45	-0.36	0.0244	0.415	CIP
group_946-group_1499-group_1846	0.0133	0.0379	5.08E-45	0.317	0.0215	0.415	TET
hemC-group_1846-group_1499	0.0133	0.0379	5.08E-45	0.317	0.0215	0.415	TET
group_1056-group_1846-pepN	0.0218	7.10E-35	3.75E-43	-0.377	0.0261	0.406	TET
group_885-group_1826-group_1827	0.0275	1.13E-45	7.11E-43	-0.273	0.019	0.405	TET
group_143-group_933-group_1118	0.0133	2.26E-10	2.84E-39	0.31	0.0226	0.388	TET
group_212-group_592-group_1675	0.0199	2.62E-31	1.90E-31	-0.256	0.0212	0.348	TET
group_644-group_801-group_505	0.0256	4.65E-42	5.58E-29	-0.257	0.0223	0.335	TET
group_1236-group_1556-group_1237	0.0598	8.52E-07	1.83E-28	0.23	0.0202	0.332	TET
group_505-group_1775-group_1232	0.0247	2.95E-40	4.42E-27	-0.251	0.0227	0.323	TET
group_1775-group_505-group_801	0.0237	1.85E-38	1.11E-23	-0.231	0.0225	0.302	TET
group_238-group_1556-group_1236	0.0949	3.87E-42	1.83E-22	-0.189	0.019	0.294	TET
group_936-group_885-group_1826	0.102	1.26E-19	3.17E-22	-0.121	0.0122	0.292	TET
group_1367-group_1237-group_1556	0.0294	0.76	1.11E-21	0.206	0.0211	0.289	TET
group_1824-group_164-group_1899	0.0123	0.0199	1.53E-20	0.176	0.0185	0.281	CIP
group_1229-group_1118-group_933	0.0218	6.11E-25	3.74E-20	0.183	0.0195	0.278	TET
group_936-group_885-group_1827	0.0417	0.971	6.45E-19	0.123	0.0136	0.269	TET
group_1335-group_1794-group_1099	0.018	0.00505	2.90E-18	0.358	0.0403	0.264	CRO
group_1720-group_75-group_1901	0.0503	7.80E-79	1.96E-17	-0.258	0.0299	0.258	SMX
group_1367-group_238-group_1556	0.102	4.08E-50	1.51E-16	-0.142	0.0169	0.251	TET
group_1393-group_1795-group_1690	0.0104	0.00778	4.41E-16	0.204	0.0247	0.247	TET
group_1032-group_1894-group_1051	0.0123	0.226	7.84E-16	0.145	0.0177	0.245	CIP
group_1630-ppsR-group_1489	0.0237	0.203	1.78E-15	-0.147	0.0182	0.242	CIP
ppsA-group_1489-ppsR	0.0237	0.203	1.78E-15	-0.147	0.0182	0.242	CIP
group_1332-group_1826-group_117	0.0114	0.0123	7.75E-15	-0.163	0.0206	0.236	CIP
