Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2024), pp. 1–13 doi: 10.1111/pbi.14545

Exploring intra- and intergenomic variation in haplotype-resolved pangenomes

Eef M. Jonkheer[1,2], Dick de Ridder[1], Theo A. J. van der Lee[2], Jorn R. de Haan[3], Lidija Berke[3] and Sandra Smit[1,]*

[1] *Bioinformatics Group, Wageningen University & Research, Wageningen, The Netherlands*
[2] *Biointeractions and Plant Health, Wageningen Plant Research, Wageningen, The Netherlands*
[3] *Genetwister Technologies B.V, Wageningen, The Netherlands*

Received 22 June 2024;
revised 21 November 2024;
accepted 23 November 2024.
**Correspondence* (Tel +31 317482036;
email sandra.smit@wur.nl)

Keywords: pangenomics, haplotype-resolved genomes, genome evolution, genome organization, genetic variation, visualization.

Summary

With advances in long-read sequencing and assembly techniques, haplotype-resolved (phased) genome assemblies are becoming more common, also in the field of plant genomics. Computational tools to effectively explore these phased genomes, particularly for polyploid genomes, are currently limited. Here we describe a new strategy adopting a pangenome approach. To analyse both intra- and intergenomic variation in phased genome assemblies, we have made the software package PanTools ploidy-aware by updating the pangenome graph representation and adding several novel functionalities to assess synteny and gene retention, profile repeats and calculate synonymous and nonsynonymous mutation rates. Using PanTools, we constructed and analysed a pangenome comprising of one diploid and four tetraploid potato cultivars, and a pangenome of five diploid apple species. Both pangenomes show high intra- and intergenomic allelic diversity in terms of gene absence/presence, SNPs, indels and larger structural variants. Our findings show that the new functionalities and visualizations are useful to discover introgressions and detect likely misassemblies in phased genomes. PanTools is available at https://git.wur.nl/bioinformatics/pantools.

Introduction

Until recently, obtaining chromosome-scale genome assemblies, let alone haplotype-phased genomes, required tremendous effort. As a result, genome assemblies of diploid or polyploid organisms used in genomic analyses are typically haploid representations, where the multiple copies of a chromosome are collapsed into a single sequence with a mosaic of alleles. Such haploid representations ignore the intragenomic variation in gene content and organization. However, differences between haplotypes may provide novel insights, not only into the evolutionary history but also in explaining certain phenotypes (Hasing *et al*., 2020). Including intragenomic variation in reference genomes would greatly facilitate their use for genetics and breeding but requires a comprehensive methodology to define haplotypes (Brinton *et al*., 2020; Leitwein *et al*., 2020). Due to various advances in sequencing technology (Jain *et al*., 2018; Wenger *et al*., 2019) and assembly algorithms (Cheng *et al*., 2021; Koren *et al*., 2018), the different haplotypes of genomes with two or more chromosome sets can now be accurately resolved. Exploring such fully phased genomes allows for a more comprehensive assessment of the genetic variation, and enables new types of analyses such as gene-dosage analysis or detection of allele-specific expression (when combined with other-omics data).

In recent years, there has been an increasing number of available (partially) phased genome assemblies for fungi (Duan *et al*., 2022; Hamlin *et al*., 2019), plants (Hoopes *et al*., 2022; Lin *et al*., 2023; Shirasawa *et al*., 2019; Sun *et al*., 2020) and animals (Garg *et al*., 2021; Han *et al*., 2022; Porubsky *et al*., 2021). In these recent publications, the haplotype-resolved genomes are generally analysed through custom pipelines that consist of various tools not specifically designed to handle (partially) phased genome assemblies. To enable efficient comparison of multiple-phased genomes, we propose pangenome representations and related tools as a possible solution. These approaches already serve to identify small and large-scale variations in large genome collections.

There has been a rise in the availability of sequence-level pangenome representations, each offering unique strengths suited to various applications. The recently published toolkits Pangenome Graph Builder (pggb) (Garrison *et al*., 2023) and Minigraph-Cactus (MC) (Hickey *et al*., 2024), stand out for their efficiency in representing genomic variation across many genomes/haplotypes. These graphs are particularly effective as pangenome references, and excel in read mapping analyses, as well as variant and structural variant (SV) calling. While these tools can handle haplotype-resolved assemblies, they offer limited support for annotations (Novak *et al*., 2024), and are not designed for gene-level analyses.

Conversely, the toolkit PanTools (Sheikhizadeh *et al*., 2016), developed by us, natively combines annotations and genome sequences in the pangenome representation, distinguishing itself from sequence-based or gene-based approaches. The hierarchical graph structure, which includes a compacted De Bruijn graph (DBG) to represent sequences, is stored in a Neo4j database. Structural annotation nodes are linked to their respective start and stop positions in the DBG. Homology relationships function

Please cite this article as: Jonkheer, E.M., de Ridder, D., van der Lee, T.A.J., de Haan, J.R., Berke, L. and Smit, S. (2024) Exploring intra- and intergenomic variation in haplotype-resolved pangenomes. *Plant Biotechnol. J.*, https://doi.org/10.1111/pbi.14545.

as an additional layer to connect annotation nodes. The heterogeneous graph can be queried through Neo4j's query language Cypher. Annotating features in the graph makes PanTools an effective framework to analyse genome content, organization and evolution (Jonkheer *et al*., 2022).

To enable the analyses of haplotype-resolved genomes, we now made PanTools 'ploidy aware' by including new functionality to perform comparative analyses within and between genomes. In this study, we demonstrate the new PanTools functionality on two datasets: tetraploid potatoes (*Solanum tuberosum*) and diploid apples (*Malus* spp.). These two crops were selected due to the availability of multiple high-quality haplotype-resolved assemblies (Bao *et al*., 2022; Chen *et al*., 2019; Daccord *et al*., 2017; Hoopes *et al*., 2022; Pham *et al*., 2020; Sun *et al*., 2020, 2022), along with a hypothesized high intragenomic variation among haplotypes. This variation is thought to arise from the occurrence of multiple ancient whole-genome duplication (WGD) events within these lineages. Potato and apple share three WGD events and each has its own lineage-specific event (Jung *et al*., 2012; Soltis *et al*., 2009; Tomato Genome Consortium, 2012). Moreover, the more recent domestication of these two crops involved extensive selective breeding and hybridization (Cornille *et al*., 2012; Gaiero *et al*., 2018).

We show the usefulness of our pangenome representation and functionalities, emphasizing the previously hidden intragenomic variation within haploid reference genomes (see Supplement S3 for reproducibility). We describe a universal approach for constructing both pangenome graphs and characterizing gene content both inter- and intragenomically, and explore different approaches to establish phylogenetic relationships between sequences. With our newly developed visualization methodology, we create visualizations that provide insights into the variation in genomic organization. Finally, we utilized our pangenome approach to identify the new allelic variation of *StCDF1*, a key regulator of maturation and tuberization in potato.

Results and discussion

Ploidy-aware pangenome analyses

PanTools has a hierarchical pangenome representation, linking divergent genomes not only through a sequence variation graph but also through structural and functional annotations and homology. To enable the analysis of haplotype-resolved assemblies, we introduced new annotations to label haplotypes, updated existing functionalities and introduced a new set of command-line tools.

PanTools allows for the incorporation of the haplotype information to control which sequences or features are compared. Pangenomes are constructed from collections of genome FASTA files; a genome layer is formed with *genome* nodes that connect to *sequence* nodes (representing contigs/scaffolds/pseudomolecules), which in turn link to the start and stop *nucleotide* nodes in a compacted De Bruijn graph (DBG) representation of the DNA sequences. The database scheme underlying the graph is shown in Figure S1 in Supplement S1. A *sequence* node can be annotated with a chromosome number (1, 2, ...) and haplotype phase (A, B, ...). By combining the two, each sequence node has a unique haplotype identifier within a genome (e.g. 1A, 1B, 2A, etc.). Sequences in a genome with the same haplotype (letter) are considered to be a subgenome. From a biological standpoint, this concept of a subgenome does not exist, as there is no genetic or physical linkage between assembled chromosomes. However, defining the subgenomes in this way enables the assessment of gene presence among multiple haplotypes within a specific chromosome (e.g. 1A, 1B, 1C, 1D). Nevertheless, it is worth to note that it is not meaningful to compare, for instance, the 'C' subgenome between two genomes, given the random composition of chromosomes within the subgenome. Finally, sequences that lack phasing information are all labelled as chromosome 0 (and do not receive a unique haplotype identifier). A schematic overview of the terminology is included as Figure S2 in Supplement S1.

Every pangenome analysis starts with collecting genome and sequence nodes, to determine which sequences and features will be compared. Phasing information enables the collection of specific sequences to perform more targeted analyses – for example, comparison among genomes, specific subgenomes or homeologous chromosomes. In genome assemblies where the chromosomal organization is still unknown, phylogenetic relatedness is determined from the number of shared *k*-mers in the DBG. This distance method was updated to count *k*-mers per sequence instead of per genome, allowing the identification of homoeologous chromosomes.

We updated several PanTools methods to work at both sequence level and genome level. Single-copy genes function as ideal markers for phylogeny inference (Li *et al*., 2017). In phased genomes, single-copy genes may have a copy per subgenome, drastically reducing the number of genes that can be detected when simply looking for a single copy using the current methods at the genome level. To address this issue, we allow for one gene copy per subgenome. In this way, applications that rely on single-copy genes, such as BUSCO and the core phylogeny, can work at a subgenome level rather than the bulked genome.

