Plant Pangenome: Impacts On Phenotypes And Evolution

Christine Tranchant-Dubreuil, Mathieu Rouard, Francois Sabot

▶ **To cite this version:**

Christine Tranchant-Dubreuil, Mathieu Rouard, Francois Sabot. Plant Pangenome: Impacts On Phenotypes And Evolution. Annual Plant Reviews, 2019, 10.1002/9781119312994.apr0664 . hal-02053647

HAL Id: hal-02053647
https://hal.science/hal-02053647

Submitted on 1 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Plant Pangenome: Impacts On Phenotypes And Evolution

Christine Tranchant-Dubreuil[1,3], Mathieu Rouard[2,3], and Francois Sabot[1,3]

[1]DIADE University of Montpellier, IRD, 911 Avenue Agropolis, 34934 Montpellier Cedex 5, France
[2]Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France
[3]South Green Bioinformatics Platform, Bioversity, CIRAD, INRA, IRD, Montpellier, France

With the emergence of low-cost high-throughput sequencing technologies, numerous studies have shown that a single genome is not enough to identify all the genes present in a species. Recently, the pangenome concept has become widely used to investigate genome composition of a collection of individuals. The pangenome consists in the core genome, which encompasses all the sequences shared by all the individuals, and the dispensable genome, composed of sequences present in only some individuals. Pangenomic analyses open new ways to investigate and compare multiple genomes of closely related individuals at once, and more broadly new opportunities for optimizing breeding and studying evolution. This emerging concept combined with the power of the third-generation sequencing technologies gives unprecedented opportunities to uncover new genes, to fully explore genetic diversity and to advance knowledge about the evolutionary forces that shape genome organization and dynamics.

Pangenome, Gene Diversity, Adaptation, Evolution, Population Genomics, Structural Variation

Correspondence: ***christine.tranchant@ird.fr, m.rouard@cgiar.org, francois.sabot@ird.fr***

Introduction

Revolutionary advances in high-throughput sequencing technology during the last two decades have offered new ways to study genome diversity and evolution. Limited initially to a few reference genomes, current capabilities allow sequencing and analysis of multiple genomes of closely related species. Indeed, for years genomic studies typically used a reference-centric approach, which relied mainly on the expensive and low-throughput Sanger sequencing, limiting large-scale population studies to a few loci, or to markers such as simple sequence repeats (SSRs) (1, 2). Since the advent of Next-Generation Sequencing (NGS), a transition has occurred from a single-genome/species to multiple-genomes/species analysis. The data deluge produced by these NGS data revealed that individuals from the same species do not systematically share the same genetic content (3).

Genetic Diversity and Structural Variations. Many genetic diversity studies have focused on single-nucleotide polymorphisms (SNPs) as the main source of genetic variation (4–7). However, larger structural variations (SVs), including copy number variation (CNV) and presence/absence variation (PAV), have been shown to play a major role on genetic variability, and are thought to contribute to phenotypic variations (3, 8, 9). Moreover, even if there are variations within genes, such as SNPs or small insertions or deletions (InDels), several studies showed that all the genes from a given species are not obtained using a single genome (10–12). In plants, evidence first from maize (13, 14) showed that only half of the genomic structure is conserved between two individuals. Similarly, a study of 18 wheat cultivars revealed the absence of 12,150 genes from the reference genome (15). Previous studies performed on rice showed that genes absent from the Asian rice *Oryza sativa japonica* subspecies are present in other rice varieties (16) and confer tolerance to submergence (*Submergence 1, Sub1*) (17), deep water (*SNORKEL1, SNORKEL2*) (18) or low-phosphorus soils (*Phosphorus-Starvation Tolerance, Pstol1*) (19). In the same species, Yao et al. (20) highlighted that 41.6 % of trait-associated SNPs (from GBS markers) were not present in the reference genome sequence. In the wild *Brachypodium distachyon*, the flowering time divergence is directly linked to SV and pangenomic variations (21).

Origin of the Pangenome Concept. Studies on bacteria benefited earlier by the NGS potential due to their small genome size and large populations, and gave rise to the Pangenome concept, first introduced by Tettelin et al in 2005 (22, 23), to refer to the full genomic content of a species. The pangenome was first defined to consist of the core genome shared among all individuals and the dispensable genome, shared only between some individuals. In plants as in bacteria, the dispensable genome turns out not to be so "dispensable" (24), and encompasses a large portion of structural variants that affect a large number of genes. The dispensable genome may contribute to phenotypic trait diversity (9) such as biotic resistance, organ size or flowering time (21) and may play a role in adaptation to various environments. The pangenome view of the genome opens new ways to investigate diversity, adaptation and evolution with strong impacts on the species concept itself.

What is a Pangenome ?

Since the pangenome concept was first proposed (22, 23), definitions and objectives fluctuated between various interpretations including (*i*) the total number of non-redundant genes that are present in a given dataset (25), (*ii*) the full gene repertoire of a species (26), (*iii*) the result of genomic comparison of different organisms of the same species or genus (27, 28), (*iv*) the similarity-based representation of the total set of genes, which are present in a group of closely related

species or strains of a single species (29) or (*v*) the sum of the genes of all living organisms, viruses, and different mobile genetic elements (30).

Different Ways to Define a Pangenome. These multiple definitions highlight the flexibility of the pangenome concept and the levels of granularity possible in relation to taxonomy (genus, species, subspecies) or composition of the core genome (e.g. single copy genes versus CNV, gene and non-genic).
Here, we propose to define the pangenome in two different ways: a function-based and a structure-based. Whatever definition is used, the considered group can be a species, a subspecies, a genera or even a family. Thus, the limits of a specific pangenome will change if the referential group changes.

Function-Based Definition. The function-based definition states that the pangenome is the sum of all genes within a given set of individuals; it can be extended at the gene family level, as in (31). This is similar to the definition used in bacteria, and relies on the identification of gene clusters (genes with close phylogenetic relationships, that may be scattered all along the genome). Highly similar sequences (e.g. recent paralogs) may be considered as the same sequence. Once all gene clusters per individual are identified, the presence/absence for each gene is scored, wherever the location. In this context, if at least one member of a gene family is present in each individual (whatever is the sequence itself), the function belongs to the core genome. Such a definition is gene-centric and does not take into account transposable elements or non-canonical genes (e.g. tRNA, miRNA). Most of the current pangenome analyses use this definition.

Structure-Based Definition. The structure-based definition states that the pangenome is defined as the complete set of non-redundant sequences approximately 100 base pairs (bp) in length or more (except for few SNP and InDels, see below) within a given group of individuals. The advantage of this definition is that it allows both genes and non-genic sequences to be taken into account. However, this definition may be difficult to apply when dealing with copies of transposable elements (TE; see below). Sequences of 100 bp can be identified and annotated with few ambiguities (e.g. the size of a small TE, miRNA locus or tRNA gene). The presence or absence of a sequence here is purely position-based. Thus, in the case of a recent duplication followed by an alternative deletion (*i.e.* individual A conserves A copy and individual B the B copy; see **Figure 1**), none of the copies are in the core genome. In the same way, genomic recombination in a portion of the population can change the location of a given region, and thus will not be included in the core genome. Transposition of a Class I (Copy-and-Paste) or of a Class II TE (Cut-and-Paste) (32) will also change the core genome content. Indeed, more and more studies show that the position of these events (recombination and transposition) will impact the expression of adjacent genes (33). Thus, the location of a given sequence may modify its impact on the phenotype, in addition to selection and evolution.

The Different Compartments of the Pangenome.

The Core Genome. The core genome is the common set of sequences shared by all individuals of the group, and is generally described as the minimal genome sequence required for a cell to live. Indeed, the core genome has been shown to include the main essential gene functions: (*i*) Maintenance of the basic functions of the organism which include DNA replication, translation and maintenance of cellular homeostasis (22), and (*ii*) Essential cellular processes (e.g. glycolysis) (21).
However, some authors (e.g. (34)) proposed that the core genome consists of two sub-compartments, one essential and the other 'persistent'. The persistent core genome sub-compartment includes genes or sequences that were perhaps necessary at one time in the life history of an organism, but have lost their necessity and have not yet been removed by the genetic drift.

The Dispensable Genome. An unexpectedly large number of sequences, including a surprising number of genes, belong to the dispensable genome (11). Thus, PAVs were identified within 38% of genes in the *Brassica napus* pangenome (35). Similarly, Zhao et al. identified 10,872 novel genes (absent from the reference genome) using 66 rice accessions (36). Among those, several genes detected in previous studies as absent in the reference genome were reported, such as *SUBMERGENCE1A* (*Sub1A*), *SNORKEL1* and *SNORKEL2* (17, 18), controlling submergence tolerance, and *PHOSPHORUS-STARVATION TOLERANCE 1* (*Pstol1*), implied in the tolerance to phosphorus-deficient soil (19), respectively.
Dispensable genes in bacteria are thought to contribute to diversity and adaptation (22). In plants, the dispensable genome seems to be enriched in abiotic and biotic stress related genes, including defense and response, and developmental genes such as those that control flowering time (16, 21, 36–39). Disease resistance-related genes are some of the most prevalent types in the dispensable genome (40, 41). In rice, Schatz et al. showed that 5 to 12% of the dispensable genes within 3 divergent genomes contain the NB-ARC domain (nucleotide-binding domain of plant R-genes), *versus* only 0.35% within the core genome. In *Arabidopsis*, the largest part of the dispensable genome assembly (absent from the Columbia reference) belongs to to nucleotide-binding site leucine-rich repeat (NBS-LRR) genes (42). Other gene families that are also enriched in the dispensable genome include auxin- or flowering-related genes, or genes that encode enzymes involved in secondary metabolism (e.g. glucosinolates) (39). Finally, more "accessory" functions are linked to the dispensable genome sequences such as telomere maintenance or negative regulation (21). However, those "accessory" functions can drive major differentiation within a species, as in the wild *Brachypodium distachyon* with which flowering time is the main population splitter (21). In this last study, almost 77.6% of the protein coding genes from the core genome has similarities with known *InterPro* domains, a much higher proportion than that in the dispensable genome

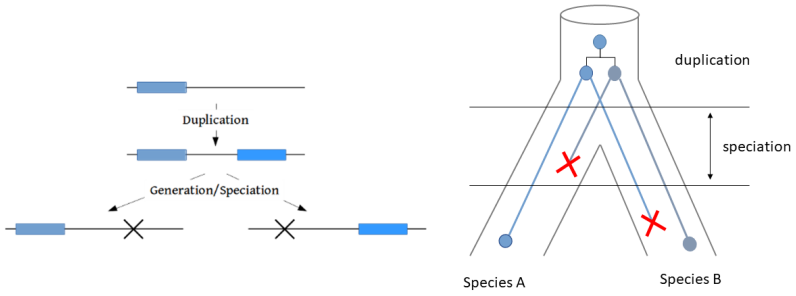

Fig. 1. After a recent duplication, through generation or speciation, one individual will conserve the dark blue paralog while the second individual will conserve the light blue paralog

set (35.8%). This observation led some authors to suspect that a portion of the PAV genes in the dispensable genome set may be just annotation artifacts or pseudogenes (43).

