SCIENCE CHINA
Life Sciences

•INSIGHT•

https://doi.org/10.1007/s11427-020-1808-0

From one linear genome to a graph-based pan-genome: a new era for genomics

Yucheng Liu[1,2] & Zhixi Tian[1,2*]

[1]*State Key Laboratory of Plant Cell and Chromosome Engineering, Center for Genome Editing, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing 100101, China;*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*

Received July 31, 2020; accepted August 25, 2020; published online September 7, 2020

Citation: Liu, Y., and Tian, Z. (2020). From one linear genome to a graph-based pan-genome: a new era for genomics. Sci China Life Sci 63, https://doi.org/10.1007/s11427-020-1808-0

The seeds for genomics were sown with the development of DNA sequencing (Sanger et al., 1977), cultivated with each advance in molecular biology, and have since grown into one of the most important aspects of the life sciences. With the sequencing of the first draft human genome (Venter et al., 2001), the era of post-genomics began. Once one genome has been sequenced, it can be used as the reference for a certain species, and sequences from individuals can be mapped onto it to compare genetic variations among different lines. This allows further population studies, such as whole genome genotyping, molecular evolution, and identification of trait-conferring loci. Construction of a reference genome has become a prerequisite for deeper functional analyses. However, the increasing number of genomics and genetics studies have shown us that a reference genome cannot fully represent the entire genetic variation of a species. Instead, we must consider to what extent a reference genome is representative (Ballouz et al., 2019; Yang et al., 2019). In this insight, we discuss the trends and challenges of constructing reference genomes (Figure 1).

From one genome to a pan-genome. In genomic studies, we always find a significant reduction in overall nucleotide diversity between the genomes of the domesticated species and the wild species. However, this reduction in diversity coexists with more divergent phenotypes for some traits in the cultivated species. Moreover, larger structural variants (SVs), which are rarely detectable by mapping short reads against one reference genome, play important roles in conferring some agronomical trait variations (Golicz et al., 2016a; Sherman and Salzberg, 2020). The use of just one or a few reference genomes in functional genomic studies may underestimate the genetic divergence and miss some key variants (Danilevicz et al., 2020). Therefore, there is a need to move towards pan-genome construction (Ameur, 2019; Golicz et al., 2016a; Sherman and Salzberg, 2020).

The prefix *pan-* comes from Greek and means “all” or “everything”. A pan-genome tries to encompass all sequenced genomes of a species to better represent the diverse regions within the genome. The first pan-genome was constructed by sequencing the genomes of 8 *Streptococcus agalactiae* strains (Tettelin et al., 2005). Through comparative genomic analyses, the pan-genome was delineated, which consists of a core set of genes shared by all isolates and a dispensable set of genes that includes partially shared and strain-specific genes. In this microbe, the core genome accounted for ~80% of any single genome. Although this work sketched an outline for pan-genomic analyses, investigations in more complexed eukaryotic species did not rapidly follow. This is partially because genomes of eukaryotes are much larger than those of bacteria, which means that pan-genomic studies in eukaryotes are more expensive, in part due to the costs for whole genome sequencing and the required computing resources. In addition, eukaryotic gen-

*Corresponding author (email: zxtian@genetics.ac.cn)

