Phenetic Comparison of Prokaryotic Genomes Using k-mers

Maxime Déraspe,[†,1,2,3] Frédéric Raymond,[†,1,2] Sébastien Boisvert,[4] Alexander Culley,[5] Paul H. Roy,[1,5] François Laviolette,[‡,2,6] and Jacques Corbeil[*,‡,1,2,3]

[1]Centre de Recherche en Infectiologie, CHU de Québec-Université Laval, Quebec City, QC, Canada

[2]Centre de Recherche en Données Massives de l'Université Laval, Quebec City, QC, Canada

[3]Département de Médecine Moléculaire, Université Laval, Quebec City, QC, Canada

[4]Gydle Inc., Quebec City, QC, Canada

[5]Département de Biochimie, Microbiologie et Bio-informatique, Université Laval, Quebec City, QC, Canada

[6]Département d'Informatique et de Génie Logiciel, Université Laval, Quebec City, QC, Canada

[†]These authors contributed equally to this work.

[‡]Shared senior authorship.

***Corresponding author:** E-mail: jacques.corbeil@genome.ulaval.ca.

Associate editor: Miriam Barlow

Abstract

Bacterial genomics studies are getting more extensive and complex, requiring new ways to envision analyses. Using the Ray Surveyor software, we demonstrate that comparison of genomes based on their k-mer content allows reconstruction of phenetic trees without the need of prior data curation, such as core genome alignment of a species. We validated the methodology using simulated genomes and previously published phylogenomic studies of *Streptococcus pneumoniae* and *Pseudomonas aeruginosa*. We also investigated the relationship of specific genetic determinants with bacterial population structures. By comparing clusters from the complete genomic content of a genome population with clusters from specific functional categories of genes, we can determine how the population structures are correlated. Indeed, the strain clustering based on a subset of k-mers allows determination of its similarity with the whole genome clusters. We also applied this methodology on 42 species of bacteria to determine the correlational significance of five important bacterial genomic characteristics. For example, intrinsic resistance is more important in *P. aeruginosa* than in *S. pneumoniae*, and the former has increased correlation of its population structure with antibiotic resistance genes. The global view of the pangenome of bacteria also demonstrated the taxa-dependent interaction of population structure with antibiotic resistance, bacteriophage, plasmid, and mobile element k-mer data sets.

Key words: **comparative genomics, microbial evolution, population structure, horizontal gene transfer, software.**

Article

Introduction

Genomic data sets are continuously increasing in size and a single study now contains hundreds to thousands of samples that must be rigorously compared and clustered (Nasser et al. 2014; Walsh et al. 2016). Large-scale genomic projects such as the 1,000 genomes project (Siva 2010), the Human Microbiome Project (Integrative HMP [iHMP] Research Network Consortium 2014) or any recent epidemiological studies of outbreaks (Editor 2011; Snitkin et al. 2012; Gire et al. 2014) rely on comparative genomics and large scale phylogenies to uncover underlying biological patterns and trends. Nowadays, sequenced genomes are compared based on conserved genes, polymorphic positions and/or annotations (16S rRNA, *rpoB*, *atpB*, etc.; Patwardhan et al. 2014). For example, multilocus sequence analysis (MLSA) uses the sequences of housekeeping genes to construct phylogenies (Glaeser and Kämpfer 2015). On a larger scale, phylogenomics often compare genomes using the conserved genes of the population under study (Pennisi 2008). Another common approach for whole genome comparison is the Average Nucleotide Identity (ANI) that relies on sequence alignments in order to determine the percentage of similarity between genomes (Konstantinidis et al. 2006). Researchers are thus often interpreting their results solely based on a comparison of the shared features of their samples, an approach that may omit important genomic determinants that could better characterize and discriminate subpopulations or phenotypes (Tu and Lin 2016). Indeed, the accessory or dispensable genome can be responsible for important phenotypes such as antibiotic resistance, adaptation to specific environments or colonization of different hosts (Medini et al. 2005). Genes acquired by horizontal gene transfer (HGT) are not measured by traditional methods that use conserved genes to compute evolutionary distance between bacteria. Given the importance of the accessory genome in pathogen traits, such as virulence and antibiotic resistance, it is of interest to have analytical tools capable of comparing thousands of genome sequences without reducing analysis to conserved features.

K-mer-based methodologies are not new and have attracted researchers' interest for quite a while now

(Vinga and Almeida 2003; Song et al. 2014; Haubold 2014). It is the gold standard for short read assemblies with De Bruijn graphs (Compeau et al. 2011; Boisvert et al. 2012) and there are several highly efficient k-mer counters, like MSPKmerCounter (Li and Yan 2015), DSK (Rizk et al. 2013), and KMC2 (Deorowicz et al. 2014). Alignment-free sequence comparisons have been studied in numerous ways and are competitive with alignment-based methods in terms of accuracy while being generally computationally more efficient (Marçais and Kingsford 2011; Gardner et al. 2015; Ondov et al. 2016). They have also been used for the comparison of assembled microbiomes (Raymond et al. 2016b) and proved to be an important tool in the phylogenetic analysis toolbox (Qi et al. 2004; Wen et al. 2014). Comparison of k-mer content can also be combined with machine learning algorithms to predict phenotypes such as antibiotic resistance (Drouin et al. 2016).

In this work, we evaluated whether k-mers can be used to rapidly and accurately compare large collections of genomes. With this approach, genomes are clustered based on the similarity of their complete sequence by counting the total number of shared k-mers, including the accessory genome. In addition, we tested the hypothesis that it is possible to characterize populations of genomes based on specific features using presence/absence of k-mers related to these features. To do so, we filtered genome sequences by selecting only k-mers that were also present in a reference sequence data set, and then compared the clustering of whole genomes against the filtered genomes. The purpose of the filtered data set is to establish a functional set of genes with common characteristics. We then suggest that if genome clustering based on specific gene functions restores the population structure based on whole genomes, this functional set of genes is linked to the structure of the population under study. This suggests that the functional set of genes could have a conserved function in the population and presumably a selective pressure similar to the whole genomes, for example. On the basis of this logic, we explored this relationship by comparing a large number of bacterial genomes with several gene sequence data sets, each one representing a different functional gene category. We used reference sequence data sets of antibiotic resistance genes (ARG), insertion sequences, plasmids, bacteriophages and biosynthetic gene clusters (BGC) and observed their relationship with genome population structure for different bacterial species. This approach is implemented in the Ray Surveyor software, which is built on top of the scalable Ray framework (Boisvert et al. 2012, 2010). The defining feature of Ray Surveyor is the ability to compare whole genomes based on their complete set of k-mers along subsets of their k-mers, filtered with other sequence data sets. Ray Surveyor allowed us to determine how the five genetic element categories tested are linked with the population structure of 42 species of bacteria.

Results and Discussion

Validation with Simulated Genome Populations

To overcome possible uncertainties introduced by real genome data sets, we started by generating random phylogenetic trees (Kuhner and Felsenstein 1994; Guindon and Gascuel 2002; Boc et al. 2012) and simulating genome sequences from these trees (Spielman and Wilke 2015). Three different branch lengths were used to simulate tree structures in order to measure the impact of this parameter on the clustering methods used in Ray Surveyor analyses. The branch lengths were computed using an exponential distribution, which yielded an average depth of $log_2(n)$ with n being the number of genomes in the tree, 100 in our case. For each average branch length, ten random trees were computed to evaluate reproducibility. Sequences of one million nucleotides were produced for each simulated genome in the phylogenies in the form of an alignment, using Pyvolve (Spielman and Wilke 2015).

The three branch lengths we examined were chosen to model bacterial populations of within-species genomes (0.001), within-genera genomes (0.005), and interspecies genomes (0.01). This assumption was based on the ANI of all simulated trees. The ANI cutoff to distinguish bacterial species is estimated to be between 93 and 96% ANI (Rossello-Mora and Amann 2015). Consequently, trees with an average branch length of 0.001 (average ANI = 98.3%) are akin to intraspecies data sets and branch lengths of 0.01 (average ANI = 85.4%) to interspecies data sets. An average branch length of 0.005 corresponds to an ANI of 92.1% between all pairs of genomes, with 56.5% of them being below 93%. Therefore, in trees with an average branch length of 0.005, half of the genomes belong to the same bacterial species whereas the other half belongs to different species from the same genera. Although these cut-offs do not apply to all bacterial species, they generally reflect the current state of the NCBI taxonomy and they allow the evaluation of the influence of strain diversity on comparative genomics methods.

To allow comparison of Ray Surveyor clusters with phylogenies, we took the distance matrices derived from the simulated trees and generated a dendrogram by hierarchical clustering with the UPGMA linkage method. Similarly, the k-mer Gram matrices generated with Ray Surveyor were transformed into distance matrices upon which hierarchical clustering dendrograms were computed. Those dendrograms are referred to as phenetic trees throughout the manuscript. The cophenetic correlation coefficient (CCC; Sokal and Rohlf 1962) was then used to assess how Ray Surveyor phenetic trees correlated with the simulated phenetic trees. The CCC in our case measures how well two phenetic trees preserve the pairwise distances between all pairs of genomes. We tested the impact of four distance metrics on the transformation of the Ray Surveyor Gram matrix using Euclidean, cosine, correlation and Canberra distances. Ray Surveyor analyses were also performed with k-mer lengths ranging from 11 to 101 nucleotides to evaluate their impact on accuracy.

