Pangenomics comes of age – from bacteria to plant and animal applications

Agnieszka A. Golicz[1], Philipp E. Bayer[2], Prem L. Bhalla[1], Jacqueline Batley[2], David Edwards[2]

[1]Plant Molecular Biology and Biotechnology Laboratory, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Melbourne, VIC, Australia

[2]School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Crawley, WA, Australia

Correspondence: David Edwards dave.edwards@uwa.edu.au, Agnieszka A. Golicz: agnieszka.golicz@unimelb.edu.au

Key words

Pangenomics, presence/absence variation, tree of life, sequence graph

Abstract

The pangenome refers to a collection of genomic sequence found in the entire species or population rather than in a single individual; the sequence can be core, present in all individuals, or accessory (variable or dispensable), found in a subset of individuals only. While pangenomic studies were first undertaken in bacterial species, developments in genome sequencing and assembly approaches have allowed construction of pangenomes for eukaryotic organisms, fungi, plants and animals, including two large scale human pangenome projects. Analysis of the these pangenomes revealed key differences, most likely stemming from divergent evolutionary histories, but also surprising similarities.

Birth and evolution of the pangenome concept

The early observation which hinted at the existence of gene presence/absence variation within bacterial species came from studies of the genus *Aeromonas*, where a puzzling incongruence between nearly identical 16S rRNA DNA sequence and low levels of DNA:DNA hybridization was observed, suggesting the existence of genomic regions that are not shared by closely related strains [1]. Over a decade later, the finding was supported by advances in sequencing technologies which allowed assembly and comparisons of entire bacterial genomes. The term pangenome (pan, from the Greek word παν, meaning whole) was first used by Tettelin et al. (Figure 1), who compared the genomes of six *Streptococcus agalactiae* strains and noticed a large number of genes which were not shared between isolates. As a result, the term 'pangenome' was introduced to describe the complete gene complement across strains [2]. The concept of the pangenome was soon adopted by plant and animal researchers, resulting in over 20 eukaryotic pangenome studies performed to date (Figure 1, Table 1).

The pangenome is divided into the **core** genome representing genes or sequence found in all individuals, and the accessory genome (also known as **dispensable** or variable) composed of genes or sequence absent from one or more individuals (Figure 2). Bacteria typically possess small genomes, dominated by the presence of coding genes, with relatively little extra-genic and regulatory sequence, hence the study of protein coding gene content was a natural choice for the analysis of bacterial pangenomes. However, as the pangenome studies extended to plants and animals, the use of the term evolved. Eukaryotic genomes have a very different structure to bacterial genomes, and host a much larger proportion of extragenic sequences, many of which are functional, for example complex gene promoters and enhancers. To accommodate the differences between bacterial and eukaryotic genomes, two definitions of the pangenome have evolved. The gene-centric approach defines the pangenome as a union of all genes (or **orthologous gene clusters**), whereas in the sequence-centric approach, the pangenome is defined as the complete, non-redundant set of sequences found in all individuals. Currently, the use of the term is mostly application and context dependant.

Regardless whether the gene-centric or sequence-centric definition is used, the pangenome can be either open or closed (Figure 3). When the pangenome is closed, the sequencing of a sufficient number of individuals would capture almost the entire gene/sequence space, and the theoretical size of the pangenome can be predicted. In contrast, when the pangenome is open, the sequencing of each new individual adds to the pangenome content, making it impossible to

predict final pangenome size. Interestingly the pangenome can be open with respect to the total sequence content but closed with respect to gene number [3]. The majority of pangenomes to date have been built on the species level (with implicit understanding that in some cases, especially in bacteria, species may be difficult to define), but in order to capture maximum diversity, pangenome analysis has been extended to higher taxonomic groups [4, 5] and the breakdown of core/pangenome ratio has even been used to define new species in bacteria [6].

When discussing the concept of the pangenome, it is important to note that although many earlier genomic studies did not aim to build pangenomes, the identification and characterization of structural and gene presence/absence variants has long been a focus of research efforts, now providing insights into our understating of the pangenome [7-9]. In this review we focus predominantly on studies which have been identified as pangenomic by the authors and aimed to estimate the size of the core and accessory genomes.

Pangenomes across the tree of life

Bacterial pangenomes

Pangenome studies in bacteria have the longest history, with hundreds of bacterial species pangenomes reported to date [10] . Some of the largest pangenome studies include more than two thousand isolates, for example, pangenomes of *Escherichia coli* and *Streptococcus pyogenes* [11, 12]. The size of the accessory genome is a function of the number of genomes used for analysis, but varies greatly, even among the well sampled species, with a core genome size range from few to over 80% [10, 13, 14] (Table 1). The size of the core and accessory genomes is also strongly correlated with lifestyle. Bacteria with **sympatric** lifestyles are in contact with other organisms, while **allopatric** species live in isolation (including obligate intracellular bacteria). When comparable numbers of genomes are used for analysis, bacterial species with **sympatric** lifestyles tend to have open pangenomes, with a much lower proportion of core genes, while **allopatric** bacterial species tend to have closed pangenomes, with a smaller number of accessory genes [15]. In bacteria, most accessory genes result from **horizontal gene transfer** (HGT), and sharing environment with other organisms facilitates this gene transfer leading to larger, open pangenomes [15]. Some of the accessory genes confer an adaptive advantage in changing environments [16], while others are responsible for pathogenicity and antibiotic resistance [17-19], however, the overall adaptive or neutral nature of accessory genes in prokaryotes has been a subject to active debate [13, 20-24] (see Box 1).

Bacterial pangenome studies have key practical applications, supporting the selection of antimicrobial targets and vaccine candidates [2]. Finding essential, core genes, which are indispensable to organism's survival, is especially promising for antibiotic development, as protein products of essential genes constitute prime antibiotic drug targets [25]. The analysis of the core genome can also guide selection of vaccine candidates, which correspond to genes found across all strains within a species, often with low sequence variation, boosting chances of antigen recognition post-immunization [11].

