CHAPTER 3

Evolutionary pan-genomics and applications

Basant K. Tiwary
Centre for Bioinformatics, Pondicherry University, Pondicherry, India

1 Introduction

The human genome was completely sequenced and assembled in the form of a reference sequence in the year 2001 [1]. The advent of next-generation sequencing methods has paved the way for the resequencing of entire populations of a particular species or a phylogenetic clade in a short span of time with minimum cost [2]. Thus, there was a paradigm shift in the concept of genome from a single reference genome to pan-genome after this technological revolution. The pan-genome represents a full set of genes in a particular species consisting of three major categories, a core genome which is present in all individuals of a species, accessory genome which is present in some individuals of a species, and singleton or unique genome restricted to one individual only (Figs. 1 and 2) [3]. The genes present in the core genome participate in the basic metabolic functions of the cell like housekeeping and conferring antibiotic resistance in bacteria. In addition, the core genome is treated as a conserved genomic unit to infer evolutionary relationships among different strains of bacteria. On the other hand, accessory genes frequently undergo gene gain/loss events and are often subjected to horizontal gene transfer to facilitate adaptations in a novel ecological niche.

The first ever concept of the pan-genome was developed by Tettelin et al. [3] during their study on a bacterial species, *Streptococcus agalactiae*. Since then, the research work on pan-genomes was extended to many prokaryotic species followed by some work on eukaryotic species. The pan-genome may also be defined as a combined analysis of a collection of genomic sequences treated as reference for particular species [4]. A pan-genome analysis can generate three types of new information; the size of the core genome, size of the accessory genome, and gene gain/loss events with addition of new samples. A successful study regarding a pan-genome is based on the quality of the reference assembly, quality of annotation, and the selection of appropriate individuals for study. In prokaryotes, the core part is associated with vertical transmission and homologous recombination whereas the variable part is related to horizontal gene transfer and site-specific recombination. Even the core part and accessory part may follow different evolutionary trajectories in a particular species. Generally, the core part provides a stable

Pan-genomics: Applications, Challenges, and Future Prospects
https://doi.org/10.1016/B978-0-12-817076-2.00003-2

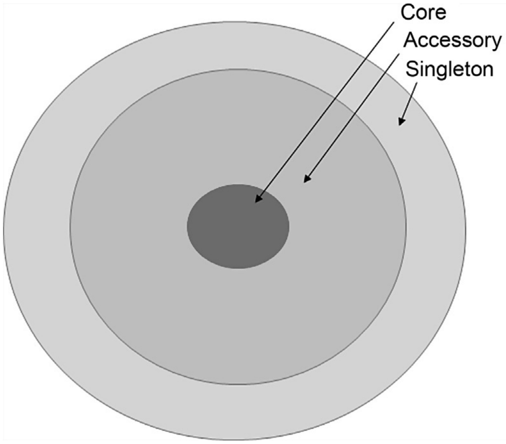

Fig. 1 A pan-genome can be classified as the core, accessory, and singleton parts.

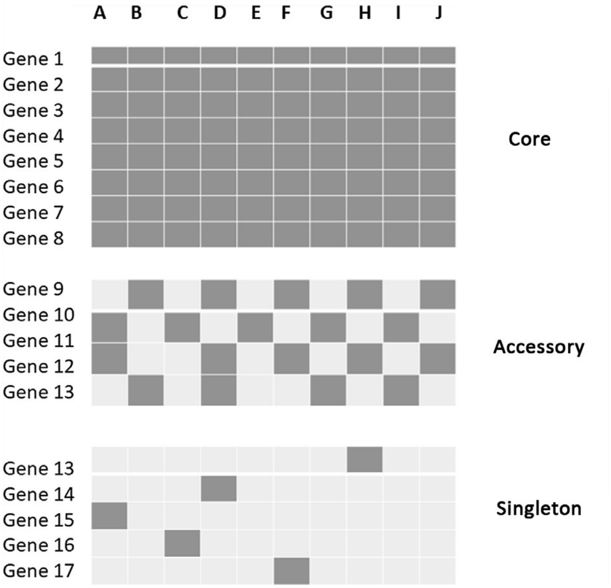

Fig. 2 Distribution of individual genes as core genes, accessory genes, and singleton genes in the pan-genome of 10 strains (A–J) of a species.

metabolic and genomic support to the species and the variable part, on the other hand, is responsible for high diversity among individuals in a population [5]. The majority of this variable part is restricted to the flexible genomic islands having size more than 10 kb [6]. Therefore, the desirable features of an ideal pan-genome are completeness (i.e., includes all functional elements), stability (i.e., unique characteristic features), comprehensibility (i.e., includes genomic information of all individuals or species), and efficiency (i.e., organized data structures) [4]. The evolutionary history of a species can be reconstructed using their genome sequences. The evolutionary signals in the genome in the form of gene content, shared marker gene or single-nucleotide polymorphisms (SNPs) across the genome may provide useful information during phylogenetic reconstruction for inferring evolutionary relationships among strains or species.

2 Computational methods in evolutionary pan-genomics

Pan-genomes are constructed from various many available resources such as the reference sequence and its variants, raw reads and haplotype reference panels. The data structure of a pan-genome is represented by a coordinate system with explicit information on all genetic variants (Fig. 3).

The simplest form of a pan-genome is a set of unaligned sequences which does not provide much useful information. A better representation of the pan-genome is multiple sequence alignment, which provides a coordinate system with many columns specifying the particular location of genes on the pan-genome [7]. However, it is only suitable for small genomic segments and does not demonstrate major genomic rearrangements like inversions and translocations. More efficiently k-mers, which are sequences with length k, provide a better representation of the pan-genome in form of de Bruijn graph (DBG) [8]. DBG is widely used as an algorithm for assembly of short reads. Further, the colored DBG suits better for the pan-genome and provides a promising method for representing the pan-genome [9]. The color of each k-mer is assigned as per the input sample in a colored DBG. A graph structure with nodes and edges can also represent a pan-genome with individual genomes as edges and coordinate system as nodes. The sequence graph may be cyclic or acyclic in nature. Even there are haplotype-centric models, where each haplotype denotes a sequence of fixed length. The positional Burrows-Wheeler Transform (PBWT) is an efficient data structure to represent a haplotype panel with compression facility [10]. Another widely used haplotype-centric model is the Li-Stephens model, which is a hidden Markov model with a matrix of states with rows indicating haplotypes and columns indicating each variant [11].