The cophenetic correlation results from Ray Surveyor analyses were affected by the average pairwise phylogenetic distance of genomes and the k-mer lengths used in the analysis (fig. 1A). Indeed, CCCs were higher for intraspecies genome populations (lower average pairwise distance) and were only slightly affected by k-mer length or distance metrics. When genome populations grew more distant,

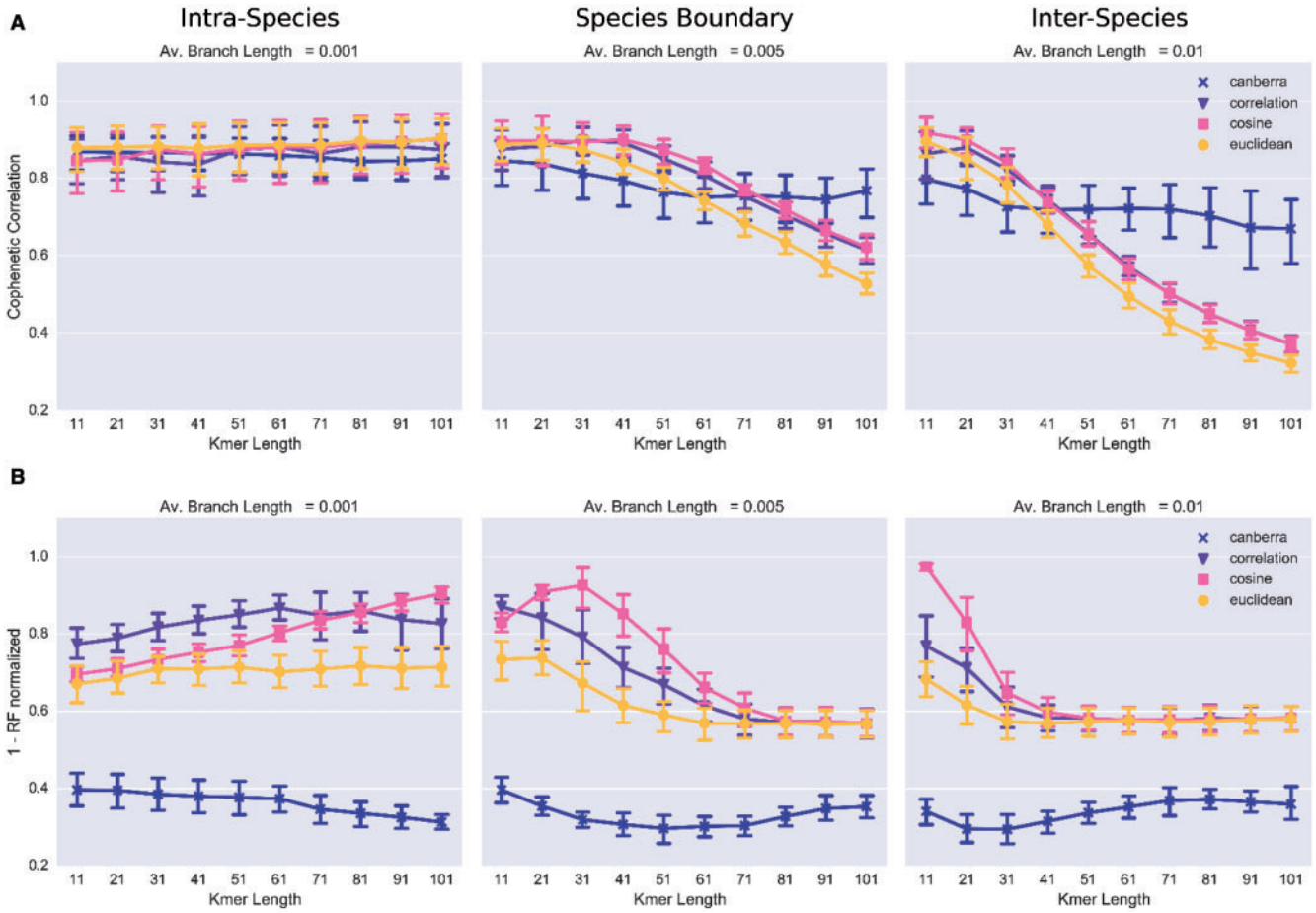

FIG. 1. Evaluation of simulated genome populations with Ray Surveyor. Colors and symbols represent the distance metrics used to transform the Ray Surveyor's Gram matrix into a distance matrix. Each column represents a different evolutionary distance between the genomes, based on the average branch length and bacterial species definition. Ten replicates were performed for each point. First row (*A*) is the cophenetic correlation between the reference phylogeny and the phenetic tree. Second row (*B*) is the Robinson–Foulds metric between the reference phylogeny and the Ray Surveyor derived tree.

crossing the species boundary, CCCs decreased with increasing k-mer length. By comparing distance metrics used to construct phenetic trees based on Ray Surveyor results, we observed that Euclidean, cosine, and correlation distances behaved similarly on simulated genome populations (see Materials and Methods). The Canberra distance provided lower CCC for more closely related genomes, but it was less affected by more heterogeneous populations when the k-mer lengths were increased. This result is likely due to the fact that the Canberra distance is more tolerant of low absolute values (the number of shared k-mers), as observed by Loureiro et al. (2004). For a control, we produced alignment-based phylogenies of the simulated sequences that had average CCCs of 0.98 for branch lengths of 0.001, 0.97 for 0.005 and 0.99 for 0.01.

In order to test the capability of Ray Surveyor to restore good topologies for phylogenetic trees, we also computed a Neighbor-Joining tree for all the distance matrices. In this comparison, we used the original simulated phylogenetic tree against those derived from Ray Surveyor. The Robinson–Foulds (RF) metric allows a comparison of unrooted phylogenetic trees, essentially by measuring the number of changes required to align two trees together by transforming one tree into the other (Robinson and Foulds 1981). Similar to the cophenetic correlation, the RF results varied with sequence diversity and k-mer length (fig. 1*B*). For the intraspecies genome populations (branch length = 0.001, average ANI = 98.3%) longer k-mer length performed better and peaked with the 101-mers and the cosine metrics. At the species boundary, the cosine distance metrics yielded the best topological trees with 31-mers. When comparing genomes of different species (branch length = 0.01, average ANI = 85.4%), a k-mer length < of 31 yielded better topological trees for the cosine, correlation and Euclidean metrics.

On the basis of these results and on the literature, the choice of k-mer length can be seen as a trade-off between sensitivity and specificity (Ondov et al. 2016). Evolutionarily distant genomes require shorter k-mers to get a good signal (sensitivity) whereas more similar genomes benefit from larger k-mer lengths for more specificity. Moreover, previous studies have shown the efficiency of 31-mers in genome clustering (Melsted and Pritchard 2011) and the robustness in bacterial metagenome profiling (Boisvert et al. 2012) when this length of k-mer is used. For the following analyses on real genome data sets, we selected a length of 31-mers, which offers a compromise between sensitivity and specificity for

both intraspecies and interspecies comparison. We also focused our analyses on the cophenetic correlation for the phenetic trees, since we needed to characterize genomes based on specific genetic elements rather than finding their ancestral history.

Population Scale Genomics with k-mers

This section aims to benchmark the application of the Ray Surveyor genome comparison in comparative genomics projects and to assess how it performs on microbial populations of different scales. As a first step, we validated that k-mer-based phenetic trees accurately reflected previously determined phylogenies based on publicly available comparative genomic studies of *Streptococcus pneumoniae* and *Pseudomonas aeruginosa* (fig. 2). For *P. aeruginosa*, 387 genomes were taken from a study by Kos et al. (2015; fig. 2A). For *S. pneumoniae*, a first data set of 616 genomes from Croucher and collaborators was used, along with a second data set comprising 173 genomes previously studied by Hilty and collaborators to investigate the difference between encapsulated and nonencapsulated pneumococci (Croucher et al. 2013; Donati et al. 2010; Hilty et al. 2014). Whole genome phylogenies were obtained from the authors for the Kos and the Croucher data sets, while the phylogeny for the Hilty collection was built using 602 conserved genes. We calculated the cophenetic correlation between the phenetic trees (hierarchical cluster dendrograms) created using Ray Surveyor and the derived phenetic trees from the phylogenies for these three data sets (fig. 2A). All four distance metrics (see Materials and Methods) performed above 0.91 CCC on *P. aeruginosa*, with the Canberra distance yielding the highest CCC of 0.97. Correlation distance had the highest CCC (0.92) compared with other distance metrics ($<$0.75) for *S. pneumoniae*. The Hilty and collaborators data set of *S. pneumoniae* genomes was tested and provided 0.89 CCC between correlation distance based on k-mers and the core genome phenetic tree. Heatmaps representing the clustering based on the distance between isolates of the Croucher and Kos data sets are shown in supplementary figures 1 and 2, Supplementary Material online.

This approach can also be used to quickly add a new genome to an existing phylogeny. For example, we added the recently sequenced genome of *P. aeruginosa* strain E6130952 to the Kos *et al.* genome collection (CP020603.1 [https://www.ncbi.nlm.nih.gov/nuccore/CP020603; last accessed July 19, 2017]; supplementary fig. 3, Supplementary Material online). This pathogenic strain was isolated from a patient with respiratory failure and was resistant to all tested antibiotics, including colistin (Xiong et al. 2017). The closest isolate in the phylogeny (AZPAE14730) was also resistant to levofloxacin, meropenem, and amikacin, but not to colistin (Kos et al. 2015). Both strains have a similar genome size and share 97% of their k-mers.

