Computational Strategies for Eukaryotic Pangenome Analyses

Zhiqiang Hu, Chaochun Wei, and Zhikang Li

Abstract Over the last few years, pangenome analyses have been applied to eukaryotes, especially to important crops. A handful of eukaryotic pangenome studies have demonstrated widespread variation in gene presence/absence among plant species and its implications on agronomically important traits. In this chapter, we focus on the methodology of pangenome analysis, which can generally be classified into two different types of approaches, a homolog-based strategy and a "map-to-pan" strategy. In a homolog-based strategy, the genomes of individuals are independently assembled, and the presence/absence of a gene family is determined by clustering protein sequences into homologs. Alternatively, in a "map-to-pan" strategy, pangenome sequences are constructed by combining a well-annotated reference genome with newly identified non-reference representative sequences, from which the presence/absence of a gene is then determined based on read coverage after individual reads are mapped to the pangenome. We highlight the advantages and limitations of the homolog-based strategy and several variant approaches to the "map-to-pan" strategy. We conclude that the "map-to-pan" strategy is highly recommended for eukaryotic pangenome analysis. However, programs and parameters for pangenome analysis need to be carefully selected for eukaryotes with different genome sizes.

Z. Hu (✉)
Department of Plant & Microbial Biology, University of California, Berkeley, CA, USA

Institute of Crop Sciences/National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Haidian District, Beijing, China
e-mail: hu.zhiqiang@berkeley.edu

C. Wei
School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

Z. Li
Institute of Crop Sciences/National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Haidian District, Beijing, China

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_13

Keywords Pangenome · Plant pangenome · Gene presence–absence variation · Gene PAV · Map-to-pan

In 2005, Tettelin et al. introduced the concept of a pangenome, namely the entire gene set of a species, in their study of eight strains of *Streptococcus agalactiae*, that causes neonatal infection in humans (Tettelin et al. 2005). The pangenome is comprised of a "core-genome" that contains genes shared by all individuals of the species, and a "dispensable genome" containing genes present in some but not all individuals of the species. The core-genome is generally believed to be responsible for functions essential to the species, such as growth and development, whereas the dispensable genome confers functions related to environmental adaptations (Vernikos et al. 2015). During the past 10 years, pangenome studies have been widely applied to bacteria and other microorganisms. However, only a handful of pangenome analyses of higher eukaryotes have been reported (Wang et al. 2018; Hu et al. 2018; Sun et al. 2017; Zhao et al. 2018; Ou et al. 2018; Darracq et al. 2018; Montenegro et al. 2017; Pinosio et al. 2016; Golicz et al. 2016; Nguyen et al. 2015; Lu et al. 2015; Yao et al. 2015; Hirsch et al. 2014; Read et al. 2013; Li et al. 2010, 2014). In this chapter, we will first review the biological insights highlighted from these studies. Then, we will introduce current challenges and strategies for performing eukaryotic pangenome analysis, and finally, we will discuss future directions in this field.

Next-generation sequencing (NGS) technologies have enabled whole-genome sequencing and comparisons of multiple individual genomes within a species. Single nucleotide variations (SNPs), small insertions and deletions (InDels), and structural variations (SVs), including copy number variations (CNVs) and presence/absence variations (PAVs), can be identified when comparing against a reference genome. A considerable number of SVs have been observed among human (Sudmant et al. 2015; Genomes Project et al. 2015; Feuk et al. 2006) and animal genomes (Bickhart and Liu 2014). For example, a typical human genome contains 2100–2500 structural variants (including ~1000 large deletions), affecting ~20 Mb sequences when comparing with a reference genome (~3 Gb) (Genomes Project et al. 2015). In contrast, SVs have been reported to be more pervasive within plant genomes (Saxena et al. 2014), such as rice (Wang et al. 2018; Hu et al. 2018), arabidopsis (Cao et al. 2011), maize (Swanson-Wagner et al. 2010), sorghum (Zheng et al. 2011), and potato (Potato Genome Sequencing C et al. 2011). For example, the total sequences affected by SV that differentiate two typical rice accessions, on average, are about 22–70 M (out of ~380 M) (Wang et al. 2018). These results imply that there might be widespread presence of gene PAVs associated with SV sequences.