rho-ppsA-group_1489	0.0247	0.145	9.58E-15	-0.14	0.0178	0.235	CIP
group_136-group_1367-group_1237	0.037	0.398	1.12E-14	0.136	0.0174	0.235	TET
group_68-group_64-group_37	0.0342	0.000752	1.25E-14	-0.126	0.0161	0.234	CIP
group_64-group_37-group_915	0.0342	0.000752	1.25E-14	-0.126	0.0161	0.234	CIP
group_1333-group_117-group_1826	0.0123	0.00691	1.92E-14	-0.157	0.0202	0.233	CIP
group_190-group_577-group_150	0.0294	0.377	2.73E-14	0.111	0.0144	0.231	TET
hemC-group_1846-pepN	0.0266	3.68E-17	6.23E-14	-0.141	0.0186	0.228	TET
rpsS-rplB-group_1157	0.829	7.48E-31	6.85E-14	0.575	0.0757	0.228	AZM
rpsS-rplB-rplW	0.171	7.48E-31	6.85E-14	-0.575	0.0757	0.228	AZM
group_676-mafB_mafB_2-group_1127	0.0266	0.000891	1.19E-13	0.675	0.0898	0.226	CRO
group_1895-group_493-group_1591	0.858	2.81E-35	1.91E-13	0.56	0.0752	0.224	AZM
group_702-group_1633-group_1869	0.0171	0.458	1.93E-13	0.142	0.019	0.224	TET
group_1051-group_460-group_1482	0.0104	0.0217	2.24E-13	0.128	0.0172	0.223	CIP
group_446-pepN-group_1846	0.0531	1.40E-10	1.20E-12	-0.107	0.0149	0.217	TET
group_972-group_1612-group_1829	0.0323	8.11E-07	1.36E-12	0.104	0.0145	0.216	CIP
group_563-group_303-group_1171	0.0218	9.75E-17	1.79E-12	-0.138	0.0193	0.215	TET
group_2008-group_1991-group_1972	0.135	1.55E-05	2.46E-12	-0.105	0.0148	0.214	TET
group_393-group_262-group_1171	0.0199	3.84E-16	4.36E-12	-0.141	0.0201	0.211	TET
group_303-group_1171-group_262	0.019	1.31E-14	6.45E-12	-0.142	0.0204	0.209	TET
group_1121-group_1531-group_1176	0.0228	0.458	8.01E-12	0.202	0.0292	0.209	AZM
group_255-group_841-group_1576	0.488	4.57E-82	8.19E-12	0.312	0.0451	0.208	PEN
metG-group_1576-group_841	0.488	4.57E-82	8.19E-12	0.312	0.0451	0.208	PEN
group_758-group_841-group_255	0.512	4.57E-82	8.19E-12	-0.312	0.0451	0.208	PEN
group_758-group_1576-metG	0.512	4.57E-82	8.19E-12	-0.312	0.0451	0.208	PEN
group_2024-group_1993-group_1985	0.14	1.43E-05	1.03E-11	-0.0985	0.0143	0.207	TET
