Justus Liebig University,
Giessen Institute of Agronomy and Plant Breeding I
Department of Agrobioinformatics

Establishing pangenome graph as a framework for the analysis of the impact of structural variation on gene expression

Inaugural Dissertation for a Doctorate Degree in Natural Sciences (Dr. rer.nat.)
In the Faculty of Agricultural Sciences, Nutritional Sciences and Environmental Management

Examiners:

Prof. Dr. Agnieszka A. Golicz
Prof. Dr. Rod J. Snowdon

Submitted by

Gözde Yildiz
Giessen, 2025

"Success is not the key to happiness. Happiness is the key to success. If you love what you are doing, you will be successful."

Albert Schweitzer

Contents

1. Introduction

Feeding a growing global population in an increasingly changing and extreme climate is a major challenge (The State of Food Security and Nutrition in the World 2024). To meet this demand, scientists are working to improve crop yields while maintaining crop diversity and quality. One key approach is using genomics and metabolomics strategies to develop crops that can better withstand environmental stress (Ma et al. 2020; Purugganan and Jackson 2021). Next-generation sequencing (NGS) has played a major role in this effort, enabling researchers to explore the genetic potential of both wild and domesticated plants and providing insights into how crops have evolved and how they can be further improved. By integrating NGS in modern breeding approaches like marker-assisted breeding, quantitative trait locus (QTL) mapping, and the construction and use of reference genomes and pangenomes, researchers can develop improved crops more efficiently (Ashraf et al. 2022; Satam et al. 2023).

Pan-genomes comprise the entire set of genetic material in a species, including both coding and non-coding sequences, and describe various types of genomic variations within individuals. Genetic diversity within a species is fundamental to evolution, adaptation, and domestication, shaping key agronomic traits such as yield, stress tolerance, and disease resistance (Yuan et al. 2021; Zanini et al. 2022; Du et al. 2024; Zhang et al. 2024). Genomic variations range from single-nucleotide polymorphisms (SNPs) to large structural variants (SVs), which include inversions, deletions, duplications, translocations, insertions, copy number variations (CNVs), and presence/absence variations (PAVs), typically 50 base pairs (bp) or longer (Sedlazeck et al. 2018; Alkan et al. 2011; IGVF Consortium 2024). SVs vary widely in type and size, affecting a larger portion of the genome per nucleotide change, and can have a greater impact on gene expression and protein function than SNPs or short insertions and deletions (INDELs) (Zanini et al. 2022; Leonard et al. 2024; Yildiz et al. 2025).

Compared to mammalian genomes, crop genomes are more complex, especially in polyploid species such as wheat, oilseed rape, and cotton. This complexity arises from

factors like polyploidy and abundant repetitive content, which make genome assembly and downstream comparative analyses more challenging. Moreover, accurately detecting large SVs remains difficult for many sequencing technologies, requiring advanced algorithms and high-performance computing to address these challenges effectively. Plant genome sequencing began with *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000) and rice (Sasaki and International Rice Genome Sequencing Project 2005), using the bacterial artificial chromosome (BAC) approach combined with Sanger sequencing. These early efforts produced high-quality assemblies at prohibitive costs. Since then, Illumina-based short-read technology has enabled the sequencing of over 200 plant genomes by drastically reducing costs and enabling massive parallel sequencing while maintaining sequence accuracy. However, many of these assemblies suffer from low contiguity and are composed of fragmented scaffolds (Belser et al. 2018). While gene-rich regions are generally well assembled, areas rich in transposable elements remain particularly difficult to resolve with short reads.

Advances in sequencing technologies such as PacBio and Oxford Nanopore delivered high-quality assemblies and have enabled the detection of variations across multiple individuals within a species, greatly expanding our understanding of SV structure and prevalence (Jiao and Schneeberger 2017; Yuan et al. 2021; Hu et al. 2024). For instance, Illumina-based short-read sequencing typically identifies a smaller number of short SVs (Abel et al. 2020; Ebert et al. 2021; Collins et al. 2020), whereas long-read sequencing technologies have significantly improved SVs detection, identifying a larger number of SVs spanning hundreds to hundreds of thousands of base pairs (Ebert et al. 2021; Chaisson et al. 2019). Long-read sequencing and chromatin conformation capture (Hi-C) technologies provide significant advantages for SV characterization. Hi-C sequencing captures chromosomal interactions, allowing for the detection of large-scale SVs across entire chromosomes (Ho et al. 2020; Yuan et al. 2021). Initially, long-read sequencing faced challenges such as high error rates (5–15%), low throughput, and high costs, limiting its widespread applications (Yuan et al. 2017). However, with improved accuracy (>99%), reduced costs, and advancements in PacBio HiFi (15-30 kb) and Oxford Nanopore R10.3 reads (up to 100 kb), long-read sequencing has become essential

for haplotype-aware genome assembly, large SVs detection, improved mapping in repetitive regions, and generating pangenome references (Wenger et al. 2019; Sedlazeck et al. 2018; Coster et al. 2021; Espinosa et al. 2024). Oxford Nanopore's sequencing technology has improved with the transition to R10.4.1, now achieving up to 99% accuracy in reading individual DNA fragments (Bogaerts Bert et al. 2024). This improvement enables more accurate variant calling, consensus accuracy, and other analyses by integrating information from multiple reads. While higher raw read accuracy contributes to better results, further improvements can be achieved through increased genome coverage (Kim et al. 2024). Recently, gapless telomere-to-telomere (T2T) haplotype-resolved assemblies have been constructed for various crop species, including wheat, sorghum, and rapeseed (Liu et al. 2025; Wang et al. 2025a; Li et al. 2024; Li et al. 2023a). Gapless T2T assemblies have deepened our understanding of genome structure by revealing the regulatory roles of repetitive elements and fully resolving centromeres, telomeres, and structural variant regions, with broad implications for pangenomics, functional genomics, breeding, and genome editing (Garg et al. 2024).

1.1 Genome Structure of *Brassica napus*

Brassica napus (rapeseed/canola) is an allotetraploid genome (AACC, 2n = 38) and the second-largest oilseed crop worldwide, contributing 13%–16% of global vegetable oil production (https://www.ers.usda.gov/). It originated from the hybridization of *B. rapa* (AA, 2n:20) and *B. oleracea* (CC, 2n = 18) approximately 7,500 years ago (Chalhoub et al. 2014; Song et al. 2020). First used in Europe for lamp oil and soap, rapeseed became globally important after double-low varieties (low erucic acid, low glucosinolates) were introduced in the 1970s (Friedt and Snowdon 2010). Today, it provides high-value vegetable oil for food and biodiesel, while its meal serves as a protein-rich animal feed. It is also a key crop in cereal-based rotations (Mason and Snowdon 2016; Tan et al. 2024). In recent years, *B. napus* breeding has focused not only on high yield but also on seed quality, nutrient efficiency, disease resistance, and suitability for mechanized farming (Raza et al. 2021; Hu et al. 2022; Li et al. 2023b; Tan et al. 2024). Achieving these

multifaceted breeding goals require a deeper understanding of gene functions in the genome and the regulatory networks underlying agronomic traits.

B. napus was formed by the merger of subgenomes with distinct evolutionary histories. This hybridization event triggered various genomic and epigenetic changes, including altered gene expression, DNA methylation, and transposable element regulation (Rigal et al. 2016; Edger et al. 2017). Structural rearrangements occur due to homoeologous exchanges (HEs), where chromosome segments are swapped between subgenomes, contributing to genome diploidization (Xiong et al. 2011; Gabur et al. 2019). The first reference genome of a cultivar 'Darmor-bzh' was sequenced and annotated with 101,040 gene models (Chalhoub et al. 2014). The Cn subgenome (525.8 Mb) is larger than the An subgenome (314.2 Mb), consistent with their progenitor genomes: *B. oleracea*'s Co genome (~630 Mb, 85% covered) and *B. rapa*'s Ar genome (~530 Mb, 59% covered) (Chalhoub et al. 2014). To date, short-read-based re-sequencing has been performed on 52 (Schmutzer et al. 2015), 588 (Lu et al. 2019), and 991 (Wu et al. 2019) *B. napus* accessions. Several long-read-based assemblies have since been generated, including updates of Darmor-bzh (v8:(Bayer et al. 2017); v10:(Rousseau-Gueutin et al. 2020)), as well as Express617 (Lee et al. 2020), Tapidor3, and ZS11 (Song et al. 2020; Chen et al. 2021). Recently, a gap-free genome assembly of Xiang5A was completed using optical network terminal (ONT) ultra-long reads, PacBio high-fidelity reads, and Hi-C data (Li et al. 2023a), comprehensively reviewed by (Tan et al. 2024).

Comparative studies across diverse *B. napus* accessions have highlighted high genomic diversity and extensive HEs, including candidate genes and outlier regions associated with key crop traits (Schmutzer et al. 2015; Lu et al. 2019; Wu et al. 2019). The highly dynamic polyploid genome of *B. napus* contains extensive SNPs and SVs, including CNVs, TEs, and PAVs. For example, Darmor-bzh *B. napus* assembly contains 34.8% TEs, less than the 40% estimated from raw reads (Chalhoub et al. 2014). Another study showed that pre-vernalization expression of BnaA02.FLC and BnaA10.FLC varied significantly among eight accessions, associated with PAVs and CNVs (Song et al. 2020). However, SVs are more difficult to detect due to challenges in accurately aligning

reads from highly similar chromosomes (Lee et al. 2020). Both diploid progenitors, *B. rapa* and *B. oleracea*, have undergone multiple paleopolyploidization events, resulting in large-scale genomic rearrangements following divergence from a common ancestor (Parkin et al. 2005). SVs frequently arise after genome duplication, contributing to genome differentiation and gene loss (Schiessl et al. 2019).

One approach to studying how genomic variants and SVs shape phenotypes is performing expression quantitative trait locus (eQTL) analysis. eQTL is a genomic locus where genetic variants, such as SNPs or SVs, are statistically associated with variation in gene expression levels across individuals. These loci help identify regulatory elements that influence gene activity and contribute to phenotypic variation. Recent studies have primarily examined SNP-based expression quantitative trait loci (SNP-eQTLs), while SV-based expression quantitative trait loci (SV-eQTLs) studies in polyploids remain limited due to the difficulty of mapping short reads across multiple homologous regions (Zhang et al. 2024). As a result, many SVs and their contributions to genomic and phenotypic diversity remain undetected. A comprehensive genomic resource linking SVs to phenotypic variation is essential for understanding their functional significance at the species level.

1.2 Structural Variants (SVs) in Crop Genomes

Structural variants (SVs) such as insertions, deletions, inversions, translocations, copy number variations (CNVs), presence/absence variations (PAVs), and transposable elements (TEs) are common in crop genomes. Alongside single-nucleotide polymorphisms (SNPs) and epigenetic modifications, these SVs contribute to heritable phenotypic diversity both within and between species (Ho et al. 2020; Yuan et al. 2021). Understanding the influence of SVs on plant traits is essential for advancing crop improvement. Therefore, further research is needed to fully uncover the roles of SVs in shaping plant genomes and their impact on key agricultural traits.

Copy number variations refer to differences in the number of copies of DNA segments among individuals. Extensive research has shown that CNVs are widespread in plant genomes, making them one of the most studied SV types. CNVs are classified by their variation type, genomic location, and size. In terms of variation type, they include gain-of-copy (increased DNA segments that may enhance gene expression) and loss-of-copy (missing DNA segments that may reduce or eliminate gene function) (Moradi et al. 2022). Based on genomic location, CNVs are intragenic (within genes, affecting function) or intergenic (in non-coding regions, potentially altering gene regulation) (Luo et al. 2022). CNVs also vary in size from small-scale events spanning a few kilobases to large-scale rearrangements that may affect multiple genes (Pös et al. 2021). Overall, these variations can lead to abnormalities in gene structure and changes in gene expression.

Recent research has expanded CNV studies beyond simple gene copy number changes to include complex mechanisms such as genome rearrangements, transposon activity, and responses to environmental stress (Silaiyiman et al. 2025). For example, duplication of the ZmLOX5 gene has been shown to enhance insect resistance in maize and incorporating this CNV into high-yield but insect-susceptible varieties improves both pest resistance and abiotic stress tolerance (Yuan et al. 2024). In soybean, QTL mapping in a recombinant inbred line (RIL) population of 460 lines identified a locus associated with a trailing-growth-and-shoot-length QTL. This region contained a CNV involving increased copies of gibberellin 2-oxidase 8A/B genes, which was found to suppress these traits during soybean domestication (Wang et al. 2021). In *Brassica napus*, a study on resistance gene analogues (RGAs) across eight lines found that CNVs are more prevalent in gene clusters than in single genes. Of the 112 disease resistance genes linked to blackleg resistance quantitative trait loci (QTLs), 25 are affected by CNVs, providing valuable information for rapeseed breeding (Dolatabadian et al. 2022). Advances in sequencing technologies have significantly improved CNV detection, emphasizing their importance in plant growth, adaptation, and disease resistance. These discoveries highlight the broad potential of CNVs in crop improvement, particularly in selective breeding and genomic selection strategies.

Presence–absence variations (PAVs) are a form of genetic variation where certain genomic sequences are present in one genome but missing in another (Wang et al. 2023a). Considered an extreme form of CNV, PAVs are increasingly recognized for their important role in crop improvement. Pangenome and gene ontology studies in crops have shown that PAVs are often enriched in genes involved in abiotic stress response and disease resistance (Bayer et al. 2022; Wang et al. 2023a; Wang et al. 2025b). In rice, PAVs at the Se locus contribute to hybrid sterility (HS) between indica and japonica varieties, acting as a reproductive barrier. Understanding this mechanism could help overcome reproductive barriers in indica-japonica hybrid rice breeding (Wang et al. 2023a). In legumes, pangenome analysis revealed 8,990 mutually exclusive and 30,272 co-occurring gene PAVs across biological pathways, providing insights into the functional complementarity of genes. These finding offers a valuable resource for future research and breeding in bean (Wang et al. 2025b). In oilseed and wheat, using pangenome graphs to understand PAVs and genomic diversity informs both conventional breeding and modern genome editing strategies (Bayer et al. 2022; Zanini et al. 2022; Yildiz et al. 2022).

Transposable elements (TEs) are another major contributor to genetic diversity, first discovered in maize (McClintock 1965). TEs are mobile genetic elements capable of moving a genome, and they are classified into two main types based on their transposition mechanism: class I retrotransposons (RTs), which follow a "copy-and-paste" mechanism via an RNA intermediate, and class II DNA transposons, which use a "cut-and-paste" method without RNA involvement (Slotkin and Martienssen 2007; Wicker et al. 2007). Retrotransposons increase in copy number more rapidly, but both types contribute to genome expansion (Ma and Bennetzen 2004). TEs make up a large portion of polyploid plant genomes, accounting for approximately 80% of the hexaploid wheat genome (Cantu et al. 2010), 67% in tetraploid cotton (Wang et al. 2016), and 62% in oilseed rape (Chen et al. 2021). The composition of TE families varies among plant species: Copia-type LTR elements are more abundant than Gypsy elements in carrot (Wang et al. 2023b) and banana (Martin et al. 2025), while Gypsy elements dominate in wheat (Xie et al. 2023), sunflower (Ventimiglia et al. 2023), and cotton (Tian et al. 2025).

In the *B. napus* genome, Helitrons represent the most numerous TE superfamily, while LTR retrotransposons account for the highest proportion of genome (Yildiz et al. 2025; Xiao et al. 2025). Generally, Class I retrotransposons contribute more nucleotides than Class II DNA transposons. However, notable exceptions with high proportions of Class II elements include rice (17.6%), *Brachypodium distachyon* (23.8%), wheat (16.4%), and kiwifruit (14.9%) (Tao et al. 2025). Recent research focuses on TE-derived coding genes and non-coding RNAs, TE-derived cis-regulatory module variations, and epigenetic regulation of TEs in plants.

TE-derived-coding genes and non-coding RNAs regulate gene expression through trans-acting mechanisms (Tao et al. 2025). TE-derived small RNAs can regulate gene expression by guiding RNA-directed DNA methylation or through post-transcriptional gene silencing (Sun et al. 2023). Additionally, TE-derived long non-coding RNAs (lncRNAs) can function as microRNA sponges or interact with RNA-binding proteins (Li et al. 2022), playing roles in plant growth and development, disease resistance, and responses to abiotic stresses. The expansion and sequence variation of TEs serve as a significant source of cis-regulatory modules (CRMs) (Long et al. 2024; Xie et al. 2023). These CRM variations influence the binding of trans-acting factors, such as transcription factors and regulatory proteins, resulting in gene expression patterns that are tissue-specific, stress-responsive, cell type-dependent, or uneven across subgenomes (Schmitz et al. 2022; Zhang et al. 2022). In *B. napus*, epigenetic regulation of TEs ensures genome stability while allowing developmental and environmental reprogramming (Xiao et al. 2025). Genome-wide methylation studies have revealed TE methylation dynamics, and recent advances now enable exploration at single-cell and population scales (Zhao et al. 2024; Frost et al. 2024), opening new paths for TE-related epihaplotype discovery and epigenome engineering in many plants (Xue et al. 2025; Tao et al. 2025).

1.3 SV Detection and Representation in Crop Genomes

Detecting SVs in plant genomes remains challenging, due to their large sizes and complexity, especially in polyploid genomes, which make up nearly 80% of crops (Ho et

al. 2020; Yuan et al. 2021). Two primary approaches are commonly used for SVs detection: de novo genome assembly comparisons and read-mapping-based methods, which analyze paired reads (PR), read depth (RD), and split reads (SR) in short-read sequencing (Escaramís et al. 2015; Ho et al. 2020). In contrast, long-read sequencing facilitates the direct alignment of contiguous reads, enabling the comprehensive resolution of SV breakpoints via alignment-based, assembly-based, and graph-based approaches, which reduce mapping ambiguity. Long-read mapping-based SV callers outperform short-read methods by leveraging extended read lengths for improved alignment accuracy and complete allele resolution (**Fig. 1**). Since the release of the *Arabidopsis thaliana* genome in 2000 (The Arabidopsis Genome Initiative 2000), the number of assembled plant genomes has increased to approximately 450, as recorded in https://www.plabipd.de/. This growth has facilitated improved SVs identification, through whole-genome comparisons using tools such as Mauve (Darling et al. 2004), MUMMER (Kurtz et al. 2004), LASTZ (Harris 2007), Assemblytics (Nattestad and Schatz 2016), paftools (Li 2018), SyRI (Goel et al. 2019), and SVIM-asm (Heller and Vingron 2021). However, whole-genome comparisons are limited to detecting SVs, making read mapping the more widely used method for SV calling.

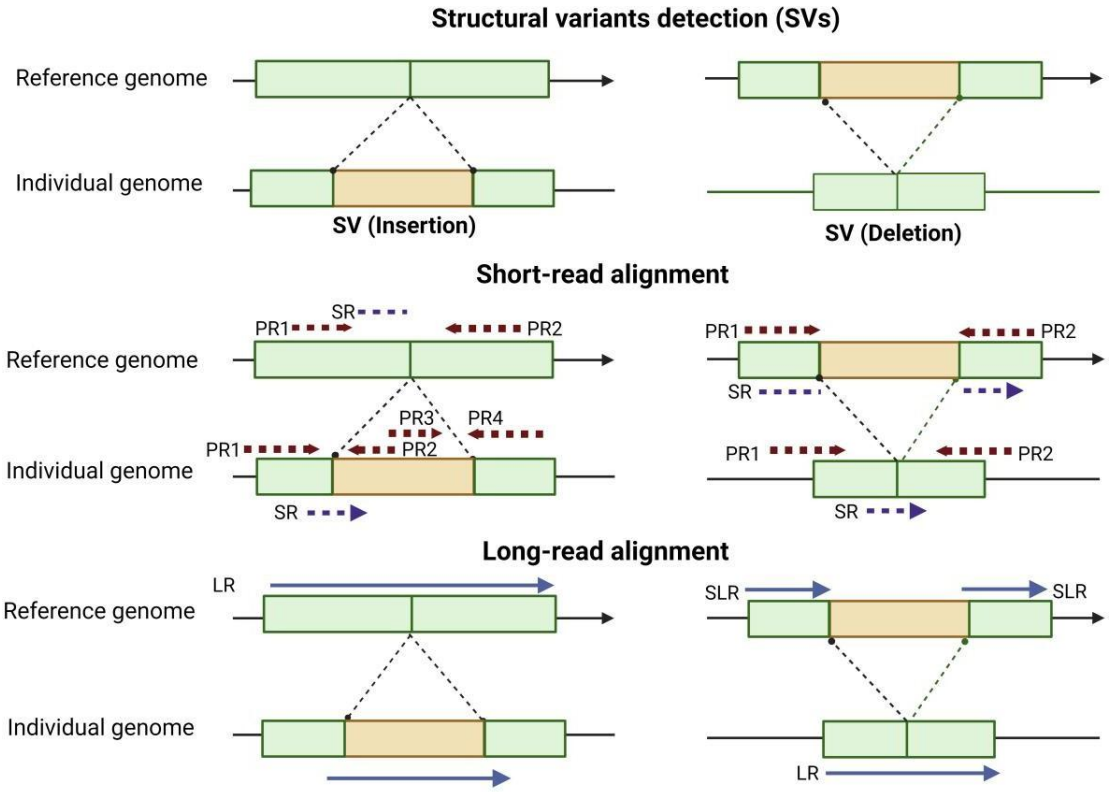


Fig. 1 Schematic illustration of two major classes of SV detection using short- and long-read alignments. An insertion (left) occurs when the individual genome contains an additional sequence segment (orange) absent from the reference genome, whereas a deletion (right) occurs when a sequence present in the reference genome is absent from the individual genome. In short-read alignments, paired-end reads (PR1-4; red arrows) and split reads (SR; blue arrows) provide evidence for insertions or deletions through discordant read-pair distances, unexpected orientations, and partially mapped reads. In long-read alignments, long reads (LR) or split long reads (SLR; purple arrows) typically span the entire variant and its flanking regions, enabling direct resolution of both insertions and deletions via contiguous alignments that reveal sequence gains or losses relative to the reference genome.

Comparing SV calls generated by de novo assembly and mapping-based approaches introduces a layer of complexity. For example, while genome alignment might reveal a novel sequence, mapping-based methods could interpret the same event as a tandem

duplication if the inserted sequence is similar to its nearby region. Altogether, these ambiguities make reconciling SV call sets from different methods particularly challenging (Mahmoud et al. 2019). Given the trade-offs between alignment- and assembly-based approaches, integrating both methods could improve SV detection accuracy and provide a more comprehensive view of genomic variation (Jiang et al. 2020). Misaligned or low-quality reads often lead to inaccurate SV size estimations and misplaced breakpoints, resulting in false negatives and misleading variant calls. The complexity increases when analyzing polyploid genomes. Although assembly-based approaches are ideally suited for detecting and resolving complex SVs, they require high sequencing coverage and integration of multiple data types (e.g., short reads, long reads, Hi-C, optical mapping), making them impractical for large-scale population studies (Mahmoud et al. 2019).

Over the past years, numerous of tools have been developed for SV calling from both short and long reads, including DELLY (Rausch et al. 2012), LUMPY (Layer et al. 2014), Manta (Chen et al. 2016), NanoSV (Cretu Stancu et al. 2017), Sniffles1& Sniffles2 (Sedlazeck et al. 2018; Smolka et al. 2024) and SVIM (Heller and Vingron 2019), cuteSV (Jiang et al. 2020), and dysgu (Cleal and Baird 2022), which have been comprehensively reviewed recently (Yuan et al. 2021; Liu et al. 2024). These tools only optimize diploid organisms and, when applied to polyploid crops, where multiple versions may exist and per-haplotype read support is lower, can lead to mis-genotyping or undetected SVs. Therefore, interpreting and comparing SVs remains challenging. In addition, generating new real datasets with specific features requires extensive and cost- and labor-demanding biological experiments and can be biased by experimental design.

In silico simulations are an inexpensive and unbiased alternative, and the available ground truth enables an accurate estimation of the precision and recall of SV calling methods. Some SV simulators have been developed, including RSVSim (Bartenhagen and Dugas 2013), SCNVSim (Qin et al. 2015), VarSim (Mu et al. 2015), BAMSurgeon (Ewing et al. 2015), SVEngine (Xia et al. 2018), and VISOR (Bolognini et al. 2020). In addition, benchmarking SV callers using ground truth datasets is essential to assess their

performance and correctness. Benchmarks show that short-read SV detection tools have limited sensitivity and precision, even when combined. While long reads offer better SV discovery, their high cost and error rates prevent them from fully replacing short reads. A promising approach is using short-read data to genotype SVs discovered from long reads, enabling population-scale studies. However, best practices for applying these methods in highly repetitive plant genomes, like soybeans, are still needed (Lemay et al. 2022).

1.4 Pangenome Graph: Trend Approach to Genomic Diversity

Reference genomes are essential for interpreting DNA sequences, but traditional linear references represent only a single version of each locus. This approach overlooks the full spectrum of genetic diversity within a population, leading to reference bias (Zanini et al. 2022; Garrison et al. 2024; Secomandi et al. 2025). This limitation affects the identification of SVs, which are key to agronomic traits (Lye and Purugganan 2019; Deng et al. 2017; Liu et al. 2020b; Alonge et al. 2020). To overcome this, pangenomes and variation graphs have emerged as powerful alternatives (Golicz et al. 2016; Tao et al. 2019). Variation graphs are bidirected sequence graphs that efficiently represent genetic diversity across populations, including large SVs like inversions and duplications (Paten et al. 2017) (**Fig. 2**). Unlike linear references, these graph-based models represent multiple alleles per locus, enabling more accurate and comprehensive genotyping (Garrison et al. 2018; Eggertsson et al. 2019; Ebler et al. 2022; Hickey et al. 2020; Chen et al. 2019b).

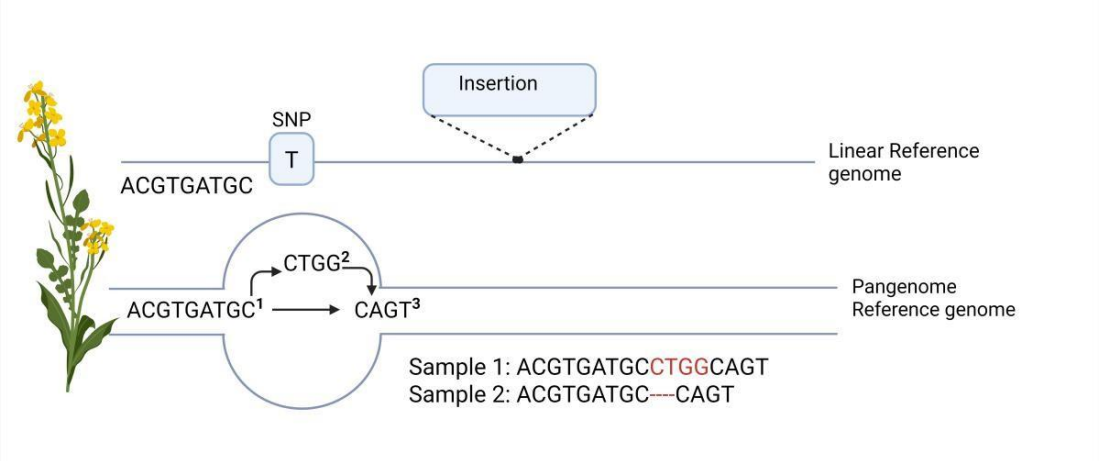


Fig. 2 In the linear reference genome, variants such as SNPs and insertions are represented along a single path, potentially limiting the representation of alternative haplotypes. In contrast, the pangenome reference genome (bottom) incorporates multiple sequence paths, enabling the representation of alternative alleles and SVs within a graph-based framework. For example: node 1 ($ACGTGATGC^{1}$) and node 3 ($CAGT^{3}$) contain the sequence, shared by all samples. Node 2 ($CTGG^{2}$) is not present in sample 2. Sample 1 follows the path 1 → 2 → 3, encoding the sequence ACGTGATGCCTGGCAGT. However, sample 2 include only the path 1→ 3. This pangenome-style graph structure captures both conserved and variable regions, allowing multiple genomes to be encoded within a single data structure among samples.

In small genomes, genetic variation has been studied by assembling entire genomes and comparing them directly (Delcher et al. 1999; Paten et al. 2017). However, in large genomes, complete and accurate de novo assembly can be difficult due to repetitive sequences and genome size. Therefore, the need for the assembly can be circumvented and the sequence data can be aligned directly to a high-quality reference genome. While this approach is faster than de novo assembly and simplifies variant identification, it introduces mapping bias, favoring reference-matching variants while missing alternative variants (Garrison et al. 2018; Garrison et al. 2024). To reduce mapping bias, an ideal approach would involve aligning data to a pangenome reference that includes the individual's variants (Yuan and Qin 2012). Since most genomic differences already exist in the population, such references offer a more inclusive representation. Graph-based models capture this diversity by encoding shared variants and alternative paths within a unified structure (Garrison et al. 2018). Mapping tools designed for pangenomic references help reduce bias when aligning DNA reads, improving accuracy in tasks like variant calling, downstream analyses, genotyping, and providing valuable advantages for RNA-seq data (Edwards and Batley 2022; Hickey et al. 2020; Sibbesen et al. 2018).

Pangenomes can be constructed from various sources, including: (1) existing linear reference genomes and their variants, (2) haplotype reference panels, and (3) raw reads

from either bulk or multi-sample sequencing. An effective data structure should support incremental updates, such as adding or removing genomes or variants, without needing complete reconstruction. In practice, moderate sequencing coverages (> 10x) for both short-and long-read sequencing can be sufficient for many pangenome applications. While recent pangenome graph tools (e.g., PGGB (Garrison et al. 2024), Minigraph-Cactus (Hickey et al. 2024), and Minigraph (Li et al. 2020)) are optimized for high-quality whole-genome assemblies, alignment-based approaches like VG enable SVs discovery even with medium-depth sequencing. These alignment-based SV graph methods are particularly valuable for studies where generating high-quality genome assemblies is not feasible in large-scale plant genomics projects. In such cases, researchers can leverage available sequencing data to identify SVs, construct graph-based pangenomes and then genotype SVs in a larger population SVs without the need for full de novo genome assembly, making the approach cost-prohibitive, scalable, and accessible.

Another current approach is pantranscriptomics, which uses reference transcriptomes from populations to enhance transcript analyses. Traditional RNA-seq mappers align reads over known splice junctions (Sibbesen et al. 2018; Wu et al. 2016). Considering population variations at splice site motifs have also been shown to improve the detection of new splice sites (Stein et al. 2015). Some existing tools integrate sequence graphs for this purpose. AERON and GraphAligner identify gene fusions using splicing graphs (Rautiainen et al. 2020; Rautiainen and Marschall 2020), while ASGAL detects novel splicing events using splicing graphs. Additionally, HISAT2, originally developed from the RNA-seq tool HISAT, enables pantranscriptomic mapping (Kim et al. 2019; Kim et al. 2015). Pantranscriptomic methods have been developed to analyze haplotype-specific gene expression using existing haplotype panels to estimate haplotype-specific gene expression (Lee et al. 2018; Aguiar et al. 2019). RPVG utilizes alignments from the vg mpmap to measure haplotype-specific transcript expression. The pantranscriptome reference incorporates population variation, allowing analysis without prior genome characterization (Sibbesen et al. 2023). The pan-transcriptomic approaches promise

improved gene expression quantification by removing reference bias in RNA-Seq read mapping.

1.5 Graph-based Structural Variant Genotyping

Once SVs are discovered in and a pangenome graph is constructed, the graph can then be used for SV genotyping in a population. Graph-based SV genotyper offers promising approaches for genotyping SVs using short-read sequencing techniques (Chen et al. 2019; Hickey et al. 2020). Short-read sequencing remains valuable for genotyping known SVs, especially in large-scale studies where high-throughput and cost-effectiveness are essential. Traditional SV genotyping tools use reads mapped to a reference genome to identify unusual mappings that might indicate the presence of SVs. For example, 68% of the maize genome consists of non-unique k-mers, compared to only 18% in humans (Du et al. 2024). This significantly reduces the number of unique, informative k-mers available for indexing and genotyping in tools. Beyond sequence repetitiveness, plant genome graphs are computationally demanding to construct and index. For instance, building a graph for just chromosome 10 from seven maize genomes using the vg map tool can require up to 90 GB of memory. To reduce high computational cost, informative k-mers strategies like BayesTyper and PanGenie are used, but they still have bottlenecks in highly repetitive crop genomes (Sibbesen et al. 2018; Ebler et al. 2022). BayesTyper relies on exact k-mer matches and is particularly sensitive to such errors. In contrast, PanGenie demonstrates better performance by incorporating haplotype-resolved pangenome references, enabling more robust genotyping in complex regions. Still, both tools require moderate to high sequencing depth for reliable results, which is often unavailable in early-stage plant studies where coverage may only reach 3–10×. Accurate SV genotyping at low coverage remains a critical challenge, particularly when distinguishing true variants from sequencing errors in heterozygous or repeat-rich regions.

Incorporating haplotype information from high-quality genome assemblies could improve genotyping accuracy under low coverage conditions. Furthermore, many graph-based

tools lack scalability. In contrast to k-mer-based methods, tools like vg giraffe use global alignment strategies and require more memory efficiency during graph construction by using minimizer-based seeding, which enhances speed and reduces resource use (Hickey et al. 2020). Vg stands out among graph-based tools for its consistently strong performance across multiple datasets. It supports simultaneous genotyping of overlapping SVs through a technique called snarl decomposition, which improves accuracy in regions with complex variation. Vg also tolerates small breakpoint inaccuracies—up to 10 bp—by fine-tuning them using read alignments. However, its performance can decline with larger breakpoint uncertainty, especially in low-depth or noisy datasets. A key strength of vg is its unified framework for jointly calling SVs, SNPs, and small indels, making it a flexible tool for comprehensive variant genotyping. With active community development and support for integration of long-read data, vg continues to evolve and is well-suited for future pangenome applications (Hickey et al. 2020).

Other tools like GraphTyper2 and Paragraph take localized approaches by refining variant calls using reads already aligned near breakpoints, which offers efficiency benefits (Eggertsson et al. 2019; Chen et al. 2019). Paragraph, specifically, constructs a small, local sequence graph for each targeted SV, rather than aligning reads to a large whole-genome graph. This design significantly reduces computational demands and makes Paragraph highly scalable for population-level studies. While whole-genome graphs can rescue misaligned reads from novel insertions, they require high computational resources and complete realignment when new variants are added. In contrast, Paragraph's localized method integrates efficiently with existing pipelines and enables rapid genotyping of large variant sets without the need to rebuild or realign entire datasets (Chen et al. 2019). Tool performance often declines while runtime and memory usage rise sharply, if variant numbers increase in population-level studies. None of the tools have been designed with polyploidy genomes in mind. Developing SV genotyping models specifically tailored to polyploid genomes would significantly improve accuracy in plant genomic research.

1.6 Understanding Expression Quantitative Trait Loci (eQTL)

Genetic variations influence gene expression, which in turn affects multiple phenotypes (Alonge et al. 2020). Expression quantitative trait locus (eQTL) studies help link genetic variations and transcriptomic datasets to gene activity, providing insights into how specific genes contribute to traits (Zhu et al. 2016). Gene expression differences can result from cis- or trans-regulatory changes (**Fig. 3**) (Wittkopp et al. 2004). Cis-regulatory variants are located near the gene they control and affect its transcription in an allele-specific manner (Wittkopp et al. 2004). These variants are often found in linkage studies as local eQTL (Rockman and Kruglyak 2006). In contrast, trans-regulatory variants affect the transcript abundance of both alleles of the target gene (Wittkopp et al. 2004). These are often referred to as distant eQTL (Rockman and Kruglyak 2006). Studies have shown that local eQTL generally have a stronger effect on gene expression than distant eQTL (Kliebenstein 2009; Cubillos et al. 2012; Wang et al. 2018). However, distant eQTLs are still valuable because they help identify trans-acting hotspots that contain key regulators that control the expression of many downstream genes involved in important biological processes (Wang et al. 2018).

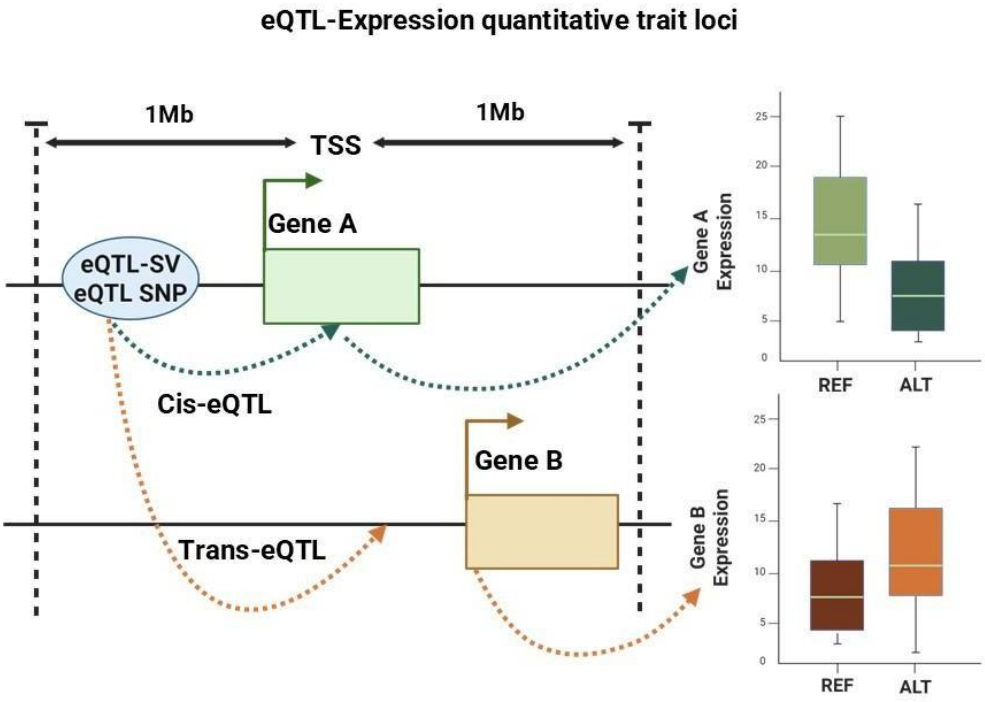


Fig. 3 Examples of SVs and SNPs can be associated with different gene expression levels. eQTL-SVs or eQTL-SNPs that map to approximately the same location as Gene A (represented by the green rectangle on the same chromosome) are referred to as cis-eQTLs. These are generally located within 1 Mb of the gene. In contrast, eQTL-SVs or eQTL-SNPs that are located far from the position of Gene B (represented by the orange rectangle, often on different chromosomes) are called trans-eQTLs.These two types of eQTLs are sometimes also called local (cis) and distal (trans) eQTLs. TSS stands for Transcription Start Site.

eQTL studies have been instrumental in identifying key genes and regulatory networks underlying plant traits. For example, in maize, eQTL mapping revealed that *Abscisic acid 8'-hydroxylase* is involved in drought tolerance (Liu et al. 2020a). In cotton, eQTL analysis identified *GhHRK1* as a regulator of heat stress response (Ma et al. 2021). Similarly, in peanut 1,207 local and 15,837 distant eQTLs were found to contribute to genome-wide transcriptomic variation, helping to uncover novel genes associated with purple seed coat pigmentation (Huang et al. 2020). In polyploid crops, the regulation of genes between subgenomes can be imbalanced (Zhuang et al. 2019; Cheng et al. 2018). In cotton, the Dn subgenome has stronger gene expression than the An subgenome, and differences in DNA methylation levels can affect how genes are expressed in each subgenome, leading to biased expression of similar gene pairs (Song and Chen 2015; Zheng et al. 2016). Similarly, in *B. napus*, the An subgenome harbors a higher number of SNP eQTLs than the Cn subgenome (An et al. 2019; Zhang et al. 2024; Tan et al. 2022). eQTL mapping is a powerful tool for understanding gene regulation, but identifying specific gene regulatory relationships from eQTL data is still challenging (Sullivan and Susztak 2020). One issue is that multiple variants can have significant association with some gene. Not all significant variants are regulatory, and finding the actual regulatory elements and variants among many significant ones is still challenging (Schaid et al. 2018). Combining eQTLs with novel machine learning approaches can aid in exploring and understanding how genes interact (Torlay et al. 2017). One case study used a large transcriptome dataset and XGBoost models to predict gene regulatory networks

(GRNs) and identify key transcription factors (TFs) in different GRNs by integrating eQTL information (Zhou et al. 2020).

Additionally, regulatory elements (REs) are more common in open chromatin regions (OCRs), and identifying these REs and significant variants helps pinpoint regulatory variants in eQTL studies (Tan et al. 2022; Fullard et al. 2017). Transposase-accessible chromatin sequencing (ATAC-seq) provides information on open chromatin patterns, and deep learning models can automatically analyze complex genomic data to predict regulatory variants in non-coding regions (Zhao et al. 2021b; Chen et al. 2019a; Zhao et al. 2021a). These advances have made it easier to identify gene regulatory relationships, which is important for understanding how genomic variations affect gene expression and contribute to traits. RNA sequencing (RNA-seq) is a crucial tool for studying cellular activity. However, if a sample's genome differs from the reference, bioinformatics tools must account for mismatches. Larger genomic variation reduces the accuracy of read alignment, making it more difficult to correctly identify transcripts and affecting the reliability of transcriptome analysis (Stevenson et al. 2013; Sibbesen et al. 2023).

