A Review of Pangenome Tools and Recent Studies

G. S. Vernikos

Abstract With the advance of sequencing technologies, the landscape of genomic analysis has been transformed, by moving from single strain to species (or even higher taxa)-wide genomic resolution, toward the direction of capturing the "totality" of life diversity; from this scientific advance and curiosity, the concept of "pangenome" was born. Herein we will review, from practical and technical implementation, existing projects of pangenome analysis, with the aim of providing the reader with a snapshot of useful tools should they need to embark on such a pangenomic journey.

Keywords Pangenome · Whole-genome · Exhaustive search · Subsampling · Regression function · Command line · Web-interface · Bayesian · Hidden Markov Models · Clustering · ORF alignment similarity · Combinatorial approach · Ortholog clusters · Reference pangenome · Finite supragenome model · Binomial mixture model · Infinitely many genes model · Gene presence/absence frequency

1 Introduction

Almost 15 years ago, Tettelin et al. (2005) conceived the concept of pangenome, in an attempt to describe and model the genomic totality of a taxa (species, serovar, phylum, kingdom, etc.) of interest. Since then the nomenclature of this concept became fairly wide to accommodate words like pangenome, core and dispensable genes, strain-specific genes (Medini et al. 2005; Tettelin et al. 2005), supragenome, distributed and unique genes (Lapierre and Gogarten 2009), and flexible regions (Rodriguez-Valera and Ussery 2012). Simply put, using the original definition, the core-genome describes the set of sequences shared by all members of the taxa of interest, the dispensable genome captures a subset of sequences shared by some

G. S. Vernikos (✉)
GlaxoSmithKline, Medical Affairs Department, Athens, Greece
e-mail: georgios.x.vernikos@gsk.com

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_4

members of the group (dictating the diversity of the group: alternative biochemical pathways, niche adaptation, antibiotic resistance, etc.) while the pangenome is simply the union of core and dispensable genomes (describing the totality of taxa at the level of sequence datasets).

The exponential growth of genomic databases started in 1995 with *Haemophilus influenzae* being the first complete genome project (Fleischmann et al. 1995). Today, as of August 2018, 110,660 complete whole-genome sequencing projects—of which 87% are bacteria—and 15,066 finished whole-genome sequencing projects (Mukherjee et al. 2017) are available in the public domain. These fueled the interest of many researchers to carry out pangenome analysis at every conceivable phylogenetic resolution level (Table 1), exploiting various modeling frameworks, assumptions, and underlying homology search engines.

A pivotal work in terms of phylogenetic resolution was carried out by Lapierre and Gogarten (2009), showing that on average in the largest bacterium group analyzed so far, the core gene set accounts only for 8% of the pangenome.

The pangenome concept can be implemented either in reverse or in forward-thinking approaches; in the first case, we are interested to capture the genomic diversity of the group of interest, while in the second case we are more interested in exploring and predicting from a pragmatic perspective what is the minimum number of genome sequences required to capture the totality of the group. Obviously, limited or sparse datasets might lead to erroneous conclusions; therefore, it was recommended (Vernikos et al. 2015) that the minimum number of genomes to analyze be at least five.

The lifestyle of the species of interest is one of the parameters strongly dictating the distribution shape of the pangenome; for example, if by recurring addition of group members, the pangenome continues to grow, we are analyzing an open pangenome (such examples include human pathogens and environmental bacteria) (Hiller et al. 2007; Tettelin et al. 2008). On the other hand, if the group complexity is exhausted very fast even from the analyses of a handful of group members then we are dealing with a closed pangenome whereby we only need few representatives to describe the totality of the sequence variability.

2 Technical Implementation

In pangenome analysis, the sequence unit for the modeling can be anything from ORFs, genes, clusters of orthologous groups COGs (Tatusov et al. 1997), coding sequences (CDS), proteins, arbitrary sequence chunks, concatenated gene or protein entities, etc.

Practical aspects of consideration that directly influence the validity of the conclusions drawn, include how quickly is expected a pangenome to grow and reach a plateau (open or close pangenome), the parameters that determine in the search engine the orthologous sequences and thereby directly affect the pool of core and dispensable sequence entities, the mathematical model and the applied

Table 1 Examples of the application of pangenome approaches at different levels of phylogenetic resolution

Level	Organism	Approach[a]	# Genomes	Core size (# genes)	Year (reference)
Species	*Streptococcus agalactiae*	ORFsim, Comb	8	1806	Tettelin et al. (2005)
	Neisseria meningitidis	ORFsim, Comb	6	1337	Schoen et al. (2008)
		ORFsim, Comb	20	1630	Budroni et al. (2011)
	Borrelia burgdoferi	ORFsim, Comb	21	1200	Mongodin et al. (2013)
	Escherichia coli	ORFsim, Comb	17	2344	Rasko et al. (2008)
	Enterococcus faecium	ORFsim, Comb	7	2172	van Schaik et al. (2010)
	Yersinia pestis	ORFsim, Comb	14	3668	Eppinger et al. (2010)
	Streptococcus pyogenes	OG, Comb	11	1376	Lefebure and Stanhope (2007)
	Clostridium difficile	OG, Comb	15	1033	Scaria et al. (2010)
	Lactobacillus paracasei	OG	34	1800	Smokvina et al. (2013)
	Campylobacter jejuni	ORFsim, Ref	130	1042	Meric et al. (2014)
	Campylobacter coli	ORFsim, Ref	62	947	Meric et al. (2014)
	Haemophilus influenzae	FSM	13	1450	Hogg et al. (2007)
	Streptococcus pneumoniae	FSM	17	1400	Hiller et al. (2007)
		ORFsim, Comb	44	1666	Donati et al. (2010)
	Staphylococcus aureus	FSM	16	2245	Boissy et al. (2011)
	Moraxella catarrhalis	FSM	12	1755	Davie et al. (2011)
	Lactobacillus casei	FSM	17	1715	Broadbent et al. (2012)
	Gardnerella vaginalis	FSM	17	746	Ahmed et al. (2012)
	Clostridium botulinum	ORFsim, Comb	13	2657	Bhardwaj and Somvanshi (2017)
Group	*Bacillus cereus*	ORFsim, Comb	4	3000	Lapidus et al. (2008)
	Bacillus subset of species	ORFsim, Comb	12	2009	Eppinger et al. (2011)

(continued)

Table 1 (continued)