rho-ppsA-ppsR	0.0285	0.00396	3.69E-11	0.0904	0.0135	0.202	CIP
group_1630-ppsR-ppsA	0.0285	0.00396	3.69E-11	0.0904	0.0135	0.202	CIP
dut-group_385-group_1465	0.0104	0.0225	4.32E-11	0.603	0.0906	0.201	CRO
group_1066-group_1531-group_1176	0.544	1.83E-09	4.49E-11	-0.173	0.026	0.201	AZM
group_1711-group_1694-group_1894	0.0266	1.94E-27	8.25E-11	-0.135	0.0205	0.198	TET
pyrG-group_1870-group_714	0.0304	0.74	1.07E-10	0.235	0.036	0.197	PEN
group_95-group_363-group_1224	0.0455	3.36E-10	1.12E-10	-0.311	0.0478	0.197	CIP
group_363-group_95-group_1299	0.0455	3.36E-10	1.12E-10	-0.311	0.0478	0.197	CIP
group_177-group_791-group_892	0.0152	7.03E-07	3.50E-10	-0.13	0.0205	0.192	TET
group_1066-group_1121-group_1558	0.97	0.327	4.70E-10	-0.162	0.0257	0.19	AZM
group_182-group_714-group_1870	0.0114	1.79E-12	8.68E-10	-0.154	0.0249	0.187	TET
group_1134-alaS-group_67	0.0304	0.00204	1.14E-09	0.0918	0.0149	0.186	TET
group_283-group_67-alaS	0.0304	0.00204	1.14E-09	0.0918	0.0149	0.186	TET
group_185-group_274-group_908	0.737	1.49E-08	1.42E-09	-0.0695	0.0114	0.185	CIP
group_2007-group_1985-group_1993	0.214	0.175	1.58E-09	-0.0818	0.0134	0.185	TET
group_1531-group_1121-group_1558	0.0275	0.637	1.60E-09	0.157	0.0258	0.184	AZM
group_1401-group_1332-group_1826	0.0161	0.343	1.92E-09	-0.0982	0.0162	0.184	CIP
group_1824-group_164-opaD_piiC	0.0332	0.0109	2.68E-09	-0.0883	0.0147	0.182	CIP
group_292-group_412-group_321	0.0342	2.07E-07	4.03E-09	0.143	0.024	0.18	AZM
group_1126-mafB_mafB_2-group_1136	0.315	1.17E-10	4.05E-09	0.205	0.0345	0.18	PEN
group_2026-group_2011-group_2005	0.225	0.201	7.69E-09	-0.0756	0.013	0.177	TET
group_1066-group_1531-group_1190	0.426	1.44E-11	7.80E-09	-0.183	0.0314	0.177	CRO
group_683-group_385-group_1738	0.0503	0.0135	8.26E-09	0.473	0.0815	0.176	CRO
group_1516-group_1056-group_1846	0.0152	1.77E-22	9.22E-09	-0.149	0.0257	0.176	TET