Fungal pangenomes

Pangenomes for several fungal species, including the model species *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans*, *Aspergillus fumigatus* [26, 27] and plant pathogens *Parastagonospora spp*. and *Zymoseptoria tritici* [28, 29] have been constructed. The largest fungal study to date investigated 1,011 genomes of *S. cerevisiae*, representing a world-wide genetic and phenotypic diversity panel [27]. The pangenome is composed of 7,796 open reading frames (ORFs) with 63.4% being core. The distribution of accessory genes was biased towards sub-telomeric regions, and gene functional analysis showed strong enrichment for cell-cell interactions, secondary metabolism and stress responses [27]. The core genome was shown to be under stronger selective constraints than the accessory genome [27]. For the other fungal species, the reported proportion of accessory genes/**orthologous gene clusters** (see Box 2 and Table 1) ranged from 60% for *Parastagonospora spp*. (33 genomes analysed across four species) to 9.4% for *C. albicans* (34 genome analysed) (Table 1). Functional analysis of the core genes of the four model species revealed overrepresentation of housekeeping functions, while accessory genes were enriched in processes linked to pathogenesis and microbial resistance. The core genes were also found to be evolutionary older, with up to 40% of accessory genes being duplicates of core genes. **Horizontal gene transfer** (HGT), which is a main driving force shaping bacterial pangenomes, was found to have limited role in fungi, which is consistent with the low frequency of HGT observed in eukaryotes [26]. The accessory genomes of two plant pathogens were enriched in effector genes [28, 29]. Effectors are molecules secreted by pathogens to facilitate host infection. Some of the effectors become avirulence (AVR) proteins, which can be recognized by plant resistance (R) proteins to trigger plant defence response [30]. While the core effector genes are expected to play essential roles in pathogenicity, the varying repertoire of effector and AVR genes correspond to differences in virulence and the ability of the fungus to evade host defence responses. Fungal pathogen pangenome studies can therefore be used track

relationships between isolates of differing virulence and to identify novel genes involved in infection and host response.

Plant pangenomes

In plants, the concept of the pangenome was first introduced in relation to transposable elements (TEs) rather than protein coding genes [31], as observations in maize suggested that the TEs were mostly responsible for the plant accessory genome. Since then, the field of plant pangenomics has expanded significantly. However, most of the recent research efforts have focused on protein coding genes. To date, plant pangenomes have been constructed for more than ten species, including *Brachypodium distachyon*, *Brassica oleracea*, *Brassica napus*, *Capsicum*, Medicago, poplar, rice, soybean, sesame, tomato, sunflower and bread wheat [4, 32-42]. The reported proportion of core genes/**orthologous gene clusters** (see Box 2 and Table 1) ranges from 33% to over 80%. Direct comparisons between plant pangenomes are challenging due to differences in ploidy and pangenome construction strategies applied. However, certain broad trends have been noted, with polyploid and **out-crossing** species being associated with larger accessory genomes [43]. Many plants, especially crop plants, are either ancient or modern polyploids [44]. Throughout evolution, successive rounds of polyploidy and subsequent diploidisation lead to gene redundancy, differential loss, neo- and sub-functionalisation [44, 45], contributing to the accessory gene content. Out-crossing (as opposed to selfing) is a reproductive strategy which involves the mating of two individuals to produce progeny and promotes genetic diversity. Out-crossing plants have larger effective population sizes and lower linkage disequilibrium, which could translate to larger accessory genomes. However, for crop plants, the effects of polyploidy and out-crossing on the accessory gene content are expected to be strongly influenced by artificial selection and breeding history (Table 1).

The functional characterisation of accessory genes in plants repeatedly points to roles in signalling and defence response [33, 34, 41], which confirmed earlier observation of extensive presence/absence variation for disease resistance genes across a range of plant species [46], and correlates well with fungal pathogen accessory genomes being enriched in effector and avirulence genes. The plant pangenome can therefore be used to identify disease resistance genes which may have been lost in elite varieties due to selection during domestication and subsequent improvement [47]. The concept of the pangenome is also linked to heterosis, which refers to superior performance of offspring when crossing two inbred, genetically distant

parents, and complementarity of accessory genomes upon crossing is considered to be one of the potential contributing factors in heterosis [48, 49].

Animal pangenomes

To date, three human pangenome studies have been reported. The first compared the existing reference human genome with newly generated *de novo* assemblies of an Asian and an African genome [50]. The study identified ~5 Mb of novel sequences in each of these assemblies but absent from the reference genome. The researchers also identified 162 human NCBI RefSeq genes that could not be mapped to the reference genome, and 53% of those corresponded to the additional sequence assembled, indicating that the existing refence was missing both coding and non-coding sequence found in the broader human population. Estimates based on analysis of these three genomes suggested that the human pangenome includes an additional ~19–40 Mb of sequence [50]. More recently, two pangenomes using sequence data from 910 individuals of African descent and 275 Han Chinese individuals were constructed. The first, identified 296 Mb of data not present in the GRCh38 reference genome, but did not analyse protein coding genes [51]. The second identified 29.5 Mb of sequence missing from the GRCh38 reference genome, and 188 novel protein coding genes [3]. The amount of novel sequence identified by these three studies differs by an order of magnitude. The difference can most likely be explained by three factors: the number of individuals sequenced, existing differences in the genomic makeup of populations, and methodologies used (‘de novo assembly’ for Chinese Han genomes and concurrent ‘mapping and assembly’ for the genomes of individuals of African descent, see Box 2). Additionally, a large proportion of the sequences reported (~85% of the 296 Mb and ~75% of 29.5 Mb) were almost exclusively made up of simple and satellite repeats [3, 52]. These pangenomes appeared to be open with respect to sequence content (adding data for more individuals will expand the pangenome), and closed with respect to protein coding gene content, suggesting that, similarly to previous findings in plants, the number of human protein coding genes is finite [3, 51].

