Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in Microbiology

Ten years of pan-genome analyses

George Vernikos[1], Duccio Medini[2], David R Riley[3] and Hervé Tettelin[3]

Next generation sequencing technologies have engendered a genome sequence data deluge in public databases. Genome analyses have transitioned from single or few genomes to hundreds to thousands of genomes. Pan-genome analyses provide a framework for estimating the genomic diversity of the dataset at hand and predicting the number of additional whole genomes sequences that would be necessary to fully characterize that diversity. We review recent implementations of the pan-genome approach, its impact and limits, and we propose possible extensions, including analyses at the whole genome multiple sequence alignment level.

Addresses
[1] Novartis (Hellas) S.A.C.I., 12th Km Athens-Lamia North Road, 14451 Metamorfossi, Athens, Greece
[2] Novartis Vaccines Research, Via Fiorentina 1, 53100 Siena, Italy
[3] Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, 801 West Baltimore Street, Baltimore, MD 21201, USA

Corresponding author: Tettelin, Hervé (tettelin@som.umaryland.edu)

Current Opinion in Microbiology 2015, **23**:148–154

This review comes from a themed issue on **Genomics**

Edited by **Neil Hall** and **Jay CD Hinton**

For a complete overview see the Issue and the Editorial

Available online 5th December 2014

http://dx.doi.org/10.1016/j.mib.2014.11.016

1369-5274/© 2014 Elsevier Ltd. All right reserved.

Introduction

The pan-genome defines the entire genomic repertoire of a given phylogenetic clade and encodes for all possible lifestyles carried out by its organisms. The phylogenetic resolution of the clade of interest is unlimited ranging from species and serovar to phylum, kingdom and beyond. The term pan-genome was first coined a decade ago by Tettelin *et al.* [1] and describes the union of sequence entities (usually genes or open reading frames, ORFs) shared by genomes of interest. The wording in the scientific literature often used to describe the union, intersection and any combination of subsets from this sequence collection is fairly variable: pan-genome, core genes, dispensable genes and strain-specific genes [1,2], supragenome, distributed genes and unique genes [3], accessory and character gene pool [4••], and flexible regions [5]. For the purpose of this review, we will use the following nomenclature (from Ref. [1]): the pan-genome that encompasses the entire repertoire of genes accessible to the clade studied; the core genome that contains genes shared by all strains within the clade and typically includes genes responsible for the basic aspects of the biology of the clade and its major phenotypic traits; the dispensable genome made of genes shared by a subset of the strains and contributes to the species diversity, it might encode supplementary biochemical pathways and functions that are not essential for growth but which confer selective advantages, such as adaptation to different niches, antibiotic resistance, or colonization of a new host [2]; and strain-specific genes.

Twenty years after the first complete genome sequence (*Haemophilus influenzae* [6]), there are, as of May 2014, 18940 complete genomes (of which 94% are bacteria) and 3087 finished genome projects [7], making exploration of the boundaries of the biological species definition via multi-genome — pan-genome — analyses tempting. Indeed, after the pioneering work on the pan-genome in 2005 [1], several other pan-genome projects have followed differing mainly on the number of analyzed genomes/strains, the phylogenetic resolution of interest, the mathematical prediction model, the model assumptions and parameters, the alignment search algorithm and associated parameters (% identity and % of pairwise aligned sequence length), threshold of orthology definition, and genome sampling order. For example, from the phylogenetic resolution point of view, there are projects focused on the species level, genus level, and even at the class, phylum or super kingdom levels (Table 1). Lapierre and Gogarten [4] showed in the largest — in terms of phylogenetic resolution — bacterial pan-genome analysis to date that on average strain-specific genes and dispensable genes shared by only a few of the strains account for 28% of a bacterial genome whereas the extended core genes, shared by all or almost all genomes, account for nearly 8% of the gene repertoire; the remaining dispensable genes account for the majority (64%) of a bacterial genome and are usually involved in specific environmental niche adaptation.