Pangenome analyses aim to study gene PAVs, providing a new functional interpretation of within-species variations. Compared to SV studies, pangenome analyses identify undiscovered genomic sequences and their associated genes and reveal the species core and dispensable genome. Early pangenome studies focused on comparisons among a small number (2–3) of well-assembled individual genomes

(Liu et al. 2007; Ma and Bennetzen 2004). These studies revealed the space of undiscovered genes and demonstrated widespread gene PAVs within a species. For instance, Li et al. assembled an Asian and an African genome, leading to the detection of 5 Mb sequences and hundreds of undiscovered genes that are absent in the human reference genome. Liu et al. sequenced ten thousand cDNAs of 93–11, a *Xian*(*indica)* rice accession, and found that >1000 genes were absent in the *Geng (japonica)* reference genome (Liu et al. 2007), which was believed to have diverged from *Xian* ~0.44 million years ago (Ma and Bennetzen 2004); later, Schatz et al. compared three assembled genomes of a *Xian* (IR64), a *Geng,* and an *aus* (DJ123) accession, and found that ~3000 genes were absent in at least one accession.

However, studying a small number of individuals cannot reveal the global landscape of gene PAVs of a species and cannot confidently identify the species core and dispensable genomes. Thus, systematic studies involving a large number of representative individuals within a species is highly desired. Large-scale plant pangenome studies involving tens to hundreds of individuals have emerged over recent years (Table 1). Many of these studies revealed that gene PAVs are a very important aspect of the genomic diversity within eukaryotic species/populations that can provide significant insights into evolutionary history of the species/populations with significant implications on the functional genomic research of important traits.

In *Emiliania huxleyi*, a marine phytoplankton important for carbon fixation in ecosystems, one-third of the genes in the reference genome are absent in the 13 sequenced individuals (Read et al. 2013). The core-genome controls inorganic nitrogen uptake/assimilation and nitrogen-rich compound acquisition/degradation, while the dispensable genome is in charge of metabolic repertoires, of which over one-fourth involve iron-binding activities and vitamin B1 and B12 synthesis (Read et al. 2013).

In rice, several studies consistently report that about ten thousand genes are missing in the widely used Nipponbare reference genome (Wang et al. 2018; Zhao et al. 2018; Yao et al. 2015), and almost all of them can be detected in wild rice (Wang et al. 2018). The dispensable genome accounts for >38% of the species pangenome and over one-fourth of a typical individual genome (Wang et al. 2018). On average, two *Xian* or *Geng* genomes differ by about 4000 (~10%) genes, respectively, whereas a *Xian* genome and a *Geng* genome differ by more than 6000 (~15%) genes (Wang et al. 2018). Although the dispensable genome is less studied, it appeared to harbor functions related to environmental adaptations, such as regulation of immune/defense responses and ethylene metabolism (Wang et al. 2018). Interestingly, the well-known Green Revolution gene, *sd-1*, coding a key enzyme, GA-oxidase20, in the biosynthesis of the important plant hormone, GA_1/GA_4, is a dispensable gene that associates with many important processes in plant growth, development, and responses to abiotic stresses (Wang et al. 2018; Zhao et al. 2018).

In *Brassica oleracea* (Golicz et al. 2016), bread wheat (Montenegro et al. 2017) and wild soybean (Li et al. 2014), it was reported that the dispensable genomes take up 18.7%, 20%, and 35.7% of the pangenomes, respectively. Although the pipelines and parameters/thresholds used to determine gene presence differed a lot in the above studies, it is well demonstrated that plants exhibit considerably large

Table 1 Representative pangenome studies

Species	Haploid genome size (bps)[a]	N	References	Strategy	Comment
Homo sapiens (human)	2991 M	3	Li et al. (2010)	Directly comparing two de novo assembled individual human genomes (an Asian and an African) with the human reference genome.	19~40 Mb sequences containing >150 genes cannot be found in the reference.
Emiliania huxleyi (coccolithophore)	168 M	14	Read et al. (2013)	Building a reference genome from an individual genome; assembling 3 additional individual genomes and comparing them with the reference genome; determining presence/ absence of reference genes by mapping short reads of additional 10 individuals to the reference.	>1300 reference genes are not present in the 3 individual genomes; the core-genome accounts for 2/3 of the reference genes.
Zea mays (maize)	2135 M	503	Hirsch et al. (2014)	Sequencing the transcriptome of 503 accessions. Assembling genes from transcriptome sequencing. A gene with FPKM > 0 is considered as present.	Identifying ~8600 representative transcript assemblies (RTAs) absent in the B73 reference; 16.4% RTAs express in all lines and 82.7% express in subsets of the lines.
Glycine soja (wild soybean)	924 M	7	Li et al. (2014)	Sequencing and de novo assembling 7 individuals' genomes. Clustering annotated genes to gene families.	Dispensable genome accounts for 20% of the pangenome, and displays greater sequence variation than the core-genome.
Oryza sativa (rice)	374 M	1483	Yao et al. (2015)	Aligning low-depth (1~3x)	Detecting ~9000 genes for the

(continued)

Table 1 (continued)