Individual Specific Genome. The individual-specific compartment contains sequences uniquely detected for one individual and therefore potentially responsible for specific features of the individual. Although this compartment may contain sequences with real biological functions (for highly divergent sequences or neogenes) (44), many of them are probably artifacts, misannotations or contaminations. It may also be the result of sampling bias; additional individual data may transfer those sequences to the dispensable genome. Consequently, this compartment might either be merged with the dispensable genome (5), or discarded for subsequent analyses as in *Brachypodium* analyses (21).

To Be or not to Be Core. The core genome is generally considered as the set of sequences common to all individuals of the considered group. However, even if in theory this is a valid definition, due to various limitations (e.g. sampling, sequencing and technical issues linked to GC-content), some sequences may not be detected in some individuals even though they are present. Thus, we propose that a sequence belongs to the core genome when 90 to 95% of the individuals harbor it, as published previously (21, 36). All sequences not included in the core genome are by default placed in the dispensable genome (**Figure 2**). Other authors proposed a less strict definition than core and dispensable genomes, using more sophisticated statistical approaches to define persistent, shell and cloud levels in the pangenome (34). Some sequences may be unique to a single individual, while some may be shared only by less than 90-95% of the group. While individual-specific sequences are most of the time artifacts or contamination, they could indeed be new genes (see below).
From a functional point of view of the pangenome, a gene family will remain in the core if any member of this family is able to perform the function and is present in each individual. As the classification in a given family will depend on the threshold used in its computation, using a functional definition may be complex and may also depend on the clustering method used (e.g. OrthoMCL(45), MCL(46), Mutual Best Hit(47), GET_HOMOLOGUES-EST(43)).

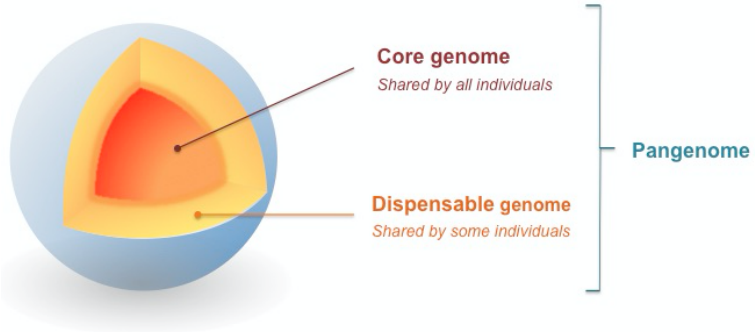

Fig. 2. The pangenome is seen as a sphere that contains all the genome of a collection of individuals. The core genome gathers all the common sequences shared by all individuals while the dispensable genome consists of sequences shared by only some individuals.

How Many Genomes to Capture the Whole Genome Content of a Given Group ?. For each pangenome analysis, recurrent questions arise: 1) How many genomes should be sequenced to maximize the diversity within a group? 2) Will there always be the same set of sequences shared by all members of a group even when newly sequenced individuals are added? 3) Will new specific sequences still be discovered with additional individual sequencing?
In order to validate whether the definitive pangenome size has been reached, Tettelin et al. (22, 48) proposed to represent the evolution of the total number of sequences found after the addition of each individual sequenced. If the number of sequences levels off to a plateau with each newly added genome, the pangenome is closed. Otherwise, the pangenome is still unlimited and defined as open (**Figure 3A**).
Tettelin et al showed that the pangenome of the bacteria *S. agalactidae* is very large and open with numerous new

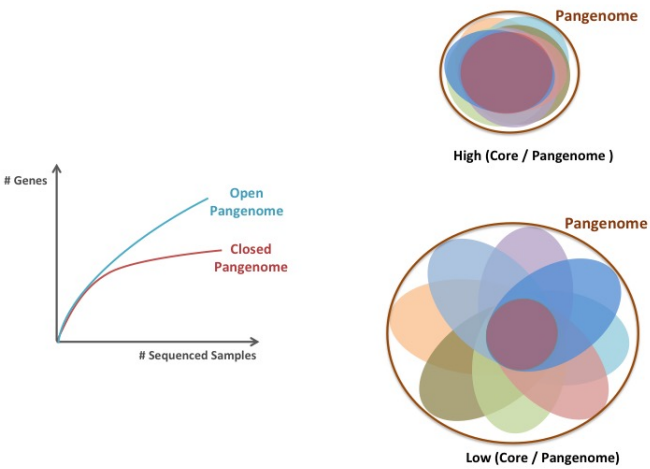

Fig. 3. A. Open and closed pangenomes (adapted from (49)). B. C/P ratio illustration.

unique genes identified even after hundreds of genomes were sequenced (48). Within plants, the pangenome was shown closed for several models such as soybean, *Brassica oleracea*, maize or *Medicago* with a small number of samples (39, 50–52). Indeed, the size of the pangenome will depend on the genome dynamics of the considered group; thus, in this regard, bacteria have a relatively larger pangenome than plants because of their higher level of gene flow.

Such observations are generally performed based on gene sequences only (*i.e.* standard protein coding genes *a fortiori*), and not on non-protein genes, neogenes and TEs. In addition, the sample choice is critically important: an underrepresentation of the diversity within a given group may indicate that the pangenome is closed, while if the population of individuals sampled is increased, the pangenome may be found to be open. In this case, the largest possible population of individuals should be targeted for sampling (15).

The Core/Pangenome ratio seems to be related to an organism's capacity of adaptation (53), with values under 85% showing a huge adaptability(**Figure 3B**). In plants, the core genome represents from 40 to 80% of the total pangenome, depending on the organism and group's structure (**Table 1**), indicating a large potential for plants.

Methods for Pangenome Assembly

Whichever pangenome definition is used, the first step is to obtain sequences per individual. Up to now, three main approaches have been used in plants to assemble pangenomes.

Assemble-Then-Map: Complete *de novo* Assembly Approach. With bacteria, pangenome studies used complete *de novo* assemblies of small genomes (and their subsequent annotations) (22, 48). With plants, most studies (16, 21, 36, 50, 54, 55) also used a similar approach (**Figure 4**A). With this method, sequences from each individual are assembled separately, then mapped all-versus-all sequences and also to a reference in order to reduce redundancy and to identify shared and non-shared sequences. This approach is time-consuming, requires costly computing and sometimes leads to errors when short read sequencing (e.g. Illumina) is used for large genomes. Indeed, repeated sequences are difficult to resolve using short reads sequences and such assemblies generally result in fragmentation of contigs, leading to a loss of collinearity of fragments. However, the recent and rapid development of long-read sequencing technologies such as Oxford Nanopore Technologies and Pacific Biosciences will allow better assemblies and longer contigs, which should resolve the main problem with this approach.

Metagenomic-Like Approach. Yao et al. (20) combined low-coverage data of around 1,500 rice genomes to perform a pangenome assembly using a metagenomic-like approach (**Figure 4**B). They assembled the whole sequence data together then re-assigned the different contigs to individuals through mapping of the single individual data on their metagenomic assembly. While this allows working with low coverage data from a large number of samples, such an approach may result in chimerical assembly of artifactual sequences.

Map-Then-Assemble: Reference-Based Approach. The map-then-assemble approach allows to perform individual assemblies after shared sequences are identified (**Figure 4**C). The idea here is to map all the sequences upon a reference sequence, then to re-assemble per individual the unmapped data (12, 42, 56). An alternative way is to re-assemble through an iterative mode (39), where samples are mapped successively on a panreference, which is updated each turn by the newly assembled sequences. In such a way, shared repeated and complex regions are resolved immediately. Assemblies per individuals are then grouped and de-duplicated to avoid redundancy. This approach is less time-consuming than the *de novo* assembly previously described; however it may impair the detection of recent duplicated sequences. Reads from the two copies that are the result of a recent duplication may map on the single target. In addition, this approach is generally performed using short reads and which may lead to short contigs as described in the sections above for the metagenomic-like or *de novo* assembly approaches.