Pan-genome analyses provide a framework to determine the genomic diversity of the dataset at hand, but also to predict, via extrapolation, how many additional whole genome sequences would be necessary to characterize the entire pan-genome or gene repertoire. It should be noted that extrapolations will only be robust if a sufficient number of genomes (data points) is considered. We recommend that at least five genomes be compared,

Table 1

Examples of the application of pan-genome approaches at different levels of phylogenetic resolution

Level	Organism	Approach[a]	# genomes	Core size (# genes)	Year (reference)
Species	*Streptococcus agalactiae*	ORFsim, Comb	8	1806	2005 [1]
	Neisseria meningitidis	ORFsim, Comb	6	1337	2008 [42]
		ORFsim, Comb	20	1630	2011 [43]
	Borrelia burgdoferi	ORFsim, Comb	21	1200	2013 [12]
	Escherichia coli	ORFsim, Comb	17	2344	2008 [26]
	Enterococcus faecium	ORFsim, Comb	7	2172	2010 [44]
	Yersinia pestis	ORFsim, Comb	14	3668	2010 [10]
	Streptococcus pyogenes	OG, Comb	11	1376	2007 [45]
	Clostridium difficile	OG, Comb	15	1033	2010 [46]
	Lactobacillus paracasei	OG	34	1800	2013 [47]
	Campylobacter jejuni	ORFsim, Ref	130	1042	2014 [27]
	Campylobacter coli	ORFsim, Ref	62	947	2014 [27]
	Haemophilus influenzae	FSM	13	1450	2007 [48]
	Streptococcus pneumoniae	FSM	17	1400	2007 [3]
		ORFsim, Comb	44	1666	2010 [49]
	Staphylococcus aureus	FSM	16	2245	2011 [50]
	Moraxella catarrhalis	FSM	12	1755	2011 [51]
	Lactobacillus casei	FSM	17	1715	2012 [52]
	Gardnerella vaginalis	FSM	17	746	2012 [53]
Group	*Bacillus cereus*	ORFsim, Comb	4	3000	2008 [54]
	Bacillus subset of species	ORFsim, Comb	12	2009	2011 [11]
Genus	*Streptococcus*	OG, Comb	26	600	2007 [45]
		ORFsim, Comb	52	522	2010 [49]
	Prochlorococcus	ORFsim, Comb	12	1273	2007 [55]
	Bifidobacterium	ORFsim, Comb	14	967	2010 [56]
	Listeria	BMM	13	2032	2010 [57]
	Salmonella	BMM	35	2811	2011 [15]
Class	Bacilli	IMGM	172	143	2012 [58•]
Phylum	*Chlamydiae*	OG	19	560	2011 [59]
Super kingdom	*Eubacteria*	Gene freq.	573	250	2009 [4••]

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

but many more are desirable and this should no longer be problem in the current climate of next generation whole genome sequencing.

Previous analyses at the bacterial species level have demonstrated that several species including human pathogens and environmental bacteria display an open pan-genome [3,8••]. An open pan-genome indicates that a very large, undetermined number of additional genomes would be needed to identify all genes accessible to the species. In contrast, for species with a closed pan-genome, additional genomes sequenced do not provide additional new genes to expand the pan-genome — the entire gene repertoire has been characterized, assuming that the sampling of strains sequenced is not biased. These observations are derived from extrapolations based on the current sample of bacterial genomes analyzed.

Technical implementation

Users interested in pan-genome analyses have the option to implement methods like whole-genome multiple sequence alignment to improve sensitivity for high-resolution comparisons at the species/sub-species or strain level, or they could use amino acid similarity, protein clustering (both ab initio and based on ortholog clusters such as COGs [9]), structural alignment, and pathway/metabolic information at higher levels to decrease noise and eliminate sequence alignment artifacts for genomes with limited primary sequence similarity.