There are many popular software packages available for evolutionary pan-genome analysis (Table 1). They are primarily used for identifications of SNPs, orthologous genes, reconstruction of phylogenetic tree and profiling of different parts of pan-genome. Panseq is the first online and most popular tool for identification of core and variable parts of the genome along with SNPs associated with the core genomic region [12]. However, functional enrichment analysis to understand the functional role of each element of the genomic region is not available in this tool. The PanCGHweb is another online tool to perform pangenomic microarray analysis for the classification of orthologs and phylogenetic reconstruction among related strains [13]. The major limitation of this algorithm is not to facilitate RNA-Seq data analysis. The CAMBer can identify multigene families

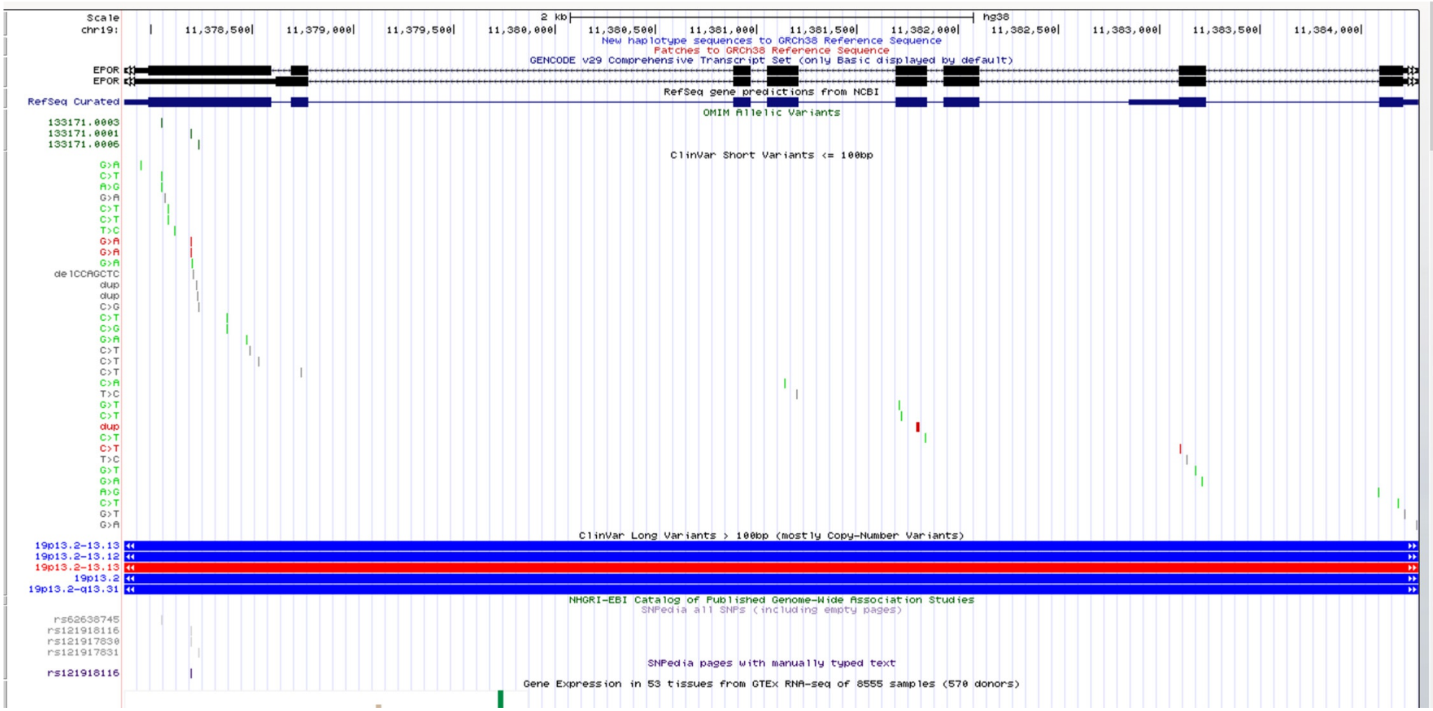
Scale
chr19:
2 kb
hg38
11,378,500
11,379,000
11,379,500
11,380,000
11,380,500
11,381,000
11,381,500
11,382,000
11,382,500
11,383,000
11,383,500
11,384,000
New haplotype sequences to GRCh38 Reference Sequence
Patches to GRCh38 Reference Sequence
GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)
EPOR
EPOR
RefSeq gene predictions from NCBI
RefSeq Curated
OMIM Allelic Variants
133171.0003
133171.0001
133171.0006
ClinVar Short Variants <= 100bp
G>A
C>T
A>G
G>A
C>T
C>T
T>C
G>A
G>A
G>A
delCCAGCTC
dup
dup
C>G
C>T
C>G
G>A
C>T
C>T
C>T
C>A
T>C
G>T
C>T
dup
C>T
C>T
T>C
G>T
G>A
A>G
C>T
G>T
G>A
ClinVar Long Variants > 100bp (mostly Copy-Number Variants)
19p13.2-13.13
19p13.2-13.12
19p13.2-13.13
19p13.2
19p13.2-q13.31
NHGRI-EBI Catalog of Published Genome-Wide Association Studies
SNPedia all SNPs (including empty pages)
rs62638745
rs121918116
rs121917830
rs121917831
SNPedia pages with manually typed text
rs121918116
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)

Fig. 3 A sequence coordinate graph generated using the UCSC browser showing genetic variants in the form of single-nucleotide polymorphism and copy number variants of the erythropoietin receptor gene in human.