PanTools' gene classification method identifies shared genes between genomes and can describe a pangenome's gene content as core (present in all genomes), accessory/dispensable (present in some but not all) and cloud (present in one genome) (Figure S3 in Supplement S1). We updated the term 'unique' to 'cloud' because the former suggests a gene is found only once, while it may actually have multiple (allelic) copies. With haplotype information incorporated in the graph, gene presence/absence can now be established for every subgenome of a genome. Accordingly, we further characterize gene content by the presence of number of subgenomes (1, 2, ...).

To further explore the information in fully phased genomes, new functionalities have been developed in PanTools, integrating bioinformatics methods into the pangenome representation. First, synteny (collinear gene blocks) between genomes, as detected by MCScanX (Wang *et al*., 2012), can be added to the pangenome (Figure S1 in Supplement S1). From the synteny block information, we calculate gene retention and visualize fractionation patterns across chromosomes. We speed up the minimap2 (Li, 2018) whole-genome alignments analysis, by making use of haplotype and chromosome information to compare only homeologous chromosomes, thereby avoiding computationally intensive all-vs-all comparisons. With PAL2NAL (Suyama *et al*., 2006), we calculate synonymous and nonsynonymous mutation rates on aligned sequences in homology groups or syntenic gene pairs, allowing to study evolutionary rates in a species or population. Finally, we developed a novel graphical representation of genomic structure in which we combine the newly integrated genomic features such as synteny, repeat density and subgenome presence of genes.

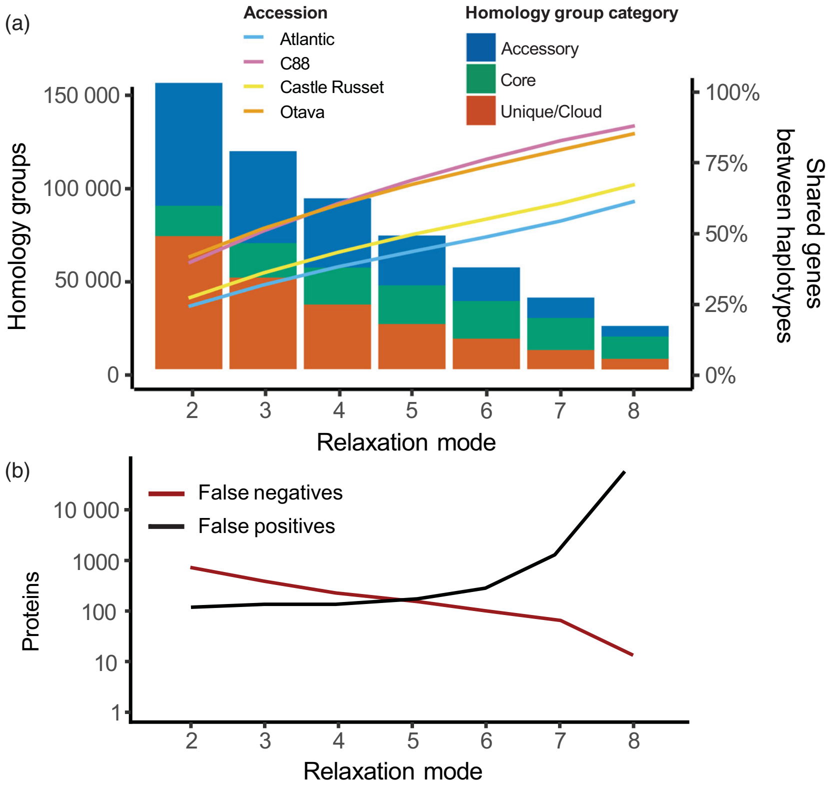


Figure 1 (a) The effect of increasing relaxation modes (lowering clustering stringency) on the *S. tuberosum* pangenome composition, in terms of total number of homology groups (bar charts) and average percentage of genes shared between subgenomes (line graphs). (b) PanTools' BUSCO benchmark results of the seven homology grouping settings (relaxation modes).

Pangenome design choices and construction

To showcase PanTools' updated and novel functionalities, we built two pangenomes exhibiting different ploidy levels. A species-level pangenome was constructed of five *S. tuberosum* (potato) cultivars: DM1-3516 R44 (DM) (Pham *et al*., 2020), Atlantic (Hoopes *et al*., 2022), Castle Russet (CR) (Hoopes *et al*., 2022), Otava (Sun *et al*., 2022) and Cooperation-88 (C88) (Bao *et al*., 2022). DM is a doubled haploid and, therefore, represented as a haploid assembly, whereas the other accessions were tetraploid with fully resolved haplotypes, leading to 17 (1 + 4 × 4) subgenomes in total. All assemblies were chromosome-scale and had 12 or 48 pseudomolecules, matching the base chromosome number of 12 in *S. tuberosum* (Pham *et al*., 2020). We found a clear dichotomy in assembly statistics as a result of different sequencing and assembly approaches. The Otava and C88 assemblies, of size 3.1–3.2 Gb, were considerably larger than CR and Atlantic at 2.5–2.7 Gb. This large difference was further reflected in the total numbers of genes annotated: 150 853–152 835 in Otava and C88 and 105 449–114 021 in Atlantic and CR. These notable discrepancies suggest the two latter assemblies have a greater number of collapsed regions or suffer from other assembly/annotation challenges.

Benchmarking universal single-copy orthologs (BUSCO) (Manni *et al*., 2021) has become the standard for assessing genome assembly quality. Unlike technical metrics, such as the number of reads mapping back to the genome or the N50 value, BUSCO is biologically meaningful and based on informed expectations of the single-copy gene content. Completeness of the *S. tuberosum* genomes according to BUSCO ranged from 93.6% to 99.5% (Figure S4 in Supplement S1). All four haplotype-resolved genomes showed high levels of duplication, especially C88 in which nearly all (98.2%) BUSCO genes were duplicated. Because we suspected that most duplicates were a result of haplotype phasing, BUSCO was run separately on each of the subgenomes (Figure S5 in Supplement S1).

This approach successfully removed gene duplicates, but also revealed large differences in subgenome completeness: 59.2%–62.9% in CR, 70.2%–77.3% in Atlantic, 88.9%–90.9% in Otava and 95.6%–96.6% in C88. BUSCO applied on only unsorted/unphased sequences showed minimal completeness (<1%) in Otava and C88, but substantial completeness in Atlantic and CR (>31.0%). This again suggests haplotypes of the latter two genomes were not fully phased. The C88 subgenomes had the highest completeness, resulting in 4870 BUSCO genes detected as single-copy on all four subgenomes (Figures S6–S9 in Supplement S1). In comparison, Atlantic had only 1307 single-copy genes found in all four subgenomes. These contrasts in statistics reflect clear differences in phasing quality between genomes.

Inferring homology is fundamental to gene-based pangenome analyses. *S. tuberosum* proteomes were clustered with different settings (so-called 'relaxation modes'), where increasing relaxation modes indicate lower clustering stringency. The most critical parameter is the minimum required sequence similarity of pairwise alignments, starting at 95% and lowered by 10% in each subsequent mode. Between relaxation modes 2 and 8, a nearly sevenfold decrease in the number of homology groups was observed (Figure 1a). The availability of different relaxation modes allows calibration to different data sets, but obviously raises the question of what the optimal setting is.

Table 1 Genome assembly and pangenome statistics

	S. tuberosum	*Malus*
Genomes	5	5
Subgenomes	17	8
Sequences	34 050	158 662
Total input size	12.2 Gb	5.3 Gb
Protein-coding genes	723 497	361 864
K-mers	1117.7 M	690.0 M
Compressed *k*-mers	221.6 M	103.2 M
Homology groups	52 240	68 751
Singleton groups	6638	15 635

We used BUSCO genes to assess each homology grouping and found a clear trade-off between recall and precision (Figure 1b). The highest F_1 scores (not shown), combining recall and precision, were obtained with mode 5 and 6, corresponding to 55% or 45% sequence similarity thresholds. After mode 6, the number of false positives rapidly increases. In addition to this benchmark, we calculated the average percentage of genes shared between subgenomes as a metric for selecting a suitable grouping (line graphs in Figure 1a). The clear division between the two pairs of genomes is in line with the earlier BUSCO completeness assessment: CR and Atlantic subgenomes share less because of lower phasing quality. As the difference in F_1 scores between modes 5 and 6 was negligible, mode 6 was selected based on the higher overlap in gene content between subgenomes.

Besides potato, a genus-level pangenome was constructed from five diploid *Malus* (apple) accessions: *M. domestica* cv. Gala (Gala), *M. domestica* 'Golden Delicious' GDDH13, *M. sieversii*, *M. sylvestris* and *M. baccata*. Haplotypes were resolved for the Gala, *M. sieversii* and *M. sylvestris* genomes. All assemblies except *M. baccata* were arranged into 17 or 34 whole-chromosome pseudomolecules, correctly representing the 17 chromosomes of the *Malus* genus (Daccord *et al*., 2017).

For constructing the *Malus* pangenome, we followed the same BUSCO clustering approach. BUSCO indicated very high completeness for all genomes, as well as for the separate subgenomes. The analysis supports the existence of a recent Maleae-specific whole-genome duplication (WGD), as nearly one third of the gene content in all *Malus* (sub)genomes was marked as duplicated. A more detailed description of the Malus pangenome construction is provided in Analysis 1 in Supplement S2. Following a universal strategy, we built the *S. tuberosum* and *Malus* pangenomes of datasets that were distinct in terms of genome size, ploidy level, phasing quality and evolutionary divergence between genomes (Table 1).

Gene content and relatedness of (sub)genomes

Pangenomes are studied by classifying genes as shared between (subsets) of genomes, that is, which genes are core, which are accessory and which are cloud. Given phased assemblies, gene content can also be assessed within and between subgenomes: where are these core/accessory/cloud genes located, and to what extent are they found in different subgenomes? Such analyses shed light on the organization and evolution of subgenomes, albeit subject to assembly and annotation quality. With homology relationships established and integrated into the graph, we next characterized the genetic composition of the *S. tuberosum* pangenome. The *Malus* analysis is included in Analysis 2 in Supplement S2.