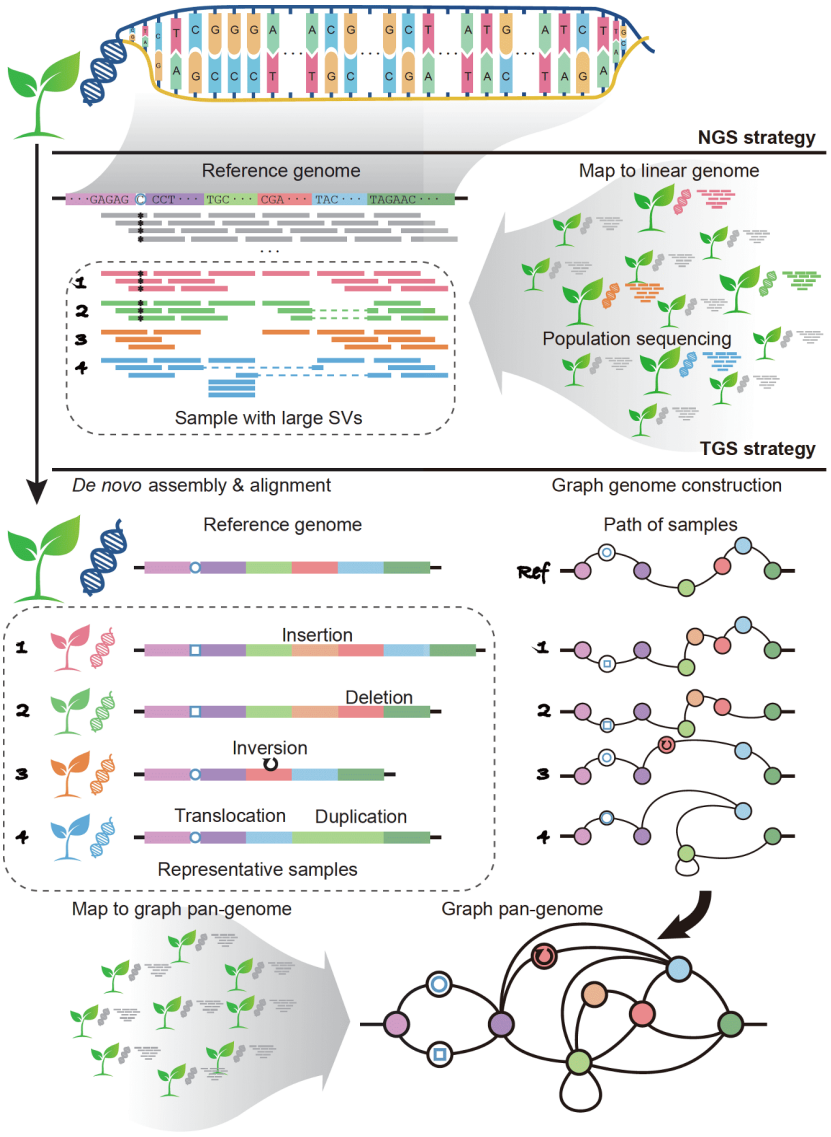

Figure 1 Trends and strategies for reference genome construction. Currently, linear reference genome is a commonly used form for genomic studies. To encompasses all sequenced genomes of a species by the most representing the diversities within the genome, there is a need to move towards graph-based pan-genome.

omes are more complex than those of bacteria. For instance, eukaryotic genomes contain more repetitive sequences, increasing the difficulty in delineating the pan-genome composition.

The quality of a pan-genome analysis is largely determined by the choice of samples, the quality of the assembly and annotation, and by the method used to detect nucleotide variation (Golicz et al., 2016a). Although the concept of the pan-genome can be easily adopted, there is no standard strategy for analysis of a high-quality pan-genome. One strategy is to map short-reads from different lines to a reference genome. With this "map-to-genome" method, some novel sequences can be identified by assembling the unmapped short-reads into new contigs. This method takes full advantage of high through-put next-generation sequencing (NGS) technology, which can investigate a larger population at a low cost. However, it largely relies on an existing reference genome, which will miss large amounts of individual genetic information from the non-reference lines, particular for the larger SVs. A more ideal strategy uses *de novo* assembly of individual genomes followed by comparative genomic analyses. While use of high-quality individual genomes makes it easier to detect genetic variants from different lines, this will cost more, which in turn restricts the population size. However, ongoing development of sequencing technologies is bringing the costs down, making large-scale "*de novo* sequence-and-compare" analyses practicable.

The first rudimentary pan-genome study in plants was reported for *Arabidopsis thaliana* in 2011. Through assembly and comparison of the single-copy genomes of 18 natural

lines, Gan et al. (2011) identified a total of ~28.3 Mb of non-redundant variations, which accounted for 4.5–7.6 Mb in each single sample. Since then, pan-genomic studies have gradually become more common in plants. Over the last several years, an increasing number of reports have indicated that the core genome of a plant accounts for 40% to 70% of its entire genome (Golicz et al., 2016a).