To our knowledge the only other mammal for which a pangenome has been reported is *Sus scrofa* (pig) [53]. The pangenome was constructed using 12 pig genomes, and 72.5 Mb of novel sequence was found, corresponding to ~3% of the genome. Interestingly, in contrast to the observations in human, the repeat content of the newly assembled pan-sequences was similar to that of the reference pig genome assembly, and newly assembled repeats were more evenly distributed across categories [53]. This study also identified TIG3, an essential regulator of

adipocyte lipolysis which displays presence/absence variation in pig populations, and could contribute to differences in physiology among pigs [53]. A previous study which analysed nine non-reference pig genomes reported 137 Mb of additional sequence harbouring 1,737 genes [54]. These findings underscore the potential impact of genes missing from the reference for clinical and agricultural applications.

Beyond mammals a pangenome of the Mediterranean mussel (*Mytilus galloprovincialis*) has also been reported. Mussels are marine bivalves with wide geographic distribution, high biotic and abiotic stress tolerance and history of lineage specific **whole genome duplication** events with gene content over three times than that reported for human. In total, 25% of genes were found to be accessory and were enriched in functions related to survival and defence response [55]. The results suggest that large scale gene presence/absence variation is likely to be a significant contributor to genomic and likely phenotypic diversity at least for some animal species.

Common and unique features of pangenomes

The results to date show that the concept of the pangenome is applicable to species across the tree of life. There are some compelling features of pangenomes which appear to be shared not only across species, but also across bacteria, fungi, plants and animals. Bacteria have been shown to have the greatest proportion of accessory genome, though more generally, the size of the pangenome and the accessory genome is strongly influenced by the number and genetic diversity of individuals used in the analysis, as well additional factors such as life style (**sympatry** vs **allopatry** for bacteria; **out-crossing** vs selfing for plants), genome evolutionary history, polyploidy and selection.

The core genes appear to be universally over-represented by house-keeping functions, and include genes essential for the life of the organism [25-27, 33, 56], while the accessory genes are often associated with communication, virulence and defence response. Bacterial pangenomes have been used to identify novel defence systems [57], and the accessory genome of gut microbiome bacteria was enriched in defence response and cell signalling [58]. Analysis of the *Klebsiella pneumoniae* pangenome revealed that many of the accessory genes were related to virulence and drug resistance [19]. Similarly, the accessory genomes of fungal plant pathogens carry a high proportion of effectors necessary for infection of the host. Plant accessory genes are in turn over-represented by functions related to signalling and disease resistance as well as abiotic stress response [34, 35, 40, 56]. Functional analysis of human

accessory genes categorized around 30% as belonging to highly variable gene families involved in defence response (mucins, major histocompatibility complex), while the remaining accessory genes had no know functions [50].

In bacteria accessory genes are mostly derived from HGT, whereas in eukaryotes they can mostly be traced to local and **whole genome duplications**, as well as *de novo* gene birth [28, 59, 60]. While accessory genes constitute a substantial proportion of bacterial, fungal and plant pangenomes, mammalian pangenomes are dominated by core genes (up to 96.88% of human genes are reported as core) [3, 53]. However, a substantial proportion (25%) of accessory genes found in Mediterranean mussel demonstrates that the current animal gene space is far from complete, especially for species with high adaptive capabilities, high levels of heterozygosity, high levels of repetitive elements and histories of whole genome duplication events [55].

Frontiers of pangenome research

The pangenome and cis-regulatory elements

Much of the analysis of presence/absence variation has focused on coding regions. However, there is a growing understanding of the importance of **cis-regulatory** and repetitive sequence content in health and disease [61-63] as well as crop domestication and improvement [41, 64]. Continued development of new technologies for regulatory region identification [65-69], brings the non-coding and regulatory regions (**cis-regulatory** elements, CREs) to the forefront of pangenome research. In addition, the increased application of long read sequencing technologies promises the delivery of high-quality, contiguous reference genomes with well resolved repetitive and other non-coding regions [70]. As a result, the classical understanding of the pangenome, which focused on the differences in coding gene content, is being expanded to include the non-coding sequences. The effect of non-coding presence/absence variation on phenotypes is only beginning to be understood, for example, the recent anchoring of accessory pangenome contigs in pig using DNA conformation capture (**Hi-C**) data revealed the existence of variable enhancers [53]. As pangenome methods continue to advance, so will our understanding of the function of accessory non-coding regions of the pangenome.

Understanding core and accessory gene networks

One of the proposed roles of accessory genes is to provide the phenotypic plasticity needed to adapt to the changing environments and new ecological niches. However, in order to perform their function, the accessory genes need to be incorporated into existing biological pathways

and regulatory networks. Previous work in *E. coli* has shown that fine-tuned integration of a horizontally transferred gene into the regulatory network can take millions of years, and requires the evolution of regulatory regions, stabilization of protein-protein interactions and optimisation of codon usage [71]. Subsequent research efforts in *Sinorhizobium fredii* demonstrated that newer pangenome members are less integrated with the core regulatory network [72].

A similar question can be asked of accessory genes in plant and animal systems. How well are these genes incorporated into regulatory networks? Accessory genes are generally under less selective constraint, and the strength of purifying selection is positively correlated with connectivity in co-expression networks, suggesting that the eukaryotic accessory genes are likely to be less networked than the core genes [73]. A better understanding of the steps necessary for the gene to become interfaced with the existing regulatory network is especially relevant in light of potential network and pathways engineering using CRISPR/Cas9 knock-ins [74].

Identification of the essential core genome

One of the key concepts linked to the analysis of the core and accessory genes is the identification of genes which are essential to the organism's survival. However, the relationship between the 'core' status of the gene and its essentiality is not straight forward, as it depends on the genetic background, for example, one accessory gene may compensate for another, but losing both would be lethal, and the environmental conditions, as some genes may only be required under certain environments [25, 75, 76]. A recent study of nine strains of *Pseudomonas aeruginosa* used the core genome to identify essential genes under five growth conditions [25]. In yeast, out of 1,072 essential genes defined in the S288C background, 89% belonged to the core genome and further 7% could be complemented by close orthologues [27].

