Bioinformatics, 2023, **39(10)**, btad621
https://doi.org/10.1093/bioinformatics/btad621
Advance Access Publication Date: 16 October 2023
Applications Note

Genome analysis

PanKmer: *k*-mer-based and reference-free pangenome analysis

Anthony J. Aylward [1], Semar Petrus[1], Allen Mamerto [1], Nolan T. Hartwick [1], Todd P. Michael [1,*]

[1]The Plant Molecular and Cellular Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, United States

*Corresponding author. The Plant Molecular and Cellular Biology Laboratory, The Salk Institute for Biological Studies. 10010 N Torrey Pines Rd, La Jolla, CA 92037, United States. E-mail: tmichael@salk.edu (T.P.M.)

Associate Editor: Can Alkan

Abstract

Summary: Pangenomes are replacing single reference genomes as the definitive representation of DNA sequence within a species or clade. Pangenome analysis predominantly leverages graph-based methods that require computationally intensive multiple genome alignments, do not scale to highly complex eukaryotic genomes, limit their scope to identifying structural variants (SVs), or incur bias by relying on a reference genome. Here, we present PanKmer, a toolkit designed for reference-free analysis of pangenome datasets consisting of dozens to thousands of individual genomes. PanKmer decomposes a set of input genomes into a table of observed *k*-mers and their presence–absence values in each genome. These are stored in an efficient *k*-mer index data format that encodes SNPs, INDELs, and SVs. It also includes functions for downstream analysis of the *k*-mer index, such as calculating sequence similarity statistics between individuals at whole-genome or local scales. For example, *k*-mers can be "anchored" in any individual genome to quantify sequence variability or conservation at a specific locus. This facilitates workflows with various biological applications, e.g. identifying cases of hybridization between plant species. PanKmer provides researchers with a valuable and convenient means to explore the full scope of genetic variation in a population, without reference bias.

Availability and implementation: PanKmer is implemented as a Python package with components written in Rust, released under a BSD license. The source code is available from the Python Package Index (PyPI) at https://pypi.org/project/pankmer/ as well as Gitlab at https://gitlab.com/salk-tm/pankmer. Full documentation is available at https://salk-tm.gitlab.io/pankmer/.

1 Introduction

Pangenomes consolidate genomic sequence data from multiple individual organisms into a single data structure representing the genomes of a population, species, or clade. Pangenomics was first applied to microbes, but has grown to include a wide variety of eukaryotes (Medini *et al.* 2005, Golicz *et al.* 2020, Li *et al.* 2022b). The advent of plant and animal pangenomes coincided with the reduced cost of NGS technologies. Crop plants were an early subject of pangenome studies, including soybean, brassica, and wheat species (Li *et al.* 2014, Golicz *et al.* 2016, Montenegro *et al.* 2017). These pangenomes focused on genic sequence only, and they were built by collecting short-read whole-genome sequencing datasets and mapping to a reference, allowing relatively simple if biased assembly of multiple individual genomes. Improvements in long-read sequencing technology and assembly methods have enabled the construction of dozens or hundreds of high-quality genomes for a single plant or animal species. This has facilitated a wave of pangenome studies based on collections of entire assembled genomes (Gui *et al.* 2022, Li *et al.* 2022a, Montenegro *et al.* 2022, Shang *et al.* 2022, Tang et al. 2022, Tong *et al.* 2022, Yang *et al.* 2022).

The transition from representation and analysis of single genomes to pangenomes presents significant challenges. Most studied populations include extensive structural variation (SV), which means only a fraction of genes or intergenic sequences are present in all individuals. This fraction is referred to as the "core" genome, while the remainder is variously called "dispensable," "variable," or "accessory" (Golicz *et al.* 2020, Lei *et al.* 2021, Aggarwal *et al.* 2022). Furthermore, the core genome cannot be defined by a simple linear coordinate system. A useful pangenomic dataset must identify the core genome and facilitate analysis of the variable regions.