Pan-genomics has found that SVs not only contribute to genome variation, but also confer phenotypic variations. A recent deep inspection of the PanSV genome in tomato revealed that almost half of the SVs overlapped with genes or regulatory sequences and that half of the SVs affecting coding sequences were associated with differential gene expression (Alonge et al., 2020). Some of the SVs that changed gene dosage and expression levels are responsible for the modification of important traits, including fruit flavor, size, and production. These findings highlight the importance and utility of SVs in crop improvement.

From linear genome to graph-based genome. Currently, the genetic information of a species is presented as a reference genome organized by linear scaffolds that list the nucleotide sequences in order. A pan-genomics analysis tends to construct an integrated genome that by the most represents the genetic variations within a species. Most recently reported pan-genome analyses more likely added multiple individual genomes to a reference genome, with some additional sequences. Nevertheless, the combination of multiple genomes does not simplify the analyses, because the data size and dimension are increased from one to multiple. Instead, we must construct new types of reference genome that can fully represent the genetic variations of a species and make subsequent genomic analyses efficient.

One strategy is to construct an integrated reference genome by iterative assembly. Beginning with one reference genome, iterative assembly first maps the reads from each sample to the reference and then modifies the reference directly with non-redundant sequence. Using this methodology, an integrated pan-genome was built in cabbage, which added 99 Mb of sequence to the reference genome (Golicz et al., 2016b). The iterative assembly genome is highly integrated and remains in a linear form that is readable by traditional approaches. However, during iterative assembly, the non-redundant sequences have to be added by continually overwriting the former genome, which results in a loss of the natural characters of the original reference genome. Therefore, an iterative assembly genome cannot represent the individual genomic sequence of each line (Sherman and Salzberg, 2020).

Another strategy is to construct an integrated reference genome using a graph methodology. First, one genome is set as the reference, then genetic variants among the different genomes are determined through comparative genomic analyses. Thereafter, based on the relationships between reference and variant sequences, the reference and the alternative genetic variants are recorded as nodes and edges, respectively, and then stored in a graph form (Iqbal et al., 2012). While the graph-based genome is less readable by traditional approaches, it compresses and better maintains the genetic information from each line. Moreover, the compressed graph form enables fast and accurate computation, such as NGS data mapping and variation calling (Ameur, 2019).

In the last few years, several tools have been developed to craft genome graphs, including vg (Garrison et al., 2018), HISAT2 (Kim et al., 2019), and GraphTyper2 (Eggertsson et al., 2019). While these techniques have been used for human genome analyses, they have seldomly been performed for other species. A recent work on soybean constructed a graph-based pan-genome for the first time in a plant (Liu et al., 2020). Through *de novo* assembly of 26 soybean genomes, ~70 kb non-redundant, low-repeat presence/absence variations (PAVs) were found. Using the ZH13 genome as the reference, a graph-based genome with the PAV dataset was constructed. Through alignments of the short reads from ~3,000 soybean accessions and a genome-wide association study, a PAV highly associated with variation in seed lustre was identified. This is an initial but meaningful attempt of graph-based pan-genome analysis in a crop species.

New start, new challenges. Over the past 40 years, as sequencing technology has rapidly developed, the amount of available sequencing data has exponentially grown, either from resequencing data over a larger population or from *de novo* assembled genomes (Shendure et al., 2017). The importance of pan-genome analyses has been increasing realized (Danilevicz et al., 2020; Golicz et al., 2016a; Sherman and Salzberg, 2020). Along with the development of new sequencing technologies, sequence quality will be further improved, while cost will be largely reduced. This will lead to a continual generation of larger amounts of high-quality sequencing data. Over time, sample size may no longer be a limitation for pan-genome analyses. Eventually, we may be able to sequence "all" the samples and generate a genome that indeed represents the entire genetics of a species.