Table 1 Popular software for evolutionary pangenomics

Name	Authors	Reference
Panseq	Laing et al. (2010)	[12]
PanCGHweb	Bayjanov et al. (2010)	[13]
CAMBer	Wozniak et al. (2011)	[14]
PGAT	Brittnacher et al. (2011)	[15]
PGAP	Zhao et al. (2012)	[16]
GET_HOMOLOGUES	Contreras-Moreira and Vinuesa (2013)	[17]
GET_HOMOLOGUES-EST	Contreras-Moreira et al. (2017)	[18]
PanTools	Sheikhizadeh et al. (2016)	[19]
EDGAR 2.0	Blom et al. (2016)	[20]
PanX	Ding et al. (2018)	[21]
Micropan	Snipen and Liland (2015)	[22]
FindMyFriends	Pedersen (2015)	[23]
Piggy	Thorpe et al. (2018)	[24]
PanViz	Pedersen et al. (2017)	[25]

and mutations in a variety of bacterial strains but does not provide evolutionary analysis of these strains [14]. The prokaryotic genome analysis tool (PGAT) is a web-based database tool with multiple functions for limited number of species [15]. The functions of this tool include identification of SNPs, comparison of gene order across the strains, association with the KEGG pathway and Cluster of Orthologous Groups of proteins (COG). PGAP is another package with standalone facility for creating a pan-genomic profile, and evolutionary analysis of different species along with functional enrichment of strains of a particular pan-genome [16]. GET_HOMOLOGUES is a standalone program that can perform a variety of tasks such as identification of homologues, profiling of pangenome with graphics and reconstruction of the phylogenetic tree of bacterial species [17]. An improved version of this program, GET_HOMOLOGUES-EST was developed for the evolutionary analysis of intraspecific eukaryotic pan-genomes [18]. PanTools is a java application-based tool both for prokaryotes and eukaryotes using de Bruijn graph algorithm for constructing, annotating, and grouping the homologous genes of the pan-genome [19]. The current version of the web server, EDGAR 2.0 provides very powerful phylogenetic analysis features such as average amino acid identity and average nucleotide identity among microbial genomes [20]. Recently, PanX was developed for evolutionary analysis of microbial pan-genomes with capability to display alignment, reconstruct the phylogenetic tree, infer gene gain/loss, and map mutations on the core genome [21]. Micropan is an R-package available in the R language and environment [26] for computing various properties of microbial pan-genome such as pan-genome size, openness or closeness of pan-genome, genomic fluidity, and pan-genome phylogenetic tree [22]. Another R-package FindMyFriends has a broader scope than the Micropan in the sense that it does alignment-free sequence-guided comparison following cosine similarity of k-mer vectors instead of depending on a tedious all-vs-all BLAST process [23]. Piggy detects highly divergent intergenic regions upstream of coding sequences in microbial pan-genomes [24]. PanViz is an interactive visualization tool for pangenomes written in JavaScript but can be accessed in the R environment using a package, PanVizGenerator [25].

3 Evolutionary pan-genomics of prokaryotes

Microbes are most widely studied organisms due to their small genome size and their clinical importance. An evolutionary study of the pan-genome may open up new avenues for diagnosis and therapy of microbial infections. Therefore, due to the availability of a large number of sequences of different strains of a particular microbe, a complete pan-genome of a microbial species can be created with full information regarding individual variations across strains. Microbes provide an extremely variable genome generated by point mutations in the form of SNPs and subsequently fixed in the population under the influence of evolutionary forces such as natural selection and genetic drift. The pan-genomic studies on various microbes have been conducted and core genome size varies widely across bacterial species (Table 2) [27–44]. The highest core genome size in the terms of number genes (3972) was observed in the pathogen for anthrax (*Bacillus anthracis*) whereas the minimum core genome (746) was found in *Gardnerella vaginalis*.

Majority of bacterial species have demonstrated an open pan-genome that needs a large number of additional genomes to further expand the pan-genome of the species. For example, the *E. coli* genome is an open genome and expanding further with the discovery of a new strain. On the other hand, the pan-genome of a species is fully saturated and characterized in the closed pan-genome. *Bacillus anthracis* is the best example of a closed genome because it became fully saturated after the sequencing of the first four genomes. The Heaps law model provides a metric called the alpha parameter to measure the openness or closeness of a pan-genome [43]. The alpha value is always more than 1 in the case of a closed pan-genome but it is less than 1 for the open pan-genome. Horizontal gene transfer (HGT) is another vital evolutionary force in microbial evolution for adaptation to ever-changing environments. HGT is a predominant force of microbial evolution supplemented by a lesser contribution of gene duplication in the evolutionary process [45]. Considering the fast pace of sequencing of microbial genomes, the size of the accessory genome is expanding with increasing number of samples whereas the size of the core genome is concomitantly shrinking with more number of sequenced samples. McInerney et al. opined that the effective population size and tendency to occupy novel ecological niches are two major factors regulating the pan-genome size in microbes [45].

Table 2 Pan-genomic features of bacterial species

Species	Core genome size (No. of genes)	Authors	Reference
Streptococcus agalactiae	1806	Tettelin et al. (2005)	[3]
Streptococcus pyogenes	1376	Lefebure et al. (2007)	[27]
Haemophilus influenzae	1450	Hogg et al. (2007)	[28]
Streptococcus pneumoniae	1400	Hiller et al. (2007)	[29]
Escherichia coli	2344	Rasko et al. (2008)	[30]
Neisseria meningitidis	1337	Schoen et al. (2008)	[31]
Enterococcus faecium	2172	van Schaik et al. (2010)	[32]
Yersinia pestis	3668	Eppinger et al. (2010)	[33]
Clostridium difficile	1033	Scaria et al. (2010)	[34]
Lactobacillus casei	1715	Broadbent et al. (2012)	[35]
Gardnerella vaginalis	746	Ahmed et al. (2012)	[36]
Borrelia burgdoferi	1200	Mongodin et al. (2013)	[37]
Lactobacillus paracasei	1800	Smokvina et al. (2013)	[38]
Campylobacter jejuni	1042	Meric et al. (2014)	[39]
Campylobacter coli	947	Meric et al. (2014)	[39]
Moritella viscosa	3737	Karlsen et al. (2017)	[44]
Pseudoalteromonas	1571	Bosi et al. (2017)	[40]
Bacillus amyloliquefaciens	2870	Kim et al. (2017)	[41]
Bacillus anthracis	3972	Kim et al. (2017)	[41]
Bacillus cereus	1656	Kim et al. (2017)	[41]
Bacillus subtilis	1022	Kim et al. (2017)	[41]
Bacillus thuringiensis	2299	Kim et al. (2017)	[41]
Lactobacillus plantarum	2144	Inglin et al. (2018)	[42]