The identification of essential genes has important implications, not only for our understating of basic biological processes but also for human health, by facilitating the design of anti-microbial and anti-cancer agents [75]. In agriculture, knowledge of essential genes and associated mutations found within breeding populations allows the prevention of livestock loss due to **embryonic lethality** [77]. The identification of essential plant genes can help our understanding of key processes such as photosynthesis [78], and assign potential functions to uncharacterized genes. As research in bacteria and yeast [25, 27] showed that core genes are more likely to be essential, future pangenome studies can be coupled with integrative analysis

of multi-omics datasets, using machine learning and CRISPR/Cas9 knock-outs to identify and characterise the functions of essential genes across the tree of life [75, 79-81].

Moving from genomics to pangenomics

The bias resulting from the use of a single reference genome on genomic analysis cannot be ignored. A recent study in maize showed that the choice of reference genome has a major impact on gene expression quantification and genome-wide association study (GWAS) results [82]. Accessory genomic regions associated with important traits in some individuals may be completely missing from the reference sequence and therefore be inadvertently excluded from association studies. The incorporation of pangenome sequence and alternative alleles in re-sequencing analyses is therefore important to improve the accuracy of trait association analysis. The abundance of non-coding disease associated variants in humans [83], and the prevalence of non-coding variable regions in the human pangenome, suggests that similar factors are important for human trait association studies.

Adopting a pangenome reference in genomic studies is non-trivial due to requirement for suitable data structures. The simplest approach would be to add alternative alleles and accessory sequences to the existing reference sequence. This strategy has already been adopted in plant research and has been shown to improve the accuracy of single nucleotide polymorphism (SNP) calling [34]. Using the entire pig pangenome as a reference was also shown to improve the accuracy of read mapping in this species [53]. However, such a linear representation of the genomic sequence is not optimal, as it does not preserve contiguity of the pangenome sequence, nor account for the subtleties of sequence presence/absence variation, for example, where some sequence combinations are always found together while others are mutually exclusive. A better approach would be to use a pangenome graph, which captures links and relationships between pangenome sequences [84-86]. To date, pangenome graphs capable of storing human-sized references have mostly been used for sequence read mapping and variant genotyping [87-89]. However, new capabilities for genome graphs are rapidly developing, for example with gene prediction being applied directly to the assembly graph rather than the linear representation of chromosomes [90].

Beyond improved variant identification and genotyping, the adoption of the pangenome as a reference will also allow inclusion of variants other than SNPs in genome wide association studies. Although currently the practice is not common, studies in both plants and human suggest that the inclusion of structural variants in association studies could help identify causal

variants [91, 92]. The inclusion of sequence presence/absence variation contributed to identification of missing QTLs associated with disease resistance in oilseed rape [93]. In *S. cerevisiae*, when both SNPs and CNVs were used for association studies, CNVs explained higher percentage of variance [27]. Taken together, the findings underscore the importance of using different variant classes in association studies.

Concluding Remarks and Future Perspectives

The last decade of genomic research was characterized by rapidly decreasing DNA sequencing costs, advanced bioinformatics tools and use of high-performance computational infrastructure, allowing for the generation of high-quality reference sequences, the re-sequencing of numerous diverse individuals, and variant identification. This data has provided valuable insights into population-level diversity, the genetics of health and disease, as well as an improved understanding of key agronomic traits in plants and animals. As our knowledge of genomic variation increased, it become apparent that a single reference sequence is insufficient to represent the extent of genomic variation found within species, resulting in the introduction and growth of the pangenome concept. In the coming decade the application of pangenomes will become commonplace likely making the single reference approach to genomic analysis obsolete. In addition, parallels in pangenomics across the tree of life present and excellent resource for interdisciplinary studies. With time, extending pangenomic studies to higher taxonomic groups will provide the resources necessary to study the combinatorial differences in genomic content between organisms, supporting further characterisation of genes, their evolutionary history and function and fuelling developments in other fields, such as synthetic biology, which strives to identify the 'minimal genome' required to support robust cellular life.

Glossary

Allopatric bacteria – live isolated from other microorganisms (for example, obligate intracellular bacteria)

Cis-regulatory element (CRE) – is a section of non-coding DNA, which regulates transcription of neighbouring genes. CREs can be proximal (promoters) or distal (enhancers/silencers)

Core genome – genes/DNA sequence found in all the individuals under study

Dispensable genome – genes/DNA sequence found in some individuals but not others. The term has increasingly been replaced by "accessory genome" or "variable genome" as sequence

which is dispensable for one individual may be important for another due to differences in total gene/sequence content (genetic background) and environment

Embryonic lethality – death within the embryonic period of development

Genome duplication – is a process in which additional copies of the entire genomes are generated. The resulting cells are polyploid – contain more than two copies of chromosomes

Hi-C – a conformation capture method used to study spatial organization of chromatin within a cell. Can be used to identify promoter-enhancer interactions

Homologous (Orthologous) gene cluster – clusters of genes in different species that are related by descent, originating from a single ancestral gene

Horizontal gene transfer (HGT) – movement of genetic material by means other than descent (not from parent to offspring). HGT allows sharing of genetic material between unrelated organisms and is especially prevalent in prokaryotes

Outcrossing – reproductive practice which requires crossing of two unrelated individuals to produce progeny. Outcrossing increases genetic diversity

Sympatric bacteria – interact with many other bacteria, often belonging to different phyla, allowing them to exchange genes

Synteny – conservation of gene order across chromosomes reflecting ancestral gene order

Box 1

Is the accessory genome adaptive or neutral?