Currently, the dominant methodology is pangenome sequence graphs, which consist of nodes representing segments of genomic sequence connected by edges which allow any individual genome to be traced as a path through the graph (Hickey *et al.* 2020, Li *et al.* 2020, Baaijens *et al.* 2022, Bradbury *et al.* 2022, Montenegro *et al.* 2022). These graphs have replaced "iterative assembly" methods to advance our understanding of genomic diversity, especially of SV, and demonstrated utility for crop breeding (Ruperao *et al.* 2021, Gui *et al.* 2022, Montenegro *et al.* 2022, Shang *et al.* 2022). However, their construction is far from a solved problem and generally relies either on computationally expensive multiple genome alignment or on biased alignment to a single reference. Their application to highly complex eukaryotic genomes, such as those of plants, is limited (Bayer *et al.* 2020, Danilevicz *et al.* 2020, Khan *et al.* 2020). Such graphs may be limited to as few as a dozen input genomes (Zhang *et al.* 2021, Li *et al.* 2022a).

Received: 31 March 2023; **Revised:** 29 August 2023; **Editorial Decision:** 29 September 2023; **Accepted:** 13 October 2023

Other methods gain efficiency by focusing on specific categories of variants (Li *et al.* 2020, Bradbury *et al.* 2022).

Several methods have adopted a format based on a De Bruijn graph representing overlaps of k-mers (genomic substrings of length k) rather than the sequence graph (Sheikhizadeh *et al.* 2016, Almodaresi *et al.* 2018, Holley and Melsted 2020, Jonkheer *et al.* 2022). One such method is PanTools, which defines the pangenome as a comprehensive representation of multiple annotated genomes and provides functions enabling gene-level analysis, sequence alignment, and phylogenomics.

k-mer decomposition is an alternative to graph-based pangenomes. In this framework, the space of genomic sequences across a population is represented by a set of k-mers. Each individual is represented by a subset: all k-mers observed in a single genome. This view does not depend on any coordinate system and therefore sidesteps the difficulties of multiple genome alignment and graph construction. Kmer-db demonstrated the use of k-mer decomposition for efficient analysis of microbial genomes (Deorowicz *et al.* 2019). k-mers have many applications in genomics and pangenomics, and they have recently been used as markers for GWAS (Holley *et al.* 2016, Sheikhizadeh *et al.* 2016, Aun *et al.* 2018, Khan *et al.* 2020, Voichek and Weigel 2020, Gupta 2021, Jayakodi *et al.* 2021, Jonkheer *et al.* 2022, Karikari *et al.* 2022, 2023).

Here we present PanKmer, a non-graphical k-mer decomposition method designed to efficiently represent and analyze many forms of variation in large pangenomic datasets, with no reliance on a reference genome and no assumption of annotation.

2 Features and implementation

2.1 k-mer index

The foundational component of PanKmer's pangenome representation is the *k-mer index*. It is constructed from a set G of input genomes by decomposing them into a set K of all unique canonical k-mers and noting for each input genome which k-mers are present and which are absent (Fig. 1A). This is similar to the content of Kmer-db's k-mer database (Deorowicz *et al.* 2019). Each k-mer is considered equivalent to its reverse complement and is recorded in canonical form. The k-mer index X is then a $|K|$ by $|G|$ table of binary values indicating presence/absence of each canonical k-mer in each genome. The index can integrate an arbitrary number of genomes from one or several species, requires no reference genome, and enables a range of downstream analyses.

The index is constructed by scanning all $|G|$ input genomes sequentially, recording newly encountered k-mers, and updating presence/absence values with each new genome scanned. To make efficient use of all available CPU's, this process is parallelized across the theoretical k-mer space. The set κ of all possible canonical k-mers has size $|\kappa| = \frac{1}{2}\kappa^4$, which is divided among n segments $\kappa_1 \ldots \kappa_n$. Index construction is then divided into n subprocesses, each of which constructs a subindex X_i skipping k-mers not included in κ_i. To efficiently store and access the k-mer index on disk, each k-mer is converted to an integer value.