In epidemiological studies, genomes are often classified based on experimentally derived categories such as multilocus sequence typing or serotypes. The Fowlkes–Mallows index (FMI) allows calculation of the similarity between two clusterings (Fowlkes and Mallows 1983) and can be used to compare clustering based on k-mers or phylogeny to categorical information of clinical relevance. Thus, we used this metric to quantify the concordance between the clusters generated with Ray Surveyor or with phylogeny to metadata associated with genomes. Therefore, we calculated the FMI between clustering based on the phylogenetic and phenetic trees of *P. aeruginosa* and *S. pneumoniae* when compared with MLST and serotype genome classification, for a range of 2 to *N* clusters (fig. 2B). Phylogenetic genome comparison and k-mer-based comparison provided similar results when compared with MLST or serotype categorization. The highest divergence in FMI between phylogeny and k-mers was $<$5%. Similarity with MLST was higher ($\geq$ 85%) than similarity with serotype ($\leq$ 67%), suggesting that MLST is more related to complete genome phylogeny than serotype. Indeed, in *S. pneumoniae*, the capsular operon can be modified through capsular switching, a process that decouples serotypes from the core and accessory genomes (Andam and Hanage 2015). In the Hilty data set, genomes from different strain types could be associated within the category of nonencapsulated *S. pneumoniae*, thus explaining the low FMI of serotypes in comparison to the near-perfect FMI obtained when benchmarking against MLST results.

In order to explore Ray Surveyor's capacity to work with a large number of distantly related genomes, we created a data set of 2,429 complete genomes from 30 phyla in the domain *Bacteria*. The 2,429 bacterial genomes from which this data set was derived were selected in order to limit the bias caused by a relative overrepresentation of certain genomes in the public database, such as laboratory strains of *Escherichia coli* or clonal isolates from epidemiological studies. We compared the phenetic tree built with these genomes using Ray Surveyor to the 16S rRNA phylogenetic tree of these strains. Canberra distance was the best performing metric (0.69 CCC compared with $<$0.10 for other distance metrics) for the tree of 2,429 bacterial genomes, most certainly because of the low number of shared k-mers between distant genomes (fig. 3A). We also used the FMI to compare 16S phylogenetic and Ray Surveyor phenetic trees to the taxonomical classification of genomes at the family rank based on the NCBI taxonomy. Although the NCBI taxonomy may not always be in line with other taxonomies, it provides a convenient way to perform taxonomy-related analyses with genomic sequences obtained from NCBI (Federhen 2012; Balvočit and Huson 2017). When comparing the classification of 2,429 genomes from 262 bacterial families to genome-based clustering, the peak FMI was 67% for k-mers (469 clusters) compared with 68% for 16S phylogeny (310 clusters; fig. 3B). While these methods had similar correlations with current NCBI taxonomy at the family rank, we also observed that the accuracy of clusters was influenced by the number of genomes within each bacterial family (fig. 3B). When considering only bacterial families represented by at least 20 genome sequences (39 families), k-mers had a maximal FMI value of 78% (at 167 clusters) compared with phylogeny which had a maximal value of 77% at 107 clusters. In contrast, when considering families represented by $<$20 genomes (223 families), FMI was 62% for k-mer analysis (378 clusters) compared with 71% for 16S rRNA phylogenetic trees (451 clusters). The discrepancies

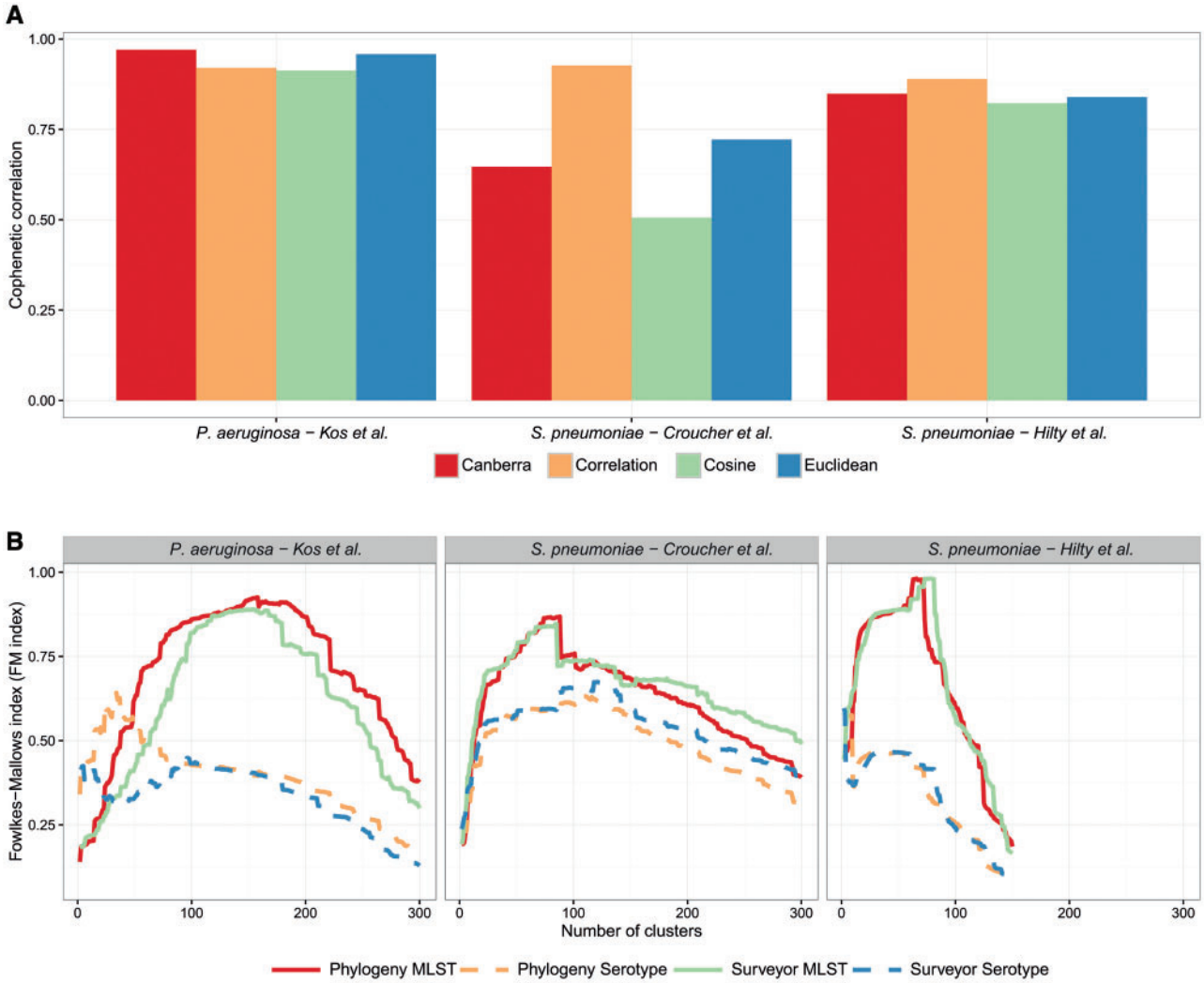

Fig. 2. Comparison of phenetic trees created using Ray Surveyor to phylogenies calculated using conserved genomes or marker genes for *Pseudomonas aeruginosa* and *Streptococcus pneumoniae*. (*A*) Cophenetic correlation between alignment-based phylogeny and phenetic trees calculated using four different distance metrics. (*B*) Fowlkes–Marlows index comparing clustering done using Ray Surveyor (correlation distance metric) and phylogeny compared with classification based on multiple locus sequence typing or serotypes.

between 16S rRNA phylogeny and k-mer-based clustering were mainly associated with regions where only a small number of genomes were included in the analysis. Additionally, the low count of shared k-mers between these small groups of genomes and the rest of the taxa makes it hard to find common ancestors and thus infer their correct placement in the final dendrogram. Hence, efficient clustering of phylogenetically distant bacteria that share a nonsignificant amount of k-mers would require more intermediate genomes to effectively drive the hierarchical clustering and a shorter k-mer length to get more signal.

To investigate the relationship between traits and genome clustering, quantitative and qualitative metadata can be plotted against a phenetic tree. For example, supplementary figure 4, Supplementary Material online, plots a phenetic tree of 2,429 bacterial genomes versus their GC-content and their taxonomic class rank. In this representation, differences in GC-content seem related to the taxonomical classification. Because phenetic trees do not rely on sequence alignments, we cannot correct for GC-content or codon bias using substitution models or other methods, as suggested in the literature (Mooers et al. 2000). Therefore, we do not expect branch lengths, generated using our k-mer approach, to be representative of evolutionary distance. The clustering of high taxonomic rank could also be biased by GC content (Mooers and Holmes 2000). At the k-mer level, differences in GC-content and codon usage should negatively affect k-mer similarity. Indeed, k-mer similarity is expected to decrease quickly as the number of mismatches increase. Previous studies have shown that the type of environment and particular lifestyles of the bacteria is related to genomic GC-content and codon usage (Foerstner et al. 2005; Botzman and Margalit 2011; Lassalle et al. 2015). Differences in ecological niches are also reflected in the accessory genome, which can lead to large differences in k-mer content (Medini et al. 2005).