4 Evolutionary pan-genomics of eukaryotes

The evolution of the eukaryote pan-genome is different from prokaryotes due to the fact that gene duplication is a predominant process in eukaryotes in contrast to HGT in prokaryotes. The genomic variations in eukaryotes are manifested in the form of SNPs, copy number variants (CNVs) (i.e., variable number of copies of a sequence in individuals), and presence or absence of variants (PAVs) (i.e., presence or absence of a sequence in individuals). Pan-genome studies on crop plants using quantitative trait loci (QTL), genome-wide association mapping and phylogenetic analysis may decipher the SNPs associated with crop productivity. Most of the genomic SNPs are not selected by natural selection and fixed by random genetic drift in the population and thus are selectively neutral. The presence of a nonsynonymous SNP changes the encoded amino acid and thereby alters the overall protein structure and function. On the other hand, a synonymous SNP does not change the encoded amino acid and contributes in retaining the

Table 3 Pan-genomic features in eukaryotes

Species	Core genome size (No. of genes)	Authors	Reference
Zymoseptoria tritici	9149	Plissonneau et al. (2018)	[48]
Glycine soja	28712	Li et al. (2014)	[49]
Oryza sativa	23914	Sun et al. (2017)	[47]
Brassica oleracia	49895	Golicz et al. (2016)	[47]

overall stability of native protein structure. The availability of a pan-genome instead of a single reference sequence may improve the efficiency of SNP discovery in crops. Further, it will discriminate SNPs located in the core and variable regions of the pan-genome. A phylogenetic study on the variable and conserved sites in different individuals of a pan-genome will provide an insight into evolutionary trend in a population. The concept of molecular clock can be implicated using SNPs to estimate the divergence time of species. Pan-genome-based phylogenetic studies have been performed in the few species of plants (Table 3) [45–48]. The crop plant *Brassica oleracia* has the maximum size of the core genome (49,895) among eukaryotes studies till date but the core genome size is comparatively smaller for a wheat plant fungal pathogen (*Zymoseptoria tritici*). More pan-genomic studies are expected for a new species of crops and their pathogen in the near future.

5 Orthology prediction and genomic plasticity in pan-genomics

Orthologous gene detection is a prerequisite evolutionary method to create the pan-genome of a species. It is useful in inferring phylogenetic trees, annotating a genome, and predicting the function of a gene. Orthologous genes are homologous genes derived from a common ancestor through the speciation process whereas paralogous genes are products of gene duplication events [49]. Orthologous genes have a common biological function but paralogous genes tend to have distinct biological functions even within a particular species. As per ortholog conjecture, orthologs are likely to have closely related function due to constant selection pressure unlike paralogs [50]. The orthology detection methods can be benchmarked using some functional similarity measures such as conservation of a protein domain or coexpression levels of genes [51]. A web-based facility is also developed to benchmark all available orthology detection tools on a large-scale basis [52]. There are several computational methods for orthologous gene detection using both graph-based and tree-based approaches. Graph-based methods heuristically search a sequence similarity score for a large number of sequences. OrthoMCL is the most popular algorithm among graph-based methods for the automated classification of eukaryotic orthologous groups [53]. First, it constructs a similarity score matrix in the form of a graph with protein sequences as nodes and relationship among protein sequences as edges.

Several subgraphs representing orthologous clusters are created from this graph using the Markov clustering algorithm (MCL). The MCL algorithm simulates random walks on a graph using Markov matrices to obtain transition probabilities among the nodes [54]. Although this algorithm is computationally efficient, it does not consider evolutionary information available on the sequences. Thus, orthology detection using this algorithm is prone to error in clustering, especially when there is a differential gene loss in the lineages under study [55]. Tree-based method is a better approach of orthology prediction, which looks for congruency between the gene tree and the species tree to infer orthologs and paralogs [56, 57]. First, a gene phylogeny is reconstructed from multiple sequence alignment of a certain gene and a particular gene phylogeny is then compared to overall species phylogeny using maximum parsimony in order to distinguish speciation and duplication processes [58]. The maximum parsimony is based on the notion that the evolutionary path showing the minimum number of mutations is the most probable path of evolution. Tree-based is although based on a powerful evolutionary concept of maximum parsimony but suffers from two disadvantages; the species phylogeny of many species is not yet resolved and large-scale phylogenetic analysis is not possible due high computational cost of this approach. However, there are some hybrid methods such as Ortholuge [59], EnsemblCompara [60], and HomoloGene [61], etc. combining the merits of both graph-based and tree-based methods.

A microbial genome can be visualized as a dynamic entity undergoing recurrent gene gain and loss processes. The genomic plasticity in a microbial species is the result of various events in which horizontal gene transfer is of primary importance [62]. Horizontal gene transfer facilitates in acquiring blocks of genes known as genomic islands in a species resulting in accelerated rate of evolution. The core genes in a microbe represent the conserved nature of evolution under high selective constraints. In fact, Koonin has advocated that these core genes provide a strong backbone structure for remaining part of the genome [63]. Although character genes constitute a major part of the bacterial genome (64%), the number of gene families represented by them is very small (~7900) [64]. However, these genes are flexible enough to adapt to novel functions in a short span of time. Although these genes show similarity at the sequence level but exhibit great diversity in specificity to different substrates. Thus, it appears that nature does not opt for creating a new gene *de novo* whenever necessity arises. Instead, new biological solutions are obtained from the existing number of gene families, although limited in number, through two evolutionary processes: gene mutations and gene duplications [65–68]. For example, ABC transporters exhibit some wide substrate specificities due to gene substitutions. In contrast, accessory genes are not strongly linked to any particular lineage and are not highly conserved unlike core genome. They are also not subjected to strong evolutionary pressure unlike core genes [69] and have high turnover rates in microbial genomes [70]. The majority of accessory genes are involved in the process of gene creation, generally leading to loss of a gene from the genome. Rarely do they get adaptive advantage during the gene creation process and ultimately transformed as a character gene in the genome.