The resultant index is robust to varying contiguity in the input genomes, so chromosome-level assemblies can be directly compared to unscaffolded contigs or unaligned reads. The implementation presented here uses $k = 31$, chosen for three reasons. First, 31-mers are short enough to be encoded as 64-bit integers (Rahman *et al.* 2018). Second, they are long enough to impose a low rate of non-unique k-mers occurring by chance (Sheikhizadeh *et al.* 2016). Finally, 31-mers have been used successfully to define variation in previous studies (Rahman *et al.* 2018, Voichek and Weigel 2020).

2.2 Adjacency matrix

Once the k-mer index is constructed, each input genome g_i is represented by $|K|$ binary values representing presence or absence of each k-mer in K. This provides a natural means of calculating pairwise similarity/adjacency values for the input genomes. PanKmer includes a function to calculate the number of shared k-mers between all pairs of input genomes and return them as an adjacency matrix. Subsequently, the adjacency values can be used to perform a hierarchical clustering of input genomes and plot adjacency values as a heatmap. The adjacency values may also be converted to Jaccard, QV as described in MerQury (Rhie *et al.* 2020), a symmetric version of QV, or average nucleotide identity (ANI) (Fig. 1B).

2.3 Genome anchoring

While the k-mer index does not rely on any specified reference genome, it can be used to contextualize individual sequences. Given a sufficient k-mer length (e.g. the default $k = 31$), we can assume that each k-mer present in an individual genome g_i occurs approximately once in g_i. Therefore, we can quantify variation across any locus in an "anchor" genome g_i by walking along the sequence, checking the k-mer that corresponds to each position, and calculating the fraction of genomes in G which share that k-mer. We refer to this fraction as the "k-mer conservation" value at each position. High k-mer conservation values indicate core loci which are conserved in many individuals across the pangenome, while low values indicate variable loci which are present only in g_i and a small number of other genomes. Hence, core sequences will have high k-mer conservation levels in all target genomes, while variable sequences will have relatively lower levels in each genome that features them.

3 Results

To demonstrate the utility of PanKmer, we constructed a cross-species pangenome of cattail downloaded from NCBI SRA: 10 *Typha domingensis* (TD) and 3 *Typha latifolia* (TL) genomes (Fig. 1B) (Supplementary Table S1). Three of the TD genomes (TD01, TD22, TD23) were highly heterozygous (Supplementary Fig. S1 and Table 1). We used k-mer profiles to compute two measures of adjacency, the number of shared k-mers and ANI (Fig. 1B). The three TL genomes were highly similar to one another, with an average ANI of 99.85% (186M shared k-mers) for TL–TL comparisons, while the average ANI of TD–TD comparisons was 98.94% (151M shared k-mers) (Supplementary Tables S2 and S3). The average ANI of TL–TD comparisons was 96.14% (64M shared k-mers). We also constructed single-species pangenomes of *T.domingensis* and *T.latifolia* (Supplementary Fig. S2 and Tables S4–S7). One TD genome, TD01, was an outlier relative to other TD, with ANI of only 98.03% (127M shared k-mers) on average. Conversely, TD01 showed relatively high ANI with TL genomes, averaging 98.60% (148M shared k-mers). To inspect TD01 more closely, we calculated average k-mer conservation in 100 kb bins across its assembled contigs and observed a mixture of TD and TL sequences. A close