Level	Organism	Approach[a]	# Genomes	Core size (# genes)	Year (reference)
Genus	*Streptococcus*	OG, Comb	26	600	Lefebure and Stanhope (2007)
		ORFsim, Comb	52	522	Donati et al. (2010)
	Prochlorococcus	ORFsim, Comb	12	1273	Kettler et al. (2007)
	Bifidobacterium	ORFsim, Comb	14	967	Bottacini et al. (2010)
	Listeria	BMM	13	2032	den Bakker et al. (2010)
	Salmonella	BMM	35	2811	Jacobsen et al. (2011)
	Shewanella	OG	24	1878	Zhong et al. (2018)
	Finegoldia	OG	12	1202	Brüggemann et al. (2018)
Class	Bacilli	IMGM	172	143	Collins and Higgs (2012)
Phylum	Chlamydiae	OG	19	560	Collingro et al. (2011)
Super kingdom	Eubacteria	Gene freq.	573	250	Lapierre and Gogarten (2009)

[a]*ORFsim* ORF alignment similarity, *Comb* combinatorial approach of adding successive genomes, *OG* ortholog clusters, *Ref* initial generation of a reference pangenome using a subset of strains, *FSM* finite supragenome model, *BMM* binomial mixture model, *IMGM* infinitely many genes model, *Gene freq* gene presence/absence frequency

distribution of forecasting the evolution of the pangenome and core-genome size. Another limiting factor, as the number of genomes becomes higher and higher, is the scalability of all possible genome addition permutations, since the total number of comparisons needed is described from the following function:

$$C = \frac{N!}{(n-1)! \cdot (N-n)!}$$

where C is the total number of comparisons, and N is the total number of genomes.

A workaround to an exhaustive approach is a method of subsampling (Vernikos et al. 2015) the total number of comparisons needed; comparisons are randomly selected making sure that each genome undergoes the same number of comparisons; the trick here is to set the number of possible comparisons to a number that will optimally balance the existing computational power and the target dataset size. Indeed, observations from limited in size datasets, showed that even extreme sampling is still able to model reliably the pangenome bypassing the need to follow an exhaustive all-against-all comparison (Fig. 1) (Vernikos et al. 2015). Additional optimizations can be achieved by exploiting alternative (to the original exponential

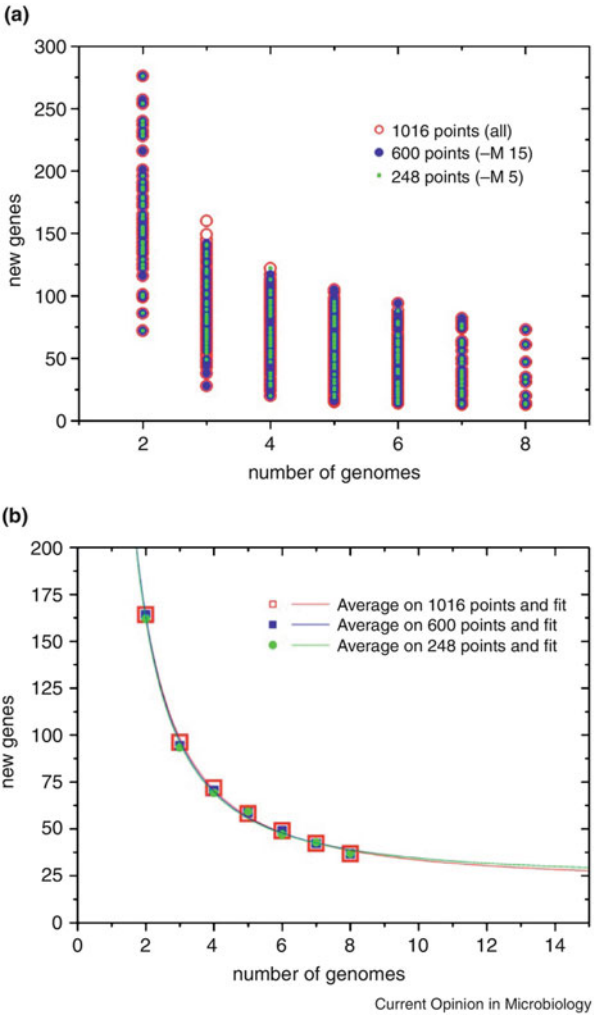

Fig. 1 Pangenome analysis plots for *Streptococcus agalactiae* genomes ($n = 8$). (**a**) Number of new genes detected for adding a genome g to $g - 1$ genomes. Red bubbles: 1016 points for the total number of comparisons (no subsampling). Blue bubbles: 600 points (subsampling, multiplicity of 15). Green bubbles: 248 points (subsampling, multiplicity of 5). (**b**) Regression curve on averages (the subsampling method has limited impact on the outcome)

decay) regressions functions; practical implementations of such optimizations are described in Tettelin et al. (2008), Eppinger et al. (2010, 2011), Mongodin et al. (2013) and Riley et al. (2012).

Recently several stand-alone or server-based suites have become available for pangenome analysis; in the next paragraphs, we will review the most promising and interesting initiatives. See also Table 2 for additional details.

Table 2 Pangenome software synopsis

	Roary	GET_HOMOLOGUES	EDGAR	ITEP	Harvest	PanOCT	panX	PGAP	PanGP	PanCGHweb	SplitMEM
Software type	Package/ Module	Package/Module		Framework/ Library	Toolkit/Suite	Package/ Module		Pipeline/ Workflow			Package/ Module
Interface	Command line interface	Command line interface	Web user interface	Command line interface	Command line interface	Command line interface	Web user interface	Command line interface	Graphical user interface	Web user interface	Command line interface
Input data	Annotated assemblies										
Input format	GFF3				GGR, FASTA, VCF						
Operating system	Unix/ Linux	Unix/Linux, Mac OS		Unix/Linux	Unix/Linux, Mac OS	Unix/ Linux		Unix/Linux	Windows/ Linux		Unix/Linux
Programming languages	Perl	Perl, R	Javascript, R	Python, Shell (Bash)	C++, Python, Shell (Bash)	Perl		Perl	C++		C++
License	GNU General Public License version 2.0	GNU General Public License		GNU General Public License version 2.0		GNU General Public License version 2.0					Apache License version 2.0
Computer skills	Advanced	Advanced	Basic	Advanced	Advanced	Advanced	Basic	Advanced	Basic	Basic	Advanced
	PanViz	EUPAN	PanTools	Spine	AGEnt	Panseq	PanWeb	PanGet	micropan	Pan-Tetris	ClustAGE
Software type	Package/ Module	Toolkit/Suite	Package/ Module	Package/Module	Package/ Module	Package/ Module		Application/ Script	Package/ Module	Framework/ Library	
Interface	Command line interface	Command line interface	Graphical user interface	Command line interface	Command line interface	Command line interface	Web user interface	Command line interface	Command line interface	Graphical user interface	Web user interface and Command line interface
Input data							An annotation files for each genome				
Input format							EMBL				
Operating system	Unix/ Linux	Unix/Linux	Unix/ Linux, Mac OS, Windows	Unix/Linux, Mac OS, Windows	Unix/Linux, Mac OS, Windows	Unix/ Linux, Mac OS, Windows		Unix/Linux, Mac OS, Windows	Unix/ Linux, Mac OS, Windows	On any machine with a Java VM installed	