The protein-coding genes of the *S. tuberosum* genomes clustered in 52 240 homology groups, 37.1% of which were core, 33.0% accessory and 29.9% cloud (Figure 2a). Nearly half of the cloud groups are present in only a single subgenome. Notably, 80% of these subgenome exclusive groups are singleton groups and do not show any homology to another sequence, causing suspicion about their realness. Interestingly, cloud genes were more abundant in the higher quality Otava and C88 genomes. On the other end of the spectrum, core genes require occurrence in a minimum of 5 subgenomes (from different accessions), but are generally found in 11–17 subgenomes. When we analyse the gene content of individual genomes, we see the majority of genes were characterized as core (Figure 2b).

The genomic distribution of C88's genes revealed an interesting pattern (Figure 2c). Most accessory and cloud genes are positioned in the pericentromeric region of the chromosome, whereas core genes generally lie in chromosome arms. The observed localization of cloud genes was most prominent in the Otava and C88 genomes but was also seen in the other genomes (Figures S10–S13 in Supplement S1). Clear differences in patterns between these visualizations suggest that due to the higher quality, a larger number of genes were likely annotated in high repetitive regions. The high frequency of cloud genes in pericentromeric regions may be partly attributed due to the lack of recombination in the heterochromatic centromere (Jiang *et al*., 2023; Marand *et al*., 2017). Another contributing factor might be the accumulation of deleterious mutations in potato genomes, disrupting open reading frames and altering the protein sequences (Zhang *et al*., 2019).

As an alternative to grouping *S. tuberosum* genes by their intergenomic presence in the pangenome, we can now also characterize genes by intragenomic presence in (1 to 4) subgenomes. In Otava and C88 nearly half of the groups have genes present in all four subgenomes, whereas in Atlantic and CR this was around a fifth of the groups (Figure 2d). We again visualized C88's gene regions but now coloured them according to the subgenome characterization (Figure 2e). Core genes in the chromosome arms are strongly correlated to the presence in all four haplotypes. We briefly discuss four chromosomal sections with distinctive patterns:

- All four haplotypes in the middle of Chr 1 show a stretch of gene regions occurring in a single haplotype, indicating a distinct gene set on each haplotype.
- The absence of genes within the same region of four Chr 3 haplotypes is indicative of the centromere, and it indeed overlaps with C88's centromere prediction, which is based on repeat arrays (Zhou *et al*., 2020).
- Chr 10 shows one disparate haplotype, indicated in orange (3 occurrences) in the remaining haplotypes.
- The pericentromeric regions of Chr 11 predominantly show two divergent alleles present on two haplotypes, suggesting two distinct sets of alleles.

This visualization demonstrates the haplotype diversity in the C88 genome. Additionally, it shows clear mosaic patterns that indicate large genomic regions can exist in one to four copies. The uneven distribution is clearly exemplified in the last two patterns: Chr 10 and 11 have differently observed relationships (2:2 versus 3:1). As the C88 potato variety derived from two distinct cultivars,

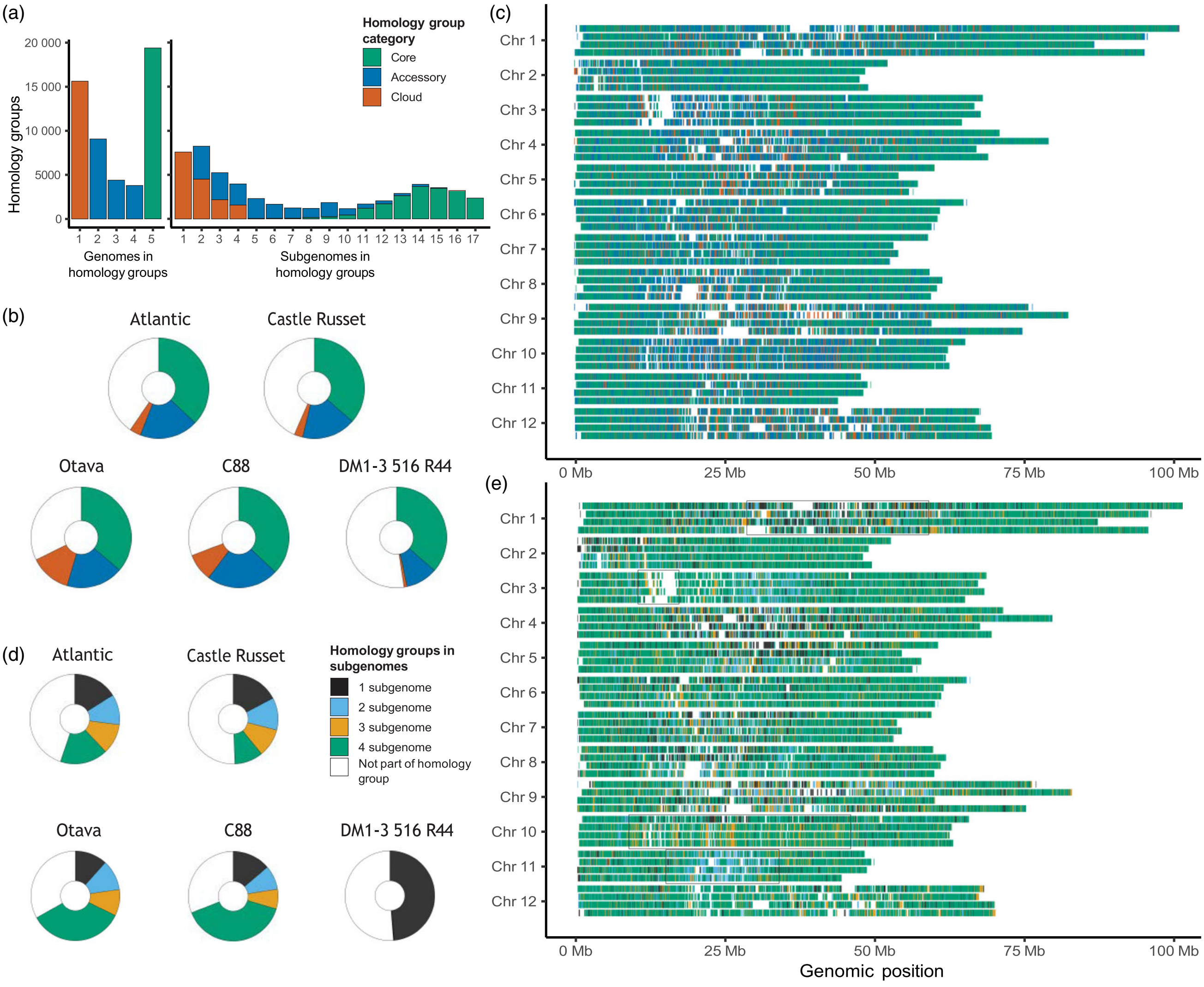


Figure 2 Characterization of the *S. tuberosum* pangenome gene content. (a) Number of genomes (left) and subgenomes (right) in the 52 240 homology groups. (b) Pie chart slices representing the proportion of groups being classified as core, accessory or cloud. Each circle represents the pangenome's 52 240 homology groups. (c) Gene regions of C88 coloured based on intergenomic variation, the pangenomic gene classification of five genomes. (d) Pie charts where the slices show the number of homology groups with genes in a total number of subgenomes. The proportion of white in these circles is slightly larger compared to plot b, because certain genes are not part of a subgenome; they are located on unphased sequences. (e) C88 gene regions coloured by intragenomic variation, their presence in 1–4 subgenomes.

backcrossing is the likely cause of this observable genomic architecture (Bao *et al*., 2022).

We generated genome visualization of the other four *S. tuberosum* genomes as well (Figures S10–S13 in Supplement S1). Otava was comparable to C88, having multiple chromosomes with a single disparate haplotype and also a chromosome with bi-allelic regions. In contrast, patterns in both CR and Atlantic mostly indicate limited phasing. Overall, these results demonstrate how subgenome-level pangenomics can help explore differences in terms of gene content, provided genome assemblies are of sufficient quality.

Establishing the evolutionary history in the pangenome

Inferring an accurate phylogeny is crucial for understanding the evolutionary history. The heterogenous pangenome graph allows us to infer phylogeny from different types of genetic variation, such as *k*-mers, SNPs or genes. The choice of input data and algorithm depends on the research objective. In this study, we were interested in comparing different methods to identify a reliable method for haplotype-resolved assemblies.

Here, we describe phylogenetic relationships in the *S. tuberosum* pangenome; the phylogenetic analysis of *Malus* is described in Analysis 3 in Supplement S2. Establishing a genome-level SNP tree on single-copy genes (occurring only once in every genome) is the most commonly used method for resolving plant phylogeny. It was not possible to directly obtain these high-resolution SNP trees for either of the haplotype-resolved pangenomes, as *S. tuberosum* had only 7 identified single-copy homology groups, whereas *Malus* had 17 groups.