Comparing Genomes Based on Specific Traits

Not all genes within a genome have the same association with the evolutionary story of a species as inferred from phylogeny (Land et al. 2015). For example, genes acquired by HGT may

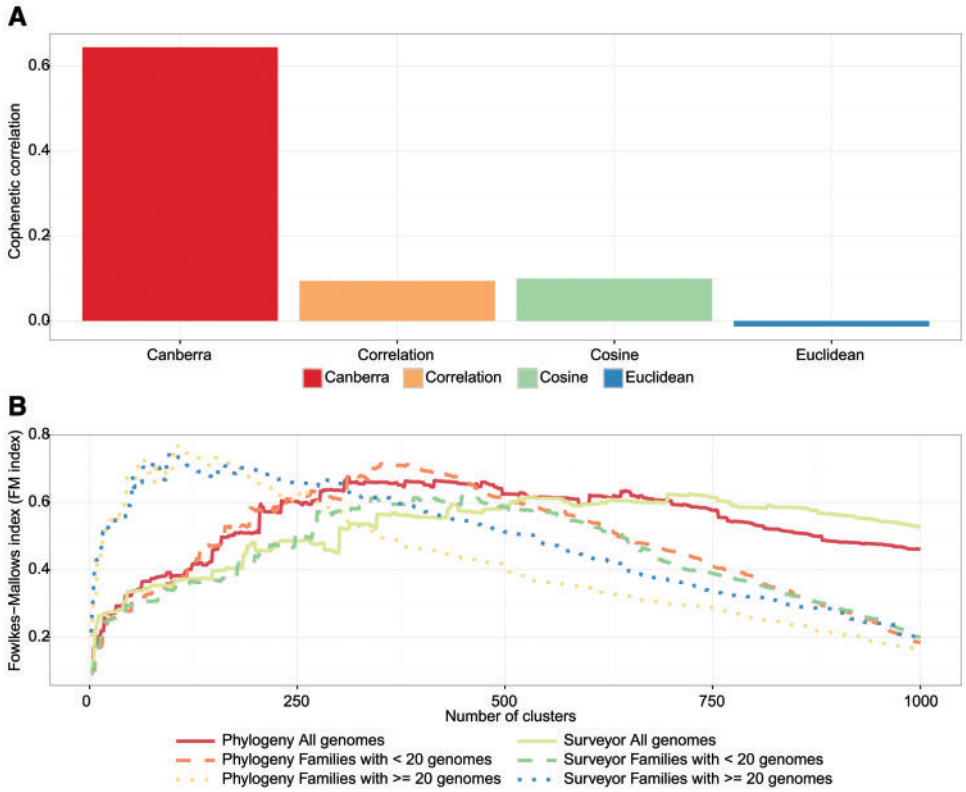

Fig. 3. Comparison of phenetic trees created using Ray Surveyor to phylogeny based on 16S gene sequence for 2,429 bacterial genomes. (*A*) Cophenetic correlation between alignment-based phylogeny and phenetic trees calculated using four different distance metrics. (*B*) Fowlkes–Marlows index comparing clustering done using Ray Surveyor (correlation distance metric) and phylogeny compared with taxonomical classification at the family rank.

not be linked to the phylogeny of a species and may have been acquired independently by different strains, for example, genes from mobile elements, bacteriophages or plasmids (Philippe and Douady 2003). Resistance genes as well as secondary metabolite operons (Dobrindt et al. 2004) can also be disseminated by HGT (Pärnänen et al. 2016).

In order to investigate HGT patterns in our data set, we developed an approach to quantify how the phenetic tree generated using a subset of k-mers reflected the tree generated using the total k-mer content of a genome. We hypothesize that if the two trees are correlated, the group of k-mers is linked to the phylogeny of the studied population. Conversely, the absence of correlation indicates independence between the whole genome population and the filtered genome population. The first steps to conduct the analyses are similar to the ones' explained in the two previous sections. We first calculated a Gram matrix of shared k-mers for all pairs of genomes. For each population two Gram matrices were produced, one with the total count of shared k-mers between the genomes and the second containing only the count of shared k-mers included in the filtering data sets. We then generated a distance matrix for the complete and filtered Gram matrices using the Canberra distance, which we chose in order to reduce bias caused by samples with a limited number of filtered k-mers. Phenetic trees were then built using UPGMA clustering on the distance matrices. We aligned the heatmaps of the clusters based on the whole genome phenetic tree to visualize its similarity with the filtered phenetic tree. In addition, the correlation between phenetic trees based on complete k-mer content and filtered k-mer sets was quantified using CCC. A coefficient of 0 indicates the absence of correlation whereas a coefficient of 1 indicates perfect cophenetic correlation between selected k-mers and complete genomes, thereby suggesting that these k-mers are associated with the phylogeny of the population.

In our initial analysis, we further investigated genome populations of *S. pneumoniae* and *P. aeruginosa* and the 2,429 bacterial genomes using subsets of k-mers that could be acquired through HGT and may have an impact on the evolution of bacterial species. We used five filtering data sets: mobile elements (insertion sequences), resistance genes, bacteriophages, plasmids, and BGC. The filtering analyses were produced using the strict inclusion of k-mers from the filtering data sets. However, for the plasmids filtering, we also excluded the k-mers from the resistance genes and mobile elements data sets as these genetic elements often co-appear on plasmids and chromosomes. As represented in figure 4, the coherence between the heatmaps based on filtering and those based on complete genomes, also expressed quantitatively by the cophenetic correlation, is different between filtering data sets and genome collections. *Streptococcus pneumoniae* showed low (0.28 CCC) correlation between antibiotic resistance k-mers and complete genome clustering. BGC (0.48 CCC) and plasmids (0.52 CCC) had moderate

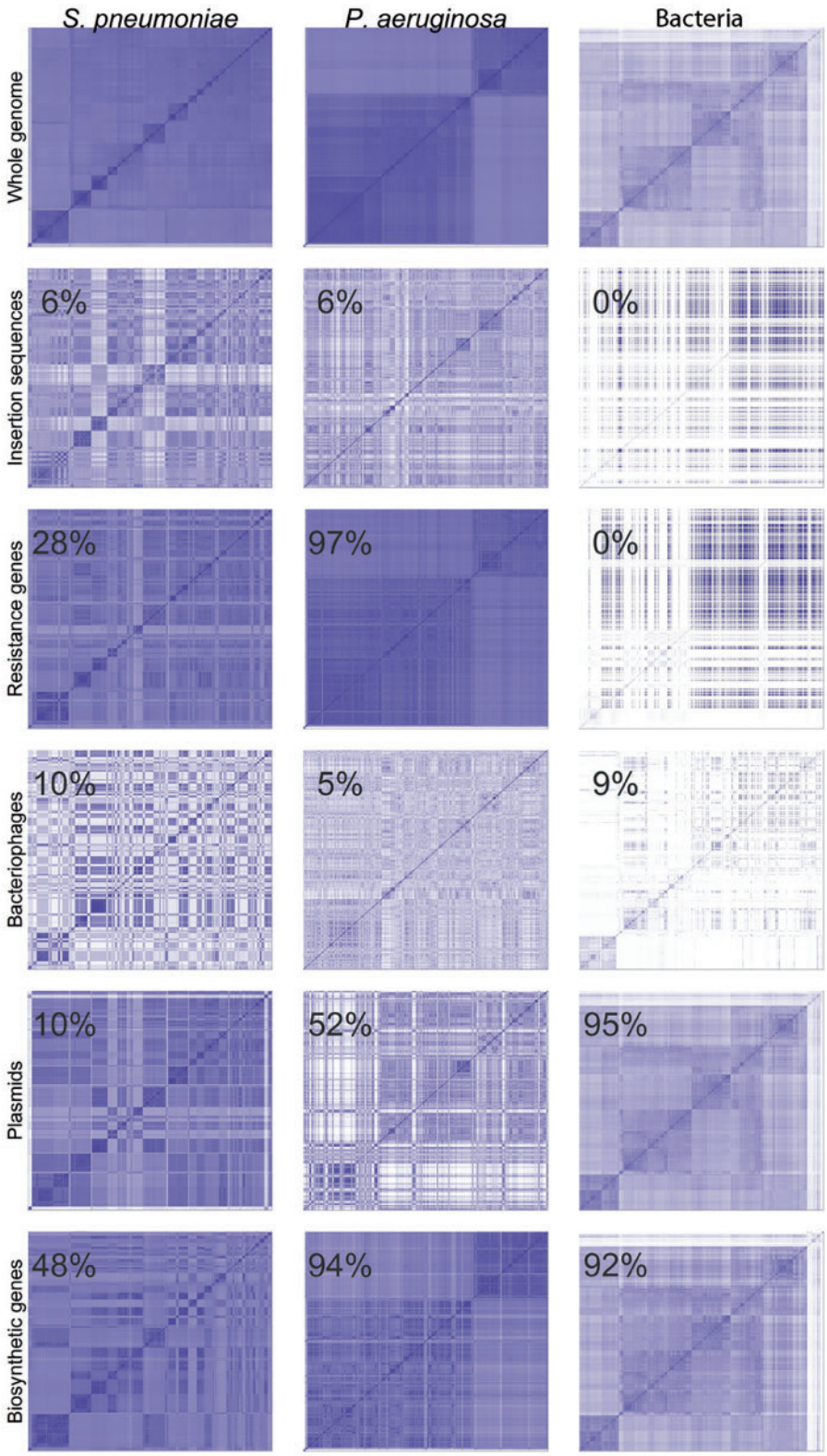

Fig. 4. Comparison of the relationship between strains when genome sequences are filtered using one of five filtering data sets for *Streptococcus pneumoniae*, *Pseudomonas aeruginosa* and the 2,429 representative bacterial genomes. The Heatmap represents the Canberra distance between genomes collated on a subset of k-mers. The *X* and *Y* axis of the heatmap are genomes ordered based on hierarchical clustering of the complete genome. The number in top left corner of heatmaps is the cophenetic distance, expressed in percentages, between filtered data sets and whole genome phenetic tree. The darker the shade of blue, the higher the similarity between samples.