Programming languages	Javascript	C++, Perl, R	Java	Perl	Perl	Perl	PHP, R	Perl	R	Java	Perl
License	GNU General Public License version 2.0			GNU General Public License version 2.0	GNU General Public License version 2.0	GNU General Public License version 3.0			GNU General Public License version 2.0		GNU General Public License version 3.0
Computer skills	Advanced	Advanced	Medium	Advanced	Advanced	Advanced	Basic	Advanced	Advanced	Medium	Basic
	DeNoGAP	PanACEA	LS-BSR	PANNOTATOR	BGDMdocker	PanCake	Panaconda	PanFunPro	Piggy	seq-seq-pan	NGSPanPipe
Software type	Pipeline/ Workflow	Toolkit/Suite	Pipeline/ Workflow		Pipeline/ Workflow	Package/ Module	Toolkit/ Suite	Application/ Script	Pipeline/ Workflow	Application/ Script	Script
Interface	Command line interface	Command line interface	Command line interface	Web user interface	Command line interface	Command line interface	Command line interface	Web user interface and command line	Command line interface	Command line interface	Command line interface
Input data							An annotation of genome features with some family designation	A list of genomes	Bacterial genome assemblies		Reads in FASTQ format, reference sequence file (FASTA format), parameter for filtering of reads and reference genome protein translation table (PTT) file
Input format								FASTA	GFF3		
Operating system	Unix/ Linux	Unix/Linux, Mac OS, Windows	Unix/ Linux		Unix/Linux, Mac OS, Windows	Unix/ Linux, Windows	Unix/ Linux		Unix/ Linux	Unix/Linux	
Programming languages	Perl	Perl				Python	Python	Perl	Perl, R, Shell (Bash)	Java, Python	Perl
License	GNU General Public License version 3.0	GNU General Public License version 2.0			GNU General Public License version 3.0				GNU General Public License version 3.0	FreeBSD	
Computer skills	Advanced	Advanced	Advanced	Basic	Advanced	Advanced	Advanced	Basic	Advanced	Advanced	Medium

3 Bayesian Decision Model

van Tonder et al. (2014) designed a methodology based on Bayesian decision model, able to analyze directly next-generation sequencing (NGS) data. The model defines the core-genome of bacterial populations allowing also the identification of novel genes. A nice caveat of this approach is that it can analyze even strains without a subset of genes since the model does not assume that all sequences have the entire core gene dataset present. The model has been benchmarked analyzing *Streptococcus pneumoniae* sequences.

4 BGDMdocker

BGDMdocker (Cheng et al. 2017) relies on docker technology to analyze and visualize bacterial pangenome and biosynthetic gene clusters. The pipeline consists of three stand-alone tools, namely Prokka v1.11 (Seemann 2014) for rapid prokaryotic genome annotation, panX (Ding et al. 2018) for pangenome analysis, and antiSMASH3.0 (Weber et al. 2015) for automatic genomic identification and analysis of biosynthetic gene clusters. The visualization supports several options, including alignment, phylogenetic trees, mutations mapped on the phylogenetic branches, and gene loss and gain mapping on the core-genome phylogeny. Benchmarking took place on 44 *Bacillus amyloliquefaciens* strains.

5 Bacterial PanGenome Analysis

Bacterial Pangenome Analysis (BPGA) (Chaudhari et al. 2016), comes with a handful of new options and features most notably that of optimizing the speed of execution. In addition, it offers various entity (core-, pangenome, and MLST) phylogeny, phyletic profile analysis (gene presence/absence), subset analysis, atypical sequence composition analysis, orthologous, and functional annotation for all gene datasets, user-selection of gene clustering algorithm, command line interface, and nice graphics. It runs both in Windows and in Linux as executables files (source code in Perl). BPGA has dependencies with other tools that require installation. In terms of input files, BPGA can "digest" the following file formats: GenBank (.gbk) files, protein sequence file (e.g.,.faa or .fsa or fasta format), binary (0,1) matrix (tab-delimited) file as output of other tools. The seven functional modules of BPGA algorithm include: Pangenome profile analysis, pangenome sequence extraction, exclusive gene family analysis, atypical GC content analysis, pangenome functional analysis, species phylogenetic analysis, and subset analysis.

6 ClustAGE

ClustAGE (Ozer 2018) suite (both online and stand-alone) clusters noncore accessory sequences within a collection of bacterial isolates implementing the BLAST algorithm. It is therefore focused on the accessory genomic dimension of pangenome; Benchmarking of this tool has taken place on *Pseudomonas aeruginosa* genome sequences.

7 DeNoGAP

DeNoGAP (Thakur and Guttman 2016) does many more than pure pangenome analysis, including functional annotation, gene prediction, protein classification, and orthology search; therefore, it is applicable both for complete and draft genomic data. To do this, it implements a big set of existing analysis algorithms. In terms of scalability, it runs linearly due to implementation of iteratively refined Hidden Markov models. Its modular structure supports easy updates and addition of new tools.

8 EDGAR

Implementing phylogenetic concepts like average amino acid and nucleotide identity indices, an online application namely "EDGAR" (Blom et al. 2009, 2016) was developed to support comparative genomic analyses of related isolates. Strong utilities of the suite include Venn diagrams and interactive synteny plots, as well as ease of access to taxa of interest and quick analyses like pangenome vs. core plot, the core-genome and singletons.