1.7 Research Gaps and Study Objectives

1.7.1 Research Gaps

Despite significant advancements in genome analysis, there remain critical gaps in our understanding of SVs and their influence on gene expression, particularly in complex crop genomes. Single-reference genome studies fail to capture the full spectrum of genetic diversity within a species, limiting our ability to accurately characterize SVs and their phenotypic consequences. Pangenomic approaches offer a more comprehensive framework by incorporating multiple genome references and new tools and methods are continuously being developed. While pangenomes have been used to enhance our understanding of genetic variability, their integration with transcriptomic data for functional studies is still in its early stages. One major gap lies in the lack of standardized methodologies for SV discovery across different sequencing platforms at low sequencing

coverages. Oxford Nanopore Technology (ONT) has revolutionized long-read sequencing, yet there is no consensus on the optimal combination of read aligners and SV callers, particularly for polyploid crops. Additionally, current benchmarking studies are often focused on model organisms like humans, making it difficult to determine the best analytical pipelines for agronomically important species with complex genomes.

Another significant challenge is the limited exploration of how SVs, including insertions, deletions, and transposable elements (TEs), contribute to gene expression diversity. Most expression quantitative trait loci (eQTL) studies rely heavily on single-nucleotide polymorphisms (SNPs), potentially overlooking key regulatory variations caused by SVs. The role of pangenome graphs in integrating SVs into transcriptomic analyses remains underexplored, leaving a gap in our ability to link structural genomic changes to functional gene expression differences in crops.

1.7.2 Aims of This Study

The main goal of this thesis is to improve our understanding of structural variations (SVs) in *Brassica napus* by combining pangenomic and transcriptomic approaches. The study aims to:

i) Evaluate the role of pangenomics in crop genomics. This research examines how pangenomics can improve the discovery and understanding of genetic variations, compared to using a single reference genome. We review existing studies to show how pangenomics has been used to study important traits like disease resistance, plant shape, and yield on different crops (**Chapter 2**).

ii) Compare read aligners and SV detection tools: We believe that certain combinations of ONT read aligners and SV detection tools will perform better at finding insertions and deletions in complex crop genomes. To test this, we compare different tools using both real and simulated ONT data from *B. napus* (allotetraploid) and *Solanum lycopersicum* (diploid) at different sequencing depths (5×, 10×, and 20×). We also validate our results

using datasets from maize and soybean to develop a reliable framework for SV detection (**Chapter 3**).

iii) Investigate how SVs affect gene expression and improve eQTL analysis with pangenome graphs: This study combines long-read ONT sequencing and short-read Illumina data with mRNA-Seq from young leaves of *B. napus* to see how different SVs influence gene expression. We explore how pangenome graphs can be used for SV genotyping and to improve transcript expression quantification We develop a first plant pangenome graph-based framework for eQTL analysis. In addition, by comparing traditional SNP-based approaches with those that include SVs, we show how SVs should be considered in understanding their effects on crop traits (**Chapter 4**). By addressing these objectives, this study seeks to advance the application of pangenomics in plant science, providing valuable insights into the genetic and transcriptomic complexity of crop genomes. The findings will contribute to the development of improved genomic tools for plant breeding and crop improvement strategies.

2. Pangenomics in Agriculture

Gözde Yildiz, Silvia Zanini, Paul Knight and Agnieszka A. Golicz

CAB International 2022. Next-Generation

Sequencing and Agriculture

(eds P.E. Bayer and D. Edwards)

https://doi.org/10.1079/9781789247848.0008

8 Pangenomics in Agriculture

Gözde Yildiz, Silvia Zanini, Paul Knight and Agnieszka A. Golicz*

Department of Plant Breeding, IFZ Research Centre for Biosystems, Land Use and Nutrition, Justus Liebig University Gießen, Gießen, Germany

Abstract

Pangenomic approaches are increasingly being applied to plant science research to facilitate the discovery and representation of crop genetic variability. Compared with single reference genomes, pangenomes can represent the entire variation repertoire of a certain species or genus, enabling faster and more accurate characterization of structural variations (SVs) and their impact on phenotype. This chapter discusses existing pangenomic studies, highlighting both their features and application potential. From disease resistance to plant morphology and yield, pangenomics transforms our understanding of the genetic variation underlying key agronomical traits.

8.1 Introduction: Importance of Crop Plant Pangenomes

The pangenome concept was first introduced in the analysis of multiple pathogenic strains of *Streptococcus agalactiae* (Tettelin *et al*., 2005). A pangenome represents all of the genes/sequences found within a species and is composed of core and variable genes/sequences. The core genome is found in all individuals and is often enriched in genes performing critical cellular functions (Li *et al*., 2014). The variable genome (also known as accessory or dispensable) differs across individuals and contributes to the variability in phenotypes observed within species. Many variable genes in crops are found to be associated with agronomically important traits (Tao *et al*., 2019). Such genes play key roles in rapid evolutionary adaptation such as acquiring biotic and abiotic stress tolerance (Sutton *et al*., 2007; Knox *et al*., 2010; Cook *et al*., 2012; Maron *et al*., 2013) but also differences in flowering time (Nitcher *et al*., 2013; Würschum *et al*., 2015) and other traits like panicle erectness (Zhou *et al*., 2009; Studer *et al*., 2011) and fruit flavour, size and production (Alonge *et al*., 2020). Pangenomes help represent genetic variations from single-nucleotide polymorphisms (SNPs) to large structural variants (SVs),

*Corresponding author: agnieszka.golicz@agrar.uni-giessen.de

 Next-Generation Sequencing and Agriculture (eds P.E. Bayer and D. Edwards)
DOI: 10.1079/9781789247848.0008

including presence/absence variants (PAVs) and copy-number variants (CNVs). Such variants affect gene function and can alter crop productivity and propensity for adaption to climate change or stress conditions (Brozynska *et al*., 2016). Early crop pangenome studies are reported in maize (*Zea mays*) (Hirsch *et al*., 2014), soybean (*Glycine soja*) (Li *et al*., 2014), rice (*Oryza sativa*) (Yao *et al*., 2015), cabbage (*Brassica rapa*) (Lin *et al*., 2014) and *Brassica oleracea* (Golicz *et al*., 2016a). In recent years, pangenomic approaches have been widely adopted, facilitating agronomic trait improvement.

8.2 From Genome to Pangenome

Multiple genome-sequencing projects have provided insights into the complexity of crop plant genomes. Some of the earliest assembled crop genomes were that of rice (Sasaki and International Rice Genome Sequencing Project, 2005), grape (Jaillon *et al*., 2007), maize (Schnable *et al*., 2009), sorghum (Paterson *et al*., 2009), soybean (Schmutz *et al*., 2010) and potato (Xu *et al*., 2011b). Species- and genus-level diversity was initially explored through extensive resequencing studies where short-read genomic data are mapped to an assembled reference genome. Among others, such resequencing approaches have been utilized for wild and cultivated soybean (Lam *et al*., 2010) and rice genomes (Xu *et al*., 2011a). However, the use of a single reference genome means such studies have reduced capacity to identify complex SVs, including PAVs and CNVs. The decrease in the cost of Illumina sequencing and improvements in short-read assembly methods have allowed for the construction of draft assemblies for multiple individuals in a species which, together with the assembly of unmapped reads, has allowed for the identification of gene sequences representing the accessory genomic sequence (Hirsch *et al*., 2016; Springer *et al*., 2018; Li *et al*., 2019). In addition to these developments, the advent of long-read sequencing has allowed for the generation of more contiguous and complete assemblies, facilitating the elucidation of non-coding and regulatory variation roles in determining key agronomic traits (Alonge *et al*., 2020; Liu *et al*., 2020a; Song *et al*., 2020a).

Several methodologies have been applied to pangenome construction, including comparative *de novo* assembly, iterative mapping and assembly, and, finally, the map-to-pan approach. The comparative *de novo* assembly approach aims to assemble full-length genomes of all accessions to improve resolution of repetitive regions and CNVs. This approach can be affected by several technical limitations including relatively high cost of data generation, high computational resource requirements and artefactual variations in assembly and annotation, which can cause spurious PAV calls. Whole-genome comparisons to determine core and variable sequences are often based on gene-level analysis. These usually rely on orthologous gene clustering, which can result in mistakes for highly duplicated crop genome assembly (Golicz *et al*., 2020). However, recently the development of long-read sequencing technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore has contributed to the generation of accurate chromosome-level

assemblies, significantly simplifying *de novo* assembly workflows (Mascher *et al*., 2017; Belser *et al*., 2018).

The iterative mapping and assembly approach is based on sequential mapping reads from all individuals to the reference genome and updating the original reference with assembled unmapped reads, resulting in a new pangenome reference (Golicz *et al*., 2016a). Iterative mapping and assembly allows PAV calls at every gene locus without using orthologous gene clustering and is applicable for analysing PAVs across large population-based short-read data sets, but it lacks precise positioning of accessory sequences and will likely result in under-representation of repetitive sequences due to inaccuracies in mapping and assembly of reads representing repeats. The complementary advantages and limitations of *de novo* and iterative mapping approaches mean comprehensive pangenome studies of a species would ideally combine both strategies. Finally, the map-to-pan approach is based on generating and mapping several low-quality *de novo* assemblies to a pre-existing reference genome (Hu *et al*., 2017). An overview of crop pangenome studies using different approaches is presented by year in Table 8.1.

8.3 Selected Crop Plant Pangenomes

Pangenome studies have now been undertaken across a range of agronomically important species, giving us insights into their evolutionary history, extent of genomic diversity and the effect of PAVs on agronomic traits (Fig. 8.1).

8.3.1 Maize pangenomes

The maize pangenome/pan-transcriptome was constructed using RNA-sequencing (RNA-seq) data from 503 inbred maize accessions to discover the complete complement of protein-coding genes expressed in seedlings. RNA-seq reads were mapped to the reference genome and unmapped reads were used for the novel transcript assembly (Hirsch *et al*., 2014). A set of essential/core transcripts consisting of 14,968 annotated reference genes and 1425 assembled transcripts were expressed in every individual. A further 18,327 annotated reference genes and 7183 assembled transcripts, constituting the dispensable portion of the maize pan-transcriptome, were instead found only in some lines (Hirsch *et al*., 2014). Further maize pangenomic studies highlighted intraspecies variation in repeat and gene content. Four European flint accessions were sequenced to obtain high-quality maize reference genomes, generating ~220–320× coverage Illumina paired-end and mate-pair sequences (Haberer *et al*., 2020). These four *Z. mays* flint and two additional *Z. mays* dent lines were then used to detect PAVs occurring within their genic regions. Approximately 46,200 to 48,000 gene models were identified for each line, with core genes expressed at higher levels compared with variable genes. The study also found 15,000 high-quality full-length

Table 8.1. Overview of crop pangenome studies presented by year and from various perspectives.

Reference/year published	Methods	Species	No. of accessions	Pangenome size	Percentage of core genes/ gene clusters	Variant/type	Coding/non-coding	Traits studied using the pangenome
Li *et al*. (2014)	*De novo*	*Glycine soja* (soybean)	7	59,080 genes	49	CNV PAV SNP	Coding	Disease resistance Flowering time Oil content Height and lodging Yield
Hirsch *et al*. (2014)	*De novo* transcriptome	*Zea mays* (maize)	503	41,903 transcripts	39	SNP CNV	Coding	Flowering time
Lin *et al*. (2014)	*De novo*	*Brassica rapa* (cabbage)	3	41,858 genes	87	PAV	Coding	Flowering time Stress resistance Lignin formation
Schatz *et al*. (2014)	*De novo*	*Oryza sativa* (rice)	3	39,891 genes	92	SNP PAV	Coding	Disease resistance Yield
Yao *et al*. (2015)	*De novo* (metagenome assembly)	*O. sativa* (rice)	1,483	52,976 sequences (*indica*) 30,349 sequences (*japonica*)	8,991 genes (*indica*) 6,366 genes (*japonica*)	PAV	Coding	Disease resistance Stress resistance Grain width and size
Golicz *et al*. (2016b)	Iterative assembly	*Brassica oleracea* *Brassica macrocarpa* (cabbage)	10	61,379 genes	81	SNP PAV CNV	Coding	Disease resistance Flowering time Secondary metabolites

Continued

Table 8.1. Continued

Reference/year published	Methods	Species	No. of accessions	Pangenome size	Percentage of core genes/ gene clusters	Variant/type	Coding/non-coding	Traits studied using the pangenome
Pinosio *et al*. (2016)	Read mapping (no assembly)	*Populus nigra* *Populus deltoides* *Populus trichocarpa* (poplar)	7	497 Mb	81	CNV PAV SNP	Coding	Disease resistance Stress resistance
Gordon *et al*. (2017)	*De novo*	*Brachypodium distachyon* (stiff brome)	54	37,886 genes	55	PAV CNV	Coding	Disease resistance Flowering time
Zhou *et al*. (2017)	*De novo*	*Medicago truncatula* (barrel clover)	15	74,700 genes	33	CNV	Coding	Disease resistance
Montenegro *et al*. (2017)	Iterative assembly	*Triticum aestivum* (bread wheat)	19	139,747 genes	64	PAV	Coding	Disease resistance Stress resistance
Ou *et al*. (2018)	Iterative assembly	*Capsicum annuum* *Capsicum baccatum* *Capsicum chinense* *Capsicum frutescens* (pepper)	383	51,757 genes	56	PAV	Coding	Carotenoid and capsaicinoid biosynthetic pathways

Continued

Table 8.1. Continued

Reference/year published	Methods	Species	No. of accessions	Pangenome size	Percentage of core genes/ gene clusters	Variant/type	Coding/non-coding	Traits studied using the pangenome
Zhao *et al*. (2018)	Iterative assembly	*O. sativa* *Oryza rufipogon* (rice)	67	42,580 genes	62	SNP PAV CNV	Coding	Flowering time Stress tolerance Grain weight Tiller angle and plant height Hull colour
Wang *et al*. (2018)	Map-to-pan	*O. sativa* *Oryza glaberrima* (rice)	3,010	48,098 genes	54–62	SNP PAV	Coding	Flowering time Grain length and width Disease resistance
Hurgobin *et al*. (2018)	Iterative assembly	*Brassica napus* (cabbage)	53	94,013 genes	62	PAV SNP HEs	Coding	Disease resistance
Hübner *et al*. (2019)	Iterative assembly	*Helianthus annuus* (sunflower)	493	61,205 genes	73	SNP PAV	Coding	Disease resistance
Gao *et al*. (2019)	Iterative assembly	*Solanum lycopersicum* *Solanum cheesmaniae* *Solanum galapagense* (tomato)	725	40,369 genes	74	PAV Promoter PAV (substitution)	Coding Non-coding	Disease resistance Fruit flavour
Yu *et al*. (2019)	*De novo*	*Sesamum indicum* (sesame)	5	26,472 gene clusters	58	PAV	Coding	Disease resistance Biosynthetic pathways

Continued

Table 8.1. Continued

Reference/year published	Methods	Species	No. of accessions	Pangenome size	Percentage of core genes/ gene clusters	Variant/type	Coding/non-coding	Traits studied using the pangenome
Dolatabadian *et al*. (2020)	Iterative assembly	*B. napus*	50	73,270 genes	73	PAV SNP HEs	Coding	Disease resistance
Song *et al*. (2020b)	*De novo*	*B. napus* (oilseed rape)	9	105,672 gene clusters	56	SNP PAV	Non-coding	Silique length Seed weight Flowering time
Zhao *et al*. (2020)	Iterative assembly	*Cajanus cajan* (pigeon pea)	89	55,512 genes	86	SNP PAV	Coding	Self-fertilization Disease resistance Seed weight
Liu *et al*. (2020b)	*De novo*, graph	*Glycine max* (soybean)	29	57,492 gene clusters	50	PAV	Coding Non-coding	Iron uptake
Jayakodi *et al*. (2020)	*De novo*	*Hordeum vulgare* (barley)	20	40,176 gene clusters	54	PAV Inversions	Coding	Yield
Sun *et al*. (2020)	*De novo*	*Malus domestica* *Malus sieversii* *Malus sylvestris* (apple)	91	69,411 gene clusters	81–87	PAV TE insertion	Coding	Fruit quality
Trouern-Trend *et al*. (2020)	*De novo*	*Juglans* spp. (walnut)	6	26,458 gene clusters	55	PAV CNV	Coding	Disease resistance Biosynthetic pathways
Haberer *et al*. (2020)	*De novo*	*Z. mays* (maize)	6	~46,200–48,000 gene clusters	94–98	PAV	Coding	Biosynthesis pathways

Continued

Table 8.1. Continued

Reference/year published	Methods	Species	No. of accessions	Pangenome size	Percentage of core genes/ gene clusters	Variant/type	Coding/non-coding	Traits studied using the pangenome
Walkowiak *et al*. (2020)	*De novo*	*T. aestivum* (bread wheat)	10	NA	NA	SNP Indels PAV CNV	Coding Non-coding	Stress resistance Yield Grain quality Disease resistance
Ruperao *et al*. (2021)	Iterative assembly	*Sorghum bicolor* (sorghum)	177	35,719 genes	47	PAV CNV	Coding	Drought resistance
Bayer *et al*. (2021)	Iterative assembly	*B. oleracea* *B. rapa* *B. napus*	87 77 79	6,715 genes 19,767 genes 5,060 genes	79 67 74	PAV	Coding Non-coding	Disease resistance Stress resistance
Li *et al*. (2021)	Iterative assembly	*Gossypium hirsutum* *Gossypium barbadense* (cotton)	1,581	102,768 genes 80,148 genes	78.4 85.8	PAV SNP CNV	Coding	Fibre development Flowering time Yield

NA, not applicable; CNV, copy-number variant; PAV, presence/absence variant; SNP, single-nucleotide polymorphism; HE, homoeologous exchange; TE, transposable element; indel, insertion/deletion.

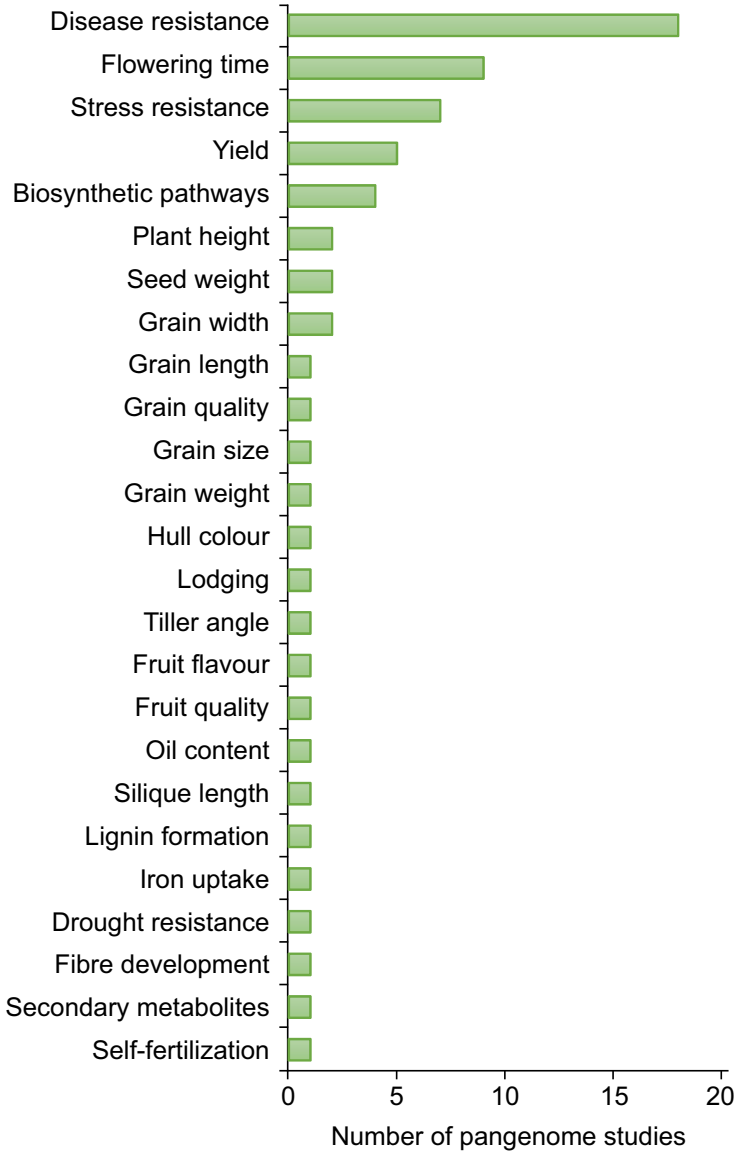


Fig. 8.1. Agronomic traits investigated in pangenome studies.

orthologous long terminal repeat (fl-LTR) retrotransposons per line, confirming their contribution to genomic variation. Only 3% of all fl-LTR locations were found to be shared between six lines, in stark contrast to genes where 76.1% retained syntenic positions (Haberer *et al.*, 2020).

8.3.2 Rice pangenomes

Genetic variability in rice was explored by constructing a variable rice genome. A metagenome-like *de novo* assembly was used to map and compare with the cv. Nipponbare reference genome data from 1483 cultivated rice accessions (*O. sativa* L.). More than 8000 dispensable protein-coding genes were predicted, including 3033 genes supported by evidence of expression and high confidence homology. The variable sequences of different rice accessions

included 0.46 million polymorphisms. Approximately 23.5% of metabolic traits under investigation had significant associations with polymorphism from variable sequences, more so than SNPs from the reference sequence (Yao *et al*., 2015). RPAN is a rice pangenome browser making use of ~3000 Rice Genomes Project resources. A total of 23,914 core genes (present in 100% of accessions), 4986 candidate core genes (present in >99% of accessions) and 22,095 distributed genes (present in <99% of accessions) were detected in 3010 accessions after mapping sequencing data to the IRGSP-1.0 genome. Nearly 12,000 novel genes were found absent in the reference genome (Sun *et al*., 2017). Complementary analysis of the population structure and diversity among 3010 Asian cultivated rice genomes identified a total of 93,683 SVs (582 SVs larger than 500 kb), with an average of 12,178 SVs per genome. The resulting *O. sativa* pangenome included 12,770 core gene families and ~9050 dispensable gene families (Wang *et al*., 2018). In a cultivated and wild rice pangenome study, 66 diverse accessions belonging to *O. sativa* and *Oryza rufipogon* species were deep sequenced and *de novo* assembled. A total of 10,872 novel genes identified across these rice accessions were absent from the cv. Nipponbare reference, while the core genome contained ~77.6% of all predicted coding genes (Zhao *et al*., 2018). To generate reference-grade assemblies for 15 rice subpopulations, 12 *O. sativa* accessions were sequenced using the PacBio platform for long-reads and the Illumina platform for short reads, together with Bionano optical maps construction, and combined with four existing reference sequences that represent the three remaining rice subpopulations. The resulting analysis identified several thousand SVs in all pairwise genome comparisons, with greater variability between accessions representing major groups (Zhou *et al*., 2020).

8.3.3 Wheat pangenomes

The first wheat pangenome study explored genome diversity in current modern elite cultivars. Genome sequences of 18 wheat cultivars were mapped to the Chinese Spring assembly and unmapped reads were assembled as new contigs. The pangenome resulted in 128,656 genes, with 89,795 being core (found in all 19 accessions) and 49,952 genes being variable (missing in at least one accession). Based on PAVs identified, the authors estimated a total pangenome size of 140,500 ± 102 genes, a core genome of 81,070 ± 1631 genes. Interestingly, 12,150 genes were not found in the Chinese Spring reference sequence but were present in all the other accessions. The annotated variable genes were associated with important agronomic traits and enriched in several critical biological processes, including response to environmental stress, defence response to biotic stress and abiotic stress tolerance (Montenegro *et al*., 2017). A later study confirmed the extensive variability among several wheat lines commonly utilized in breeding programmes (Walkowiak *et al*., 2020). By generating and comparing 15 reference-quality hexaploid wheat assemblies, the authors highlighted several SV classes, including 26% of predicted genes found in tandem duplications (CNVs).

Furthermore, *de novo* annotation of loci containing NLR motifs, often corresponding to disease resistance genes, identified almost 8000 unique NLR signatures. Among these, only 31–34% were shared across all genomes, confirming the need for multiple high-quality genome assemblies to successfully capture intraspecific genetic variations (Walkowiak *et al*., 2020).

8.3.4 Barley pangenome

To accurately represent barley's species-wide heterogeneity, chromosome-scale assemblies were constructed for 20 distinct *Hordeum vulgare* accessions, including landraces, elite cultivars and one wild ecotype (Jayakodi *et al*., 2020). The assemblies were then used to detect a total of 1.5 million PAVs, including several affecting resistance gene homologues. Due to the high proportion of repetitive sequences, a so-called 'single-copy pangenome' was generated by clustering only single-copy regions from each of the 20 assemblies, removing on average 90% of the original sequences. The resulting single-copy pangenome was 638.6 Mb, with 402 Mb core and 235.9 Mb variable, a significant reduction in size and complexity compared with barley haploid genome assembly (5 Gb).

8.3.5 Cabbage pangenomes

The genetic diversity in cabbage (*B. rapa*) morphotypes was determined by comparing the genomes of a turnip, a rapid cycling and Chiifu, a Chinese cabbage variety. The genomes of turnip and rapid cycling were sequenced, assembled and annotated, while the Chiifu reference sequence was re-annotated with 41,052 gene models, 11,715 of which were novel predictions. The three accessions shared 38,186 gene models, while 1090, 1118 and 1464 genes were unique to turnip, rapid cycling and Chiifu, respectively (Lin *et al*., 2014). The *B. oleracea* pangenome was constructed by sequencing eight cultivated accessions and one wild genotype (*Brassica macrocarpa*) and using an available rapid cycling line genome with iterative mapping and assembly approach. The assembled pangenome covered 587 Mb and a total of 61,379 genes. The core genome constituted the majority of the pangenome, representing 49,895 genes (81.3%), whereas 11,484 genes (18.7%) were variable, with 1322 (2.2%) being present only in one line. A total of 188 variable and 251 core genes were identified as resistance genes. Variable genes were also enriched for functions predicted to be involved in stress responses, water homeostasis, amino acid phosphorylation and signal transduction (Golicz *et al*., 2016b). The *Brassica napus* pangenome was built using 20 synthetic and 33 non-synthetic lines using iterative mapping and assembly, resulting in 1044 Mb (94,013 genes) compared with the Darmor-*bzh* v8.1 assembly of 850 Mb (80,382 genes). The pangenome included 58,532 core (62%) and 35,481 variable (38%) genes. In addition, extensive variations were demonstrated: an average of 22 uniquely present and 435 uniquely absent genes were found in the synthetic accessions, while two uniquely present and

49 uniquely absent genes were found in the non-synthetic accessions. A total of 94 *R* genes were identified in the core genome and 213 in the variable, with 47.6% of total *R* genes located in the new assembly contigs (Hurgobin *et al*., 2018). A later study identified additional disease resistance genes among the 50 *B. napus* accessions, resulting in a total of 1749 resistance genes analysed, 996 of which were core, 753 variable and 368 absent from the reference. Additionally, SNPs within the *R* genes were characterized, with 10,584 SNPs found within 731 core genes and 4734 SNPs within 299 variable genes (Dolatabadian *et al*., 2020). Finally, a recent study expanded on the previous knowledge regarding *Brassica* species' pangenomics by generating and comparing new pangenomes for both *B. napus* and its two diploid progenitors, *B. rapa* and *B. oleracea* (Bayer *et al*., 2021). These assemblies resulted in 58,315 gene models in *B. oleracea*, 59,864 in *B. rapa* and 108,580 in *B. napus*, a slightly different estimate compared with the aforementioned studies due to differences in annotation methods and the previous inclusion of wild relatives and/or synthetic lines. In addition to confirming the association of dispensable genes with biotic and abiotic stress responses in all three pangenomes, the authors investigated the underlying causes of gene PAVs. Interestingly, modelling of gene loss propensity in the three species highlighted differences between diploids and polyploids, with proximity to transposable elements (TEs) being most strongly linked with gene dispensability in *B. oleracea* and *B. rapa*, compared with homoeologous exchanges (HEs) in *B. napa* (Hurgobin *et al*., 2018; Bayer *et al*., 2021).

8.3.6 Tomato pangenomes

Seven hundred and twenty-five phylogenetically and geographically representative genotypes were used to construct the modern tomato pangenomes resulting in 1179 Mb of sequence and 40,369 protein-coding genes (4873 absent from the reference genome). A total of 586 accessions were used for PAV analysis, identifying 29,938 core (shared among all), 3232 soft core (present in ≥99% of accessions), 5912 shell (present in 1–99% of accessions) and 1287 cloud genes (present in ≤1% accessions) (Gao *et al*., 2019). To better understand the impact of SVs on quantitative traits, 100 diverse tomato lines were sequenced using long-read Nanopore technology (~40× genome coverage). Available short-read sequencing data from over 800 tomato accessions were used to select genotypes representing maximum SV diversity. A total of 238,490 SVs (>30 bp) were identified among the long-read sequencing of 100 tomato genomes. Approximately 50% of these SVs affecting coding and *cis*-regulatory regions were significantly associated with differences in expression. Also, a nearly 50 kbp tandem duplication at the *fw3.2* locus, containing three genes including two identical copies of the cytochrome P450 gene *SlKLUH*, was associated with both fruit weight and size traits (Alonge *et al*., 2020). A recent study focused on the impact of TEs on 602 cultivated and wild tomato accessions, investigating 6906 TE insertion polymorphisms (TIPs) resulting from the movement of 337 distinct TE families. Most of the TIPs were low-frequency variants and were disproportionately found close to genes involved in environmental responses.

Additionally, long-read Nanopore transcriptomics-based analyses discovered several genic TE insertions resulting in transcriptional changes and production of new isoforms (Domínguez *et al*., 2020).

8.3.7 Pepper pangenomes

A pepper pangenome browser, called PepperPan, was released in 2018 containing data for 383 cultivated peppers belonging to *Capsicum annuum*, *Capsicum baccatum*, *Capsicum chinense* and *Capsicum frutescens* species. The pangenome consisted of 89,181 total genes, with 51,757 high-quality genes and 37,424 low-quality genes depending on their Annotation Edit Distance (AED) metric and the relationship with TEs. Among these four species, 55.7% of high-quality genes were core, while 8.9% were variable. Genetic variations were detected in multiple essential genes related to critical agronomic traits, including the biosynthetic pathways of capsaicinoids and carotenoids. A 2.5 kb deletion in the Pungent gene 1 (*Pun1*) region was observed in several pepper cultivars with known low capsaicin content (Ou *et al*., 2018).

8.3.8 Soybean pangenomes

The first crop pangenome was reported for the wild soybean, *G. soja*, and the cultivated soybean, *Glycine max* (Li *et al*., 2014). The genetic variability among wild and cultivated soybean accessions was investigated by comparing *de novo* assembled genomes of seven wild *G. soja* accessions with the domesticated *G. max* reference genome. Most of the pangenome sequence (80%) was identified as core, corresponding to 28,716 gene families in the orthologous gene clustering-based comparisons (Li *et al*., 2014). A later study compared 26 representative wild and cultivated soybean genomes (Liu *et al*., 2020b). The pangenome was built using *de novo* assembled genomes and short-read sequencing data from 2898 individuals were analysed to detect the SVs using a pangenome graph as a reference. RNA-seq data from the initial 26 accessions supported the association between the structural variation and gene expression changes. A total of 27,175 genes from the newly sequenced individuals were absent from the ZH13 reference genome, and 48,249 genes were not identified in at least one of the 26 *de novo* assemblies. Overall, 723,862 PAVs, 27,531 CNVs, 21,886 translocations and 3120 inversion events were identified, with more than 90% of the size variation detected in the assembled genomes resulting from only PAVs (Liu *et al*., 2020b).

8.3.9 Sunflower pangenome

Whole-genome sequences of 287 cultivated lines, 17 Native American landraces and 189 wild accessions were used to characterize genetic diversity in

sunflower and determine its pangenome. The cultivated sunflower pangenome contained 61,205 genes and a total of 45,302 high-confidence genes were used to examine PAVs across cultivated lines. The core genome was represented by 32,917 (72.7%) genes, which appeared in >95% of the accessions, while 2464 (5.4%) dispensable genes were found in <5% of the accessions. Gene ontology (GO) analysis highlighted 25 biological processes significantly enriched, including terms related to biotic stress response, such as response to biotic stimulus and chitinase activity (Hübner *et al*., 2019).

8.3.10 Sesame pangenome

Sesame (*Sesamum indicum* L.) pangenome was constructed with two landraces and three modern cultivars. The pangenome (554.05 Mb) contained 26,742 orthologous gene clusters, with 15,409 core (58.2%) and 15,890 variable genes (41.8%). Gene number variations in the five sesame accessions provide insights into potential links between PAVs and phenotypic diversity. The comparative evolutionary analysis identified several examples of environmental adaptation and selection for high seed oil content, including plant–pathogen interaction and lipid metabolism (Yu *et al*., 2019).

8.3.11 Cotton pangenome

To characterize the tetraploid cotton (*Gossypium hirsutum* and *Gossypium barbadense*) pangenome, 1961 accessions were resequenced at an average depth of ~14.8× (Li *et al*., 2021). A subset of 742 accessions with depth >10× was used to identify 32,099 deletions, 7576 duplications, 1112 inversions, 357 translocations and 173,166 CNVs. The pangenome was built using an iterative mapping and assembly approach. Sequencing data from 1581 *G. hirsutum* and 226 *G. barbadense* accessions were aligned to the respective reference genome. The unmapped reads were assembled, producing 704 Mb and 1422 Mb non-redundant, non-reference sequences encoding 32,569 *G. hirsutum* genes and 8851 *G. barbadense* genes. The total pangenome sizes were 102,768 genes (63,489 core genes) for *G. hirsutum* and 80,148 genes (68,789 core genes) for *G. barbadense*. Core genes were reported to be involved in cellular metabolic processes and development, whereas the variable genes were involved in defence, stress responses and signal transduction. Further analysis identified 124 PAVs which overlapped with 89 quantitative trait loci, including ones associated with yield and fibre quality.

8.3.12 Apple pangenome

The apple pangenome was built using three phased *de novo* assemblies as a backbone representing cultivated apple (*Malus domestica* cv. Gala) and its two major wild progenitors, *Malus sieversii* and *Malus sylvestris* (Sun *et al*.,

2020). The assemblies were constructed using a combination of Illumina short reads and long, low-error, PacBio High Fidelity (HiFi) reads. Apple is known to be highly heterozygous, and the assemblies were twice the size of the haploid genome, suggesting that the two haplotypes were successfully resolved. Further 91 accessions were sequenced using Illumina technology at ~60× coverage. The resequencing data were used for individual *de novo* assemblies, which then were utilized to construct the pangenome. A total of 89, 212 and 141 Mb of new sequence containing 1736, 3438 and 2104 genes were identified for *M. sylvestris*, *M. sieversii* and *M. domestica*, respectively. The core genome size ranged from 81 to 87% for all three species. The new genes were enriched in functions related to pollination, signal transduction and response to stress.

8.3.13 Poplar pangenomes

Seven genome assemblies from *Populus nigra*, *Populus deltoides* and *Populus trichocarpa* individuals were used to identify SVs and build a poplar pangenome. The poplar pangenome was 497 Mb in size, with 401 Mb (81%) of sequences identified in all individuals and 96 Mb (19%) represented in some accessions but not others. In total, 7889 deletions (33.2 Mb), 10,586 insertions (62.9 Mb) and 3230 genes affected by CNV (relative to the *P. trichocarpa* reference) were detected. The insertions and deletions were mostly associated with TE activity, while genes affected by CNV showed enrichment in biological processes related to resistance to stresses and pathogens (Pinosio *et al*., 2016). Further poplar pangenome studies provided an understanding of evolutionary history among the *Populus* genus (Zhang *et al*., 2019). The genomes of ten species from five sections of the genus *Populus* were sequenced and highlighted 71 million genetic variations (SNPs and SVs). Among 200,501 SVs, deletion was more common than insertion and other kinds of SVs. *R* genes of these species, with heterozygous loss of function (LOF), significantly affected the diversity during poplar evolution. LOF mutations in the self-incompatibility genes were associated with genomic control of self-fertilization (Zhang *et al*., 2019).

8.4 Common and Distinct Features of Plant Pangenomes

The multiple pangenome studies discussed above allow us to observe some common themes, for example the association of TEs with the variable genome. TEs, also known as jumping genes, commonly contribute to genome size variation and dynamic regions between plant species' genomes. TEs are associated with extensive variations in both intergenic and genic regions in several crop genomes. They contribute to intra- and interspecies variation and may play an essential functional role for the variable genome (Lisch, 2013). TEs include class I (LTR retrotransposons) and class II (DNA transposons of different superfamilies) and can contribute to dramatic differences

in local sequence content among individuals belonging to the same species (Morgante *et al*., 2007). TE activity can mediate large-scale chromosomal rearrangements, such as seen for the *Ac/Ds* TEs (Yu *et al*., 2011) and *Helitron* transposons in maize (Lai *et al*., 2005) and the mutator-like elements (MULEs) in the rice genome (Hanada *et al*., 2009). An enrichment of TEs in the variable genome was reported in *Brachypodium distachyon* (Gordon *et al*., 2017) and *B. oleracea* (Golicz *et al*., 2016b). A higher proportion of the haT superfamily transposons are reported to reside in adjacent variable genes. The haT superfamily transposons, which include the *Ac/Ds* TE system, are suggested to mediate the formation of SVs due to alternative transposition (Golicz *et al*., 2016b; Pinosio *et al*., 2016). Besides TEs, recombination of non-allelic homologous sequences affects the formation of SVs. Nucleotide-binding site (NBS) or leucine-rich repeat (LRR) domains present in resistance genes of plants together with repetitive elements facilitate recombination among non-allelic homologous sequences (Yao *et al*., 2015; Golicz *et al*., 2016b).

8.5 Variable Genomes and Agronomic Traits

Key biological processes in plants such as development and reproduction are mainly regulated by core genes. However, dispensable genes constitute a significant portion of crop pangenomes (~5–50%) and are enriched in functions related to signalling response to environmental stress and defence against biotic agents (Li *et al*., 2014; Yao *et al*., 2015). The most critical SVs are CNVs and PAVs, which drive the majority of gene content and phenotypic diversity (Saxena *et al*., 2014). Understanding the role of SVs in the control of agricultural traits will have an impact on crop breeding. Multiple examples exist of the contribution of the variable genome to the control of agronomic traits and are reported in this section.

8.5.1 Flowering time

Flowering time differences in maize were associated with a miniature inverted-repeat TE (MITE) insertion and a 2 bp CGindel587 variation found in the *Vegetative to generative transition 1* (*Vgt1*) chromosome region (Salvi *et al*., 2007; Ducrocq *et al*., 2008). Additionally, insertion of a CACTA-like TE in the *ZmCCT10* promoter and Harbinger-like TE insertion at *ZmCCT9* gene supported early flowering at higher latitudes and photoperiod sensitivity due to acting as a *cis*-regulatory element (Yang *et al*., 2013; Huang *et al*., 2018). Similarly, in tomato, a *Rider* Ty1-*copia*-like retrotransposon insertion 27 bp downstream of the stop codon in *terminating flower* gene caused early flowering and the switch from multi-flowered inflorescence to single flowers (MacAlister *et al*., 2012). In barley, the increased copy number of *HvFT1* is associated with the dominant spring growth habit (Nitcher *et al*., 2013). In *B. oleracea*, 14 variable genes were predicted to regulate flowering time affected by PAVs (Golicz *et al*., 2016b). Finally, in a study of wheat flowering times,

higher CNVs at *Photoperiod-B1* (*Ppd-B1*) and the *Vernalization-A1* (*Vrn-A1*) were linked to divergence of photosensitivity and flowering time (Würschum *et al*., 2015).

8.5.2 Tolerance to biotic stress

Many studies in crop genomes have focused on the discovery and characterization of SVs which are known to play a significant role in a plant defence mechanism against biotic stress. In potato, presence of a resistance gene to late blight (*R1*) and an elicitin response (*ELR*) gene is associated with enhanced disease resistance to potato blight (Ballvora *et al*., 2002; Du *et al*., 2015). In wheat, leaf rust resistance locus (*Lr10*) encodes a single-copy gene on chromosome 1AS providing enhanced resistance to fungal pathogen *Puccinia triticina* and is absent in susceptible individuals (Feuillet *et al*., 2003). Likewise, high-temperature stripe rust resistance gene *Yr36* is present in wild wheat, resulting in resistance to stripe rust, while it is absent in modern pasta and bread wheat varieties (Fu *et al*., 2009). In rice, a blast-resistant genotype (Tsuyuake) carries the rice blast resistance gene *Pikm* which is absent in the susceptible genotype (Nipponbare) (Ashikawa *et al*., 2008). In soybean, more copies of the 31 kb repeat (CNV) at *Rhg1* mediate resistance to the cyst nematode (*Heterodera glycines*) (Cook *et al*., 2012). In maize, a major quantitative resistance locus (*qHSR1*) to head smut pathogen carries the *ZmWAK* gene which is absent in susceptible lines (Zuo *et al*., 2015). In soybean, the 11 lineage-specific *R* gene domain architectures were found in the dispensable portion of *G. soja* pangenome, possibly reflecting adaptation to biotic stresses (Li *et al*., 2014).