6 Phylogenomics and genomic epidemiology in pan-genomics

A phylogenetic tree based on the genome (Phylogenomics) is reconstructed using a set of genes in the genome rather than a single gene. A species or genus can be characterized based on a pan-genomic study on all available strains. This diversity within a genome across different strains can be visualized in the form of a tree. There are two major approaches, namely sequence based and gene content based for reconstructing phylogenomic trees using whole genome data [71]. In a sequence-based tree approach, we first align the sequences using multiple sequence alignment and a phylogenomic tree is reconstructed based on evolutionary distances. On the other hand, we use binary data of presence and absence of a gene in different genomes in a gene content-based tree and then a phylogenomic tree is reconstructed using a derived distance matrix from the data. Two types of distances between pan-genome profiles are commonly used in the pan-genomic tree reconstruction: Manhattan distance and Jaccard distance. Manhattan distance is defined as the sum of the differences between each element of two genomes. Jaccard distance between two genomes, on the other hand, measures the degree of similarity between two genomes in each element with respect to the presence or absence of a gene cluster. Genomic fluidity is another measure of a similar kind but it computes the population diversity of the whole population by taking the average of each pair [72]. A pan-genomic tree can be reconstructed based on hierarchical clustering using distance-based UPGMA or neighbor joining methods on these distances (Fig. 4). Such a tree will demonstrate the differences in gene content between genomes. Different gene family weights are necessary for core, accessory, and singleton genes due to wide variation in the degree of their conservation. For example, core genes are highly conserved across the pan-genome and provide no signal for differences between genomes. Therefore, zero weights are assigned to the core genes. Similarly, genes present in a single genome (singleton or ORFans) are often doubtful and therefore, given zero weights as well. The R package micropan is commonly used for reconstructing the pan-tree from the central genome after partitioning into the medoide genome [22]. The bcgTree is an automatic pipeline for reconstruction of the pan-tree both from genomic databases or in-house generated sequences in the laboratory [73]. It retrieved automatically 107 single copy bacterial core genes using hidden Markov models and subsequently reconstructed a pan-tree using partitioned maximum likelihood analysis.

Genome-based molecular epidemiology or genomic epidemiology is a powerful tool of public health investigations of bacterial infections [74]. Alternatively, different subtypes of pathogenic bacteria were identified using some common laboratory techniques like pulse-field gel electrophoresis and multi-sequence typing. These techniques are

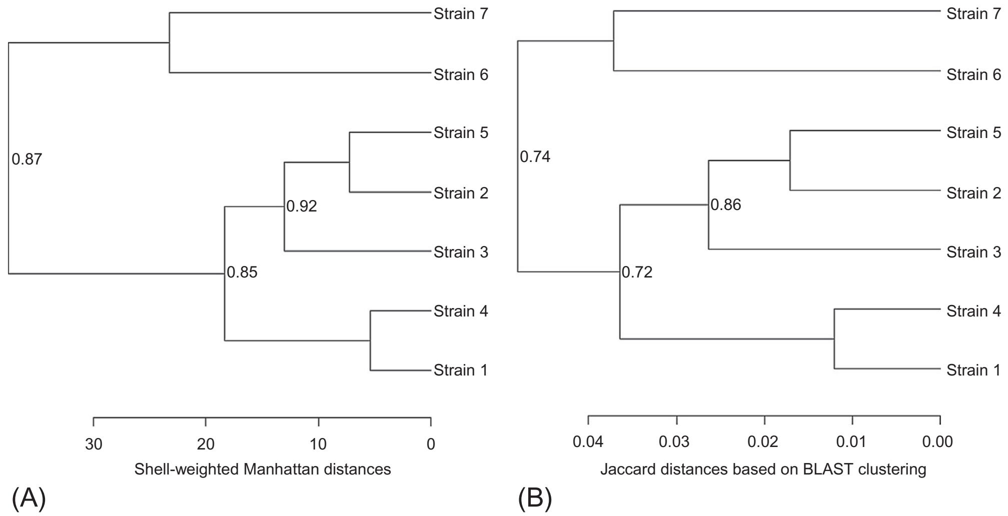

Fig. 4 A pan-tree showing evolutionary relationship between seven strains of a bacterial species based on Manhattan distances (A) and Jaccard distances (B). The values at the node indicate bootstrap values for each clade.

although tedious and time consuming generate limited genetic information regarding the pathogen. However, next-generation whole genome sequencing methods can uncover all SNPs spanning the genome present in different strains of a pathogen within a short span of time. Different strains of *Legionella* were classified into outbreak and non-outbreak groups based whole genome sequencing [75]. It was found that the persistence and virulence of *Legionella pneumophila* were encoded by the core genes [76]. However, some pathogens such as *Yersinia pestis* and *Bacillus anthracis* are found in the soil in dormant state and becomes active and proliferates only in the host. Thus, they do not get an opportunity to exchange genes, and therefore have a closed genome. In fact, the core/pan-genome ratio reaches to an extreme value of 99% in the *B. anthracis* [77]. Therefore, pan-genomic study on a pathogen in an environmental sample will reveal the genomic details of different strains of a pathogen and thereby further help us controlling the outbreak of any epidemic disease.

7 Future directions

There are successful examples of pan-genomic evolutionary studies in various species of prokaryotes and eukaryotes. Concomitantly, appropriate data structures and suitable computational algorithms are being developed for better data analysis of the pan-genome across genera and species. However, there is an urgent need to develop qualitatively better data structure and new computational methods to analyze the fast expanding pan-genomic data. Another major challenge in this area is a better annotation of the pan-genome with relevant functional and phenotypic information. Biochemical modifications on the sequences such as hyper-methylated regions will be a useful additional feature of future pan-genomes. Some additional features like SNPs, non-coding RNA, and indels need special attention in future. There is also a significant development of orthology prediction methods till date. A statistically robust method is needed to discriminate the orthologs and the paralogs with minimal false positives. The evolutionary mechanism regulating genomic plasticity is not yet clear and needs further investigation. Distance-based phylogenomic analysis is a powerful tool to infer evolutionary relationship between different taxa. Character-based methods need more emphasis in their implementation in phylogenomic analysis of the pan-genome for better results.