The question of the overall adaptive, neutral or even deleterious nature of the accessory genome remains open, for both bacteria and eukaryotes. Recent applications of evolutionary theory to study bacterial pangenomes arrived at contrasting conclusions. On the one hand, modelling work shows that gene acquisition is largely an adaptive process [22], with niche adaptation as one of prime examples [23]. On the other hand, it was suggested that the larger, more fluid pangenome is reflective of a larger effective population size and mostly dictated by neutral evolution [24, 94].

The question is further complicated in large eukaryotic genomes by the presence of additional factors including homologous relationships between genes, allowing for functional redundancy, and extensive linkage disequilibrium resulting in inheritance of entire haplotypes,

possibly containing multiple accessory genes of unrelated function. Some accessory genes can be considered adaptive, having been associated with important traits including biotic and abiotic stress response and flowering time in plants [43]. The over-representation of effectors and disease resistance genes within the accessory genes of plant pathogens and crop plants appear to be prime examples of their adaptive role in infection and defence response. It should also be noted that, in parallel to essentiality, whether the genomic region is considered adaptive, deleterious or neutral can be a function of time and environment. In addition, the role of many accessory genes remains elusive, especially since they appear to evolve under reduced evolutionary constraints and have overall lower gene expression levels [27, 28, 32, 33]. Even for the relatively small bacterial genomes, the function of many accessory genes remains unknown, with variable sequences derived from mobile genetic elements making inference about the role of the accessory genome challenging [11, 13, 15]. Previously methodologies have been developed, aiming to link properties of genes such as expression level, connectivity in biological networks, and sequence conservation to their function [79, 81, 95]. Similar approaches could be adopted to understand which accessory genes are likely to be functional and have phenotypic effects.

Box 2

Pangenome analysis methods

Two methodologies ('*de novo* assembly' and 'mapping and assembly') are wildly in used in pangenome studies.

De novo genome assembly

Genomes are assembled *de novo* and annotated individually, then comparisons are made using whole genome alignments and gene orthology detection tools (using orthologous gene clustering or **synteny**) [3, 26, 28, 33, 39, 96]. This approach allows recovery of full-length genomes of all individuals and has a potential to resolve repetitive regions and copy number variants, but also suffers from several shortcomings. *De novo* assemblies require the generation of large, often expensive datasets, and technical errors and variation in assembly and annotation can result in spurious presence/absence variation calls. In addition, whole genome comparisons (identification of core and variable sequences using whole-genome alignments) are often de-coupled from gene level analysis, usually performed using sequence based orthologous gene clustering, and which can be error-prone, especially for highly duplicated genomes. Expected

technological improvements along with the reducing cost and increasing accuracy of long read sequencing will allow for the production of higher quality chromosome-level assemblies, and the replacement of sequence-only based orthologous gene detection by more accurate synteny-based approaches.

Read mapping and assembly

The read mapping and assembly approach was developed to accommodate significant quantities of publicly available short read sequencing datasets which are not of a scale required for accurate *de novo* genome assembly. However, these short reads can still be used for pangenome construction [34, 35, 40, 51]. In the 'iterative mapping and assembly' approach [34] reads from individuals are mapped to the reference, with the unmapped reads subsequently assembled and added to the growing pangenome reference. The pangenome reference is then used to call gene presence/absence variation by re-mapping the reads from all individuals to the reference and examining coverage of each gene for each individual. In the 'concurrent mapping and assembly' protocol reads are mapped and assembled in parallel and the resulting assemblies are processed to remove redundancy [51]. 'Mapping and assembly' has the advantage of calling presence/absence at every gene locus without the need for orthologous gene clustering, and given the low cost of data generation, is suitable for the analysis of presence/absence variation across large populations of individuals. However, the placement of the newly identified genes within the genome is not always possible. Given the complementary benefits and limitations of the *de novo* and read mapping approaches, a thorough pangenome study of a species would ideally include a combination of both.

Figures

Figure 1. **Timeline of developments in pangenomic research.** The term was first introduced in 2005 by Tettelin at al., but the concept was quickly taken up by plant, and then human researchers. To date, over 20 eukaryotic pangenomes have been constructed for organisms with genome sizes ranging from 12 Mb (baker's yeast) to 17 Gb (bread wheat).

Figure 2. **Construction of pangenome using diverse genotypes.** Comparisons can be performed using whole genomes, coding genes or both.

Figure 3. **Pangenome size as a function of the number of individuals used in the analysis.** For closed pangenomes, the theoretical size of the pangenome can be predicted (dashed line). When the pangenome is open, its size increases indefinitely with each added individual.

Bacterial pangenomes can be either open or closed depending on the lifestyle and the extent of horizontal gene transfer (HGT); extensive HGT being associated with open pangenomes. Eukaryotic pangenomes are expected to be closed with respect to coding gene number but can be open when total sequence content is considered.

Tables

Table 1. Summary of selected prokaryotic and the eukaryotic pangenomes reported to date. a – only additional accessory sequence/genes were reported; b – pan-transcriptome analysis; HGC – homologous (orthologous) gene clustering. The allopatric and sympatric categorization of bacterial species was obtained from [15]. Outcrossing rates were obtained from [43].

	Species	**Ploidy**	**~Genome size**	**Number of individuals**	**Core genome (% genes/gene clusters)**	**Core gene/gene cluster detection method**	**Core genome (% sequence)**	**Additional information**	**Ref.**
Bacteria	*Chlamydia trachomatis*	Haploid	1.04 Mb	85	80	HGC	N/A	Allopatric	[10]
	Rickettsia prowazekii	Haploid	1.1 Mb	10	8	HGC	N/A	Allopatric	[10]
	Mycobacterium tuberculosis	Haploid	4.4 Mb	168	78	HGC	N/A	Allopatric	[10]
	Yersinia pestis	Haploid	4.7 Mb	36	56	HGC	N/A	Allopatric	[10]
	Bacillus anthracis	Haploid	5.2 Mb	50	51	HGC	N/A	Allopatric	[10]
	Streptococcus agalactiae	Haploid	2.2 Mb	57	28	HGC	N/A	Sympatric	[10]
	Streptococcus pneumoniae	Haploid	2.2 Mb	52	31	HGC	N/A	Sympatric	[10]
	Haemophilus influenzae	Haploid	1.8 Mb	55	33	HGC	N/A	Sympatric	[10]
	Escherichia coli	Haploid	4.6 Mb	633	8	HGC	N/A	Sympatric	[10]
	Clostridium botulinum	Haploid	3.9 Mb	46	5	HGC	N/A	Sympatric	[10]
	Prochlorococcus marinus	Haploid	1.8 Mb	10	18	HGC	N/A	Sympatric	[10]
Fungi	*Saccharomyces cerevisiae* (baker's yeast)	Mostly diploid	12 Mb	1,011	63	ORF sequence similarity	N/A	Single cell organism Domesticated and wild	[27]