8.5.3 Tolerance to abiotic stress

Tolerance to abiotic stress such as extreme weather and environmental conditions, mineral deficiency, a multidrug and toxic compound has been associated with SVs in many crop species. In rice, PAVs of the *Submergence 1* (*Sub1A*) gene regulate submergence tolerance (Xu *et al*., 2006) and the presence of *phosphorus-starvation tolerance 1* (*PSTOL1*) gene in the *Phosphorus uptake1* (*Pup1*) locus is associated with phosphorus starvation tolerance (Schatz *et al*., 2014). Sequencing of the traditional *aus*-type rice Kasalath revealed that the *Pup1* locus carried a ~90 kb transposon-rich insertion/deletion that is absent from the Nipponbare reference genome and other rice varieties intolerant to phosphorus starvation (Gamuyao *et al*., 2012). Additionally, the ethylene response factors *SNORKEL1* (*SK1*) and *SNORKEL2* (*SK2*) genes control the deep-water responses of wild rice species. *O. rufipogon* carried both *SK1* and *SK2*, whereas *Oryza nivara* had *SK1* and a new stop codon in exon 2 of *SK2* caused by insertion of a transposon. The *SK2* gene may be non-essential in dry areas and adaptation to this environment may have caused the loss of *SK2* in *O. nivara* (Hattori *et al*., 2009). Boron-toxicity tolerance in barley

was conferred by multiple copies of the *Bot1* gene (Sutton *et al*., 2007), and copy number of the *HvCBF4* and *HvCBF2* genes was reported to influence frost resistance (Francia *et al*., 2016). High copy number of the *multidrug and toxic compound extrusion 1* (*MATE1*) gene increased aluminium tolerance in maize (Maron *et al*., 2013). Three maize lines (carrying the three-*MATE1* copy) had high *MATE1* gene expression (Maron *et al*., 2013). In sorghum, polymorphisms in regulatory regions of the aluminium tolerance locus Alt_{SB} increased Alt_{SB} expression, contributing large allelic effects in the root of aluminium-tolerant individuals. In addition, a tourist-like MITE located in the *SbMATE* coding region increased aluminium tolerance (Magalhaes *et al*., 2007). In wheat, copy number of *C-repeat Binding Factor* (*CBF*) gene at the *Frost Resistant-2* (*FR-H2*) locus led to increased freezing tolerance and winter hardiness (Knox *et al*., 2010). In the sensitive soybean accessions, the gain of Ty1/*copia* retrotransposon in the coding region of novel ion transporter (*GmCHX1*) gene was identified as causing low salt tolerance (Qi *et al*., 2014).

8.5.4 Grain yield and quality of crops

SVs were also shown to affect yield and quality. The 1.2 kb PAV, including transposon insertions, was detected in the *KRN4* locus related to increased kernel row number in different maize lines (Liu *et al*., 2015). In rice, CNV containing tandem duplication of 17 kb at the *Grain Length on Chromosome 7* (*GL7*) improved grain length and grain quality (Wang *et al*., 2015). Similarly, CNV of 18 bp duplication was identified at the upstream silencer of an *FZP* (*FRIZZY PANICLE*) locus and resulted in gene repression and increased grain size in rice (Bai *et al*., 2017). Furthermore, a 1212 bp deletion found ~5 kb upstream of the *GW5* (*grain width and weight on chromosome 5*) gene affected its expression levels and regulated grain width and weight in rice (Liu *et al*., 2017).

8.5.5 Other agronomic traits

Other agronomic traits including fruit shape, plant architecture and fertility are also affected by SVs. In tomato, 24.7 kb gene duplication moderated by the LTR retrotransposon *Rider* increased the *SUN* gene expression controlling the elongated fruit shape (Xiao *et al*., 2008). In rice, the *qPE9-1* locus regulates rice plant architecture, containing erect panicle, contributing to increasing plant yield. The LOF mutation in the *qPE9-1* gene causes more erect panicles (Zhou *et al*., 2009). In rice, structural changes and CNVs at the *Sc* locus confer *japonica–indica* hybrid male sterility (Shen *et al*., 2017). Duplicated pollen essential genes such as *DPL1/DPL2* and *S27/S28* locus produced ~25% sterile pollen grains in male meiosis of the hybrids (Mizuta *et al*., 2010). In cucumber, CNVs of a 30.2 kb region including four genes defining the *Female* (*F*) locus cause plants carrying only female flowers (Zhang *et al*., 2015).

8.6 Future Perspectives

With the increasing number and pronounced importance of crop plant pangenomes, new data structures including pangenome graphs are being developed to aid data analysis and visualization and will likely replace single reference genomes in the near future (Eizenga *et al*., 2020). Recent developments in highly accurate long-read sequencing technologies (e.g. Pacific Biosciences HiFi reads) will further promote construction of multiple chromosome-level assemblies for individuals from the same species, allowing for even more comprehensive pangenomic studies (Garrison *et al*., 2018; Eggertsson *et al*., 2019; Rabbani *et al*., 2020). One key remaining challenge is the meaningful functional annotation of pangenomes to include both genes and regulatory regions, in order to fully understand the impact of observed variations on phenotype. Integrative genomic methodologies, which improve understanding of expression level, biological and functional networks, can contribute significant value to pangenome research (Golicz *et al*., 2018). Altogether, the different classes of genomic variants identified by pangenome studies will have a strong impact on association studies and our ability to identify causal variants responsible for key agronomic traits, while extending pangenome studies to genus level and including crop wild relatives will provide new sources of variation.

References

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L. *et al*. (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182(1), 145–161. DOI: 10.1016/j.cell.2020.05.021.

Ashikawa, I., Hayashi, N., Yamane, H., Kanamori, H., Wu, J. *et al*. (2008) Two adjacent nucleotide-binding site-leucine-rich repeat class genes are required to confer *Pikm*-specific rice blast resistance. *Genetics* 180(4), 2267–2276. DOI: 10.1534/genetics.108.095034.

Bai, X., Huang, Y., Hu, Y., Liu, H., Zhang, B. *et al*. (2017) Duplication of an upstream silencer of *FZP* increases grain yield in rice. *Nature Plants* 3(11), 885–893. DOI: 10.1038/s41477-017-0042-4.

Ballvora, A., Ercolano, M.R., Weiss, J., Meksem, K., Bormann, C.A. *et al*. (2002) The *R1* gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. *The Plant Journal* 30(3), 361–371. DOI: 10.1046/j.1365-313x.2001.01292.x.

Bayer, P.E., Scheben, A., Golicz, A.A., Yuan, Y., Faure, S. *et al*. (2021) Modelling of gene loss propensity in the pangenomes of three *Brassica* species suggests different mechanisms between polyploids and diploids. *Plant Biotechnology Journal* 19(12), 2488–2500. DOI: 10.1111/pbi.13674.

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C. *et al*. (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants* 4(11), 879–887. DOI: 10.1038/s41477-018-0289-4.

Brozynska, M., Furtado, A. and Henry, R.J. (2016) Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnology Journal* 14(4), 1070–1085. DOI: 10.1111/pbi.12454.

Cook, D.E., Lee, T.G., Guo, X., Melito, S., Wang, K. *et al*. (2012) Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* 338(6111), 1206–1209. DOI: 10.1126/science.1228746.

Dolatabadian, A., Bayer, P.E., Tirnaz, S., Hurgobin, B., Edwards, D. *et al*. (2020) Characterization of disease resistance genes in the *Brassica napus* pangenome reveals significant structural variation. *Plant Biotechnology Journal* 18(4), 969–982. DOI: 10.1111/pbi.13262.

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J.M. *et al*. (2020) The impact of transposable elements on tomato diversity. *Nature Communications* 11(1), 4058. DOI: 10.1038/s41467-020-17874-2.

Du, J., Verzaux, E., Chaparro-Garcia, A., Bijsterbosch, G., Keizer, L.C.P. *et al*. (2015) Elicitin recognition confers enhanced resistance to *Phytophthora infestans* in potato. *Nature Plants* 1(4), 15034. DOI: 10.1038/nplants.2015.34.

Ducrocq, S., Madur, D., Veyrieras, J.-B., Camus-Kulandaivelu, L., Kloiber-Maitz, M. *et al*. (2008) Key impact of *Vgt1* on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. *Genetics* 178(4), 2433–2437. DOI: 10.1534/genetics.107.084830.

Eggertsson, H.P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A. *et al*. (2019) GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications* 10(1), 5402. DOI: 10.1038/s41467-019-13341-9.

Eizenga, J.M., Novak, A.M., Sibbesen, J.A., Heumos, S., Ghaffaari, A. *et al*. (2020) Pangenome graphs. *Annual Review of Genomics and Human Genetics* 21(1), 139–162. DOI: 10.1146/annurev-genom-120219-080406.

Feuillet, C., Travella, S., Stein, N., Albar, L., Nublat, A. *et al*. (2003) Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum* L.) genome. *Proceedings of the National Academy of Sciences USA* 100(25), 15253–15258. DOI: 10.1073/pnas.2435133100.

Francia, E., Morcia, C., Pasquariello, M., Mazzamurro, V., Milc, J.A. *et al*. (2016) Copy number variation at the *HvCBF4–HvCBF2* genomic segment is a major component of frost resistance in barley. *Plant Molecular Biology* 92(1–2), 161–175. DOI: 10.1007/s11103-016-0505-4.

Fu, D., Uauy, C., Distelfeld, A., Blechl, A., Epstein, L. *et al*. (2009) A kinase-START gene confers temperature-dependent resistance to wheat stripe rust. *Science* 323(5919), 1357–1360. DOI: 10.1126/science.1166289.

Gamuyao, R., Chin, J.H., Pariasca-Tanaka, J., Pesaresi, P., Catausan, S. *et al*. (2012) The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. *Nature* 488(7412), 535–539. DOI: 10.1038/nature11346.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K. *et al*. (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics* 51(6), 1044–1051. DOI: 10.1038/s41588-019-0410-2.

Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M. *et al*. (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology* 36(9), 875–879. DOI: 10.1038/nbt.4227.

Golicz, A.A., Batley, J. and Edwards, D. (2016a) Towards plant pangenomics. *Plant Biotechnology Journal* 14(4), 1099–1105. DOI: 10.1111/pbi.12499.

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H. *et al*. (2016b) The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications* 7, 13390. DOI: 10.1038/ncomms13390.

Golicz, A.A., Bhalla, P.L. and Singh, M.B. (2018) MCRiceRepGP: a framework for the identification of genes associated with sexual reproduction in rice. *The Plant Journal* 96(1), 188–202. DOI: 10.1111/tpj.14019.

Golicz, A.A., Bayer, P.E., Bhalla, P.L., Batley, J. and Edwards, D. (2020) Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics* 36(2), 132–145. DOI: 10.1016/j.tig.2019.11.006.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D. *et al*. (2017) Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications* 8(1), 2184. DOI: 10.1038/s41467-017-02292-8.

Haberer, G., Kamal, N., Bauer, E., Gundlach, H., Fischer, I. *et al*. (2020) European maize genomes highlight intraspecies variation in repeat and gene content. *Nature Genetics* 52(9), 950–957. DOI: 10.1038/s41588-020-0671-9.

Hanada, K., Vallejo, V., Nobuta, K., Slotkin, R.K., Lisch, D. *et al*. (2009) The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. *The Plant Cell* 21(1), 25–38. DOI: 10.1105/tpc.108.063206.

Hattori, Y., Nagai, K., Furukawa, S., Song, X.-J., Kawano, R. *et al*. (2009) The ethylene response factors *SNORKEL1* and *SNORKEL2* allow rice to adapt to deep water. *Nature* 460(7258), 1026–1030. DOI: 10.1038/nature08258.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G. *et al*. (2014) Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell* 26(1), 121–135. DOI: 10.1105/tpc.113.119982.

Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I. *et al*. (2016) Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *The Plant Cell* 28(11), 2700–2714. DOI: 10.1105/tpc.16.00353.

Hu, Z., Sun, C., Lu, K.-C., Chu, X., Zhao, Y. *et al*. (2017) EUPAN enables pan-genome studies of a large number of eukaryotic genomes. *Bioinformatics* 33(15), 2408–2409. DOI: 10.1093/bioinformatics/btx170.

Huang, C., Sun, H., Xu, D., Chen, Q., Liang, Y. *et al*. (2018) *ZmCCT9* enhances maize adaptation to higher latitudes. *Proceedings of the National Academy of Sciences USA* 115(2), E334–E341. DOI: 10.1073/pnas.1718058115.

Hübner, S., Bercovich, N., Todesco, M., Mandel, J.R., Odenheimer, J. *et al*. (2019) Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nature Plants* 5(1), 54–62. DOI: 10.1038/s41477-018-0329-0.

Hurgobin, B., Golicz, A.A., Bayer, P.E., Chan, C.-K.K., Tirnaz, S. *et al*. (2018) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal* 16(7), 1265–1274. DOI: 10.1111/pbi.12867.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C. *et al*. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449(7161), 463–467. DOI: 10.1038/nature06148.

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H. *et al*. (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* 588(7837), 284–289. DOI: 10.1038/s41586-020-2947-8.

Knox, A.K., Dhillon, T., Cheng, H., Tondelli, A., Pecchioni, N. *et al*. (2010) *CBF* gene copy number variation at *Frost Resistance-2* is associated with levels of freezing tolerance in temperate-climate cereals. *Theoretical and Applied Genetics* 121(1), 21–35. DOI: 10.1007/s00122-010-1288-7.

Lai, J., Li, Y., Messing, J. and Dooner, H.K. (2005) Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. *Proceedings of the National Academy of Sciences* 102(25), 9068–9073. DOI: 10.1073/pnas.0502923102.

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G. *et al*. (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* 42(12), 1053–1059. DOI: 10.1038/ng.715.

Li, C., Song, W., Luo, Y., Gao, S., Zhang, R. *et al*. (2019) The HuangZaoSi maize genome provides insights into genomic variation and improvement history of maize. *Molecular Plant* 12(3), 402–409. DOI: 10.1016/j.molp.2019.02.009.

Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M. *et al*. (2021) Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biology* 22(1), 119. DOI: 10.1186/s13059-021-02351-w.

Li, Y., Zhou, G., Ma, J., Jiang, W., Jin, L. *et al*. (2014) *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology* 32(10), 1045–1052. DOI: 10.1038/nbt.2979.

Lin, K., Zhang, N., Severing, E.I., Nijveen, H., Cheng, F. *et al*. (2014) Beyond genomic variation - comparison and functional annotation of three *Brassica rapa* genomes: a turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics* 15(1), 250. DOI: 10.1186/1471-2164-15-250.

Lisch, D. (2013) How important are transposons for plant evolution? *Nature Reviews Genetics* 14(1), 49–61. DOI: 10.1038/nrg3374.

Liu, J., Chen, J., Zheng, X., Wu, F., Lin, Q. *et al*. (2017) *GW5* acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. *Nature Plants* 3(5), 17043. DOI: 10.1038/nplants.2017.43.

Liu, J., Seetharam, A.S., Chougule, K., Ou, S., Swentowsky, K.W. *et al*. (2020a) Gapless assembly of maize chromosomes using long-read technologies (preprint posted January 15, 2020). *BioRxiv* 2020.01.14.906230. DOI: 10.1186/s13059-020-02029-9.

Liu, L., Du, Y., Shen, X., Li, M., Sun, W. *et al*. (2015) *KRN4* controls quantitative variation in maize kernel row number. *PLoS Genetics* 11(11), e1005670. DOI: 10.1371/journal.pgen.1005670.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H. *et al*. (2020b) Pan-genome of wild and cultivated soybeans. *Cell* 182(1), 162–176. DOI: 10.1016/j.cell.2020.05.023.

MacAlister, C.A., Park, S.J., Jiang, K., Marcel, F., Bendahmane, A. *et al*. (2012) Synchronization of the flowering transition by the tomato *TERMINATING FLOWER* gene. *Nature Genetics* 44(12), 1393–1398. DOI: 10.1038/ng.2465.

Magalhaes, J.V., Liu, J., Guimarães, C.T., Lana, U.G.P., Alves, V.M.C. *et al*. (2007) A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. *Nature Genetics* 39(9), 1156–1161. DOI: 10.1038/ng2074.

Maron, L.G., Guimarães, C.T., Kirst, M., Albert, P.S., Birchler, J.A. *et al*. (2013) Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. *Proceedings of the National Academy of Sciences USA* 110(13), 5241–5246. DOI: 10.1073/pnas.1220766110.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O. *et al*. (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544(7651), 427–433. DOI: 10.1038/nature22043.

Mizuta, Y., Harushima, Y. and Kurata, N. (2010) Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. *Proceedings of the National Academy of Sciences USA* 107(47), 20417–20422. DOI: 10.1073/pnas.1003124107.

Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H. *et al*. (2017) The pangenome of hexaploid bread wheat. *The Plant Journal* 90(5), 1007–1013. DOI: 10.1111/tpj.13515.

Morgante, M., De Paoli, E. and Radovic, S. (2007) Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology* 10(2), 149–155. DOI: 10.1016/j.pbi.2007.02.001.

Nitcher, R., Distelfeld, A., Tan, C., Yan, L. and Dubcovsky, J. (2013) Increased copy number at the *HvFT1* locus is associated with accelerated flowering time in barley. *Molecular Genetics and Genomics* 288(5–6), 261–275. DOI: 10.1007/s00438-013-0746-8.

Ou, L., Li, D., Lv, J., Chen, W., Zhang, Z. *et al*. (2018) Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses. *New Phytologist* 220(2), 360–363. DOI: 10.1111/nph.15413.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J. *et al*. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457(7229), 551–556. DOI: 10.1038/nature07723.

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V. *et al*. (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular Biology and Evolution* 33(10), 2706–2719. DOI: 10.1093/molbev/msw161.

Qi, X., Li, M.-W., Xie, M., Liu, X., Ni, M. *et al*. (2014) Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. *Nature Communications* 5(1), 4340. DOI: 10.1038/ncomms5340.

Rabbani, L., Müller, J. and Weigel, D. (2020) An algorithm to build a multi-genome reference (preprint posted April 13, 2020). *BioRxiv* 2020.04.11.036871. DOI: 10.1101/2020.04.11.036871.

Ruperao, P., Thirunavukkarasu, N., Gandham, P., Selvanayagam, S., Govindaraj, M. *et al*. (2021) Sorghum pan-genome explores the functional utility to accelerate the genetic gain (preprint posted February 03, 2021). *bioRxiv* 2021.02.02.429137. DOI: 10.1101/2021.02.02.429137.

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Niu, X. *et al*. (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proceedings of the National Academy of Sciences USA* 104(27), 11376–11381. DOI: 10.1073/pnas.0704145104.

Sasaki, T. and International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436, 793–800. DOI: 10.1038/nature03895.

Saxena, R.K., Edwards, D. and Varshney, R.K. (2014) Structural variations in plant genomes. *Briefings in Functional Genomics* 13(4), 296–307. DOI: 10.1093/bfgp/elu016.

Schatz, M.C., Maron, L.G., Stein, J.C., Hernandez Wences, A., Gurtowski, J. *et al*. (2014) Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. *Genome Biology* 15(11), 506. DOI: 10.1186/PREACCEPT-2784872521277375.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T. *et al*. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463(7278), 178–183. DOI: 10.1038/nature08670.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F. *et al*. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326(5956), 1112–1115. DOI: 10.1126/science.1178534.

Shen, R., Wang, L., Liu, X., Wu, J., Jin, W. *et al*. (2017) Genomic structural variation-mediated allelic suppression causes hybrid male sterility in rice. *Nature Communications* 8(1), 1310. DOI: 10.1038/s41467-017-01400-y.

Song, B., Wang, H., Wu, Y., Rees, E., Gates, D.J. *et al*. (2020a) Constrained non-coding sequence provides insights into regulatory elements and loss of gene expression in maize (preprint posted July 13, 2020). *BioRxiv* 2020.07.11.192575. DOI: 10.1101/2020.07.11.192575.

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z. *et al*. (2020b) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nature Plants* 6(1), 34–45. DOI: 10.1038/s41477-019-0577-7.

Springer, N.M., Anderson, S.N., Andorf, C.M., Ahern, K.R., Bai, F. *et al*. (2018) The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nature Genetics* 50(9), 1282–1288. DOI: 10.1038/s41588-018-0158-0.

Studer, A., Zhao, Q., Ross-Ibarra, J. and Doebley, J. (2011) Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nature Genetics* 43(11), 1160–1163. DOI: 10.1038/ng.942.

Sun, C., Hu, Z., Zheng, T., Lu, K., Zhao, Y. *et al*. (2017) RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Research* 45(2), 597–605. DOI: 10.1093/nar/gkw958.

Sun, X., Jiao, C., Schwaninger, H., Chao, C.T., Ma, Y. *et al*. (2020) Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nature Genetics* 52(12), 1423–1432. DOI: 10.1038/s41588-020-00723-9.

Sutton, T., Baumann, U., Hayes, J., Collins, N.C., Shi, B.-J. *et al*. (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science* 318(5855), 1446–1449. DOI: 10.1126/science.1146853.

Tao, Y., Zhao, X., Mace, E., Henry, R. and Jordan, D. (2019) Exploring and exploiting pan-genomics for crop improvement. *Molecular Plant* 12(2), 156–169. DOI: 10.1016/j.molp.2018.12.016.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D. *et al*. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences USA* 102(39), 13950–13955. DOI: 10.1073/pnas.0506758102.

Trouern-Trend, A.J., Falk, T., Zaman, S., Caballero, M., Neale, D.B. *et al*. (2020) Comparative genomics of six *Juglans* species reveals disease-associated gene family contractions. *The Plant Journal* 102(2), 410–423. DOI: 10.1111/tpj.14630.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T. *et al*. (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nature* 588(7837), 277–283. DOI: 10.1038/s41586-020-2961-x.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S. *et al*. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557(7703), 43–49. DOI: 10.1038/s41586-018-0063-9.

Wang, Y., Xiong, G., Hu, J., Jiang, L., Yu, H. *et al*. (2015) Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nature Genetics* 47(8), 944–948. DOI: 10.1038/ng.3346.

Würschum, T., Boeven, P.H.G., Langer, S.M., Longin, C.F.H. and Leiser, W.L. (2015) Multiply to conquer: copy number variations at *Ppd-B1* and *Vrn-A1* facilitate global adaptation in wheat. *BMC Genetics* 16, 96. DOI: 10.1186/s12863-015-0258-0.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E.J. and van der Knaap, E. (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* 319(5869), 1527–1530. DOI: 10.1126/science.1153040.

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R. *et al*. (2006) *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* 442(7103), 705–708. DOI: 10.1038/nature04920.

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F. *et al*. (2011a) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology* 30(1), 105–111. DOI: 10.1038/nbt.2050.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D. *et al*. (2011b) Genome sequence and analysis of the tuber crop potato. *Nature* 475(7355), 189–195. DOI: 10.1038/nature10158.

Yang, Q., Li, Z., Li, W., Ku, L., Wang, C. *et al*. (2013) CACTA-like transposable element in *ZmCCT* attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. *Proceedings of the National Academy of Sciences USA* 110(42), 16969–16974. DOI: 10.1073/pnas.1310949110.

Yao, W., Li, G., Zhao, H., Wang, G., Lian, X. *et al*. (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biology* 16(1), 187. DOI: 10.1186/s13059-015-0757-3.

Yu, C., Zhang, J. and Peterson, T. (2011) Genome rearrangements in maize induced by alternative transposition of reversed Ac/Ds termini. *Genetics* 188(1), 59–67. DOI: 10.1534/genetics.111.126847.

Yu, J., Golicz, A.A., Lu, K., Dossa, K., Zhang, Y. *et al*. (2019) Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnology Journal* 17(5), 881–892. DOI: 10.1111/pbi.13022.

Zhang, B., Zhu, W., Diao, S., Wu, X., Lu, J. *et al*. (2019) The poplar pangenome provides insights into the evolutionary history of the genus. *Communications Biology* 2, 215. DOI: 10.1038/s42003-019-0474-7.

Zhang, Z., Mao, L., Chen, H., Bu, F., Li, G. *et al*. (2015) Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *The Plant Cell* 27(6), 1595–1604. DOI: 10.1105/tpc.114.135848.

Zhao, J., Bayer, P.E., Ruperao, P., Saxena, R.K., Khan, A.W. *et al*. (2020) Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). *Plant Biotechnology Journal* 18(9), 1946–1954. DOI: 10.1111/pbi.13354.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A. *et al*. (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics* 50(2), 278–284. DOI: 10.1038/s41588-018-0041-z.

Zhou, P., Silverstein, K.A.T., Ramaraj, T., Guhlin, J., Denny, R. *et al*. (2017) Exploring structural variation and gene family architecture with *de novo* assemblies of 15 *Medicago* genomes. *BMC Genomics* 18(1), 261. DOI: 10.1186/s12864-017-3654-1.

Zhou, Y., Zhu, J., Li, Z., Yi, C., Liu, J. *et al*. (2009) Deletion in a quantitative trait gene *qPE9-1* associated with panicle erectness improves plant architecture during rice domestication. *Genetics* 183(1), 315–324. DOI: 10.1534/genetics.109.102681.

Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S. *et al*. (2020) A platinum standard pan-genome resource that represents the population structure of Asian rice. *Scientific Data* 7(1), 113. DOI: 10.1038/s41597-020-0438-2.

Zuo, W., Chao, Q., Zhang, N., Ye, J., Tan, G. *et al*. (2015) A maize wall-associated kinase confers quantitative resistance to head smut. *Nature Genetics* 47(2), 151–157. DOI: 10.1038/ng.3170.

3. Benchmarking Oxford Nanopore read alignment-based insertion and deletion detection in crop plant genomes

Gözde Yildiz, Silvia F. Zanini, Nazanin P. Afsharyan, Christian Obermeier, Rod J. Snowdon, and Agnieszka A. Golicz

The Plant Genome, 16, e20314.

https://doi.org/10.1002/tpg2.20314

Received: 22 September 2022 | Accepted: 15 January 2023

DOI: 10.1002/tpg2.20314

The Plant Genome

ORIGINAL ARTICLE

Benchmarking Oxford Nanopore read alignment-based insertion and deletion detection in crop plant genomes

Gözde Yildiz | **Silvia F. Zanini** | **Nazanin P. Afsharyan** | **Christian Obermeier** | **Rod J. Snowdon** | **Agnieszka A. Golicz**

Department of Plant Breeding, Justus Liebig University Giessen, Giessen, Germany

Correspondence
Agnieszka A. Golicz, Department of Plant Breeding, Justus Liebig University Giessen, Giessen, Germany.
Email: Agnieszka.Golicz@agrar.uni-giessen.de

Assigned to Associate Editor Hon-Ming Lam.

Funding information
Alexander von Humboldt Foundation; German Research Foundation, Grant/Award Number: 458716530

Abstract

Structural variations (SVs) are larger polymorphisms (> 50 bp in length), which consist of insertions, deletions, inversions, duplications, and translocations. They can have a strong impact on agronomical traits and play an important role in environmental adaptation. The development of long-read sequencing technologies, including Oxford Nanopore, allows for comprehensive SV discovery and characterization even in complex polyploid crop genomes. However, many of the SV discovery pipeline benchmarks do not include complex plant genome datasets. In this study, we benchmarked insertion and deletion detection by popular long-read alignment-based SV detection tools for crop plant genomes. We used real and simulated Oxford Nanopore reads for two crops, allotetraploid *Brassica napus* (oilseed rape) and diploid *Solanum lycopersicum* (tomato), and evaluated several read aligners and SV callers across 5×, 10×, and 20× coverages typically used in re-sequencing studies. We further validated our findings using maize and soybean datasets. Our benchmarks provide a useful guide for designing Oxford Nanopore re-sequencing projects and SV discovery pipelines for crop plants.

1 | INTRODUCTION

Structural variations (SVs) are a major type of polymorphisms, which consist of insertions, deletions, inversions, duplications, and translocations. SVs are larger polymorphisms (> 50 bp) compared with single nucleotide polymorphisms (SNPs) and small indels (insertions and deletions). Copy number variations (CNVs) and presence/absence variations (PAVs) occur due to these genomic polymorphisms (Alkan et al., 2011; Sedlazeck et al., 2018a). Insertions and deletions are the most abundant type of SV (Alonge et al., 2020; Fuentes et al., 2019; Goel et al., 2019), can have a strong effect on crop traits, and have been shown to play a role in domestication and environmental adaptation (Gill et al., 2021; Tao et al., 2019; Yildiz et al., 2022; Zanini et al., 2022). Until recently, the lack of high-quality reference assemblies and the complex nature of often large, polyploid genomes made comprehensive SV exploration challenging in crop genomic research (Meyers & Levin, 2006; Yuan et al., 2021).

Development of long-read sequencing technologies such as Oxford Nanopore Technologies (ONT) (Jain et al., 2016) and Pacific Biosciences (PacBio) (Roberts et al., 2013) provided new opportunities for comprehensive SV discovery in crop plants. The sequencing accuracy of these technologies is continuously improving. Currently, PacBio HiFi

Abbreviations: CNV, copy number variant; ONT, Oxford Nanopore Technologies; PacBio, Pacific Biosciences; PAV, presence/absence variant; SNP, single nucleotide polymorphism; SV, structural variant.

consensus reads exceed 99% accuracy (Wenger et al., 2019) while ONT R10.3 raw reads accuracy exceeds 95% (Delahaye & Nicolas, 2021). The reduction in error rates facilitates downstream applications, including the production of high-quality genome assemblies, and SV detection. ONT sequencing in particular is being adopted in crop plant research for large scale re-sequencing projects of tens to hundreds of individuals (Alonge et al., 2020; Chawla et al., 2021; Lemay et al., 2022; Vollrath et al., 2021; Zhang et al., 2022). Despite the constant decrease in sequencing error rate, long-read technologies require specialized computational approaches to take advantage of them efficiently.

The two main approaches for SV discovery are *de novo* assembly-based and read alignment-based. *De novo* assembly-based approaches assemble reads into longer contigs and identify SVs by aligning assemblies (Wenger et al., 2019). Read alignment-based approaches directly align reads to reference genomes to discover SVs. *De novo* assembly-based methods perform better at finding larger variants (tens to hundreds of kbp long; exceeding the length of individual reads) but require sufficient amount of data to produce high-quality assemblies, which leads to substantial increase in cost of the experiments for larger crop genomes. However, read alignment-based approaches can perform well even at modest sequencing depths of 5× to 10× and use less computational resources, but the discovered SVs are limited to differences with the reference genome which makes this approach more suitable for larger re-sequencing projects (Coster et al., 2021). Several algorithms were developed for SV discovery from long-reads including Sniffles (Sedlazeck et al., 2018b), NanoVar (Tham et al., 2019), SVIM (Heller & Vingron, 2019), cuteSV (Jiang et al., 2020), and dysgu (Cleal & Baird, 2022), which have been comprehensively reviewed recently (Mahmoud et al., 2019; Yuan et al., 2021). Additionally, several long-read aligners are available such as minimap2 (Li, 2018), NGMLR (Sedlazeck et al., 2018a), Vulcan (Fu et al., 2021), and lra (Ren & Chaisson, 2021). Considering the continued development and improvement in read-alignment and SV detection algorithms and multitude of their possible combinations, their combined performances in SV detection demand realistic and up-to-date benchmarks to guide the selection of SV discovery tools.

In this study, we hypothesized that certain combination(s) of read aligners and SV discovery software will have superior performance in datasets representing complex crop genomes. We used real and simulated ONT reads for two crop plant genomes and evaluated several mappers and SV callers across coverages including 5×, 10×, and 20× typically utilized in re-sequencing studies. We chose to perform benchmarking on allotetraploid *Brassica napus* (oilseed rape) and diploid *Solanum lycopersicum* (tomato) as these two species represent different ploidy, have different SV profiles, and were already studied using Oxford Nanopore Technology. We further validated our findings using maize and soybean datasets. Our benchmarks provide a guide for choosing insertion and deletion discovery tools for low to medium coverage sequencing projects.

Core Ideas

- Structural variants (SVs) have strong impact on crop traits and play an important role in environmental adaptation.
- Long read based SV discovery tools have not been comprehensively evaluated in crops.
- We benchmarked popular SV discovery tools using real and simulated data for two contrasting crop genomes.
- Our benchmarks provide a guide for choosing insertion and deletion discovery tools for low to medium sequencing coverage experiments.

2 | MATERIALS AND METHODS

2.1 | Read aligners, SV callers, and benchmarking datasets

The SV callers included in the study were selected using several criteria: (1) citation count (adjusted by number of years since publication and used as a proxy for popularity in the research community); (2) publication date and maintenance status (excluding older tools that were no longer maintained); (3) ability to detect both insertion and deletion SVs from ONT data. The benchmarking approach involved four long-read aligners, including minimap2 (Li, 2018), NGMLR (Sedlazeck et al., 2018a), lra (Ren & Chaisson, 2021), and Vulcan (Fu et al., 2021) as well as five SV calling software namely Sniffles (v2) (Sedlazeck et al., 2018b), NanoVar (Tham et al., 2019), SVIM (Heller & Vingron, 2019), cuteSV (Jiang et al., 2020), and dysgu (Cleal & Baird, 2022). All aligners and SV caller versions are provided in detail in (Table S1). Three simulated datasets (Sim_ONT_Bn1, Sim_ONT_Bn2, and Sim_ONT_Sl) and publicly available data, for *B. napus* and *S. lycopersicum* genomes, were used. The real-world datasets for whole genome Nanopore sequencing of *B. napus* cv. King 10 (accession number: SRR15731030) (Vollrath et al., 2021), *S. lycopersicum* cv. M82 (accession number: SRR16966224) (Alonge et al., 2021), *Zea mays* cv. Mo17 (accession number: SRR15447413), and *Glycine max* cv. Maple Isle (accession number: SRR15342671 and SRR15342672) were downloaded from NCBI Sequencing Read Archive. All but soybean datasets were randomly subsampled to 5×, 10×, and 20×

coverages using Rasusa (Hall, 2022) to test the effect of sequencing depth on SV discovery.

2.2 | Simulated dataset generation

For three simulated datasets (workflow for all simulations is presented in (Figure S1), new haplotypes including SVs were generated, and synthetic ONT reads were simulated using VISOR v1.1 (Bolognini et al., 2020). For simulation one (Sim_ONT_Bn1), 20,000 genomic intervals (mean: 750 bp, SD: 500 bp) were randomly drawn from the *B. napus* genome (Express 617 v1). A subset of 10,000 was denoted as deletions. For the remaining 10,000, denoted as insertions, the genomic start coordinate was retained, while the sequences corresponding to the genomic intervals were extracted, randomly re-assigned to the coordinates, and served as insertion sequences at those coordinates (Figure S1).

Simulations two and three, denoted Sim_ONT_Bn2 and Sim_ONT_Sl, were designed to reflect SVs found in real-world datasets. For Sim_ONT_Bn2, the assembled *B. napus* genomes Express 617 v1 (Lee et al., 2020) and Westar (Song et al., 2020) were aligned using minimap2 v2.24. SVs were detected using SVIM-asm v1.0.2 (Heller & Vingron, 2020). To reduce the effect of using minimap2 for benchmarking dataset generation, the SV locations were shifted by a randomly selected number in the (−5000, 5000) interval. This changed the exact SV site while maintaining the realistic distribution of SV sizes and locations along the genome. A random subset of 10,000 insertions and 10,000 deletions was drawn from all SVs to create the benchmarking dataset. SNPs discovered from short reads using bcftools v1.15.1 were also included. The SVs and SNPs were provided to VISOR to generate a new haplotype, which in turn was used for Oxford Nanopore read simulation. Sim_ONT_Sl was generated using the same strategy as for Sim_ONT_Bn2 but designed to reflect SVs of the *S. lycopersicum* genome. Heinz 1706 (Slycopersicum_691_SL4.0) and M82 (Alonge et al., 2021) assemblies were used for whole genome alignments. Due to smaller number of SVs, a random subset of 2500 insertions and 2500 deletions were drawn from all SVs. For maize, we used Zmays_493_APGv4 (B73) and ZmaysB84_681 (B84) (Bornowski et al., 2021).

To test the effect of sequencing depth on SV discovery, the datasets were simulated at 5×, 10×, and 20× coverage. The simulations provided the objective truth sets, which could be used to calculate SV precision, recall, and combined F1-scores. Precision describes the proportion of correct positive predictions among all positive predictions. It is calculated by dividing the true positives by overall positives. Recall describes the proportion of positive predictions made out of all positive elements in the dataset. It is calculated by dividing true positives by total number of relevant elements. F1-score combines precision and recall by taking their harmonic mean. Its value ranges from 0 to 1. F1-score close to 1 indicates high precision and recall. Using two different strategies for generating simulated datasets will make it possible to minimize analytical bias. If the same combination of tools performed best on all simulated datasets, this will likely reflect true superior performance.

2.3 | Comparative analyses

Express 617 v1 for *B. napus* (Lee et al., 2020) and Slycopersicum_691_SL4.0 for *S. lycopersicum* (Hosmani et al., 2019) were used as reference sequences. Simulated datasets and real subsampled reads at each coverage depth were aligned to respective reference genomes. The SV call sets were filtered using the following criteria: (1) number of minimum supporting reads: 5×: 3, 10×: 5, and 20×: 8; (2) SV type: INS or DEL (the most abundant SVs supported by all the benchmarked tools); (3) minimum SV length: 50 bp; (4) SV quality: SVs flagged as "PASS"; (5) genotype: homozygous genotype for alternative allele ('1/1'). For simulated data, precision, recall, and F1-scores of the SVs were computed for each combination of coverage depth, read aligner, and SV caller using Truvari v3.0.0 (English et al., 2022). Comparisons between results from the same tool combination across different coverages and different tool combinations across the same coverages were performed using surpyvor v0.8.1 (Jeffares et al., 2017). For real datasets, where no truth sets were available, we focused on within-dataset comparisons and how those compared to the results from simulated data. All the relevant commands for simulated data generation and SV discovery are available in the Supporting Information. To ensure that the datasets were comparable, soybean SV calls were filtered using the same criteria as described in Lemay et al. (2022).

3 | RESULTS

3.1 | Selecting the benchmarking datasets

We chose to focus on two crop plant species *B. napus* (oilseed rape; genome size ~1.1 Gbp) and *S. lycopersicum* (tomato; genome size ~900 Mbp) because they are both important crops and their structural variation was previously studied using Oxford Nanopore Technologies (Alonge et al., 2020; Chawla et al., 2021). Whole Genome Alignment (WGA)-based SV discovery also suggested that they have quite different SV profiles with 38,666 SVs (Real_WGA_Bn, mean size: 2068 bp, median size: 593 bp, 19,450 insertions and 19,216 deletions) discovered for *B. napus* and 7108 SVs (Real_WGA_Sl, mean size: 3029 bp, median size:

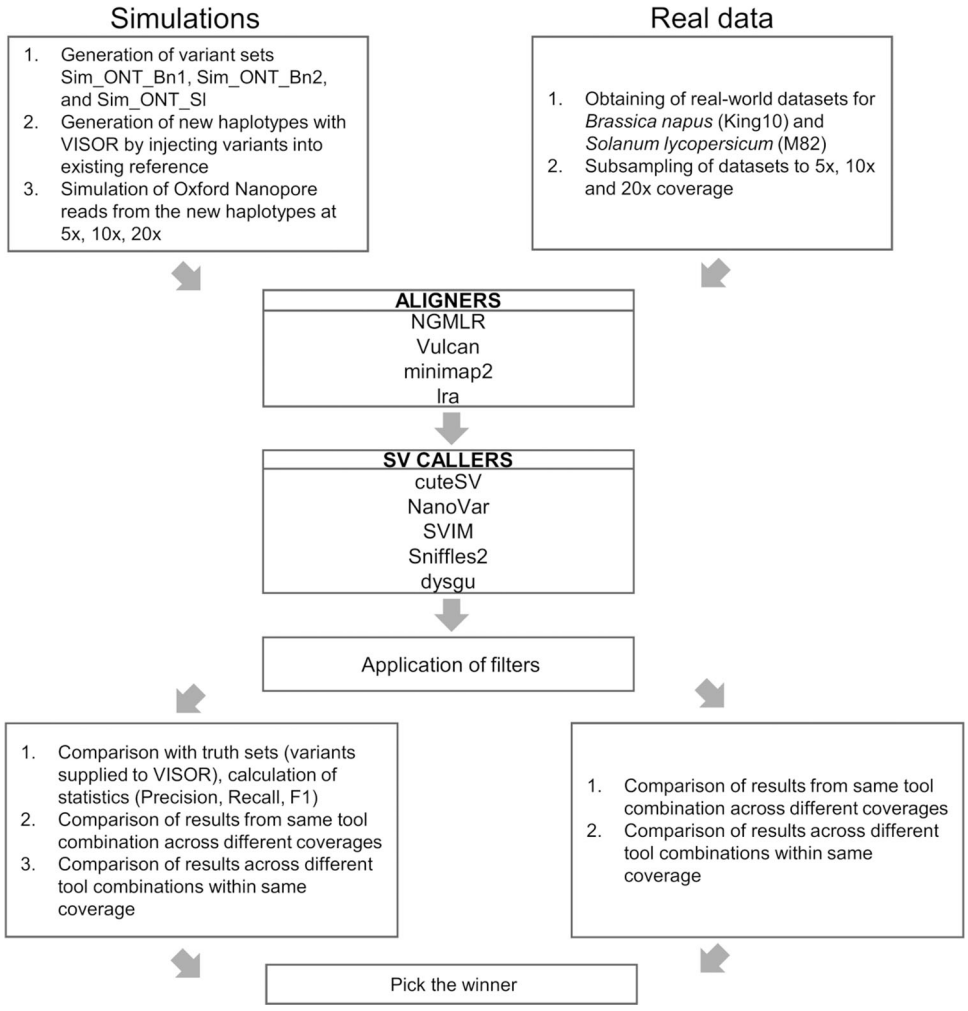


FIGURE 1 Graphical overview of the benchmarking workflow. SV, structural variant.

178 bp, 4159 insertions and 2949 deletions) discovered for *S. lycopersicum*.