Creating a Panreference. Whatever approach is selected to identify the core and dispensable sequences, the dispensable genome is generally anchored into a reference sequence in order to create a panreference needed for subsequent analyses. Anchoring to a reference sequence may be performed by gene synteny (21) using the collinearity of nearby core genes to identify the position of the dispensable sequences. This method allows to anchor the data precisely, but only if the dispensable sequence is located close to core genes (annotation-based). Another approach is the use of linkage disequilibrium (LD) between genetic markers (e.g. SNPs) on dispensable sequences and on core markers (20). It can be faster than gene synteny methods and allows working with non-genic data, but the anchoring is not precise and generally depends on the local LD value. Similarity-based approaches can also be used

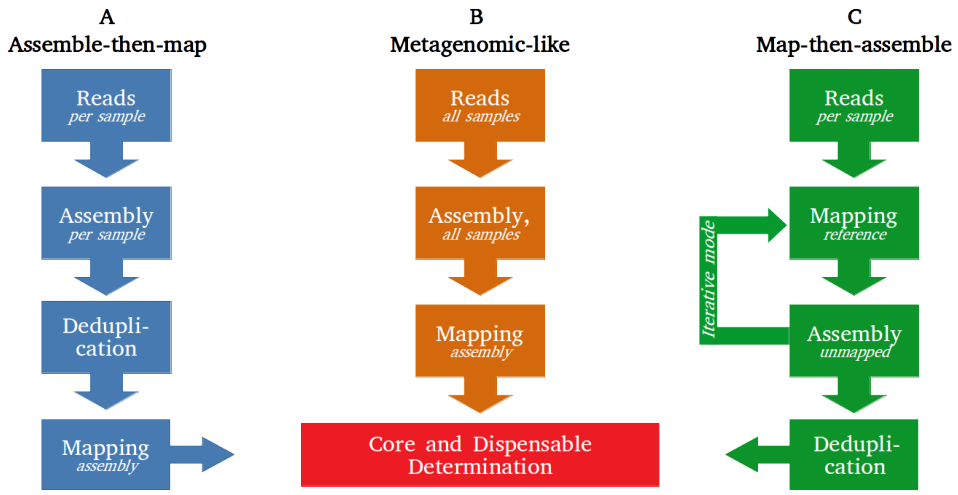

Fig. 4. The three approaches for pangenome sequence data assembly: left, assemble-then-map; center, metagenomic-like; right, map-then-assemble. For assemble-then-map and map-the-assemble methods, reduction of redundancy is performed (deduplication step) to identify the common sequences of different individuals.

to identify where the border of dispensable sequences are located within the reference sequence. While this approach can be precise at the single-base scale, in the case of repeated sequences, the similarity can occur with multiple regions and the exact location between all these similar regions may be difficult to distinguished.

Annotation of the Core and Dispensable Genome. As for any classical single-sequence genome annotation, sequences can be annotated to provide functions. Dispensable and core gene sequences are generally clustered using methods coming from comparative genomics: pair-wise BLASTP or MCL tools (e.g. OrthoMCL (57), OrthoFinder (58)) and clustering with GET_HOMOLOGUES-EST approaches (43). The stringency of the clustering level used here will heavily influence the results. Different studies used different thresholds, and these thresholds for clustering will mainly depend on the genetic diversity of the considered group. For instance, a highly recently diverged group will be analyzed using a high threshold (up to 95% of similarity), while an older diverged group will use a more relaxed threshold (80% e.g.). For non-genic elements, such as TEs or miRNAs, the annotation will also be performed as with classical genomic annotation, using state-of-the-art tools (59, 60).
With bacteria, the pangenomic analyses generally rely on gene annotation and gene family clustering. With plants, no specific trend (gene family or structure or synteny) has been clearly adopted by the community, and authors tend to combine several approaches within the same study (21).

Dynamics of Pangenome Compartments

The ability of the pangenome size to increase or to be stable, as well as switching from core to dispensable and reversely, is strongly connected to the balance between gain and loss events and the ability to adapt to diverse environments (25). Different factors and forces can impact on the pangenome structure, including gene birth and death, horizontal transfers and TE activity (**Figure 5**).

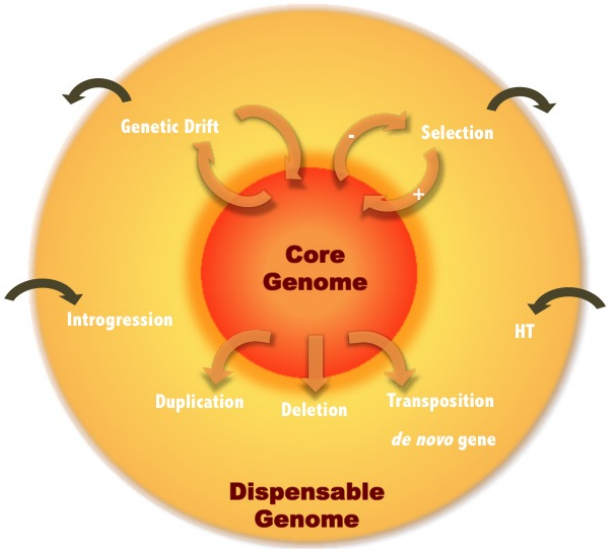

Fig. 5. Dynamic overview of the pangenome structure shaped by different events and forces. New sequences are added to the dispensable genome through mutations, duplications, deletions and transpositions, while the core genome content may decrease by deletion and transposition. Horizontal transfer and introgression also impact on the dispensable genome compartment (sequence gain). Moreover, positive and purifying selection as well as genetic drift impact on the core and dispensable genomes (sequence gain and loss) as well as on the pangenome (sequence loss).

Gene Birth-and-Death Processes. Gene creation and elimination can occur as a result of different processes, including errors during recombination that eliminate genes, TEs that mediate gene duplication, duplication events that result in gene gains, followed by diversification and neofunctionalization (21). There is evidence that most of the dispensable genes may arise from these gene birth-and-death mechanisms. Unique protein-coding genes may emerged from (*i*) non-coding DNA (*de novo* genes), (*ii*) an older coding sequence by a combination of mechanisms such as duplication followed by rapid divergence, horizontal gene transfer, or ancient gene lost with important sequence variation followed by neofunction, or exapted transposon (domesticated by the host genome to provide a new biological function) (61, 62).

Dispensable genes identified in several studies tend to display common features similar to young genes: short gene, weak Interpro homology, low expression, rapid evolution and turnover (16, 39, 63, 64). Several studies showed a regulatory role of these genes in response to numerous varying environmental conditions, biotic (65) or abiotic stresses (44), and potentially also in death gene processes (52).

Transposable Elements, Umpires and Players. Ubiquitous in all eukaryotic genomes, TEs represent a major part of many plant genomes. TEs are endogenous genomic elements able to duplicate themselves and to insert elsewhere in a host genome. They use different strategies to move, including RNA (retrotransposons, Class I) or DNA intermediates (DNA transposons, Class II) (32). The Copy-and-Paste amplification strategy used by retrotransposons allows them to accumulate in the genome at a high-copy numbers, with the result that some TE families can represent a predominant part of a genome. For instance, 85% of the maize (*Zea mays*) genome consists of TEs, mainly LTR Retrotransposons (66). TE movements are at the origin of numerous genomic variations within species (14, 67). For instance, in maize, the activity of TEs, especially Helitron-like elements, was able to modify up to 50% of the genome structure in a vast majority of the collinear BACs analyzed (13, 14). Due to their ability to change location within the host genome (32), they are the first candidates for dispensable genome creation (14, 16), as every new insertion will belong to this compartment. Finally, their ability to spread through a population at a high rate allows them to invade even the core genome of species, such as the *P* element in different *Drosophila* species (68).
Beyond their own intrinsic activity, the presence of TEs at a given position may alter the pangenomic structure. Golicz et al. observed a higher TE density surrounding variable genes in *Brassica oleacerea* (39). In the same way, Gordon et al. (21) showed that non-core genes are more linked to TE activity than core genes. TEs can alter the expression of surrounding genes (69, 70), but also the global genome structure by serving as anchor for illegitimate recombination (71).

Horizontal Transfers. It has been shown within bacteria the importance of Horizontal Gene Transfer (HGT)(72) and its impact on the pangenome (73). HGT has been shown in eukaryotes (74, 75), sometimes with a high success and essential functions, and can be selected to become a core gene. Horizontal Transfers (HT) from non-genic elements such as TEs (76, 77) may also impact the dispensable genome and could invade the host genome in a very short period of time (68), and on a large array of species (78, 79).

Challenges, Perspectives and the Way Forward

Links and Impacts on Phenotype. Dispensable genes are thought to be responsible for considerable phenotypic variation that could be suitable for breeding improved crop varieties and evolutionary studies of adaptive traits (21). Structural variations (including CNV and PAV) can significantly have an impact on phenotypic variation in plants. For instance, Lu et al, (80) investigated the contribution of PAVs to phenotypic variance using GWAS on 4 traits in maize and SNPs located in non-reference sequences were found enriched in the significant GWAS hits compared to reference-based SNPs, indicating their possible role in such variation. In the same way, in a study with rice, more than 40% of the agronomical traits were linked to non-reference sequences, thus dispensable (20). In addition, in wild African rice *O. barthii*, the PAV of the PROSTATE GROWTH 1 (*PROG1*) gene directly impacts global plant phenotype: when absent, plants are erect, while when present plants are not erect (7). The absent state seems to have been selected in the cultivated relative *O. glaberrima*. Many other examples exist in rice and other crops that show numerous phenotypes of interest are not linked to SNPs but to PAVs.

Adaptation, Selection and Speciation. The pangenome concept offers new perspectives to increase our knowledge about evolutionary mechanisms that allow organisms to adapt quickly to new environments. Indeed, more and more studies show adaptive phenotypic changes in plants for various traits due to CNVs (e.g. flowering time (81, 82), pest and diseases resistance (83, 84), herbicide resistance (85), plant height (86)). The ability to acquire new genes and to generate gene allelic diversity has various potential effects including neutral, adaptive or not on fitness (49, 87). Pangenome analysis offers new ways to investigate the adaptation processes and to understand their impacts on the core and dispensable genomes. It would be particularly interesting to focus on different periods of divergence within a given group, such as speciation, when effective population size is small, genetic drift effect is important, and events such as the reproductive isolation is occurring. It was shown for some species that speciation will impact drastically the pangenomic structure (12, 39, 64). This will lead to additional questions and possibilities, such as what is a species in perspective of the pangenome? Is having the same core genome enough to be from the same species, or is it too restrictive or relaxed?