The original implementation of the algorithm or workflow pipeline for a pan-genome analysis [1,8], although conceptually intuitive, has several potential technical pitfalls, some of which are pivotal enough to directly impact the conclusions drawn. Issues include the prediction of an open vs. closed pan-genome, a fast or slowly growing pan-genome (the rate at which new genes identified from additional genomes expand the pan-genome), genes that are assigned to the core vs. the dispensable genomes (the choice of parameters affects whether genes are considered shared/core or not core), and the determination of the core genome size (the asymptote for the extrapolation of the trend of the decreasing core genome as more genomes are added).

In addition, the combinatorial aspect of the approach, whereby all permutations of adding a genome to a set of genomes previously analyzed are considered, does not scale to large numbers of genomes. The number of comparisons used to calculate the new, core and shared genes at the nth genome can be modeled with the following function, where C is the total number of comparisons and N is the total number of genomes:

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

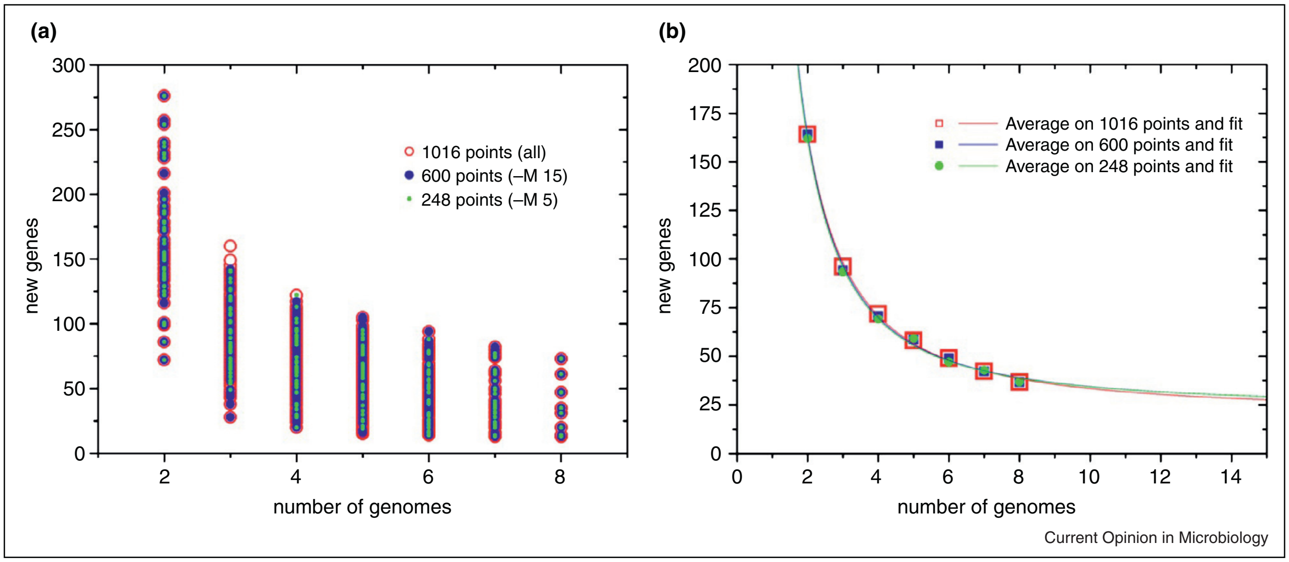
To circumvent the scalability problem, we developed a method for sub-sampling the number of comparisons to be performed between N genomes. The sampling approach is controlled random in that for each strain, at each value of n, comparisons are randomly selected while ensuring that each strain undergoes the same number of comparisons. Each comparison represents adding a target strain to a random sampling of $n - 1$ genomes and counting the new, core and dispensable genes. The number of comparisons per strain, or multiplicity, is configurable such that a balance can be struck between dataset size and available compute power.