Complete subgenomes cannot be directly compared, as the haplotype assignments per chromosome are ambiguous when genomes are assembled without parental data. Therefore, rather than making a phylogeny of complete subgenomes, we inferred separate trees per chromosome. PanTools includes a range of

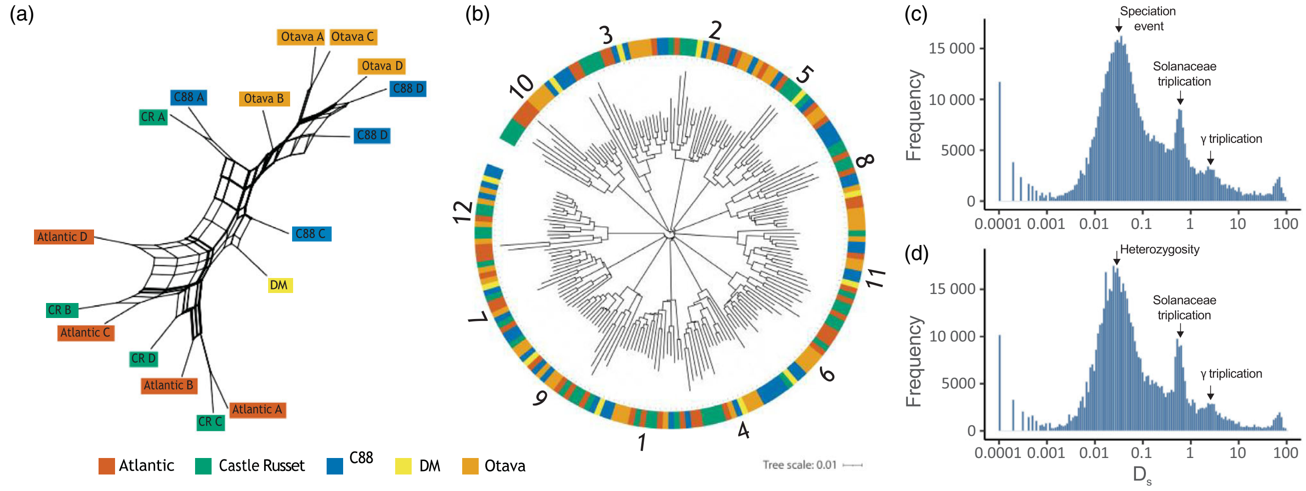


Figure 3 Evolutionary history of *S. tuberosum*. (a) Splits graph of *S. tuberosum* Chr 1 core SNP phylogeny. Tree labels are coloured by accession name. (b) *K*-mer distance phylogenetic tree of *S. tuberosum* sequences with at least 100 gene annotations. Clades are marked by a chromosome number. The tree is rooted at the midpoint. (c, d) Distribution of synonymous substitution rates (D_S) derived from homologous sequences between Otava and C88 (c), and within the single C88 genome (d). Evolutionary events are highlighted by arrows and labels.

phylogenomic methods that were adjusted to this end. Two methods use a distance based on either the number of shared *k*-mers or genes. Two more sophisticated methods, a core genome SNP tree and a consensus tree of core gene trees, were strongly hampered by the phasing quality as they require core gene content.

Here, we explore the tree topologies of *S. tuberosum* Chr 1 as a representative example for the other chromosomes. Tree topologies of the 11 remaining potato chromosomes show similar trends as Chr 1. As input for the Chr1 core SNP tree, we collected homology groups with one gene copy per haplotype. Genes of the individual Chr1 haplotypes cluster into 2155–4010 homology groups, but only 296 were shared among all haplotypes, with 52 groups identified as single-copy. From these single-copy groups, a low-resolution SNP tree with high ambiguity was inferred. A Splitstree representation (Huson, 1998) of this phylogeny in Figure 3a reveals clear conflicting signals. The consensus tree (Figure S14A in Supplement S1) was inferred from the 296 groups present in every (Chr 1) haplotype. Equivalent to the SNP tree, the consensus tree shows minimal branch support.

To obtain the *k*-mer and gene trees, we utilized the pangenome graph to extract shared entities and used these to calculate pairwise distances between sequences. From shared *k*-mers in the DBG we established a sequence-level *k*-mer distance tree. The tree shows 12 clades (Figure 3b), corresponding to the number of *S. tuberosum* chromosomes. None of the chromosome numbers assigned to the sequences conflicted with another, supporting a correct topology. The fourth tree was based on gene absence/presence identified from homology groups (Figure S14B in Supplement S1). This tree stands out as it shows that all Chr 1 haplotypes of a genome cluster together.

The hybrid origin of *S. tuberosum* explains the extensive conflict observed in the phylogenetic relationships (Tang *et al.*, 2022). Removing the lower quality genomes will certainly improve the phylogenetic signal. Nevertheless, truly resolving the potato taxonomy calls for more complex models such as phylogenetic networks that consider reticulation events (Blair and Ané, 2020).

To explore the WGD history of *S. tuberosum* we calculated synonymous (D_S) substitution rates of homologous genes, between and within genomes (Figure S15 in Supplement S1). The distribution of the intergenomic synonymous substitution rates between Otava and C88 reveals three clearly visible peaks that can be linked to evolutionary events (Figure 3c). The youngest and highest peak (D_S 0.001–0.01) derived from orthologous pairs indicates the speciation time of the two *S. tuberosum* species. Paralogous regions that originated in the *Solanaceae* paleohexaploidy appear as a second peak (D_S 0.6–0.9). A third, weak peak (D_S 2–3) provides evidence of the eudicot paleohexaploidy (γ) event.

The distribution of D_s substitutions is generally only reported between genomes. Therefore, we were interested in observing the intragenomic mutation patterns, and comparing haplotypes within a single genome. The intragenomic D_S distribution of C88 (Figure 3d) shows three peaks that look nearly identical to the distribution between C88 and Otava. This triple-peak pattern of mutation rates was not specific to C88 but was found in all intragenomic comparisons of the phased genomes (Figure S15 in Supplement S1). This was notable because, in unphased (haploid) genome assemblies, synonymous substitution plots reveal only whole-genome duplication events. A plausible explanation for the visibility of the youngest peak (D_S 0.001–0.01) is, that as a result of phasing, alleles located on the different haplotypes are now aligned whereas in unphased assemblies only duplicated genes are aligned. Although the first peak in intergenomic comparisons is associated with speciation, in intragenomic haplotype-resolved assemblies it reflects heterozygosity between subgenomes. These results demonstrate that pangenomic evolutionary analyses offer far more insight when performed at the subgenome level.

Extensive haplotype-specific variation revealed in ancient polyploids

Comparing the genomic organization of organisms uncovers the genomic conservation and rearrangements, which provide insights into the evolutionary dynamics of genomes. The homology grouping, together with established phylogenetic relationships in the pangenome, serves as a framework to analyse genome organization. Here, we discuss several PanTools methods to examine structural conservation and changes among *Malus* chromosomes; in Analysis 4 in Supplement S2, we apply these same approaches to the potato pangenome.

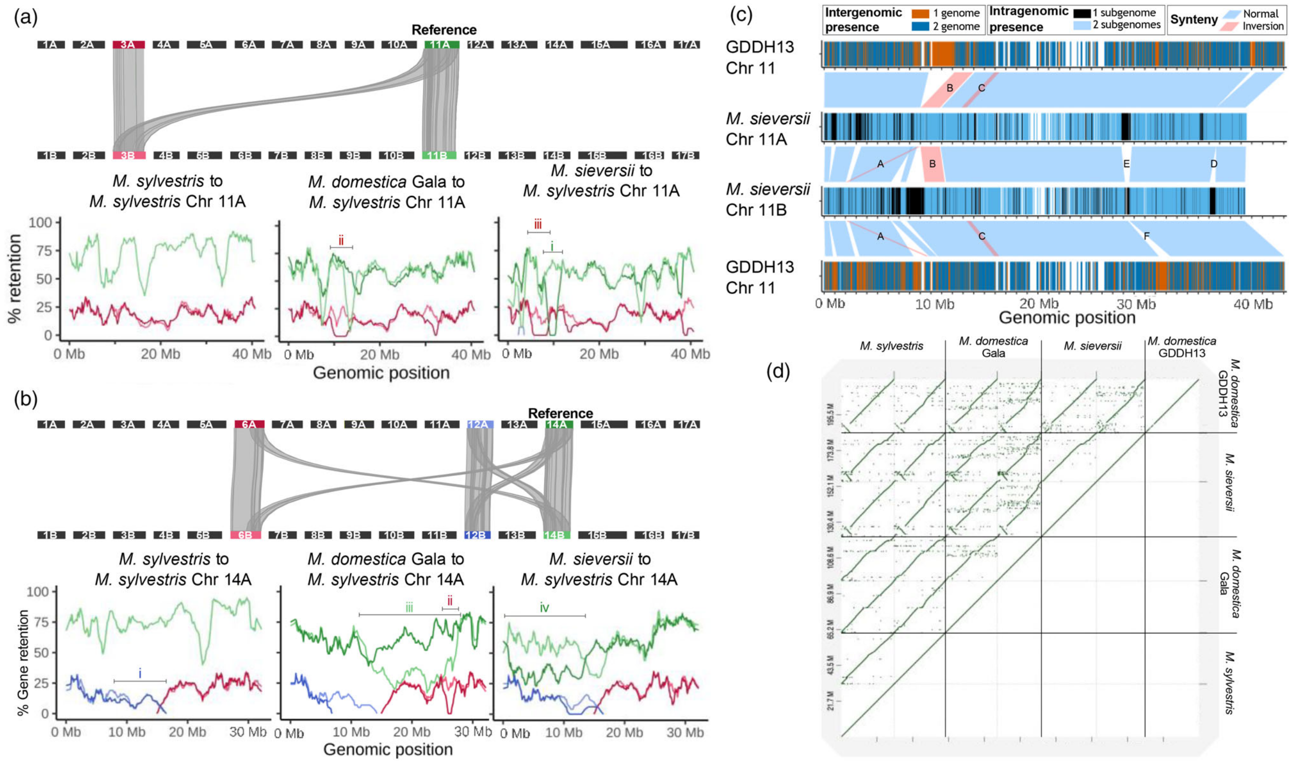


Figure 4 Structural variant visualizations on the *Malus* pangenome. (a, b) Syntenic gene retention of *M. sylvestris* Chr 11A (a), 14A (b) to every sequence of the pangenome. Some regions showing a divergent retention pattern were numbered and are discussed in the text. A schematic representation of macrosynteny between *M. sylvestris*'s chromosomes above each trio of retention plots shows the underlying genomic organization. Synteny relations are drawn between the selected reference sequence and all collinear regions and the corresponding chromosome in the other haplotype. (c) Genetic and structural variation in *M. sieversii* Chr 11, in relation to GDDH13. There are three different types of annotation bars. Starting from the top: GDDH13 Chr 11 gene regions shared with *M. sieversii* 11A (blue) or not (red); syntenic blocks; and gene occurrence in one (black) or both (blue) haplotypes. (d) Dot plot visualization of *Malus* Chr 1 alignments.