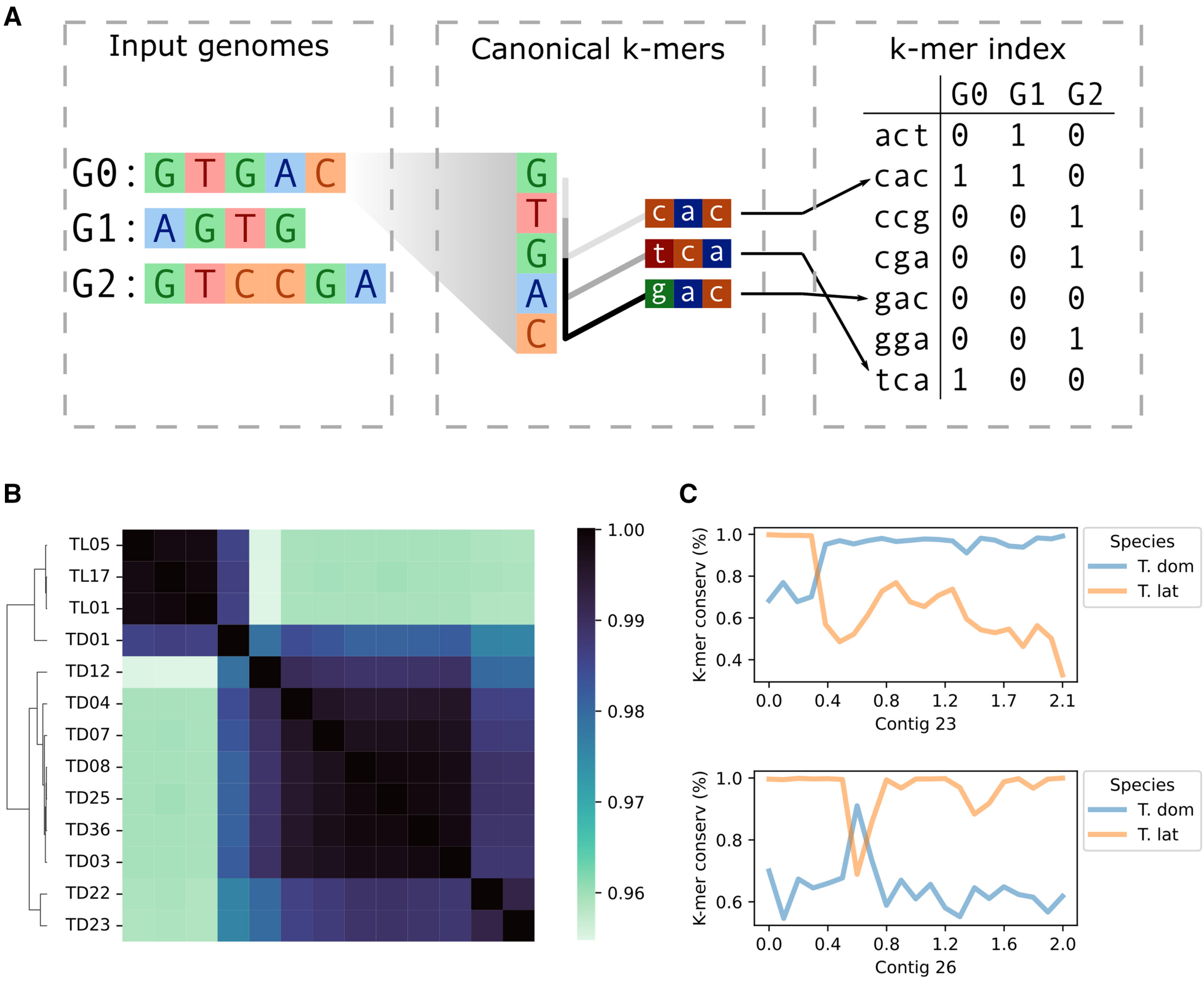

Figure 1. PanKmer enables the rapid estimation of relatedness across the pangenome as well as analysis of specific loci. (A) Schematic of procedure for constructing the *k*-mer index. In the example, each genome G0–G2 is decomposed into canonical 3-mers. Each 3-mer is equivalent to its reverse complement, and the lexicographically first is the canonical form. Each 3-mer is assigned an integer value, and its presence/absence is recorded for each genome in the index. (B) Relatedness heatmap of *Typha* pangenome, ANI values shown. (C) Genome anchoring plots of representative contigs in TD01. Average *k*-mer conservation of 100-kb bins shown, where *k*-mer conservation is the fraction of TD or TL genomes that include each *k*-mer along the contig.