Two simulated *B. napus* haplotypes (Sim_ONT_Bn1 and Sim_ONT_Bn2) and one simulated *S. lycopersicum* haplotype (Sim_ONT_Sl) were used to generate Oxford Nanopore reads at 5×, 10×, and 20× to test the effect of sequencing depth on SV discovery. The two publicly available real-world datasets, from *B. napus* (38×) and *S. lycopersicum* (68×), were subsampled with the same logic (Real_ONT_Bn, Real_ONT_Sl). The available graphical representation of a workflow for simulation and real data are shown in Figure 1.

3.2 | Characteristics of structural variant truth sets

The SVs supplied to VISOR to generate Sim_ONT_Bn1, Sim_ONT_Bn2, and Sim_ONT_Sl haplotypes served as three truth sets for our comparisons. The truth sets included deletions and insertions. The length distribution of truth set SVs is presented in Figure 2. Sim_ONT_Bn1 is unbiased in terms of the bioinformatics tools used, as the regions representing SVs were entirely randomly drawn from the *B. napus* genome. For any simulated dataset to reflect realistic SV distribution, SVs have to be discovered first and provided to the simulation software. Any relationship between tools used for SV identification for long-read dataset simulation and tools used for SV detection from these simulated reads (for example use of similar/same mapping algorithm) can result in inflated performance and biased results. However, Sim_ONT_Bn1 does not reflect realistic SV length and genomic distribution. To mitigate that, Sim_ONT_Bn2 and Sim_ONT_Sl were created using SVs derived from real-world datasets. The two simulation strategies are complementary and should allow both unbiased and realistic assessment of SV calls. The median (mean) sizes (bp) for insertions and deletions were 800 (834) and 795 (825) for Sim_ONT_Bn1, 629 (1959) and 594 (1904) for Sim_ONT_Bn2 and 162 (3178) and 165 (2477) for Sim_ONT_Sl. Overall, the Sim_ONT_Bn2 and Sim_ONT_Sl truth sets had a wider range of insertion and deletion sizes.

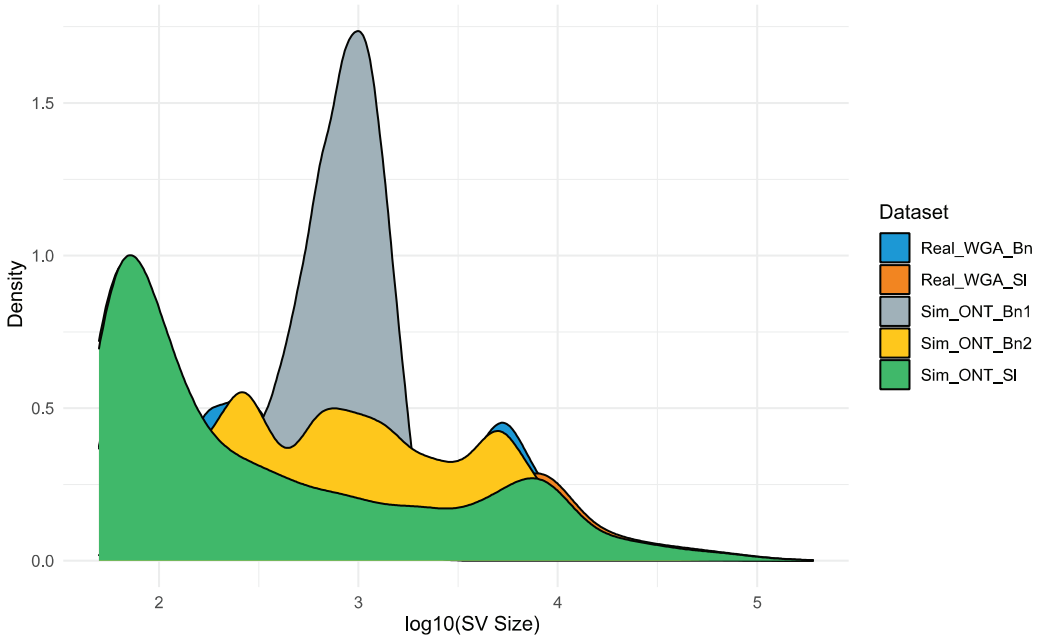


FIGURE 2 Size distribution of the real-world structural variants (SVs) and SVs from three benchmarking datasets.

They were more reflective of true biological variation, making them more realistic than the Sim_ONT_Bn1 truth set.

3.3 | Performance of long read aligners

Subsampled *S. lycopersicum, B. napus*, and simulated reads were aligned using lra, minimap2, Vulcan, and NGMLR to the Slycopersicum_691_SL4.0, and Express 617 v1 reference genomes. Mapping statistics and run times of alignment against relevant reference genomes with different coverages of Sim_ONT_Bn1, Sim_ONT_Bn2, Sim_ONT_Sl, *B. napus* (Real_ONT_Bn), and *S. lycopersicum* (Real_ONT_Sl) real-world datasets are given in Table S2. Minimap2 had the shortest run time across all coverages. Conversely, NGMLR had the longest run time and also the lowest mapping rate. Figure 3 shows mapping runtime (h:mm:ss or m:ss) for both simulation and real-world datasets with eight CPUs. Real_ONT_Bn dataset with 20× coverage was aligned ~220 h by NGMLR and ~119 h by Vulcan, compared to ~4 h by minimap2 and ~5 h by lra. Therefore, minimap2 and lra provided a greater speed advantage than NGMLR and Vulcan. The run times increased with the higher coverages (Figure 3). Processing of real data took substantially longer than processing of simulated data. Moreover, Vulcan and minimap2 produced the highest proportion of mapped reads in Real_ONT_Bn (> 96%), Real_ONT_Sl (96%−98%), and all simulated data (> 98%) (Table S2). NGMLR reported the lowest proportion of mapped reads for Real_ONT_Bn (~81%) and Real_ONT_Sl (~76%), while lra and NGMLR resulted in similar statistics (96%−97%) for Sim_ONT_Bn1, Sim_ONT_Bn2, and Sim_ONT_Sl at each coverage. The combination of fast run time, good mapping rate, and the SV calling results presented below suggest that minimap2 is the top-performing aligner for simulated and real reads.

3.4 | Performance of SV callers on simulated data

3.4.1 | Performance using Sim_ONT_Bn1 as benchmark

We calculated the precision, recall, and F1-score of the SVs generated using different mapper and SV caller combinations using the Sim_ONT_Bn1 truth set. Table S3 shows comparison of the precision, recall, and F1-scores for all mapper/SV caller combinations at the 5×, 10×, and 20× coverages. Each aligner/SV caller combination was evaluated with respect to total SVs, deletions, and insertions. Figure 4 presents the corresponding F1-scores at 5× to 20× coverages. CuteSV after minimap2 alignment reached the highest F1-scores 5×:~0.90, 10×:~0.97, and 20×:~0.99 for total SVs, 5×:~0.91, 10×:~0.97, and 20×:~0.99 for deletions, and 5×:~0.89, 10×:~0.96, and 20×:~0.99 for insertions. At the lower end of coverage (5×), the combination of minimap2/cuteSV provided a better advantage when compared to other mapper/SV caller combinations, especially in capturing insertions. Minimap2/Sniffles2 had second-best F1-scores (Figure 4). SVs detection by NanoVar was obtained directly from reads as NanoVar has its own internal mapping

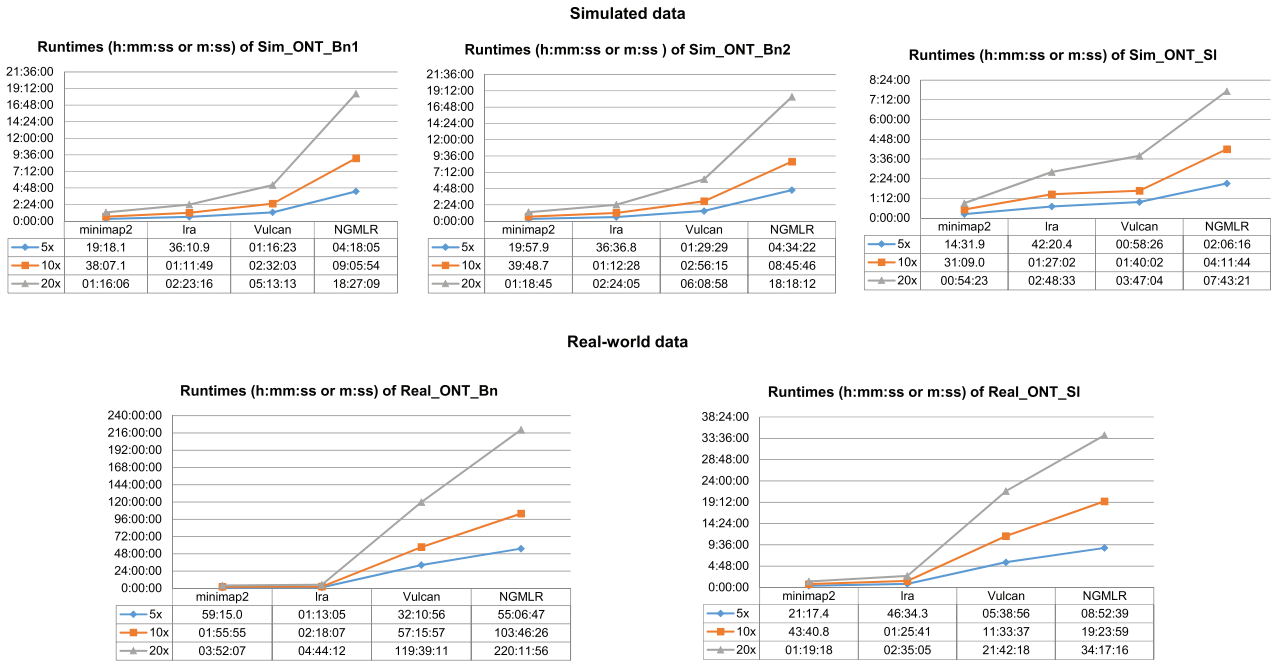


FIGURE 3 Read aligner run time (h:mm:ss or m:ss) for both simulation and real-world datasets with 5×, 10×, and 20× coverages (8 CPU). The reads were simulated with a mean length of 15,000 bp. Unplaced contigs were not included in simulations, which may reduce run time for simulated reads. Read-world reads had higher N50 (~29 Kbp for *B. napus* and ~42 Kbp for *S. lycopersicum*) compared to simulated data (~22 Kbp). In addition, *B. napus* real world data could contain non-reciprocal homeologous exchanges (HEs) uncounted for in simulations. Higher N50 and presence of HEs could increase run time for real-world data.

algorithm; therefore, the precision, recall, and F1-scores for different aligners are not included.

We also compared the total number of SVs, insertions, and deletions for all tested aligner/SV caller combinations. Table S4 summarizes the number of SVs found at 5×, 10×, and 20× coverages. There were more discovered deletions than insertions regardless of coverage. The combinations of minimap2/cuteSV and minimap2/Sniffles2 detected the highest number of SVs at each coverage. We also analyzed how many of the SVs overlapped across different coverages while using the same tool combination and how many of the SVs overlapped across different tool combinations within the same coverage. Data S1 shows the number of overlapping and unique SVs across coverages. Minimap2/cuteSV combination had the highest number of overlapping SVs. It also resulted in the highest proportion of overlapping SVs; 76.99% for all SVs, 79.19% for deletions, and 74.79% for insertions, while the minimap2/Sniffles2 combination (second best according to F1-scores) had the second highest percentage overlap; 75.35% for all SVs, 78.35% for deletions, and 72.33% for insertions (Table S5). In addition, we performed comparisons across different tool combinations within the same coverage. Data S2 displays the overlap, including the intersection sizes between SV calls and the Sim_ONT_Bn1 truth set. The highest number of overlapping SVs was found at 20x coverage, following minimap2 aligner. Our Sim_ONT_Bn1 results suggest that the combination of cuteSV and Sniffles2 with minimap2 alignment gave the best results achieving high F1-scores and capturing the highest number of overlapping SVs across coverages.

3.4.2 | Performance using Sim_ONT_Bn2 as benchmark

While Sim_ONT_Bn1 represents relatively short SVs randomly distributed along the genome, Sim_ONT_Bn2 reflects true biological variation in *B. napus*. Table S6 presents comparison of the precision, recall, and F1-scores for all mapper/SV caller combinations at the 5×, 10×, and 20× coverages. Figure 5 presents the F1-scores of SVs (total, insertions, and deletions) obtained using different combinations of aligners and variant callers across coverages. CuteSV following minimap2 alignment again was the top performing combination with the highest overall F1-score values 5×:~0.87, 10×:~0.93, and 20×:~0.96 for total SVs, 5×:~0.90, 10×:~0.96, and 20×:~0.98 for deletions, and 5×:~0.83, 10×:~0.90, and 20×:~0.94 for insertions. Especially, at low 5× coverage, this combination performed better than others. Minimap2/Sniffles2 had the second highest F1-scores at 20× coverage as in Sim_ONT_Bn1. However, minimap2/dysgu F1-score for insertions at 5× and 10× was higher than Sniffles2 after the minimap2 alignment.

Total			
minimap2			
	5x-F1	10x-F1	20x-F1
cuteSV	0.9003	0.9676	0.9955
Sniffles2	0.8928	0.9635	0.9948
SVIM	0.7825	0.9645	0.9869
dysgu	0.8618	0.9417	0.9776
lra			
cuteSV	0.8665	0.9417	0.9829
Sniffles2	0.8578	0.9352	0.9801
SVIM	0.7291	0.9354	0.9793
dysgu	0.7593	0.8718	0.9148
Vulcan			
cuteSV	0.8495	0.9256	0.9751
Sniffles2	0.8000	0.8787	0.9463
SVIM	0.6864	0.9024	0.9345
dysgu	0.7441	0.8253	0.8639
NGMLR			
cuteSV	0.8001	0.8691	0.9220
Sniffles2	0.6524	0.7174	0.7689
SVIM	0.6295	0.8496	0.8980
dysgu	0.6275	0.7120	0.7415
	5x-F1	10x-F1	20x-F1
NanoVar	0.8950	0.9593	0.9848

Deletions			
minimap2			
	5x-F1	10x-F1	20x-F1
cuteSV	0.9074	0.9733	0.9961
Sniffles2	0.9037	0.9724	0.9963
SVIM	0.7970	0.9715	0.9957
dysgu	0.9057	0.9721	0.9952
lra			
cuteSV	0.8836	0.9562	0.9860
Sniffles2	0.8821	0.9557	0.9865
SVIM	0.7696	0.9563	0.9857
dysgu	0.8783	0.9552	0.9852
Vulcan			
cuteSV	0.8707	0.9441	0.9823
Sniffles2	0.8544	0.9323	0.9780
SVIM	0.7325	0.9389	0.9809
dysgu	0.8695	0.9484	0.9814
NGMLR			
cuteSV	0.8490	0.9152	0.9496
Sniffles2	0.8198	0.8924	0.9338
SVIM	0.7116	0.9110	0.9477
dysgu	0.8428	0.9260	0.9534
	5x-F1	10x-F1	20x-F1
NanoVar	0.9012	0.9676	0.9913

Insertions			
minimap2			
	5x-F1	10x-F1	20x-F1
cuteSV	0.8931	0.9620	0.9948
Sniffles2	0.8818	0.9544	0.9933
SVIM	0.7676	0.9574	0.9778
dysgu	0.8140	0.9092	0.9593
lra			
cuteSV	0.8488	0.9267	0.9798
Sniffles2	0.8324	0.9138	0.9736
SVIM	0.6857	0.9135	0.9728
dysgu	0.6112	0.7735	0.8336
Vulcan			
cuteSV	0.8275	0.9065	0.9678
Sniffles2	0.7401	0.8191	0.9124
SVIM	0.6367	0.8632	0.8834
dysgu	0.5866	0.6689	0.7150
NGMLR			
cuteSV	0.7465	0.8187	0.8927
Sniffles2	0.4282	0.4753	0.5424
SVIM	0.5358	0.7805	0.8431
dysgu	0.3116	0.3894	0.4193
	5x-F1	10x-F1	20x-F1
NanoVar	0.8886	0.9509	0.9784

FIGURE 4 F1-scores of Sim_ONT_Bn1 including total structural variants (SVs), deletions, and insertions at 5×, 10×, and 20× coverages for different combinations of read aligners and SV callers.

Total			
minimap2			
	5x-F1	10x-F1	20x-F1
cuteSV	0.8709	0.9301	0.9628
Sniffles2	0.8589	0.9182	0.9545
SVIM	0.7481	0.9195	0.9527
dysgu	0.8602	0.9316	0.9528
lra			
cuteSV	0.8059	0.8732	0.9203
Sniffles2	0.8032	0.8768	0.9237
SVIM	0.6726	0.8616	0.9068
dysgu	0.7045	0.8295	0.8686
Vulcan			
cuteSV	0.8000	0.8635	0.9122
Sniffles2	0.7553	0.8240	0.8759
SVIM	0.6448	0.8382	0.8870
dysgu	0.7240	0.8391	0.8770
NGMLR			
cuteSV	0.7762	0.8408	0.8885
Sniffles2	0.7219	0.7895	0.8442
SVIM	0.6137	0.8095	0.8608
dysgu	0.6703	0.7998	0.8436
NanoVar	0.7987	0.8583	0.8964

Deletions			
minimap2			
	5x-F1	10x-F1	20x-F1
cuteSV	0.9060	0.9609	0.9825
Sniffles2	0.9011	0.9580	0.9827
SVIM	0.7942	0.9549	0.9791
dysgu	0.8968	0.9576	0.9756
lra			
cuteSV	0.8592	0.9254	0.9648
Sniffles2	0.8473	0.9189	0.9578
SVIM	0.7334	0.9178	0.9553
dysgu	0.8125	0.9085	0.9466
Vulcan			
cuteSV	0.8469	0.9101	0.9524
Sniffles2	0.8136	0.8832	0.9300
SVIM	0.7005	0.8919	0.9358
dysgu	0.7788	0.8914	0.9310
NGMLR			
cuteSV	0.8272	0.8887	0.9302
Sniffles2	0.7857	0.8531	0.9029
SVIM	0.6825	0.8731	0.9204
dysgu	0.7541	0.8749	0.9160
NanoVar	0.8399	0.9030	0.9432

Insertions			
minimap2			
	5x-F1	10x-F1	20x-F1
cuteSV	0.8335	0.8973	0.9422
Sniffles2	0.8132	0.8752	0.9248
SVIM	0.6984	0.8816	0.9250
dysgu	0.8214	0.9045	0.9292
lra			
cuteSV	0.7474	0.8155	0.8715
Sniffles2	0.7556	0.8313	0.8874
SVIM	0.6056	0.7992	0.8535
dysgu	0.5752	0.7382	0.7784
Vulcan			
cuteSV	0.7490	0.8126	0.8686
Sniffles2	0.6910	0.7581	0.8160
SVIM	0.5841	0.7790	0.8336
dysgu	0.6642	0.7819	0.8175
NGMLR			
cuteSV	0.7201	0.7882	0.8429
Sniffles2	0.6509	0.7182	0.7785
SVIM	0.5372	0.7382	0.7943
dysgu	0.5743	0.7143	0.7612
NanoVar	0.7550	0.8108	0.8471

FIGURE 5 F1-scores of Sim_ONT_Bn2 including total structural variants (SVs), deletions, and insertions at 5×, 10×, and 20× coverages for different combinations of read aligners and SV callers.

In addition, the total number of SVs, the total number of insertions, and deletions for all combinations of tested aligners and SV callers were compared. Table S7 summarizes the total number of SVs detected at 5×, 10×, and 20× coverages. Minimap2/cuteSV found the highest number of SVs at each coverage like in Sim_ONT_Bn1. Again, more deletions than insertions were found for all aligner and SV caller combinations across different coverages. We also analyzed how many of the SVs overlapped across different coverages while using the same tool combination and how many of the SVs overlapped across different tool combinations within the same coverage. Data S3 lists the number of overlapping SVs

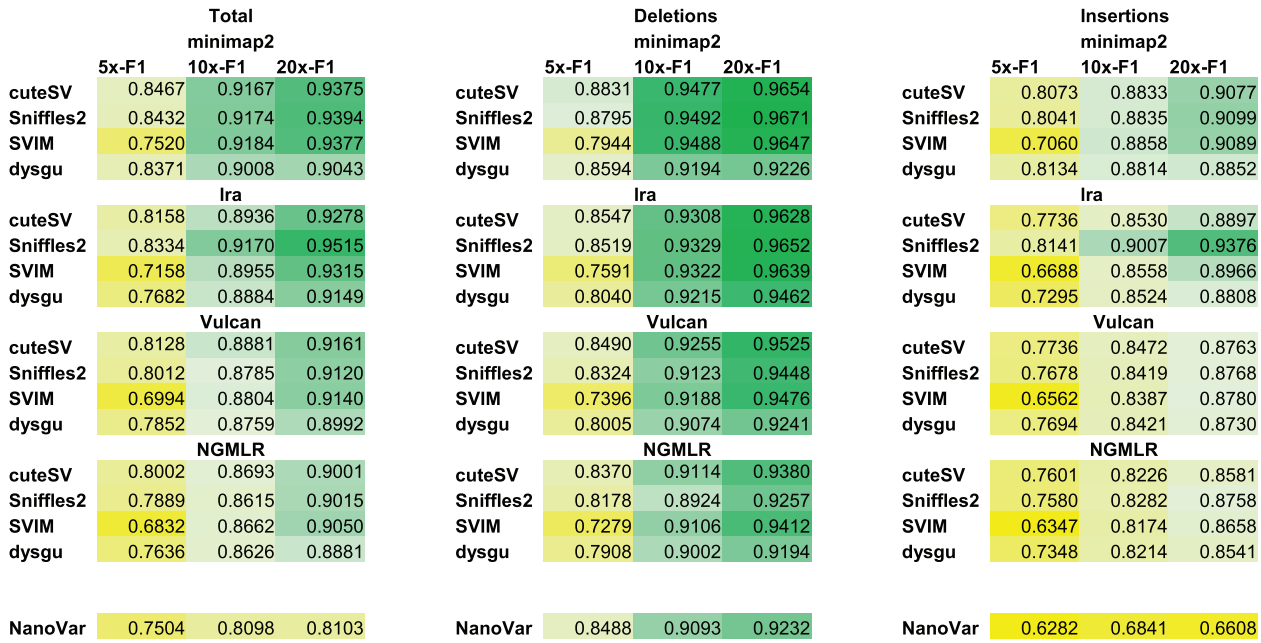


FIGURE 6 F1-scores of Sim_ONT_Sl including total structural variants (SVs), deletions, and insertions at 5×, 10×, and 20× coverages for different combinations of read aligners and SV callers.

across different coverages using the same tool combination. Minimap2/cuteSV combination had the highest number of overlapping SVs. It also had the highest proportion of overlapping SVs; 73.95% for all SVs, 80.05% for deletions, and 67.44% for insertions. The minimap2/dysgu combination was second best detecting 73.23% for all SVs, and 67.28% for insertions. Minimap2/Sniffles2 combination was the second best for deletions with 79.14% overlap (Table S5). Data S4 displays overlap between results from different SV callers within the same coverage after each aligner, including the intersection with the Sim_ONT_Bn2 truth set. The highest number of overlapping SVs was found at 20x coverage, following minimap2 aligner. Overall, in Sim_ONT_Bn2, the combination of cuteSV after minimap2 alignment gave the best results both in terms of F1-Scores and concordance across coverages.

3.4.3 | Performance using Sim_ONT_Sl as benchmark

Sim_ONT_Sl represents the true biological variation of *S. lycopersicum*. Table S8 presents comparison of the precision, recall, and F1-scores for all mapper/SV caller combinations at the 5×, 10×, and 20× coverages. Figure 6 shows the F1-score of SVs (total, insertions, and deletions) identified using combinations of the different aligners and variant callers. CuteSV and Sniffles2 with minimap2 alignment were top performers with the highest F1-score values (5×:~0.85, 10×:~0.92, and 20×:~0.94) for total SVs, (5×:~0.88, 10×:~0.95, and 20×:~0.97) for deletions, and (5×:~0.81, 10×:~0.88, and 20×:~0.91) for insertions. Lra/Sniffles2 combination had the best F1-score for insertions for each coverage.

In addition, the total number of SVs, the total number of insertions, and deletions for all tested aligner/SV caller combinations were compared. Table S9 summarizes the total number of SVs at 5×, 10×, and 20× coverages. Again, more deletions than insertions were found for all aligner and SV caller combinations across coverages like in the previous simulated datasets. The number of SVs overlapping across coverages while using the same tool combination and the number of SVs overlapping across different tool combinations but within the same coverage were also calculated. Data S5 shows the number of overlapping SVs across different coverages using the same tool combination. Minimap2/dysgu combination had the highest number of overlapping SVs. However, minimap2/cuteSV combination found the highest proportion of overlap; 73.49% for all SVs, 77.52% for deletions, and 68.98% for insertions, while the minimap2/Sniffles2 combination was second best detecting 72.73% for all SVs, 76.32% for deletions, and 68.72% for insertions (Figure 7 and Table S5). Although minimap2/dysgu found the highest number of SVs at each coverage in Sim_ONT_Sl, the proportion of overlapped SVs was reported as 68.82%. Data S6 displays overlap between results from different SV callers within the same coverage after each aligner, including the intersection with Sim_ONT_Sl truth set. The highest number of overlapping SVs was found at 20x coverage, following minimap2 aligner. Overall, in Sim_ONT_Sl, the combination of cuteSV and Sniffles2 after minimap2 alignment gave the best

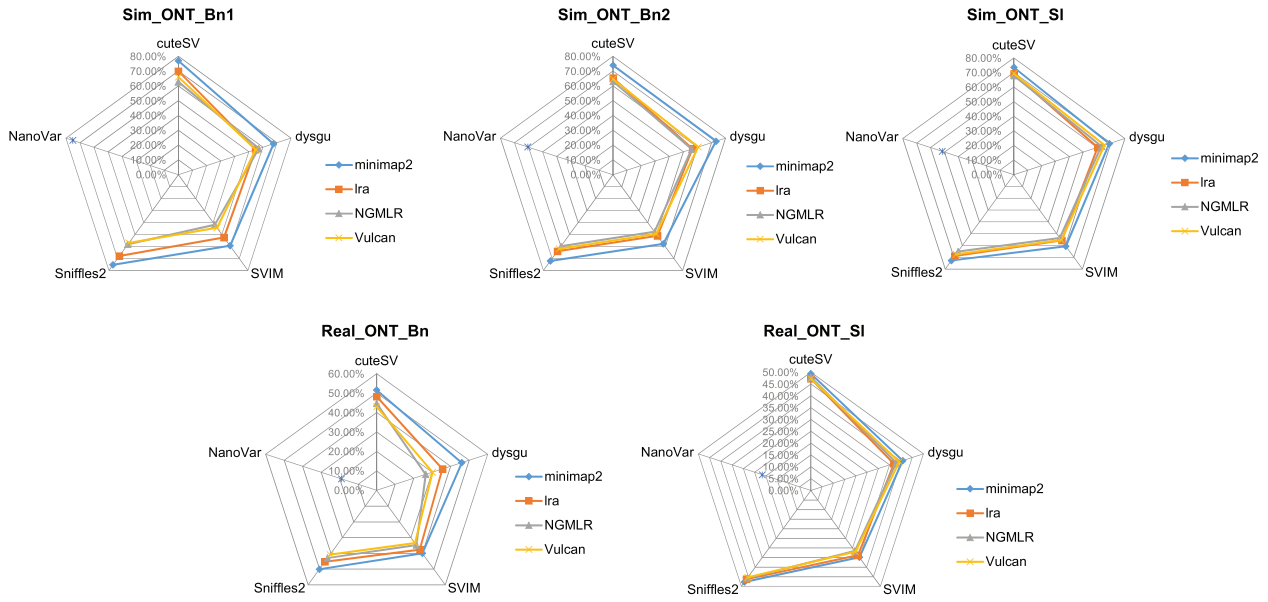


FIGURE 7 Proportion of overlapped structural variant (SVs) (%), across 5×, 10×, and 20× coverages for simulated and real-world datasets.

results both in terms of F1-Scores and concordance across coverages.

3.5 | Performance of SV callers on real-world data

While tool performance on simulated data provides a useful guide, real-world datasets usually provide additional unaccounted-for complexity and challenges. After finding the best combinations in simulated data, we investigated whether the pattern would be similar in real-world datasets. Since for the real-world data we do not have an objective truth set, they were only evaluated from two perspectives which are the congruence of results when using the same tool combination across different coverages and when using different tool combinations within the same coverage.

3.5.1 | Performance on *B. napus* real-world ONT data

B. napus ONT real dataset (Real_ONT_Bn) was evaluated using the above-described strategy. Table S10 shows the number of SVs from all tested combinations at different coverages in *B. napus*. The minimap2/cuteSV and minimap2/dysgu combinations within all coverages captured the highest number of total SVs, deletions, and insertions. Overall, a higher number of deletions than insertions was detected for all aligner and SV caller combinations at different coverages. The number of overlapped SVs across coverages for the same SVs caller/aligner combinations was calculated (Data S7). Minimap2/cuteSV combination found the highest proportion of overlapping SVs discovered at different coverages using the same combination of tools (51.53% of total SVs, 54.52% of deletions, and 47.91% of insertions), while the minimap2/Sniffles2 combination was second best, detecting overlap of 50.1% for all SVs, 54.56% for deletions, and 44.92% for insertions across coverages (Figure 7). Although the minimap2/dysgu combination found more SVs, the percentage of intersecting SV was low. NanoVar detected the lowest proportion of overlapping SVs across coverages (19.04% of total SVs, 25.07% of deletions, and 10.21% of insertions) and discovered more unique SVs. Surprisingly we noticed a high proportion of heterozygous genotypes (0/1) in SV calling results for Real_ONT_Bn, considering that the data represented a highly inbred elite line (Vollrath et al., 2021). Tables S11 and S12 show the number of SVs genotyped as homozygous and heterozygous in simulated and real-world data, respectively. As our SV filtering required the genotypes to be homozygous for the alternative allele (1/1), these heterozygous calls were removed prior to analysis. We also investigated the overlap in SV calls across different tool combinations within the same coverage (Data S8). We observed that a substantial proportion of deletions and insertions were shared by most SV callers, with the largest number of overlapping SVs at 20×, following minimap2 alignment.

3.5.2 | Performance on *S. lycopersicum* real-world ONT data

We performed a similar evaluation for the real-world dataset of *Solanum lycopersicum* (Real_ONT_Sl). Table S13 shows

the number of SVs found from all tested combinations at different coverages. The minimap2/dysgu combinations at 5×, 10×, and 20× captured the most SVs. Additionally, for *S. lycopersicum* all tool combinations with the exception of NanoVar found more insertions than deletions at each coverage. We also calculated the number of overlapping SVs while using the same tool combination across different coverages (Data S9). Minimap2/cuteSV combination found the highest proportion of overlapping SVs; 49.34% for all SVs, 49.63% for deletions, and 49.16% for insertions, while the minimap2/Sniffles2 combination detected 47.80% for all SVs, 49.41% for deletions, and 46.61% for insertions. Even though the minimap2/dysgu combination found more SVs, the percentage of common SVs (40.82%) was low like Real_ONT_Bn data. NanoVar again detected the lowest proportion of overlapping SVs (21.57% for all SVs, 31.20% for deletions, and 12.16% for insertions), and it discovered more unique SVs like for the Real_ONT_Bn dataset (Table S14 and Figure 7). Again, we also tested overlaps between SV calls within the same coverage, but across different tool combinations (Data S10). The largest number of overlapping SVs was found at 20×, following minimap2 alignment.

3.5.3 | Performance of Minimap2 and cuteSV/Sniffles2 combination in other crops

To assess whether our observations are robust for other crops, we performed similar benchmarking analysis for maize and compared already published SV calls in soybean, discovered using a combination of NGMLR and Sniffles1, with our results obtained from minimap2/cuteSV and minimap2/Sniffles2 combinations (Lemay et al., 2022). For maize simulated data, we found that the combination of minimap2/cuteSV had the best performance for deletions while the combination of minimap2/dysgu had the best performance for insertions (Figures S2 and S3). However, as for *B. napus* and *S. lycopersicum*, minimap2/cuteSV combination had much higher overlap across coverages in real world data (Figure S4). For soybean, we found that minimap2/cuteSV and minimap2/Sniffles discovered over 3500 new SVs, while recovering a vast majority of existing calls (Figures S5–S9).

3.5.4 | The Unique features of real-world datasets

We found a surprisingly high proportion of heterozygous calls in the real-world datasets given the highly inbred nature of the material used for sequencing. A high proportion of those is therefore likely SV discovery/genotyping errors. More heterozygous calls were found in the *B. napus* than in the *S. lycopersicum* dataset. *B. napus* is an allotetraploid species, which undergoes reciprocal and non-reciprocal homeologous exchanges (HEs; exchanges of large corresponding chromosome segments between subgenomes). Non-reciprocal HEs could potentially cause erroneous SV calls if there are HE present in the reference, but absent in the sample. As a result, reads will have no corresponding mapping location and may be mis-mapped. To test such a scenario, we used the Sim_ONT_Bn2 dataset (20×, minimap2 for mapping, and cuteSV for SV detection) and two versions of the modified Express 617 reference. In the first version, we replaced chromosome A01 by C01 (two C01 chromosomes and no A01). In the second version, we replaced chromosome C01 by A01 (two A01 chromosomes and no C01). In both cases, the use of the modified reference resulted in an increased number of heterozygous (162.3% for reference with A01 missing, and 237.1% for reference with C01 missing), but not homozygous calls across all chromosomes (Figure 8), suggesting the non-reciprocal HEs can contribute to produce erroneous heterozygous calls.

4 | DISCUSSION

Many of the SV detection tools are benchmarked primarily on human/animal datasets (Bolognini & Magi, 2021; Coster et al., 2019; Dierckxsens et al., 2021; Jiang et al., 2020, 2021; Zhou et al., 2019); however, the complexity and different SV profiles of crop plant genomes might bring unique challenges. Therefore, to guide the design of large-scale long-read re-sequencing studies, this study performed comprehensive benchmarking of popular SV calling tools with a focus on tool performance at lower sequencing coverage. For this purpose, we designed two data simulation strategies representing both unbiased and realistic benchmarking datasets reflecting structural variation for two major crops, oilseed rape (*B. napus*) and tomato (*S. lycopersicum*). We further validated our findings using maize and soybean datasets.

Four long-read aligners (minimap2, NGMLR, lra, and Vulcan) and five SV callers (Sniffles2, SVIM, cuteSV, dysgu, and NanoVar) were tested to detect SVs, particularly deletions and insertions. Our analysis focused on deletions and insertions as they are by far the most abundant SV types. Alignment time varied widely between the four aligners, while differences in the proportion of mapped reads were moderate. As expected, higher sequencing coverage and reference genome size length increased the run time of the mapping algorithms. The real-world datasets required more time at the same coverage, which most likely reflected several factors: exclusion of unplaced contigs from simulations, higher N50 of real world reads, potential presence of homeologous exchanges in *B. napus* dataset, and additional complexity not captured in simulations. Overall, the results found minimap2 to be the best performing aligner for SV calling applications, which also

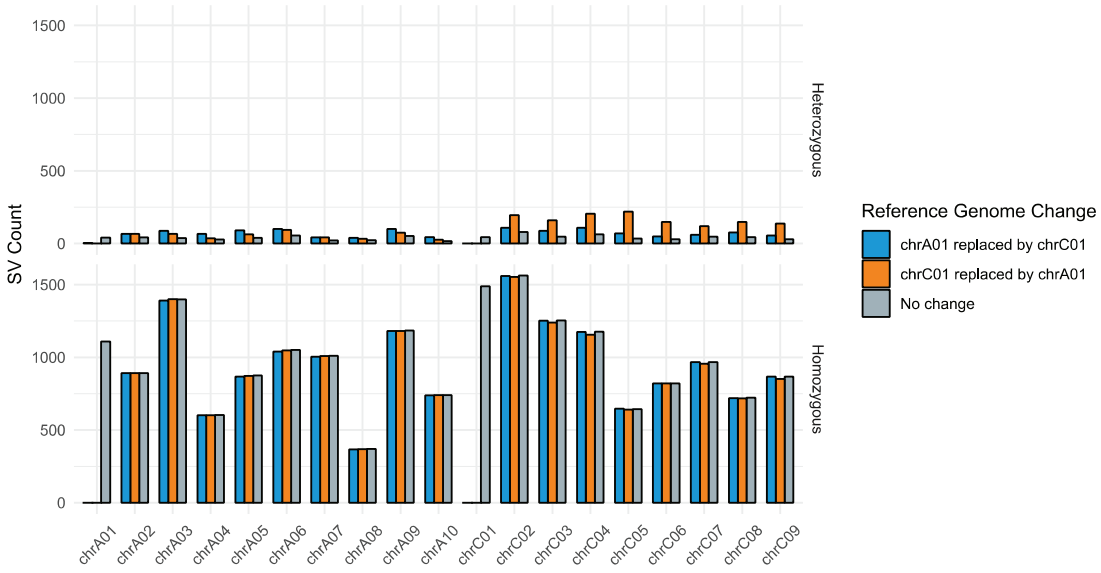


FIGURE 8 The effect of non-reciprocal homeologous exchanges on structural variant (SV) discovery. Nonreciprocal homeologous exchanges were simulated by replacing chromosome A01 by C01 and C01 by A01.

had the fastest run time and the most mapped bases. Recent benchmarking studies on human data also recommended minimap2 among tested aligners such as GraphMap, LAST, and NGMLR (Bolognini & Magi, 2021; Coster et al., 2019; Zhou et al., 2019).

We found that similar tool combinations (especially cuteSV, followed closely by Sniffles2 and dysgu after minimap2 alignment) had superior performance across all the simulated datasets. The findings are in line with a recent study reporting that cuteSV performed better than other tested SV tools such as Sniffles1, SVIM, and pbsv for precision and recall at both SV calling and genotyping in human datasets (Bolognini & Magi, 2021). Increasing coverage improved recall and F1-scores for all tested SVs calling combinations, confirming that the probability of detecting quality SVs increases with more sequencing coverage (Jiang et al., 2021). However, even at low coverages (5×) using cuteSV, Sniffles2, and dysgu for SV detection from reads aligned by minimap2 achieved > 0.8 F1-scores on simulated datasets, suggesting that Oxford Nanopore technology might be suitable for large-scale low coverage re-sequencing projects. While the lack of objective truth sets for real-world datasets precludes similar comparisons, the results revealed that tool combinations with best performance for simulated datasets also had the most consistent outcome across the range of coverages.

The criteria for filtering SV in this study were quite stringent, including retaining only SV genotyped as homozygous for alternative allele (1/1). While in simulated datasets the number of SV genotyped as heterozygous was relatively low, the proportion was much higher for real-world datasets, especially in *B. napus*. We found that in *B. napus*, the presence of homeologous exchanges will likely contribute to the erroneous discovery of heterozygous SV. *B. napus* is well known to harbor wide-spread nonreciprocal homeologous chromosomal exchanges even extending to whole chromosomes, for example, for chromosomes A01 and C01 as simulated here (Udall et al., 2005). The finding underlies the importance of species-specific consideration when interpreting SV discovery results. The presence of HEs likely explains only a proportion of the observed heterozygous calls and other factors need to be considered as well, including other sources of mis-mappings, genotyping errors, and residual heterozygosity in samples.

In conclusion, we found that for homozygous/inbred genotypes often used in crop studies, a substantial proportion of SVs can be discovered/genotyped at coverages as low as 5×, making Oxford Nanopore technology a suitable option for larger-scale re-sequencing studies. At this time, following our benchmarks, we recommend using the minimap2 aligner in combination with either cuteSV or Sniffles2, as it achieves good precision and recall at insertion and deletion calling and found the highest overlap between SVs across coverages.

AUTHOR CONTRIBUTIONS

Gözde Yildiz: Formal analysis; Methodology; Writing – original draft; Writing – review & editing. **Silvia F. Zanini**: Conceptualization; Methodology; Writing – original draft; Writing – review & editing. **Nazanin P. Afsharyan**: Methodology; Writing – review & editing. **Christian Obermeier**: Methodology; Writing – review & editing. **Rod J. Snowdon**: Methodology; Writing – review & editing. **Agnieszka A. Golicz**: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing – original draft; Writing – review & editing.

ACKNOWLEDGMENTS

This work was supported by the Alexander von Humboldt Foundation in the framework of Sofja Kovalevskaja Award to Agnieszka A. Golicz and the German Research Foundation (DFG) project number 458716530 to Rod J. Snowdon. This work was performed with support from Justus Liebig University Bioinformatics Core Facility (BCF).

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ORCID

Gözde Yildiz https://orcid.org/0000-0003-0407-1829
Silvia F. Zanini https://orcid.org/0000-0002-9137-8783
Nazanin P. Afsharyan https://orcid.org/0000-0003-0298-988X
Rod J. Snowdon https://orcid.org/0000-0001-5577-7616
Agnieszka A. Golicz https://orcid.org/0000-0002-9711-4826

REFERENCES

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*, 363–376.