Graphical Visualization. Graphical tools have been developed to handle and display bacterial pangenome datasets such as PanX (88), pan-Tetris (89), PanViz (90), seq-seq-pan (91) and PanACEAE (92). However, fewer have been proposed for plant genomes including Rpan (93) and Brachypan (21). Generally, publications on the topic have revisited Venn diagram or flower plot like representations (94–97) (**Figure 6A**) to illustrate PAV, but this representation has limitations with increased sample size, leading to the possibility of alternate visualization tools, such as Upset (98). Various approaches are emerging using graph-based structures (99, 100) (genome and variation graphes; Figure 6B), for example using the de Bruijn Graph Algorithm.
The main challenge remains to design scalable solutions for large panels of samples able to support PAV-based functional analyses that allow to zoom into chromosome segments to visualize individual SVs and SNPs for structural-based analyses. However, such comprehensive systems are still in their

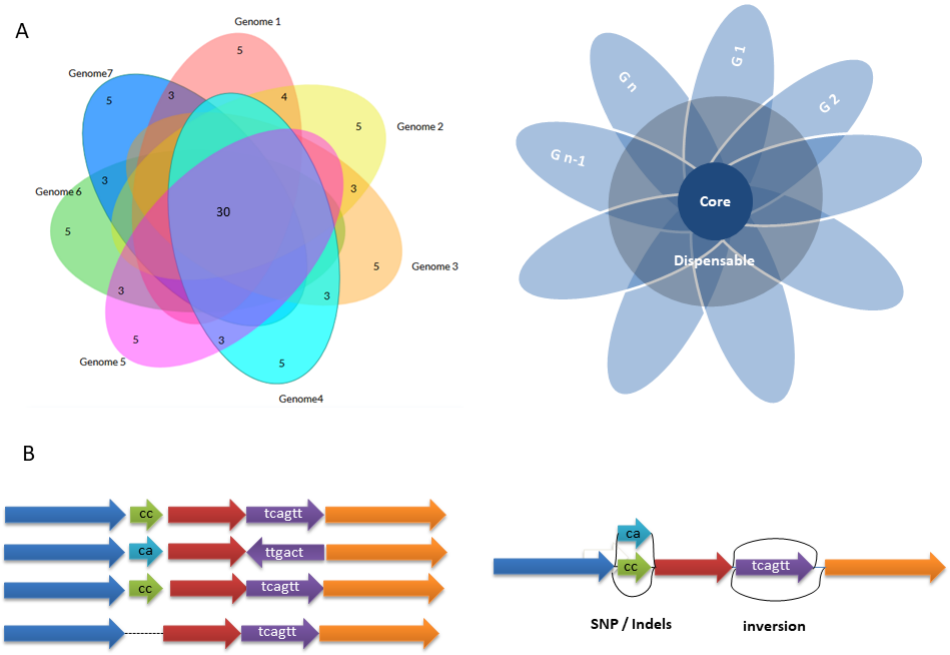

Fig. 6. A. Frequent static representations of Pangenomes. B. Cartoon of a Graph-based structure, adapted from (99)

infancy and are not yet operational. Whatever solutions will be developed as reference tools, it would be recommended to build them upon existing systems with powerful capacity to explore genomes and variants, or at least to enable interoperability between them.

Expected Contribution of recent sequencing approaches. Conformation capture methods such as Hi-C, mate-pairs libraries or 10x synthetic long-reads are second generation technology based approaches that can be used in the near future to resolve panreference assemblies. Besides, third-generation sequencing technologies such as Pacific Bio-Sciences SMRT or Oxford Nanopore Technologies (ONT) offer single-strand long-reads sequences (up to 2 Mb, the current ONT record so far). These methods, and especially the low-cost Minion from ONT, will change the paradigm of one high-quality reference sequence and many draft sequence samples. Indeed, a golden-standard quality sequence can be performed for less than 1,000 USD for genomes of around 150Mb (101), and 1,300 USD for a rice-sized genome (400Mb; F. Sabot, unpublished data). Thus, the Assemble-then-map approach (see **Figure 4**) may become the standard for future pangenomic approaches. Indeed, the capacity of long-read assemblies to overcome the repeat sequence paradox and to solve the scaffolding difficulties will make this technology the best tool for pangenome analysis. Advantages of a portable solution such as Minion will allow rapid sequence capacities in any lab with any sample of interest to identify PAVs or CNVs of interest.

Conclusions and Future Perspectives. The pangenome concept combined with high-throughput third-generation sequencing will probably allow access to large gene repertoires of the wild relatives of cultivated plants, particularly interesting for crucial agronomic traits such as drought and salinity tolerance. Genome portals will have to evolve from a reference genome centric view to adopt a pangenome reference view, or to manage multiple reference assemblies with a granular level of display with standardized genome assembly and gene models nomenclature. Using dispensable genome data will allow identifying the genetic basis of phenotypes of interest in dedicated lines.

Summary points

1. Pangenome view of a genome opens new challenging ways to explore genetic diversity, adaptation and evolutionary mechanisms within a group.

2. Pangenome analyses give access to a surprisingly large reservoir of genes/sequences never identified when working on a single reference genome.

3. Increased knowledge of dispensable genes may be of high importance for breeding applications.

Disclosure Statement

None

ACKNOWLEDGEMENTS
Authors want to thank their respective host institutions for funding as well as the CGIAR Research Programs (CRP) on Roots, Tubers and Bananas (RTB) and RICE.

Table 1. Current overview of current plant pangenome studies.

Organism	Sample number	Method	Total Pangenes	Core[a]	Dispensable[a]	References
Arabidopsis thaliana	19	assemble-then-map	37,789	69.7	30.3	(43)
Brachypodium distachyon	54	assemble-then-map	37,886	54	46	(21)
Brassica oleacera	10[b]	map-then-assemble	61,379	81.3	18.7	(39)
Capsicum	355	map-then-assemble	51,757	55.7	44.3	(102)
Medicago	15	assemble-then-map	74,700	41.9	58.1	(103)
Poplar	22	mapping only	-	80.7	19.3	(104)
Asian Rice	66[b]	assemble-then-map	42,580	61.9	38.1	(36)
Asian Rice	453/3,000	assemble-then-map	46,115[c]/47,288[d]	52.9/61.3	47.1/28.7	(105)
African Rice	120 cultivated / 74 wild	map-the-assemble	35,198/36,252	86.5/98.6	13.5/1.4	(12)
Soybean	7[b]	assemble-then-map	59,080	80.1	19.9	(5)
Bread Wheat	18	map-then-assemble	128,656	64.3	33.7	(15)

[1a] Percentage of total pangenes; [b] Wild and cultivated ; [c] *indica* subspecies ; [d] *japonica* subspecies.