For values of N low enough to allow for all combinations to be calculated, we observed that even aggressive sub-sampling still provides a representative set of data points with average or median values nearly identical to those of the entire set of data points (all combinations). Figure 1 illustrates the application of three multiplicity levels of sub-sampling on $N = 8$ *Streptococcus agalactiae* genomes, the set of genomes we used originally to develop the pan-genome concept [1].

We also began fitting regression curves using a power law model (Heap's law [8]) instead of an exponential decay. The power law model should be fitted only on the tail of the distribution and as such exclude low values of n. Examples of power law regression fitting as well as sub-sampling of comparisons include those presented in Tettelin *et al.* [8] and more recent analyses we performed on *Yersinia* [10], *Bacillus* [11], *Borrelia* [12] and the Strepneumo Sybil website dedicated to the comparative analysis of 34 *Streptococcus pneumoniae* genomes ([13], http://strepneumo-sybil.igs.umaryland.edu/pangenome).

In the recent past, in an effort to computationally standardize pan-genome analyses, several online tools and software suites have been developed. For example, GET_HOMOLOGUES [14•] is a customizable and detailed pan-genome analysis platform for microorganisms addressed to non-bioinformaticians. BLAST atlas [15] intuitively visualizes which genes from the reference genome are present in other genomes. Mugsy-Annotator [16] identifies syntenic orthologs

Figure 1



New gene discovery plots for the pan-genome analysis of eight *Streptococcus agalactiae* genomes. **(a)** Distribution of data points for the number of new genes identified with all combinations of adding a genome n to $n - 1$ genomes (see Ref. [1]). The total number of comparisons for 8 genomes without sub-sampling results in 1016 data points represented by red circles. Blue dots: sub-sampling of the comparisons at a multiplicity of 15, resulting in 600 data points. Green dots: sub-sampling of the comparisons at a multiplicity of 5, resulting in 248 data points. The plot shows that the controlled random sampling ensures the lack of bias in the distribution of comparisons at different multiplicities. **(b)** As a consequence of the lack of bias, the averages and regression curves are not significantly affected by the sub-sampling approach.

and evaluates annotation quality in prokaryotic genomes using whole genome multiple alignment.

PANNOTATOR [17•], a web-based suite, automates the transfer of annotation onto closely related genomes, a cumbersome task in pan-genome analysis projects. Panseq [18•] is another software suite that supports core/dispensable gene mapping and classification of a collection of genome sequences. This tool identifies both unique stretches of DNA and conserved regions within a group of sequences more or less in a similar philosophy as the MAUVE whole genome aligner [19] but more focused on the pan-genome and automation. PGAP [20•] executes five analysis modules: cluster analysis of functional genes (the core module), pan-genome profile analysis, genetic variation analysis of functional genes, species evolution analysis and function enrichment analysis of gene clusters. Hence the deliverables include the pan-genome profile and curve, genome variation and SNP data, pan-genome and SNP-based phylogenetic trees and additional functional information for each gene/protein.

SplitMEM [21•] generates a compressed de Bruijn graph of the pan-genome by traversing a suffix tree of the genomes. Within the graph, sequences that are shared or unique in the population are represented as nodes, and edges represent branch points between shared and strain-specific sequences. PanOCT [22•] is a graph-based ortholog clustering tool for pan-genome analysis of closely related prokaryotic genomes exploiting conserved gene neighborhood information to separate recently diverged paralogs into distinct clusters of orthologs.