Species	Haploid genome size (bps)[a]	N	References	Strategy	Comment
				reads to a pangenome; building the dispensable genome by assembling the pool of unaligned reads from each individual. *Indica* and *japonica* accessions are separately studied.	*indica* dispensable genome and >6000 genes for *japonica* dispensable genome.
Brassica oleracea	514 M	9	Golicz et al. (2016)	Using a reference-based iterative strategy to assemble the pangenome: (1) mapping reads to the reference sequence; (2) assembling unmapped reads; (3) and updating the reference. Determine gene PAV by mapping short reads to the pangenome.	Dispensable genome accounts for 18.7% of the pangenome.
Triticum aestivum (bread wheat)	13,672 M	18	Montenegro et al. (2017)	Building a reference genome; Constructing the pangenome sequences by combining the reference genome and non-reference sequences, which are assembled from the pool of unaligned reads from each individual. Determining gene presence/ absence by mapping short reads to the pangenome.	Dispensable genome accounts for 35.7% of the pangenome.

(continued)

Table 1 (continued)

Species	Haploid genome size (bps)[a]	N	References	Strategy	Comment
Oryza sativa (rice)	374 M	453	Wang et al. (2018), Hu et al. (2018), Sun et al. (2017)	Assembling 3010 individual genomes independently; building representative non-reference sequences by removing the redundant sequences from the pool of contigs that are unaligned to the reference. Constructing a pangenome by combining the reference genome and representative non-reference sequences. Determining gene presence/absence for 453 individuals with sequencing depth >20 by mapping short reads to the pangenome.	Discover 283 M non-reference sequences with >10,000 genes; Dispensable genome accounts for 35.7% of the pangenome. Dispensable genes tend to be younger, shorter, exhibiting higher level of SNPs.
Capsicum (including 4 species) (pepper)	3095 M	383	Ou et al. (2018)	Using the same strategy as the above rice study.	Discover 956 M non-reference sequences with >50,000 genes; 55.7% of the pangenome show >50% presence frequencies in all the 4 species.
Oryza sativa and *Oryza rufipogon* (rice and wild rice)	374 M	66	Zhao et al. (2018)	Sequencing and de novo assembling 66 individual genomes. Clustering annotated genes to gene families.	Discover >10,000 non-reference genes; 62% of the pangenome can be found in ≥60 individuals.

[a] The genome size was obtained from NCBI genome database. It can be the size of a reference genome or the average size of several independent assemblies

dispensable genomes, harboring functions related to many agronomically important traits. Moreover, several studies consistently demonstrate that dispensable genes tend to be younger (Wang et al. 2018; Chen et al. 2012; Bush et al. 2013), shorter (Wang et al. 2018; Bush et al. 2013; Schatz et al. 2014), have less exons (Wang et al. 2018; Bush et al. 2013; Schatz et al. 2014), harbor a much higher level of sequence variations (Wang et al. 2018; Li et al. 2014), and have fewer paralogs (Wang et al. 2018; Bush et al. 2013).

1 Eukaryotic Pangenome Analysis Strategy

Because pangenome is a property of a species/population, any desirable pangenome study should seriously consider its sampling strategy such that the maximum gene PAVs can be detected with a minimum number of samples. According to the core collection concept in plant genetic resources (Frankel and Brown 1984), a core collection of a plant species germplasm consisting of limited but well-sampled accessions of a plant species would represent the whole spectrum of its total within-species diversity. In practice, a well-established semi-stratified sampling strategy considering both the center(s) of diversity/origin and geographic distribution of a plant species has demonstrated that the core collection containing only 5% of the total collected accessions of a species would cover ~95% of the total species diversity (Jia et al. 2017). Obviously, this concept should equally be applicable to pangenome research of animal species.

For the analytic methodology, almost all bacterial pangenome analyses follow a homolog-based strategy (Fig. 1) involving (1) de novo assembly of individual genomes; (2) independent annotation of protein-coding genes in each assembled genome; and (3) pooling all protein sequences together and clustering them into homologs (gene families) or orthologs using protein clustering tools (Steinegger and Söding 2018; Fu et al. 2012) or ortholog grouping tools (Emms and Kelly 2015; Li et al. 2003). Gene family presence/absence in each individual can be retrieved from