Table 1. Basic statistics of *T.domingensis* and *T.latifolia* genome assemblies.

Name	Predicted genome size (Mb)	Heterozygosity (%)	Contig N50 (Mb)	GC content (%)
TD01	214	4.12	1.0	37.7
TD03	222	0.13	11.5	37.6
TD04	226	0.45	4.8	37.7
TD07	263	0.14	7.7	37.6
TD08	213	0.05	13.8	37.7
TD12	224	0.22	6.1	37.5
TD22	218	2.60	1.0	37.8
TD23	213	2.60	0.7	37.8
TD25	210	0.12	8.3	37.7
TD36	215	0.15	8.9	37.6
TL01	234	0.07	14.5	37.9
TL05	218	0.07	14.5	37.8
TL17	215	0.07	13.5	37.7

examination of variability anchored in TD01 revealed the presence of large and small introgressions (Fig. 1C and Supplementary Fig. S5). High heterozygosity together with the presence of introgressions suggest TD01 is the descendant of a recent TD–TL hybridization event.

To explore suitability for larger eukaryote genomes and sample sizes, we benchmarked PanKmer on several published pangenomes and super-pangenomes, including *Solanum*, *Zea*, *Homo sapiens*, *Arabidopsis thaliana* (Table 2 and Supplementary Fig. S7 and Table S12) (Alonso-Blanco *et al.* 2016, Woodhouse *et al.* 2021, Montenegro *et al.* 2022, Liao *et al.* 2023). PanKmer successfully built *k*-mer indexes for all input pangenomes (Supplementary Fig. S8). We compared the performance of PanKmer to Kmer-db (Supplementary Fig. S9 and Table S13). We found that PanKmer had a smaller memory footprint than Kmer-db, due in part to its ability to divide *k*-mer decomposition into multiple rounds (Supplementary Material).

4 Conclusion

PanKmer enables *k*-mer-based analysis of pangenome datasets. It accepts as input a collection of genome assemblies or unaligned reads in FASTA format, and produces a *k*-mer

Table 2. Benchmarking results of PanKmer configured for moderate memory usage and moderate runtime (wall clock time) (16 rounds of *k*-mer decomposition).

Clade	No. of genomes	Peak memory (GB)	Time (min)
Typha	4	6	5.56
Typha	13	8	18.61
Solanum	4	14	16.26
Solanum	16	20	62.87
Solanum	46	20	215.19
Zea	4	20	50.12
Zea	16	71	239.38
Zea	54	80	783.1
H.sapiens	4	34	66.45
H.sapiens	16	35	248.5
H.sapiens	64	38	707.39
H.sapiens	94	40	1364.33
A.thaliana	4	6	2.08
A.thaliana	16	6	7.98
A.thaliana	64	7	35.9
A.thaliana	256	7	142.93
A.thaliana	1135	15	648.9

index which is similar to the database of kmer-db (Deorowicz *et al.* 2019). The index is agnostic to the contiguity of assemblies. A powerful feature of PanKmer is the ability to anchor the index in any individual genome and identify core or variable loci. This allows users to explore the full scope of genetic variation, without incurring bias from the choice of a single reference. PanKmer includes tools for downstream processing of the index which provide data visualizations and biological insights, such as identifying a hybridization event between *T.latifolia* and *T.domingensis*.

The primary advantage of PanKmer over other pangenome analysis tools is its ability to capture all forms of presence–absence variation, including SNPs, INDELs, SVs, and any variant that adds or removes a *k*-mer from the genome. Our reference-free and alignment-free algorithm is also more computationally tractable than graph-based methods. On the other hand, PanKmer is limited by inability to detect copy number variants in repetitive sequences (Supplementary Fig. S6), and by the loss of spatial relationships between *k*-mers in the index. However, their spatial context can be rescued by projecting the index onto an anchor genome. Currently, PanKmer does not have genotyping functions, but these are planned for future releases.