9 EUPAN

EUPAN (Hu et al. 2017) is one of the first concrete attempts to analyze eukaryotic pangenomes, even at a relatively low sequencing depth supporting gene annotation of pangenomic dataset, genome assembly, and identification of core and accessory gene datasets exploiting read coverage. The tool has been benchmarked using 453 rice genomes.

10 GET_HOMOLOGUES

GET_HOMOLOGUES (Contreras-Moreira and Vinuesa 2013) is a customizable and detailed pangenome analysis platform (open source written in Perl and R) for microorganisms addressed to non-bioinformaticians. GET_HOMOLOGUES can cluster homologous gene families using bidirectional best-hit clustering algorithms. The cluster granularity can be adjusted by the user based on various filtering strategies (e.g., by controlling key blast parameters such as percentage overlap and identity of pairwise alignments and E-score cutoff value). To estimate the size of the core- and pangenome, the tool supports both exponential and binomial mixture models to fit the data.

11 Harvest

Harvest (Treangen et al. 2014) is suitable for the analysis of (up to thousands of) microbial genomes. It hosts three modules, namely *Parsnp* (core-genome analysis), *Gingr* (output visualization), and *HarvestTools* (meta-analysis). Parsnp exploits jointly whole-genome alignment and read mapping to optimize accuracy and scalability aspects of sequence alignment; this approach can accommodate scalability for up to thousands of genomic datasets. For indexing purposes, it implements directed acyclic graph improving the identification of unique matches (anchors). The input of Parsnp is a directory of MultiFASTA files; the output includes core-genome alignment, variant calls, and a SNP tree, all of which can be visualized via Gingr. Broadly speaking, this tool represents a compromise between whole-genome alignment and read mapping. Parsnp performance has been evaluated on simulated and real data.

12 ITEP

ITEP (Benedict et al. 2014) is a suite of BASH scripts and Python libraries that interface with an SQLite database backend and a large number of tools for the comparison of microbial genomes. ITEP hosts several de novo prediction tools such as sequence alignment, metabolic, clustering, and protein prediction. Users can develop their own customized comparative analysis workflows.

13 LS-BSR

LS-BSR (large-scale BLAST score ratio) (Sahl et al. 2014), calculates a score ratio (BSR value = query/reference bit score) per coding sequence (matrix) within a pangenome dataset using BLAST (Altschul et al. 1997) or BLAT (Kent 2002) for

all-against-all alignment purposes. The output (bit score per CDS) can be visualized as a heatmap. Benchmarking has taken place on *Escherichia coli* and *Shigella* datasets.

14 micropan

micropan (Snipen and Liland 2015) is an R package for the pangenome study of prokaryotes. The R computing environment supports several options of statistical analyses (e.g., principal component analysis), pangenome models (e.g., Heaps' law), and graphics. External free software (e.g., HMMER3) is used for the heavy computations involved. Benchmarking has been carried out on 342 *Enterococcus faecalis* genomes.

15 NGSPanPipe

NGSPanPipe (Kulsum et al. 2018) supports microbial pangenome analysis directly from experimental reads. Benchmarking has been carried out using simulated reads of *Mycobacterium tuberculosis*. The pipeline expects as input experimental reads and outputs three files, one of which is a binary matrix showing the presence/absence of genes in each strain; this matrix can be used as input to other pangenome tools like PanOCT (Fouts et al. 2012) and PGAP (Zhao et al. 2012).

16 PanACEA

PanACEA (Clarke et al. 2018) is an open source stand-alone computer program written in Perl that supports users to create an interconnected set of html, javascript, and json files visualizing prokaryotic pan-chromosomes (core and variable regions) generated by PanOCT (Fouts et al. 2012) or other pangenome clustering tools. PanACEA was developed to serve as an intuitive, easy-to-use, stand-alone viewer. Regions and genes can be functionally annotated to allow for visual identification of regions of interest. PanACEA's memory and time requirements are within the capacities of standard laptops. Benchmarking took place on 219 *Enterobacter hormaechei* genomes.

17 Panaconda

Panaconda (Warren et al. 2017) creates whole-genome multiple sequence comparisons and provides a model for representing the relationship among sequences as a graph of syntenic gene families, by discovering collision points within a group of genomes. The first step is to create a de Bruijn graph and use its traversal to build a pan-synteny graph; the alphabet used is based on gene families (instead of nucleotide alphabet). This approach is novel in the context of generating a graph, wherein all sequences are fully represented as paths.

18 PanCake

PanCake (Ernst and Rahmann 2013) is another tool for pangenome analysis (core and unique regions) relying exclusively on sequence data and pairwise alignments (nucmer or BLAST), which makes it annotation independent (i.e., it processes pure whole-genome content). It hosts a command line interface with several subcommands, allowing to add chromosomes, to specify a genome for each chromosome, to add alignments, to compute core and unique regions, and to output selected regions of the analyzed chromosomes. Benchmarking took place on three genera, namely *Pseudomonas*, *Yersinia*, and *Burkholderia*. PanCake is written in Python.

19 PanFunPro

PanFunPro (Lukjancenko et al. 2013) exploits functional information (profiles) for pangenome analysis. The suite supports among others calculation of core, and accessory gene datasets, homology search (all-against-all and pairwise sub-querying), functional annotation (HMM-based), and gene-ontology information analysis. PanFunPro is available both as a standalone (Perl) tool and as a web server. Benchmarking took place on 21 *Lactobacillus* genomes.

20 PanGeT

PanGeT (Yuvaraj et al. 2017) can digest both genomic and proteomic data in order to construct the pangenome for a selection of taxa, exploiting BLASTN or BLASTP, respectively. In terms of performance, it has been benchmarked using a set of 11 *Streptococcus pyogenes* strains. The output is given in the form of a flower plot (core, dispensable, and strain-specific genes).

21 PanGFR-HM

PanGFR-HM (Chaudhari et al. 2018), is putting an interesting view point on the "table" of pangenome, by analyzing exclusively microbes from the Human Microbiome Project; it is a web-based platform integrating functional and genomic analysis for a collection of ~1300 complete human-associated microbial genomes exploiting a novel dimensionality of analysis that of body site (location of the bug in the human body) when comparing different groups of organisms.

