Credit: Art of Food / Alamy Stock Photo

> a single reference genome for a species does not adequately represent the genetic diversity across the population

MILESTONE 15

Pan-genomes: moving beyond the reference

The pan-genome represents the entire set of genes within a species, consisting of a core genome — containing sequences shared between all individuals of the species — and the 'dispensable' genome. The idea of a pan-genome was first conceived for bacterial species in 2005, when the genomes of six strains of *Streptococcus agalactiae* were sequenced, revealing a core genome containing 80% of *S. agalactiae* genes. Since then, there have been efforts to elucidate the pan-genome of many species beyond bacteria. Assembling and studying pan-genomes has shown that relying on a single reference genome for a species can have an adverse effect on our understanding of the genomic basis of diverse traits. For example, many agronomically important genes in plant species are most often found in the dispensable genome.

Putting together a pan-genome for a genome more complex than those of bacterial species was facilitated by improvements in genome sequencing technologies, particularly long-read sequencing. Larger genomes contain higher proportions of repetitive sequences (up to 50% of the human genome and up to 90% of plant genomes consist of repetitive DNA), which are more difficult to analyse using short reads.

The first plant pan-genome was published in 2014, in a study by Li et al. The authors sequenced seven accessions of *Glycine soja*, a wild relative of cultivated soybean (*Glycine max*). Cultivated soybean has lost much of its genetic diversity through domestication, and so *G. soja* represents a potential source of new genes for soybean improvement. The seven accessions used represent 87% of the genetic diversity found in *G. soja*. Performing de novo assembly of the genomes rather than resequencing, the authors found that approximately 80% of the pan-genome is present in all seven accessions, representing the core genome of this species. However, the dispensable genome of *G. soja* contains more than 51% of gene families. Ultimately, this study concluded that having a single reference genome for a species does not adequately represent the genetic diversity across the population.

Subsequent plant pan-genomes have equally shown the importance of looking at the entire gene repertoire in a species. A study of 54 lines of the grass *Brachypodium distachyon* yielded a pan-genome containing twice the number of genes found in any single individual. Many of the genes found in the dispensable genome are involved in functions such as biotic stress response and development. Indeed, disease resistance genes are among the 4,873 genes in the tomato dispensable genome. As climate change and decreases in arable land worsen, these pools of genetic diversity in the dispensable genome represent a promising avenue for introducing beneficial genes into important crop species.

The inadequacy of single reference genomes is not reserved to plants. A study by Sherman et al., published in 2018, sequenced a dataset of 910 human individuals of African descent, working towards assembling a human pan-genome. The authors estimated that up to 10% of the sequences in the total genome are missing from the reference, many of which fall within protein-coding genes. Having a human pan-genome — or the pan-genome of a subset of the human population — allows the discovery of variants that are missing from the reference genome but may be associated with specific phenotypes. Attempts to create a human pan-genome are relatively rare compared with other species, although efforts are underway to capture global diversity.

Major challenges remain. The studies by Li et al. and Sherman et al. were conducted using short-read sequencing. This requires increased sequencing coverage to ensure sufficient coverage to identify variants with confidence. The complexity of human and plant genomes makes assembly of deep sequencing reads time-consuming and computationally expensive. For example, no computational tool exists that is powerful enough to assemble a pan-genome representing all human sequence variation. Advances in sequencing technology and computational tools to assemble genomes should facilitate the construction and study of pan-genomes from humans, plants and many other species.

Dominique Morneau,
Nature Reviews Methods Primers

ORIGINAL ARTICLES Li, Y.-h. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014) | Gordon, S. P. et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017) | Sherman, R. M. et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019)

FURTHER READING Tettelin, H. et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005) | Li, R. et al. Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010) | Golicz, A. A. et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* **7**, 13390 (2016) | Montenegro, J. D. et al. The pangenome of hexaploid bread wheat. *Plant J.* **90**, 1007–1013 (2017) | Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018) | Gao, L. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019) | Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell.* **182**, 145–161 (2020) | Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176 (2020)