group_702-group_1633-group_1902	0.0209	0.00462	9.51E-09	0.0917	0.0158	0.176	CIP
ispH-lspA-group_1875	0.0588	0.000396	1.14E-08	0.106	0.0184	0.175	AZM
rplB-group_1157-rplD	0.824	1.14E-30	1.53E-08	0.327	0.0574	0.173	AZM
rplC-rplD-group_1157	0.824	1.14E-30	1.53E-08	0.327	0.0574	0.173	AZM
rplC-rplD-rplW	0.176	1.14E-30	1.53E-08	-0.327	0.0574	0.173	AZM
group_976-group_494-group_1552	0.278	7.66E-09	1.65E-08	-0.109	0.0192	0.173	AZM
group_494-group_1552-tbpA	0.278	7.66E-09	1.65E-08	-0.109	0.0192	0.173	AZM
pilS1_pilE_pilE_1_pilE1_pilS6_2_pilS5_pilS2_pilS6_pilS6_1_pilE_2_pilE_3-pilE_1_pilS6_pilE_pilE1-pilE_pilE_1_pilS6_pilE1_pilE_2	0.0408	0.432	2.87E-08	0.178	0.0317	0.17	CRO
pilS5_pilE_pilS2-pilE_1_pilE1_pilS6-pilS6_pilE_pilS2	0.0104	7.70E-11	3.06E-08	0.161	0.0288	0.17	TET
group_205-pyrG-group_1870	0.0361	0.734	3.16E-08	0.191	0.0343	0.169	PEN
group_903-group_1759-group_1716	0.0142	0.0115	3.39E-08	-0.177	0.0319	0.169	CRO
group_1404-group_321-group_412	0.038	1.37E-07	3.51E-08	0.126	0.0226	0.169	AZM
group_1751-group_1224-group_363	0.0465	1.84E-10	4.12E-08	-0.226	0.041	0.168	CIP
group_1951-group_46-group_886	0.0123	0.597	4.21E-08	-0.351	0.0636	0.168	PEN
group_451-group_714-group_1870	0.0218	0.136	4.52E-08	0.246	0.0446	0.167	PEN
group_1761-group_1801-group_1749	0.0664	6.69E-05	5.18E-08	-0.0652	0.0119	0.167	TET
group_1103-group_953-group_899	0.0152	0.00122	6.37E-08	0.16	0.0294	0.166	CIP
group_185-group_274-group_167	0.0408	9.82E-10	7.58E-08	0.0633	0.0117	0.165	CIP
group_536-group_940-group_252	0.0199	1.41E-21	8.45E-08	0.162	0.03	0.164	TET
group_940-group_252-group_757	0.0199	1.41E-21	8.45E-08	0.162	0.03	0.164	TET
group_1514-group_679-porB	0.229	5.98E-14	8.51E-08	-0.16	0.0296	0.164	PEN
lpxC-pilS5_pilE_2_pilS5_2_pilE_pilE1_pilE_1_pilE_4-group_1778	0.0133	0.00407	8.87E-08	0.133	0.0247	0.164	CRO
group_136-group_1367-group_238	0.166	5.32E-37	1.11E-07	-0.0617	0.0115	0.163	TET