Bibliography

1. Karl J Schmid, Sebastian Ramos-Onsins, Henriette Ringys-Beckstein, Bernd Weisshaar, and Thomas Mitchell-Olds. A multilocus sequence survey in arabidopsis thaliana reveals a genome-wide departure from a neutral model of dna sequence polymorphism. *Genetics*, 169(3):1601—1615, March 2005. ISSN 0016-6731. doi: 10.1534/genetics.104.033795.
2. Zhang De-Xing and Godfrey M. Hewitt. Nuclear dna analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, 12(3):563–584, 2003. doi: 10.1046/j.1365-294X.2003.01773.x.
3. Richard Redon, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, Michael H Shapero, Andrew R Carson, Wenwei Chen, Eun Kyung Cho, Stephanie Dallaire, Jennifer L Freeman, Juan R González, Mònica Gratacòs, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R MacDonald, Christian R Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol, Donald F Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P Carter, Hiroyuki Aburatani, Charles Lee, Keith W Jones, Stephen W Scherer, and Matthew E Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–54, nov 2006. ISSN 1476-4687. doi: 10.1038/nature05329.
4. Jianjian Qi, Xin Liu, Di Shen, Han Miao, Bingyan Xie, Xixiang Li, Peng Zeng, Shenhao Wang, Yi Shang, Xingfang Gu, Yongchen Du, Ying Li, Tao Lin, Jinhong Yuan, Xueyong Yang, Jinfeng Chen, Huiming Chen, Xingyao Xiong, Ke Huang, Zhangjun Fei, Linyong Mao, Li Tian, Thomas Städler, Susanne S Renner, Sophien Kamoun, William J Lucas, Zhonghua Zhang, and Sanwen Huang. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nature Genetics*, 45:1510, oct 2013.
5. Jia-Yang Li, Jun Wang, and Robert S Zeigler. The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience*, 3(1):8, dec 2014. ISSN 2047-217X. doi: 10.1186/2047-217X-3-8.
6. Tao Lin, Guangtao Zhu, Junhong Zhang, Xiangyang Xu, Qinghui Yu, Zheng Zheng, Zhonghua Zhang, Yaoyao Lun, Shuai Li, Xiaoxuan Wang, Zejun Huang, Junming Li, Chunzhi Zhang, Taotao Wang, Yuyang Zhang, Aoxue Wang, Yancong Zhang, Kui Lin, Chuanyou Li, Guosheng Xiong, Yongbiao Xue, Andrea Mazzucato, Mathilde Causse, Zhangjun Fei, James J Giovannoni, Roger T Chetelat, Dani Zamir, Thomas Städler, Jingfu Li, Zhibiao Ye, Yongchen Du, and Sanwen Huang. Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics*, 46:1220, oct 2014.
7. P. Cubry, C. Tranchant-Dubreuil, A.-C. Thuillet, C. Monat, M.-N. Ndjiondjop, K. Labadie, C. Cruaud, S. Engelen, N. Scarcelli, B. Rhoné, C. Burgarella, C. Dupuy, P. Larmande, P. Wincker, O. François, F. Sabot, and Y. Vigouroux. The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes. *Current Biology*, 2018. ISSN 09609822. doi: 10.1016/j.cub.2018.05.066.
8. Nathan M Springer, Kai Ying, Yan Fu, Tieming Ji, Cheng-Ting Yeh, Yi Jia, Wei Wu, Todd Richmond, Jacob Kitzman, Heidi Rosenbaum, A Leonardo Iniguez, W Brad Barbazuk, Jeffrey A Jeddeloh, Dan Nettleton, and Patrick S Schnable. Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLOS Genetics*, 5(11):e1000734, nov 2009.
9. Rachit K Saxena, David Edwards, and Rajeev K Varshney. Structural variations in plant genomes. *Briefings in functional genomics*, 13(4):296–307, jul 2014. ISSN 2041-2657. doi: 10.1093/bfgp/elu016.
10. Bhavna Hurgobin and David Edwards. SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete? *Biology*, 6(1), mar 2017. ISSN 2079-7737. doi: 10.3390/biology6010021.
11. Cécile Monat, Bérengère Pera, Marie-Noelle Ndjiondjop, Mounirou Sow, Christine Tranchant-Dubreuil, Leila Bastianelli, Alain Ghesquière, and Francois Sabot. De novo assemblies of three oryza glaberrima accessions provide first insights about pan-genome of african rices. *Genome Biology and Evolution*, 9(1):1–6, 2017. doi: 10.1093/gbe/evw253.
12. Cecile Monat, Christine Tranchant-Dubreuil, Stefan Engelen, Karine Labadie, Emmanuel Paradis, Ndomassi Tando, and Francois Sabot. Comparison of two african rice species through a new pan-genomic approach on massive data. *bioRxiv*, 2018. doi: 10.1101/245431.
13. Michele Morgante, Stephan Brunner, Giorgio Pea, Kevin Fengler, Andrea Zuccolo, and Antoni Rafalski. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature genetics*, 37(9):997—1002, September 2005. ISSN 1061-4036. doi: 10.1038/ng1615.
14. Michele Morgante, Emanuele De Paoli, and Slobodanka Radovic. Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology*, 10(2):149 – 155, 2007. ISSN 1369-5266. doi: https://doi.org/10.1016/j.pbi.2007.02.001. Genome Studies and Molecular Genetics / Edited by Stefan Jansson and Edward S Buckler.
15. Juan D. Montenegro, Agnieszka A. Golicz, Philipp E. Bayer, Bhavna Hurgobin, HueyTyng Lee, Chon-Kit Kenneth Chan, Paul Visendi, Kaitao Lai, Jaroslav Doležel, Jacqueline Batley, and David Edwards. The pangenome of hexaploid bread wheat. *The Plant Journal*, 90(5):1007–1013, jun 2017. ISSN 09607412. doi: 10.1111/tpj.13515.
16. Michael C Schatz, Lyza G Maron, Joshua C Stein, Alejandro Hernandez Wences, James Gurtowski, Eric Biggers, Hayan Lee, Melissa Kramer, Eric Antoniou, Elena Ghiban, Mark H Wright, Jer-ming Chia, Doreen Ware, Susan R McCouch, and W Richard McCombie. 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, 2014. ISSN 1474-760X. doi: 10.1186/PREACCEPT-2784872521277375.
17. Kenong Xu, Xia Xu, Takeshi Fukao, Patrick Canlas, Reycel Maghirang-Rodriguez, Sigrid Heuer, Abdelbagi M. Ismail, Julia Bailey-Serres, Pamela C. Ronald, and David J. Mackill. Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature*, 442(7103):705–708, aug 2006. ISSN 0028-0836. doi: 10.1038/nature04920.
18. Yoko Hattori, Keisuke Nagai, Shizuka Furukawa, Xian-Jun Song, Ritsuko Kawano, Hitoshi Sakakibara, Jianzhong Wu, Takashi Matsumoto, Atsushi Yoshimura, Hidemi Kitano, Makoto Matsuoka, Hitoshi Mori, and Motoyuki Ashikari. The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. *Nature*, 460(7258):1026–1030, aug 2009. ISSN 0028-0836. doi: 10.1038/nature08258.
19. Rico Gamuyao, Joong Hyoun Chin, Juan Pariasca-Tanaka, Paolo Pesaresi, Sheryl Catausan, Cheryl Dalid, Inez Slamet-Loedin, Evelyn Mae Tecson-Mendoza, Matthias Wissuwa, and Sigrid Heuer. The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. *Nature*, 488(7412):535–539, aug 2012. ISSN 0028-0836. doi: 10.1038/nature11346.
20. Wen Yao, Guangwei Li, Hu Zhao, Gongwei Wang, Xingming Lian, and Weibo Xie. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome biology*, 16:187, sep 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0757-3.
21. Sean P. Gordon, Bruno Contreras-Moreira, Daniel P. Woods, David L. Des Marais, Diane Burgess, Shengqiang Shu, Christoph Stritt, Anne C. Roulin, Wendy Schackwitz, Ludmila Tyler, Joel Martin, Anna Lipzen, Niklas Dochy, Jeremy Phillips, Kerrie Barry, Koen Geuten, Hikmet Budak, Thomas E. Juenger, Richard Amasino, Ana L. Caicedo, David Goodstein, Patrick Davidson, Luis A. J. Mur, Melania Figueroa, Michael Freeling, Pilar Catalan, and John P. Vogel. Extensive gene content variation in the Brachypodium distachyon pangenome correlates with population structure. *Nature Communications*, 8(1):2184, dec 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-02292-8.
22. Hervé Tettelin, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, Jonathan Crabtree, Amanda L. Jones, A. Scott Durkin, Robert T. DeBoy, Tanja M. Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D. Peterson, Christopher R. Hauser, Jaideep P. Sundaram, William C. Nelson, Ramana Madupu, Lauren M. Brinkac, Robert J. Dodson, Mary J. Rosovitz, Steven A. Sullivan, Sean C. Daugherty, Daniel H. Haft, Jeremy Selengut, Michelle L. Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J. B. O'Connor, Shannon Smith, Teresa R. Utterback, Owen White, Craig E. Rubens, Guido Grandi, Lawrence C. Madoff, Dennis L. Kasper, John L. Telford, Michael R. Wessels, Rino Rappuoli, and Claire M. Fraser. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102(39):13950–13955, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506758102.
23. Duccio Medini, Claudio Donati, Herve Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6):589 – 594, 2005. ISSN 0959-437X. doi: https://doi.org/10.1016/j.gde.2005.09.006. Genomes and evolution.
24. Fabio Marroni, Sara Pinosio, and Michele Morgante. Structural variation and genome complexity: is dispensable really dispensable? *Current Opinion in Plant Biology*, 18:31–36, apr 2014. ISSN 1369-5266. doi: 10.1016/J.PBI.2014.01.003.
25. Luis Carlos Guimarães, Jolanta Florczak-Wyspianska, Leandro Benevides de Jesus, Marcus Vinícius Canário Viana, Artur Silva, Rommel Thiago Jucá Ramos, Siomar de Castro Soares, and Siomar de Castro Soares. Inside the pan-genome - methods and software overview. *Current genomics*, 16(4):245—252, August 2015. ISSN 1389-2029. doi: 10.2174/1389202916666150423002311.
26. Clémence Plissonneau, Fanny E Hartmann, and Daniel Croll. Pangenome analyses of the wheat pathogen zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. *BMC biology*, 16(1):5, January 2018. ISSN 1741-7007. doi: 10.1186/s12915-017-0457-4.
27. Lars Snipen, Trygve Almøy, and David W Ussery. Microbial comparative pan-genomics using binomial mixture models. *BMC genomics*, 10:385, 2009. ISSN 1471-2164. doi: 10.1186/1471-2164-10-385.
28. Luis David Alcaraz, Gabriel Moreno-Hagelsieb, Luis E Eguiarte, Valeria Souza, Luis Herrera-Estrella, and Gabriela Olmedo. Understanding the evolutionary relationships and major traits of bacillus through comparative genomics. *BMC genomics*, 11:332, May 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-332.
29. David A Rasko, MJ Rosovitz, Garry S A Myers, Emmanuel F Mongodin, W Florian Fricke, Pawel Gajer, Jonathan Crabtree, Mohammed Sebaihia, Nicholas R Thomson, Roy Chaudhuri, Ian R Henderson, Vanessa Sperandio, and Jacques Ravel. The pangenome structure of escherichia coli: comparative genomic analysis of e. coli commensal and pathogenic isolates. *Journal of bacteriology*, 190(20):6881—6893, October 2008. ISSN 0021-9193. doi: 10.1128/jb.00619-08.
30. Victor V Tetz. The pangenome concept: a unifying view of genetic information. *Medical science monitor*, 11(7):HY24–HY29, 2005.
31. Chen Sun, Zhiqiang Hu, Tianqing Zheng, Kuangchen Lu, Yue Zhao, Wensheng Wang, Jianxin Shi, Chunchao Wang, Jinyuan Lu, Dabing Zhang, Zhikang Li, and Chaochun Wei. Rpan: rice pan-genome browser for 3000 rice genomes. *Nucleic Acids Research*, 45(2):597–605, 2017. doi: 10.1093/nar/gkw958.
32. Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H Schulman. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8:973, dec 2007.
33. Sarah C.R. Elgin and Gunter Reuter. Position-effect variegation, heterochromatin formation, and gene silencing in drosophila. *Cold Spring Harbor Perspectives in Biology*, 5(8), 2013. doi: 10.1101/cshperspect.a017780.
34. R. Eric Collins and Paul G. Higgs. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Molecular Biology and Evolution*, 29(11):3413–3425, 2012. doi: 10.1093/molbev/mss163.
35. Hurgobin Bhavna, Golicz Agnieszka A., Bayer Philipp E., Chan Chon-Kit Kenneth, Tirnaz Soodeh, Dolatabadian Aria, Schiessl Sarah V., Samans Birgit, Montenegro Juan D., Parkin Isobel A. P., Pires J. Chris, Chalhoub Boulos, King Graham J., Snowdon Rod, Batley Jacqueline, and Edwards David. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid brassica napus. *Plant Biotechnology Journal*, 16(7):1265–1274, 2017. doi: 10.1111/pbi.12867.
36. Qiang Zhao, Qi Feng, Hengyun Lu, Yan Li, Ahong Wang, Qilin Tian, Qilin Zhan, Yiqi Lu, Lei Zhang, Tao Huang, Yongchun Wang, Danlin Fan, Yan Zhao, Ziqun Wang, Congcong Zhou, Jiaying Chen, Chuanrang Zhu, Wenjun Li, Qijun Weng, Qun Xu, Zi-Xuan Wang, Xinghua Wei, Bin Han, and Xuehui Huang. Pan-genome analysis highlights the extent of genomic

variation in cultivated and wild rice. *Nature Genetics*, page 1, jan 2018. ISSN 1061-4036. doi: 10.1038/s41588-018-0041-z.