Through synteny analyses, we determined pairwise conserved collinearity between all sequences in the pangenome. The macrosynteny suggests high collinearity, with blocks spanning nearly entire chromosomes. However, at the microsynteny level, between syntenic blocks, additional genes were not collinear but rather just fall between two syntenic anchors. To this end, we developed a visualization of both large-scale genome structure analysis (macrosynteny) and conservation of local gene content and order (microsynteny). Retention is calculated in sliding windows based on homology (conserved gene sequence) or synteny (conserved gene order). Two examples on apple demonstrate the use of these visualization methods and show the high diversity between the haplotypes and reveal the fate of duplicated chromosomes post-polyploidization.

Retention patterns based on *M. sylvestris* Chr 11A (Figure 4a) as a query were representative of the majority of *Malus* visualizations; one chromosome pair (green lines) showing high retention of syntenic genes, with another pair (red lines) having lost most collinear gene pairs. The two most similar haplotype sequences represent chromosomes homeologous to the selected query. Less retained sequences are remnants of the most recent WGD. Overall, gene retention w.r.t. the homeologous chromosomes was high, although with strong local fluctuations. The strongest loss of synteny was observed in *M. sieversii* Chr 11A around genomic position 10 Mb (region i), where collinearity was fully lost. The observable retention pattern of the WGD-duplicated regions relative to the reference is highly similar between all three retention plots. Two prominent exceptions of regions displaying disparate synteny levels are Gala 10–13 Mb (region ii) and *M. sieversii* 7–9 MB (region iii), as no syntenic gene pairs were found within these regions. The visualizations support hypotheses of fractionation initially quickly degrades duplicated regions, but further advances at a diminishing rate (Zhang *et al*., 2021).

In a second example, we explore the retention plots using *M. sylvestris* Chr 14A as query. The WGD-duplicated regions were fragmented and located on two different chromosomes, indicated by the red and blue lines in Figure 4b. Such rearrangements are frequently seen in WGDs, after which polyploids may gradually return to a diploid state (Mandáková and Lysak, 2018). Different models involving diploidization have been proposed to explain the chromosomal organization of diploid *Malus* species (Considine *et al*., 2012; Velasco *et al*., 2010). The level of gene retention in these WGD-derived segments was highly similar across genomes and displayed just two outlying patterns. Only *M. sylvestris* retained the collinearity of duplicated genes on Chr 12 (region i, 10–15 Mb) on both haplotypes. Conversely, Gala Chr 6B (region ii, 25–26 Mb) was the only region that lost all genes syntenic with the reference sequence. Aside from this rearrangement, there was substantial fractionation in the homeologous chromosomes (green lines) illustrated by a high loss of synteny in nearly half the chromosome in both Gala (region iii) and *M. sieversii* (region iv). In Analysis 5 in Supplement S2 we examined another *Malus* chromosome query that showed the strongest reduction of duplicated regions and also revealed a translocated region.

We developed a novel visualization functionality that combines genomic and pangenomic features extracted from the graph database. In Figure 4c we show an example of Chr 11 from *M. sieversii* and GDDH13 where we combine intra- and intergenomic gene absence/presence variation with synteny annotations. *M. sieversii*'s Chr 11 was selected because of the high number of intragenomic rearrangements. Earlier, Chr 11 of *M. domestica* GDDH13 was used to assist in phasing the *M. sieversii* assembly (Sun *et al*., 2020); therefore, we included it here to place the variation in the context of a reference. The visualization shows the large blocks of unshared genes correlate to the intra- and intergenomic synteny breakpoints. Synteny blocks further reveal three major inversions. The leftmost inversion (marked by letter A) was specific to sequence 11B and appeared to be translocated as well. The second haplotype-specific inversion (B) in 11A (9–11 Mb, 130 genes) was the longest identified structural variation between any *Malus* chromosomes. The third inversion (C) was intergenomic and, therefore, only visible (around position 13–14 Mb) between *M. sieversii* and GDDH13.

Aside from inversions, synteny relationships display multiple breakpoint regions. All synteny breakpoints were due to haplotype-specific insertions/deletions, emphasizing the importance of intragenomic variation. Synteny was lost around position 37 Mb (D) because of a nearly 1 Mb-sized region in which no gene is shared. Perhaps even more intriguing is the two-sided breakpoint (E) identified at position 28 Mb, where both haplotypes show a distinct region of only non-homologous genes. The second half of this non-syntenic region (F) on Chr 11B was the only block of haplotype-specific genes that broke synteny to the GDDH13 reference. Possibly, this could be a genomic fragment introgressed into the *M. sieversii* genome, or it could have been lost from all other haplotypes. Considering all other haplotype-specific regions in *M. sieversii* are syntenic to GDDH13, it is most likely these synteny breakpoints are the result of collapsed haplotypes.

With PanTools' visualization utility, we created an image for each haplotype-resolved chromosome set of apple and potato to provide a comprehensive view of the intragenomic variation (Supplement S3). These visualizations display gene regions with colouring according to the presence in number of subgenomes, with synteny relationships drawn between the chromosomes. In addition, the apple visualizations display the percentage of the genome that overlaps with gene and repeat annotations, typically revealing an anti-correlation between the two features. Each image offers a clear overview of gene-absence variation and gene order conservation in sets of two (apple) or four (potato) homologous chromosomes.

Another perspective on intragenomic and intergenomic variation across a set of chromosomes is offered by a multi-genome dot plot. Dot plots are popular visualizations to identify large-scale deletions, inversions and repeats. Using the earlier established phylogenetic relationships, only homeologous chromosomes were aligned to another. In Figure 4d, we show the dot plot visualization of all *Malus* Chr 1 haplotypes in the pangenome. Notably, both GDDH13 as *M. sieversii* have an 5 MB inversion with regard to the other *Malus* genomes. This inverted region in *M. sieversii* displays another small inversion, that is, most clearly visible against *M. sylvestris*'s chromosomes.

Graphically representing the genomic organization with structural variation supports a better understanding of the complexity of genomes. The introduction of new PanTools features allows for users to create both novel and more traditional plots to display chromosomal rearrangements. The presented examples demonstrated extremely high variation between haplotypes. Regardless of whether the observed variations reflect true biology or assembly artefacts, these visualizations can provide valuable support for comparative genomic analyses.

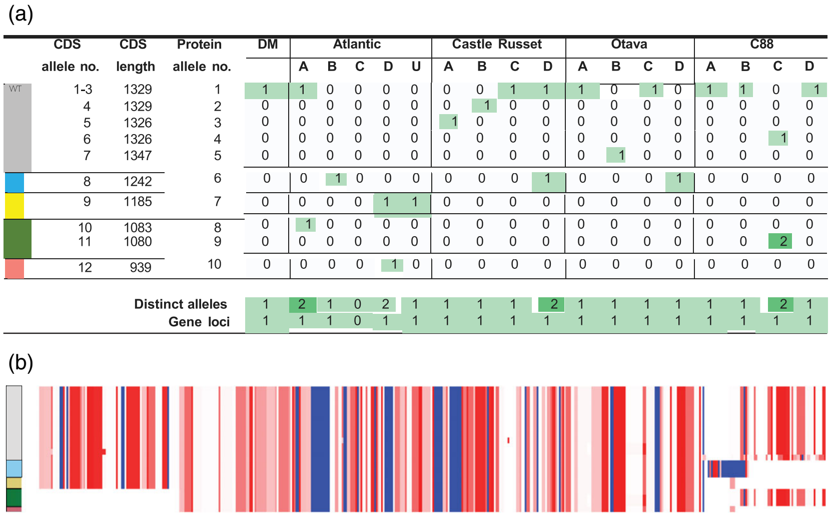
Exploring allelic diversity

A desirable feature of pangenomes is the ability to identify all allelic variants of genes for functional selection and breeding. We demonstrate novel functionalities for this purpose on the potato gene *StCDF1*. *S. tuberosum* originates from a region close to the equator and its adaption to short-day growing conditions prevents tuber formation in the long-day conditions during spring and summer in higher latitude locations. The transcription factor CYCLING DOF FACTOR 1 (*StCDF1*) is a key regulator for reaching maturity and tuber formation (Kloosterman *et al*., 2013). Potato plants adapted to longer day lengths have specific *StCDF1* allelic variants (Kloosterman *et al*., 2013).

The pangenome database was utilized to find all allelic variation in *StCDF1* among the potato cultivars. First, *StCDF1* was identified in DM Chr 5, where it clustered in a homology group of 22 proteins. The hierarchical pangenome annotations allowed the extraction of not only the protein sequences but also the encoding gene, transcript and CDS sequences. Twenty-two proteins were derived from 17 genes. Manual BLAST against the genome assemblies verified the presence of 17 *StCDF1* loci. Figure 5a provides an overview of protein sequences within the homology group, showing which alleles are present in each subgenome.

Through protein sequence alignment five major *StCDF1* allelic groups were distinguished based on truncations and insertions (Figure 5b). The wild-type (WT) allele (*StCDF1.1*, protein allele no. 1) in the grey group coding for full-length *StCDF1* proteins was the most abundant and found at least once per genome. In the C88 cultivar, the WT is present on 3 subgenomes, while in the Atlantic it was limited to one. Blue and yellow groups encode proteins with a 3′ (C-terminal) truncation; blue group genes had new coding sequences inserted. The green group was characterized by a 5′ (N-terminal) truncation and consists of one Atlantic gene and two C88 genes. Lastly, the pink group gene in Atlantic had both 5′ and 3′ truncations.