ileS-group_1875-lspA	0.0408	0.0324	1.13E-07	0.116	0.0217	0.162	AZM
group_1126-mafB_mafB_2-group_1100	0.766	4.67E-24	1.21E-07	0.171	0.0321	0.162	PEN
group_976-group_494-group_1864	0.0787	5.06E-15	1.25E-07	0.101	0.019	0.162	AZM
group_1136-mafB_mafB_2-group_1127	0.0607	0.191	1.30E-07	0.247	0.0465	0.162	PEN
group_1265-group_1890-group_1901	0.0209	3.42E-06	1.39E-07	-0.367	0.0691	0.161	PEN
pip-opaD_piiC-group_164	0.0275	0.11	1.54E-07	-0.0839	0.0159	0.161	CIP
group_992-group_128-group_1602	0.018	3.33E-21	2.32E-07	0.158	0.0304	0.158	TET
group_1562-group_1190-group_1176	0.319	1.63E-09	2.53E-07	0.109	0.021	0.158	SMX
lpxC-pilS5_pilE_2_pilS5_2_pilE_pilE1_pilE_1_pilE_4-pilS5_pilE_pilS2	0.0142	0.0115	2.57E-07	0.123	0.0238	0.158	CRO
group_861-group_485-group_935	0.0266	3.94E-05	2.91E-07	0.0586	0.0113	0.157	CIP
group_1541-group_68-group_914	0.108	2.47E-14	3.67E-07	-0.0443	0.00866	0.156	TET
group_68-group_914-group_37	0.108	2.47E-14	3.67E-07	-0.0443	0.00866	0.156	TET
group_1374-pilE1_pilE_1_pilE__pilE_2_pilS3_pilS6-pilS5_pilE_2_pilS5_2_pilE_pilE1_pilE_1_pilE_4	0.0427	0.000119	4.00E-07	0.105	0.0206	0.155	TET
group_914-group_37-group_915	0.694	5.88E-10	4.11E-07	0.0588	0.0115	0.155	CIP
group_495-ileS-group_1875	0.0636	0.000141	4.79E-07	0.104	0.0205	0.154	AZM
group_1613-group_504-group_412	0.0114	0.0924	5.28E-07	-0.259	0.0513	0.154	AZM

DRAFT

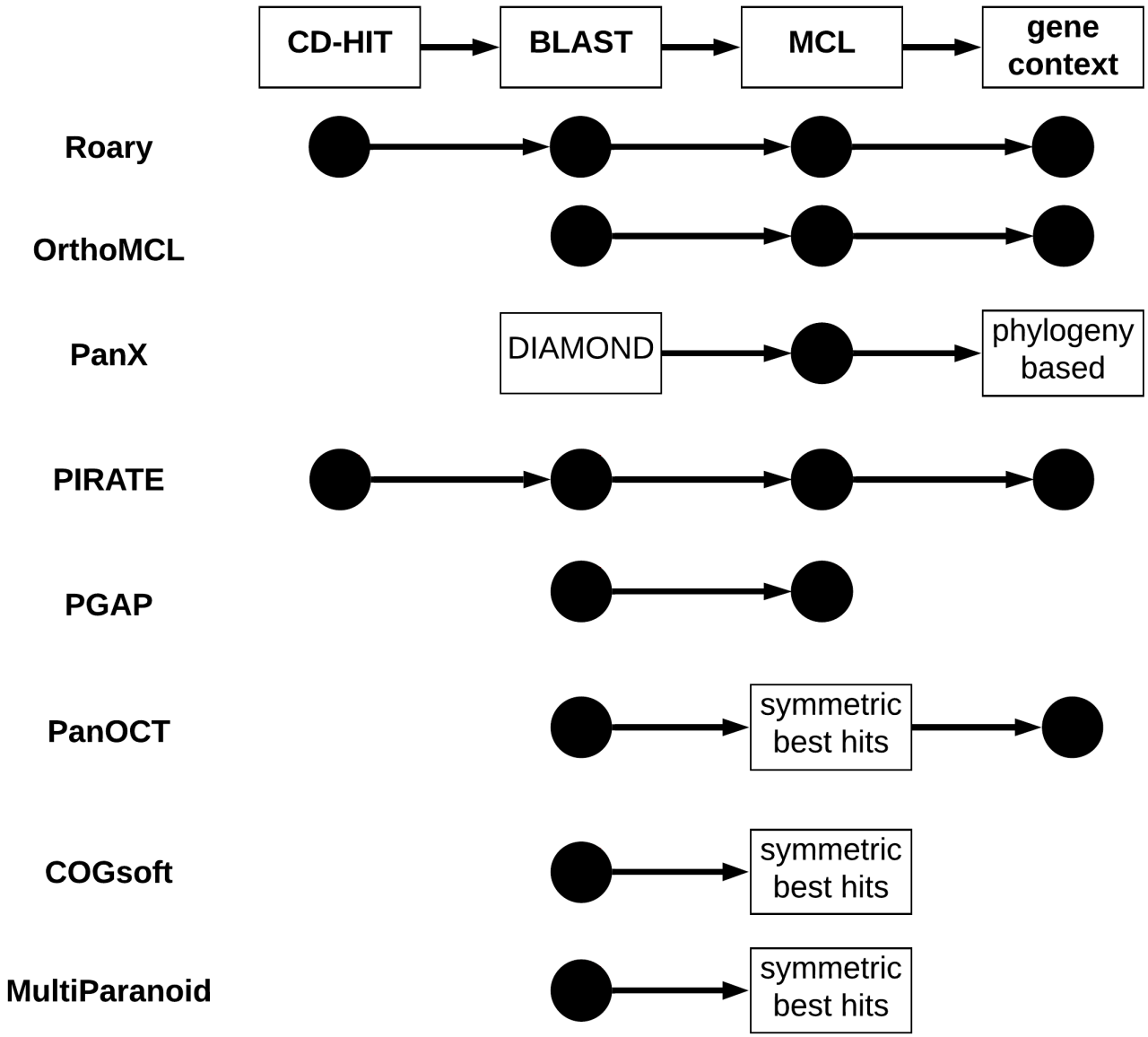
bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

variant	af	filter-pvalue	lrt-pvalue	beta	beta-std-err	variant_h2	antibiotic
group_303-group_1450-group_262	0.0142	1.14E-05	6.64E-07	-0.113	0.0226	0.152	TET
group_1131-group_795-group_1138	0.743	0.123	6.87E-07	0.322	0.0645	0.152	CRO
group_771-group_1002-group_1138	0.743	0.123	6.87E-07	0.322	0.0645	0.152	CRO
group_1002-group_795-group_1131	0.257	0.123	6.87E-07	-0.322	0.0645	0.152	CRO
group_771-group_1002-group_795	0.257	0.123	6.87E-07	-0.322	0.0645	0.152	CRO