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2021). Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *bioRxiv*, 2021.11.18.469135.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., Amsellem, Z., Razifard, H., Caicedo, A. L., Tieman, D. M., Klee, H., Kirsche, M., … & Lippman, Z. B. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, *182*, 145–161.e23. https://doi.org/10.1016/j.cell.2020.05.021

Bolognini, D., & Magi, A. (2021). Evaluation of germline structural variant calling methods for nanopore sequencing data. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.761791

Bolognini, D., Sanders, A., Korbel, J. O., Magi, A., Benes, V., & Rausch, T. (2020). VISOR: A versatile haplotype-aware structural variant simulator for short- and long-read sequencing. *Bioinformatics*, *36*, 1267–1269. https://doi.org/10.1093/bioinformatics/btz719

Bornowski, N., Michel, K. J., Hamilton, J. P., Ou, S., Seetharam, A. S., Jenkins, J., Grimwood, J., Plott, C., Shu, S., Talag, J., Kennedy, M., Hundley, H., Singan, V. R., Barry, K., Daum, C., Yoshinaga, Y., Schmutz, J., Hirsch, C. N., Hufford, M. B., … & Buell, C. R. (2021). Genomic variation within the maize stiff-stalk heterotic germplasm pool. *Plant Genome*, *14*, e20114. https://doi.org/10.1002/tpg2.20114

Chawla, H. S., Lee, H., Gabur, I., Vollrath, P., Tamilselvan-Nattar-Amutha, S., Obermeier, C., Schiessl, S. V., Song, J. -M., Liu, K., Guo, L., Parkin, I. A. P., & Snowdon, R. J. (2021). Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. *Plant Biotechnology Journal*, *19*, 240–250. https://doi.org/10.1111/pbi.13456

Cleal, K., & Baird, D. M. (2022). Dysgu: Efficient structural variant calling using short or long reads. *Nucleic Acids Research*, *50*, e53. https://doi.org/10.1093/nar/gkac039

Coster, W. d., Rijk, P. d., Roeck, A. d., Pooter, T. d., D'Hert, S., Strazisar, M., Sleegers, K., & van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, *29*, 1178–1187. https://doi.org/10.1101/gr.244939.118

Coster, W. d., Weissensteiner, M. H., & Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nature Reviews Genetics*, *22*, 572–587. https://doi.org/10.1038/s41576-021-00367-3

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE*, *16*, e0257521. https://doi.org/10.1371/journal.pone.0257521

Dierckxsens, N., Li, T., Vermeesch, J. R., & Xie, Z. (2021). A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biology*, *22*, 342. 10.1186/s13059-021-02551-4

English, A. C., Menon, V. K., Gibbs, R., Metcalf, G. A., & Sedlazeck, F. J. (2022). Truvari: Refined structural variant comparison preserves allelic diversity. *BioRxiv*, 2022.02.21.481353.

Fu, Y., Mahmoud, M., Muraliraman, V. V., Sedlazeck, F. J., & Treangen, T. J. (2021). Vulcan: Improved long-read mapping and structural variant calling via dual-mode alignment. *Gigascience*, *10*, giab063. https://doi.org/10.1093/gigascience/giab063

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., La Hoz, J. F. d., Mohiyuddin, M., Wing, R. A., McNally, K. L., Tatarinova, T., Grigoriev, A., Mauleon, R., & Alexandrov, N. (2019). Structural variants in 3000 rice genomes. *Genome Research*, *29*, 870–880. https://doi.org/10.1101/gr.241240.118

Gill, R. A., Scossa, F., King, G. J., Golicz, A. A., Tong, C., Snowdon, R. J., Fernie, A. R., & Liu, S. (2021). On the role of transposable elements in the regulation of gene expression and subgenomic interactions in crop genomes. *Critical Reviews in Plant Sciences*, *40*, 157–189. https://doi.org/10.1080/07352689.2021.1920731

Goel, M., Sun, H., Jiao, W. -B., & Schneeberger, K. (2019). SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, *20*, 277. https://doi.org/10.1186/s13059-019-1911-0

Hall, M. B. (2022). Rasusa: Randomly subsample sequencing reads to a specified coverage. *Journal of Open Source Software*, *7*, 3941. https://doi.org/10.21105/joss.03941

Heller, D., & Vingron, M. (2019). SVIM: Structural variant identification using mapped long reads. *Bioinformatics*, *35*, 2907–2915. https://doi.org/10.1093/bioinformatics/btz041

Heller, D., & Vingron, M. (2020). SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics*, *36*, 5519–5521. https://doi.org/10.1093/bioinformatics/btaa1034

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J., Mueller, L. A., & Saha, S. (2019). An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *BioRxiv*, 767764.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore minion: Delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*, 239. https://doi.org/10.1186/s13059-016-1103-0

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*, 14061. https://doi.org/10.1038/ncomms14061

Jiang, T., Liu, S., Cao, S., Liu, Y., Cui, Z., Wang, Y., & Guo, H. (2021). Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation. *BMC Bioinformatics [Electronic Resource]*, *22*, 552. https://doi.org/10.1186/s12859-021-04422-y

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., & Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, *21*, 189. https://doi.org/10.1186/s13059-020-02107-y

Lee, H., Chawla, H. S., Obermeier, C., Dreyer, F., Abbadi, A., & Snowdon, R. (2020). Chromosome-Scale assembly of winter oilseed rape *Brassica napus*. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00496

Lemay, M. -A., Sibbesen, J. A., Torkamaneh, D., Hamel, J., Levesque, R. C., & Belzile, F. (2022). Combined use of Oxford Nanopore and illumina sequencing yields insights into soybean structural variation biology. *BMC Biology*, *20*, 53. https://doi.org/10.1186/s12915-022-01255-w

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biology*, *20*, 246. https://doi.org/10.1186/s13059-019-1828-7

Meyers, L. A., & Levin, D. A. (2006). On the abundance of polyploids in flowering plants. *Evolution; Internation Journal of Organic Evolution*, *60*, 1198–1206.

Ren, J., & Chaisson, M. J. P. (2021). lra: A long read aligner for sequences and contigs. *PLOS Computational Biology*, *17*, e1009078. https://doi.org/10.1371/journal.pcbi.1009078

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology*, *14*, 405. https://doi.org/10.1186/gb-2013-14-6-405

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018a). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, *19*, 329–346. https://doi.org/10.1038/s41576-018-0003-4

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Haeseler, A. v., & Schatz, M. C. (2018b). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*, 461–468. https://doi.org/10.1038/s41592-018-0001-7

Song, J. -M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., Xie, W.-Z., Cheng, Y., Zhang, Y., Liu, K., Yang, Q.-Y., Chen, L.-L., & Guo, L. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nature Plants*, *6*, 34–45. https://doi.org/10.1038/s41477-019-0577-7

Tao, Y., Zhao, X., Mace, E., Henry, R., & Jordan, D. (2019). Exploring and exploiting Pan-genomics for crop improvement. *Molecular Plant*, *12*, 156–169. https://doi.org/10.1016/j.molp.2018.12.016

Tham, C. Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M. J., Koh, B. T., Wang, W., Ng, C. H., Chng, W. J., Thiery, A., Tenen, D. G., & Benoukraf, T. (2019). NanoVar: Accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *BioRxiv*, 662940.

Udall, J. A., Quijada, P. A., & Osborn, T. C. (2005). Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of *Brassica napus* L. *Genetics*, *169*, 967–979. https://doi.org/10.1534/genetics.104.033209

Vollrath, P., Chawla, H. S., Schiessl, S. V., Gabur, I., Lee, H., Snowdon, R. J., & Obermeier, C. (2021). A novel deletion in FLOWERING LOCUS t modulates flowering time in winter oilseed rape. *Theoretical and Applied Genetics*, *134*, 1217–1231. https://doi.org/10.1007/s00122-021-03768-4

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. -C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. -S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … & Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Yildiz, G., Zanini, S. F., Knight, P., & Golicz, A. A. (2022). *Pangenomics in agriculture*. *CABI Biotechnology Series*. CABI.

Yuan, Y., Bayer, P. E., Batley, J., & Edwards, D. (2021). Current status of structural variation studies in plants. *Plant Biotechnology Journal*, *19*, 2153–2163. https://doi.org/10.1111/pbi.13646

Zanini, S. F., Bayer, P. E., Wells, R., Snowdon, R. J., Batley, J., Varshney, R. K., Nguyen, H. T., Edwards, D., & Golicz, A. A. (2022). Pangenomics in crop improvement—from coding structural variations to finding regulatory variants with pangenome graphs. *Plant Genome*, *15*, e20177. https://doi.org/10.1002/tpg2.20177

Zhang, F., Xue, H., Dong, X., Li, M., Zheng, X., Li, Z., Xu, J., Wang, W., & Wei, C. (2022). Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Research*, *32*, 853–863.

Zhou, A., Lin, T., & Xing, J. (2019). Evaluating nanopore sequencing data processing pipelines for structural variation identification. *Genome Biology*, *20*, 237. https://doi.org/10.1186/s13059-019-1858-1

SUPPORTING INFORMATION

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

How to cite this article: Yildiz, G., Zanini, S. F., Afsharyan, N. P., Obermeier, C., Snowdon, R. J., & Golicz, A. A. (2023). Benchmarking Oxford Nanopore read alignment-based insertion and deletion detection in crop plant genomes. *The Plant Genome*, *16,* e20314. https://doi.org/10.1002/tpg2.20314

4. Graphical pangenomics-enabled characterization of structural variant impact on gene expression in *Brassica napus*

Gözde Yildiz, Silvia F. Zanini, Sven Weber, Venkataramana Kopalli, Tobias Kox, Amine Abbadi, Rod J. Snowdon, and Agnieszka A. Golicz

Theor Appl Genet 138, 91 (2025).

https://doi.org/10.1007/s00122-025-04867-2

Theoretical and Applied Genetics (2025) 138:91
https://doi.org/10.1007/s00122-025-04867-2

ORIGINAL ARTICLE

Graphical pangenomics-enabled characterization of structural variant impact on gene expression in *Brassica napus*

Gözde Yildiz[1] · Silvia F. Zanini[1] · Sven Weber[2] · Venkataramana Kopalli[1] · Tobias Kox[3] · Amine Abbadi[3] · Rod J. Snowdon[2] · Agnieszka A. Golicz[1]

Received: 19 November 2024 / Accepted: 25 February 2025

Abstract

Key message **Pangenome graphs enable population-scale genotyping and improve expression analysis, revealing that structural variations (SVs), particularly transposable elements (TEs), significantly contribute to gene expression variation in winter oilseed rape.**

Abstract Structural variations (SVs) impact important traits, from yield to flowering behaviour and stress responses. Pangenome graphs capture population-level diversity, including SVs, within a single data structure and provide a robust framework for downstream applications. They have the potential to serve as unbiased references for SV genotyping, pan-transcriptomic analyses, and association studies, offering significant advantages over single reference genomes. However, their full potential for expression quantitative trait locus (eQTL) analysis is yet to be explored. We combined long and short-read whole genome sequencing data with expression profiling of *Brassica napus* (oilseed rape) to assess the impact of SVs on gene expression regulation and explored the utility of pangenome graphs for eQTL analysis. Over 90,000 SVs were discovered from 57 long-read datasets. Pangenome graph as reference was evaluated and used for SV genotyping with short reads and transcript expression quantification. Using SVs genotyped from the graph and 100 expression datasets, we identified 267 gene proximal (cis) SV-eQTLs. Over 70% of eQTL-SVs had similarity to transposable elements (TEs), especially Helitrons. The highest proportion of cis-eQTL-SVs were found in promoter regions. About a third of transcripts whose expression was associated with SVs, had no associated SNPs, suggesting that including SVs allows capturing of relationship which would be missed in SNP-only analyses. This study demonstrated that pangenome graphs provide a unifying framework for eQTL analysis by allowing population-scale SV genotyping and gene expression quantification. We also showed that SVs make an appreciable contribution to gene expression variation in winter oilseed rape.

Communicated by Isobel AP Parkin.

✉ Silvia F. Zanini
silvia.f.zanini@agrar.uni-giessen.de

✉ Agnieszka A. Golicz
agnieszka.golicz@agrar.uni-giessen.de

Gözde Yildiz
goezde.yildiz@agrar.uni-giessen.de

Sven Weber
sven.weber@agrar.uni-giessen.de

Venkataramana Kopalli
venkataramana.kopalli@agrar.uni-giessen.de

Tobias Kox
t.kox@npz.de

Amine Abbadi
A.Abbadi@npz-innovation.de

Rod J. Snowdon
rod.snowdon@agrar.uni-giessen.de

[1] Department of Agrobioinformatics, IFZ Research Center for Biosystems, Land Use and Nutrition, Justus Liebig University, Heinrich Buff Ring 26-32, 35392 Giessen, Germany

[2] Department of Plant Breeding, IFZ Research Center for Biosystems, Land Use and Nutrition, Justus Liebig University, Heinrich Buff Ring 26-32, 35392 Giessen, Germany

[3] NPZ Innovation GmbH, Hohenlieth-Hof, 24363 Holtsee, Germany

Published online: 03 April 2025

Introduction

Structural variations (SVs) are genomic alterations over 50 bp in length, with insertions and deletions representing the most common forms (Alonge et al. 2020; Yildiz et al. 2023). SVs are prevalent in the complex genomes of major crops including wheat (Walkowiak et al. 2020), barley (Jayakodi et al. 2020), and oilseed rape (Chawla et al. 2021). They are associated with key traits such as yield and flowering time in oilseed rape (Song et al. 2020), fruit flavour in tomato (Li et al. 2023), and quality traits in cotton (Jin et al. 2023). SVs can impact gene function by altering protein-coding sequences, splicing patterns, gene expression levels, or any combination thereof (Chiang et al. 2017; Zanini et al. 2022). Expression quantitative trait loci (eQTL) analysis maps associations between genomic variation and gene expression. Results from eQTL studies are often used in conjunction with classical QTL mapping or genome-wide association studies (GWAS) to pinpoint causal or candidate genes (Druka et al. 2010). They can however also be used to help understand the regulatory architecture of gene expression and complex phenotypic traits. The most common variants used in eQTL studies are single nucleotide polymorphisms (SNPs), however, due to increasing capacity for population-scale SV discovery (Alonge et al. 2020; Chawla et al. 2021; Zhang et al. 2022), the impact of SVs on genome-wide expression patterns can now also be investigated in large scale eQTL analyses (Leonard et al. 2024).

Recently, pangenome graphs have emerged as a robust framework for genomic data analysis, capturing species-wide genomic diversity within a single data structure (Yildiz et al. 2022; Zanini et al. 2022; Hu et al. 2024). The main methods for constructing plant pangenomes are de novo assembly and comparison, reference genome-based iterative assembly, and graph-based pangenome approach (Hu et al. 2024). In the de novo assembly method, individual genomes are assembled from scratch to identify shared and unique regions. Subsequent analyses commonly focus on comparing gene annotations across genomes, emphasizing the species' pangene set. The iterative mapping and assembly method starts by aligning reads to an existing reference genome. Reads that don't align are then assembled, and the resulting annotated contigs are integrated into a linear pangenome reference, allowing for the representation of all sequences but compromising on their positional relationships (Golicz et al. 2016; Jain and Garg 2020). The third approach, graph-based pangenomes, represents all genomic sequences and variants as nodes and edges, offering major advantages over reference-based genomes, including:(1) reduced redundancy, by integrating multiple genome sequences into a single graph structure that preserves linear proximity of nodes, even in the presence of complex rearrangements; (2) improved read mapping accuracy and variant detection, by capturing large SVs and unique alleles that may not be represented in single reference genomes or linear pangenomes, and (3) provides a more comprehensive and unbiased reference for association studies (Edwards and Batley 2022).

A necessary prerequisite of association studies is that genomic variations across large populations need to be genotyped accurately and rapidly (Wang et al. 2018; Fuentes et al. 2019). Traditional genotyping methods align short reads to a single reference genome (Alkan et al. 2011; DePristo et al. 2011). However, read alignment errors caused by single reference bias result in inaccurate genotypes, especially for alternative alleles (Cameron et al. 2019). Therefore, graph-based SV genotyping methods using short reads emerged as a powerful alternative (Liu et al. 2020; Lemay et al. 2022; Li et al. 2022; Leonard et al. 2024). Graph-based genotyping algorithms use either read alignment or k-mer matching against the variation/sequence graphs to genotype variants using short reads (Chen et al. 2019; Hickey et al. 2020). However, these methods still have some limitations, being mainly optimized for human genomes, with only limited benchmarking on crop genomes (Lemay et al. 2022; Du et al. 2024). Additionally, crop genomes can present unique challenges for SV genotyping due to their complexity, including differences in genome size, high repeat content, heterozygosity, and polyploidy. Beyond its utility for SV genotyping, pangenome graphs can also be utilized for pan-transcriptomic analyses (Sibbesen et al. 2023), where genomic variation is accounted for during mRNA-Seq read mapping and subsequent quantification.

In this study, we combined long-read Oxford Nanopore (ONT) and short-read Illumina genome sequencing data with mRNA-Seq data from young leaves of *B. napus* (oilseed rape) to assess the impact of SVs on gene expression regulation and explore the utility of pangenome graphs for eQTL analysis in plants. We assessed the effectiveness of graph-based SV genotyping using state-of-the-art approaches and further tested the utility of pangenome graphs for transcript expression quantification. We found that insertions, deletions and especially transposable elements (TEs) contribute to gene expression diversity and that some of the associations could not be detected using only SNPs, highlighting the importance of integrating SVs in association studies to understand the impact of different types of mutations on crop traits.

Materials and methods

Material selection

A total of 100 genetically diverse, elite inbred winter oilseed rape breeding lines from the commercial breeding

programme of Norddeutsche Pflanzenzucht HG Lembke (NPZ KG, Hohenlieth, Germany) were used in the study. All 100 lines were used for short-read sequencing. Based on genetic diversity analysis using genome-wide SNPs called from the short-read data, a subset of 57 lines representing the total genetic diversity of the full collection was selected for long-read sequencing. Single plants from each inbred line were harvested for the short and long-read sequencing, respectively.

Short-read genomic and RNA-Seq sample preparation and sequencing

Plants were grown in a climate-controlled growth chamber with 16-h day (16 °C) and 8-h night (12 °C). Leaf samples were harvested simultaneously for all genotypes after 30 days at the 5–6 leaf stage, immediately shock-frozen in liquid nitrogen, and stored at − 80 °C until DNA/RNA extraction. Leaf material was then ground to a fine powder in liquid nitrogen and separated into aliquots for DNA and RNA extraction. Total genomic DNA was extracted from each sample using the CTAB extraction method of Doyle (1990). Total RNA was extracted using the RNeasy Mini Kit (Qiagen, Hilden, Germany) and treated using RNase-free DNase (Qiagen, Hilden, Germany) to remove DNA. Quantity and quality of RNA samples were checked using a Fragment Analyzer Automated Capillary Electrophoresis system (Advanced Analytical, Heidelberg, Germany). Equimolar RNA/DNA samples were shipped on dry ice to BGI Tech Solutions (Hong Kong, China) for library preparation and sequencing. Whole-genome DNA sequencing was performed with 150nt paired-end reads on the Illumina Hiseq XTen platform. RNA-Seq was performed on the Illumina HiSeq 4000 platform with 100nt paired-end sequencing.

Long-read genomic sample preparation and sequencing

Plants were grown in the same conditions as for short-read sequencing, leaves were harvested from plants at the 4–6 leaf stage, flash frozen, and ground to a fine powder using a mortar and pestle. High-molecular-weight DNA was isolated and sequenced using a modified protocol from Chawla et al (2021). Briefly, 11 mL of pre-heated lysis buffer (1% w/v PVP40, 1% w/v PVP10, 500 mm NaCl, 100 mm TRIS pH8, 50 mm EDTA, 1.25% w/v SDS, 1% (w/v) $Na_2S_2O_5$, 5 mm $C_4H_{10}O_2S_2$, 1% v/v Triton X-100) were added to 1.2–1.5 g of tissue and incubated for 30 min at 37 °C in a rotator. 11 µl RnaseCocktail (ThermoFisher, ref AM2288) were added and the lysate was incubated in a rotator at 37 °C for 20 min. 110 µl of ProteinaseK (ThermoFisher, ref QS0511) were then added and samples were incubated in a rotator at 37 °C for a further 20 min. 4 mL of 5 M potassium acetate were added to the cooled-down lysate, mixed by inversion 20 times, incubated for 10 min on ice, and pelleted by centrifugation at 4 °C, 4250 g for 10 min.

Finally, magnetic beads were used to recover the HMW-DNA, washed twice with 70% ethanol, and incubated with TE buffer for 10 min at 37 °C to release the DNA from the beads into the buffer. 1 to 3 ug of DNA were used for library preparation with the ligation sequencing kit SQK-LSK109, according to the manufacturer's recommendations, and loaded onto an Oxford Nanopore MinION flow cell for sequencing.

SV calling from long reads

In our previous work, we established optimal combinations of alignment and SV calling methods for low to medium sequencing depths (Yildiz et al. 2023). Long-read datasets (*n*: 57) were aligned against reference *B. napus* genome (Express 617 v1) (Lee et al. 2020) using minimap2 v2.24-r1122 (Li 2018), followed by sorting and indexing of the aligned reads with samtools v1.9 (Li et al. 2009). Subsequently, cuteSV v1.0.13 (Jiang et al. 2020) was used to detect SVs with varying coverages; 5x (13 lines), 10x (25 lines), and 20x (19 lines) designated as –min_support values 3, 5, and 8, respectively. SVs genotyping option (–genotype) was enabled and calls across samples were merged using Jasmine v1.0.2 (Kirsche et al. 2023). Merged SVs were re-genotyped with cuteSV (-lvcf). Variants were further processed to only retain insertions and deletions with genotype missing call rate: < 5%, heterozygous genotype call rate: < 5% and remove variants > 20 kb. This approach was based on our earlier findings (Yildiz et al. 2023), which highlighted insertions and deletions as the most prevalent variant types in *B. napus* and were associated with lowest detection errors. Additionally, heterozygous SVs were excluded from analysis, due to potentially erroneous genotype calls in the highly inbred lines used in the analysis.

SNP calling from short reads

Short reads (*n*: 57) were aligned to the reference genome (Express 617 v1) using bwa-mem2 v2.2.1 (Vasimuddin et al. 2019). SNP calling was performed with bcftools v1.15.1 mpileup –skip-indels –min-MQ 10 (minimum read mapping quality) and bcftools call -mv -Ov –ploidy 2 (Danecek et al. 2021). SNPs were filtered using similar criteria as for SVs calling, retaining variants with genotype missing call rate: < 5%, heterozygous genotype call rate: < 5% and minor allele frequency: > 5%.

Graph-based SV genotyping

We performed graph-based SV genotyping using Paragraph v.2.4a (Chen et al. 2019), vg toolkit: v1.43.0 Giraffe/vg (Hickey et al. 2020; Sirén et al. 2021), and v1.1.8 Ensemble Variant Genotyper (EVG) (Du et al. 2024) on 57 short-read datasets. SV genotyping tools included in the study were selected using several criteria: (1) all of them perform graph-based SV genotyping, (2) Giraffe/vg appears to be the most popular genotyper in literature to date (Liu et al. 2020; Sirén et al. 2021; Li et al. 2023), (3) Paragraph is the best performing genotyper based on benchmarking in soybean (Lemay et al. 2022), (4) EVG combines multiple graph-based SV genotyping algorithms. For Giraffe/vg genotyping short reads were aligned to the pangenome graph using vg giraffe (Sirén et al. 2021). SVs from long reads (*n*: 57) were used in vg autoindex v1.43.0 –workflow giraffe. SVs were genotyped using vg pack and vg call with default parameters (read support with -Q 5, ignore mapping and base quality below 5, -s 5, ignore first and last 5 bp from each read). For Paragraph, SVs from long reads (*n*: 57) were provided along with Express 617 v1 reference genome and genotyping was done using default parameters. For EVG, graph SVs (*n*: 57) and Express 617 v1 reference genome were provided, and genotyping was performed with default parameters.

F1-score calculation

We calculated F1-scores for each variant using SV genotypes obtained from different short-read genotypers and long-read SV genotypes used as the truth set.

$$F1 = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

F1-scores were calculated both from the perspective of alternative (ALT: non-reference) and reference (REF) alleles as it can lead to somewhat different results. For example, for the 57 lines, a variant which in the truth set had alternative allele call in two lines, but was genotyped from short reads as alternative allele in 3 lines, will have an F1-score of 0.8 ((2*TP:2)/(2*TP:2 + FP:1 + FN:0)) for the ALT allele, but F1-score of 0.99 ((2*TP:54)/(2*TP:54 + FP:0 + FN:1)) for the REF allele. Heterozygous calls were treated as missing and not included in the calculation, and variants which had > 20% missing rate were designated F1-score of zero.

Comparison of SV with low and high F1-scores

Properties of SVs with low and high F1-scores were compared with respect to length, location, copy number and initial SV calling accuracy. SVs were considered gene proximal if they were within 1 kb of protein-coding genes. Sequences of all variants were extracted and used as query in a BLAST search (with -blastn -evalue 1e-5 -outfmt 6) against Express 617 v1 genome to see if the sequences corresponding to variants with low F1-score have a higher genome-wide copy number.

Transcript expression quantification from RNA-Seq reads

A pangenome graph was built using vg v1.4.30 (Garrison et al. 2018) autoindex, based on the Express 617 v1 reference genome sequence and using SNPs and SVs which passed the quality control filtering steps described above. RNA-Seq reads were mapped to the graph using vg mpmap. The mappings were passed to rpvg for quantification. For each sample, rpvg outputs quantification results along with haplotype probabilities above a certain threshold. Per-sample results were filtered to retain only haplotypes with the highest probability for each gene. Further, only genes for which the haplotype could be assigned in all samples were retained. Transcripts per million (TPM) values were extracted directly from the rpvg output. Kallisto v0.44.0 (Bray et al. 2016) was used for quantification using transcripts extracted from the Express 617 v1 assembly. TPM values were extracted directly from Kallisto outputs. Transcripts quantified with Kallisto, which could not be assigned a haplotype by rpvg for all samples, were removed prior to comparisons. Pearson and Spearman correlations were calculated for each gene across 50 samples. Transcripts with Pearson correlation below 0.75 were tested for over-representation of SNPs and SVs using a permutation test implemented in regioneR v1.26.1 (Gel et al. 2016). All transcripts quantified by rpvg in all samples were used as a universe for resampling in 100 iterations.

Simulation of RNA-Seq reads from pangenome graph

RNA-Seq reads were simulated from the pangenome graph using a previously described approach (Sibbesen et al. 2023). In short, haplotype-specific transcripts for one of the samples and their corresponding sequences were extracted from the graph. These served as a reference for gene expression quantification using RNA-Seq data for the same sample with RSEM v1.3.3 (Li and Dewey 2011), generating expression levels from paired-end RNA sequencing reads. These were in turn provided to vg sim v1.57 to simulate corresponding expression levels.

eQTL identification

Gene expression quantification was performed using rpvg. Only transcripts with mean TPM ≥ 0.1 and expression ≥ 1 TPM in at least two samples were retained for eQTL analysis. The expression matrix was transformed using inverse normal transformation. Five top principal components (PCs) identified from SNP data and top components identified from expression data using the Elbow method of PCAforQTL v0.1.0 (Zhou et al. 2022) were used as covariates in matrixEQTL v2.3 (Shabalin 2012). eQTL analysis was performed jointly for SNPs and SVs. For comparison between short ($n = 100$, $n = 57$) and long ($n = 57$) reads, prior to eQTL analysis variants were further filtered to remove variants with minor allele frequency < 10%, ensuring that minor allele is found in at least five samples. For the final eQTL analysis ($n = 100$), a more relaxed MAF threshold of 5% (also ensuring that minor allele was found in at least five samples) was applied and lead variants were identified by lowest *p*-value. When SNPs and SVs had equal lowest *p*-value, both were retained.

cis-eQTL variants were defined as +/− 3,000 bp from the target gene body, encompassing promoter (3 kb upstream), transcription start site (TSS), exons, introns and 3 kb downstream from the transcription termination site (TTS).

Throughout the manuscript text, "eQTL", "SV-eQTL" and "SNP-eQTL" refer to a variant-locus pair, while "eQTL-SNP" and "eQTL-SV" refer to the variant only.

Transposable element annotation

Transposable element library for the Express 617 genome assembly was generated using EDTA v2.0.1 (Ou et al. 2019). Transposable elements were annotated using RepeatMasker v4.1.2 (Smit et al. 2015). Sequences of insertions and deletions were extracted and compared against the TE library using BLASTn v2.13.0 (Camacho et al. 2009) (e-value cutoff 1e-5). Top BLAST matches were used to assign SV sequences to TE families (Lemay et al. 2022). EDTA/ classification of Helitrons was further confirmed comparing it to the output of a Helitron-specific annotation tool, Heliano (v 1.2.1) (Li et al. 2024). Positional overlaps were computed with bedtools intersect (v2.30, -f 0.5). Above 90% of EDTA families classified as Helitron were confirmed by positional overlap with Heliano annotations, therefore EDTA annotation was used for downstream analysis.

Arabidopsis homologue identification and GO term annotation

Homologue identification was performed using a previously developed method (Golicz et al. 2021; Sessa et al. 2023). Briefly, protein sequences of *B. napus* transcripts were compared against *Arabidopsis thaliana* proteome database using BLASTp v2.13.0 (e-value cutoff 1e-5). Top BLAST matches of *B. napus* transcripts were identified as homologues. GO annotation was performed by transferring TAIR GO annotation of *A. thaliana* to the *B. napus* homologous genes.

Results

Winter oilseed rape harbours extensive structural variation

Structural variant (SV) discovery was performed using Oxford Nanopore (ONT) long-read sequencing data for 57 lines of German winter oilseed rape. The average coverage for long-read data was 16.5x. Following removal of variants with excessive heterozygous genotype calls (unexpected in highly inbred material and indicative of SV calling/genotyping errors) and excessive genotype missing rate we discovered total of 94,824 structural variations, including 48,396 insertions (INS) and 46,428 deletions (DEL) (Fig. 1A). Deletions averaged 1,745 base pairs in length (median: 561.0), while insertions averaged 1,724 base pairs (median: 727.0). These resulted in a total of 164 Mb of SV space, of which 83 Mb consisted in insertions and as such were not represented in a single reference genome. The length distribution of SVs is represented in Fig. 1B. Regarding their genomic locations, 12.80% of SVs were found within genes (inside), 21.22% of SVs within 1 kb of genes (withinkb), and the majority 65.98% of SVs, in intergenic regions (notwithinkb) (Fig. 1C). Insertions and deletions had similar distribution across genic and non-genic regions (Fig. 1C).

Graph-based approach allows population-scale SV genotyping

SV genotyping from population-level datasets, for example using short Illumina WGS data, is a prerequisite for association analyses. Graph-based SV genotyping from short reads has been shown to be the leading approach, however results from different pipelines vary (Chen et al. 2019; Hickey et al. 2020; Du et al. 2024). We tested three graph-based SV genotyping methods, including Paragraph (Chen et al. 2019), Giraffe/vg (Hickey et al. 2020), and EVG (Du et al. 2024). SVs discovered and genotyped from long reads across 57 samples were used a truth set (Fig. 2A). We then genotyped SVs using short reads derived from the same 57 samples (average coverage 12x). Because matched samples sequenced using ONT and Illumina technology were available, F1-scores could be calculated for each SV individually, checking long- and short-read genotype concordance across samples.

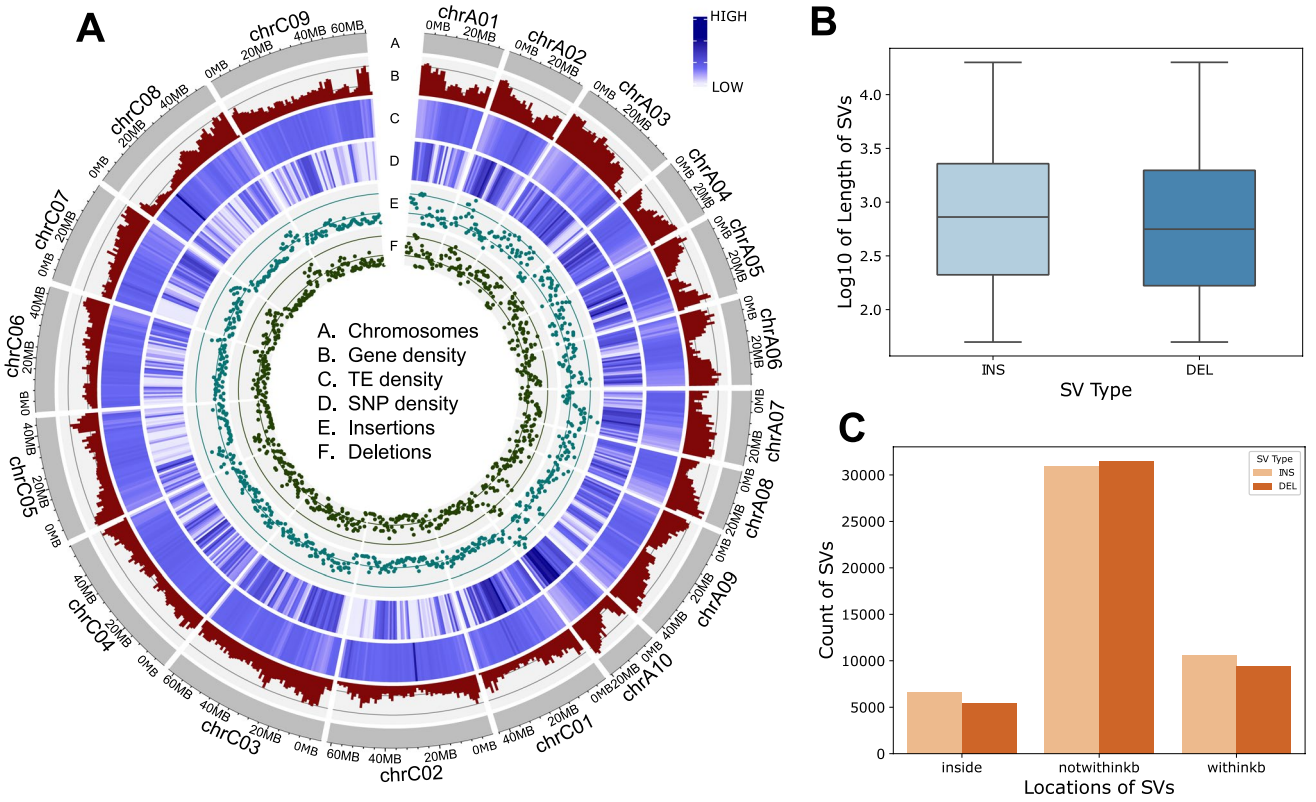


Fig. 1 **A** Distribution of various genomic features of the *B. napus* genome, Chromosomes (**A**), Gene density (**B**), TE density (**C**), SNP density (**D**), Insertions (**E**) and Deletions (**F**). Densities were calculated using 1 Mb window size. **B** Distribution of insertion (INS) and deletion (DEL) lengths. (**C**) Distribution of locations of insertion and deletion SVs relative to annotated genes (inside, not within one kb distance and within one kb distance from genes)

We observed a bimodal distribution of F1-scores (Fig. 2B, C), either close to one (well genotyped, ≥ 0.8) or close to zero (poorly genotyped, ≤ 0.2), in both Paragraph and Giraffe/vg results. However, Giraffe/vg had poorer performance compared to Paragraph, especially for genotyping insertion alternative alleles (Fig. 2B). Giraffe/vg correctly genotyped 34.14% of SVs (SVs with REF allele F1 ≥ 0.8 AND ALT allele F1 ≥ 0.8), while 45.27% of SVs (SVs with REF F1 ≤ 0.2 OR ALT F1 ≤ 0.2) were incorrectly genotyped (Fig. 2B). Paragraph correctly genotyped 43.48% of SVs (SVs with REF allele F1 ≥ 0.8 AND ALT allele F1 ≥ 0.8), while 36.02% of SVs (SVs with REF F1 ≤ 0.2 OR ALT F1 ≤ 0.2) were incorrectly genotyped (Fig. 2C). Overall, Paragraph had a more balanced performance especially for genotyping insertions for both reference and alternative alleles (Fig. 2C). We did not observe improved performance with EVG, likely because of lack of agreement between different genotyping methods. Our results are concordant with previous findings in soybean (Lemay et al. 2022), suggesting that Paragraph is the best performing short-read graph-based genotyper also for *B. napus*. Consequently, we selected the Paragraph results for further analysis.

Variants with good and poor genotyping outcomes have different features

We further explored the reasons behind differences in F1-scores for Paragraph genotyped SVs, to understand why some variants can be genotyped with short reads while others cannot. Variants with high F1-scores for both alleles (REF allele F1 ≥ 0.8 AND ALT allele F1 ≥ 0.8) were considered correctly genotyped, while those with low F1 -scores (SVs with REF F1 ≤ 0.2 OR ALT F1 ≤ 0.2) were considered incorrectly genotyped (Fig. 2C). The correctly genotyped SVs were longer (mean: 1834.25 bp and median: 776.0 bp) compared to incorrectly genotyped SVs (mean: 1522.98 bp and median: 432.0 bp) (Fig. 3A). The correctly genotyped SVs were slightly more likely to be found in proximity of coding genes (34.1% of correctly genotyped SVs were inside or within 1 kb of protein-coding genes, compared to 33.2% for incorrectly genotyped SVs), however the overall distribution of positions relative to genes was very similar for both groups (Fig S1). The incorrectly genotyped SVs on average occurred in a higher copy number (mean: 53.81, median: 5) than the correctly genotyped SVs (mean: 44.21, median:

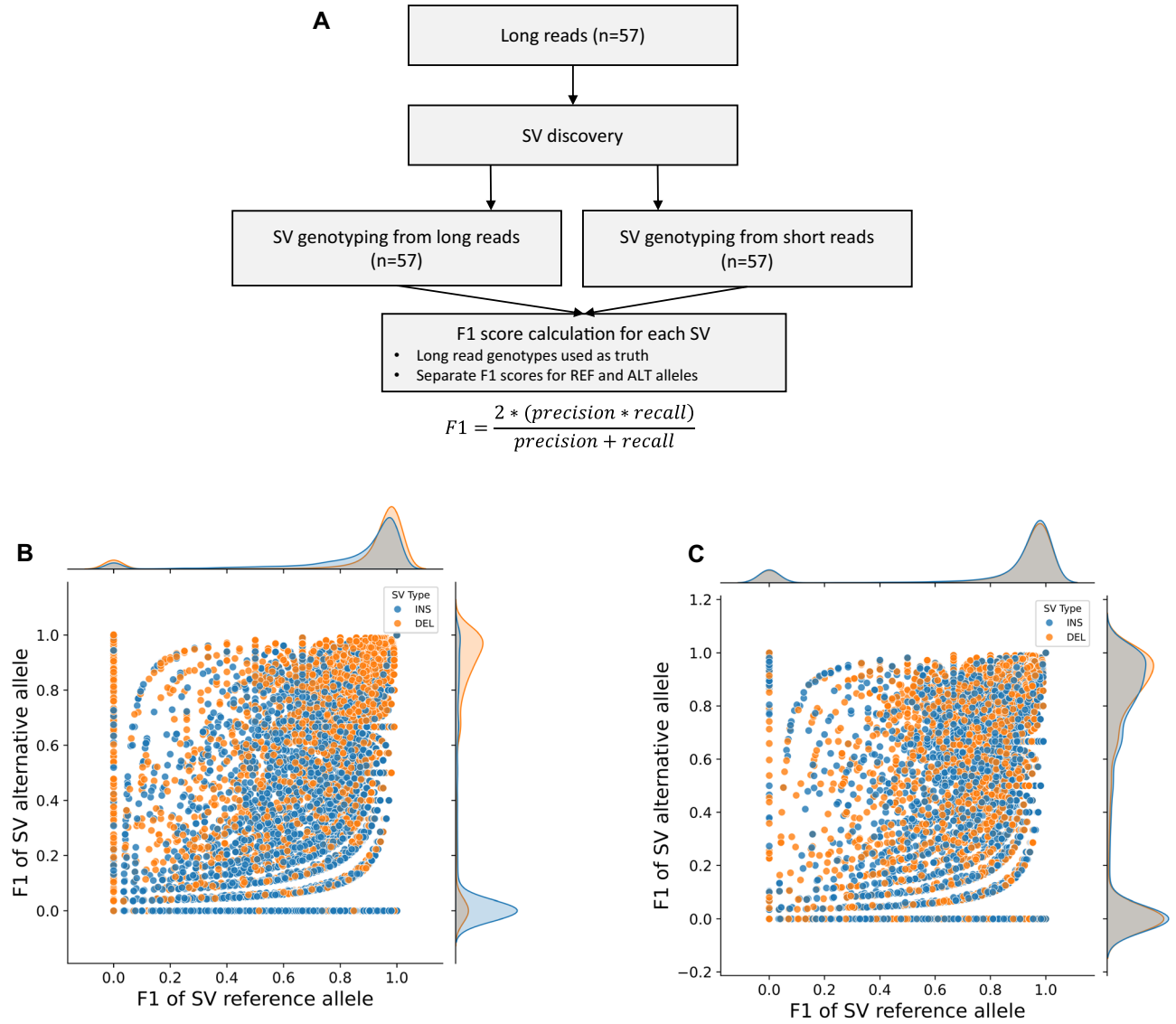


Fig. 2 **A** Procedure for F1-score calculation for genotypes obtained from short reads with Giraffe/vg and Paragraph. **B** F1-scores from Giraffe/vg graph-based genotyping for reference (REF) and alternative (ALT) alleles. **C** F1-scores from Paragraph graph-based genotyping for reference (REF) and alternative (ALT) alleles. F1-scores ≤ 0.2 for either alternate or reference alleles are indicative of poor genotyping outcomes, SVs with F1 ≥ 0.8 for both alternate and reference alleles are considered correctly genotyped

4) (Fig. 3B). We also found that the incorrectly genotyped SVs were associated with higher error during initial SV calling from long reads (Fig. 3C). Specifically, the incorrectly genotyped SVs had wider confidence intervals for positions (CIPOS) and lengths (CILEN) compared to correctly genotyped SVs both for deletions and insertion (Fig. 3C, D).