37. Weiya Xue, Yongzhong Xing, Xiaoyu Weng, Yu Zhao, Weijiang Tang, Lei Wang, Hongju Zhou, Sibin Yu, Caiguo Xu, Xianghua Li, and Qifa Zhang. Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nature Genetics*, 40:761, may 2008.
38. Cui Jiajun, Fan Shengci, Shao Tian, Huang Zejun, Zheng Dali, Tang Ding, Li Ming, Qian Qian, and Cheng Zhukuan. Characterization and fine mapping of the ibf mutant in rice. *Journal of Integrative Plant Biology*, 49(5):678–685, 2007. doi: 10.1111/j.1744-7909.2007.00467.x.
39. Agnieszka A. Golicz, Philipp E. Bayer, Guy C. Barker, Patrick P. Edger, HyeRan Kim, Paula A. Martinez, Chon Kit Kenneth Chan, Anita Severn-Ellis, W. Richard McCombie, Isobel A. P. Parkin, Andrew H. Paterson, J. Chris Pires, Andrew G. Sharpe, Haibao Tang, Graham R. Teakle, Christopher D. Town, Jacqueline Batley, and David Edwards. The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications*, 7:13390, nov 2016. ISSN 2041-1723. doi: 10.1038/ncomms13390.
40. Xun Xu, Xin Liu, Song Ge, Jeffrey D Jensen, Fengyi Hu, Xin Li, Yang Dong, Ryan N Gutenkunst, Lin Fang, Lei Huang, Jingxiang Li, Weiming He, Guojie Zhang, Xiaoming Zheng, Fumin Zhang, Yingrui Li, Chang Yu, Karsten Kristiansen, Xiuqing Zhang, Jian Wang, Mark Wright, Susan McCouch, Rasmus Nielsen, Jun Wang, and Wen Wang. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology*, 30:105, dec 2011.
41. Leah K. McHale, William J. Haun, Wayne W. Xu, Pudota B. Bhaskar, Justin E. Anderson, David L. Hyten, Daniel J. Gerhardt, Jeffrey A. Jeddeloh, and Robert M. Stupar. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology*, 159(4):1295–1308, 2012. ISSN 0032-0889. doi: 10.1104/pp.112.194605.
42. Jun Cao, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, Xi Wang, Felix Ott, Jonas Müller, Carlos Alonso-Blanco, Karsten Borgwardt, Karl J Schmid, and Detlef Weigel. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43:956, aug 2011.
43. Bruno Contreras-Moreira, Carlos P. Cantalapiedra, María J. García-Pereira, Sean P. Gordon, John P. Vogel, Ernesto Igartua, Ana M. Casas, and Pablo Vinuesa. 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, 2017. ISSN 1664-462X. doi: 10.3389/fpls.2017.00184.
44. Li Ling, Foster Carol M., Gan Qinglei, Nettleton Dan, James Martha G., Myers Alan M., and Wurtele Eve Syrkin. Identification of the novel protein qqs as a component of the starch metabolic network in arabidopsis leaves. *The Plant Journal*, 58(3):485–498, 2009. doi: 10.1111/j.1365-313X.2009.03793.x.
45. Li Li, Christian J Stoeckert, and David S Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189, September 2003. ISSN 1088-9051. doi: 10.1101/gr.1224503.
46. A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.*, 30(7):1575–1584, 2002. doi: 10.1093/nar/30.7.1575.
47. Roman L. Tatusov, Eugene V. Koonin, and David J. Lipman. A Genomic Perspective on Protein Families. *Science*, 278(5338):631 –637, October 1997. doi: 10.1126/science.278.5338.631.
48. Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5):472 – 477, 2008. ISSN 1369-5274. doi: https://doi.org/10.1016/j.mib.2008.09.006. Antimicrobials/Genomics.
49. James O McInerney, Alan McNally, and Mary J O'Connell. Why prokaryotes have pangenomes. *Nature Microbiology*, 2:17040, mar 2017.
50. Ying-hui Li, Guangyu Zhou, Jianxin Ma, Wenkai Jiang, Long-guo Jin, Zhouhao Zhang, Yong Guo, Jinbo Zhang, Yi Sui, Liangtao Zheng, Shan-shan Zhang, Qiyang Zuo, Xue-hui Shi, Yan-fei Li, Wan-ke Zhang, Yiyao Hu, Guanyi Kong, Hui-long Hong, Bing Tan, Jian Song, Zhang-xiong Liu, Yaoshen Wang, Hang Ruan, Carol K L Yeung, Jian Liu, Hailong Wang, Li-juan Zhang, Rong-xia Guan, Ke-jing Wang, Wen-bin Li, Shou-yi Chen, Ru-zhen Chang, Zhi Jiang, Scott A Jackson, Ruiqiang Li, and Li-juan Qiu. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, 32(10):1045–1052, 2014. ISSN 1087-0156. doi: 10.1038/nbt.2979.
51. Candice N. Hirsch, Jillian M. Foerster, James M. Johnson, Rajandeep S. Sekhon, German Muttoni, Brieanne Vaillancourt, Francisco Peñagaricano, Erika Lindquist, Mary Ann Pedraza, Kerrie Barry, Natalia de Leon, Shawn M. Kaeppler, and C. Robin Buell. Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell*, 26(1):121–135, 2014. ISSN 1040-4651. doi: 10.1105/tpc.113.119982.
52. Peng Zhou, Kevin A. T. Silverstein, Thiruvarangan Ramaraj, Joseph Guhlin, Roxanne Denny, Junqi Liu, Andrew D. Farmer, Kelly P. Steele, Robert M. Stupar, Jason R. Miller, Peter Tiffin, Joann Mudge, and Nevin D. Young. Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. *BMC Genomics*, 18(1):261, dec 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-3654-1.
53. Aurélia Caputo, Vicky Merhej, Kalliopi Georgiades, Pierre-Edouard Fournier, Olivier Croce, Catherine Robert, and Didier Raoult. Pan-genomic analysis to redefine species and subspecies based on quantum discontinuous variation: the Klebsiella paradigm. *Biology Direct*, 10(1):55, 2015. ISSN 1745-6150. doi: 10.1186/s13062-015-0085-2.
54. Xiangchao Gan, Oliver Stegle, Jonas Behr, Joshua G Steffen, Philipp Drewe, Katie L Hildebrand, Rune Lyngsoe, Sebastian J Schultheiss, Edward J Osborne, Vipin T Sreedharan, André Kahles, Regina Bohnert, Géraldine Jean, Paul Derwent, Paul Kersey, Eric J Belfield, Nicholas P Harberd, Eric Kemen, Christopher Toomajian, Paula X Kover, Richard M Clark, Gunnar Rätsch, and Richard Mott. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477:419, aug 2011.
55. Hiroaki Sakai, Hiroyuki Kanamori, Yuko Arai-Kichise, Mari Shibata-Hatta, Kaworu Ebana, Youko Oono, Kanako Kurita, Hiroko Fujisawa, Satoshi Katagiri, Yoshiyuki Mukai, Masao Hamada, Takeshi Itoh, Takashi Matsumoto, Yuichi Katayose, Kyo Wakasa, Masahiro Yano, and Jianzhong Wu. Construction of pseudomolecule sequences of the aus rice cultivar Kasalath for comparative genomics of Asian cultivated rice. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 21(4):397–405, aug 2014. ISSN 1756-1663. doi: 10.1093/dnares/dsu006.
56. Veronika Laine, Toni I Gossmann, Kees van Oers, Marcel E Visser, and Martien A.M. Groenen. Exploring the unmapped dna and rna reads in a songbird genome. *bioRxiv*, 2018. doi: 10.1101/371963.