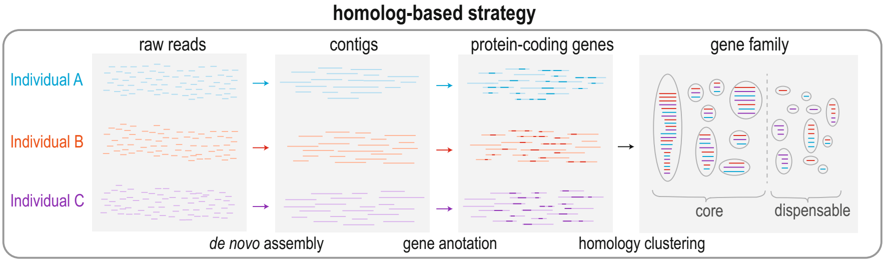


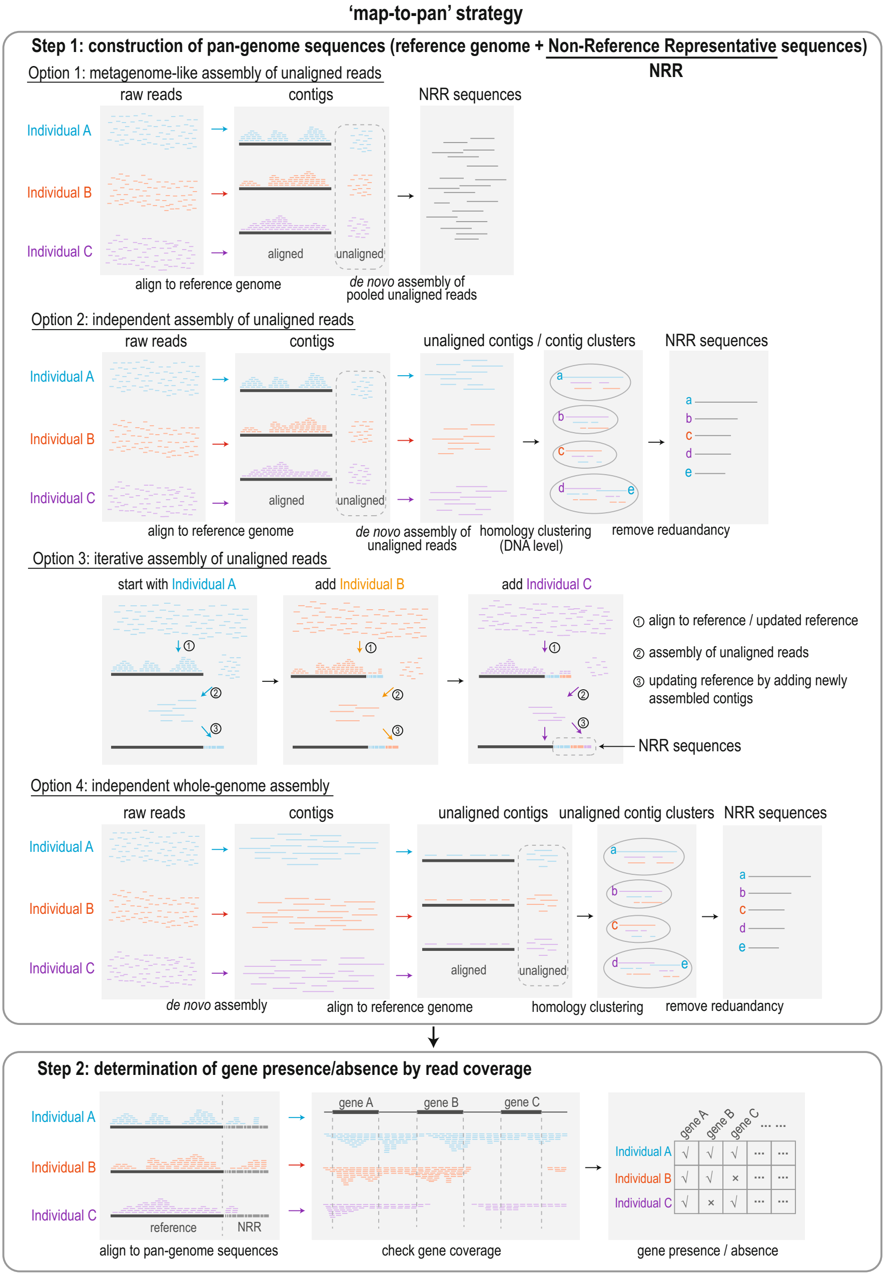
Fig. 1 Homolog-based strategy for pangenome analyses. This strategy is widely used for bacterial pangenome analyses. It includes the following steps: (1) independent assemblies of individual whole genomes; (2) annotation of protein-coding genes for each genome; and (3) clustering genes to homologs (gene families) to determine the presence/absence of each gene family

the clustering results. This strategy is highly dependent on the completeness of the whole-genome assembly. Failure in assembling a sequence segment will lead to calling the absence of all genes located on this sequence segment. Moreover, the protein similarity threshold for gene family determination may impact the size and even the relative size of the core-genome and pangenome.