22 PanGP

PanGP (Zhao et al. 2014) supports scalable pangenome analysis by analyzing clusters of orthologs pre-computed by OrthoMCL (Li et al. 2003), PGAP (Zhao et al. 2012), Mugsy-Annotator (Angiuoli et al. 2011), or PanOCT (Fouts et al. 2012). In order to predict core and accessory gene datasets, the suite implements random or distance-guided sampling; in the latter, the genomic diversity (GD) drives the sampling of strain permutations. GD is modeled relying on three alternative assumptions: GD is determined by the evolutionary distance on phylogenetic trees, the difference in gene numbers per strain, or by the discrepancy among gene clusters; among the three models the third seems more reliable (preferred model for PanGP).

23 PANINI

PANINI (Abudahab et al. 2018) is a web browser implementation for rapid online visualization and analysis of the core and accessory genome content, implementing unsupervised machine learning with stochastic neighbour embedding based on the t-SNE (t-distributed stochastic neighbour embedding) algorithm; this algorithm calculates first the similarities between the data (in high dimensional space) and then it minimizes the divergence between the two probability matrices over the embedding coordinates. PANINI expects as input the output of Roary (Page et al. 2015).

24 PANNOTATOR

PANNOTATOR (Santos et al. 2013) supports the efforts of automatic annotation transfer onto related unannotated genomes exploiting the existing annotation of a curated genome. From this perspective, it is not a main pangenome analysis tool, but rather as a side-product of cross-comparison it provides pangenomic-related

information. Its main contribution though to pangenome analysis is to accelerate the functional annotation of closely related isolates. For this task, it implements a relational database, interactive tools, several SQL reports, and a web-based interface. The expected input is the DNA strand, the gene prediction plus the reference annotated genome.

25 PanOCT

PanOCT (Fouts et al. 2012) is a graph-based ortholog clustering tool for pangenome analysis of closely related prokaryotic genomes exploiting conserved gene neighborhood information to separate recently diverged paralogs into distinct clusters of orthologs where homology-only clustering methods cannot. PanOCT is utilizing BLAST (Altschul et al. 1997) and conserved gene neighborhood information. Four input files are expected including a tabular file of all-versus-all BLASTP searches and the actual protein fasta sequences. PanOCT is specifically designed for pangenome analysis of closely related taxa (in order to be able to distinguish groups of paralogs into separate clusters of orthologs). In terms of memory requirements, PanOCT is greedier than other tools used to benchmark its performance; the memory usage is unchanged until the sixth genome, with a usage of 0.25 GB per genome, maxing out at 0.5 GB per genome by the 25th genome.

26 Panseq

Panseq (Laing et al. 2010) builds pangenomes and identifies single nucleotide polymorphisms (SNPs) using genomic data as input. In addition, it produces files for further phylogenetic analysis exploiting both the information of SNPs as well as the phyletic profile of accessory sequences; all these wrapped-up with a user-friendly graphical user interface.

27 Pan-Tetris

Pan-Tetris (Hennig et al. 2015) is a Java-based tool that exploits an aggregation technique inspired by the Tetris game, to provide an interactive and dynamic visualization of the gene content in a pangenome table with the option of editing and on-the-fly modification of user-defined (pan) gene groups. The suite has been tested on 32 *Staphylococcus aureus* genomes. Pan-Tetris is one of the first attempts that enable modification of the computed pangenome. The computation of whole genome alignment exploits progressiveMAUVE (Darling et al. 2010) algorithm.

28 PanTools

PanTools (Sheikhizadeh et al. 2016) suite supports the construction and visualization of pangenomes hosting online tools and algorithms; the visual representation of the pangenome is based on generalized De Bruijn graphs. The pangenome construction algorithm scales nicely even with large eukaryotic datasets. In addition to the basic pangenome tasks (construction and visualization), the suite supports other handy utilities such as adding, retrieving and grouping of sequences as well as annotating, reconstructing, and comparing genomes or pangenomes. Overall, it can easily support multi-genome read mapping, pangenome browsing, structure-based variation detection and comparative genomics. It has been benchmarked on *E. coli*, yeast, and *Arabidopsis thaliana* genomes.

29 PanViz

PanViz (Pedersen et al. 2017) is a pangenome visualization tool with some analysis options. It can generate dynamic visualizations supporting both pangenome subset selection as well as mapping of new genomes to existing pangenomes. The input data needed is a pangenome matrix (gene group presence/absence across the included genomes), as well as a gene ontology-based functional annotation of each gene group.

30 PanWeb

PanWeb (Pantoja et al. 2017) is a web application that performs pangenome analyses based on PGAP pipeline, providing in addition a user-friendly graphical interface supporting multiple user-defined analysis queries. It can be implemented by users without computational skills. As input, it receives the annotation files for each genome in EMBL format. A complete set of graphs (e.g., pangenome, accessory, core-genome, and unique genes) is provided.

31 panX

panX (Ding et al. 2018) identifies orthologous gene clusters in pangenomes via a user-friendly and interactive web-based visualization. The visualization consists of connected components that allow further analysis. The suite provides alignment and, phylogenetic tree, it maps mutations of each gene cluster and infers gene gain and loss in the core-genome phylogeny. The pipeline breaks annotated genomes into

genes and then clusters them into orthologous groups. To identify homologous proteins, panX performs an all-against-all similarity search, while the actual clustering of orthologous genes is carried out by a Markov clustering algorithm.

32 PGAdb-Builder

PGAdb-builder (Liu et al. 2016), constructs a pangenome allele database (PGAdb) to empower whole genome multilocus sequence typing (wgMLST) analyses and operates as a web service suite. Two modules are implemented, namely *Build_PGAdb* for building a PGAdb database and *Build_wgMLSTtree* for constructing a wgMLST tree and determine the genetic relatedness of the input sequences; both modules "digest" genome contigs in FASTA format. PGAdb-builder, has however dependencies with other existing suites like Prokka (Seemann 2014) and Roary (Page et al. 2015).

33 PGAP

PGAP (Zhao et al. 2012) supports pangenome analysis and in addition analysis of functional gene clusters, species evolution, genetic variation, and functional enrichment of query sequences. It outputs the basic pangenome structure and growth curve and in addition SNP and genomic variation information, phylogenetic, and functional annotation metadata. Benchmarking has taken place on *Streptococcus pyogenes* datasets.

34 PGAP-X

Building on PGAP, and in order to more effectively interpret and visualize the results, PGAP-X (Zhao et al. 2018) was developed. The visualization utility can intuitively lead to conclusions on pangenomic structure, conserved regions and overall on genetic variability throughout the pangenomic datasets at hand. Benchmarking has taken place on *S. pneumoniae* and *Chlamydia trachomatis* datasets. One current limitation of PGAP-X (that is not present in PGAP) is that it expects as input only complete genomes.