Nevertheless, we may not be satisfied with intra-species pan-genome construction. Further explorations could include 3D pan-genomics, a pan-transcriptome, or inter-species pan-genomes. In fact, several attempts have been initiated. One project is the "Earth BioGenome Project", which aims to assemble representative genomes for all known eukaryotic species (Lewin et al., 2018). Similarly, the "10,000 Plant Genomes Project" plans to sequence and characterize representative genomes from every major clade of embryophytes, green algae, and protists (excluding fungi) within the next 5 years (Cheng et al., 2018; https://db.cngb.org/10kp/).

As sequencing technology advances, the sequencing of the

genomes of all species, subspecies, strains, and lines may no longer be a technical problem. However, we still face some challenges. First, we must solve how to store and efficiently present the sequencing data. Currently, handling the datasets from hundreds of samples has been a huge task, particularly for species with large genomes, such as wheat. To deal with 10K or more genomes from different species will be a challenge. While development of computing technology may partially solve this problem, the formatting of sequence data can also be improved to reduce its storage size. Secondly, we must consider which type of genome presentation to use. So far, graph-based genomes have shown better performance (Ameur, 2019). However, it is possible that we further improve the current genome format, or even innovate a more readable and compressible type of genome representation. Thirdly, we must develop tools that meet the requirements of new genomic analyses. Although several tools related to graph-based genomes have been developed (Garrison et al., 2018; Kim et al., 2019; Eggertsson et al., 2019), they are not yet as mature as those for linear reference genome analyses. The strain of using multiple datasets for each species and of sequencing additional species also must be met with more efficient tools. Machine learning may be another option for solving these problems (Danilevicz et al., 2020).

Human curiosity has no boundary, and we must develop the tools to best store and explore our ever-growing body of knowledge. In the post-genome era, the pan-genome has proven to be a valuable and necessary resource for capturing and sharing the discoveries that are missed from traditional reference genomes. This next step for genomic analyses comes with new challenges. We must develop novel technologies that can work together to address these problems. Eventually, we may be able to answer the question: what are the core and dispensable genomes of life?

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the Ministry of Agriculture of China (2016ZX08009-003), the Chinese Academy of Sciences (ZDRWZS-2019-2) and the National Natural Science Foundation of China (31788103). We thank Ms. Anita K. Snyder for revising the manuscript.*

References

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182, 145–161.e23.

Ameur, A. (2019). Goodbye reference, hello genome graphs. Nat Biotechnol 37, 866–868.

Ballouz, S., Dobin, A., and Gillis, J.A. (2019). Is it time to change the reference genome? Genome Biol 20, 159.

Cheng, S., Melkonian, M., Smith, S.A., Brockington, S., Archibald, J.M., Delaux, P.M., Li, F.W., Melkonian, B., Mavrodiev, E.V., Sun, W., et al. (2018). 10KP: A phylodiverse genome sequencing plan. GigaScience 7, 1–9.

Danilevicz, M.F., Tay Fernandez, C.G., Marsh, J.I., Bayer, P.E., and Edwards, D. (2020). Plant pangenomics: approaches, applications and advancements. Curr Opin Plant Biol 54, 18–25.

Eggertsson, H.P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M.T., Gudbjartsson, D.F., Stefansson, K., Halldorsson, B.V., and Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nat Commun 10, 5402.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477, 419–423.

Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol 36, 875–879.

Golicz, A.A., Batley, J., and Edwards, D. (2016a). Towards plant pangenomics. Plant Biotechnol J 14, 1099–1105.

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H.R., Martinez, P. A., Chan, C.K.K., Severn-Ellis, A., McCombie, W.R., Parkin, I.A.P., et al. (2016b). The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun 7, 13390.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). *De novo* assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44, 226–232.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907–915.

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., et al. (2018). Earth BioGenome Project: sequencing life for the future of life. Proc Natl Acad Sci USA 115, 4325–4333.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M., et al. (2020). Pan-genome of wild and cultivated soybeans. Cell 182, 162–176.e13.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74, 5463–5467.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. Nature 550, 345–353.

Sherman, R.M., and Salzberg, S.L. (2020). Pan-genomics in the human genome era. Nat Rev Genet 21, 243–254.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., 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.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G. G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science 291, 1304–1351.

Yang, X., Lee, W.P., Ye, K., and Lee, C. (2019). One reference genome is not enough. Genome Biol 20, 104.