Acknowledgements

We thank members of the *Typha* genomes team specifically Suzanne Thomas, James J. La Clair, and Justin Pacheco from the Noel lab, Bradley W. Abramson, Kelly Colt and Mariele Lensink from the Michael lab, Theryn K. Henkel, Eva Hillman, and Alexander S. Kolker. The plant collection, propagation and genome work were funded by the following awards to Joseph P. Noel, the Howard Hughes Medical Institute, the Salk Institute Arthur and Julie Woodrow Chair, the Chambers Medical Foundation, the Kitzmiller-Moran Challenge Funds, and the Joe W. and Dorothy Dorsett Brown Foundation. We also thank Edward S. Buckler, Bethany F. Econopouly, Zachary R. Miller, Peter J. Bradbury, and Matt Wiese of the United States Department of Agriculture-Agricultural Research Service and Institute for Genomic Diversity at Cornell University for helpful discussion and collaboration on software features and implementation.

Author contributions

A.J.A.: software (main developer, testing), writing (original draft, review and editing) funding acquisition. S.P.: software (original developer), writing (review and editing). A.M.: software (pipeline management, testing), writing (review and editing). N.T.H.: software (optimization). T.P.M.: supervision, funding acquisition, writing (review and editing).

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by Bill & Melinda Gates Foundation (BMGF) (INV-040541 to T.P.M.); and the Tang Genomics Fund.

Data availability

The raw reads have been deposited in the NCBI short read archive (SRA) under BioProject PRJNA742003. Accession numbers for individual experiments are provided in the Supplementary Material. The genome assemblies are available for download from Michael lab AWS storage at: https://salk-tm-pub.s3.us-west-2.amazonaws.com/pub-supplementary/PRJNA742003-ASSEMBL.tar. Complete tutorials and documentation are available in the Supplementary Material and online: https://salk-tm.gitlab.io/pankmer/.

References

Aggarwal SK, Singh A, Choudhary M *et al.* Pangenomics in microbial and crop research: progress, applications, and perspectives. *Genes (Basel)* 2022;**13**:598.

Almodaresi F, Sarkar H, Srivastava A *et al.* A space and time-efficient index for the compacted colored De Bruijn graph. *Bioinformatics* 2018;**34**:i169–77.

Alonso-Blanco C, Andrade J, Becker C *et al.* 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 2016; **166**:481–91.

Aun E, Brauer A, Kisand V *et al.* A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput. Biol* 2018;**14**: e1006434.

Baaijens JA, Bonizzoni P, Boucher C *et al.* Computational graph pangenomics: a tutorial on data structures and their applications. *Nat Comput* 2022;**21**:81–108.

Bayer PE, Golicz AA, Scheben A *et al.* Plant pan-genomes are the new reference. *Nat Plants* 2020;**6**:914–20.

Bradbury PJ, Casstevens T, Jensen SE *et al.* The practical haplotype graph, a platform for storing and using pangenomes for imputation. *Bioinformatics* 2022;**38**:3698–702.

Danilevicz MF, Tay Fernandez CG, Marsh JI *et al.* Plant pangenomics: approaches, applications and advancements. *Curr Opin Plant Biol* 2020;**54**:18–25.

Deorowicz S, Gudys A, Dlugosz M *et al.* Kmer-db: instant evolutionary distance estimation. *Bioinformatics* 2019;**35**:133–6.

Golicz AA, Bayer PE, Barker GC *et al.* The pangenome of an agronomically important crop plant brassica oleracea. *Nat Commun* 2016;**7**: 13390.

Golicz AA, Bayer PE, Bhalla PL *et al.* Pangenomics comes of age: from bacteria to plant and animal applications. *Trends Genet* 2020;**36**: 132–45.

Gui S, Wei W, Jiang C *et al.* A pan-zea genome map for enhancing maize improvement. *Genome Biol* 2022;**23**:178.