Several challenges hinder applying a homolog-based strategy to eukaryotic genomes. First, eukaryotic genomes are usually large, ranging from hundred millions of bases to billions of bases, and possess a high level of repetitive sequences, making whole-genome assembly challenging. Several approaches can help improve the assembly, including increasing the sequencing depth, sequencing multiple libraries with diverse insertion sizes, and integrating long-read sequencing technologies (Rhoads and Au 2015; Schneider and Dekker 2012). However, all of these approaches significantly increase the cost of whole-genome assembly, thus limiting the number of individuals involved in a study. Second, eukaryotes have split gene structures. Automatic gene annotation may be inaccurate and lead to biased results. Even with these challenges, there are several studies following the homolog-based strategy. Li et al. sequenced seven wild soybean genomes using Illumina technology, each with three libraries (insertion sizes of 180 bp, 500 bp, and 2000 bp) (Li et al. 2014). The average overall sequencing depth was 112x. Based on this data, they were able to assemble ~89% of the genome. Recently, Zhao et al. sequenced 66 rice and wild rice accessions, each with two libraries (insertion sizes of 400 bp and 700 bp) (Zhao et al. 2018). The average sequencing depth reached 115x, and they were able to assemble ~85% of the genomes. Notably, a significant portion of individual genomes were not assembled in both studies. The associated genes were labeled as "absent" in the corresponding individuals. However, given that these false negatives repeatedly happen for certain genes within repeat-rich regions, they can be treated as systematic errors. The overall results may be still meaningful.

Reference-based genomic studies are prevalent in eukaryotes. Researchers have been taking tremendous efforts to build more complete reference genomes and providing confident gene annotations for important species. These reference genomes and their annotated genes are the basis for modern genomics studies. Moreover, reference-based genomic variants show a great power in explaining phenotypic variations when used as markers for genome-wide association analyses. Therefore, when introducing the pangenome concept to eukaryotic genomic analyses, taking advantage of a pre-existing well-annotated reference genome is a straightforward choice. Following this idea, the "map-to-pan" strategy became prevalent for eukaryotic pangenome studies, especially when the target genome is extremely large or the study involves a large number of individuals (Fig. 2). The "map-to-pan" strategy includes two main steps: construction of pangenome sequences by combining the reference genome and non-reference representative (NRR) sequences (upper panel of Fig. 2) and determination of the presence/absence of each gene in each individual by mapping reads to the pangenome and examining the gene coverage (lower pane of Fig. 2).

Several approaches for detecting NRR sequences have been reported (Wang et al. 2018; Ou et al. 2018; Montenegro et al. 2017; Yao et al. 2015; Read et al. 2013; Li


'map-to-pan' strategy
Step 1: construction of pan-genome sequences (reference genome + Non-Reference Representative sequences)
NRR
Option 1: metagenome-like assembly of unaligned reads
raw reads
contigs
NRR sequences
Individual A
Individual B
Individual C
aligned
unaligned
align to reference genome
de novo assembly of pooled unaligned reads
Option 2: independent assembly of unaligned reads
raw reads
contigs
unaligned contigs / contig clusters
NRR sequences
Individual A
Individual B
Individual C
aligned
unaligned
a
b
c
d
e
align to reference genome
de novo assembly of unaligned reads
homology clustering (DNA level)
remove reduandancy
Option 3: iterative assembly of unaligned reads
start with Individual A
add Individual B
add Individual C
① align to reference / updated reference
② assembly of unaligned reads
③ updating reference by adding newly assembled contigs
NRR sequences
Option 4: independent whole-genome assembly
raw reads
contigs
unaligned contigs
unaligned contig clusters
NRR sequences
Individual A
Individual B
Individual C
aligned
unaligned
a
b
c
d
e
de novo assembly
align to reference genome
homology clustering
remove reduandancy
Step 2: determination of gene presence/absence by read coverage
Individual A
Individual B
Individual C
reference
NRR
align to pan-genome sequences
gene A
gene B
gene C
check gene coverage
gene presence / absence

	gene A	gene B	gene C	···	···
Individual A	√	√	√	···	···
Individual B	√	√	×	···	···
Individual C	√	×	√	···	···