	Candida albicans	Diploid	15 Mb	34	91	Synteny	N/A	Single cell organism Animal commensal	[26]
	Cryptococcus neoformans	Haploid	19 MB	25	81	Synteny	N/A	Single cell soil organism	[26]
	Aspergillus fumigatus	Haploid	29 Mb	12	83	Synteny	N/A	Environmental filamentous fungus	[26]
	Parastagonospora spp	Haploid	75 Mb	33	~40	HGC	N/A	Plant pathogen	[29]
	Zymoseptoria tritici	Haploid	40 Mb	5	58	HGC	N/A	Plant pathogen	[28]
Protists	*Emiliania huxleyi*	Diploid	142 Mb	14	<70	Read mapping	N/A	Single cell marine phytoplankton Wold-wide distribution	[97]
Plants	*Brachypodium distachyon* (stiff brome)	Diploid	355 Mb	54	35	HGC	N/A	~5% outcrossing Not domesticated	[33]
	Brassica oleracea	Diploid	650 Mb	10	81	Read mapping	N/A	>95% outcrossing Crop plant	[34]
	Brassica napus (oil seed rape)	Tetraploid	1.1 Gb	53	62	Read mapping	N/A	~30% outcrossing Crop plant	[35]
	Brassica rapa	Diploid	490Mb	3	87	Synteny	N/A	>95% outcrossing Crop plant	[96]
	Capsicum (pepper)	Diploid	3.5 Gb	383	56	Read mapping	N/A	10%-90% outcrossing Crop plant	[36]

	Glycine soya (soybean)	Tetraploid (diploidized)	1 Gb	7	49	HGC	80	2%-20% outcrossing Wild crop relative	[39]
	Helianthus annuus (sunflower)	Diploid	3 Gb	493	83	Sequence similarity	N/A	Mostly self-incompatible (modern varieties self-compatible)	[42]
	Medicago truncatula	Diploid	465 Mb	15	33	HGC	58	1%-35% outcrossing Non-crop legume	[37]
	Oryza sativa (rice)	Diploid	430 Mb	3	92	Intersection of gene coordinates	89	1%-2% outcrossing Crop plant	[98]
	O. sativa[a] (indica/japonica)	Diploid	430 Mb	1483	8,991/6,366 additional genes found	N/A	52,976/30,349 additional contigs found	1%-2% outcrossing Crop plant	[99]
	O. sativa	Diploid	430 Mb	3,010	54-62	Read mapping and HGC	N/A	1%-2% outcrossing Crop plant	[56]
	O. sativa/O. rifupogon	Diploid	430 Mb	67	62	Sequence similarity	N/A	1-2%/10%-50% outcrossing Crop plant	[38]
	Populus (poplar)	Diploid	500 Mb	7	<90	Read mapping	81	Genus of trees	[100]
	Sesamum indicum (sesame)	Diploid	350 Mb	5	58	HGC	N/A	4%-23% Orphan crop	[60]
	Solanum lycopersicum (tomato)	Diploid	950 Mb	725	74	Read mapping	N/A	0%-5% Crop plant	[41]
	Triticum aestivum (bread wheat)	Hexaploid	17 Gb	19	64	Read mapping	N/A	<1% outcrossing	[40]

								Crop plant	
	Zea mays[b] (maize)	Tetraploid (diploidized)	2.4 Gb	503	39	Read mapping	N/A	~95% outcrossing Crop plant	[101]
	Mytilus galloprovincialis (mussel)	Diploid	1.6 Gb	16	75	Read mapping	N/A	Marine filter-feeder	[55]
Animals	*Homo sapiens*[a] (human)	Diploid	3.2 Gb	2	86 additional genes found	N/A	19-40 Mb additional sequence predicted	Asian and African genomes	[50]
	H. sapiens[a]	Diploid	3.2 Gb	910	N/A	N/A	296 Mb additional sequence found	Individuals of African descent	[51]
	H. sapiens	Diploid	3.2 Gb	185	97	Read mapping	N/A	Han Chinese Individuals	[3]
	Sus scrofa[a] (pig)	Diploid	2.7 Gb	9	1,737 additional genes found	N/A	137 Mb of additional sequence found	European and Chinese breeds	[54]
	S scrofa[a] (pig)	Diploid	2.7 Gb	12	N/A	N/A	72.5 Mb additional sequence found	European and Chinese breeds	[53]