PanGP [23•] builds upon clusters of orthologs such as those computed with OrthoMCL [24], PGAP [20] or PanOCT [22•] and performs scalable pan-genome analyses. PanGP implements two sampling algorithms — totally random and distance guide — on combinations of *N* strains and generates pan-genome, core genome and new gene graphs similar to Tettelin *et al.* [1]. van Tonder *et al.* [25] developed a Bayesian decision model to define the estimated core gene pool of bacterial populations directly from next-generation whole genome sequencing data, enabling the identification of putative novel genes associated with key biological functions. The model does not require that every single isolate sequenced harbor all core genes. This accommodates for the possible presence of rare strain variants that may be missing some genes that would otherwise be considered core. As a case study, the Tatusov *et al.* COGs [9] and the Bayesian model were applied to the core genome of *S. pneumoniae*. The methods identified 1194 and 980 core genes, respectively, with a common set of 840 core genes.

The pan-genome implementation is influenced by six major aspects: (A) the alignment algorithm (e.g., BLAST or FASTA) and parameters (% identity and % sequence length) used to define similarity (orthologs, xenologs, and paralogs); (B) the phylogenetic resolution of the target clade (narrow vs. wide); (C) the sample of input genomes selected or available to represent the target clade; (D) the model used to estimate the of number of new genes vs. the number of genomes; (E) the type and quality of sequence annotation (genes, ORFs, CDSs); and (F) the all-against-all level of comparison (e.g., sequence similarity vs. phyletic profile of gene presence/absence regardless of sequence similarity).

For example, Tettelin *et al.* [1] used a similarity threshold of 50% identity over 50% of the sequence lengths, whereas Hiller *et al.* [3] used a more stringent threshold of ≥70% sequence conservation over 70% of the sequence length. Rasko *et al.* [26] used the BLAST score ratio with a strict threshold for inclusion of >80% over the length of the proteins, while recently Meric *et al.* [27] exploited a BLAST match of ≥70% identity over ≥50% of the sequence lengths. Bentley *et al.* [28], although not geared towards pan-genome analysis, used a threshold of 30% identity over 80% of sequence lengths to define orthologous sets of genes via an all-against-all reciprocal best FASTA hit search.

Broad taxonomic groupings (e.g., at the phylum or kingdom level) or inherent sequence variability as observed for surface protein antigens (driven by immune selection) or substrate specificity of transporters [29] can in theory increase the ambiguity of genuine orthology to a point where high-resolution algorithms such as PSI-BLAST or phyletic profile strategies (gene presence/absence instead of sequence similarity) have to be implemented [4].

The starting level of annotation for a pan-genome project also deserves thoughtful consideration at the pre-implementation, design level since the analysis is annotation-dependent [29]. The key point here is the definition of sequence entity targeted; for example ORFs (defined as any sequence between a start and a stop codon) vs. predicted protein-coding genes (that were subsequently manually curated or not). If ORFs are used, what is the minimum sequence length of start-to-stop codon (e.g., 100 bp, 500 bp, 1000 bp) used? If predicted genes are used instead, do we trust the in silico prediction without manual curation especially at routinely problematic sites like translation initiation sites, frame shifts, internal premature stop codons, or intragenic low complexity repeats? What about missed, un-annotated genes, or ORFans [30]? Such annotation inconsistencies can greatly impact the core and dispensable genomes in favor of the former or the latter, influencing in turn whether the pan-genome at hand will be predicted to be open or closed.

Species phylogeny

The pan-genome concept is so profound in comparative genomics that it is sometimes hard to reconcile it with the classical definition of species [31] or to effectively model it using strictly bifurcating tree-like structures [32]. In terms of phylogenetic resolution, traditional classification systems analyze a handful of genetically distinct, often non-overlapping species representative features and capture only a tiny fraction of the species variation [33]; as such they struggle to cope with the increasingly complex structure, the overlapping (fuzzy) boundaries, and the dynamic nature of bacterial populations. Moving from single-gene, for example, 16s rRNA [34], phylogenies which exploit only a tiny fraction (~0.07%) of a genome to approaches using a larger sequence sample, for example, multilocus sequence typing — MLST (~0.2%) [35], and recently to pan-genomes (100% coverage) [1,2], brings us closer to understanding and more reliably reconstructing the phylogenetic history of bacterial populations.