correlation with complete genome clustering. The genomes harbor on average 403 and 5,636 k-mers for BGC and plasmids, respectively, suggesting sequences from these origins are not widely abundant in the species, although they are correlated with the structure of the population. *Streptococcus pneumoniae* does not frequently harbor plasmids, which is reflected in the count of k-mers related to these genetic elements (Romero et al. 2007). The lack of characterized BGC from the species in the filtering data set could also have an impact on the moderate correlation. In contrast, the *P. aeruginosa* phenetic trees based on resistance genes (0.98 CCC) and BGC (0.92 CCC) were highly correlated with phenetic trees based on the whole genome. The number of shared k-mers associated with the two filtering data sets was on average 41,240 and 63,820 k-mers, respectively. Similar results were obtained on 71 genomes of *P. aeruginosa* downloaded from the PATRIC database (Wattam et al. 2014), which included some environmental samples, and on 500 *P. aeruginosa* genomes randomly selected from NCBI (supplementary fig. 5, Supplementary Material online). In the case of the 2,429 bacterial genomes data set, the whole genome phylogeny was highly correlated with plasmids and BGC. The overall relationship between representative taxa in the domain *Bacteria* was not distinctively defined by resistance genes, which are broadly distributed in the microbial tree of life and can be associated with HGT (Metcalf et al. 2014)

In order to dissect the relationship between bacterial pathogens and the five filtering data sets, we applied the methodology described above to 42 bacterial species for which at least 100 genomes were available in the NCBI RefSeq database (fig. 5). These taxa are associated with human infections, with the exception of *Lactobacillus plantarum* which is found in fermented food (van den Nieuwboer et al. 2016). Our hypothesis is that high cophenetic correlation of clustering between complete and filtered k-mer content is a good indicator of how the tested elements are related to the phylogeny of the species.

The majority of the gammaproteobacteria had strong correlations with the ARG data set, especially species from *Klebsiella*, *Escherichia*, *Enterobacter*, *Vibrio*, *Pseudomonas*, and *Acinetobacter*. This could be related to the large number of intrinsic resistance determinants characterized in those species, especially the drug efflux systems (Rodionov et al. 2001). Other studies have put into evidence the importance of bacteriophages and plasmids in the ongoing evolution of the *Vibrio* genus (Hazen et al. 2010), as reflected in figure 5. *Shigella flexneri* is the Proteobacteria with the highest correlation with bacteriophages (0.83 CCC). Indeed, their O-antigens were often modified by serotype-converting bacteriophages (Allison and Verma 2000; Sun et al. 2013). To further investigate this question, we used alignments to validate which bacteriophages used for filtering would be found in the 147 *Shigella* genomes. Interestingly, we found some specific prophage sequences that could delineate the clusters seen with clustering based only on phage k-mers (supplementary fig. 6, Supplementary Material online). Polysaccharides-related BGC, which encode capsular antigens and O-antigens, could thus explain the high CCC of BGC for *S. flexneri* and *Vibrio cholerae* (Cimermancic et al. 2014). On the other hand, *E. coli* has several characterized BGC in the MIBiG database while showing moderate correlation with the whole genome (0.49 CCC; Medema et al. 2015b). Comparison of clustering between whole genome and BGC of *E. coli* indicate that a portion of the population can be delineated by BGC while others seem unrelated (supplementary fig. 7, Supplementary Material online). The *Francisella tularensis* genome can contain over 100 insertion sequence genes (Larsson et al. 2009), which could explain its high correlation with mobile elements. In opposition to most of the tested species, *F. tularensis* was also significantly correlated with plasmids. This high correlation could be related to a misannotated 100 kb plasmid that is in fact part of the *F. tularensis* genome (CP010448.1 which was replaced by CP010446.2). This large chromosomal region could indeed have boosted the impact of plasmids in the correlation observed, as it is integrated to the genome. It is important to consider that for most genomes in RefSeq, the plasmid sequences are found under a different accession number than the genome, therefore it is not considered in the clustering. In whole genome shotgun sequencing, plasmid sequences are generally included in the assemblies, thus plasmid filtering could prove useful to exclude these sequences from whole genome comparisons.

Six species from the *Firmicutes* phylum had correlation >0.70 CCC with ARG. *Bacillus anthracis*, *B. cereus* and *B. subtilis* were all above 0.85 CCC for ARG. This high correlation could originate from the chromosome-encoded β-lactamases harbored by the species (Colombo et al. 2004; Fenselau et al. 2008; Materon et al. 2003). The other members of the phylum, *Firmicutes* having good correlation with ARG, were *Listeria monocytogenes* 0.93 CCC, *Enterococcus faecalis* 0.82 CCC, and *S. pneumoniae* 0.70 CCC. The three *Bacillus* species also had CCC >0.70 for BGC. Bacilli are known to produce several types of secondary metabolites (Sansinenea and Ortiz 2011). All the *Firmicutes* analyzed were below 0.55 CCC with the mobile elements data set. In *Firmicutes*, bacteriophages had best correlations with *Streptococcus suis* (0.87 CCC) and *B. anthracis* (0.98 CCC). The 500 *S. suis* genomes had an important number of k-mers associated with phages (18,642 in average), that along with the correlation, supported the idea that prophage sequences in the species are linked to the whole genome phylogeny. It was also shown in previous observations that remnants of phage sequences are distributed throughout *S. suis* genomes (Tang et al. 2013). *Bacillus anthracis* had a high correlation with bacteriophages compared with the other *Bacillus* species, although shared k-mers from the filtering data set were not numerous (3,936 in average). The correlation could be related to four defective and conserved prophages harbored by the species as reported in Sozhamannan et al. (2006). In agreement with our results, they suggested that these prophages could be used as a chromosomal signature of the species. Bacteriophages could also be associated with ecological adaptation in *B. anthracis* (Schuch and Fischetti 2009).

Overall, the interpretation of the results represented in figure 5 supports our hypothesis that correlation between filtered genomes and complete genomes indicates a