57. Li Li, Christian J. Stoeckert, and David S. Roos. Orthomcl: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189, 2003. doi: 10.1101/gr.1224503.
58. David M. Emms and Steven Kelly. Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1):157, Aug 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0721-2.
59. Adam D. Ewing. Transposable element detection from whole genome sequence data. *Mobile DNA*, 6(1):24, dec 2015. ISSN 1759-8753. doi: 10.1186/s13100-015-0055-3.
60. Lavanya Rishishwar, Leonardo Mariño-Ramírez, and I. King Jordan. Benchmarking computational tools for polymorphic transposable element detection. *Briefings in Bioinformatics*, 18(6):bbw072, aug 2016. ISSN 1467-5463. doi: 10.1093/bib/bbw072.
61. Zebulun W. Arendsee, Ling Li, and Eve Syrkin Wurtele. Coming of age: orphan genes in plants. *Trends in Plant Science*, 19(11):698–708, nov 2014. ISSN 1360-1385. doi: 10.1016/J.TPLANTS.2014.07.003.
62. Christian Schlötterer. Genes from scratch–the evolutionary fate of de novo genes. *Trends in genetics : TIG*, 31(4):215–9, apr 2015. ISSN 0168-9525. doi: 10.1016/j.tig.2015.02.007.
63. Joshua C. Stein, Yeisoo Yu, Dario Copetti, Derrick J. Zwickl, Li Zhang, Chengjun Zhang, Kapeel Chougule, Dongying Gao, Aiko Iwata, Jose Luis Goicoechea, Sharon Wei, Jun Wang, Yi Liao, Muhua Wang, Julie Jacquemin, Claude Becker, Dave Kudrna, Jianwei Zhang, Carlos E. M. Londono, Xiang Song, Seunghee Lee, Paul Sanchez, Andrea Zuccolo, Jetty S. S. Ammiraju, Jayson Talag, Ann Danowitz, Luis F. Rivera, Andrea R. Gschwend, Christos Noutsos, Cheng-chieh Wu, Shu-min Kao, Jhih-wun Zeng, Fu-jin Wei, Qiang Zhao, Qi Feng, Moaine El Baidouri, Marie-Christine Carpentier, Eric Lasserre, Richard Cooke, Daniel da Rosa Farias, Luciano Carlos da Maia, Railson S. dos Santos, Kevin G. Nyberg, Kenneth L. McNally, Ramil Mauleon, Nickolai Alexandrov, Jeremy Schmutz, Dave Flowers, Chuanzhu Fan, Detlef Weigel, Kshirod K. Jena, Thomas Wicker, Mingsheng Chen, Bin Han, Robert Henry, Yue-ie C. Hsing, Nori Kurata, Antonio Costa de Oliveira, Olivier Panaud, Scott A. Jackson, Carlos A. Machado, Michael J. Sanderson, Manyuan Long, Doreen Ware, and Rod A. Wing. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nature Genetics*, page 1, jan 2018. ISSN 1061-4036. doi: 10.1038/s41588-018-0040-0.
64. Agnieszka A. Golicz, Jacqueline Batley, and David Edwards. Towards plant pangenomics. *Plant Biotechnology Journal*, 14(4):1099–1105, apr 2016. ISSN 14677644. doi: 10.1111/pbi.12499.
65. Wenfei Xiao, Hongbo Liu, Yu Li, Xianghua Li, Caiguo Xu, Manyuan Long, and Shiping Wang. A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLOS ONE*, 4(2):1–12, 02 2009. doi: 10.1371/journal.pone.0004603.
66. Patrick S. Schnable, Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A. Graves, Patrick Minx, Amy Denise Reily, Laura Courtney, Scott S. Kruchowski, Chad Tomlinson, Cindy Strong, Kim Delehaunty, Catrina Fronick, Bill Courtney, Susan M. Rock, Eddie Belter, Feiyu Du, Kyung Kim, Rachel M. Abbott, Marc Cotton, Andy Levy, Pamela Marchetto, Kerri Ochoa, Stephanie M. Jackson, Barbara Gillam, Weizu Chen, Le Yan, Jamey Higginbotham, Marco Cardenas, Jason Waligorski, Elizabeth Applebaum, Lindsey Phelps, Jason Falcone, Krishna Kanchi, Thynn Thane, Adam Scimone, Nay Thane, Jessica Henke, Tom Wang, Jessica Ruppert, Neha Shah, Kelsi Rotter, Jennifer Hodges, Elizabeth Ingenthron, Matt Cordes, Sara Kohlberg, Jennifer Sgro, Brandon Delgado, Kelly Mead, Asif Chinwalla, Shawn Leonard, Kevin Crouse, Kristi Collura, Dave Kudrna, Jennifer Currie, Ruifeng He, Angelina Angelova, Shanmugam Rajasekar, Teri Mueller, Rene Lomeli, Gabriel Scara, Ara Ko, Krista Delaney, Marina Wissotski, Georgina Lopez, David Campos, Michele Braidotti, Elizabeth Ashley, Wolfgang Golser, HyeRan Kim, Seunghee Lee, Jinke Lin, Zeljko Dujmic, Woojin Kim, Jayson Talag, Andrea Zuccolo, Chuanzhu Fan, Aswathy Sebastian, Melissa Kramer, Lori Spiegel, Lidia Nascimento, Theresa Zutavern, Beth Miller, Claude Ambroise, Stephanie Muller, Will Spooner, Apurva Narechania, Liya Ren, Sharon Wei, Sunita Kumari, Ben Faga, Michael J. Levy, Linda McMahan, Peter Van Buren, Matthew W. Vaughn, Kai Ying, Cheng-Ting Yeh, Scott J. Emrich, Yi Jia, Ananth Kalyanaraman, An-Ping Hsia, W. Brad Barbazuk, Regina S. Baucom, Thomas P. Brutnell, Nicholas C. Carpita, Cristian Chaparro, Jer-Ming Chia, Jean-Marc Deragon, James C. Estill, Yan Fu, Jeffrey A. Jeddeloh, Yujun Han, Hyeran Lee, Pinghua Li, Damon R. Lisch, Sanzhen Liu, Zhijie Liu, Dawn Holligan Nagel, Maureen C. McCann, Phillip SanMiguel, Alan M. Myers, Dan Nettleton, John Nguyen, Bryan W. Penning, Lalit Ponnala, Kevin L. Schneider, David C. Schwartz, Anupma Sharma, Carol Soderlund, Nathan M. Springer, Qi Sun, Hao Wang, Michael Waterman, Richard Westerman, Thomas K. Wolfgruber, Lixing Yang, Yeisoo Yu, Lifang Zhang, Shiguo Zhou, Qihui Zhu, Jeffrey L. Bennetzen, R. Kelly Dawe, Jiming Jiang, Ning Jiang, Gernot G. Presting, Susan R. Wessler, Srinivas Aluru, Robert A. Martienssen, Sandra W. Clifton, W. Richard McCombie, Rod A. Wing, and Richard K. Wilson. The b73 maize genome: Complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115, 2009. ISSN 0036-8075. doi: 10.1126/science.1178534.
67. Benoit Piegu, Romain Guyot, Nathalie Picault, Anne Roulin, Abhijit Sanyal, Hyeran Kim, Kristi Collura, Darshan S Brar, Scott Jackson, Rod A Wing, and Olivier Panaud. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in oryza australiensis, a wild relative of rice. *Genome research*, 16(10):1262—1269, October 2006. ISSN 1088-9051. doi: 10.1101/gr.5290206.
68. Robert Kofler, Tom Hill, Viola Nolte, Andrea J Betancourt, and Christian Schlötterer. The recent invasion of natural drosophila simulans populations by the p-element. *Proceedings of the National Academy of Sciences of the United States of America*, 112(21):6659—6663, May 2015. ISSN 0027-8424. doi: 10.1073/pnas.1500758112.
69. Cory D. Hirsch and Nathan M. Springer. Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1860