Apart from the wild-type allele *StCDF1.1*, alleles *StCDF1.2* to *StCDF1.5* have been identified earlier at the locus (Achakkagari *et al*., 2022). These alleles carry specific insertions, leading to FKF domain truncation in the 3′ region, thereby avoiding ubiquitination (Kloosterman *et al*., 2013). Even though *StCDF1* is intensively studied, to our knowledge, so far no *StCDF1* alleles were reported with a truncated 5′ region. While our overview shows a total of 3 truncated 5′ alleles, we remain cautious as the prediction of gene models is highly complex and often results in incorrect annotations. The Atlantic and C88 genomes, in which these transcripts were found, were annotated following comprehensive strategies supported with sufficient transcript evidence (Bao *et al*., 2022; Hoopes *et al*., 2022). Upon inspecting the gene models we found that the 3 alleles with 5′ truncations are protein isoforms. These isoforms arose from variations in the exon–intron boundary predictions of gene regions, resulting in different models. As a follow-up, the gene models were examined in the Spud DB Jbrowse instance (http://spuddb.uga.edu). Mapped leaf and tuber ONT cDNA reads showed very minimal support for alternative splicing in the two Atlantic genes (Figure S16 in Supplement S1). Moreover, C88 lacked transcript/cDNA data for validation.



	CDS allele no.	CDS length	Protein allele no.	DM	Atlantic					Castle Russet				Otava				C88			
					A	B	C	D	U	A	B	C	D	A	B	C	D	A	B	C	D
WT	1-3	1329	1	1	1	0	0	0	0	0	0	1	1	1	0	1	0	1	1	0	1
	4	1329	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	5	1326	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	6	1326	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	7	1347	5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	8	1242	6	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
	9	1185	7	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	10	1083	8	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	11	1080	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
	12	939	10	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
			Distinct alleles	1	2	1	0	2	1	1	1	1	2	1	1	1	1	1	1	2	1
			Gene loci	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 5 (a) Occurrence of *StCDF1* alleles in *S. tuberosum* subgenomes. Below each column, the total number of unique alleles and genes in a subgenome is given. (b) Alignment of 22 *StCDF1* protein sequences (visualized via https://alignmentviewer.org). Amino acid residues are coloured by hydrophobicity, where (dark) red is the most hydrophobic and blue is the most hydrophilic. Assigned groups (colours) based on truncations and insertions are shown on the left of the alignment.

A potato plant reaches maturity when its tubers are fully developed and ready for harvest, the timing of which varies strongly between cultivars. Early maturity of the Atlantic cultivar is attributed to multiple 3′ truncated alleles (Hoopes *et al*., 2022), not necessarily the 5′ truncation. The C88 variety, despite having four *StCDF1.1* alleles, can still reach maturity in long-day conditions (Li *et al*., 2011; Myrick *et al*., 2021). C88 has two non-WT alleles. One is still considered *StCDF1.1* and shows a 3 bp deletion outside the FKF domain, that is, unlikely to affect gene functionality. This leaves the truncated 5′ allele as a potential candidate variant for C88's long-day acclimation.

To conclude, PanTools helped identify allelic diversity in *StCDF1*, distinguished by major truncations in the 3′ and 5′ regions. Further validation of the diversity found is necessary, requiring greater sequencing depth or more extensive experimental validation of the gene.

Concluding discussion

Genomic analyses provide an important foundation for our understanding of biology. Our current ability to resolve genomes at the haplotype level provides a representation, that is, far more accurate than the collapsed genomes studied thus far. However, methodology to easily analyse collections of such genomes is still scarce. We updated the PanTools pangenomics platform and added new functionalities to represent phased genome assemblies and enable the identification of intragenomic variation. We demonstrated these functionalities on tetraploid potato and diploid apple pangenomes, showing both practical applications and the potential for plant breeding (e.g. maturity in potato).

The most critical factor for accurate pangenome analysis is the quality of genome assemblies. We showed that BUSCO completeness should be assessed at the subgenome level to assess phasing quality, and that visualization of subgenomes with polymorphisms is essential to decide whether certain variation actually reflects biology or results from an assembly/phasing artefact. The defined subgenomes lack any biological linkage, making them only applicable to assess gene presence and frequencies in random haplotype combinations. Our analyses further revealed high heterozygosity in potato and apple, characterized by gene absence/presence variation and structural rearrangements. Overall, our results demonstrate the usefulness of combining a pangenome representation with state-of-the-art bioinformatics tools for detailed intra- and intergenomic analyses.

In the past decade, we have seen a rise in many pangenomic software toolkits (Naithani *et al*., 2023). PanTools distinguishes itself through its hierarchical pangenome graph and by providing an extensive set of comparative genomics functionalities, partially through a connection to existing tools. PanTools facilitates comprehensive pangenome analyses, from the initial quality checks to the downstream analyses. In this study, we introduced a new set of functionalities specific for haplotype-resolved assemblies. These were based on typical comparative genomic analyses, but we adopted a pangenomic approach for their implementation.

A pangenome represents all variations found in a population and provides a valuable overview of all available alleles. The application of pangenomics methodology as currently implemented in PanTools to haplotype-resolved genomes is already promising, but further development will allow a fuller exploration of this rich source of data. More interactive visualization methods will facilitate visual analytics, that is, user-guided exploration of (sub)genomic architecture and structural variations. This approach should allow users to choose any chromosome as a reference and zoom into specific regions of interest, enabling the visualization of complex genetic patterns.

We can also use the current pangenomic representation framework for analysing additional sources of data. PanTools was designed to easily include such diverse data types to enable the study of complex biological systems. We observed that only half of the genes in genomes are found on all subgenomes, often with high variation among alleles. Our pangenome representation can overcome limitations imposed by a reference bias and facilitate analyses hampered by a reference bias, such as the identification of allele-specific expression.

PanTools' hierarchical pangenome graph holds potential for exploring the evolutionary history. In this study, we observed extensive fractionation after whole-genome duplication events in two autopolyploid pangenomes. In contrast, allopolyploids have highly divergent parental genomes, often leading to one subgenome becoming dominant over the other (Bird *et al*., 2018). PanTools can currently identify biases in gene content and fractionation; however, studying subgenome dominance requires the integration of expression and epigenetic data (Bird *et al*., 2018). The advent of more haplotype-resolved genomes will drive the methodological development needed to perform such analyses, which will help obtain an increasingly detailed picture of pangenome content, organization and evolution.

Methods

Genome and annotation data collection

We focus on two use cases, a potato (*Solanum tuberosum*) and apple (*Malus*) pangenome. Both use cases were built upon publicly available datasets. For the potato use case, all data was downloaded directly from Spud DB (http://spuddb.uga.edu) in January 2022. Two phased Atlantic and Castle Russet assemblies were obtained from the study of Hoopes et al. (Hoopes *et al*., 2022). Genomic data of Otava were derived from the Sun *et al*. (2022) paper. C88 data was accompanied by the publication from Bao *et al*. (2022). The only unphased potato genome was the most recent (v6.1) assembly of DM 1-3516 R44 created by Pham *et al*. (2020). For apple, all three haplotype-resolved genomes *M. domestica* (v1.0), *M. sieversii* (v1.0) and *M. sylvestris* (v1.0) were obtained from the study of Sun et al. (Sun *et al*., 2020). Two additional phased genomes were obtained; the apple reference genome *M. domestica* GDDH13 (v1.1) collected from the publication by Daccord et al. (Daccord *et al*., 2017) and a wild apple genome *M. baccata* assembly (v1.0) by Chen et al. (Chen *et al*., 2019). Original GFF annotation files of *M. baccata* and DM were invalid and were updated with the AGAT toolkit (agat_convert_sp_gxf2gxf.pl) (Dainat, 2022) to include missing 'gene' features.

Pangenome construction and annotation

Potato and apple pangenomes were built and analysed using the 'phased_pangenomics' development branch in the PanTools repository. Here, we briefly discuss the PanTools functions and their respective arguments used to perform the analysis. For more detailed explanations of the underlying algorithms, we refer to the online manual (https://pantools.readthedocs.io). The pangenomes were constructed with the 'build_pangenome' function and a *k*-mer size *k* of 19. Structural annotations were connected to the De Bruijn graph (DBG) with 'add_annotations'. The gene, transcript and protein sequences created by PanTools were compared to the extracted sequences of AGAT (agat_sp_extract_sequences.pl) (Dainat, 2022). Chromosome and haplotype information was added to the database with 'add_phasing'. Species names were included as metadata using 'add_phenotypes'. Transposable element annotations derived from EDTA (v2.0.0) (Ou *et al*., 2019) with default settings were included in the pangenome database through the 'add_repeats' function.

Determining the optimal homology grouping

PanTools' 'busco_protein' function assessed the completeness of the pangenome using BUSCO (v5.3.2) (Manni *et al*., 2021) with the most specific lineage datasets; *Solanales* odb10 for potato and eudicots odb10 for the apple genomes. We assessed the completeness of genomes and separate subgenomes. First, 'busco_protein' was run without any additional arguments to evaluate the entire proteome of each genome assembly. Second, by setting the '–phasing' and '–longest-transcripts' arguments, BUSCO was performed against proteome subgenome subsets only including the longest protein-coding transcripts of genes.