35 Piggy

Piggy (Thorpe et al. 2018) is a tool for analyzing the intergenic component of bacterial genomes and it is designed to be used in conjunction with Roary (Page et al. 2015). The latter works by analyzing protein-coding sequences thus excluding nonprotein-coding intergenic regions (IGRs) which typically account for approximately 15% of the genome. Piggy matches Roary except that it is based only on IGRs. Benchmarking took place on *Staphylococcus aureus* and *Escherichia coli* using large genome datasets. In terms of input and output, Piggy uses the same format as in Roary and has similar running time requirements. Piggy provides a means to rapidly identify IGR switches, with many evolutionary applications including analysis of the role of horizontal transfer in shaping the bacterial regulome.

36 pyseer

pyseer (Lees et al. 2018), is geared toward genome-wide association studies in the "world" of microbes with the task at hand to identify potential genetic variation linked with certain phenotypic aspects. Pyseer is actually a python implementation of a previous initiative written in C++, namely SEER (Lees et al. 2016). The foundation of pyseer is the use of K-mers (words) of variable length (input) coming from draft assemblies, while using a generalized linear model for each word their link with a potential phenotype is evaluated. In addition, multidimensional scaling of a pairwise distance matrix is implemented in order to control for population structure (embedded in the regression analysis).

37 Roary

Roary (Page et al. 2015) enables the construction of large pangenomes even on a typical desktop machine, yielding fairly accurate output. For example, it can digest up to 1000 strains (13 GB of RAM) building the pangenome in ~4 h. Roary achieves high accuracy which is attributable to utilization of the context of conserved gene neighborhood information. A suite of command line tools is provided to interrogate the dataset providing union, intersection, and complement.

38 seq-seq-pan

seq-seq-pan (Jandrasits et al. 2018) is a workflow for the sequential alignment of sequences to build a pangenome data structure and a whole-genome alignment. seq-seq-pan builds a pangenome data structure allowing editing (addition or removal) of genomes from a set of aligned sequences and subsequent re-alignment of the whole-genome sequences; for whole-genome alignments it relies on progressiveMauve (Darling et al. 2010). The alignment is optimized for generating a representative linear presentation of the aligned set of genomes.

39 Spine and AGEnt

Spine (Ozer et al. 2014) determines the core-genome from a group of genomic sequences and AGEnt (Ozer et al. 2014) identifies the accessory genome in draft genomic sequences. They both use nucmer to align sequences. The pipeline has been tested on genome sequences of *Pseudomonas aeruginosa*. However, as mentioned by the authors, whole genome alignment of reference genomes and core-genome identification with Spine can be time-consuming.

40 SplitMEM

SplitMEM (Marcus et al. 2014) scales linearly in terms of time and space in relation to the number of genomes of interest. To do this, it traverses suffix trees (for the genomes) and builds compressed de Bruijn graphs of pangenomes. In terms of notation, nodes within the graph represent conserved or strain-specific sequences of the pangenome. Benchmarking has taken place on *Bacillus anthracis* and *E. coli* datasets.

41 Highlights

Pangenome analysis has today many options when it comes to practical implementation. Depending on the analysis focus, the desired input and output, the dependability on other algorithms, as well as the modeling parametrization, users have many options to choose from. In the current review, we highlight the following five tools: BPGA (Chaudhari et al. 2016) for its very fast execution time, the intuitive handling and the user-defined clustering algorithm, Roary (Page et al. 2015) due to its internal processing (clustering of high similarity sequences) that results in linear memory consumption, LS-BSR (Sahl et al. 2014) that similarly to Roary performs pre-clustering reducing substantially the running time, PanOCT (Fouts et al. 2012), which takes into account both homology and positional gene neighborhood

information and PGAP (Zhao et al. 2012) that can work also with draft forms of genomic data such as annotated assemblies.

42 Food for Thought

The final results and conclusions of a pangenome analysis, among others, massively depend on the following aspects, that need thoughtful consideration prior to embarking any such project: Homology search algorithm, the phylogenetic sample at hand, the pangenome model implemented and the type and quality of sequence entities (e.g., DNA, protein, presence/absence—phyletic profile, and SNPs).

For example, when it comes to homology definition based on sequence similarity there is a wide range of similarity thresholds used in previous attempts: $i = 50\%$, $L = 50\%$ (Tettelin et al. 2005), $i = 70\%$, $L = 70\%$ (Hiller et al. 2007), $i = 70\%$, $L = 50\%$ (Meric et al. 2014), $i = 30\%$, $L = 80\%$ (Bentley et al. 2007), where i stands for sequence identity and L for sequence length.

The starting level (ORFs, CDSs, genes, proteins, SNPs) and the quality (in silico, manual curation) of annotation as well as inherent bacterial genomic complexity at the sequence level such as low complexity repeats, recombination hot spots, horizontally acquired genomic fragments constitute other important aspects of consideration. Such information variability can massively affect the predicted conserved and unique genes in favor of the former or the latter; this might also determine the structure of pangenome (open or closed).

43 Conclusions

Being able algorithmically to digest the largest possible pool of data available is critical in order to approach more reliably the phylogenetic history of bacterial populations. Indeed such comparative genomic analyses started by exploiting ~0.07% of a genome (16s rRNA) (Woese 1987), latter on using up to ~0.2% of the genomic information (MLST) (Maiden et al. 1998), and recently up to 100% of the information exploiting the pangenome wealth of data (Medini et al. 2005; Tettelin et al. 2005).

The recent explosion of sequencing projects replaced the limiting factor of *data sparsity* with the *immense data dimensionality* (Vernikos 2010) and we are now in the middle of a transformation moving from top-down (trying to fit the limited data to the model) to bottom-up approaches in an attempt to move from the "infant" stage of single-strain genomics to the post pangenome era of "adulthood." The model assumptions therefore become less and less pivotal as the pace of primary data generation continues to grow exponentially, asking not for modeling superpower but instead interpretation and connecting the dots super skills.