The current recognition of increased microbial genome fluidity indicates that the fundamental definition of a biological species [31] fails in some cases to provide a realistic description of the dynamic relationships that shape microbial evolution. These findings do not support the strictly bifurcating tree of life as a means of phylogenetic analysis and instead favor a phylogenetic network [36], which better represents the true relationships among species that are characterized by high rates of DNA exchange [37–40].

In the case of *S. agalactiae* (and many other bacteria) housekeeping genes comprise the majority of the core dataset, whereas strain-specific genes are often part of long mobile elements or genomic islands that may originate from horizontal gene transfer events such as conjugation, transformation, and transduction. In *S. agalactiae* and *S. pyogenes*, 10% of the strain-specific genes are of phage origin and were acquired via transduction [29].

Moreover, the pan-genome can even challenge traditional and widely used typing systems [2]. Strains of different serotypes or serogroups can be more closely related than those within the same serogroup, or strains of the same sequence type (MLST) can be genetically distant at the whole genome level. The collapse of the relationship between serotypes and genetic diversity is due to the dispensable nature of the capsular polysaccharide operon that indulges in high rates of genetic exchange between strains of different genetic landscapes, blurring the phylogenetic boundaries of species with open pan-genomes. On the contrary, traditional typing methods, which are based on a handful of genes belonging to the core genome, have a very limited genome resolution (~0.2%) ignoring the gene content of the dispensable dataset that often encodes important functions such niche adaptation or pathogenic and virulence properties.

Conclusions

High-throughput next generation sequencing projects have paved the way from single-genome studies to pan-genome analyses. This enabled revisiting of top-down — data-limiting — theories, models, and fundamental biological definitions by re-designing algorithmic methods and toolkits. Today, the limiting factor is no longer data sparsity but instead immense data dimensionality [41]. The main drawback of top-down analyses was the huge dependency on model parameters and hypotheses built on a limited amount of data. As more data became available, it became easier to generalize and draw more realistic conclusions. Theoretically speaking, a model becomes uninformative once the sample of the available data approximates the totality of the data complexity. Although we are not there yet, bottom-up definitions stemming from big data provide the chance for biology to mature from its embryonic stage of single-genome studies to the post pan-genome — *insights* — era of realization. It is conceivable that pan-genome studies for closely related taxa could be performed at the nucleotide sequence rather than the gene level, using whole genome multiple alignment (locally collinear blocks) or raw read datasets, revealing not only all protein coding sequences, but also non-protein coding features including promoters, small RNAs, and repeat structures.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest
•• of outstanding interest

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

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

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

4. Lapierre P, Gogarten JP: **Estimating the size of the bacterial**
•• **pan-genome**. *Trends Genet* 2009, **25(3)**:107-110.
The largest (in terms of phylogenetic resolution) bacterial pan-genome analysis at the super kingdom level.

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

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

7. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL: **Alignment of whole genomes**. *Nucleic Acids Res* 1999, **27(11)**:2369-2376.

8. Tettelin H, Riley D, Cattuto C, Medini D: **Comparative genomics:**
•• **the bacterial pan-genome**. *Curr Opin Microbiol* 2008, **11(5)**:472-477.
A previous review of pan-genome analyses that introduces new regression models.

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

10. Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, Achtman M, Lindler LE, Ravel J: **Genome sequence of the deep-rooted *Yersinia pestis* strain Angola reveals new insights into the evolution and pangenome of the plague bacterium**. *J Bacteriol* 2010, **192(6)**:1685-1699.

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

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

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

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

A tool available for pan-genome analysis.

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

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

17. • Santos AR, Barbosa E, Fiaux K, Zurita-Turk M, Chaitankar V, Kamapantula B, Abdelzaher A, Ghosh P, Tiwari S, Barve N *et al.*: **PANNOTATOR: an automated tool for annotation of pan-genomes**. *Genet Mol Res* 2013, **12(3)**:2982-2989.