8 Conclusion

In summary, the emergence of evolutionary pan-genomics is a major advance in understanding the diversity of genomes and inferring the full picture of their variability. I expect that with the development of new computational tools and techniques, we will have some better insights into the regulatory mechanisms generating and governing biodiversity in nature under multiple evolutionary forces in action. Future evolutionary studies are all poised to be focussed on the ever expanding pan-genome instead of a single genome sequencing representing a taxon.

References

[1] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.

[2] H.P.J. Buermans, J.T. den Dunnen, Next generation sequencing technology: advances and applications, Biochim. Biophys. Acta 1842 (2014) 1932–1941.

[3] H. Tettelin, V. Masignani, M.J. Cieslewicz, C. Donati, D. Medini, N.L. Ward, S.V. Angiuoli, J. Crabtree, A.L. Jones, A.S. Durkin, et al., Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pangenome”, Proc. Natl. Acad. Sci. U.S.A. 102 (39) (2005) 13950–13955.

[4] The Computational Pan-Genomics Consortium, Computational pan-genomics: status, promises and challenges, Brief. Bioinform. 19 (1) (2016) 118–135.

[5] F. Rodriguez-Valera, D.W. Ussery, Is the pan-genome also a pan-selectome? F1000Res. 1 (2012) 16.

[6] M. López-Pérez, F. Rodriguez-Valera, Pangenome evolution in the marine bacterium Alteromonas, Genome Biol. Evol. 8 (5) (2016) 1556–1570.

[7] C. Notredame, Recent evolutions of multiple sequence alignment algorithms, PLoS Comput. Biol. 3 (8) (2007) e123.

[8] J.R. Miller, S. Koren, G. Sutton, Assembly algorithms for next generation sequencing data, Genomics 95 (6) (2010) 315–327.

[9] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, G. McVean, De novo assembly and genotyping of variants using colored de Bruijn graphs, Nat. Genet. 44 (2) (2012) 226–232.

[10] R. Durbin, Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT), Bioinformatics 30 (9) (2014) 1266–1272.

[11] N. Li, M. Stephens, Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data, Genetics 165 (4) (2003) 2213–2233.

[12] C. Laing, C. Buchanan, E.N. Taboada, Y.X. Zhang, A. Kropinski, A. Villegas, J.E. Thomas, V. P. Gannon, Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions, BMC Bioinform. 11 (2010) 461.

[13] J.R. Bayjanov, R.J. Siezen, S.A. van Hijum, PanCGHweb: a web tool for genotype calling in pangenome CGH data, Bioinformatics 26 (9) (2010) 1256–1257.

[14] M. Wozniak, L. Wong, J. Tiuryn, CAMBer: an approach to support comparative analysis of multiple bacterial strains, BMC Genomics 12 (2011) S6.

[15] M.J. Brittnacher, C. Fong, H.S. Hayden, M.A. Jacobs, M. Radey, L. Rohmer, PGAT: a multistrain analysis resource for microbial genomes, Bioinformatics 27 (17) (2011) 2429–2430.

[16] Y. Zhao, J. Wu, J. Yang, S. Sun, J. Xiao, J. Yu, PGAP: pan-genomes analysis pipeline, Bioinformatics 28 (3) (2012) 416–418.

[17] B. Contreras-Moreira, P. Vinuesa, GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis, Appl. Environ. Microbiol. 79 (24) (2013) 7696–7701.

[18] B. Contreras-Moreira, C.P. Cantalapiedra, M.J. García-Pereira, S.P. Gordon, J.P. Vogel, E. Igartua, A.M. Casas, P. Vinuesa, Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species, Front. Plant Sci. (2017), https://doi.org/10.3389/fpls.2017.00184.

[19] S. Sheikhizadeh, M.E. Schranz, M. Akdel, D. De Ridder, S. Smit, PanTools: representation, storage and exploration of pan-genomic data, Bioinformatics 32 (17) (2016) i487–i493.

[20] J. Blom, J. Kreis, S. Spänig, T. Juhre, C. Bertelli, C. Ernst, A. Goesmann, EDGAR 2.0: an enhanced software platform for comparative gene content analyses, Nucleic Acids Res. 44 (W1) (2016) W22–W28.

[21] W. Ding, F. Baumdicker, R.A. Neher, panX: pan-genome analysis and exploration, Nucleic Acids Res. 46 (1) (2018) e5.

[22] L. Snipen, K.H. Liland, micropan: an R-package for microbial pan-genomics, BMC Bioinform. 16 (2015) 79.

[23] T.L. Pedersen, FindMyFriends: Microbial Comparative Genomics in R, R package version 1.12.0, http://bioconductor.org/packages/FindMyFriends, 2015.

[24] H.A. Thorpe, S.C. Bayliss, S.K. Sheppard, E.J. Feil, Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria, Gigascience 7 (4) (2018) 1–11.

[25] T.L. Pedersen, I. Nookaew, D.W. Ussery, M. Månsson, PanViz: interactive visualization of the structure of functionally annotated pangenomes, Bioinformatics 33 (7) (2017) 1081–1082.

[26] R Core Team, R: A Language and Environment for Statistical Computing, version 3.5, second ed., R Foundation for Statistical Computing, Vienna, Austria, 2018.

[27] T. Lefebure, M.J. Stanhope, Evolution of the core and pangenome of Streptococcus: positive selection, recombination, and genome composition, Genome Biol. (5) (2007) R71.

[28] J.S. Hogg, F.Z. Hu, B. Janto, R. Boissy, J. Hayes, R. Keefe, J.C. Post, G.D. Ehrlich, Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains, Genome Biol. 8 (6) (2007) R103.