Proteins were clustered seven times with different strictness with the 'optimal grouping' (Jonkheer *et al*., 2022) functionality. Clustering strictness is altered by changing the minimum required (normalized) similarity score between two sequences and tweaking the parameters controlling MCL (Markov clustering) (Enright *et al*., 2002). We calculated the *F*1 score (the harmonic mean of precision p and recall r, $F1 = 2(p \cdot r)/(p + r)$) for the seven groupings based on BUSCO-identified single-copy genes that are present in every subgenome. Assuming these genes are truly single-copy, a perfect clustering would place them in a separate homology group with one representative protein per subgenome. Following this assumption, we scored the grouping based on the BUSCO genes that actually cluster together and whether other genes cluster with them.

As a second measure to assess the protein clustering, we quantified the degree of overlap among gene sets between haplotypes of the same subgenome. First, a Jaccard index is calculated from shared genes within homology groups for the possible sequence combinations in a subgenome. Subsequently, the average distance in the subgenome was calculated from all combinations, followed by taking the mean of the individual averages.

Phylogenetic analyses

PanTools facilitates multiple methods to represent genomic distances among genomes or individual sequences in the pangenome. A *k*-mer distance tree was created for all chromosome-length sequences in the pangenome using the 'kmer_classification' method. MASH distance (Ondov *et al*., 2016) is calculated between two sequences by counting the shared *k*-mers in the DBG. The pairwise distances were stored in a matrix and served as input for inferring a Neighbour-Joining (NJ) tree using ape (v5.0) (Paradis and Schliep, 2019). The tree topology was validated by checking if chromosome numbers (included in genome FASTA headers) conflict within clades.

The 'core_phylogeny' method was used to create one sequence-level tree per chromosome. Single-copy groups were identified through 'gene_classification', in which only sequences belonging to a specific chromosome were included. The sequences of single-copy groups were aligned with MAFFT (v7.453) (Katoh and Toh, 2010) and trimmed to avoid noisy regions near the end of the sequence. The multiple sequence alignment (MSA) was performed in two steps. After an initial protein alignment, the longest start and end gaps were used to trim the nucleotide sequences. These trimmed sequences were input for the second alignment. A concatenated sequence of parsimony-informative single-nucleotide polymorphisms (SNPs) was created per haplotype. IQ-tree (v1.6.12) (Nguyen *et al*., 2015) was applied to the collection of concatenated sequences with 10 000 bootstrap replications.

With the 'consensus_tree' function, one tree per chromosome was generated that summarizes all gene trees associated with a chromosome. Homology groups shared by all sequences of a specific chromosome were first identified as input. To obtain gene trees, sequences were aligned as described in 'core_phylogeny'. FastTree (2.1.10) was applied to the homology group MSAs using default parameters.

The gene trees were combined in a file, from which ASTRAL-Pro (version March 2022) (Zhang *et al*., 2020) with preset configurations estimated a consensus tree.

Gene distance calculated by 'gene_classification' is the third type of distance to create a sequence tree for a single chromosome. Jaccard distances were obtained by counting shared genes and total genes between two genomes using the homology groups. Only unique elements were considered, and additional gene copies were ignored. Gene distance matrices were visualized as NJ trees created by ape (v5.0) (Paradis and Schliep, 2019). All phylogenetic tree visualizations were created with iTOL v6 (Letunic and Bork, 2019).

Gene-based analyses

PanTools' 'gene_classification' function with the '–phasing' argument was used to characterize the pangenome gene content. The homology grouping serves as the foundation for all analyses of this section. Genes were intergenomically categorized as follows: core genes were present in all genomes, accessory genes were absent in some genomes, and cloud (formerly called unique) genes were found in a single genome. For the intragenomic characterization we estimate if gene presence is in line with the ploidy of an organism, and count its occurrence in each chromosome set (1 to 2 in apple and 1 to 4 in potato). Furthermore, separate countings were performed for the individual chromosomes. For instance, a gene is found in 2 out of 4 Otava (potato) Chr 1 haplotypes. This intragenomic counting required incorporated haplotype information via 'add_phasing'. Several examples of how these classification rules were applied are included as Figure S3 in Supplement S1.

Homology groups were further used to determine the frequency of genes and alleles. Gene copies were directly counted from these groups. For the allele frequency, we considered every nucleotide or amino acid polymorphism within a group to represent a distinct allele. The nucleotide and protein sequences of a group were collected in two separate sets from which the unique elements were counted.

Synteny estimation and graph integration

Synteny blocks were computed with the 'calculate_synteny' function. First, the GFF and homology input files required for MCScanX (version October 2020) (Wang *et al*., 2012) were generated. The (highly simplified) GFF files only contain the sequence identifiers together with gene start and stop coordinates, belonging to a single sequence. The homology files state which genes are homologous to another between two sequences, and were created for every possible sequence combination. Iterating over the input files, MCScanX was employed in parallel using default settings. Subsequently, separate output files were combined into a single collinearity file, which was included in the database using 'add_synteny'. Genes part of the same syntenic block were connected in the graph through 'synteny' nodes, syntenic gene pairs gained a direct 'is_syntenic_with' relationship to each other.

Estimate synonymous substitutions rates

Synonymous (D_s) and nonsynonymous (D_n) mutation rates were calculated for homology group MSAs. The alignment was performed in two rounds, as described in the 'Phylogenetic analyses' section above. PAL2NAL (v14) (Suyama *et al*., 2006) was used to convert protein alignments into corresponding codon alignments, codons with gaps and inframe stop codons were excluded. Sequences shorter than 30 amino acids were excluded to minimize artefacts caused by short alignments. D_n and D_s values were calculated in the codon alignments with codeML (part of PAML package (Yang, 2007)).

Gene retention visualizations

The retention pattern visualization was created with PanTools' 'gene_retention' function. To calculate the retention of all sequences to a selected query sequence, the following steps are performed. First, all gene nodes of the query sequence are collected and ordered based on their genomic position. Then, a sliding window of 100 genes moves over the nodes in steps of 10. The window stops when it no longer can move 10 genes to the right, resulting in the visualization of full-sized windows only. At each window position, the percentage of retention is calculated based on shared homologues and syntelogs between the query to every other sequence. Genes are considered homologues when part of the same homology group, while syntelogs are required to be part of the same synteny block and form a syntenic pair. To ensure retention does not exceed 100%, syntenic depth is ignored, as it is highly influenced by gene duplications. Window positions were transformed into genomic coordinates of the query sequence, at which the retention values were plotted with ggplot2 (Wickham, 2016).

Whole genome alignment

Sequences belonging to the same chromosome (number) were aligned with minimap2 (v2.24) (Li, 2018) using the '-x asm5' parameter. Variants were called between two sequences with paftools.js and filtered out when the quality was below 10. The genome alignments were visualized with D-GENIES (v1.4) (Cabanettes and Klopp, 2018), where up to the 100 000 best matches per alignment were plotted.

Sequence visualizations

The chromosome visualizations based on various annotations were created with PanTools 'sequence_visualization'. The function can generate five types of annotation bars. First, gene regions were coloured by their presence in other genomes, in concordance with the homology group category: core, accessory or cloud. Second, gene regions were coloured based on their presence in the number of subgenomes within a single genome. Third, gene regions were coloured grey if a gene was found on any other chromosome. Still, gene copies on the same chromosome (number) but different haplotype do not allow the region to be coloured grey. In the fourth annotation bar, a line graph plots the coverage (percentage) of the repeat and gene regions (annotations) in 1 Mb windows. Repeat coverage of 100% indicates every nucleotide in the window overlapped with at least one repeat annotation. The fifth and final annotation represents synteny blocks that allow to connect two sequences. Annotation bars were plotted individually with ggplot2 (Wickham, 2016) and stacked horizontally using the cowplot R package (https://github.com/wilkelab/cowplot).

Author contributions

EJ and SS designed the study. EJ and SS extended the functionality of PanTools, and EJ performed the computational analyses. Result and interpretation, shaping the manuscript, were discussed with the domain experts: EJ, TL, LB, JH and SS. The manuscript was structured and written by EJ and DR. All authors critically reviewed the manuscript.

Acknowledgements

We would like to thank Richard Finkers for the helpful feedback and for testing the new PanTools functionalities. We also want to thank Christian Bachem for the insightful discussions.

Funding

This work was jointly funded by Biointeractions and Plant Health, Wageningen Plant Research, Genetwister Technologies B.V., and the Netherlands eScience Center, project ETEC.2019.019.

Data availability statement

PanTools v4 is available at https://git.wur.nl/bioinformatics/pantools, released under the GNU GPLv3 licence. The novel functionalities were developed in the 'phased_pangenomics' branch, which was merged into 'pantools_v4' at commit d5eca936 (https://git.wur.nl/bioinformatics/pantools/-/commits/) and released in v4.3.0 (https://git.wur.nl/bioinformatics/pantools/-/releases/v4.3.0). Instructions for downloading the publicly available datasets and reproducing the experiments are in Supplement S3.

References

Achakkagari, S.R., Kyriakidou, M., Gardner, K.M., De Koeyer, D., De Jong, H., Strömvik, M.V. and Tai, H.H. (2022) Genome sequencing of adapted diploid potato clones. *Front. Plant Sci.* **13**, 954933. https://doi.org/10.3389/fpls.2022.954933

Bao, Z., Li, C., Li, G., Wang, P., Peng, Z., Cheng, L., Li, H. *et al*. (2022) Genome architecture and tetrasomic inheritance of autotetraploid potato. *Mol. Plant* **15**, 1211–1226.

Bird, K.A., VanBuren, R., Puzey, J.R. and Edger, P.P. (2018) The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytol.* **220**, 87–93.

Blair, C. and Ané, C. (2020) Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic´ data. *Syst. Biol.* **69**, 593–601.

Brinton, J., Ramirez-Gonzalez, R.H., Simmonds, J., Wingen, L., Orford, S., Griffiths, S., 10 Wheat Genome Project *et al*. (2020) A haplotype-led approach to increase the precision of wheat breeding. *Communications Biology* **3**, 712.