SV genotyping errors are unlikely to have substantial impact on association studies

To examine whether genotyping errors could affect association studies, we performed eQTL analysis using genotypes derived from 57 long-read samples and 100 short-read samples (which included the 57 samples sequenced with long reads). We compared SV-eQTL variants found within 100 kb of target genes identified in the two analyses, identifying 8,940 eQTL-SVs detected from short reads only, 11,752 eQTL-SVs from long reads only, and 12,409 overlapped eQTL-SVs (Fig. 4B). Two main sources of eQTLs unique to short-read datasets could be either the increased power of the study (57 vs 100 samples) or SV genotyping errors. Importantly, 67.02% eQTL-SVs unique to short-read analysis were determined as correctly genotyped (Fig. 4A), while only 18.67% eQTL-SVs unique to long-read analysis were determined as correctly

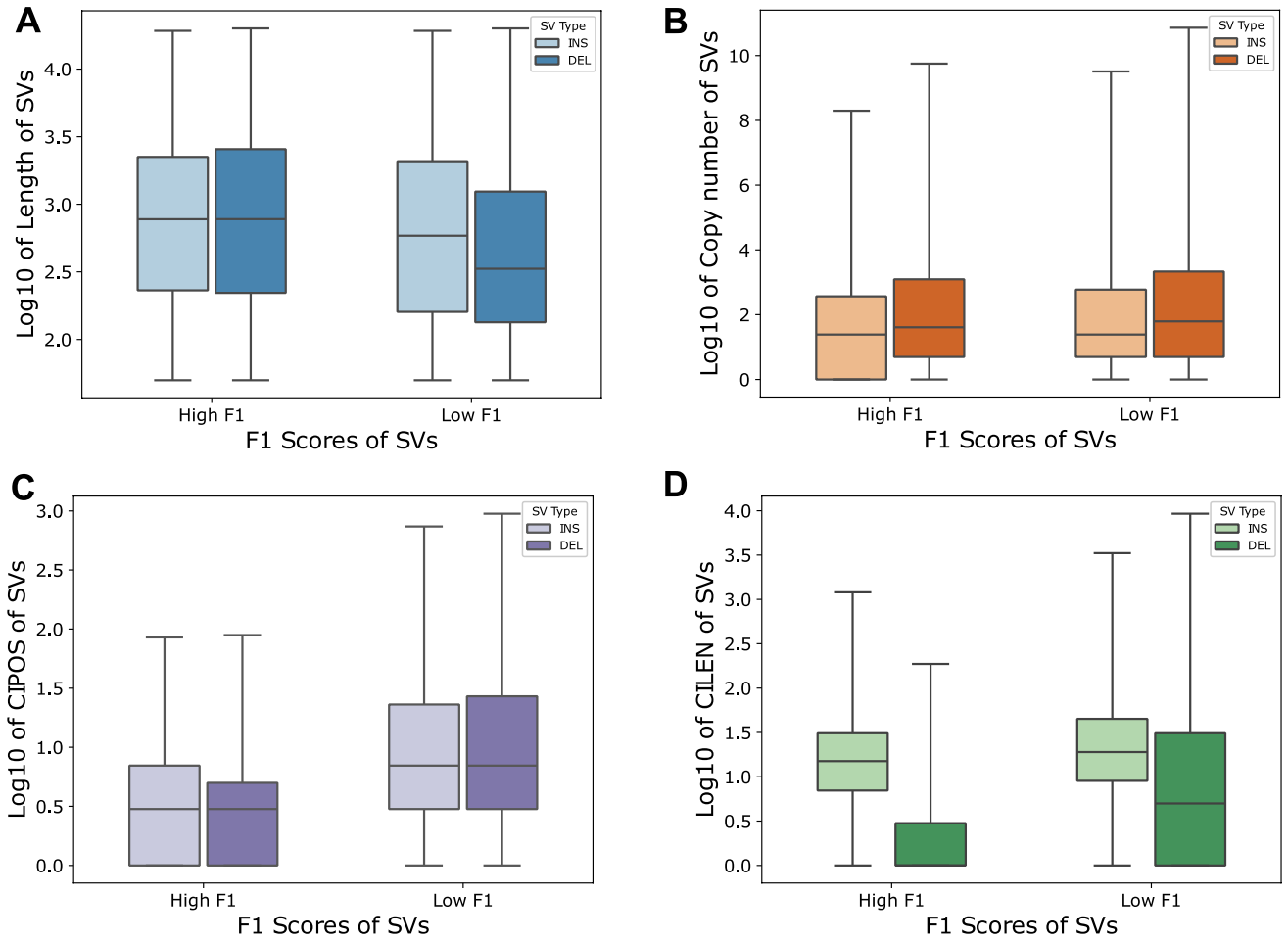


Fig. 3 Comparison of insertion and deletion SV types with low and high F1-scores. F1-scores ≤ 0.2 are indicative of poor genotyping outcomes (Low F1), SVs with F1 ≥ 0.8 are considered correctly genotyped (High F1). **A** Comparison of length of SVs with low and high F1-score. **B** Comparison of copy number of SVs with low and high F1-score. **C** Comparison of confidence interval for position (CIPOS) of SVs with low and high F1-score. **D** Comparison of confidence interval for length (CILEN) of SVs with low and high F1-score

genotyped (Fig. 4C). Overlapping eQTL-SVs from long and short reads were largely correctly genotyped: 86.69% eQTL-SVs (Fig. 4B, Fig S2). These results suggest that the eQTL-SVs identified from short reads only were mostly correctly genotyped and represented new associations found due to the increased sample size and corresponding increased power of the analysis. Conversely, eQTL-SVs found from long reads only could not be correctly genotyped from short reads and would therefore be missing from short-read only analyses. To assess the impact of the increased sample size in the short-read analysis, we downsampled the 100 short-read samples to 57 and compared variants found from the same sample size but different sequencing approaches (Fig S3). Downsampled short-read-based analysis resulted in fewer associations overall compared to the original 100 datasets and fewer associations found in SVs genotyped from short reads only.

Genomic variation affects transcript expression quantification

One of the key steps in eQTL analysis is the accurate quantification of gene expression. Sequence variation between reference genomes and the actual genotypes used for the generation of expression data can lead to quantification errors, a phenomenon often referred to as 'reference sequence bias' (Sibbesen et al. 2023). To test the effect of potential bias on gene expression quantification in *Brassica napus* and its impact on eQTL analysis, we compared transcript abundance derived from a linear reference based (Kallisto) and a pangenome graph-based (rpvg) approaches. The pangenome graph reference was constructed using SVs identified and genotyped from long-read data combined with SNPs called from short reads. For each transcript, we calculated the correlation between read counts estimated by the two methods across 50 samples

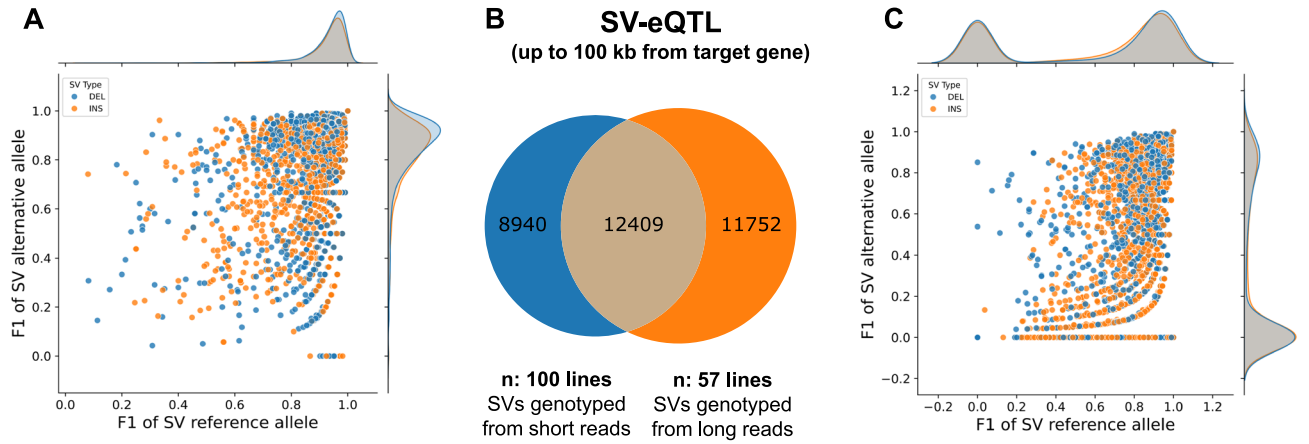


Fig. 4 **A** Distribution of F1-scores of SVs genotyped from short reads, which were unique to SV-eQTL analysis with short-read-derived genotypes. Most of the SVs are correctly genotyped, suggesting that additional associations results from increased power ($n=100$ for short-read genotypes vs $n=57$ for long-read genotypes). **B** Overlap between eQTL-SVs discovered using genotyping with short ($n=100$) and long ($n=57$) reads. **C** Distribution of F1-scores of SVs genotyped from short reads for eQTL-SVs unique to analysis with long-read-derived genotypes. F1-scores ≤ 0.2 for either alternate or reference alleles are indicative of poor genotyping outcomes, SVs with F1 ≥ 0.8 for both alternate and reference alleles are considered correctly genotyped

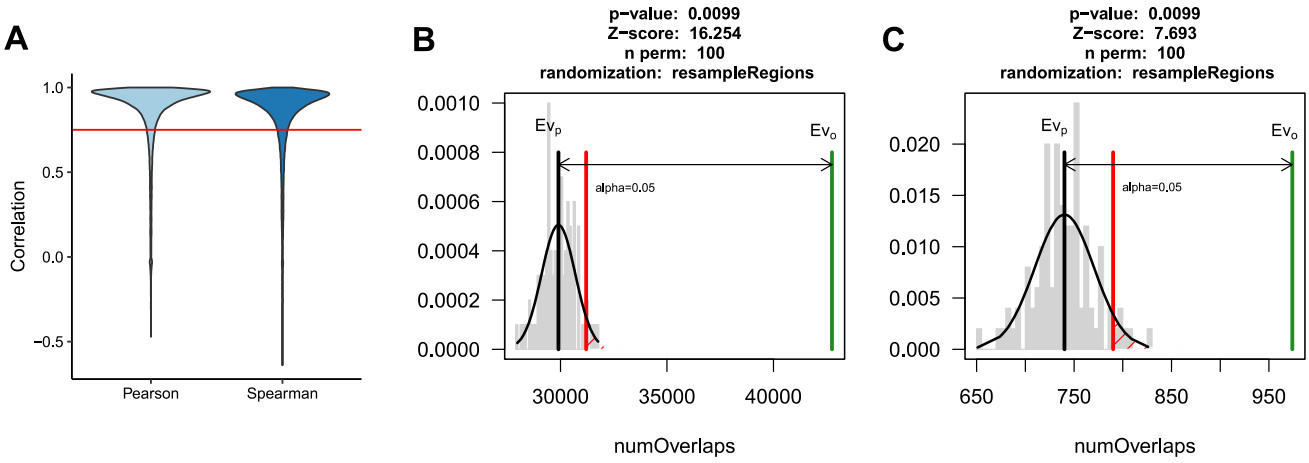


Fig. 5 Comparison of linear reference and graph transcript expression quantification approaches. Transcripts with low concordance between Kallisto and RPVG results and overrepresented in genomic variants. **A** Pearson and Spearman correlation between Kallisto and RPVG quantification across 50 samples. Red line—0.75 cutoff used to define transcripts tested for over-representation of variants. Permutation test results: **B** Transcripts with correlation coefficient below 0.75 are significantly overrepresented in SNPs and **C** Transcripts with correlation coefficient below 0.75 are significantly overrepresented in SVs. Green line—observed value, grey line—mean of permutation results, red line—significance threshold

(Fig. 5A). We then extracted transcripts with Pearson correlation below 0.75 and an equal number of transcripts with the highest correlation coefficients. If genomic variation had an appreciable effect on expression quantification, we would expect transcripts with low measurement concordance across methods to be overrepresented in variants. Indeed, we observed a statistically significant enrichment of variants in transcripts with correlation below 0.75 with a permutation test (Fig. 5B, C). Conversely, the highly correlated transcripts were depleted in variants (Figs S4A-B). A very similar result was obtained when we used transcripts per million (TPM) instead of counts as a measure of expression. To further support our observations, we simulated RNA-Seq reads for one of the samples and compared quantification results between Kallisto and rpvg quantification and expected counts for genes, which have been identified as challenging across 50 samples (correlation below 0.75). We found that the quantification results from rpvg were closer to our simulated ground truth (Fig S5). We concluded that using a pangenome graph reference could improve quantification, therefore rpvg-based expression was selected for the subsequent analysis.

Gene-proximal structural variants are linked to gene expression regulation

Final graph-based eQTL analysis was performed using SNPs, SVs genotyped from short reads and gene expression data representing young leaves at 5–6 leaf stage from 100 homozygous inbred lines (Fig. 6A, Fig S6). In total 39,546 SVs and 2,396,948 SNPs were used in the analysis. We focused the analysis on lead eQTL variants (identified by lowest *p*-value) found in proximity of their target genes (cis-eQTLs). Due to high density of genes in the *B. napus* genome (mean distance between adjacent genes is ~ 3,500 bp) we defined cis-eQTL variants as variants located in/overlapping promoter (3,000 bp upstream from the transcription start site (TSS)), exons, introns or regions immediately downstream (3,000 bp downstream from the transcription termination site (TTS) of their target genes). Using these criteria we identified 267 SV- and 5,668 SNP-eQTLs (Supplementary Data). The proportion of SVs among eQTL variants was higher (4.7%) than among all variants used for the analysis (1.6%). For 35.1% of SV-eQTL transcripts, no significant associations between any SNP and the transcript were detected, suggesting that these eQTL-SVs are not in high enough linkage disequilibrium with SNPs to be detected in SNP-only analyses.

Within our datasets, we identified more SNP-eQTLs on the A subgenome compared to the C subgenome, while a higher proportion of SV-eQTLs was found on the C subgenome (Fig S7, Chi-squared test < 0.001).

Majority of cis-eQTL-SVs have similarity to transposable elements

We investigated the distributions of eQTL-SNPs and SVs in relation to transcript feature locations. Compared to SNPs, a higher proportion of eQTL-SV were found in promoters (Fig. 6B, Chi-Square $p < 0.01$). Overall, a high relative prevalence of SVs upstream of the TSS was previously observed and linked to Class II (DNA) transposable element activity, which can perhaps be explained by easier accessibility of these regions (Han et al. 2013; Fuentes et al. 2019). Indeed, we observed that 71% of eQTL-SVs have similarity to DNA transposable elements (Fig. 7A).

We found that 56% of eQTL-SVs have similarity to Class II (DNA) transposons, 15% to Class I (RNA) transposons and 29% had no detectable similarity to TEs identified in the *B. napus* genome. Among the TE-related eQTL-SVs, the most common TE family was Helitron. The proportion of Helitrons among eQTL-SVs was higher than observed genome wide, but similar to all SVs (Fig. 7B, Fig S8). Helitrons were also the most abundant class of TEs found in promoter-located eQTL-SVs (Fig. 7C). Overall, TE insertions had a greater negative impact on gene expression than deletions (Fig. 8A, Fig S9). Together these results suggest that transposable elements, especially Helitrons, contribute to gene expression diversity in *B. napus*.

Previous eQTL studies reported a relationship between effect size (Beta) and allele frequency, with SVs associated with higher effect sizes found at lower frequencies in the population (Uzunović et al. 2019; Castanera et al. 2023). We observed a similar pattern in our data for both SVs (Fig. 8B, Fig S10A-D) and SNPs (Fig S11). These results are in line with the expected deleterious effects of rare alleles (Lye

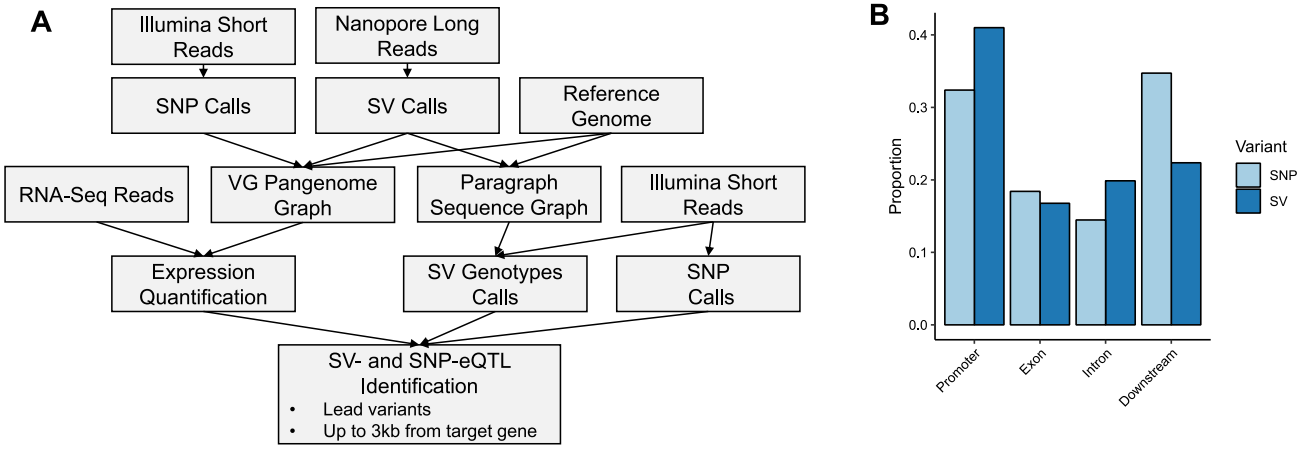


Fig. 6 **A** Procedure for graph-based eQTL analysis: long reads are used for SV identification. SVs identified from long reads along with SNPs identified from short reads are used for graph construction. Graphs are used for transcript expression quantification and SV genotyping using a larger collection of short-read samples. **B** Distribution of eQTLs relative to genomic features highlights a higher proportion of eQTL-SVs in promoter regions compared to genic and downstream regions

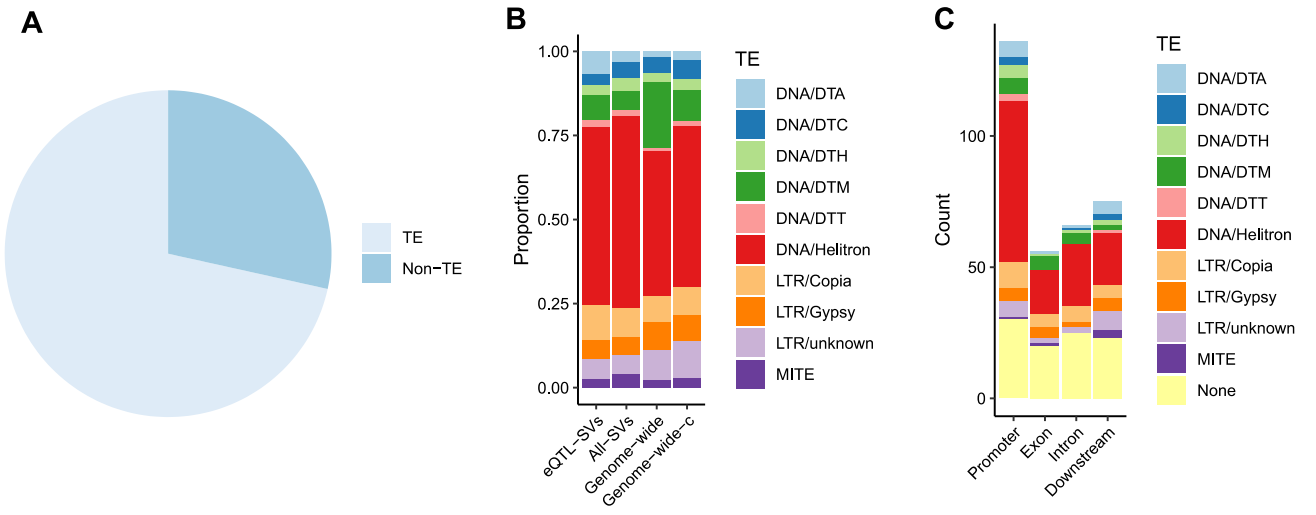


Fig. 7 Similarity of SVs to known transposable elements. **A** Almost 70% of eQTL-SVs variants have similarity to TEs. **B** A high proportion of TE-related eQTL-SVs have similarity to Helitrons compared to Genome-wide (based on counts of TEs annotated by GenomeMasker) and Genome-wide-c (based on counts of TEs annotated by GenomeMasker after merging overlapping elements of the same family). **C** A high number of promoter-associated eQTL-SVs has similarity to Helitrons compared to eQTL-SVs related to other genomic features

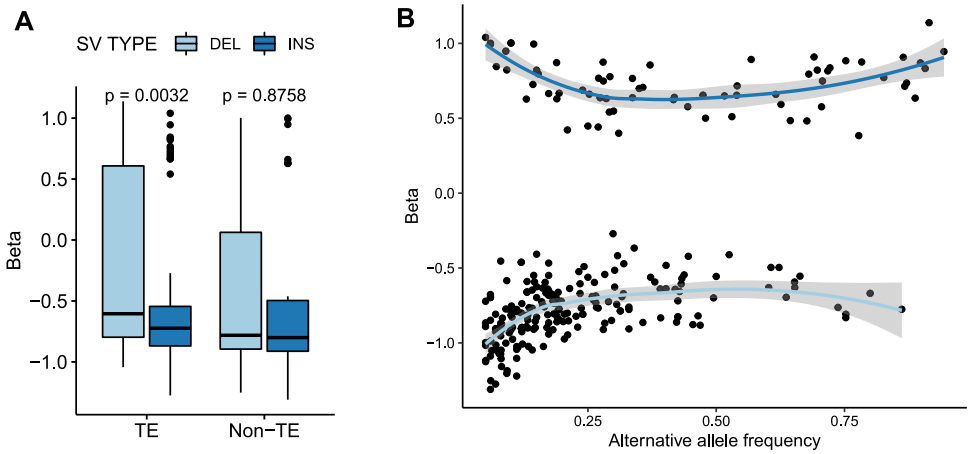


Fig. 8 Effect size of eQTL-SVs. **A** TE insertions are associated with decreased expression compared to deletions. **B** SVs with higher effect size have a lower frequency in the population

et al. 2022) and further support the high quality of our variant and eQTL calls. We observed no significant difference in Beta (Fig S12) and variance (Fig S13) explained by lead eQTL-SNPs and SVs.

Selected examples of genes affected by eQTL-SVs

Out of 259 SV-eQTL transcripts identified, 92% had homologues in the Arabidopsis genome. Gene ontology enrichment analysis did not indicate over-representation in specific processes or functions. However, some transcripts were annotated with functions related to important traits, including stress response (Fig. 9A) and morphogenesis (Fig. 9B). These results suggest that SV-driven gene expression variation could contribute to the phenotypic diversity observed in the field.

Discussion

We used a pangenome graph approach to discover SNPs and SVs associated with differences in gene expression in young leaves of winter oilseed rape. The pangenome graph was used for structural variant genotyping, but also gene expression quantification. We showed that SVs discovered from long reads mostly either genotype correctly from short reads or fail to genotype altogether, and genotyping errors are therefore unlikely to lead to false associations. However, a failure to genotype does reduce the pool of SVs available for association studies, as exemplified here, with approximately half of the initially discovered SVs being successfully genotyped from short reads and included in the downstream eQTL analysis. Failure to genotype was

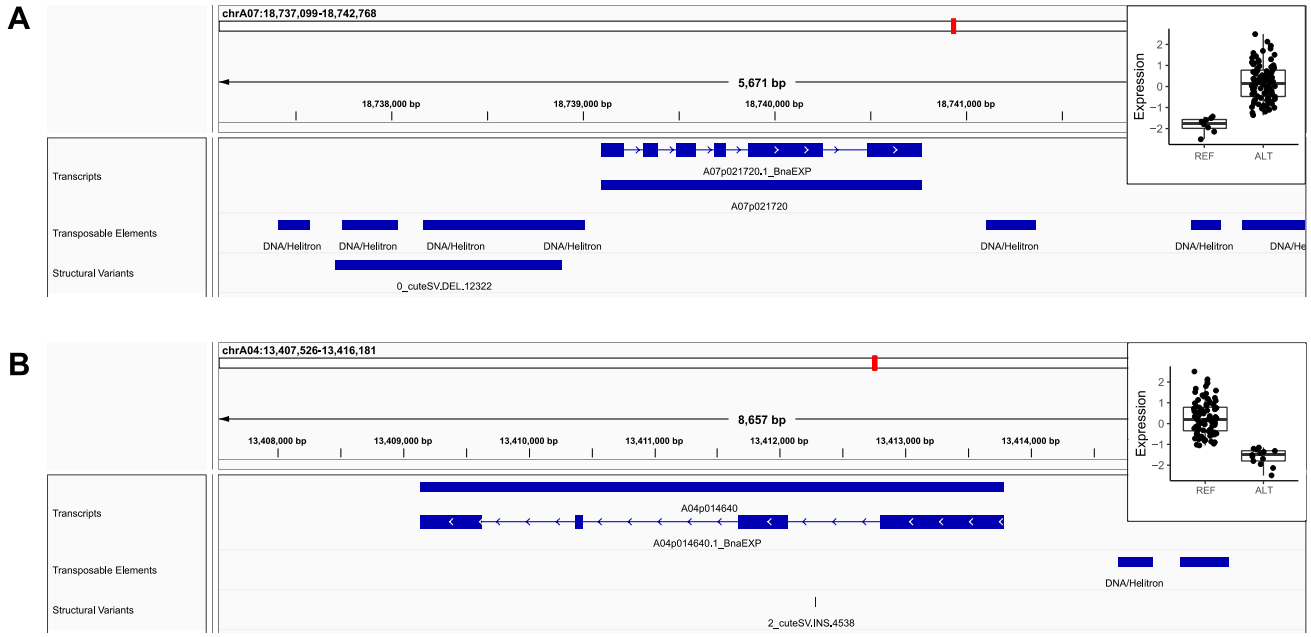


Fig. 9 Example of SVs associated with different gene expression levels. **A** Deletion of Helitron TE in the promoter region of A07p021720 is associated with an increase in gene expression. The corresponding Arabidopsis homologue (SNRK2.8) is known to be involved in response to osmotic stress (TAIR). **B** Insertion of Helitron TE in the first intron of A04p014640 is associated with decreased expression. Arabidopsis homologues (SAW2) is involved in leaf morphogenesis (TAIR). Expression is reported after inverse normal transformation. REF = reference allele, ALT = alternate allele

associated with features such as SV length (shorter SV were more difficult to genotype) and higher uncertainty of the initial SV call. It is important to note that, due to its allotetraploid genome, *B. napus* represents a particularly challenging case of SV genotyping and graph construction, as tools for these analyses were predominantly tested on diploids. Our results are in line with previous reports that SV genotyping had lower performance in paleopolyploid soybean compared to diploids, likely due to ambigious mappings of short reads across sub-genomes (Chen et al. 2019; Sirén et al. 2021; Lemay et al. 2022; Du et al. 2024).

We used medium sequencing coverages (> 10x) for both short and long reads. While this coverage is insufficient for the recent pangenome graphs building pipeline (PGGB, Garrison et al. 2018, Kopalli et al., in preparation), which requires multiple whole genome assemblies, it is appropriate for alignment-based SV discovery combined with VG pangenome graphs. We achieved reasonably good precision and recall at SV calling despite the limited coverages, as anticipated based on our previous study (Yildiz et al. 2023), which demonstrated the identification of SVs using medium-depth (5x-20x) Oxford Nanopore reads. The main limitation of using mid-coverage ONT data is a limited accuracy in the identification of exact break points and deriving consensus sequences, which appears to be reflected in genotyping results. Even using a graph-based approach, SVs with less confident breakpoints and insertions were more difficult to genotype. With the latest improvement in sequencing technologies, resulting in reduced error rates for both ONT and PacBio long reads, high quality calls will be achieved even at low to moderate coverages.

We found that the majority of identified *B. napus* eQTL-SVs sequences have similarity to transposable elements. The finding is in line with reports in other crops, including rice and *B. rapa,* where transposable element insertion polymorphisms (TIPs) were shown to contribute to phenotypic and gene expression variation (Cai et al. 2022; Castanera et al. 2023). Approximately 29% of eQTL-SVs were not annotated as TEs in this analysis. While this could be partly due to current limitations of TE detection tools (Loreto et al. 2023), we observed a more negative effect of insertions annotated as TEs compared to non-TE ones (Fig. 8), suggesting that the latter are truly not TE derived.

Compared to eQTL-SNPs, eQTL-SVs are more likely to be found in promoter regions of genes. The preference of certain transposable elements for insertion into open chromatin regions and especially promoters could make them particularity suited for the rewiring of regulatory networks (Fuentes et al. 2019; Cao et al. 2023; Barro-Trastoy and Köhler 2024). In *B. napus,* the highest number of eQTL-SVs had sequence similarity to Helitrons, likely reflecting their overall high abundance in the genome, where they cover approximately 20% of the genome and represent approximately 50% of all annotated TEs. However, many

of SVs with similarity to Helitrons appear to represent TE fragments rather than intact elements. Previous studies confirmed that Brassicas carry a high abundance of Helitrons relative to other tested species (Hu et al. 2019). While retrotransposons are more abundant in centromeric regions, distribution of DNA elements including Helitrons mirrors more closely the distribution of genes (Fig S8), reflecting their potential for altering gene activity. For example, Helitrons have been shown to play important roles in modifying gene regulation in genes involved in endosperm development and response to herbivory (Barro-Trastoy and Köhler 2024). In addition, the higher prevalence of eQTL-SVs upstream of the TSS can perhaps be explained by easier accessibility of these regions resulting in preferential TE insertion (Han et al. 2013; Fuentes et al. 2019).

Overall, TE insertions had a more negative impact on gene expression than TE deletions. This pattern was not observed for non-TE SVs. The presence of TEs is known to be associated with transcription factors activity disruption and increased DNA methylation, which can have a silencing effect on gene expression (Hollister and Gaut 2009). The stronger negative effect of TE insertions suggests that, at least to some extent, epigenetic silencing mechanisms may be at play.

Functional annotation of SV-eQTL transcripts suggests the involvement of some SVs in modulating important biological processes such as stress responses, flowering and morphogenesis. Due to the highly duplicated nature of the *B. napus* genome, owing to whole genome triplication in the ancestral species of *Brassica* and a more recent allopolyploidization (Cheng et al. 2014), predicting the impact of SVs on traits is not straightforward, even when associated with gene expression differences. Encouraging examples nevertheless exist. For example, despite the presence of multiple homologues, a deletion within the second intron of a *B. napus* FLOWERING LOCUS T homologue was associated with altered flowering time (Vollrath et al. 2021).

Our study highlights the contribution of structural variations to gene expression regulation and the utility of pangenome graph for eQTL analyses in crops. Even using a moderate sample size ($n = 100$) we identified an appreciable number of SVs associated with differences in gene expression. Expanding the sample size and including additional organs and developmental stages will likely result in the identification of many more SVs affecting gene regulation and, potentially, favourable agronomical traits.

Conclusion

In this study, we combined long- and short-read whole genome sequencing data with expression profiling of *Brassica napus* leaves to assess the impact of structural variants (SVs) on gene expression regulation and explore the utility of pangenome graphs for expression quantitative trait locus (eQTL) mapping. Using the graphical pangenome reference for both expression quantification and SV genotyping, we found that insertions, deletions and especially transposable elements (TEs) contribute to gene expression diversity in *B. napus* and that a high proportion of potentially functionally important SVs are not in linkage disequilibrium with SNPs. These SVs affect expression of genes related to important traits and represent diversity unaccounted for in classical SNP-based analyses, highlighting the still largely untapped potential of SVs in eQTL studies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-025-04867-2.

Author contribution statement GY: performed research, wrote the manuscript. SFZ: assisted in the analysis, co-supervised research, edited the manuscript. SW: performed research. VK: performed research. TK: performed research. AA: provided critical comments. RJS: provided critical comments, edited the manuscript. AAG: conceived research, supervised research, performed research, wrote the manuscript, acquired funding.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Alexander von Humboldt Foundation in the framework of Sofja Kovalevskaja Award to AAG. Data generation and analysis by SV, TK, AA and RJS were supported by grant 031B0187 from the German Federal Ministry of Education and Research (BMBF) within the project BreedPatH and DFG grant 458716530. This project was supported by the LOEWE Start Professorship from the Hessian Ministry of Higher Education, Research, Science and the Arts. VK was supported by GRK 2843 from the German Research Foundation (DFG).This work was supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022) and Justus Liebig University Bioinformatics Core Facility (BCF).

Data availability All raw data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1086556. It is made available under a CC-BY-NC-ND 4.0 International licence. Supplementary Data: eQTL results and corresponding variants can be accessed under: https://osf.io/gfphb/

Declarations

Conflict of interest Author RJS is editor in chief of Theoretical and Applied Genetics.

References

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363–376. https://doi.org/10.1038/nrg2958

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, Levy Y, Harel TH, Shalev-Schlosser G, Amsellem Z, Razifard H, Caicedo AL, Tieman DM, Klee H, Kirsche M, Aganezov S, Ranallo-Benavidez TR, Lemmon ZH, Kim J, Robitaille G, Kramer M, Goodwin S, McCombie WR, Hutton S, van Eck J, Gillis J, Eshed Y, Sedlazeck FJ, van der Knaap E, Schatz MC, Lippman ZB (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182:145-161.e23. https://doi.org/10.1016/j.cell.2020.05.021

Barro-Trastoy D, Köhler C (2024) Helitrons: genomic parasites that generate developmental novelties. Trends Genet. https://doi.org/10.1016/j.tig.2024.02.002

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525–527. https://doi.org/10.1038/nbt.3519

Cai X, Lin R, Liang J, King GJ, Wu J, Wang X (2022) Transposable element insertion: a hidden major source of domesticated phenotypic variation in Brassica rapa. Plant Biotechnol J 20:1298–1310. https://doi.org/10.1111/pbi.13807

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinf 10:421. https://doi.org/10.1186/1471-2105-10-421

Cameron DL, Di Stefano L, Papenfuss AT (2019) Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun 10:3240. https://doi.org/10.1038/s41467-019-11146-4

Cao J, Yu T, Xu B, Hu Z, Zhang X, Theurkauf WE, Weng Z (2023) Epigenetic and chromosomal features drive transposon insertion in Drosophila melanogaster. Nucleic Acids Res 51:2066–2086. https://doi.org/10.1093/nar/gkad054

Castanera R, Morales-Díaz N, Gupta S, Purugganan M, Casacuberta JM (2023) Transposons are important contributors to gene expression variability under selection in rice populations. Elife. https://doi.org/10.7554/eLife.86324

Chawla HS, Lee H, Gabur I, Vollrath P, Tamilselvan-Nattar-Amutha S, Obermeier C, Schiessl SV, Song J-M, Liu K, Guo L, Parkin IAP, Snowdon RJ (2021) Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. Plant Biotechnol J 19:240–250. https://doi.org/10.1111/pbi.13456

Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, Kirsche M, Bentley DR, Schatz MC, Sedlazeck FJ (2019) Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol 20:1–13

Cheng F, Wu J, Wang X (2014) Genome triplication drove the diversification of Brassica plants. Hortic Res 1:14024. https://doi.org/10.1038/hortres.2014.24

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, Battle A, Conrad DF, Hall IM (2017) The impact of structural variation on human gene expression. Nat Genet 49:692–699. https://doi.org/10.1038/ng.3834

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. Gigascience 10:giab008. https://doi.org/10.1093/gigascience/giab008

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. https://doi.org/10.1038/ng.806

Doyle JJ (1990) A rapid total DNA preparation procedure for fresh plant tissue. Focus 12:13–15

Druka A, Potokina E, Luo Z, Jiang N, Chen X, Kearsey M, Waugh R (2010) Expression quantitative trait loci analysis in plants. Plant Biotechnol J 8:10–27. https://doi.org/10.1111/j.1467-7652.2009.00460.x

Du Z-Z, He J-B, Jiao W-B (2024) A comprehensive benchmark of graph-based genetic variant genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. Genome Biol 25:91. https://doi.org/10.1186/s13059-024-03239-1

Edwards D, Batley J (2022) Graph pangenomes find missing heritability. Nat Genet 54:919–920. https://doi.org/10.1038/s41588-022-01099-8

Fuentes RR, Chebotarov D, Duitama J, Smith S, La Hoz JF, de, Mohiyuddin M, Wing RA, McNally KL, Tatarinova T, Grigoriev A, Mauleon R, Alexandrov N, (2019) Structural variants in 3000 rice genomes. Genome Res 29:870–880. https://doi.org/10.1101/gr.241240.118

Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol 36:875–879. https://doi.org/10.1038/nbt.4227

Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R (2016) regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics 32:289–291. https://doi.org/10.1093/bioinformatics/btv562

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IAP, Paterson AH, Pires JC, Sharpe AG, Tang H, Teakle GR, Town CD, Batley J, Edwards D (2016) The pangenome of an agronomically important crop plant Brassica oleracea. Nat Commun 7:13390. https://doi.org/10.1038/ncomms13390

Golicz AA, Allu AD, Li W, Lohani N, Singh MB, Bhalla PL (2021) A dynamic intron retention program regulates the expression of several hundred genes during pollen meiosis. Plant Reprod 34:225–242. https://doi.org/10.1007/s00497-021-00411-6

Han Y, Qin S, Wessler SR (2013) Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. BMC Genom 14:71. https://doi.org/10.1186/1471-2164-14-71

Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B (2020) Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol 21:35. https://doi.org/10.1186/s13059-020-1941-7

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19:1419–1428. https://doi.org/10.1101/gr.091678.109

Hu K, Xu K, Wen J, Yi B, Shen J, Ma C, Fu T, Ouyang Y, Tu J (2019) Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. BMC Bioinf 20:354. https://doi.org/10.1186/s12859-019-2945-8

Hu H, Li R, Zhao J, Batley J, Edwards D (2024) Technological development and advances for constructing and analyzing plant pangenomes. Genome Biol Evol 16:evae081. https://doi.org/10.1093/gbe/evae081

Jain M, Garg R (eds) (2020) Legume genomics: methods and protocols, 1st edn. methods in molecular biology, vol 2107. Springer

US; Imprint Humana, New York, NY. https://doi.org/10.1007/978-1-0716-0235-5

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, Ens J, Zhang X-Q, Angessa TT, Zhou G, Tan C, Hill C, Wang P, Schreiber M, Boston LB, Plott C, Jenkins J, Guo Y, Fiebig A, Budak H, Xu D, Zhang J, Wang C, Grimwood J, Schmutz J, Guo G, Zhang G, Mochida K, Hirayama T, Sato K, Chalmers KJ, Langridge P, Waugh R, Pozniak CJ, Scholz U, Mayer KFX, Spannagl M, Li C, Mascher M, Stein N (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588:284–289. https://doi.org/10.1038/s41586-020-2947-8

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y (2020) Long-read-based human genomic structural variation detection with cuteSV. Genome Biol 21:189. https://doi.org/10.1186/s13059-020-02107-y

Jin S, Han Z, Hu Y, Si Z, Dai F, He L, Cheng Y, Li Y, Zhao T, Fang L, Zhang T (2023) Structural variation (SV)-based pan-genome and GWAS reveal the impacts of SVs on the speciation and diversification of allotetraploid cottons. Mol Plant 16:678–693. https://doi.org/10.1016/j.molp.2023.02.004

Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, Schatz MC (2023) Jasmine and Iris: population-scale structural variant comparison and analysis. Nat Methods 20:408–417. https://doi.org/10.1038/s41592-022-01753-3

Lee H, Chawla HS, Obermeier C, Dreyer F, Abbadi A, Snowdon R (2020) Chromosome-scale assembly of winter oilseed Rape Brassica napus. Front Plant Sci 11

Lemay M-A, Sibbesen JA, Torkamaneh D, Hamel J, Levesque RC, Belzile F (2022) Combined use of Oxford Nanopore and Illumina sequencing yields insights into soybean structural variation biology. BMC Biol 20:53. https://doi.org/10.1186/s12915-022-01255-w

Leonard AS, Mapel XM, Pausch H (2024) Pangenome-genotyped structural variation improves molecular phenotype mapping in cattle. Genome Res 34:300–309. https://doi.org/10.1101/gr.278267.123

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf 12:323. https://doi.org/10.1186/1471-2105-12-323

Li H, Wang S, Chai S, Yang Z, Zhang Q, Xin H, Xu Y, Lin S, Chen X, Yao Z, Yang Q, Fei Z, Huang S, Zhang Z (2022) Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. Nat Commun 13:682. https://doi.org/10.1038/s41467-022-28362-0

Li N, He Q, Wang J, Wang B, Zhao J, Huang S, Yang T, Tang Y, Yang S, Aisimutuola P, Xu R, Hu J, Jia C, Ma K, Li Z, Jiang F, Gao J, Lan H, Zhou Y, Zhang X, Huang S, Fei Z, Wang H, Li H, Yu Q (2023) Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nat Genet 55:852–860. https://doi.org/10.1038/s41588-023-01340-y

Li Z, Gilbert C, Peng H, Pollet N (2024) Discovery of numerous novel Helitron-like elements in eukaryote genomes using HELIANO. Nucleic Acids Res 52:e79–e79. https://doi.org/10.1093/nar/gkae679

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z (2020) Pan-genome of wild and cultivated soybeans. Cell 182:162-176.e13. https://doi.org/10.1016/j.cell.2020.05.023

Loreto ELS, Melo ES de, Wallau GL, Gomes, Tiago M. F. F. (2023) The good, the bad and the ugly of transposable elements annotation tools. Genet Mol Biol 46

Lye Z, Choi JY, Purugganan MD (2022) Deleterious mutations and the rare allele burden on rice gene expression. Mol Biol Evol 39:msac193. https://doi.org/10.1093/molbev/msac193

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, Jiang N, Hirsch CN, Hufford MB (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 20:275. https://doi.org/10.1186/s13059-019-1905-y

Sessa EB, Masalia RR, Arrigo N, Barker MS, Pelosi JA (2023) GOgetter: a pipeline for summarizing and visualizing GO slim annotations for plant genetic data. Appl Plant Sci 11:e11536. https://doi.org/10.1002/aps3.11536

Shabalin AA (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28:1353–1358. https://doi.org/10.1093/bioinformatics/bts163

Sibbesen JA, Eizenga JM, Novak AM, Sirén J, Chang X, Garrison E, Paten B (2023) Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. Nat Methods 20:239–247. https://doi.org/10.1038/s41592-022-01731-9

Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang P-C, Carroll A, Gupta N, Gabriel S, Blackwell TW, Ratan A, Taylor KD, Rich SS, Rotter JI, Haussler D, Garrison E, Paten B (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science 374:abg8871. https://doi.org/10.1126/science.abg8871

Smit AF, Hubley R, Green P (2015) RepeatMasker Open-4.0. 2013–2015

Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, Xie W-Z, Cheng Y, Zhang Y, Liu K, Yang Q-Y, Chen L-L, Guo L (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants 6:34–45. https://doi.org/10.1038/s41477-019-0577-7

Uzunović J, Josephs EB, Stinchcombe JR, Wright SI (2019) Transposable elements are important contributors to standing variation in gene expression in capsella grandiflora. Mol Biol Evol 36:1734–1745. https://doi.org/10.1093/molbev/msz098

Vasimuddin M, Misra S, Li H, Aluru S Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: 2019 IEEE international parallel and distributed processing symposium (IPDPS). IEEE, pp 314–324

Vollrath P, Chawla HS, Schiessl SV, Gabur I, Lee H, Snowdon RJ, Obermeier C (2021) A novel deletion in FLOWERING LOCUS T modulates flowering time in winter oilseed rape. Theor Appl Genet 134:1217–1231. https://doi.org/10.1007/s00122-021-03768-4

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D, Klymiuk V, Byrns B, Gundlach H, Bandi V, Siri JN, Nilsen K, Aquino C, Himmelbach A, Copetti D, Ban T, Venturini L, Bevan M, Clavijo B, Koo D-H, Ens J, Wiebe K, N'Diaye A, Fritz AK, Gutwin C, Fiebig A, Fosker C, Fu BX, Accinelli GG, Gardner KA, Fradgley N, Gutierrez-Gonzalez J, Halstead-Nussloch G, Hatakeyama M, Koh CS, Deek J, Costamagna AC, Fobert P, Heavens D, Kanamori H, Kawaura K, Kobayashi F, Krasileva K, Kuo T, McKenzie N, Murata K, Nabeka Y, Paape T, Padmarasu S, Percival-Alwyn L, Kagale S, Scholz U, Sese J, Juliana P, Singh R, Shimizu-Inatsugi R, Swarbreck D, Cockram J, Budak H, Tameshige T, Tanaka T, Tsuji H, Wright J, Wu J, Steuernagel B, Small I, Cloutier S, Keeble-Gagnère G, Muehlbauer G, Tibbets J, Nasuda S, Melonek J, Hucl PJ, Sharpe AG, Clark M, Legg E,

Bharti A, Langridge P, Hall A, Uauy C, Mascher M, Krattinger SG, Handa H, Shimizu KK, Distelfeld A, Chalmers K, Keller B, Mayer KFX, Poland J, Stein N, McCartney CA, Spannagl M, Wicker T, Pozniak CJ (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283. https://doi.org/10.1038/s41586-020-2961-x

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann J-C, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49. https://doi.org/10.1038/s41586-018-0063-9

Yildiz G, Zanini SF, Afsharyan NP, Obermeier C, Snowdon RJ, Golicz AA (2023) Benchmarking Oxford Nanopore read alignment-based insertion and deletion detection in crop plant genomes. Plant Genome 16:e20314. https://doi.org/10.1002/tpg2.20314

Yildiz G, Zanini S, Knight P, Golicz AA (2022) Pangenomics in agriculture. In: Next-generation sequencing and agriculture. CABI GB, pp 163–187

Zanini SF, Bayer PE, Wells R, Snowdon RJ, Batley J, Varshney RK, Nguyen HT, Edwards D, Golicz AA (2022) Pangenomics in crop improvement—from coding structural variations to finding regulatory variants with pangenome graphs. Plant Genome 15:e20177. https://doi.org/10.1002/tpg2.20177

Zhang F, Xue H, Dong X, Li M, Zheng X, Li Z, Xu J, Wang W, Wei C (2022) Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. Genome Res 32:853–863. https://doi.org/10.1101/gr.276015.121

Zhou HJ, Li L, Li Y, Li W, Li JJ (2022) PCA outperforms popular hidden variable inference methods for molecular QTL mapping. Genome Biol 23:210. https://doi.org/10.1186/s13059-022-02761-4

5. Discussion

5.1 Overcoming Barriers in Structural Variation Calling for Crop Genomics

From human genetics to plant breeding, many fields are struggling to keep up with the rapidly growing number of genomes that need to be analyzed. As sequencing costs continue to fall, the number of assembled genomes is expected to reach hundreds of thousands in the coming years. Simply scaling up current bioinformatics approaches won't be sufficient to handle this influx of data. A wide range of sequencing technologies and SV detection tools are now available, each offering its strengths and limitations. Over the years, benchmarking efforts have played a key role in improving these tools, especially when performance is measured using standard metrics like precision and recall (Mahmoud et al. 2019). Benchmarks often rely on simulated data, which allows for controlled evaluation but may not fully represent real-world conditions. To overcome this gap, the development of reliable gold-standard or truth datasets is essential to assess and compare SV detection methods. In addition, standardized and reference databases are critical for leveraging SVs in evolutionary and agricultural studies.