[29] N.L. Hiller, B. Janto, J.S. Hogg, R. Boissy, S. Yu, E. Powell, R. Keefe, N.E. Ehrlich, K. Shen, J. Hayes, et al., Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains:insights into the pneumococcal supragenome, J. Bacteriol. 189 (22) (2007) 8186–8195.

[30] D.A. Rasko, M.J. Rosovitz, G.S. Myers, E.F. Mongodin, W.F. Fricke, P. Gajer, J. Crabtree, M. Sebaihia, N.R. Thomson, R. Chaudhuri, et al., The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates, J. Bacteriol. 190 (20) (2008) 6881–6893.

[31] C. Schoen, J. Blom, H. Claus, A. Schramm-Gluck, P. Brandt, T. Muller, A. Goesmann, B. Joseph, S. Konietzny, O. Kurzai, et al., Whole genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*, Proc. Natl. Acad. Sci. U.S.A. 105 (9) (2008) 3473–3478.

[32] W. van Schaik, J. Top, D.R. Riley, J. Boekhorst, J.E. Vrijenhoek, C.M. Schapendonk, A. P. Hendrickx, I.J. Nijman, M.J. Bonten, H. Tettelin, et al., Pyrosequencing-based comparative

genome analysis of the nosocomial pathogen *Enterococcus faecium* and identification of a large transferable pathogenicity island, BMC Genomics 11 (2010) 239.

[33] M. Eppinger, P.L. Worsham, M.P. Nikolich, D.R. Riley, Y. Sebastian, S. Mou, M. Achtman, L. E. Lindler, J. Ravel, Genome sequence of the deep-rooted *Yersinia pestis* strain Angola reveals new insights into the evolution and pangenome of the plague bacterium, J. Bacteriol. 192 (6) (2010) 1685–1699.

[34] J. Scaria, L. Ponnala, T. Janvilisri, W. Yan, L.A. Mueller, Y.F. Chang, Analysis of ultra low genome conservation in *Clostridium difficile*, PLoS One 5 (12) (2010).

[35] J.R. Broadbent, E.C. Neeno-Eckwall, B. Stahl, K. Tandee, H. Cai, W. Morovic, P. Horvath, J. Heidenreich, N.T. Perna, R. Barrangou, et al., Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation, BMC Genomics 13 (2012) 533.

[36] A. Ahmed, J. Earl, A. Retchless, S.L. Hillier, L.K. Rabe, T.L. Cherpes, E. Powell, B. Janto, R. Eutsey, N.L. Hiller, et al., Comparative genomic analyses of 17 clinical isolates of *Gardnerella vaginalis* provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars, J. Bacteriol. 194 (15) (2012) 3922–3939.

[37] E.F. Mongodin, S.R. Casjens, J.F. Bruno, Y. Xu, E.F. Drabek, D.R. Riley, B.L. Cantarel, P. E. Pagan, Y.A. Hernandez, L.C. Vargas, et al., Inter- and intra-specific pan-genomes of *Borrelia burgdorferi sensu lato*: genome stability and adaptive radiation, BMC Genomics 14 (2013) 693.

[38] T. Smokvina, M. Wels, J. Polka, C. Chervaux, S. Brisse, J. Boekhorst, J.E. van Hylckama Vlieg, R. J. Siezen, *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity, PLoS One 8 (7) (2013).

[39] G. Meric, K. Yahara, L. Mageiros, B. Pascoe, M.C. Maiden, K.A. Jolley, S.K. Sheppard, A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic campylobacter, PLoS One 9 (3) (2014).

[40] E. Bosi, M. Fondi, V. Orlandini, E. Perrin, I. Maida, D. de Pascale, M.L. Tutino, E. Parrilli, A. Lo Giudice, A. Filloux, R. Fani, The pangenome of (Antarctic) Pseudoalteromonas bacteria: evolutionary and functional insights, BMC Genomics 18 (2017) 93.

[41] Y. Kim, I. Koh, L.M. Young, W.H. Chung, M. Rho, Pan-genome analysis of Bacillus for microbiome profiling, Sci. Rep. 7 (1) (2017).

[42] R.C. Inglin, L. Meile, M.J.A. Stevens, Clustering of pan- and core-genome of lactobacillus provides novel evolutionary insights for differentiation, BMC Genomics 19 (1) (2018) 284.

[43] H. Tettelin, D. Riley, C. Cattuto, D. Medini, Comparative genomics: the bacterial pan-genome, Curr. Opin. Microbiol. 12 (2008) 472–477.

[44] C.R. Karlsen, E. Hjerde, T. Klemetsen, N.P. Willassen, Pan genome and CRISPR analyses of the bacterial fish pathogen *Moritella viscosa*, BMC Genomics 18 (2017) 313.

[45] J.O. McInerney, A. McNally, M.J. O'Connell, Why prokaryotes have pangenomes, Nat. Microbiol. 2 (2017) 17040.

[46] C. Sun, Z. Hu, T. Zheng, K. Lu, Y. Zhao, W. Wang, J. Shi, C. Wang, J. Lu, D. Zhang, Z. Li, C. Wei, RPAN: rice pan-genome browser for ∼3000 rice genomes, Nucleic Acids Res. 45 (2) (2017) 597–605.

[47] A.A. Golicz, P.E. Bayer, G.C. Barker, P.P. Edger, H. Kim, P.A. Martinez, C.K. Chan, A. Severn-Ellis, W.R. McCombie, I.A. Parkin, A.H. Paterson, J.C. Pires, A.G. Sharpe, H. Tang, G. R. Teakle, C.D. Town, J. Batley, D. Edwards, The pangenome of an agronomically important crop plant *Brassica oleracea*, Nat. Commun. 7 (2016).

[48] C. Plissonneau, F.E. Hartmann, D. Croll, Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome, BMC Biol. 16 (1) (2018) 5.