Gupta PK. GWAS for genetics of complex quantitative traits: genome to pangenome and SNPs to SVs and k-mers. *Bioessays* 2021;**43**: e2100109.

Hickey G, Heller D, Monlong J *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol* 2020;**21**: 35.

Holley G, Melsted P. Bifrost: highly parallel construction and indexing of colored and compacted De Bruijn graphs. *Genome Biol* 2020;**21**: 249.

Holley G, Wittler R, Stoye J. Bloom filter trie: an alignment-free and reference-free data structure for pan-genome storage. *Algorithms Mol. Biol* 2016;**11**:3.

Jayakodi M, Schreiber M, Stein N *et al.* Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res* 2021;**28**:dsaa030.

Jonkheer EM, van Workum DM, Anari SS *et al.* Pantools v3: functional annotation, classification and phylogenomics. *Bioinformatics* 2022; **38**:4403–5.

Karikari B, Lemay M, Belzile F. k-mer-based genome-wide association studies in plants: advances, challenges, and perspectives. *Nat. Genet* 2022;**54**:518–25.

Karikari B, Lemay M, Belzile F. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Genes* 2023;**14**:1439.

Khan AW, Garg V, Roorkiwal M *et al.* Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends Plant Sci* 2020;**25**:148–58.

Lei L, Goltsman E, Goodstein D *et al.* Plant pan-genomics comes of age. *Annu Rev Plant Biol* 2021;**72**:411–35.

Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* 2020;**21**:265.

Li H, Wang S, Chai S *et al.* Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat Commun* 2022a;**13**:682.

Li W, Liu J, Zhang H *et al.* Plant pan-genomics: recent advances, new challenges, and roads ahead. *J Genet Genomics* 2022b;**49**:833–46.

Li Y, Zhou G, Ma J *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* 2014;**32**:1045–52.

Liao W-W, Asri M, Ebler J *et al.* A draft human pangenome reference. *Nature* 2023;**617**:312–24.

Medini D, Donati C, Tettelin H *et al.* The microbial pan-genome. *Curr Opin Genet Dev* 2005;**15**:589–94.

Montenegro JD, Golicz AA, Bayer PE *et al.* The pangenome of hexaploid bread wheat. *Plant J* 2017;**90**:1007–13.

Montenegro JD, Zhou Y, Zhang Z *et al.* Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* 2022; **606**:527–34.

Rahman A, Hallgrímsdóttir I, Eisen M *et al.* Association mapping from sequencing reads using k-mers. *Elife* 2018;**7**:e32920.

Rhie A, Walenz BP, Koren S *et al.* Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* 2020;**21**:245.

Ruperao P, Thirunavukkarasu N, Gandham P *et al.* Sorghum pangenome explores the functional utility for genomic-assisted breeding to accelerate the genetic gain. *Front Plant Sci* 2021;**12**.

Shang L, Li X, He H *et al.* A super pan-genomic landscape of rice. *Cell Res* 2022;**32**:878–96.

Sheikhizadeh S, Schranz ME, Akdel M *et al.* Pantools: representation, storage and exploration of pan-genomic data. *Bioinformatics* 2016; **32**:i487–93.

Tang D, Jia Y, Zhang J *et al.* Genome evolution and diversity of wild and cultivated potatoes. *Nature* 2022;**606**:535–41.

Tong X, Han M, Lu K *et al.* High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. *Nat Commun* 2022;**13**:5619.

Voichek Y, Weigel D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat Genet* 2020;**52**: 534–40.

Woodhouse MR, Cannon EK, Portwood JLI *et al.* A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biol* 2021;**21**:385.

Yang T, Liu R, Luo Y *et al.* Improved pea reference genome and pangenome highlight genomic features and evolutionary characteristics. *Nat Genet* 2022;**54**:1553–63.

Zhang X, Liu T, Wang J *et al.* Pan-genome of *Raphanus* highlights genetic variation and introgression among domesticated, wild, and weedy radishes. *Mol Plant* 2021;**14**:2032–55.