Fig. 2 "Map-to-pan" strategy for pangenome analyses. This strategy is mostly used for eukaryotic pangenome analyses. It includes two main steps: (1) construction of pangenome sequences by integrating a reference genome and assembled non-reference sequences; (2) determination of presence/absence of each gene (both reference genes and non-reference predicted genes) based on read coverage. Four strategies for obtaining non-reference representative sequences are introduced

et al. 2014). Yao et al. utilized a metagenome-like assembly of mixed unaligned reads from 1483 rice accessions with extremely low sequencing depth (1~3x) (Yao et al. 2015) (Option 1 in Fig. 2), enabling the detection of ~9000 non-reference genes. This approach assembled NRR sequences using heterozygous reads and may generate chimeric contigs, especially when considering that non-reference sequences may exhibit higher levels of repetitive sequences. A variant of this option (Option 2 in Fig. 2) is to assemble the unaligned reads from each individual separately and retrieve NRR sequences using DNA homology clustering strategies, such as CD-HIT-EST (Fu et al. 2012), UCLUST (Edgar 2010), MeShClust (James et al. 2018), etc. Golicz et al. utilized an iterative assembly approach (Option 3 in Fig. 2), iteratively conducting the following three steps: mapping of the reads to a pseudo pangenome (starting with the reference genome); assembling the unmapped reads; and updating a new pseudo pangenome with new sequences added (Golicz et al. 2016). They demonstrated that the sizes of final assemblies were similar regardless of the order of individuals added into the iterative process. However, an improper ordering may lead to fragmented assemblies. Alternatively, Hu et al. proposed an integrated approach (implemented in EUPAN toolkit (Hu et al. 2017)) (Option 4 in Fig. 2): (1) independent assembly of individual genomes; (2) generation of NRR sequences from homology clustering of all unaligned contigs. This approach has the benefit of not involving chimeric sequences as well as keeping better sequence completeness. This approach has also been recently applied to hundreds of rice genomes (Wang et al. 2018; Hu et al. 2018; Sun et al. 2017) and the 383 Capsicum genomes (Ou et al. 2018). This strategy will perform better than Option 2 in the scenario where a novel sequence contains a short reference segment (likely to be repetitive sequences) in the middle; option 2 will assemble two segmented segments instead. However, the process of whole-genome assembly is computationally intensive, hindering its application to extremely large genomes. In summary, pooling of low-depth sequenced genomes may also contribute to pangenome construction (Option 1). Options 2–4 are preferable if sequencing depth is high enough for independent assemblies. Options 2–3 are extremely useful for eukaryotes with very large genomes (e.g., the bread wheat with a haploid genome of >13Gb).

After the construction of pangenome sequences, gene presence/absence can be determined by examining gene coverage when raw reads are mapped to the pangenome (lower panel of Fig. 2). Remarkably, very different thresholds have been applied to determine a gene's presence. For example, Wang et al. considered a gene's presence as CDS coverage ($\geq$1 read) over 0.95 and gene body coverage over 0.85 (Wang et al. 2018); Ou et al. treated a gene's presence as CDS coverage ($\geq$1 read) over 0.6 and gene body coverage over 0.5 (Ou et al. 2018); Read et al. considered a gene's presence as gene body coverage ($\geq$1 read) over 0.5 (Read et al. 2013; Montenegro et al. 2017; Golicz et al. 2016) used a threshold of exon coverage over 0.05. Unfortunately, such divergent thresholds make the quantitative cross-species comparisons of gene PAV-related features meaningless. Theoretically, with a high-enough sequencing depth, a gene's presence is equal to that the gene, at least the CDS, should be fully covered. Loss of partial sequences of a gene, defined as a "functional unit," may cause a loss of gene function. Setting up gene body coverage cutoffs will help distinguish retro-transcribed pseudo-genes from their original ancestries. In reality, certain genomic regions may be not covered due to both insufficient sequencing depth and unevenness of the sequencing. One plausible solution is to lower the thresholds. However, the sequencing depth difference may further lead to inconsistencies in sensitivities of gene presence determination among individuals; individuals with higher sequencing depth would contain more genes. Another possible solution is to study the presence/absence of gene families instead of genes by calculating "gene presence" using a low threshold and determining gene family presence based on "gene presence." In this scenario, the unbalanced sequencing depths also need to be fixed either by sampling to equal depths or setting up dynamic thresholds based on the sequencing depth. Nevertheless, it is not recommended to determine gene presence/absence from low-depth sequencing data. Gene presence/absence should only be studied and compared for individuals with sufficient sequencing data, that is, when mapping to the pangenome, the coverage of the genome should be saturated. For example, Wang et al. mapped raw reads of ~3000 rice accessions to the reference genome and found that genome coverage is stable when sequencing depth exceeds 20x; therefore, gene presence/absence was only studied for a selected set of 453 accessions with sequencing depth >20 (Wang et al. 2018).