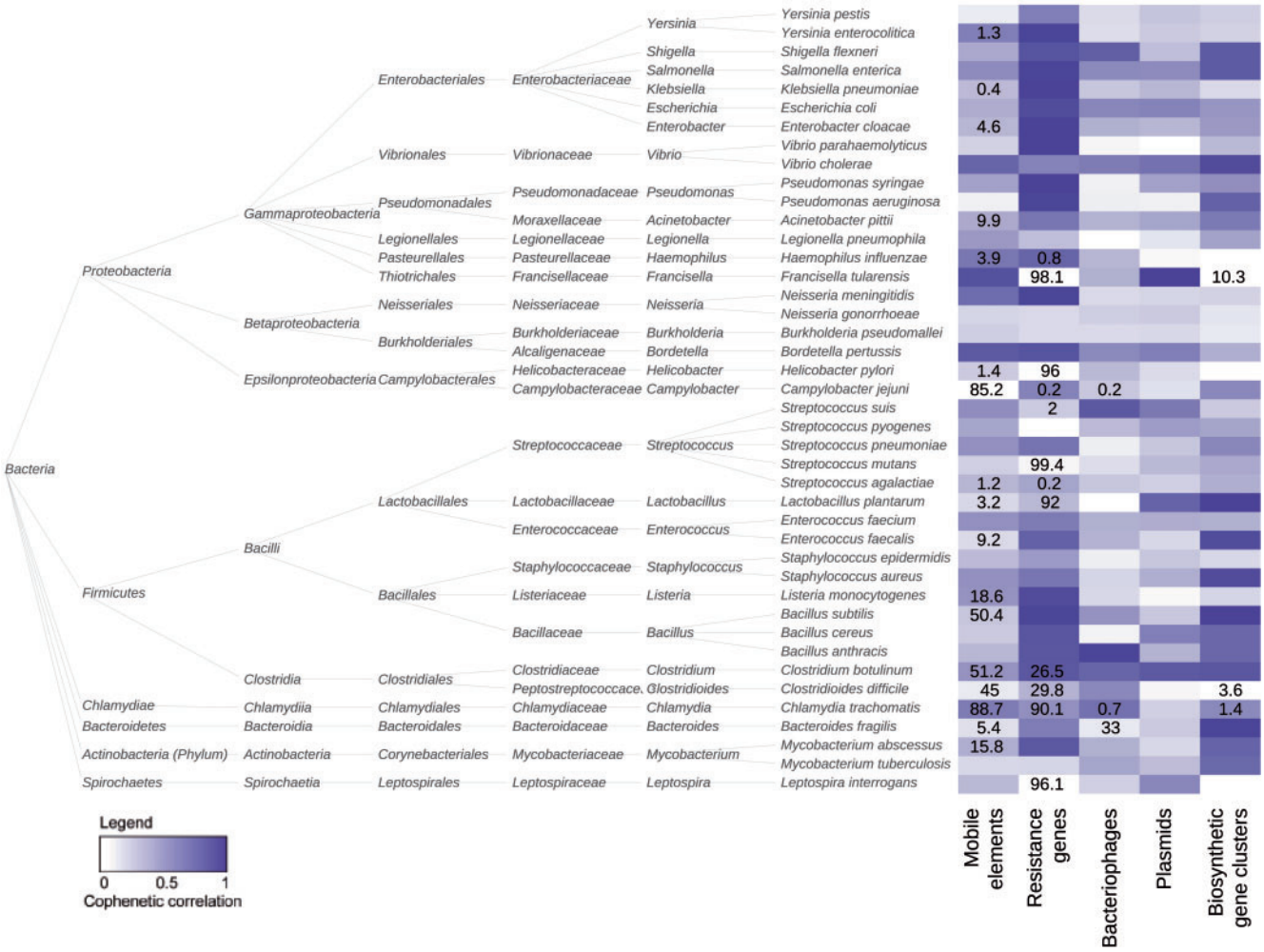

Fig. 5. Cophenetic distance between phenetic trees based on whole genome and filtered data sets for 42 bacterial species from RefSeq that included at least 100 genomes. Intensity of heatmap represents the cophenetic correlation as shown in the legend. Numbers in the heatmap are percentages of genomes with zero k-mers associated with relevant filtering data set.

relationship between selected k-mers and a species. In many cases, we observed that a cophenetic correlation occurred in species where potentially mobile genetic elements were integrated in the genome. Thus, this methodology could potentially indicate integration and conservation of these elements in the genome of a particular species, or at least their phylotype dependence.

Conclusion

By comparing the k-mer composition of genomes, we were able to reconstruct the phenetic tree of large bacterial epidemiological genomics data sets, as we demonstrated with the *S. pneumoniae* and *P. aeruginosa* data sets. We also evaluated the accuracy of the methods on synthetic genome data sets by testing different parameters that influence this kind of analysis. The methodology is based on whole genome analysis rather than on a subset of core genes, which has been shown to introduce bias (Shapiro et al. 2012; Biek et al. 2015). The use of k-mers allows comparison of genomes based on characteristics that are either conserved or specific. We also applied the method to a data set of 2,429 bacterial genomes spanning the whole bacterial tree of life, without a selection of features such as conserved genes or ribosomal RNA. This approach makes Ray Surveyor an effective tool for scalable analyses in comparative genomics research, among other applications. Using k-mers to build phenetic trees could be used to easily position newly sequenced genomes in the microbial tree of life and infer classification or to determine which branches of the tree of life are not well represented in terms of genome sequences relative to internal taxa diversity.

Analysis of population structures can further be partitioned by filtering subsets of k-mers associated with gene categories or functions. Our results demonstrate that comparison of genomes based on specific subsets of k-mers can reveal their relationship at the population scale. Indeed, without being specific about the genetic determinants involved, the method allows easy determination of strain clusters with similar potential regarding the functions of the filtered data set, such as antibiotic resistance or HGT as shown in this study. A limitation of the filtering approach is that it involves the gathering of sequence data that adequately represents the diversity of the genes or functional category under study. For example, using only reference resistance genes instead of a

large collection of orthologs, paralogs, and variants would underestimate the abundance of resistance genes in genomes containing variants of the reference gene. Still, some sequence types, such as bacteriophages or BGC, could be underrepresented in the databases used in this study. Such sequences, could have potentially resulted in more significant results, provided the availability of a more exhaustive and diverse sequence data set. As seen in figure 4 for the 2,429 bacterial genomes, some clusters of genomes show high bacteriophage signals in comparison to other regions of the heatmap. Indeed, of the 262 bacterial families included in the 2,429 genome analysis, 147 families had ≤ 100 k-mers associated with phage sequences, suggesting that some families could suffer from a lack of characterized phages in the database used for profiling (EBI). This issue should be alleviated by better filtering data sets as more sequences and better annotations become available in public databases.

Ray Surveyor is a powerful tool that allows the reconstruction and interpretation of the phenetic relationships underlying populations of bacterial species. By taking into account clinical or environmental context with the sequence filtering capabilities, this method could allow an intuitive representation of population structures and the genomic features related to their differentiation or phenotype. It is thus a hypothesis-generating tool that could be applied to investigate the importance of specific gene categories not only in pathogens but also in environmental microbial communities and in the analysis of transcriptomic and metagenomic-based research.

Materials and Methods

Theoretical Background and Software Implementation

Ray Surveyor is built on top of the highly scalable Ray framework, which includes the Ray assembler and RayPlatform (Boisvert et al. 2012, 2010). It uses the message-passing interface (MPI) to scale analysis on supercomputers. However, depending on their size, data sets can be analyzed on smaller servers or personal computers. Components of the software include, among others, a sparse distributed hash table to store the k-mers on each computer across a cluster, as well as a graph coloring scheme that associates each k-mer vertex of the de Bruijn graph with its profiling data sets. Ray Surveyor is also based on the actor model (Hewitt 1977); each actor takes care of its own task such as reading and k-merizing input sequences, gathering k-mers into a store keeper, counting the k-mers and building the Gram matrix. supplementary figure 8, Supplementary Material online, provides further details on the actors' roles and their ways of communicating.

The first step of Ray Surveyor is to split the genome sequences into k-mers and build a graph of the pangenome. The k-mer length is set by the user. We recommend using a length between 21 and 61 nucleotides, usually 31 for the comparison of bacterial genomes. The workflow then proceeds with graph coloring, which assigns a virtual color for each k-mer according to the combination of genomes or functional data sets that carry it. The next step is to iterate over each k-mer and increment the count of shared k-mers between each pair of genomes of that color and store them in the Gram matrix. Formally, each pair of genome comparisons can be seen as a simple D_2 statistic (Reinert et al. 2009; Wan et al. 2010) with a binary count (presence/absence) of their k-mers. Since our counts are dichotomic, we can formally define the Ray Surveyor mechanics based on set theory.

Let $A_i = \{k_1, k_2, \ldots, k_{l_A}\}$ be the set of all the k-mers of genome i, and similarly $B_j = \{k_1, k_2, \ldots, k_{l_B}\}$ the set of all the k-mers of genome j. Then, the Gram matrix (K) is defined such that $k_{i,j} = |A_i \cap B_j|$. Let $Z = \{z_1, z_2, \ldots, z_m\}$ be m filtering data sets and $Y = \cup_{n=1}^{m} Z_m$ their union. To filter in (include only) the k-mer set Y, $k_{i,j} = |A_i \cap B_j \cap Y|$ and to filter out (exclude) the k-mer set Y, then $k_{ij} = |(A_i \cap B_j) \backslash Y|$. The resulting matrix K is then normalized to have values in the range [0, 1], with the diagonal entries equal to 1. Consequently, the entries of the normalized matrix K' are given by $k'_{i,j} = \frac{k_{i,j}}{\sqrt{k_{i,i} * k_{j,j}}}$. However, when filtering is used, we recommend division of the entries $k_{i,j}$ by the $k_{i,i}$ and $k_{j,j}$ of the full k-mer matrix, rather than the filtered version. The reason is that the diagonal of the filtered matrix no longer represents the total number of k-mers per genome, but only the number of filtered k-mers, a subset of the genome. This renders the matrices more comparable, as they are all normalized with respect to the same total k-mer content.

After normalization, the matrix is transformed into a distance matrix with a chosen metric. We focused our experiments on four metrics that are the cosine, correlation, Euclidean and Canberra. Below, we formally define the distance formulae by using u and v and the normalized vectors of shared k-mers between a genome and all the other genomes in the population. For instance, the entry $d_{1,2}$ in the distance matrix D, would be defined as $d_{1,2} = 1 - \frac{\mathbf{k}'_1 \cdot \mathbf{k}'_2}{||\mathbf{k}'_1||_2 ||\mathbf{k}'_2||_2}$ for the cosine distance metric. With the vectors $u = \mathbf{k}'_1$ and $v = \mathbf{k}'_2$, here are the formula of the four distance metrics tested in our study:

- cosine:

$$1 - \frac{u \cdot v}{||u||_2 ||v||_2} \tag{1}$$

- correlation:

$$1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{||(u - \bar{u})||_2 ||(v - \bar{v})||_2} \tag{2}$$

- Euclidean

$$||u - v||_2 \tag{3}$$

- Canberra:

$$\sum_i \frac{|u_i - v_i|}{|u_i| + |v_i|}. \tag{4}$$

An important limitation of the cosine and correlation distances is that they cannot be evaluated if one of the vectors only contain zeros. This means that if a genome does not

share any k-mer with all the other genomes, the two metrics will fail with an undefined behavior due to the division by zero (from $\|u\|_2$ or $\|v\|_2$). This may also happen when we filter the comparison with a functional data set and there is one genome that doesn't harbor any k-mer from it. The two other metrics (Euclidean and Canberra) are robust to those outliers without shared k-mers but their results are still influenced by them. Hence, species with a large proportion of genomes containing no k-mer of the filtering data set should not be interpreted with this methodology. Undefined distances with cosine and correlation metrics were set to zero in our experiments. For this reason, in the manuscript, figures showing cophenetic distance of filtered data sets used the Canberra distance.