References

Abudahab K, Prada JM, Yang Z, Bentley SD, Croucher NJ, Corander J, Aanensen DM (2018) PANINI: pangenome neighbour identification for bacterial populations. Microb Genom 5(4). https://doi.org/10.1099/mgen.0.000220

Ahmed A, Earl J, Retchless A, Hillier SL, Rabe LK, Cherpes TL, Powell E, Janto B, Eutsey R, Hiller NL et al (2012) 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):3922–3937

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Angiuoli SV, Dunning Hotopp JC, Salzberg SL, Tettelin H (2011) Improving pan-genome annotation using whole genome multiple alignment. BMC Bioinf 12:272

Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND (2014) ITEP: an integrated toolkit for exploration of microbial pan-genomes. BMC Genomics 15:8

Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T, Cronin A, Davis PH, Holroyd NE, Jagels K, Maddison M, Moule S, Rabbinowitsch E, Sharp S, Unwin L, Whitehead S, Quail MA, Achtman M, Barrell B, Saunders NJ, Parkhill J (2007) Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. PLoS Genet 3(2):e23

Bhardwaj T, Somvanshi P (2017) Pan-genome analysis of Clostridium botulinum reveals unique targets for drug development. Gene 623:48–62. https://doi.org/10.1016/j.gene.2017.04.019

Blom J, Albaum SP, Doppmeier D, Puhler A, Vorholter FJ, Zakrzewski M, Goesmann A (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinf 10:154

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

Boissy R, Ahmed A, Janto B, Earl J, Hall BG, Hogg JS, Pusch GD, Hiller LN, Powell E, Hayes J et al (2011) Comparative supragenomic analyses among the pathogens Staphylococcus aureus, Streptococcus pneumoniae, and Haemophilus influenzae using a modification of the finite supragenome model. BMC Genomics 12:187

Bottacini F, Medini D, Pavesi A, Turroni F, Foroni E, Riley D, Giubellini V, Tettelin H, van Sinderen D, Ventura M (2010) Comparative genomics of the genus Bifidobacterium. Microbiology 156(Pt 11):3243–3254

Broadbent JR, Neeno-Eckwall EC, Stahl B, Tandee K, Cai H, Morovic W, Horvath P, Heidenreich J, Perna NT, Barrangou R et al (2012) Analysis of the Lactobacillus casei supragenome and its influence in species evolution and lifestyle adaptation. BMC Genomics 13:533

Brüggemann H, Jensen A, Nazipi S, Aslan H, Meyer RL, Poehlein A, Brzuszkiewicz E, Al-Zeer MA, Brinkmann V, Söderquist B (2018) Pan-genome analysis of the genus Finegoldia identifies two distinct clades, strain-specific heterogeneity, and putative virulence factors. Sci Rep 8 (1):266. https://doi.org/10.1038/s41598-017-18661-8

Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV et al (2011) Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A 108(11):4494–4499

Chaudhari NM, Gupta VK, Dutta C (2016) BPGA- an ultra-fast pan-genome analysis pipeline. Sci Rep 6:24373

Chaudhari NM, Gautam A, Gupta VK, Kaur G, Dutta C, Paul S (2018) PanGFR-HM: a dynamic web resource for pan-genomic and functional profiling of human microbiome with comparative features. Front Microbiol 9:2322

Cheng G, Quan L, Zhou Z, Ma L, Zhang G, Wu Y, Chen C (2017) BGDMdocker: an workflow base on Docker for analysis and visualization pan-genome and biosynthetic gene clusters of bacterial. bioRxiv:098392

Clarke TH, Brinkac LM, Inman JM, Sutton G, Fouts DE (2018) PanACEA: a bioinformatics tool for the exploration and visualization of bacterial pan-chromosomes. BMC Bioinf 19(1):246

Collingro A, Tischler P, Weinmaier T, Penz T, Heinz E, Brunham RC, Read TD, Bavoil PM, Sachse K, Kahane S et al (2011) Unity in variety — the pan-genome of the Chlamydiae. Mol Biol Evol 28(12):3253–3270

Collins RE, Higgs PG (2012) Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol 29(11):3413–3425

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

Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5(6):e11147

Davie JJ, Earl J, de Vries SP, Ahmed A, Hu FZ, Bootsma HJ, Stol K, Hermans PW, Wadowsky RM, Ehrlich GD et al (2011) Comparative analysis and supragenome modeling of twelve Moraxella catarrhalis clinical isolates. BMC Genomics 12:70

den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M (2010) Comparative genomics of the bacterial genus Listeria: genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 11:688

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

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

Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, Achtman M, Lindler LE, Ravel J (2010) 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):1685–1699

Eppinger M, Bunk B, Johns MA, Edirisinghe JN, Kutumbaka KK, Koenig SS, Creasy HH, Rosovitz MJ, Riley DR, Daugherty S et al (2011) Genome sequences of the biotechnologically important Bacillus megaterium strains QM B1551 and DSM319. J Bacteriol 193(16):4199–4213

Ernst C, Rahmann S (2013) PanCake: a data structure for pangenomes. German Conference on Bioinformatics, Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269(5223):496–512

Fouts DE, Brinkac L, Beck E, Inman J, Sutton G (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res 40(22):e172

Hennig A, Bernhardt J, Nieselt K (2015) Pan-Tetris: an interactive visualisation for pan-genomes. BMC Bioinf 16(Suppl 11):S3

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

Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD (2007) 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):R103

Hu Z, Sun C, Lu KC, Chu X, Zhao Y, Lu J, Shi J, Wei C (2017) EUPAN enables pan-genome studies of a large number of eukaryotic genomes. Bioinformatics 33(15):2408–2409

Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C (2011) The Salmonella enterica pan-genome. Microb Ecol 62(3):487–504

Jandrasits C, Dabrowski PW, Fuchs S, Renard BY (2018) Seq-seq-pan: building a computational pan-genome data structure on whole genome alignment. BMC Genomics 19(1):47

Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12(4):656–664

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J et al (2007) Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet 3(12):e231

Kulsum U, Kapil A, Singh H, Kaur P (2018) NGSPanPipe: a pipeline for pan-genome identification in microbial strains from experimental reads. Adv Exp Med Biol 1052:39–49

Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, Thomas JE, Gannon VP (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinf 11:461

Lapidus A, Goltsman E, Auger S, Galleron N, Segurens B, Dossat C, Land ML, Broussolle V, Brillard J, Guinebretiere MH et al (2008) Extending the Bacillus cereus group genomics to putative food-borne pathogens of different toxicity. Chem Biol Interact 171(2):236–249

Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends Genet 25 (3):107–110

Lees JA, Vehkala M, Valimaki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies MR, Steer AC, Tong SY, Honkela A, Parkhill J, Bentley SD, Corander J (2016) Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun 7:12797

Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J (2018) Pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics 34(24):4310–4312

Lefebure T, Stanhope MJ (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol 8(5):R71

Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9):2178–2189

Liu YY, Chiou CS, Chen CC (2016) PGAdb-builder: a web service tool for creating pan-genome allele database for molecular fine typing. Sci Rep 6:36213

Lukjancenko O, Thomsen M, Voldby Larsen M, Ussery D (2013) PanFunPro: PAN-genome analysis based on FUNctional PROfiles [version 1; referees: 3 approved with reservations]. F1000Res 2:265

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95(6):3140–3145

Marcus S, Lee H, Schatz MC (2014) SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics 30(24):3476–3483

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15(6):589–594

Meric G, Yahara K, Mageiros L, Pascoe B, Maiden MC, Jolley KA, Sheppard SK (2014) A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic campylobacter. PLoS One 9(3):e92798

Mongodin EF, Casjens SR, Bruno JF, Xu Y, Drabek EF, Riley DR, Cantarel BL, Pagan PE, Hernandez YA, Vargas LC et al (2013) Inter- and intra-specific pan-genomes of Borrelia burgdorferi sensu lato: genome stability and adaptive radiation. BMC Genomics 14:693

Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, Thomas AD, Ali R, Sharma K, Kyrpides NC, Reddy TB (2017) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res 45(D1):D446–D456

Ozer EA (2018) ClustAGE: a tool for clustering and distribution analysis of bacterial accessory genomic elements. BMC Bioinf 19(1):150

Ozer EA, Allen JP, Hauser AR (2014) Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. BMC Genomics 15:737

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31 (22):3691–3693

Pantoja Y, Pinheiro K, Veras A, Araujo F, Lopes de Sousa A, Guimaraes LC, Silva A, Ramos RTJ (2017) PanWeb: a web interface for pan-genomic analysis. PLoS One 12(5):e0178154

Pedersen TL, Nookaew I, Wayne Ussery D, Mansson M (2017) PanViz: interactive visualization of the structure of functionally annotated pangenomes. Bioinformatics 33(7):1081–1082

Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R et al (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol 190(20):6881–6893

Riley DR, Angiuoli SV, Crabtree J, Dunning Hotopp JC, Tettelin H (2012) Using Sybil for interactive comparative genomics of microbes on the web. Bioinformatics 28(2):160–166

Rodriguez-Valera F, Ussery DW (2012) Is the pan-genome also a pan-selectome? F1000Res 1:16

Sahl JW, Caporaso JG, Rasko DA, Keim P (2014) The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ 2:e332

Santos AR, Barbosa E, Fiaux K, Zurita-Turk M, Chaitankar V, Kamapantula B, Abdelzaher A, Ghosh P, Tiwari S, Barve N, Jain N, Barh D, Silva A, Miyoshi A, Azevedo V (2013) PANNOTATOR: an automated tool for annotation of pan-genomes. Genet Mol Res 12 (3):2982–2989

Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang YF (2010) Analysis of ultra low genome conservation in Clostridium difficile. PLoS One 5(12):e15147

Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T, Goesmann A, Joseph B, Konietzny S, Kurzai O et al (2008) 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):3473–3478

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14):2068–2069

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

Snipen L, Liland KH (2015) Micropan: an R-package for microbial pan-genomics. BMC Bioinf 16:79

Smokvina T, Wels M, Polka J, Chervaux C, Brisse S, Boekhorst J, van Hylckama Vlieg JE, Siezen RJ (2013) Lactobacillus paracasei comparative genomics: towards species pan-genome definition and exploitation of diversity. PLoS One 8(7):e68731

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278(5338):631–637

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS et al (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102 (39):13950–13955

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11(5):472–477

Thakur S, Guttman DS (2016) A De-Novo Genome Analysis Pipeline (DeNoGAP) for large-scale comparative prokaryotic genomics studies. BMC Bioinf 17(1):260

Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ (2018) Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. Gigascience 7(4):1–11

Treangen TJ, Ondov BD, Koren S, Phillippy AM (2014) The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 15 (11):524

van Schaik W, Top J, Riley DR, Boekhorst J, Vrijenhoek JE, Schapendonk CM, Hendrickx AP, Nijman IJ, Bonten MJ, Tettelin H et al (2010) Pyrosequencingbased comparative genome

analysis of the nosocomial pathogen Enterococcus faecium and identification of a large transferable pathogenicity island. BMC Genomics 11:239
van Tonder AJ, Mistry S, Bray JE, Hill DM, Cody AJ, Farmer CL, Klugman KP, von Gottberg A, Bentley SD, Parkhill J, Jolley KA, Maiden MC, Brueggemann AB (2014) Defining the estimated core genome of bacterial populations using a Bayesian decision model. PLoS Comput Biol 10(8):e1003788
Vernikos GS (2010) The pyramid of knowledge. Nat Rev Microbiol 8(2):91
Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154
Warren AS, Davis JJ, Wattam AR, Machi D, Setubal JC, Heath L (2017) Panaconda: application of pan-synteny graph models to genome content analysis. bioRxiv:215988
Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Muller R, Wohlleben W, Breitling R, Takano E, Medema MH (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43(W1): W237–W243
Woese CR (1987) Bacterial evolution. Microbiol Rev 51(2):221–271
Yuvaraj I, Sridhar J, Michael D, Sekar K (2017) PanGeT: pan-genomics tool. Gene 600:77–84
Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J (2012) PGAP: pan-genomes analysis pipeline. Bioinformatics 28(3):416–418
Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, Yu J, Wu J, Xiao J (2014) PanGP: a tool for quickly analyzing bacterial pan-genome profile. Bioinformatics 30(9):1297–1299
Zhao Y, Sun C, Zhao D, Zhang Y, You Y, Jia X, Yang J, Wang L, Wang J, Fu H, Kang Y, Chen F, Yu J, Wu J, Xiao J (2018) PGAP-X: extension on pan-genome analysis pipeline. BMC Genomics 19(Suppl 1):36
Zhong C, Han M, Yu S, Yang P, Li H, Ning K (2018) Pan-genome analyses of 24 Shewanella strains re-emphasize the diversification of their functions yet evolutionary dynamics of metal-reducing pathway. Biotechnol Biofuels 11:193. https://doi.org/10.1186/s13068-018-1201-1