The "map-to-pan" strategy also exhibits better accuracy. A pangenome study can be technically evaluated at two levels: (1) the accuracy of pangenome (gene annotation and gene completeness) and (2) the accuracy of gene presence/absence calling. The "map-to-pan" strategy utilizes reference sequences and their annotations directly. Strategies using a whole-genome assembly (homolog-based, and option 4 of the "map-to-pan" strategy) will have a higher possibility of detecting complete gene sequences. At the gene presence/absence level, the homolog-based strategy has a bottleneck in assembling a complete genome, and "map-to-pan" strategies definitely show better accuracy when sequencing depth is high enough (Hu et al. 2017).

After determination of gene presence/absence, similar analyses as seen in bacterial pangenome studies can be performed for eukaryotes, including but not limited to (1) simulating the pangenome and core-genome sizes; (2) constructing phylogenic relationships based on gene presence/absence; and (3) exploring functions related to the dispensable genome or to a specific dispensable gene.

2 Future Directions

In summary, the pangenome is an important property of any eukaryotic species/populations and gene PAVs represent a very important dimension of within-species/population diversity that remains uncharacterized in most eukaryotic species. As the costs in genome sequencing decrease, one would expect the pangenome analyses to be carried out in more and more species, firstly in most important and/or model plant and animal species, and then to natural populations of wild species. Thus, eukaryotic pangenome research in the next several years should focus on revealing within-species/population gene PAVs and building the pan-references for species of interest. The pan-reference of a species should include the reference illustrating (1) all the sequences within the species, (2) the connections of alternative sequence segments and (3) the genotype likelihoods (allele frequencies) such that all possible mechanisms (SVs and distribution/activities of transposable elements) potentially responsible for pangenome expansion and generation of gene PAVs can be clearly represented and understood. As pangenomes and gene PAVs are revealed in more and more plant and animal species, the eukaryotic pangenome research will be naturally extended to the comparative pangenome analyses, focusing on comparisons of the pangenome constitution between or among related species. Results from this kind of research are expected to provide new insights into the evolutionary history of eukaryotic species. For example, comparisons between related species or between different populations of the same species in portions of the core and dispensable genes/gene families in their pangenomes and their patterns how new gene emerged will provide important information on their evolutionary history. Expectedly, emergences of new species would be accompanied with bursts of new gene emergences, while major distinctions with massive gene losses in evolution. Also, it would be of great interest to compare the core-genome constitution between related species and to compare the dispensable genome constitution between different populations of the same species. In the former cases, one may see the differences in key genes and their functionalities between related species. In the latter cases, one may discover important sets of genes contributing to adaptations to specific environments important for future plant and animal improvements. In this respect, genome-wide association analyses of important traits based on pangenome SNPs or based on gene PAVs should be widely adopted (Hu et al. 2018).

As more eukaryotic pangenome analyses are expected to emerge, the technical strategy and methodology in analyses of eukaryotic pangenomes need to be improved. Because of the relatively high genome sequencing and analytic costs in eukaryotic pangenomes, the NGS technology will remain the primary technology for the pangenome studies of most eukaryotes in the short term, particularly for those species of very large genomes, and so for the “map-to-pan” strategy elaborated in detail here. However, before applying this strategy, specific attentions should be paid to the sampling strategy to make sure representative individuals of minimum sample size of the target species or population to be used, and to the selection and evaluation of parameters of the map-to-pan methodology. In the presentation and storage of results from the eukaryotic pangenome analyses, graph-based data structures are highly desirable and should be widely used in pan-reference storage and visualization (Zekic et al. 2018; Marschall et al. 2018; Baier et al. 2016). Pioneer work has been done in the human genome research, where the NRR sequences might be of a small size. Alternative sequences of highly variable regions were added to human reference genome, starting with GRCh37 (Church et al. 2011). Alternative sequences were anchored to locations along the primary assembly. Besides the limited NRR sequences, a large number of SNPs, InDels, and SVs (deletions, duplications, and translocations) can also be integrated into the pan-reference (Zekic et al. 2018; Marschall et al. 2018; Baier et al. 2016). What is more, read alignment tools and variant-calling tools working on the graph-based pan-reference will be required. However, for plant species of high within-species sequence diversity, the challenge is how to anchor large numbers of NRR sequences, whose sizes may be as large as half of the reference genome. Finally, considering the prediction of "new" or novel genes based on simple thresholds of sequence homology without detailed information on gene functionality is always somewhat arbitrary, the pangenome results based on the NGS technology can be validated and improved greatly if high-quality reference genomes of relatively few representative individuals are included in a pangenome study, particularly for important model species of relatively small genome sizes.