The matrix computation in Ray Surveyor uses the SciPy python package (Jones et al. 2001). Computation of distance metrics can also be performed with R software. Moreover, the Ray Surveyor scripts allow computation of a Newick tree from the distance matrix either with the Neighbor-Joining or UPGMA method (unweighted pair group method with arithmetic mean) based on the scikit-bio and BioPython packages (Cock et al. 2009).

Phenetic and Phylogenetic Analysis

Simulated Data Sets

Simulated trees with three different average branch lengths (0.001, 0.005, 0.01) were randomly produced to represent different evolutionary distances of 100 genomes (Kuhner and Felsenstein 1994; Guindon and Gascuel 2002). For each of the three average branch lengths, we generated 10 trees to evaluate reproducibility. Sequence alignments of 1,000,000 sites were derived from the 100 genomes' trees based on a simple nucleotide model (equal equilibrium frequencies and equal mutation rates) from the Pyvolve python package (Spielman and Wilke 2015). The sequences obtained from the gapless alignments were used for subsequent Ray Surveyor analyses. The four distance metrics (Euclidean, cosine, correlation, Canberra) were tested in our simulation to transform Ray Surveyor's similarity matrix into a distance matrix. We also tested ten different k-mer lengths—ranging from 11 to 101 with an increment of 10—to evaluate their performance. To ensure the validity of our tree and sequence models, an alignment-based phylogeny with the FastTree NT-GTR model (Price et al. 2010) was made for all the trees. The alignment-based phylogenies were also compared with the reference phylogeny using the same methods as for Ray Surveyor clusters (phenetic trees) or Neighbor-Joining trees. Two evaluations were made to test how well our method would replicate the reference simulated trees. First, the simulated tree distance matrices were compared with Ray Surveyor's distance matrices with the CCC using the ape (Paradis et al. 2004) and dendextend (Galili 2015) R packages. CCC indicates how similar the pairwise distances are between two dendrograms obtained by hierarchical clustering from the distance matrix. Secondly, the topology of the trees was compared with the RF metric with the ETE3 python package (Huerta-Cepas et al. 2016). RF counts the minimal number of branch operations required to change one tree into the other. The ANI was also computed for all the simulated alignment sequences. The ANI statistics for all the trees are reported in supplementary table 1, Supplementary Material online.

Real Prokaryotic Genome Data Sets

The phylogenies and metadata for the Croucher et al. *S. pneumoniae* data set and the Kos et al. *P. aeruginosa* data set were obtained from the authors (Croucher et al. 2013, 2015; Kos et al. 2015; Donati et al. 2010). The phylogeny of the Hilty et al. *S. pneumoniae* data set was obtained using 602 conserved genes aligned with MAFFT v7.221 (Katoh and Standley 2013). A maximum likelihood phylogeny was the performed on the 602 concatenated genes with RAxML version 8.1.20 (Stamatakis 2014). In order to compare phylogenetic trees with the clusters of Ray Surveyor, the trees were converted from their Newick format into a cophenetic distance matrix using the R package: Ape (Paradis et al. 2004). Hierarchical clustering was performed using the UPGMA (average) method. The 2,429 bacteria genome phylogenetic tree was based on the 16S rRNA gene and taxonomical annotation was based on the established NCBI taxonomy. Initially, 2,429 bacterial genomes were obtained from NCBI (see supplementary table 2, Supplementary Material online for a list). To build the phylogeny of the bacterial tree of life, the 16S rRNA gene sequences were extracted from each genome. Then, the 2,429 16S rRNA genes were aligned using MAFFT v7.221 (Katoh and Standley 2013) and a maximum likelihood phylogeny was produced with RAxML version 8.1.20 (Stamatakis 2014). CCC and FMI were calculated with the dendextend R package (Galili 2015). Ray Surveyor was run with a k-mer length of 31 to keep a high stringency in the coloring of the graph (Boisvert et al. 2012). The 2,429 bacterial genomes similarity matrix was produced with Ray Surveyor on a computer cluster using four nodes of 48 cores with 256GB of RAM for a total compute time of <6 h.

Source of Tools and Data Sets

Ray Surveyor is freely available under the GPLv3 license at https://github.com/zorino/ray (last accessed July 19, 2017). A tutorial on how to run an analysis is available at https://github.com/zorino/raysurveyor-tutorial (last accessed July 19, 2017). The 2,429 bacterial genomes were downloaded from the NCBI GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/; last accessed July 19, 2017) in September 2015. Only the sequences marked either as a representative or a reference genome in the assembly reports were selected. The goal was to compute phylogenetic trees and clustering from a limited number of genomes that represented a broad taxonomical overview of the domain Bacteria. Since the NCBI GenBank genome database has an inherent bias towards certain taxa (Tatusova et al. 2015), such as clinically relevant pathogens, it allowed us to discard a large number of similar genomes. The total number of nucleotides analyzed in this data set was 11.4 billion with an average of 3.9 million per genome. The targeted analyses of *S. pneumoniae* and *P. aeruginosa* were extracted from the literature (Croucher et al. 2013; Kos et al. 2015) and downloaded

from NCBI GenBank or ENA. The data sets of resistance genes and mobile elements were taken from the MERGEM database (http://mergem.genome.ulaval.ca; last accessed July 19, 2017; Raymond et al. 2016a), the plasmids were taken from the NCBI Plasmids collection in June 2015, the bacteriophage from the EBI collection in June 2015 (http://www.ebi.ac.uk/genomes/phage.html; last accessed July 19, 2017) and the BGC from the MIBIG v1.0 database (Medema et al. 2015a).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Author Contributions

M.D. and F.R. performed bioinformatics analyses. M.D. and S.B. programmed the Ray Surveyor software. M.D., S.B., and F.L. designed algorithms. M.D., F.R., A.C., P.H.R., and J.C. interpreted biological results. M.D., F.R., A.C., and J.C. contributed to the preparation of the manuscript. All authors critically reviewed the manuscript.

Acknowledgments

This study was financed by the Canada Research Chair in Medical Genomics (J.C.). F.R. was supported by a Mitacs post-doctoral fellowship. M.D. was supported by the Fonds de recherche du Québec—Santé. The authors thank Pier-Luc Plante, Alexandre Drouin, Pascal Belleau, and Maurice Boissinot for their comments. Computations were performed under the auspices of Calcul Québec and Compute Canada. The operations of Compute Canada are funded by the Canada Foundation for Innovation (CFI), the National Science and Engineering Research Council (NSERC), NanoQuébec, and the Fonds Québécois de Recherche sur la Nature et les Technologies (FQRNT).

References

Allison GE, Verma NK. 2000. Serotype-converting bacteriophages and O-antigen modification in *Shigella flexneri*. *Trends Microbiol*. 8(1):17–23.

Andam CP, Hanage WP. 2015. Mechanisms of genome evolution of Streptococcus. *Infect Genet Evol*. 33:334–342.

Balvočit M, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT – how do these taxonomies compare? *BMC Genomics*. 18(Suppl 2):114.

Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. 2015. Measurably evolving pathogens in the genomic era. *Trends Ecol Evol*. 30(6):306–313.

Boc A, Diallo AB, Makarenkov V. 2012. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res*. 40:(Web Server issue):W573–W579.

Boisvert S, Laviolette F, Corbeil J. 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol*. 17(11):1519–1533.

Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol*. 13(12): R122.

Botzman M, Margalit H. 2011. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol*. 12(10): R109.

Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, et al. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 158(2):412–421.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxf, Engl)* 25(11):1422–1423.

Colombo M-L, Hanique S, Baurin SL, Bauvois C, De Vriendt K, Van Beeumen JJ, Frère J-M, Joris B. 2004. The *ybxI* gene of *Bacillus subtilis* 168 encodes a class D beta-lactamase of low activity. *Antimicrob Agents Chemother*. 48(2):484–490.

Compeau PEC, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol*. 29(11):987–991.

Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M. 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet*. 45(6):656–663.

Croucher NJ, Finkelstein JA, Pelton SI, Parkhill J, Bentley SD, Lipsitch M, Hanage WP. 2015. Population genomic datasets describing the post-vaccine evolutionary epidemiology of *Streptococcus pneumoniae*. *Sci Data*. 2:150058.

Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A. 2014. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics* 31(10):1569–1576.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol*. 2(5):414–424.

Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR, et al. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol*. 11(10): R107.

Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, Bourgault A-M, Laviolette F, Corbeil J. 2016. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*. 17(1):754.

Editor. 2011. Outbreak genomics. *Nat Biotechnol*. 29(9):769.

Federhen S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res*. 40:(Database issue): D136–D143.

Fenselau C, Havey C, Teerakulkittipong N, Swatkoski S, Laine O, Edwards N. 2008. Identification of beta-lactamase in antibiotic-resistant *Bacillus cereus* spores. *Appl Environ Microbiol*. 74(3):904–906.

Foerstner KU, von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. *EMBO Rep*. 6(12):1208–1213.

Fowlkes EB, Mallows CL. 1983. A method for comparing two hierarchical clusterings. *J Am Stat Assoc*. 78(383):553.

Galili T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics (Oxf, Engl)* 31(22):3718–3720.