(1):157 – 165, 2017. ISSN 1874-9399. doi: https://doi.org/10.1016/j.bbagrm.2016.05.010. Plant Gene Regulatory Mechanisms and Networks.

70. Eugenio Butelli, Concetta Licciardello, Yang Zhang, Jianjun Liu, Steve Mackay, Paul Bailey, Giuseppe Reforgiato-Recupero, and Cathie Martin. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *The Plant Cell*, 24(3): 1242–1255, 2012. ISSN 1040-4651. doi: 10.1105/tpc.111.095232.
71. Nathalie Chantret, Jérôme Salse, François Sabot, Sadequr Rahman, Arnaud Bellec, Bastien Laubin, Ivan Dubois, Carole Dossat, Pierre Sourdille, Philippe Joudrier, Marie-Françoise Gautier, Laurence Cattolico, Michel Beckert, Sébastien Aubourg, Jean Weissenbach, Michel Caboche, Michel Bernard, Philippe Leroy, and Boulos Chalhoub. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (triticum and aegilops). *The Plant Cell*, 17(4):1033–1045, 2005. ISSN 1040-4651. doi: 10.1105/tpc.104.029181.
72. Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472—482, August 2015. ISSN 1471-0056. doi: 10.1038/nrg3962.
73. Eugene V Koonin. Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research*, 5, 2016. ISSN 2046-1402. doi: 10.12688/f1000research.8737.1.
74. Patrick J Keeling and Jeffrey D Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605—618, August 2008. ISSN 1471-0056. doi: 10.1038/nrg2386.
75. Dapeng Zhang, Lakshminarayan M. Iyer, and L. Aravind. Bacterial GRAS domain proteins throw new light on gibberellic acid response mechanisms. *Bioinformatics (Oxford, England)*, 28(19):2407–2411, October 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts464.
76. John F Y Brookfield. The ecology of the genome - mobile dna elements and their hosts. *Nature reviews. Genetics*, 6(2):128—136, February 2005. ISSN 1471-0056. doi: 10.1038/nrg1524.
77. Moaine El Baidouri, Marie-Christine Carpentier, Richard Cooke, Dongying Gao, Eric Lasserre, Christel Llauro, Marie Mirouze, Nathalie Picault, Scott A. Jackson, and Olivier Panaud. Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Research*, 24(5):831–838, 2014. doi: 10.1101/gr.164400.113.
78. Roulin Anne, Piegu Benoît, Wing Rod A., and Panaud Olivier. Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon rire1 within the genus oryza. *The Plant Journal*, 53(6):950–959, 2007. doi: 10.1111/j.1365-313X.2007.03388.x.
79. Anne Roulin, Benoit Piegu, Philippe M. Fortune, François Sabot, Angélique D'Hont, Domenica Manicacci, and Olivier Panaud. Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the ltr-retrotransposon route66 in poaceae. *BMC Evolutionary Biology*, 9(1):58, Mar 2009. ISSN 1471-2148. doi: 10.1186/1471-2148-9-58.
80. Fei Lu, Maria C. Romay, Jeffrey C. Glaubitz, Peter J. Bradbury, Robert J. Elshire, Tianyu Wang, Yu Li, Yongxiang Li, Kassa Semagn, Xuecai Zhang, Alvaro G. Hernandez, Mark A. Mikel, Ilya Soifer, Omer Barad, and Edward S. Buckler. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications*, 6:6914, April 2015. ISSN 2041-1723. doi: 10.1038/ncomms7914.
81. Aurora Díaz, Meluleki Zikhali, Adrian S. Turner, Peter Isaac, and David A. Laurie. Copy Number Variation Affecting the Photoperiod-B1 and Vernalization-A1 Genes Is Associated with Altered Flowering Time in Wheat (Triticum aestivum). *PLoS ONE*, 7(3):e33234, mar 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0033234.
82. Tobias Würschum, Philipp H. G. Boeven, Simon M. Langer, C. Friedrich H. Longin, and Willmar L. Leiser. Multiply to conquer: Copy number variations at ppd-b1 and vrn-a1 facilitate global adaptation in wheat. *BMC Genetics*, 16(1):96, Jul 2015. ISSN 1471-2156. doi: 10.1186/s12863-015-0258-0.
83. David E Cook, Tong Geon Lee, Xiaoli Guo, Sara Melito, Kai Wang, Adam M Bayless, Jianping Wang, Teresa J Hughes, David K Willis, Thomas E Clemente, et al. Copy number variation of multiple genes at rhg1 mediates nematode resistance in soybean. *Science*, 338(6111):1206–1209, 2012.
84. Michael Alan Hardigan, Emily Crisovan, John P Hamiltion, Jeongwoon Kim, Parker Laimbeer, Courtney P Leisner, Norma C Manrique-Carpintero, Linsey Newton, Gina M Pham, Brieanne Vaillancourt, et al. Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated solanum tuberosum. *The Plant Cell*, pages TPC2015–00538, 2016.
85. Eric L Patterson, Dean J Pettinga, Karl Ravet, Paul Neve, and Todd A Gaines. Glyphosate resistance and epsps gene duplication: Convergent evolution in multiple plant species. *Journal of Heredity*, 109(2):117–125, 2017.
86. Yiyuan Li, Jianhui Xiao, Jiajie Wu, Jialei Duan, Yue Liu, Xingguo Ye, Xin Zhang, Xiuping Guo, Yongqiang Gu, Lichao Zhang, et al. A tandem segmental duplication (tsd) in green revolution gene rht-d1b region underlies plant height variation. *New Phytologist*, 196(1): 282–291, 2012.
87. Michiel Vos and Adam Eyre-Walker. Are pangenomes adaptive or not? *Nature Microbiology*, 2(12):1576, 2017. ISSN 2058-5276. doi: 10.1038/s41564-017-0067-5.
88. Wei Ding, Franz Baumdicker, and Richard A. Neher. panX: pan-genome analysis and exploration. *Nucleic Acids Research*, 46(1):e5, January 2018. ISSN 1362-4962. doi: 10.1093/nar/gkx977.
89. André Hennig, Jörg Bernhardt, and Kay Nieselt. Pan-Tetris: an interactive visualisation for Pan-genomes. *BMC bioinformatics*, 16 Suppl 11:S3, 2015. ISSN 1471-2105. doi: 10.1186/1471-2105-16-S11-S3.
90. Thomas Lin Pedersen, Intawat Nookaew, David Wayne Ussery, and Maria Månsson. PanViz: interactive visualization of the structure of functionally annotated pangenomes. *Bioinformatics*, 33(7):1081–1082, April 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw761.
91. Christine Jandrasits, Piotr W. Dabrowski, Stephan Fuchs, and Bernhard Y. Renard. seqseq-pan: building a computational pan-genome data structure on whole genome alignment. *BMC genomics*, 19(1):47, 2018. ISSN 1471-2164. doi: 10.1186/s12864-017-4401-3.
92. Thomas H. Clarke, Lauren M. Brinkac, Jason M. Inman, Granger Sutton, and Derrick E. Fouts. PanACEA: a bioinformatics tool for the exploration and visualization of bacterial pan-chromosomes. *BMC Bioinformatics*, 19(1):246, June 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2250-y.
93. Chen Sun, Zhiqiang Hu, Tianqing Zheng, Kuangchen Lu, Yue Zhao, Wensheng Wang, Jianxin Shi, Chunchao Wang, Jinyuan Lu, Dabing Zhang, Zhikang Li, and Chaochun Wei. RPAN: rice pan-genome browser for 3000 rice genomes. *Nucleic Acids Research*, 45(2): 597–605, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw958.
94. Astrid Collingro, Patrick Tischler, Thomas Weinmaier, Thomas Penz, Eva Heinz, Robert C. Brunham, Timothy D. Read, Patrik M. Bavoil, Konrad Sachse, Simona Kahane, Maureen G. Friedman, Thomas Rattei, Garry S. A. Myers, and Matthias Horn. Unity in Variety—The Pan-Genome of the Chlamydiae. *Molecular Biology and Evolution*, 28(12): 3253–3270, December 2011. ISSN 0737-4038. doi: 10.1093/molbev/msr161.
95. Ravi Kant, Johanna Rintahaka, Xia Yu, Pia Sigvart-Mattila, Lars Paulin, Jukka-Pekka Mecklin, Maria Saarela, Airi Palva, and Ingemar von Ossowski. A Comparative Pan-Genome Perspective of Niche-Adaptable Cell-Surface Protein Phenotypes in Lactobacillus rhamnosus. *PLOS ONE*, 9(7):e102762, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0102762.
96. Sandip Paul, Archana Bhardwaj, Sumit K. Bag, Evgeni V. Sokurenko, and Sujay Chattopadhyay. PanCoreGen — Profiling, detecting, annotating protein-coding genes in microbial genomes. *Genomics*, 106(6):367–372, December 2015. ISSN 0888-7543. doi: 10.1016/j.ygeno.2015.10.001.
97. Guillermo Nourdin-Galindo, Patricio Sánchez, Cristian F. Molina, Daniela A. Espinoza-Rojas, Cristian Oliver, Pamela Ruiz, Luis Vargas-Chacoff, Juan G. Cárcamo, Jaime E. Figueroa, Marcos Mancilla, Vinicius Maracaja-Coutinho, and Alejandro J. Yañez. Comparative Pan-Genome Analysis of Piscirickettsia salmonis Reveals Genomic Divergences within Genogroups. *Frontiers in Cellular and Infection Microbiology*, 7:459, 2017. ISSN 2235-2988. doi: 10.3389/fcimb.2017.00459.
98. Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet: Visualization of Intersecting Sets. *IEEE transactions on visualization and computer graphics*, 20(12):1983–1992, December 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346248.
99. Benedict Paten, Adam M. Novak, Jordan M. Eizenga, and Erik Garrison. Genome graphs and the evolution of genome inference. *Genome Research*, 27(5):665–676, May 2017. ISSN 1088-9051. doi: 10.1101/gr.214155.116.
100. Erik Garrison, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones, Michael F. Lin, Benedict Paten, and Richard Durbin. Sequence variation aware genome references and read mapping with the variation graph toolkit. *bioRxiv*, page 234856, December 2017. doi: 10.1101/234856.
101. Danny E. Miller, Cynthia Staber, Julia Zeitlinger, and R. Scott Hawley. High-quality genome assemblies of 15 drosophila species generated using nanopore sequencing. *bioRxiv*, 2018. doi: 10.1101/267393.
102. Lijun Ou, Dong Li, Junheng Lv, Wenchao Chen, Zhuqing Zhang, Xuefeng Li, Bozhi Yang, Shudong Zhou, Sha Yang, Weiguo Li, Hongzhen Gao, Qin Zeng, Huiyang Yu, Bo Ouyang, Feng Li, Feng Liu, Jingyuan Zheng, Yuhua Liu, Jing Wang, Bingbing Wang, Xiongze Dai, Yanqing Ma, and Xuexiao Zou. Pan-genome of cultivated pepper (Capsicum) and its use in gene presence–absence variation analyses. *New Phytologist*, 0(0). ISSN 1469-8137. doi: 10.1111/nph.15413.
103. Peng Zhou, Kevin A. T. Silverstein, Thiruvarangan Ramaraj, Joseph Guhlin, Roxanne Denny, Junqi Liu, Andrew D. Farmer, Kelly P. Steele, Robert M. Stupar, Jason R. Miller, Peter Tiffin, Joann Mudge, and Nevin D. Young. Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. *BMC Genomics*, 18(1):261, dec 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-3654-1.
104. Sara Pinosio, Stefania Giacomello, Patricia Faivre-Rampant, Gail Taylor, Veronique Jorge, Marie Christine Le Paslier, Giusi Zaina, Catherine Bastien, Federica Cattonaro, Fabio Marroni, and Michele Morgante. Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Molecular biology and evolution*, 33(10):2706–19, oct 2016. ISSN 1537-1719. doi: 10.1093/molbev/msw161.
105. Wensheng Wang, Ramil Mauleon, Zhiqiang Hu, Dmytro Chebotarov, Shuaishuai Tai, Zhichao Wu, Min Li, Tianqing Zheng, Roven Rommel Fuentes, Fan Zhang, Locedie Mansueto, Dario Copetti, Millicent Sanciangco, Kevin Christian Palis, Jianlong Xu, Chen Sun, Binying Fu, Hongliang Zhang, Yongming Gao, Xiuqin Zhao, Fei Shen, Xiao Cui, Hong Yu, Zichao Li, Miaolin Chen, Jeffrey Detras, Yongli Zhou, Xinyuan Zhang, Yue Zhao, Dave Kudrna, Chunchao Wang, Rui Li, Ben Jia, Jinyuan Lu, Xianchang He, Zhaotong Dong, Jiabao Xu, Yanhong Li, Miao Wang, Jianxin Shi, Jing Li, Dabing Zhang, Seunghee Lee, Wushu Hu, Alexander Poliakov, Inna Dubchak, Victor Jun Ulat, Frances Nikki Borja, John Robert Mendoza, Jauhar Ali, Jing Li, Qiang Gao, Yongchao Niu, Zhen Yue, Ma Elizabeth B. Naredo, Jayson Talag, Xueqiang Wang, Jinjie Li, Xiaodong Fang, Ye Yin, Jean-Christophe Glaszmann, Jianwei Zhang, Jiayang Li, Ruaraidh Sackville Hamilton, Rod A. Wing, Jue Ruan, Gengyun Zhang, Chaochun Wei, Nickolai Alexandrov, Kenneth L. McNally, Zhikang Li, and Hei Leung. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, 557(7703):43–49, May 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0063-9.