References

Baier U, Beller T, Ohlebusch E (2016) Graphical pan-genome analysis with compressed suffix trees and the Burrows-Wheeler transform. Bioinformatics 32:497–504

Bickhart DM, Liu GE (2014) The challenges and importance of structural variation detection in livestock. Front Genet 5:37

Bush SJ, Castillo-Morales A, Tovar-Corona JM, Chen L, Kover PX, Urrutia AO (2013) Presence–absence variation in *A. thaliana* is primarily associated with genomic signatures consistent with relaxed selective constraints. Mol Biol Evol 31:59–69

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43:956–963

Chen W-H, Trachana K, Lercher MJ, Bork P (2012) Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol Biol Evol 29:1703–1706

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R, McLaren WM, Ritchie GR (2011) Modernizing reference genome assemblies. PLoS Biol 9: e1001091

Darracq A, Vitte C, Nicolas S, Duarte J, Pichon JP, Mary-Huard T, Chevalier C, Berard A, Le Paslier MC, Rogowsky P et al (2018) Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. BMC Genomics 19:119

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

Frankel O, Brown A (1984) Current plant genetic resources – a critical appraisal. In: Chopra VL et al (eds) Genetics: new frontiers: proceedings of the XV international congress of genetics. Oxford & IBH Publishing Co., c1984, New Delhi

Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526:68–74

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA et al (2016) The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun 7:13390

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Penagaricano F, Lindquist E, Pedraza MA, Barry K et al (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26:121–135

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

Hu Z, Wang W, Wu Z, Sun C, Li M, Lu J, Fu B, Shi J, Xu J, Ruan J et al (2018) Novel sequences, structural variations and gene presence variations of Asian cultivated rice. Sci Data 5:180079

James BT, Luczak BB, Girgis HZ (2018) MeShClust: an intelligent tool for clustering DNA sequences. Nucleic Acids Res 46(14):e83

Jia J, Li H, Zhang X, Li Z, Qiu L (2017) Genomics-based plant germplasm research (GPGR). Crop J 5:166–174

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

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J et al (2010) Building the sequence map of the human pan-genome. Nat Biotechnol 28:57–63

Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L et al (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32:1045–1052

Liu XH, Lu TT, Yu SL, Li Y, Huang YC, Huang T, Zhang L, Zhu JJ, Zhao Q, Fan DL et al (2007) A collection of 10,096 indica rice full-length cDNAs reveals highly expressed sequence divergence between *Oryza sativa indica* and *japonica* subspecies. Plant Mol Biol 65:403–415

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X et al (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun 6:6914

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404–12410

Marschall T, Marz M, Abeel T, Dijkstra L, Dutilh BE, Ghaffaari A, Kersey P, Kloosterman WP, Makinen V, Novak AM et al (2018) Computational pan-genomics: status, promises and challenges. Brief Bioinform 19:118–135

Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan CK, Visendi P, Lai K, Dolezel J, Batley J, Edwards D (2017) The pangenome of hexaploid bread wheat. Plant J 90:1007–1013

Nguyen N, Hickey G, Zerbino DR, Raney B, Earl D, Armstrong J, Kent WJ, Haussler D, Paten B (2015) Building a pan-genome reference for a population. J Comput Biol 22:387–401

Ou L, Li D, Lv J, Chen W, Zhang Z, Li X, Yang B, Zhou S, Yang S, Li W (2018) Pan-genome of cultivated pepper (Capsicum) and its use in gene presence-absence variation analyses. New Phytol 220:360

Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, Zaina G, Bastien C, Cattonaro F, Marroni F, Morgante M (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. Mol Biol Evol 33:2706–2719

Potato Genome Sequencing C, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R et al (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195

Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A, Salamov A et al (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499:209–213

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289

Saxena RK, Edwards D, Varshney RK (2014) Structural variations in plant genomes. Brief Funct Genomics 13:296–307

Schatz MC, Maron LG, Stein JC, Hernandez Wences A, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E 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 Biol 15:506

Schneider GF, Dekker C (2012) DNA sequencing with nanopores. Nat Biotechnol 30:326

Steinegger M, Söding J (2018) Clustering huge protein sequence sets in linear time. Nat Commun 9:2542

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH et al (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Sun C, Hu Z, Zheng T, Lu K, Zhao Y, Wang W, Shi J, Wang C, Lu J, Zhang D et al (2017) RPAN: rice pan-genome browser for approximately 3000 rice genomes. Nucleic Acids Res 45:597–605

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20:1689–1699

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

Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F et al (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49

Yao W, Li G, Zhao H, Wang G, Lian X, Xie W (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol 16:187

Zekic T, Holley G, Stoye J (2018) Pan-genome storage and analysis techniques. Methods Mol Biol 1704:29–53

Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T et al (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50:278–284

Zheng LY, Guo XS, He B, Sun LJ, Peng Y, Dong SS, Liu TF, Jiang SY, Ramachandran S, Liu CM, Jing HC (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). Genome Biol 12:R114