References

1. Martinez-Murcia, A.J. et al. (1992) Phylogenetic interrelationships of members of the genera Aeromonas and Plesiomonas as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations. International Journal of Systematic and Evolutionary Microbiology 42 (3), 412-421.
2. Tettelin, H. et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America 102 (39), 13950-13955.
3. Duan, Z. et al. (2019) HUPAN: a pan-genome analysis pipeline for human genomes. Genome Biology 20 (1), 149.
4. Zhang, B. et al. (2019) The poplar pangenome provides insights into the evolutionary history of the genus. Commun Biol 2, 215.
5. Lapierre, P. and Gogarten, J.P. (2009) Estimating the size of the bacterial pan-genome. Trends in Genetics 25 (3), 107-110.
6. Caputo, A. et al. (2019) Genome and pan-genome analysis to classify emerging bacteria. Biology Direct 14 (1), 5.
7. McCarroll, S.A. et al. (2006) Common deletion polymorphisms in the human genome. Nature Genetics 38 (1), 86-92.
8. Saxena, R.K. et al. (2014) Structural variations in plant genomes. Briefings in Functional Genomics 13 (4), 296-307.
9. Golicz, A.A. et al. (2016) Towards plant pangenomics. Plant Biotechnology Journal 14 (4), 1099-1105.
10. Ding, W. et al. (2017) panX: pan-genome analysis and exploration. Nucleic Acids Research 46 (1), e5-e5.
11. Davies, M.R. et al. (2019) Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. Nat Genet 51 (6), 1035-1043.
12. Land, M. et al. (2015) Insights from 20 years of bacterial genome sequencing. Functional & integrative genomics 15 (2), 141-161.
13. McInerney, J.O. et al. (2017) Why prokaryotes have pangenomes. Nat Microbiol 2, 17040.
14. Freschi, L. et al. (2018) The Pseudomonas aeruginosa pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. Genome Biology and Evolution 11 (1), 109-120.
15. Rouli, L. et al. (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect 7, 72-85.
16. Vernikos, G. et al. (2015) Ten years of pan-genome analyses. Current Opinion in Microbiology 23, 148-154.
17. Obert, C. et al. (2006) Identification of a Candidate Streptococcus pneumoniae core genome and regions of diversity correlated with invasive Pneumococcal disease. Infection and Immunity 74 (8), 4766.
18. Rasko, D.A. et al. (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. Journal of bacteriology 190 (20), 6881-6893.
19. Holt, K.E. et al. (2015) Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. Proceedings of the National Academy of Sciences of the United States of America 112 (27), E3574-E3581.
20. Vos, M. and Eyre-Walker, A. (2017) Are pangenomes adaptive or not? Nat Microbiol 2 (12), 1576.

21. Shapiro, B.J. (2017) The population genetics of pangenomes. Nat Microbiol 2 (12), 1574.
22. Sela, I. et al. (2016) Theory of prokaryotic genome evolution. Proceedings of the National Academy of Sciences 113 (41), 11399.
23. Niehus, R. et al. (2015) Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nature Communications 6, 8924.
24. Andreani, N.A. et al. (2017) Prokaryote genome fluidity is dependent on effective population size. The Isme Journal 11, 1719.
25. Poulsen, B.E. et al. (2019) Defining the core essential genome of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A 116 (20), 10072-10080.
26. McCarthy, C.G.P. and Fitzpatrick, D.A. (2019) Pan-genome analyses of model fungal species. Microbial Genomics 5 (2).
27. Peter, J. et al. (2018) Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature 556 (7701), 339-344.
28. Plissonneau, C. et al. (2018) Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. BMC Biology 16 (1), 5.
29. Syme, R.A. et al. (2018) Pan-Parastagonospora comparative genome analysis—effector prediction and genome evolution. Genome Biology and Evolution 10 (9), 2443-2457.
30. Petit-Houdenot, Y. and Fudal, I. (2017) Complex interactions between fungal avirulence genes and their corresponding plant resistance genes and consequences for disease resistance management. Frontiers in plant science 8, 1072-1072.
31. Morgante, M. et al. (2007) Transposable elements and the plant pan-genomes. Current Opinion in Plant Biology 10 (2), 149-155.
32. Contreras-Moreira, B. et al. (2017) Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. Frontiers in Plant Science 8, 184.
33. Gordon, S.P. et al. (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. Nature Communications 8 (1).
34. Golicz, A.A. et al. (2016) The pangenome of an agronomically important crop plant Brassica oleracea. Nature Communications 7.
35. Hurgobin, B. et al. (2018) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnology Journal 16 (7), 1265-1274.
36. Ou, L. et al. (2018) Pan-genome of cultivated pepper (Capsicum) and its use in gene presence–absence variation analyses. New Phytologist 220 (2), 360-363.
37. Zhou, P. et al. (2017) Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. BMC Genomics 18 (1), 261.
38. Zhao, Q. et al. (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nature Genetics 50 (2), 278-284.
39. Li, Y.H. et al. (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nature Biotechnology 32 (10), 1045-1052.
40. Montenegro, J.D. et al. (2017) The pangenome of hexaploid bread wheat. Plant Journal 90 (5), 1007-1013.
41. Gao, L. et al. (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nature Genetics 51 (6), 1044-1051.
42. Hübner, S. et al. (2019) Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nature Plants 5 (1), 54-62.
43. Tao, Y. et al. (2019) Exploring and exploiting pan-genomics for crop improvement. Molecular Plant 12 (2), 156-169.