group_1986-group_1998-group_2014	0.22	0.155	7.61E-07	-0.0673	0.0135	0.152	TET
group_1570-group_1673-group_668	0.0133	0.0188	9.18E-07	0.0905	0.0183	0.151	CIP
group_1821-group_1622-group_1145	0.018	2.29E-24	9.75E-07	0.118	0.024	0.15	TET
pilS1-pilS1_pilE_pilE_1_pilE1_pilS6_2_pilS5_pilS2_pilS6_pilS6_1_pilE_2_pilE_3-pilE_1_pilS6_pilE_pilE1	0.0104	0.324	9.80E-07	0.331	0.0673	0.15	PEN
group_1329-group_1882-group_1779	0.0342	6.56E-27	1.02E-06	0.0822	0.0167	0.15	TET
group_945-group_1522-group_816	0.952	2.05E-07	1.07E-06	0.301	0.0613	0.15	PEN
coaD-group_1369-group_1909	0.0493	3.17E-08	1.14E-06	0.102	0.0208	0.149	AZM
group_1514-group_679-group_464	0.763	3.19E-12	1.15E-06	0.157	0.032	0.149	PEN
group_447-group_1761-group_1801	0.0674	4.34E-08	1.20E-06	0.0616	0.0126	0.149	TET
group_484-group_956-group_829	0.0503	3.22E-10	1.41E-06	-0.0827	0.017	0.148	TET
rpoD-group_528-group_1522	0.0446	8.89E-10	1.54E-06	-0.402	0.0831	0.147	PEN
group_816-group_1522-group_528	0.0446	8.89E-10	1.54E-06	-0.402	0.0831	0.147	PEN
tbpA-group_1552-group_1864	0.0417	1.51E-08	1.62E-06	0.111	0.0231	0.147	AZM
group_494-group_1864-group_1552	0.0417	1.51E-08	1.62E-06	0.111	0.0231	0.147	AZM
group_947-group_1749-group_1801	0.0835	5.84E-07	1.70E-06	-0.059	0.0123	0.147	TET
group_1596-group_1720-group_75	0.0427	3.75E-55	1.89E-06	-0.11	0.0229	0.146	SMX
group_393-group_262-group_1450	0.0152	3.16E-05	1.94E-06	-0.104	0.0217	0.146	TET
group_171-group_1672-group_1229	0.107	3.42E-27	2.43E-06	0.228	0.048	0.145	CRO
group_1126-mafB_mafB_2-group_1187	0.756	4.03E-17	2.46E-06	0.131	0.0277	0.145	PEN

Supplementary Table 3: Significant sv-pan-GWAS results for antibiotic resistance in the EuroGASP collection

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

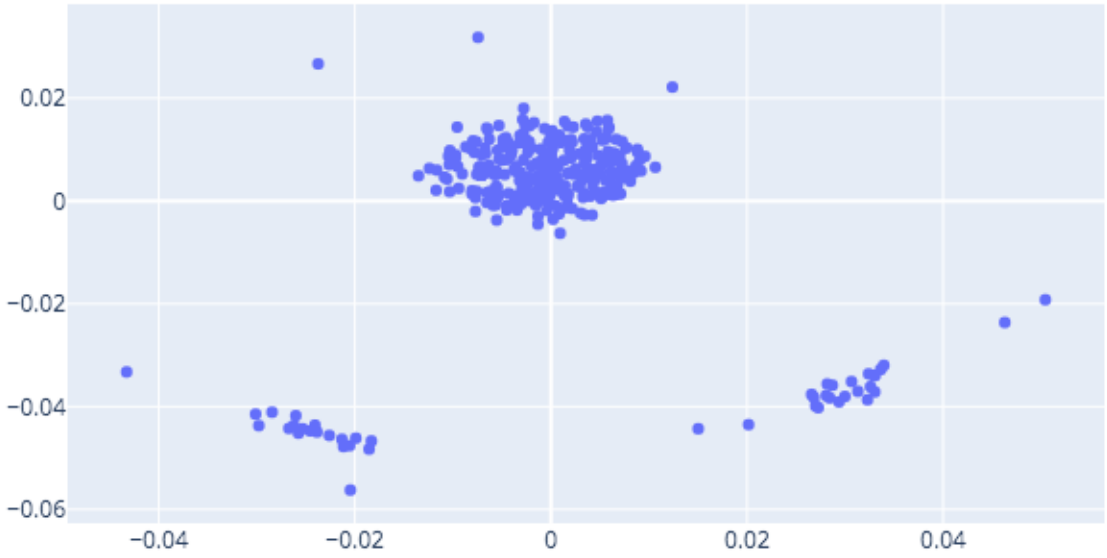


Supplementary Figure 1: A comparison of the pipelines used by different pangenome analysis tools.