Gardner SN, Slezak T, Hall BG. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* 31(17):2877–2878.

Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, et al. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 345(6202):1369–1372.

Glaeser SP, Kämpfer P. 2015. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst Appl Microbiol*. 38(4):237–245.

Guindon S, Gascuel O. 2002. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Mol Biol Evol*. 19(4):534–543.

Haubold B. 2014. Alignment-free phylogenetics and population genetics. *Brief Bioinform*. 15(3):407–418.

Hazen TH, Pan L, Gu J-D, Sobecky PA. 2010. The contribution of mobile genetic elements to the evolution and ecology of Vibrios. *FEMS Microbiol Ecol*. 74(3):485–499.

Hewitt CE. 1977. Viewing control structures as patterns of message passing. *Artif Intell*. 8(3):323–364.

Hilty M, Wüthrich D, Salter SJ, Engel H, Campbell S, Sá-Leão R, De Lencastre H, Hermans P, Sadowy E, Turner P, et al. 2014. Global

phylogenomic analysis of nonencapsulated *Streptococcus pneumoniae* reveals a deep-branching classic lineage that is distinct from multiple sporadic lineages. *Genome Biol Evol*. 6(12):3281–3294.
Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 33(6):1635–1638.
Integrative HMP (iHMP) Research Network Consortium 2014. The integrative human microbiome project: dynamic analysis of microbiome–host omics profiles during periods of human health and disease corresponding author. *Cell Host Microbe*. 16(3):276–289.
Jones E, Oliphant T, Peterson P, et al. 2001. {SciPy}: open source scientific tools for {Python}.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.
Konstantinidis KT, Ramette A, Tiedje JM. 2006. Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. *Appl Environ Microbiol*. 72(11):7286–7293.
Kos VN, Déraspe M, McLaughlin RE, Whiteaker JD, Roy PH, Alm RA, Corbeil J, Gardner H. 2015. The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrob Agents Chemother*. 59(1):427–436.
Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*. 11(3):459–468.
Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, et al. 2015. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics*. 15(2):141–161.
Larsson P, Elfsmark D, Svensson K, Wikström P, Forsman M, Brettin T, Keim P, Johansson A. 2009. Molecular evolutionary consequences of niche restriction in *Francisella tularensis*, a facultative intracellular pathogen. *PLoS Pathog*. 5(6): e1000472.
Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet*. 11(2): e1004941.
Li Y, Yan X. 2015. MSPKmerCounter: a fast and memory efficient approach for k-mer counting. *Cs.Ucsb.Edu*. p. 1–7.
Loureiro A, Torgo L, Soares C. 2004. Outlier detection using clustering methods: a data cleaning application. *Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector*.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764–770.
Materon IC, Queenan AM, Koehler TM, Bush K, Palzkill T. 2003. Biochemical characterization of beta-lactamases Bla1 and Bla2 from *Bacillus anthracis*. *Antimicrob Agents Chemother*. 47(6):2040–2042.
Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, et al. 2015a. Minimum information about a biosynthetic gene cluster. *Nat Chem Biol*. 11(9):625–631.
Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, et al. 2015b. The Minimum Information about a Biosynthetic Gene cluster (MIBiG) specification. *Nat Chem Biol*. 11(9):625–631.
Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. *Curr Opin Genet Dev*. 15(6):589–594.
Melsted P, Pritchard JK. 2011. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics*. 12(1):333.
Metcalf JA, Funkhouser-Jones LJ, Brileya K, Reysenbach A-L, Bordenstein SR. 2014. Antibacterial gene transfer across the tree of life. *eLife* 3:e04266.
Mooers H. 2000. The evolution of base composition and phylogenetic inference. *Trends Ecol Evol*. 15(9):365–369.
Nasser W, Beres SB, Olsen RJ, Dean MA, Rice KA, Long SW, Kristinsson KG, Gottfredsson M, Vuopio J, Raisanen K, et al. 2014. Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3,615 genome sequences. *Proc Natl Acad Sci U S A*. 111(17):E1768–E1776.
Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 17(1):132.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics (Oxf, Engl)* 20(2):289–290.
Pärnänen K, Karkman A, Tamminen M, Lyra C, Hultman J, Paulin L, Virta M. 2016. Evaluating the mobility potential of antibiotic resistance genes in environmental resistomes without metagenomics. *Sci Rep*. 6:35790.
Patwardhan A, Ray S, Roy A. 2014. Molecular markers in phylogenetic studies – a review. *J Phylogenet Evol Biol*. 2:131.
Pennisi E. 2008. Evolution. Building the tree of life, genome by genome. *Science (New York, N.Y.)* 320(5884):1716–1717.
Philippe H, Douady CJ. 2003. Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol*. 6(5):498–505.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*. 5(3):e9490.
Qi J, Luo H, Hao B. 2004. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res*. 32(Web Server issue):W45–W47.
Raymond F, Ouameur AA, Déraspe M, Iqbal N, Gingras H, Dridi B, Leprohon P, Plante P-L, Giroux R, Bérubé È, et al. 2016a. The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J*. 10(3):707–720.
Raymond F, Déraspe M, Boissinot M, Bergeron MG, Corbeil J. 2016b. Partial recovery of microbiomes after antibiotic treatment. *Gut Microb*. 7(5):428–434.
Reinert G, Chew D, Sun F, Waterman MS. 2009. Alignment-free sequence comparison (I): statistics and power. *J Comput Biol*. 16(12):1615–1634.
Rizk G, Lavenier D, Chikhi R. 2013. DSK: K-mer counting with very low memory usage. *Bioinformatics* 29(5):652–653.
Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci*. 53(1–2):131–147.
Rodionov DA, Gelfand MS, Mironov AA, Rakhmaninova AB. 2001. Comparative approach to analysis of regulation in complete genomes: multidrug resistance systems in gamma-proteobacteria. *J Mol Microbiol Biotechnol*. 3(2):319–324.
Romero P, Llull D, García E, Mitchell TJ, López R, Moscoso M. 2007. Isolation and characterization of a new plasmid pSpnP1 from a multidrug-resistant clone of *Streptococcus pneumoniae*. *Plasmid* 58(1):51–60.
Rossello-Mora R, Amann R. 2015. Past and future species definitions for Bacteria and Archaea. *Syst Appl Microbiol*. 38(4):209–216.
Sansinenea E, Ortiz A. 2011. Secondary metabolites of soil *Bacillus* spp. *Biotechnol Lett*. 33(8):1523–1538.
Schuch R, Fischetti VA. 2009. The secret life of the anthrax agent *Bacillus anthracis*: bacteriophage-mediated ecological adaptations. *PLoS ONE*. 4(8): e6532.
Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science (New York, N.Y.)*. 336(6077):48–51.
Siva N. 2010. 1000 genomes project. *ATLA Altern Lab Anim*. 38(6):445.
Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA. 2012. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med*. 4(148):148ra116–148ra116.
Sokal R, Rohlf F. 1962. The comparisons of dendrograms by objective methods. *Taxon* 11:33–40.
Song K, Ren J, Reinert G, Deng M, Waterman MS, Sun F. 2014. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief Bioinf*. 15(3):343–353.
Sozhamannan S, Chute MD, McAfee FD, Fouts DE, Akmal A, Galloway DR, Mateczun A, Baillie LW, Read TD. 2006. The *Bacillus anthracis* chromosome contains four conserved, excision-proficient, putative prophages. *BMC Microbiol*. 6:34.
Spielman SJ, Wilke CO. 2015. Pyvolve: a flexible python module for simulating sequences along phylogenies. *PLoS ONE*. 10(9): e0139047.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Sun Q, Lan R, Wang Y, Wang J, Wang Y, Li P, Du P, Xu J. 2013. Isolation and genomic characterization of Sfl, a serotype-converting bacteriophage of *Shigella flexneri*. *BMC Microbiol*. 13:39.

Tang F, Bossers A, Harders F, Lu C, Smith H. 2013. Comparative genomic analysis of twelve *Streptococcus suis* (pro)phages. *Genomics* 101(6):336–344.

Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. 2015. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res*. 43(7):3872.

Tu Q, Lin L. 2016. Gene content dissimilarity for subclassification of highly similar microbial strains. *BMC Genomics*. 17:647.

van den Nieuwboer M, van Hemert S, Claassen E, de Vos WM. 2016. *Lactobacillus plantarum* WCFS1 and its host interaction: a dozen years after the genome. *Microb Biotechnol*. 9(4):452–465.

Vinga S, Almeida J. 2003. Alignment-free sequence comparison – a review. *Bioinformatics (Oxf, Engl)* 19(4):513–523.

Walsh R, Thomson KL, Ware JS, Funke BH, Woodley J, McGuire KJ, Mazzarotto F, Blair E, Seller A, Taylor JC, et al. 2016. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med*. 19(2):192–203.

Wan L, Reinert G, Sun F, Waterman MS. 2010. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *J Comput Biol*. 17(11):1467–1490.

Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D. 2013. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*. 42(D1):D581–D591.

Wen J, Chan RHF, Yau SC, He RL, Yau SST. 2014. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* 546(1):25–34.

Xiong J, Déraspe M, Iqbal N, Krajden S, Chapman W, Dewar K, Roy PH. 2017. Complete genome of a pan-resistant *P. aeruginosa* isolated from a patient with respiratory failure in a Canadian Community Hospital. *Genome Announc*. 5(22):e00458–17.