A tool available for pan-genome analysis.

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

A tool available for pan-genome analysis.

19. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements**. *Genome Res* 2004, **14(7)**:1394-1403.

20. • Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J: **PGAP: pan-genomes analysis pipeline**. *Bioinformatics* 2012, **28(3)**:416-418.

A tool available for pan-genome analysis.

21. • Marcus S, Lee H, Schatz MC: **SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips**. *Bioinformatics* 2014.

A tool available for pan-genome analysis.

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

A tool available for pan-genome analysis.

23. • Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, Yu J, Wu J, Xiao J: **PanGP: a tool for quickly analyzing bacterial pan-genome profile**. *Bioinformatics* 2014, **30(9)**:1297-1299.

A tool available for pan-genome analysis.

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

25. van Tonder AJ, Mistry S, Bray JE, Hill DM, Cody AJ, Farmer CL, Klugman KP, von Gottberg A, Bentley SD, Parkhill J *et al.*: **Defining the estimated core genome of bacterial populations using a Bayesian decision model**. *PLoS Comput Biol* 2014, **10(8)**:e1003788.

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

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

28. Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T, Cronin A, Davis PH, Holroyd NE, Jagels K *et al.*: **Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18**. *PLoS Genet* 2007, **3(2)**:e23.

29. Bentley S: **Sequencing the species pan-genome**. *Nat Rev Microbiol* 2009, **7(4)**:258-259.

30. Daubin V, Ochman H: **Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli***. *Genome Res* 2004, **14(6)**:1036-1042.

31. Mayr E: *Systematics and the Origin of Species*. New York: Columbia University Press; 1942, .

32. Darwin C: *On the Origin of Species by Means of Natural Selection*. London: J. Murray; 1859, .

33. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R: **Microbiology in the post-genomic era**. *Nat Rev Microbiol* 2008, **6(6)**:419-430.

34. Woese CR: **Bacterial evolution**. *Microbiol Rev* 1987, **51(2)**:221-271.

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

36. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies**. *Mol Biol Evol* 2006, **23(2)**:254-267.

37. Doolittle WF: **Phylogenetic classification and the universal tree**. *Science* 1999, **284(5423)**:2124-2129.

38. Doolittle WF: **Lateral genomics**. *Trends Cell Biol* 1999, **9(12)**:M5-M8.

39. Gogarten JP, Townsend JP: **Horizontal gene transfer, genome innovation and evolution**. *Nat Rev Microbiol* 2005, **3(9)**:679-687.

40. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA: **The net of life: reconstructing the microbial phylogenetic network**. *Genome Res* 2005, **15(7)**:954-959.

41. Vernikos GS: **The pyramid of knowledge**. *Nat Rev Microbiol* 2010, **8(2)**:91.

42. Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T, Goesmann A, Joseph B, Konietzny S, Kurzai O *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* 2008, **105(9)**:3473-3478.

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

44. van Schaik W, Top J, Riley DR, Boekhorst J, Vrijenhoek JE, Schapendonk CM, Hendrickx AP, Nijman IJ, Bonten MJ, Tettelin H *et al.*: **Pyrosequencing-based comparative genome analysis of the nosocomial pathogen *Enterococcus faecium* and identification of a large transferable pathogenicity island**. *BMC Genomics* 2010, **11**:239.

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

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

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

48. Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD: **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* 2007, **8(6)**:R103.

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

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

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

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

53. Ahmed A, Earl J, Retchless A, Hillier SL, Rabe LK, Cherpes TL, Powell E, Janto B, Eutsey R, Hiller NL *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* 2012, **194(15)**:3922-3937.

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

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

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

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

58. • Collins RE, Higgs PG: **Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome**. *Mol Biol Evol* 2012, **29(11)**:3413-3425.
A recent example of pan-genome analysis at the bacterial class level.

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