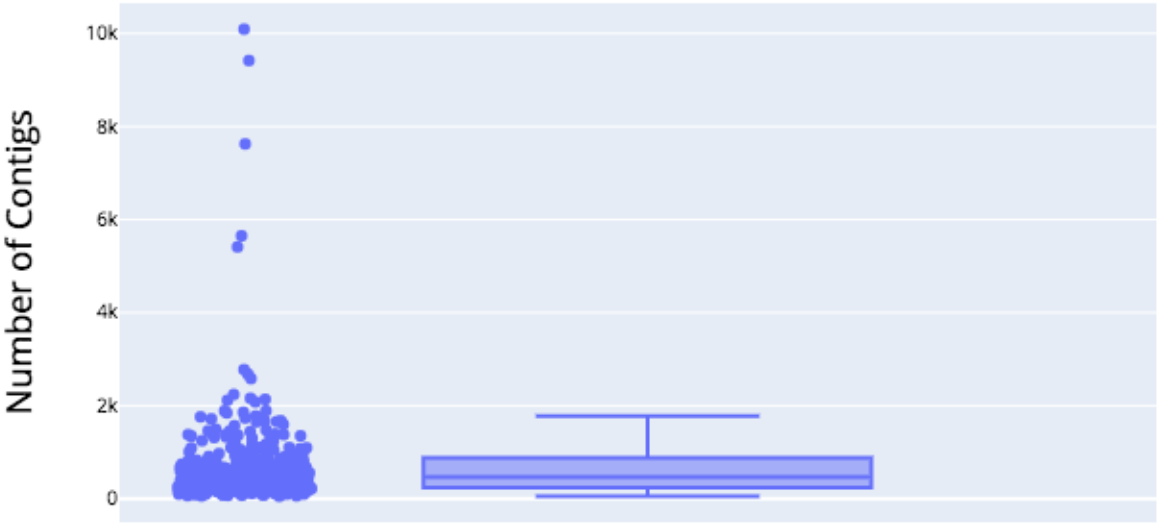
bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 2: Inferred parameters for the Finitely Many Genes model for each of the 51 clusters of the global pneumococcal sequencing project for which reliable dated phylogenies could be inferred. The log odds ratio of invasive disease, number of unique serotypes and recombination rate given in (21) are also plotted for each cluster.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

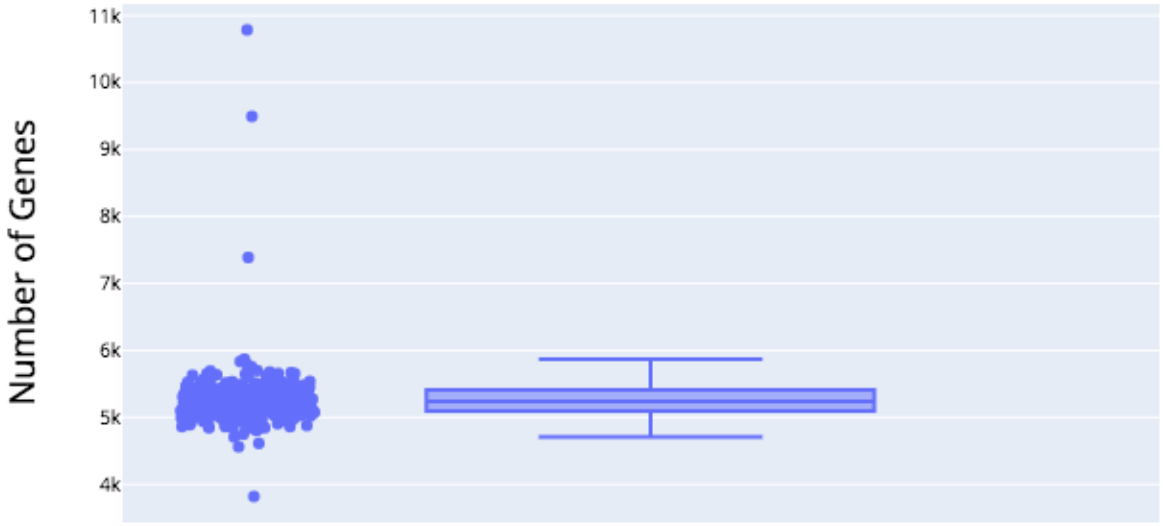


Supplementary Figure 3: Multi Dimensional Scaling (MDS) plot of pairwise mash distances between isolates in the global *K. penumoniae* dataset. This plot is produced by the Panaroo quality control script.



Supplementary Figure 4: Boxplot produced by the Panaroo quality control script indicating the number of contigs in each of the *K. pneumoniae* assemblies.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.28.922989; this version posted January 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 5: Boxplot produced by the Panaroo quality control script indicating the number of gene annotations in each of the *K. pneumoniae* assemblies.