Our benchmarking in Chapter 3 provides a significant contribution by evaluating popular SV calling and mapping pipelines specifically for crop genomes, using both simulated and real datasets from key species like *Brassica napus* (oilseed rape), *Solanum lycopersicum* (tomato), maize, and soybean. Among the aligners tested, minimap2 consistently outperformed others in terms of mapping efficiency and computational speed in plant genomes, which often have larger sizes, higher repeat content, and polyploidy. Among the SV callers, cuteSV and Sniffles2 stood out as the most robust across different coverage levels, with cuteSV showing a slight edge in performance. Notably, even at a sequencing depth of just 5×, both tools achieved F1-scores above 0.8 when paired with minimap2. This is a promising result, especially for large-scale crop research projects, where sequencing hundreds or thousands of accessions would otherwise be prohibitively expensive. These findings suggest that high-accuracy SV detection is feasible without the cost burden of deep sequencing.

The performance of Oxford Nanopore Technology (ONT) in this context is particularly promising, offering a cost-effective platform for high-throughput studies. A notable aspect of our study is the stringent filtering criteria applied during SV calling, particularly the focus on homozygous alternative genotypes. While this approach increases confidence in variant calls, especially in inbred lines typically used in crop research, it also reveals species-specific challenges. For instance, in *B. napus*, a relatively high proportion of heterozygous SVs was detected, even in presumed homozygous samples. We attribute this partly to homeologous exchanges (HEs), a well-documented feature of *B. napus* resulting from its allopolyploid origin. These exchanges can mimic heterozygosity and complicate SV interpretation, highlighting the necessity for species-specific awareness when analyzing SV data. Additionally, we observed that performance varied between simulated and real datasets. Real biological data come with added complexity, such as sequence biases, mapping artifacts, and residual heterozygosity that are not fully captured in simulations. The longer run times and lower mapping precision seen in real datasets reflect these challenges and must be considered when scaling SV pipelines for broader use.

These findings carry important implications. First, the ability to reliably detect insertions and deletions at low sequencing depths opens the door for cost-effective genome-wide association studies (GWAS), genomic selection, and diversity studies in crops. Second, our benchmarking provides clear, practical guidance for researchers: a pipeline combining minimap2 with cuteSV or Sniffles2 offers a solid starting point for SV analysis in crops. This simplifies tool selection and supports more consistent, reliable workflows.

Beyond benchmarking, our study highlights the importance of context-specific tool evaluation. SV detection strategies must be tailored to account for sequencing depth, genome complexity, and species-specific features. As ONT and other long-read platforms become more accessible, particularly in agricultural research, these benchmarking efforts will help ensure that SV analysis keeps pace with evolving technologies. Before SV

calling becomes standard in agricultural research, these challenges need to be addressed. In crop breeding and genomics, the need for standardized formats and metadata is even more urgent to ensure consistency and reliability. The current state of SV calling in agricultural research is similar to where SNP calling was a decade ago—its importance is clear, but the technology and methods are still rapidly evolving. The lack of standardized protocols, benchmarks, and reference databases means that SV analysis requires careful interpretation. With advancements in long-read sequencing and the increasing demand for SV characterization in crop improvement and plant breeding, SV analysis is expected to become a routine part of agricultural genomics in the near future.

5.2 Harnessing Pangenomic Diversity for Trait Discovery and Plant Breeding

Pangenomics is driving a major paradigm shift in genomic research by providing a more comprehensive framework to represent genetic diversity across populations. Unlike traditional linear reference genomes, pangenome graph structures offer a platform for analyzing both conserved and variable genetic regions, enabling the discovery of previously inaccessible genetic variation. This capability is becoming increasingly critical, particularly for plant species with complex and polyploid genomes.

In this thesis (Chapter 4), we developed a graph-based SVs pangenome approach for *B. napus* by integrating short- and long-read sequencing data. SVs were discovered from long reads and genotyped by short reads allowing for the investigation of their effects on gene expression. Using 57 long read samples were identified 94,824 SVs in winter oilseed rape. The SVs then genotypes in larger number of samples using short reads. Our results demonstrate that thousands of SVs can be reliably genotyped across numerous samples using low- to medium-depth short read sequencing data, without the need for complete genome assemblies. However, it is also important to notice that a proportion of SVs (36.02%) could not be genotyped and remained inaccessible to short read based analysis. A key contribution of our study is the integration of SV data with gene expression profiles using RNA-reads through pangenome graph-based eQTL analysis. We found that certain SVs, particularly those located in promoter regions, significantly

affect gene expression. Interestingly, a large proportion of these SVs were associated with transposable elements (TEs), especially Helitrons. Given their tendency to reside in open chromatin regions, these elements may influence gene regulation via epigenetic or structural mechanisms. However, approximately 29% of the expression-associated SVs lacked annotation for known TEs, suggesting that these variants may either involve currently uncharacterized elements or function through alternative mechanisms.

Our findings highlight that SVs not only contribute to sequence-level variation but may also play an important role in the evolution of gene regulatory networks. The polyploid nature of the *B. napus* genome introduces additional complexity to SV genotyping. Ambiguous alignments within homoeologous regions often reduce genotyping accuracy, underscoring the need for SV detection and pangenome analysis tools specifically designed for polyploid species. Most existing tools were originally developed for diploid organisms and exhibit limited performance in polyploid systems. This points to a critical need for new bioinformatics solutions tailored to pangenome-based SV–eQTL studies in plant genomes.

Looking forward, improving the accuracy and sensitivity of SV detection and genotyping remains essential. Enhancements in breakpoint resolution, graph granularity, and alignment algorithms will enable more precise and reliable analyses. Additionally, integrating multi-omics datasets, including methylation profiles, chromatin accessibility, and transcriptomics with pangenome structures, will provide deeper insights into gene regulation and functional variation. However, for these analyses to be possible dedicated tools will need to be developed. The findings of this thesis demonstrate that pangenomic approaches are effective not only for variant discovery but also for functional genomic analyses such as SV–eQTL mapping. Due to their cost-effectiveness and scalability, pangenome-based SV analysis holds strong promise for elucidating gene regulation and informing genome-based breeding strategies in the future.

5.3 Conclusion

This thesis illustrates both the promise and the limitations of using graph-based pangenomes for SV genotyping and eQTL mapping in crop species. Our results support the integration of long- and short-read data through graph-based methods to uncover regulatory variants often missed in SNP-focused analyses. Despite challenges related to polyploidy and mid-range coverage, we were able to identify a substantial number of biologically relevant SVs. Even with moderate sequencing coverages (>10×), our pipeline was able to achieve robust SV detection and eQTL mapping. Future studies incorporating more tissues, developmental stages, and environmental conditions will be essential to fully understand how SVs shape the regulatory architecture of crop genomes.

6. Summary

This thesis addresses key challenges and opportunities in structural variation (SV) detection, genotyping, and downstream analyses using pangenome variation graphs for *Brassica napus*. By integrating different sequencing technologies and available bioinformatics tools, it demonstrates strategies to optimize the analysis of complex plant genomes, which are typically large, repetitive, and polyploid.

Popular mapping and SV calling pipelines were evaluated using both simulated and real datasets from major crops, including rapeseed, tomato, maize, and soybean, across low to medium sequencing depths. The results demonstrate the feasibility of cost-effective SV detection, identifying the most efficient aligners and callers that achieve robust performance even at low coverage (≥5×). These findings provide a practical framework for population-scale crop studies, where sequencing costs and coverages are often a limiting factor.

A graph-based pangenome approach was developed by combining long-read SV discovery with pangenome reference, allowing comparison with existing references to assess and reduce reference bias. This strategy enabled the identification of SVs, the construction of graph-based pangenomes, and subsequent SV genotyping in larger populations using short-read data, eliminating the need for costly de novo assemblies. The approach is therefore scalable, accessible, and suitable for high-throughput crop

genomics. Importantly, integration with gene expression data revealed that many SVs, particularly those linked to transposable elements, significantly affect gene regulation and may underlie key agronomic traits.

Overall, this thesis demonstrates that SVs are not only a major source of genetic diversity but also critical drivers of gene regulatory variation in crops. By providing benchmarking guidelines, novel graph-based pipelines, and functional insights into SVs, it lays the foundation for incorporating structural variation into future genome-informed breeding and trait discovery, ultimately supporting the development of more resilient and productive crop varieties.

7. Zusammenfassung

Diese Dissertation befasst sich mit den zentralen Herausforderungen und Chancen bei der Erkennung, Genotypisierung und Analyse struktureller Variationen (SVs) mithilfe von Pangenom-Variationsgraphen in Brassica napus. Durch die Integration verschiedener Sequenzierungstechnologien und verfügbarer bioinformatischer Werkzeuge werden Strategien zur Optimierung der Analyse komplexer Pflanzengenome aufgezeigt, die typischerweise groß, repetitiv und polyploid sind.

Beliebte Mapping- und SV-Calling-Pipelines wurden anhand sowohl simulierter als auch realer Datensätze wichtiger Kulturpflanzen, darunter Raps, Tomate, Mais und Sojabohne, bei niedriger bis mittlerer Sequenziertiefe evaluiert. Die Ergebnisse zeigen die Machbarkeit einer kosteneffizienten SV-Detektion und identifizieren die effizientesten Mapper und Caller, die selbst bei geringer Abdeckung (≥5×) robuste Ergebnisse liefern. Diese Erkenntnisse bieten einen praktischen Rahmen für populationsweite Studien, bei denen Sequenzierkosten und Abdeckung oft limitierende Faktoren sind.

Ein graphbasiertes Pangenom-Verfahren wurde entwickelt, das die Entdeckung von SVs mit Langreads und die Nutzung von Pangenom-Referenzen kombiniert, wodurch Vergleiche mit bestehenden Referenzen möglich sind und Referenz-Bias reduziert werden kann. Diese Strategie ermöglichte die Identifizierung von SVs, den Aufbau

graphbasierter Pangenome sowie die anschließende Genotypisierung von SVs in größeren Populationen mit Kurzread-Daten, ohne dass teure de-novo-Assemblierungen erforderlich waren. Der Ansatz ist somit skalierbar, zugänglich und für die Hochdurchsatz-Genomik geeignet. Die Integration mit Genexpressionsdaten zeigte zudem, dass viele SVs, insbesondere solche, die mit Transposons assoziiert sind, die Genregulation maßgeblich beeinflussen und wichtige agronomische Merkmale bedingen können.

Insgesamt zeigt diese Arbeit, dass SVs nicht nur eine wichtige Quelle genetischer Diversität darstellen, sondern auch entscheidende Treiber der Genregulationsvariation in Kulturpflanzen sind. Durch die Bereitstellung von Benchmarking-Leitlinien, neuartigen graphbasierten Pipelines und funktionellen Einblicken in SVs legt diese Dissertation die Grundlage für die zukünftige Integration struktureller Variationen in die genomgestützte Züchtung und Merkmalsentdeckung und unterstützt damit die Entwicklung widerstandsfähigerer und ertragreicherer Pflanzensorten.

8. References

Abel, Haley J.; Larson, David E.; Regier, Allison A.; Chiang, Colby; Das, Indraniel; Kanchi, Krishna L. et al. (2020): Mapping and characterization of structural variation in 17,795 human genomes. In *Nature* 583 (7814), pp. 83–89. DOI: 10.1038/s41586-020-2371-0.

Aguiar, Vitor R. C.; César, Jônatas; Delaneau, Olivier; Dermitzakis, Emmanouil T.; Meyer, Diogo (2019): Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. In *PLOS Genetics* 15 (4), e1008091.

Alkan, Can; Coe, Bradley P.; Eichler, Evan E. (2011): Genome structural variation discovery and genotyping. In *Nature Reviews Genetics* 12 (5), pp. 363–376. DOI: 10.1038/nrg2958.

Alonge, Michael; Wang, Xingang; Benoit, Matthias; Soyk, Sebastian; Pereira, Lara; Zhang, Lei et al. (2020): Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. In *Cell* 182 (1), 145-161.e23. DOI: 10.1016/j.cell.2020.05.021.

An, Hong; Qi, Xinshuai; Gaynor, Michelle L.; Hao, Yue; Gebken, Sarah C.; Mabry, Makenzie E. et al. (2019): Transcriptome and organellar sequencing highlights the

complex origin and diversification of allotetraploid Brassica napus. In *Nature Communications* 10 (1), p. 2878. DOI: 10.1038/s41467-019-10757-1.

Ashraf, Muhammad F.; Hou, Dan; Hussain, Quaid; Imran, Muhammad; Pei, Jialong; Ali, Mohsin et al. (2022): Entailing the Next-Generation Sequencing and Metabolome for Sustainable Agriculture by Improving Plant Tolerance. In *International Journal of Molecular Sciences* 23 (2). DOI: 10.3390/ijms23020651.

Bartenhagen, Christoph; Dugas, Martin (2013): RSVSim: an R/Bioconductor package for the simulation of structural variations. In *Bioinformatics (Oxford, England)* 29 (13), pp. 1679–1681. DOI: 10.1093/bioinformatics/btt198.

Bayer, Philipp E.; Hurgobin, Bhavna; Golicz, Agnieszka A.; Chan, Chon-Kit Kenneth; Yuan, Yuxuan; Lee, HueyTyng et al. (2017): Assembly and comparison of two closely related Brassica napus genomes. In *Plant Biotechnol J* 15 (12), pp. 1602–1610. DOI: 10.1111/pbi.12742.

Bayer, Philipp E.; Petereit, Jakob; Durant, Éloi; Monat, Cécile; Rouard, Mathieu; Hu, Haifei et al. (2022): Wheat Panache: A pangenome graph database representing presence–absence variation across sixteen bread wheat genomes. In *Plant Genome* 15 (3), e20221. DOI: 10.1002/tpg2.20221.

Belser, Caroline; Istace, Benjamin; Denis, Erwan; Dubarry, Marion; Baurens, Franc-Christophe; Falentin, Cyril et al. (2018): Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. In *Nature plants* 4 (11), pp. 879–887. DOI: 10.1038/s41477-018-0289-4.

Bogaerts Bert; Van den Bossche An; Verhaegen Bavo; Delbrassinne Laurence; Mattheus Wesley; Nouws Stéphanie et al. (2024): Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. In *Journal of Clinical Microbiology* 62 (5), e01576-23. DOI: 10.1128/jcm.01576-23.

Bolognini, Davide; Sanders, Ashley; Korbel, Jan O.; Magi, Alberto; Benes, Vladimir; Rausch, Tobias (2020): VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. In *Bioinformatics (Oxford, England)* 36 (4), pp. 1267–1269. DOI: 10.1093/bioinformatics/btz719.

Cantu, Dario; Vanzetti, Leonardo S.; Sumner, Adam; Dubcovsky, Martin; Matvienko, Marta; Distelfeld, Assaf et al. (2010): Small RNAs, DNA methylation and transposable elements in wheat. In *BMC Genomics* 11, p. 408. DOI: 10.1186/1471-2164-11-408.

Chaisson, Mark J. P.; Sanders, Ashley D.; Zhao, Xuefang; Malhotra, Ankit; Porubsky, David; Rausch, Tobias et al. (2019): Multi-platform discovery of haplotype-resolved structural variation in human genomes. In *Nature Communications* 10 (1), p. 1784. DOI: 10.1038/s41467-018-08148-z.

Chalhoub, Boulos; Denoeud, France; Liu, Shengyi; Parkin, Isobel A. P.; Tang, Haibao; Wang, Xiyin et al. (2014): Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. In *Science (New York, N.Y.)* 345 (6199), pp. 950–953. DOI: 10.1126/science.1253435.

Chen, Kathleen M.; Cofer, Evan M.; Zhou, Jian; Troyanskaya, Olga G. (2019a): Selene: a PyTorch-based deep learning library for sequence data. In *Nature Methods* 16 (4), pp. 315–318. DOI: 10.1038/s41592-019-0360-8.

Chen, Sai; Krusche, Peter; Dolzhenko, Egor; Sherman, Rachel M.; Petrovski, Roman; Schlesinger, Felix et al. (2019b): Paragraph: a graph-based structural variant genotyper for short-read sequence data. In *Genome Biology* 20 (1), p. 291. DOI: 10.1186/s13059-019-1909-7.

Chen, Xiaoyu; Schulz-Trieglaff, Ole; Shaw, Richard; Barnes, Bret; Schlesinger, Felix; Källberg, Morten et al. (2016): Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. In *Bioinformatics (Oxford, England)* 32 (8), pp. 1220–1222. DOI: 10.1093/bioinformatics/btv710.

Chen, Xuequn; Tong, Chaobo; Zhang, Xingtan; Song, Aixia; Hu, Ming; Dong, Wei et al. (2021): A high-quality Brassica napus genome reveals expansion of transposable elements, subgenome evolution and disease resistance. In *Plant Biotechnol J* 19 (3), pp. 615–630. DOI: 10.1111/pbi.13493.

Cheng, Feng; Wu, Jian; Cai, Xu; Liang, Jianli; Freeling, Michael; Wang, Xiaowu (2018): Gene retention, fractionation and subgenome differences in polyploid plants. In *Nature plants* 4 (5), pp. 258–268. DOI: 10.1038/s41477-018-0136-7.

Cleal, Kez; Baird, Duncan M. (2022): Dysgu: efficient structural variant calling using short or long reads. In *Nucleic Acids Res* 50 (9), e53-e53. DOI: 10.1093/nar/gkac039.

Collins, Ryan L.; Brand, Harrison; Karczewski, Konrad J.; Zhao, Xuefang; Alföldi, Jessica; Francioli, Laurent C. et al. (2020): A structural variation reference for medical and population genetics. In *Nature* 581 (7809), pp. 444–451. DOI: 10.1038/s41586-020-2287-8.

Coster, Wouter de; Weissensteiner, Matthias H.; Sedlazeck, Fritz J. (2021): Towards population-scale long-read sequencing. In *Nature reviews. Genetics* 22 (9), pp. 572–587. DOI: 10.1038/s41576-021-00367-3.

Cretu Stancu, Mircea; van Roosmalen, Markus J.; Renkens, Ivo; Nieboer, Marleen M.; Middelkamp, Sjors; Ligt, Joep de et al. (2017): Mapping and phasing of structural variation in patient genomes using nanopore sequencing. In *Nature Communications* 8 (1), p. 1326. DOI: 10.1038/s41467-017-01343-4.

Cubillos, Francisco A.; Coustham, Vincent; Loudet, Olivier (2012): Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. In *Current opinion in plant biology* 15 (2), pp. 192–198.

Darling, Aaron C. E.; Mau, Bob; Blattner, Frederick R.; Perna, Nicole T. (2004): Mauve: multiple alignment of conserved genomic sequence with rearrangements. In *Genome research* 14 (7), pp. 1394–1403. DOI: 10.1101/gr.2289704.

Delcher, A. L.; Kasif, S.; Fleischmann, R. D.; Peterson, J.; White, O.; Salzberg, S. L. (1999): Alignment of whole genomes. In *Nucleic Acids Res* 27 (11), pp. 2369–2376. DOI: 10.1093/nar/27.11.2369.

Deng, Yiwen; Zhai, Keran; Xie, Zhen; Yang, Dongyong; Zhu, Xudong; Liu, Junzhong et al. (2017): Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. In *Science (New York, N.Y.)* 355 (6328), pp. 962–965.

Dolatabadian, Aria; Yuan, Yuxuan; Bayer, Philipp Emanuel; Petereit, Jakob; Severn-Ellis, Anita; Tirnaz, Soodeh et al. (2022): Copy number variation among resistance genes analogues in Brassica napus. In *Genes* 13 (11), p. 2037.

Du, Ze-Zhen; He, Jia-Bao; Jiao, Wen-Biao (2024): A comprehensive benchmark of graph-based genetic variant genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. In *Genome Biology* 25 (1), p. 91. DOI: 10.1186/s13059-024-03239-1.

Ebert, Peter; Audano, Peter A.; Zhu, Qihui; Rodriguez-Martin, Bernardo; Porubsky, David; Bonder, Marc Jan et al. (2021): Haplotype-resolved diverse human genomes and integrated analysis of structural variation. In *Science (New York, N.Y.)* 372 (6537), eabf7117. DOI: 10.1126/science.abf7117.

Ebler, Jana; Ebert, Peter; Clarke, Wayne E.; Rausch, Tobias; Audano, Peter A.; Houwaart, Torsten et al. (2022): Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. In *Nature Genetics* 54 (4), pp. 518–525. DOI: 10.1038/s41588-022-01043-w.

Edger, Patrick P.; Smith, Ronald; McKain, Michael R.; Cooley, Arielle M.; Vallejo-Marin, Mario; Yuan, Yaowu et al. (2017): Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. In *Plant Cell* 29 (9), pp. 2150–2167. DOI: 10.1105/tpc.17.00010.

Edwards, David; Batley, Jacqueline (2022): Graph pangenomes find missing heritability. In *Nature Genetics* 54 (7), pp. 919–920. DOI: 10.1038/s41588-022-01099-8.

Eggertsson, Hannes P.; Kristmundsdottir, Snaedis; Beyter, Doruk; Jonsson, Hakon; Skuladottir, Astros; Hardarson, Marteinn T. et al. (2019): GraphTyper2 enables

population-scale genotyping of structural variation using pangenome graphs. In *Nature Communications* 10 (1), p. 5402.

Escaramís, Geòrgia; Docampo, Elisa; Rabionet, Raquel (2015): A decade of structural variants: description, history and methods to detect structural variation. In *Brief Funct Genomics* 14 (5), pp. 305–314. DOI: 10.1093/bfgp/elv014.

Espinosa, Elena; Bautista, Rocio; Larrosa, Rafael; Plata, Oscar (2024): Advancements in long-read genome sequencing technologies and algorithms. In *Genomics* 116 (3), p. 110842. DOI: 10.1016/j.ygeno.2024.110842.

Ewing, Adam D.; Houlahan, Kathleen E.; Hu, Yin; Ellrott, Kyle; Caloian, Cristian; Yamaguchi, Takafumi N. et al. (2015): Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. In *Nature Methods* 12 (7), pp. 623–630. DOI: 10.1038/nmeth.3407.

Friedt, Wolfgang; Snowdon, Rod (2010): Oilseed rape. In *Oil crops*, pp. 91–126.

Frost, Jennifer M.; Rhee, Ji Hoon; Choi, Yeonhee (2024): Dynamics of DNA methylation and its impact on plant embryogenesis. In *Current opinion in plant biology* 81, p. 102593.

Fullard, John F.; Giambartolomei, Claudia; Hauberg, Mads E.; Xu, Ke; Voloudakis, Georgios; Shao, Zhiping et al. (2017): Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci. In *Hum Mol Genet* 26 (10), pp. 1942–1951. DOI: 10.1093/hmg/ddx103.

Gabur, Iulian; Chawla, Harmeet Singh; Snowdon, Rod J.; Parkin, Isobel A. P. (2019): Connecting genome structural variation with complex traits in crop plants. In *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 132 (3), pp. 733–750. DOI: 10.1007/s00122-018-3233-0.

Garg, Vanika; Bohra, Abhishek; Mascher, Martin; Spannagl, Manuel; Xu, Xun; Bevan, Michael W. et al. (2024): Unlocking plant genetics with telomere-to-telomere genome assemblies. In *Nature Genetics* 56 (9), pp. 1788–1799. DOI: 10.1038/s41588-024-01830-7.

Garrison, Erik; Guarracino, Andrea; Heumos, Simon; Villani, Flavia; Bao, Zhigui; Tattini, Lorenzo et al. (2024): Building pangenome graphs. In *Nature Methods* 21 (11), pp. 2008–2012. DOI: 10.1038/s41592-024-02430-3.

Garrison, Erik; Sirén, Jouni; Novak, Adam M.; Hickey, Glenn; Eizenga, Jordan M.; Dawson, Eric T. et al. (2018): Variation graph toolkit improves read mapping by representing genetic variation in the reference. In *Nature Biotechnology* 36 (9), pp. 875–879. DOI: 10.1038/nbt.4227.

Goel, Manish; Sun, Hequan; Jiao, Wen-Biao; Schneeberger, Korbinian (2019): SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. In *Genome Biology* 20 (1), p. 277. DOI: 10.1186/s13059-019-1911-0.

Golicz, Agnieszka A.; Batley, Jacqueline; Edwards, David (2016): Towards plant pangenomics. In *Plant Biotechnol J* 14 (4), pp. 1099–1105.

Harris, Robert S. (2007): Improved pairwise alignment of genomic DNA: The Pennsylvania State University.

Heller, David; Vingron, Martin (2019): SVIM: structural variant identification using mapped long reads. In *Bioinformatics (Oxford, England)* 35 (17), pp. 2907–2915. DOI: 10.1093/bioinformatics/btz041.

Heller, David; Vingron, Martin (2021): SVIM-asm: structural variant detection from haploid and diploid genome assemblies. In *Bioinformatics (Oxford, England)* 36 (22-23), pp. 5519–5521. DOI: 10.1093/bioinformatics/btaa1034.

Hickey, Glenn; Heller, David; Monlong, Jean; Sibbesen, Jonas A.; Sirén, Jouni; Eizenga, Jordan et al. (2020): Genotyping structural variants in pangenome graphs using the vg toolkit. In *Genome Biology* 21 (1), p. 35. DOI: 10.1186/s13059-020-1941-7.

Hickey, Glenn; Monlong, Jean; Ebler, Jana; Novak, Adam M.; Eizenga, Jordan M.; Gao, Yan et al. (2024): Pangenome graph construction from genome alignments with Minigraph-Cactus. In *Nature Biotechnology* 42 (4), pp. 663–673. DOI: 10.1038/s41587-023-01793-w.

Ho, Steve S.; Urban, Alexander E.; Mills, Ryan E. (2020): Structural variation in the sequencing era. In *Nature reviews. Genetics* 21 (3), pp. 171–189. DOI: 10.1038/s41576-019-0180-9.

Hu, Haifei; Li, Risheng; Zhao, Junliang; Batley, Jacqueline; Edwards, David (2024): Technological Development and Advances for Constructing and Analyzing Plant Pangenomes. In *Genome Biol Evol* 16 (4), evae081. DOI: 10.1093/gbe/evae081.

Hu, Jihong; Chen, Biyun; Zhao, Jing; Zhang, Fugui; Xie, Ting; Xu, Kun et al. (2022): Genomic selection and genetic architecture of agronomic traits during modern rapeseed breeding. In *Nature Genetics* 54 (5), pp. 694–704.

Huang, Li; Liu, Xia; Pandey, Manish K.; Ren, Xiaoping; Chen, Haiwen; Xue, Xiaomeng et al. (2020): Genome-wide expression quantitative trait locus analysis in a recombinant inbred line population for trait dissection in peanut. In *Plant Biotechnol J* 18 (3), pp. 779–790. DOI: 10.1111/pbi.13246.

IGVF Consortium (2024): Deciphering the impact of genomic variation on function. In *Nature* 633 (8028), pp. 47–57. DOI: 10.1038/s41586-024-07510-0.

Jiang, Tao; Liu, Yongzhuang; Jiang, Yue; Li, Junyi; Gao, Yan; Cui, Zhe et al. (2020): Long-read-based human genomic structural variation detection with cuteSV. In *Genome Biology* 21 (1), p. 189. DOI: 10.1186/s13059-020-02107-y.

Jiao, Wen-Biao; Schneeberger, Korbinian (2017): The impact of third generation genomic technologies on plant genome assembly. In *Current opinion in plant biology* 36, pp. 64–70. DOI: 10.1016/j.pbi.2017.02.002.

Kim, Bernard Y.; Gellert, Hannah R.; Church, Samuel H.; Suvorov, Anton; Anderson, Sean S.; Barmina, Olga et al. (2024): Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. In *PLoS biology* 22 (7), e3002697.

Kim, Daehwan; Langmead, Ben; Salzberg, Steven L. (2015): HISAT: a fast spliced aligner with low memory requirements. In *Nature Methods* 12 (4), pp. 357–360.

Kim, Daehwan; Paggi, Joseph M.; Park, Chanhee; Bennett, Christopher; Salzberg, Steven L. (2019): Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. In *Nature Biotechnology* 37 (8), pp. 907–915.

Kliebenstein, Dan (2009): Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. In *Annual review of plant biology* 60 (1), pp. 93–114.

Kurtz, Stefan; Phillippy, Adam; Delcher, Arthur L.; Smoot, Michael; Shumway, Martin; Antonescu, Corina; Salzberg, Steven L. (2004): Versatile and open software for comparing large genomes. In *Genome Biology* 5 (2), R12. DOI: 10.1186/gb-2004-5-2-r12.

Layer, Ryan M.; Chiang, Colby; Quinlan, Aaron R.; Hall, Ira M. (2014): LUMPY: a probabilistic framework for structural variant discovery. In *Genome Biology* 15 (6), R84. DOI: 10.1186/gb-2014-15-6-r84.

Lee, HueyTyng; Chawla, Harmeet Singh; Obermeier, Christian; Dreyer, Felix; Abbadi, Amine; Snowdon, Rod (2020): Chromosome-Scale Assembly of Winter Oilseed Rape Brassica napus. In *Frontiers in Plant Science* 11. Available online at https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2020.00496.

Lee, Wanseon; Plant, Katharine; Humburg, Peter; Knight, Julian C. (2018): AltHapAlignR: improved accuracy of RNA-seq analyses through the use of alternative haplotypes. In *Bioinformatics (Oxford, England)* 34 (14), pp. 2401–2408.

Lemay, Marc-André; Sibbesen, Jonas A.; Torkamaneh, Davoud; Hamel, Jérémie; Levesque, Roger C.; Belzile, François (2022): Combined use of Oxford Nanopore and Illumina sequencing yields insights into soybean structural variation biology. In *BMC Biology* 20 (1), p. 53. DOI: 10.1186/s12915-022-01255-w.

Leonard, Alexander S.; Mapel, Xena M.; Pausch, Hubert (2024): Pangenome-genotyped structural variation improves molecular phenotype mapping in cattle. In *Genome research* 34 (2), pp. 300–309. DOI: 10.1101/gr.278267.123.

Li, Bao; Yang, Qian; Yang, Lulu; Zhou, Xing; Deng, Lichao; Qu, Liang et al. (2023a): A gap-free reference genome reveals structural variations associated with flowering time in rapeseed (Brassica napus). In *Horticulture research* 10 (10), uhad171.

Li, Heng (2018): Minimap2: pairwise alignment for nucleotide sequences. In *Bioinformatics (Oxford, England)* 34 (18), pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191.

Li, Heng; Feng, Xiaowen; Chu, Chong (2020): The design and construction of reference pangenome graphs with minigraph. In *Genome Biology* 21 (1), p. 265. DOI: 10.1186/s13059-020-02168-z.

Li, Meng; Chen, Chunhai; Wang, Haigang; Qin, Huibin; Hou, Sen; Yang, Xukui et al. (2024): Telomere-to-telomere genome assembly of sorghum. In *Scientific Data* 11 (1), p. 835. DOI: 10.1038/s41597-024-03664-8.

Li, Qin; Luo, Tao; Cheng, Tai; Yang, Shuting; She, Huijie; Li, Jun et al. (2023b): Evaluation and screening of rapeseed varieties (Brassica napus L.) suitable for mechanized harvesting with high yield and quality. In *Agronomy* 13 (3), p. 795.

Li, Wei; Chen, Yudong; Wang, Yali; Zhao, Jia; Wang, Yijun (2022): Gypsy retrotransposon-derived maize lncRNA GARR2 modulates gibberellin response. In *The Plant Journal* 110 (5), pp. 1433–1446.

Liu, Shengxue; Li, Cuiping; Wang, Hongwei; Wang, Shuhui; Yang, Shiping; Liu, Xiaohu et al. (2020a): Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. In *Genome Biology* 21 (1), p. 163. DOI: 10.1186/s13059-020-02069-1.

Liu, Shoucheng; Li, Kui; Dai, Xiuru; Qin, Guochen; Lu, Dongdong; Gao, Zhaoxu et al. (2025): A telomere-to-telomere genome assembly coupled with multi-omic data provides insights into the evolution of hexaploid bread wheat. In *Nature Genetics* 57 (4), pp. 1008–1020. DOI: 10.1038/s41588-025-02137-x.

Liu, Yichen Henry; Luo, Can; Golding, Staunton G.; Ioffe, Jacob B.; Zhou, Xin Maizie (2024): Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data. In *Nature Communications* 15 (1), p. 2447. DOI: 10.1038/s41467-024-46614-z.

Liu, Yucheng; Du, Huilong; Li, Pengcheng; Shen, Yanting; Peng, Hua; Liu, Shulin et al. (2020b): Pan-Genome of Wild and Cultivated Soybeans. In *Cell* 182 (1), 162-176.e13. DOI: 10.1016/j.cell.2020.05.023.

Long, Yuexuan; Wendel, Jonathan F.; Zhang, Xianlong; Wang, Maojun (2024): Evolutionary insights into the organization of chromatin structure and landscape of transcriptional regulation in plants. In *Trends in Plant Science* 29 (6), pp. 638–649.

Lu, Kun; Wei, Lijuan; Li, Xiaolong; Wang, Yuntong; Wu, Jian; Liu, Miao et al. (2019): Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. In *Nature Communications* 10 (1), p. 1154. DOI: 10.1038/s41467-019-09134-9.

Luo, Xizhi; Cai, Guoshuai; Mclain, Alexander C.; Amos, Christopher I.; Cai, Bo; Xiao, Feifei (2022): BMI-CNV: a Bayesian framework for multiple genotyping platforms detection of copy number variants. In *Genetics* 222 (4), iyac147.

Lye, Zoe N.; Purugganan, Michael D. (2019): Copy number variation in domestication. In *Trends in Plant Science* 24 (4), pp. 352–365.

Ma, Jianxin; Bennetzen, Jeffrey L. (2004): Rapid recent growth and divergence of rice nuclear genomes. In *Proceedings of the National Academy of Sciences of the United States of America* 101 (34), pp. 12404–12410. DOI: 10.1073/pnas.0403715101.

Ma, Xinwei; Su, Zhao; Ma, Hong (2020): Molecular genetic analyses of abiotic stress responses during plant reproductive development: Oxford University Press UK.

Ma, Yizan; Min, Ling; Wang, Junduo; Li, Yaoyao; Wu, Yuanlong; Hu, Qin et al. (2021): A combination of genome-wide and transcriptome-wide association studies reveals genetic elements leading to male sterility during high temperature stress in cotton. In *The New phytologist* 231 (1), pp. 165–181. DOI: 10.1111/nph.17325.

Mahmoud, Medhat; Gobet, Nastassia; Cruz-Dávalos, Diana Ivette; Mounier, Ninon; Dessimoz, Christophe; Sedlazeck, Fritz J. (2019): Structural variant calling: the long and the short of it. In *Genome Biology* 20 (1), p. 246. DOI: 10.1186/s13059-019-1828-7.

Martin, Guillaume; Istace, Benjamin; Baurens, Franc-Christophe; Belser, Caroline; Hervouet, Catherine; Labadie, Karine et al. (2025): Unravelling genomic drivers of speciation in Musa through genome assemblies of wild banana ancestors. In *Nature Communications* 16 (1), p. 961. DOI: 10.1038/s41467-025-56329-4.

Mason, A. S.; Snowdon, R. J. (2016): Oilseed rape: learning about ancient and recent polyploid evolution from a recent crop species. In *Plant Biol J* 18 (6), pp. 883–892. DOI: 10.1111/plb.12462.

McClintock, B. (1965): Components of Action of the Regulators Spm and Ac. Carnegie Institution of Washington Year Book 64: 527-536.

Moradi, Mohammad Hossein; Mahmodi, Roqiah; Farahani, Amir Hossein Khaltabadi; Karimi, Mohammad Osman (2022): Genome-wide evaluation of copy gain and loss variations in three Afghan sheep breeds. In *Scientific Reports* 12 (1), p. 14286.

Mu, John C.; Mohiyuddin, Marghoob; Li, Jian; Bani Asadi, Narges; Gerstein, Mark B.; Abyzov, Alexej et al. (2015): VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. In *Bioinformatics (Oxford, England)* 31 (9), pp. 1469–1471. DOI: 10.1093/bioinformatics/btu828.