Cabanettes, F. and Klopp, C. (2018) D-genies: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958.

Chen, X., Li, S., Zhang, D., Han, M., Jin, X., Zhao, C., Wang, S. *et al*. (2019) Sequencing of a wild apple (*Malus baccata*) genome unravels the differences between cultivated and wild apple species regarding disease resistance and cold tolerance. *G3: Genes, Genomes, Genetics* **9**, 2051–2060.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. and Li, H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**(2), 170–175.

Considine, M.J., Wan, Y., D'Antuono, M.F., Zhou, Q., Han, M., Gao, H. and Wang, M. (2012) Molecular genetic features of polyploidization and aneuploidization reveal unique patterns for genome duplication in diploid malus. *PLoS One* **7**, e29449.

Cornille, A., Gladieux, P., Smulders, M.J., Roldán-Ruiz, I., Laurens, F., Le Cam, B., Nersesyan, A. *et al*. (2012) New insight into the history of domesticated apple: Secondary contribution of the european wild apple to the genome of cultivated varieties. *PLoS Genet.* **8**, e1002703.

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H. *et al*. (2017) High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106.

Dainat, J. (2022) *AGAT: another Gff analysis toolkit to handle annotations in any GTF/GFF format*. https://doi.org/10.5281/zenodo.3552717

Duan, H., Jones, A.W., Hewitt, T., Mackenzie, A., Hu, Y., Sharp, A., Lewis, D. *et al*. (2022) Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in nanopore and hifi assemblies with hi-c data. *Genome Biol.* **23**, 84.

Enright, A.J., Dongen, S.V. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584.

Gaiero, P., Speranza, P. and de Jong, H. (2018) Introgressive hybridization in potato revealed by novel cytogenetic and genomic technologies. *Am. J. Potato Res.* **95**, 607–621.

Garg, S., Fungtammasan, A., Carroll, A., Chou, M., Schmitt, A., Zhou, X., Mac, S. *et al*. (2021) Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**, 309–312.

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J. *et al*. (2023) *Building pangenome graphs. bioRxiv*. https://doi.org/10.1101/2023.04.05.535718

Hamlin, J.A.P., Dias, G.B., Bergman, C.M. and Bensasson, D. (2019) Phased diploid genome assemblies for three strains of candida albicans from oak trees. *G3 Genes—Genomes—Genetics* **9**, 3547–3554.

Han, R., Han, L., Zhao, X., Wang, Q., Xia, Y. and Li, H. (2022) Haplotype-resolved genome of sika deer reveals allele-specific gene expression and chromosome evolution. *Genom Proteom Bioinform* **21**, 470–482.

Hasing, T., Tang, H., Brym, M., Khazi, F., Huang, T. and Chambers, A. (2020) A phased vanilla planifolia genome enables genetic improvement of flavour and production. *Nature Food* **1**, 811–819.

Hickey, G., Monlong, J., Ebler, J., Novak, A.M., Eizenga, J.M., Gao, Y., Human Pangenome Reference Consortium *et al*. (2024) Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673.

Hoopes, G., Meng, X., Hamilton, J.P., Achakkagari, S.R., de Alves Freitas Guesdes, F., Bolger, M.E., Coombs, J.J. *et al*. (2022) Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. *Mol. Plant* **15**, 520–536.

Huson, D.H. (1998) Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics (Oxford, England)* **14**, 68–73.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R. *et al*. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345.

Jiang, X., Li, D., Du, H., Wang, P., Guo, L., Zhu, G. and Zhang, C. (2023) Genomic features of meiotic crossovers in diploid potato. *Horticulture Research* **10**, uhad079.

Jonkheer, E.M., van Workum, D.-J.M., Sheikhizadeh Anari, S., Brankovics, B., de Haan, J.R., Berke, L., van der Lee, T.A.J. *et al*. (2022) PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics* **38**, 4403–4405.

Jung, S., Cestaro, A., Troggio, M., Main, D., Zheng, P., Cho, I., Folta, K.M. *et al*. (2012) Whole genome comparisons of Fragaria, Prunus and Malus reveal different modes of evolution between Rosaceous subfamilies. *BMC Genomics* **13**, 129.

Katoh, K. and Toh, H. (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**, 1899–1900.

Kloosterman, B., Abelenda, J.A., Gomez, M.D.M.C., Oortwijn, M., de Boer, J.M., Kowitwanich, K., Horvath, B.M. *et al*. (2013) Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature* **495**, 246–250.

Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S. *et al*. (2018) De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182.

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P.A. and Bernatchez, L. (2020) Using haplotype information for conservation genomics. *Trends Ecol. Evol.* **35**, 245–258.

Letunic, I. and Bork, P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100.

Li, C., Wang, J., Chien, D.H., Chujoy, E., Song, B. and VanderZaag, P. (2011) Cooperation-88: A High Yielding, Multi-Purpose, Late Blight Resistant Cultivar Growing in Southwest China. *Am. J. Potato Res.* **88**, 190–194.

Li, Z., De La Torre, A.R., Sterck, L., Cánovas, F.M., Avila, C., Merino, I., Cabezas, J.A. *et al*. (2017) Single-copy genes as molecular markers for phylogenomic studies in seed plants. *Genome Biol. Evol.* **9**, 1130–1147.

Lin, H., Yao, Y., Sun, P., Feng, L., Wang, S., Ren, Y., Yu, X. *et al*. (2023) Haplotype-resolved genomes of two buckwheat crops provide insights into their contrasted rutin concentrations and reproductive systems. *BMC Biol.* **21**, 1–14.

Mandáková, T. and Lysak, M.A. (2018) Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.* **42**, 55–65.

Manni, M., Berkeley, M.R., Seppey, M. and Zdobnov, E.M. (2021) BUSCO: assessing genomic data quality and beyond. *Current Protocols* **1**, e323.

Marand, A.P., Jansky, S.H., Zhao, H., Leisner, C.P., Zhu, X., Zeng, Z., Crisovan, E. *et al*. (2017) Meiotic crossovers are associated with open chromatin and enriched with stowaway transposons in potato. *Genome Biol.* **18**, 1–16.

Myrick, S., Pradel, W., Li, C., Suarez, V., Hareau, G., Larochelle, C., Norton, G.W. *et al*. (2021) The curious case of C-88: impacts of a potato variety on farmers in Yunnan, China. *CABI Agric. Biosci.* **2**, 3. https://doi.org/10.1186/s43170-020-00022-7

Naithani, S., Deng, C.H., Sahu, S.K. and Jaiswal, P. (2023) Exploring pan-genomes: an overview of resources and tools for unraveling structure, function, and evolution of crop genes and genomes. *Biomol. Ther.* **13**, 1403.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274.

Novak, A.M., Chung, D., Hickey, G., Djebali, S., Yokoyama, T.T., Garrison, E., Narzisi, G. *et al*. (2024) *Efficient indexing and querying of annotations in a pangenome graph. bioRxiv*. https://doi.org/10.1101/2024.10.12.618009

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S. and Phillippy, A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B. *et al*. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275.

Paradis, E. and Schliep, K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528.

Pham, G.M., Hamilton, J.P., Wood, J.C., Burke, J.T., Zhao, H., Vaillancourt, B., Ou, S. *et al*. (2020) Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience* **9**, giaa100. https://doi.org/10.1093/gigascience/giaa100

Porubsky, D., Ebert, P., Audano, P.A., Vollger, M.R., Harvey, W.T., Marijon, P., Ebler, J. *et al*. (2021) Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308.

Sheikhizadeh, S., Schranz, M.E., Akdel, M., de Ridder, D. and Smit, S. (2016) PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics* **32**, i487–i493.

Shirasawa, K., Esumi, T., Hirakawa, H., Tanaka, H., Itai, A., Ghelfi, A., Nagasaki, H. *et al*. (2019) Phased genome sequence of an interspecific hybrid flowering cherry, 'Somei-Yoshino' (Cerasus × yedoensis). *DNA Res.* **26**, 379–389.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D. *et al*. (2009) Polyploidy and angiosperm diversification. *Am. J. Bot.* **96**, 336–348.

Sun, X., Jiao, C., Schwaninger, H., Chao, C.T., Ma, Y., Duan, N., Khan, A. *et al*. (2020) Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* **52**, 1423–1432.

Sun, H., Jiao, W.-B., Krause, K., Campoy, J.A., Goel, M., Folz-Donahue, K., Kukat, C. *et al*. (2022) Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **54**, 342–348.

Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612.

Tang, D., Jia, Y., Zhang, J., Li, H., Cheng, L., Wang, P., Bao, Z. *et al*. (2022) Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541.

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P. *et al*. (2010) The genome of the domesticated apple (Malus × domestica Borkh.). *Nat. Genet.* **42**, 833–839.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H. *et al*. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J. *et al*. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Zhang, C., Wang, P., Tang, D., Yang, Z., Lu, F., Qi, J., Tawari, N.R. *et al*. (2019) The genetic basis of inbreeding depression in potato. *Nat. Genet.* **51**, 374–378.

Zhang, C., Scornavacca, C., Molloy, E.K. and Mirarab, S. (2020) ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* **37**, 3292–3307.

Zhang, Y., Yu, Z., Zheng, C. and Sankoff, D. (2021) Integrated synteny- and similarity-based inference on the polyploidization-fractionation cycle. *Interface Focus* **11**, 20200059. https://doi.org/10.1098/rsfs.2020.0059

Zhou, Q., Tang, D., Huang, W., Yang, Z., Zhang, Y., Hamilton, J.P., Visser, R.G.F. *et al*. (2020) Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* **52**, 1018–1023.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplement S1 Supplementary Figures.
Supplement S2 Supplementary Analyses.
Supplement S3 Supplementary Data.