[49] Y.H. Li, G. Zhou, J. Ma, W. Jiang, L.G. Jin, Z. Zhang, Y. Guo, J. Zhang, Y. Sui, L. Zheng, S.S. Zhang, Q. Zuo, X.H. Shi, Y.F. Li, W.K. Zhang, Y. Hu, G. Kong, H.L. Hong, B. Tan, J. Song, Z.X. Liu, Y. Wang, H. Ruan, C.K. Yeung, J. Liu, H. Wang, L.J. Zhang, R.X. Guan, K.J. Wang, W.B. Li, S.Y. Chen, R.Z. Chang, Z. Jiang, S.A. Jackson, R. Li, L.J. Qiu, De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits, Nat. Biotechnol. 32 (10) (2014) 1045–1052.

[50] E.V. Koonin, Orthologs, paralogs, and evolutionary genomics, Annu. Rev. Genet. 39 (2005) 309–338.

[51] A.M. Altenhoff, R.A. Studer, M. Robinson-Rechavi, C. Dessimoz, Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs, PLoS Comput. Biol. 8 (2012).

[52] T. Hulsen, M.A. Huynen, J. de Vlieg, P.M. Groenen, Benchmarking ortholog identification methods using functional genomics data, Genome Biol. 7 (2006) R31.

[53] A. Altenhoff, B. Boeckmann, S. Capella-Gutierrez, D.A. Dalquen, T. DeLuca, K. Forslund, J. Huerta-Cepas, B. Linard, C. Pereira, L.P. Pryszcz, et al., Standardized benchmarking in the quest for orthologs, Nat. Methods 13 (2016) 425–430.

[54] L. Li, C.J. Stoeckert, D.S. Roos, Orthomcl: identification of ortholog groups for eukaryotic genomes, Genome Res. 13 (9) (2003) 2178–2189.

[55] A.J. Enright, S.V. Dongen, C.A. Ouzounis, An efficient algorithm for large-scale detection of protein families, Nucleic Acids Res. 30 (7) (2002) 1575–1584.

[56] D.R. Scannell, K.P. Byrne, J.L. Gordon, S. Wong, K.H. Wolfe, Multiple rounds of speciation associated with reciprocal gene loss in polyploidy yeasts, Nature 440 (7082) (2006) 341–345.

[57] B. Mirkin, I. Muchnik, T.F. Smith, A biologically consistent model for comparing molecular phylogenies, J. Comput. Biol. 2 (4) (1995) 493–507.

[58] R.D.M. Page, M.A. Charleston, From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem, Mol. Phylogenet. Evol. 7 (2) (1997) 231–240.

[59] M. Goodman, J. Czelusniak, G.W. Moore, A.E. Romero-Herrera, G. Matsuda, Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences, Syst. Biol. 28 (2) (1979) 132–163.

[60] D.L. Fulton, Y.Y. Li, M.R. Laird, B.G.S. Horsman, F.M. Roche, F.S.L. Brinkman, Improving the specificity of high-throughput ortholog prediction, BMC Bioinform. 7 (1) (2006) 270.

[61] A.J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, E. Birney, EnsemblcomparaGeneTrees: complete, duplication-aware phylogenetic trees in vertebrates, Genome Res. 19 (2) (2009) 327–335.

[62] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, et al., Database resources of the national center for biotechnology information, Nucleic Acids Res. 36 (Suppl 1) (2007) D13–D21.

[63] H. Schmidt, M. Hensel, Pathogenicity islands in bacterial pathogenesis, Clin. Microbiol. Rev. 17 (2004) 14–56.

[64] E.V. Koonin, Comparative genomics, minimal gene-sets and the last universal common ancestor, Nat. Rev. Microbiol. 1 (2003) 127–136.

[65] P. Lapierre, J.P. Gogarten, Estimating the size of the bacterial pan-genome, Trends Genet. 25 (3) (2009) 107–110.

[66] A.L. Davidson, J. Chen, ATP-binding cassette transporters in bacteria, Annu. Rev. Biochem. 73 (2004) 241–268.

[67] D.M. Nanavati, T.N. Nguyen, K.M. Noll, Substrate specificities and expression patterns reflect the evolutionary divergence of maltose ABC transporters in *Thermotoga maritima*, J. Bacteriol. 187 (6) (2005) 2002–2009.

[68] K. Fukami-Kobayashi, Y. Tateno, K. Nishikawa, Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins, Mol. Biol. Evol. 20 (2003) 267–277.

[69] V. Daubin, H. Ochman, Start-up entities in the origin of new genes, Curr. Opin. Genet. Dev. 14 (2004) 616–619.

[70] J.P. Gogarten, J.P. Townsend, Horizontal gene transfer, genome innovation and evolution, Nat. Rev. Microbiol. 3 (2005) 679–687.

[71] J.G. Lawrence, H. Ochman, Amelioration of bacterial genomes: rates of change and exchange, J. Mol. Evol. 44 (1997) 383–397.

[72] A.O. Kislyuk, B. Haegeman, N.H. Bergman, J.S. Weitz, Genomic fluidity: an integrative view of gene diversity within microbial populations, BMC Genomics 12 (2011) 32.

[73] M.J. Ankenbrand, A. Keller, bcgTree: automatized phylogenetic tree building from bacterial core genomes, Genome 59 (10) (2016) 783–791.

[74] M.W. Gilmour, M. Graham, A. Reimer, G. Van Domselaar, Public health genomics and the new molecular epidemiology of bacterial pathogens, Public Health Genomics 16 (2013) 25–30.

[75] S. Reuter, T.G. Harrison, C.U. Koser, M.J. Ellington, G.P. Smith, J. Parkhill, A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak, BMJ Open 3 (2013).

[76] G. D'Auria, N. Jimenez-Hernandez, F. Peris-Bondia, A. Moya, A. Latorre, *Legionella pneumophila* pan-genome reveals strain-specific virulence factors, BMC Genomics 11 (2010) 181.

[77] L. Rouli, V. Merhej, P.E. Fournier, D. Raoult, The bacterial pangenome as a new tool for analysing pathogenic bacteria, New Microbes New Infect. 7 (2015) 72–85.

Further reading

[78] L. Snipen, D.W. Ussery, Standard operating procedure for computing pangenome trees, Stand. Genomic Sci. 2 (1) (2010) 135–141.