44. Salman-Minkov, A. et al. (2016) Whole-genome duplication as a key factor in crop domestication. Nature Plants 2, 16115.
45. Jiao, Y. and Paterson, A.H. (2014) Polyploidy-associated genome modifications during land plant evolution. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 369 (1648), 20130355.
46. Dolatabadian, A. et al. (2017) Copy number variation and disease resistance in plants. Theoretical and Applied Genetics 130 (12), 2479-2490.
47. Bayer, P.E. et al. (2019) Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. Plant Biotechnology Journal 17 (4), 789-800.
48. Zhang, J. et al. (2016) Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. Proceedings of the National Academy of Sciences 113 (35), E5163.
49. Lai, J. et al. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nature Genetics 42 (11), 1027-1030.
50. Li, R. et al. (2009) Building the sequence map of the human pan-genome. Nature Biotechnology 28, 57.
51. Sherman, R.M. et al. (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nature Genetics 51 (1), 30-35.
52. Miga, H.K. (2019) Centromeric satellite DNAs: hidden sequence variation in the human population. Genes 10 (5).
53. Tian, X. et al. (2019) Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. Science China Life Sciences.
54. Li, M. et al. (2017) Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. Genome research 27 (5), 865-874.
55. Gerdol, M. et al. (2019) Massive gene presence/absence variation in the mussel genome as an adaptive strategy: first evidence of a pan-genome in Metazoa. bioRxiv, 781377.
56. Wang, W. et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557 (7703), 43-49.
57. Doron, S. et al. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. Science (New York, N.Y.) 359 (6379), eaar4120.
58. Zou, Y. et al. (2019) 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nature Biotechnology 37 (2), 179-185.
59. Zhang, L. et al. (2019) Rapid evolution of protein diversity by de novo origination in Oryza. Nature Ecology & Evolution 3 (4), 679-690.
60. Yu, J. et al. (2019) Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnology Journal 17 (5), 881-892.
61. Hannan, A.J. (2018) Tandem repeats mediating genetic plasticity in health and disease. Nature Reviews Genetics 19, 286.
62. Epstein, D.J. (2009) Cis-regulatory mutations in human disease. Briefings in Functional Genomics 8 (4), 310-316.
63. Gao, L. et al. (2018) Identifying noncoding risk variants using disease-relevant gene regulatory networks. Nature Communications 9 (1), 702.
64. Swinnen, G. et al. (2016) Lessons from domestication: targeting cis-regulatory elements for crop improvement. Trends in Plant Science 21 (6), 506-515.
65. Weber, B. et al. (2016) Plant enhancers: a call for discovery. Trends in Plant Science 21 (11), 974-987.
66. Wang, J. et al. (2018) HACER: an atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Research 47 (D1), D106-D112.

67. Maher, K.A. et al. (2018) Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. The Plant cell 30 (1), 15-36.
68. Ron, G. et al. (2017) Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nature communications 8 (1), 2237-2237.
69. Fullard, J.F. et al. (2018) An atlas of chromatin accessibility in the adult human brain. Genome research 28 (8), 1243-1252.
70. van Dijk, E.L. et al. (2018) The third revolution in sequencing technology. Trends in Genetics 34 (9), 666-681.
71. Lercher, M.J. and Pál, C. (2007) Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Molecular Biology and Evolution 25 (3), 559-567.
72. Jiao, J. et al. (2018) Coordinated regulation of core and accessory genes in the multipartite genome of Sinorhizobium fredii. PLOS Genetics 14 (5), e1007428.
73. Mähler, N. et al. (2017) Gene co-expression network connectivity is an important determinant of selective constraint. PLOS Genetics 13 (4), e1006402.
74. Ding, Y. et al. (2016) Recent advances in genome editing using CRISPR/Cas9. Frontiers in plant science 7, 703-703.
75. Rancati, G. et al. (2017) Emerging and evolving concepts in gene essentiality. Nature Reviews Genetics 19, 34.
76. Marroni, F. et al. (2014) Structural variation and genome complexity: is dispensable really dispensable? Current Opinion in Plant Biology 18, 31-36.
77. Derks, M.F.L. et al. (2019) Loss of function mutations in essential genes cause embryonic lethality in pigs. PLOS Genetics 15 (3), e1008055.
78. Rubin, B.E. et al. (2015) The essential gene set of a photosynthetic organism. Proceedings of the National Academy of Sciences 112 (48), E6634.
79. Lloyd, J.P. et al. (2015) Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes. The Plant cell 27 (8), 2133-2147.
80. Minkenberg, B. et al. (2017) Discovery of rice essential genes by characterizing a CRISPR-edited mutation of closely related rice MAP kinase genes. The Plant Journal 89 (3), 636-648.
81. Chen, H. et al. (2019) New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. Briefings in Bioinformatics.
82. Gage, J.L. et al. (2019) Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. The Plant Genome 12 (2).
83. Zhang, F. and Lupski, J.R. (2015) Non-coding genetic variants in human disease. Human Molecular Genetics 24 (R1), R102-R110.
84. Paten, B. et al. (2017) Genome graphs and the evolution of genome inference. Genome research 27 (5), 665-676.
85. Marcus, S. et al. (2014) SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics 30 (24), 3476-3483.
86. Computational Pan-Genomics, C. (2018) Computational pan-genomics: status, promises and challenges. Briefings in bioinformatics 19 (1), 118-135.
87. Eggertsson, H.P. et al. (2017) Graphtyper enables population-scale genotyping using pangenome graphs. Nat Genet 49 (11), 1654-1660.
88. Kim, D. et al. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology 37 (8), 907-915.

89. Garrison, E. et al. (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol 36 (9), 875-879.
90. Shlemov, A. and Korobeynikov, A. (2019) PathRacer: racing profile HMM paths on assembly graph. bioRxiv, 562579.
91. Chiang, C. et al. (2017) The impact of structural variation on human gene expression. Nature Genetics 49, 692.
92. Fuentes, R.R. et al. (2019) Structural variants in 3000 rice genomes. Genome Research 29 (5), 870-880.
93. Gabur, I. et al. (2018) Finding invisible quantitative trait loci with missing data. Plant Biotechnology Journal 16 (12), 2102-2112.
94. Bobay, L.-M. and Ochman, H. (2018) Factors driving effective population size and pan-genome evolution in bacteria. BMC Evolutionary Biology 18 (1), 153.
95. Golicz, A.A. et al. (2018) MCRiceRepGP: a framework for the identification of genes associated with sexual reproduction in rice. The Plant Journal 96 (1), 188-202.
96. Lin, K. et al. (2014) Beyond genomic variation - comparison and functional annotation of three Brassica rapa genomes: A turnip, a rapid cycling and a Chinese cabbage. BMC Genomics 15 (1).
97. Read, B.A. et al. (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499, 209.
98. Schatz, M.C. et al. (2014) Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. Genome biology 15 (11), 506.
99. Yao, W. et al. (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biology 16 (1), 187.
100. Pinosio, S. et al. (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. Molecular Biology and Evolution 33 (10), 2706-2719.
101. Hirsch, C.N. et al. (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26 (1), 121-135.