Nattestad, Maria; Schatz, Michael C. (2016): Assemblytics: a web analytics tool for the detection of variants from an assembly. In *Bioinformatics (Oxford, England)* 32 (19), pp. 3021–3023. DOI: 10.1093/bioinformatics/btw369.

Parkin, Isobel A. P.; Gulden, Sigrun M.; Sharpe, Andrew G.; Lukens, Lewis; Trick, Martin; Osborn, Thomas C.; Lydiate, Derek J. (2005): Segmental Structure of the Brassica napus Genome Based on Comparative Analysis With Arabidopsis thaliana. In *Genetics* 171 (2), pp. 765–781. DOI: 10.1534/genetics.105.042093.

Paten, Benedict; Novak, Adam M.; Eizenga, Jordan M.; Garrison, Erik (2017): Genome graphs and the evolution of genome inference. In *Genome research* 27 (5), pp. 665–676. DOI: 10.1101/gr.214155.116.

Pös, Ondrej; Radvanszky, Jan; Buglyó, Gergely; Pös, Zuzana; Rusnakova, Diana; Nagy, Bálint; Szemes, Tomas (2021): DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. In *Biomedical Journal* 44 (5), p. 548.

Purugganan, Michael D.; Jackson, Scott A. (2021): Advancing crop genomics from lab to field. In *Nature Genetics* 53 (5), pp. 595–601.

Qin, Maochun; Liu, Biao; Conroy, Jeffrey M.; Morrison, Carl D.; Hu, Qiang; Cheng, Yubo et al. (2015): SCNVSim: somatic copy number variation and structure variation simulator. In *BMC Bioinformatics* 16 (1), p. 66. DOI: 10.1186/s12859-015-0502-7.

Rausch, Tobias; Zichner, Thomas; Schlattl, Andreas; Stütz, Adrian M.; Benes, Vladimir; Korbel, Jan O. (2012): DELLY: structural variant discovery by integrated paired-end and split-read analysis. In *Bioinformatics (Oxford, England)* 28 (18), i333-i339. DOI: 10.1093/bioinformatics/bts378.

Rautiainen, Mikko; Durai, Dilip A.; Chen, Ying; Xin, Lixia; Low, Hwee Meng; Göke, Jonathan et al. (2020): AERON: Transcript quantification and gene-fusion detection using long reads. In *BioRxiv*, 2020-01.

Rautiainen, Mikko; Marschall, Tobias (2020): GraphAligner: rapid and versatile sequence-to-graph alignment. In *Genome Biology* 21 (1), p. 253.

Raza, Ali; Razzaq, Ali; Mehmood, Sundas Saher; Hussain, Muhammad Azhar; Wei, Su; He, Huang et al. (2021): Omics: The way forward to enhance abiotic stress tolerance in Brassica napus L. In *GM crops & food* 12 (1), pp. 251–281.

Rigal, Mélanie; Becker, Claude; Pélissier, Thierry; Pogorelcnik, Romain; Devos, Jane; Ikeda, Yoko et al. (2016): Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in Arabidopsis thaliana F1 epihybrids. In *Proceedings of the National Academy of Sciences* 113 (14), E2083-E2092. DOI: 10.1073/pnas.1600672113.

Rockman, Matthew V.; Kruglyak, Leonid (2006): Genetics of global gene expression. In *Nature reviews. Genetics* 7 (11), pp. 862–872.

Rousseau-Gueutin, Mathieu; Belser, Caroline; Da Silva, Corinne; Richard, Gautier; Istace, Benjamin; Cruaud, Corinne et al. (2020): Long-read assembly of the Brassica napus reference genome Darmor-bzh. In *GigaScience* 9 (12), giaa137. DOI: 10.1093/gigascience/giaa137.

Sasaki, Takuji; International Rice Genome Sequencing Project (2005): The map-based sequence of the rice genome. In *Nature* 436 (7052), pp. 793–800. DOI: 10.1038/nature03895.

Satam, Heena; Joshi, Kandarp; Mangrolia, Upasana; Waghoo, Sanober; Zaidi, Gulnaz; Rawool, Shravani et al. (2023): Next-Generation Sequencing Technology: Current Trends and Advancements. In *Biology* 12 (7). DOI: 10.3390/biology12070997.

Schaid, Daniel J.; Chen, Wenan; Larson, Nicholas B. (2018): From genome-wide associations to candidate causal variants by statistical fine-mapping. In *Nature reviews. Genetics* 19 (8), pp. 491–504. DOI: 10.1038/s41576-018-0016-z.

Schiessl, Sarah-Veronica; Katche, Elvis; Ihien, Elizabeth; Chawla, Harmeet Singh; Mason, Annaliese S. (2019): The role of genomic structural variation in the genetic improvement of polyploid crops. In *The Crop Journal* 7 (2), pp. 127–140. DOI: 10.1016/j.cj.2018.07.006.

Schmitz, Robert J.; Grotewold, Erich; Stam, Maike (2022): Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. In *Plant Cell* 34 (2), pp. 718–741.

Schmutzer, Thomas; Samans, Birgit; Dyrszka, Emmanuelle; Ulpinnis, Chris; Weise, Stephan; Stengel, Doreen et al. (2015): Species-wide genome sequence and nucleotide polymorphisms from the model allopolyploid plant Brassica napus. In *Scientific Data* 2 (1), p. 150072. DOI: 10.1038/sdata.2015.72.

Secomandi, Simona; Gallo, Guido Roberto; Rossi, Riccardo; Rodríguez Fernandes, Carlos; Jarvis, Erich D.; Bonisoli-Alquati, Andrea et al. (2025): Pangenome graphs and their applications in biodiversity genomics. In *Nature Genetics* 57 (1), pp. 13–26. DOI: 10.1038/s41588-024-02029-6.

Sedlazeck, Fritz J.; Rescheneder, Philipp; Smolka, Moritz; Fang, Han; Nattestad, Maria; Haeseler, Arndt von; Schatz, Michael C. (2018): Accurate detection of complex structural

variations using single-molecule sequencing. In *Nature Methods* 15 (6), pp. 461–468. DOI: 10.1038/s41592-018-0001-7.

Sibbesen, Jonas A.; Eizenga, Jordan M.; Novak, Adam M.; Sirén, Jouni; Chang, Xian; Garrison, Erik; Paten, Benedict (2023): Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. In *Nature Methods* 20 (2), pp. 239–247. DOI: 10.1038/s41592-022-01731-9.

Sibbesen, Jonas Andreas; Maretty, Lasse; Krogh, Anders (2018): Accurate genotyping across variant classes and lengths using variant graphs. In *Nature Genetics* 50 (7), pp. 1054–1059. DOI: 10.1038/s41588-018-0145-5.

Silaiyiman, Saimire; Liu, Jiaxuan; Wu, Jiaxin; Ouyang, Lejun; Cao, Zheng; Shen, Chao (2025): A Systematic Review of the Advances and New Insights into Copy Number Variations in Plant Genomes. In *Plants* 14 (9). DOI: 10.3390/plants14091399.

Slotkin, R. Keith; Martienssen, Robert (2007): Transposable elements and the epigenetic regulation of the genome. In *Nature reviews. Genetics* 8 (4), pp. 272–285. DOI: 10.1038/nrg2072.

Smolka, Moritz; Paulin, Luis F.; Grochowski, Christopher M.; Horner, Dominic W.; Mahmoud, Medhat; Behera, Sairam et al. (2024): Detection of mosaic and population-level structural variants with Sniffles2. In *Nature Biotechnology* 42 (10), pp. 1571–1580. DOI: 10.1038/s41587-023-02024-y.

Song, Jia-Ming; Guan, Zhilin; Hu, Jianlin; Guo, Chaocheng; Yang, Zhiquan; Wang, Shuo et al. (2020): Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. In *Nature plants* 6 (1), pp. 34–45. DOI: 10.1038/s41477-019-0577-7.

Song, Qingxin; Chen, Z. Jeffrey (2015): Epigenetic and developmental regulation in plant polyploids. In *Current opinion in plant biology* 24, pp. 101–109. DOI: 10.1016/j.pbi.2015.02.007.

Stein, Shayna; Lu, Zhi-xiang; Bahrami-Samani, Emad; Park, Juw Won; Xing, Yi (2015): Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. In *Nucleic Acids Res* 43 (22), pp. 10612–10622.

Stevenson, Kraig R.; Coolon, Joseph D.; Wittkopp, Patricia J. (2013): Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. In *BMC Genomics* 14, pp. 1–13.

Sullivan, Katie Marie; Susztak, Katalin (2020): Unravelling the complex genetics of common kidney diseases: from variants to mechanisms. In *Nature reviews. Nephrology* 16 (11), pp. 628–640. DOI: 10.1038/s41581-020-0298-1.

Sun, Xiaopeng; Xiang, Yanli; Dou, Nannan; Zhang, Hui; Pei, Surui; Franco, Arcadio Valdes et al. (2023): The role of transposon inverted repeats in balancing drought tolerance and yield-related traits in maize. In *Nature Biotechnology* 41 (1), pp. 120–127. DOI: 10.1038/s41587-022-01470-4.

Tan, Zengdong; Han, Xu; Dai, Cheng; Lu, Shaoping; He, Hanzi; Yao, Xuan et al. (2024): Functional genomics of Brassica napus: Progress, challenges, and perspectives. In *J. Integr. Plant Biol.* 66 (3), pp. 484–509. DOI: 10.1111/jipb.13635.

Tan, Zengdong; Peng, Yan; Xiong, Yao; Xiong, Feng; Zhang, Yuting; Guo, Ning et al. (2022): Comprehensive transcriptional variability analysis reveals gene networks regulating seed oil content of Brassica napus. In *Genome Biology* 23 (1), p. 233. DOI: 10.1186/s13059-022-02801-z.

Tao, Xiao-Yuan; Feng, Shou-Li; Yuan, Lu; Li, Yan-Jun; Li, Xin-Jia; Guan, Xue-Ying et al. (2025): Harnessing transposable elements for plant functional genomics and genome engineering. In *Trends in Plant Science*. DOI: 10.1016/j.tplants.2025.03.007.

Tao, Yongfu; Zhao, Xianrong; Mace, Emma; Henry, Robert; Jordan, David (2019): Exploring and exploiting pan-genomics for crop improvement. In *Molecular Plant* 12 (2), pp. 156–169.

The Arabidopsis Genome Initiative (2000): Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. In *Nature* 408 (6814), pp. 796–815. DOI: 10.1038/35048692.

The State of Food Security and Nutrition in the World (2024): FAO; IFAD; UNICEF; WFP; WHO.

Tian, Xuehan; Wang, Ruipeng; Liu, Zhenping; Lu, Sifan; Chen, Xinyuan; Zhang, Zeyu et al. (2025): Widespread impact of transposable elements on the evolution of post-transcriptional regulation in the cotton genus Gossypium. In *Genome Biology* 26 (1), p. 60. DOI: 10.1186/s13059-025-03534-5.

Torlay, L.; Perrone-Bertolotti, M.; Thomas, E.; Baciu, M. (2017): Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. In *Brain Informatics* 4 (3), pp. 159–169. DOI: 10.1007/s40708-017-0065-7.

Ventimiglia, Maria; Marturano, Giovanni; Vangelisti, Alberto; Usai, Gabriele; Simoni, Samuel; Cavallini, Andrea et al. (2023): Genome-wide identification and characterization of exapted transposable elements in the large genome of sunflower (Helianthus annuus L.). In *The Plant Journal* 113 (4), pp. 734–748. DOI: 10.1111/tpj.16078.

Wang, Daiqi; Wang, Hongru; Xu, Xiaomei; Wang, Man; Wang, Yahuan; Chen, Hong et al. (2023a): Two complementary genes in a presence-absence variation contribute to indica-japonica reproductive isolation in rice. In *Nature Communications* 14 (1), p. 4531. DOI: 10.1038/s41467-023-40189-x.

Wang, Fulin; Bao, Jiandong; Zhang, Heng; Zhai, Guowei; Song, Tao; Liu, Zhijian et al. (2025a): A telomere-to-telomere genome assembly of Chinese grain sorghum 654. In *Scientific Data* 12 (1), p. 460. DOI: 10.1038/s41597-025-04791-6.

Wang, Kun; Huang, Gai; Zhu, Yuxian (2016): Transposable elements play an important role during cotton genome evolution and fiber cell development. In *Science China. Life sciences* 59 (2), pp. 112–121. DOI: 10.1007/s11427-015-4928-y.

Wang, Xin; Li, Man-Wah; Wong, Fuk-Ling; Luk, Ching-Yee; Chung, Claire Yik-Lok; Yung, Wai-Shing et al. (2021): Increased copy number of gibberellin 2-oxidase 8 genes reduced trailing growth and shoot length during soybean domestication. In *The Plant Journal* 107 (6), pp. 1739–1755.

Wang, Xu; Yan, Ming; Cui, Shanshan; Li, Fang; Zhao, Qingqing; Wang, Qingnan et al. (2025b): Common bean pan-genome reveals abundant variation patterns and relationships of stress response genes and pathways. In *BMC Genomics* 26 (1), p. 495. DOI: 10.1186/s12864-025-11662-2.

Wang, Xufeng; Chen, Qiuyue; Wu, Yaoyao; Lemmon, Zachary H.; Xu, Guanghui; Huang, Cheng et al. (2018): Genome-wide Analysis of Transcriptional Variability in a Large Maize-Teosinte Population. In *Molecular Plant* 11 (3), pp. 443–459. DOI: 10.1016/j.molp.2017.12.011.

Wang, Ya-Hui; Liu, Pei-Zhuo; Liu, Hui; Zhang, Rong-Rong; Liang, Yi; Xu, Zhi-Sheng et al. (2023b): Telomere-to-telomere carrot (Daucus carota) genome assembly reveals carotenoid characteristics. In *Horticulture research* 10 (7), uhad103. DOI: 10.1093/hr/uhad103.

Wenger, Aaron M.; Peluso, Paul; Rowell, William J.; Chang, Pi-Chuan; Hall, Richard J.; Concepcion, Gregory T. et al. (2019): Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. In *Nature Biotechnology* 37 (10), pp. 1155–1162. DOI: 10.1038/s41587-019-0217-9.

Wicker, Thomas; Sabot, François; Hua-Van, Aurélie; Bennetzen, Jeffrey L.; Capy, Pierre; Chalhoub, Boulos et al. (2007): A unified classification system for eukaryotic transposable elements. In *Nature reviews. Genetics* 8 (12), pp. 973–982. DOI: 10.1038/nrg2165.

Wittkopp, Patricia J.; Haerum, Belinda K.; Clark, Andrew G. (2004): Evolutionary changes in cis and trans gene regulation. In *Nature* 430 (6995), pp. 85–88.

Wu, Dezhi; Liang, Zhe; Yan, Tao; Xu, Ying; Xuan, Lijie; Tang, Juan et al. (2019): Whole-Genome Resequencing of a Worldwide Collection of Rapeseed Accessions Reveals the Genetic Basis of Ecotype Divergence. In *Molecular Plant* 12 (1), pp. 30–43. DOI: 10.1016/j.molp.2018.11.007.

Xia, Li Charlie; Ai, Dongmei; Lee, Hojoon; Andor, Noemi; Li, Chao; Zhang, Nancy R.; Ji, Hanlee P. (2018): SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. In *GigaScience* 7 (7). DOI: 10.1093/gigascience/giy081.

Xiao, Yafang; Li, Mengdi; Wang, Jianbo (2025): Epigenetic modification brings new opportunities for gene capture by transposable elements in allopolyploid Brassica napus. In *Hortic Res* 12 (5), uhaf028. DOI: 10.1093/hr/uhaf028.

Xie, Yilin; Ying, Songbei; Li, Zijuan; Zhang, Yu'e; Zhu, Jiafu; Zhang, Jinyu et al. (2023): Transposable element-initiated enhancer-like elements generate the subgenome-biased spike specificity of polyploid wheat. In *Nature Communications* 14 (1), p. 7465. DOI: 10.1038/s41467-023-42771-9.

Xiong, Zhiyong; Gaeta, Robert T.; Pires, J. Chris (2011): Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. In *Proceedings of the National Academy of Sciences* 108 (19), pp. 7908–7913. DOI: 10.1073/pnas.1014138108.

Xue, Yan; Cao, Xiaofeng; Chen, Xiangsong; Deng, Xian; Deng, Xing Wang; Ding, Yong et al. (2025): Epigenetics in the modern era of crop improvements. In *Science China. Life sciences*, pp. 1–40.

Yildiz, Gözde; Zanini, Silvia F.; Knight, Paul; Golicz, Agnieszka A. (2022): Pangenomics in Agriculture. In *CABI Biotechnology Series*. DOI: 10.1079/9781789247848.0008.

Yildiz, Gözde; Zanini, Silvia F.; Weber, Sven; Kopalli, Venkataramana; Kox, Tobias; Abbadi, Amine et al. (2025): Graphical pangenomics-enabled characterization of structural variant impact on gene expression in Brassica napus. In *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 138 (4), p. 91. DOI: 10.1007/s00122-025-04867-2.

Yuan, Peiguo; Huang, Pei-Cheng; Martin, Timothy K.; Chappell, Thomas M.; Kolomiets, Michael V. (2024): Duplicated Copy Number Variant of the Maize 9-Lipoxygenase ZmLOX5 Improves 9,10-KODA-Mediated Resistance to Fall Armyworms. In *Genes* 15 (4). DOI: 10.3390/genes15040401.

Yuan, Shuai; Qin, Zhaohui (2012): Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression. In *IEEE International Conference on Bioinformatics and Biomedicine workshops. IEEE International Conference on Bioinformatics and Biomedicine* 2012, pp. 718–724. DOI: 10.1109/BIBMW.2012.6470225.

Yuan, Yuxuan; Bayer, Philipp E.; Batley, Jacqueline; Edwards, David (2017): Improvements in Genomic Technologies: Application to Crop Genomics. In *Trends in biotechnology* 35 (6), pp. 547–558. DOI: 10.1016/j.tibtech.2017.02.009.

Yuan, Yuxuan; Bayer, Philipp E.; Batley, Jacqueline; Edwards, David (2021): Current status of structural variation studies in plants. In *Plant Biotechnol J* 19 (11), pp. 2153–2163. DOI: 10.1111/pbi.13646.

Zanini, Silvia F.; Bayer, Philipp E.; Wells, Rachel; Snowdon, Rod J.; Batley, Jacqueline; Varshney, Rajeev K. et al. (2022): Pangenomics in crop improvement-from coding structural variations to finding regulatory variants with pangenome graphs. In *Plant Genome* 15 (1), e20177. DOI: 10.1002/tpg2.20177.

Zhang, Yuanyuan; Yang, Zhiquan; He, Yizhou; Liu, Dongxu; Liu, Yueying; Liang, Congyuan et al. (2024): Structural variation reshapes population gene expression and trait variation in 2,105 Brassica napus accessions. In *Nature Genetics* 56 (11), pp. 2538–2550. DOI: 10.1038/s41588-024-01957-7.

Zhang, Yuyun; Li, Zijuan; Liu, Jinyi; Zhang, Yu'e; Ye, Luhuan; Peng, Yuan et al. (2022): Transposable elements orchestrate subgenome-convergent and -divergent transcription in common wheat. In *Nature Communications* 13 (1), p. 6940. DOI: 10.1038/s41467-022-34290-w.

Zhao, Hu; Li, Jiacheng; Yang, Ling; Qin, Gang; Xia, Chunjiao; Xu, Xingbing et al. (2021a): An inferred functional impact map of genetic variants in rice. In *Molecular Plant* 14 (9), pp. 1584–1599. DOI: 10.1016/j.molp.2021.06.025.

Zhao, Hu; Tu, Zhuo; Liu, Yinmeng; Zong, Zhanxiang; Li, Jiacheng; Liu, Hao et al. (2021b): PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. In *Nucleic Acids Res* 49 (W1), W523-W529. DOI: 10.1093/nar/gkab383.

Zhao, Ting; Guan, Xueying; Hu, Yan; Zhang, Ziqian; Yang, Han; Shi, Xiaowen et al. (2024): Population-wide DNA methylation polymorphisms at single-nucleotide resolution in 207 cotton accessions reveal epigenomic contributions to complex traits. In *Cell Research*, pp. 1–14.

Zheng, Dewei; Ye, Wenxue; Song, Qingxin; Han, Fangpu; Zhang, Tianzhen; Chen, Z. Jeffrey (2016): Histone Modifications Define Expression Bias of Homoeologous Genomes in Allotetraploid Cotton. In *Plant Physiol* 172 (3), pp. 1760–1771. DOI: 10.1104/pp.16.01210.

Zhou, Peng; Li, Zhi; Magnusson, Erika; Gomez Cano, Fabio; Crisp, Peter A.; Noshay, Jaclyn M. et al. (2020): Meta Gene Regulatory Networks in Maize Highlight Functionally Relevant Regulatory Interactions[OPEN]. In *Plant Cell* 32 (5), pp. 1377–1396. DOI: 10.1105/tpc.20.00080.

Zhu, Zhihong; Zhang, Futao; Hu, Han; Bakshi, Andrew; Robinson, Matthew R.; Powell, Joseph E. et al. (2016): Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. In *Nature Genetics* 48 (5), pp. 481–487. DOI: 10.1038/ng.3538.

Zhuang, Weijian; Chen, Hua; Yang, Meng; Wang, Jianping; Pandey, Manish K.; Zhang, Chong et al. (2019): The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. In *Nature Genetics* 51 (5), pp. 865–876. DOI: 10.1038/s41588-019-0402-2.

9. Appendix

9.1 Appendix I: Supplementary Information

Benchmarking Oxford Nanopore read alignment-based insertion and deletion detection in crop plant genomes

The Plant Genome, 16, e20314.

https://doi.org/10.1002/tpg2.20314

Figure S1. Read simulation methodology.
Figure S2. Read aligner run time (h:mm: ss or m: ss) for maize simulation and real-world datasets with 5×, 10×, and 20× coverages (8 CPU).
Figure S3. F1-scores for maize simulated datasets including total SVs, deletions, and insertions at 5×, 10×, and 20× coverages for different combinations of read aligners and SV callers.
Figure S4. Proportion of overlapped SVs (%), across 5×, 10×, and 20× coverages for maize simulated and real-world datasets.
Figure S5. Overlap between variants found by Lemay et al., 2022 and by minimap2/cuteSV/Sniffles2 combination for genotype Maple Isle.
Figure S6-S9. IGV screen shots of variants not present Lemay et al., 2022 dataset, but found by minimap2/cuteSV/Sniffles2 combination.

Figure S1

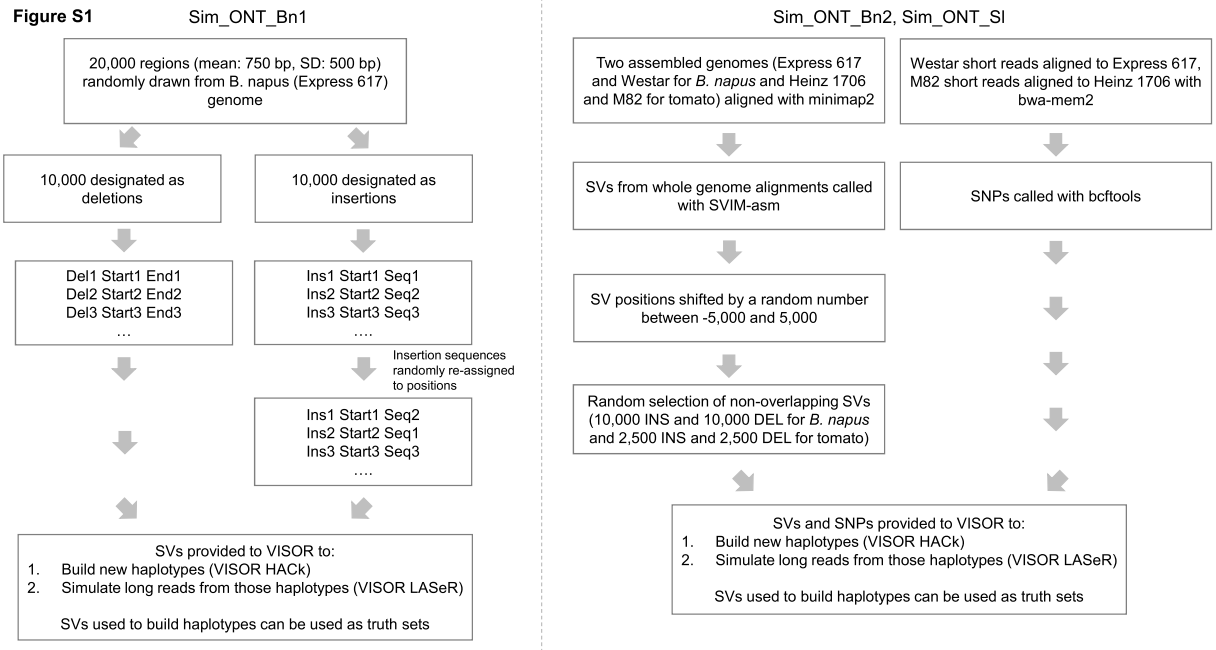
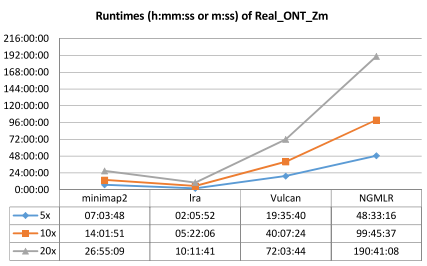
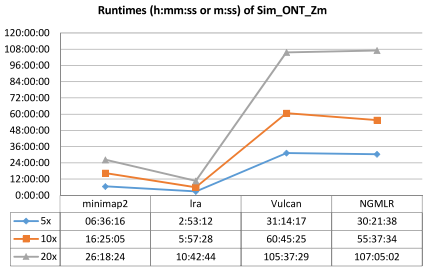


Figure S2



	minimap2	lra	Vulcan	NGMLR
5x	07:03:48	02:05:52	19:35:40	48:33:16
10x	14:01:51	05:22:06	40:07:24	99:45:37
20x	26:55:09	10:11:41	72:03:44	190:41:08



	minimap2	lra	Vulcan	NGMLR
5x	06:36:16	2:53:12	31:14:17	30:21:38
10x	16:25:05	5:57:28	60:45:25	55:37:34
20x	26:18:24	10:42:44	105:37:29	107:05:02

Figure S3

Mapping based/Sim_ONT_Zm

Total

	5x-F1	10x-F1	20x-F1
Minimap2			
cuteSV	0.8383	0.8957	0.8768
Sniffles2	0.8185	0.8774	0.8573
SVIM	0.7082	0.8771	0.8566
dysgu	0.8425	0.9265	0.8954
lra			
cuteSV	0.7682	0.8336	0.8283
Sniffles2	0.8050	0.8857	0.8812
SVIM	0.6413	0.8308	0.8235
dysgu	0.6828	0.8114	0.8001
Vulcan			
cuteSV	0.6970	0.7621	0.7648
Sniffles2	0.6480	0.7130	0.7182
SVIM	0.5450	0.7384	0.7388
dysgu	0.6336	0.7378	0.7372
NGMLR			
cuteSV	0.6104	0.6602	0.6715
Sniffles2	0.5398	0.6061	0.6276
SVIM	0.4624	0.6575	0.6730
dysgu	0.5385	0.6614	0.6753

	5x-F1	10x-F1	20x-F1
NanoVar	0.6933	0.7443	0.7147

Deletions

	5x-F1	10x-F1	20x-F1
Minimap2			
cuteSV	0.8944	0.9511	0.9215
Sniffles2	0.8867	0.9490	0.9214
SVIM	0.7695	0.9366	0.9095
dysgu	0.8790	0.9479	0.9136
lra			
cuteSV	0.8291	0.8964	0.8926
Sniffles2	0.8236	0.8955	0.8912
SVIM	0.6963	0.8958	0.8892
dysgu	0.7877	0.8925	0.8874
Vulcan			
cuteSV	0.7347	0.7994	0.8045
Sniffles2	0.6736	0.7392	0.7441
SVIM	0.5639	0.7663	0.7691
dysgu	0.6519	0.7619	0.7664
NGMLR			
cuteSV	0.6551	0.7070	0.7131
Sniffles2	0.5581	0.6187	0.6352
SVIM	0.4927	0.6953	0.7049
dysgu	0.5837	0.7058	0.7179

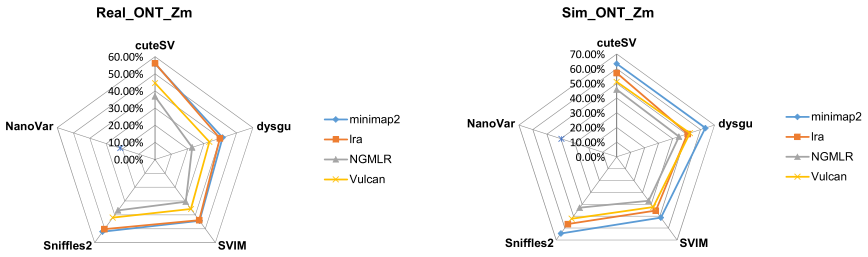
	5x-F1	10x-F1	20x-F1
NanoVar	0.8380	0.9044	0.8830

Insertions

	5x-F1	10x-F1	20x-F1
Minimap2			
cuteSV	0.7759	0.8338	0.8280
Sniffles2	0.7410	0.7953	0.7850
SVIM	0.6403	0.8106	0.7982
dysgu	0.8038	0.9043	0.8766
lra			
cuteSV	0.7005	0.7628	0.7558
Sniffles2	0.7858	0.8757	0.8711
SVIM	0.5813	0.7575	0.7494
dysgu	0.5583	0.7175	0.6981
Vulcan			
cuteSV	0.6569	0.7225	0.7223
Sniffles2	0.6214	0.6857	0.6912
SVIM	0.5256	0.7093	0.7071
dysgu	0.6148	0.7128	0.7067
NGMLR			
cuteSV	0.5625	0.6097	0.6271
Sniffles2	0.5211	0.5933	0.6199
SVIM	0.4309	0.6174	0.6395
dysgu	0.4904	0.6139	0.6298

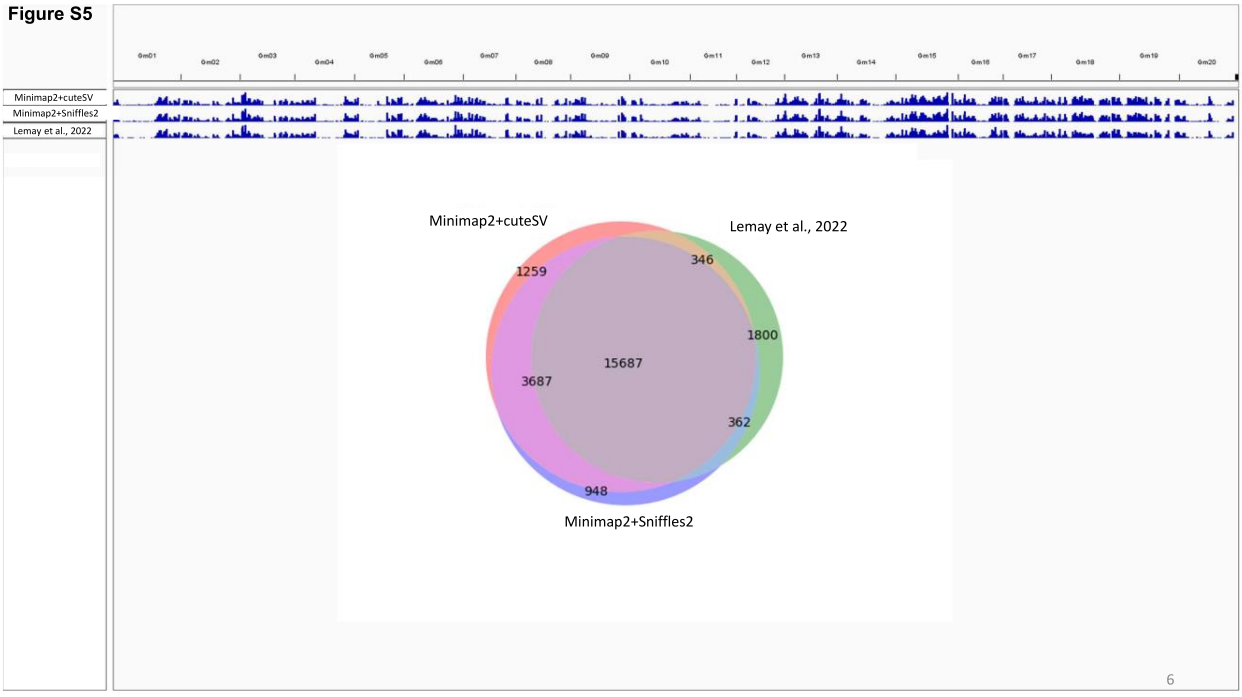
	5x-F1	10x-F1	20x-F1
NanoVar	0.5176	0.5429	0.5004

Figure S4



5

Figure S5



6

Figure S6

7

Figure S7

Figure S8

Figure S9

9.2 Appendix II: Supplementary Information

Graphical pangenomics-enabled characterization of structural variant impact on gene expression in *Brassica napus*

Theor Appl Genet 138, 91 (2025).

https://doi.org/10.1007/s00122-025-04867-2

Supplementary material

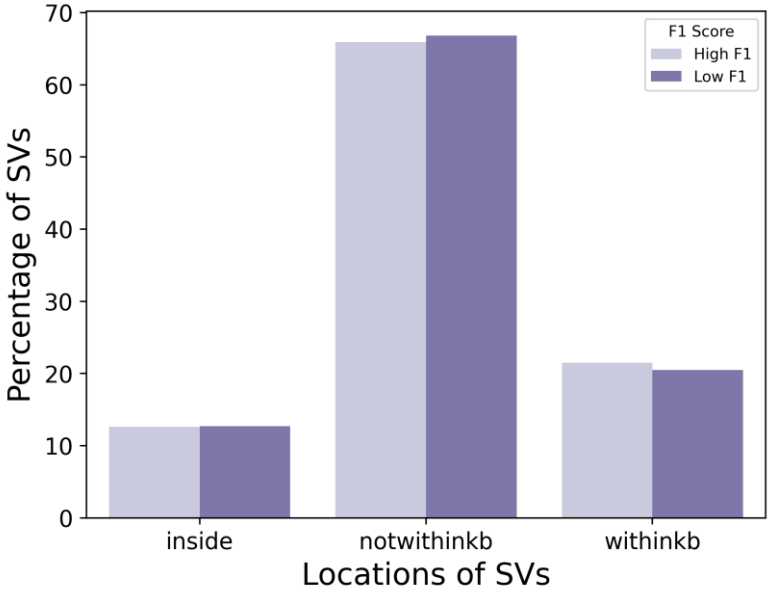


Figure S1. Gene-Proximity patterns for variants with high and low F1 scores with Paragraph.

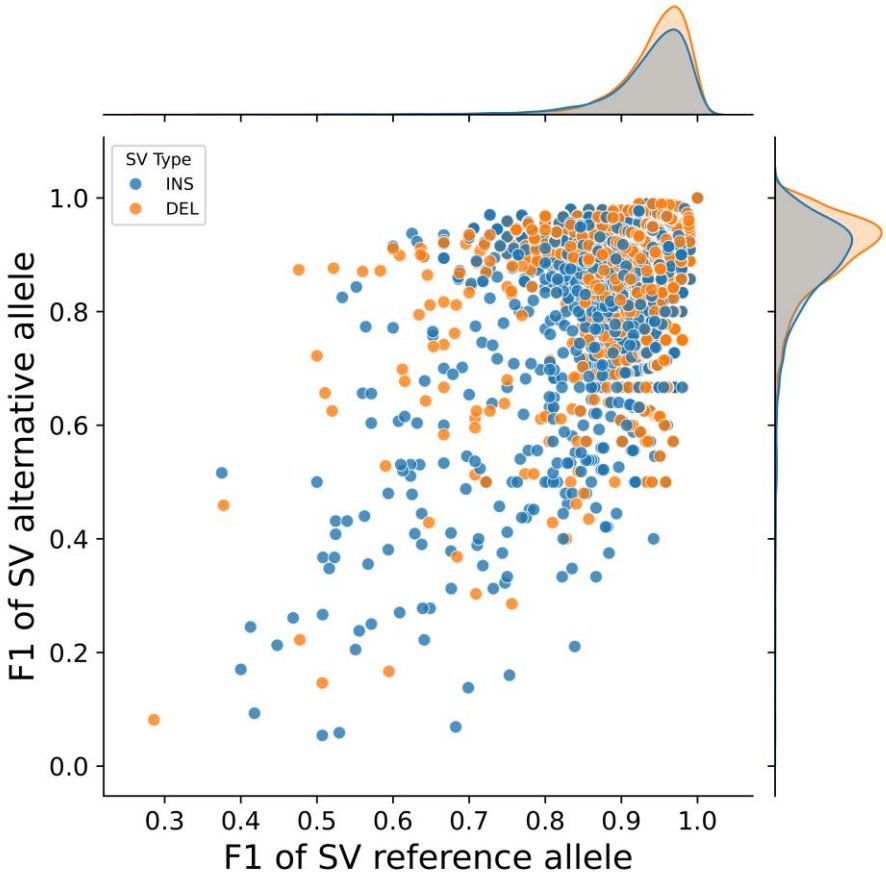


Figure S2. Distribution of F1-scores of overlapped SVs between eQTL-SVs discovered using genotyping with short (n=100) and long (n=57) reads.

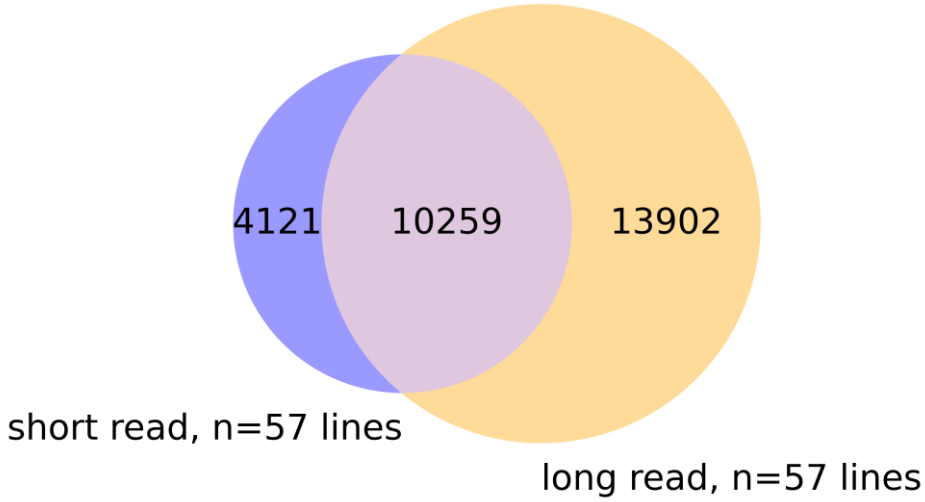


Figure S3A: Venn diagram illustrating the overlap and unique eQTL-SVs identified through genotyping using short-read (n=57) and long-read (n=57) sequencing.

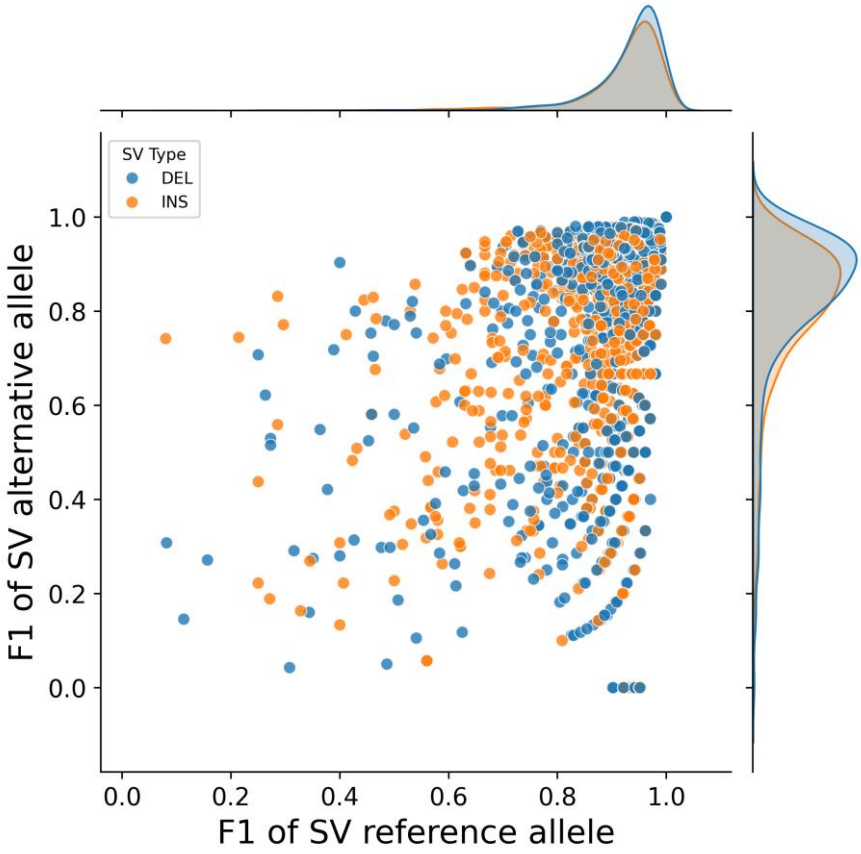


Figure S3B: Distribution of F1-scores of SVs genotyped from short reads, which were unique to SV-eQTL analysis with short read-derived genotypes (n:57 short reads).

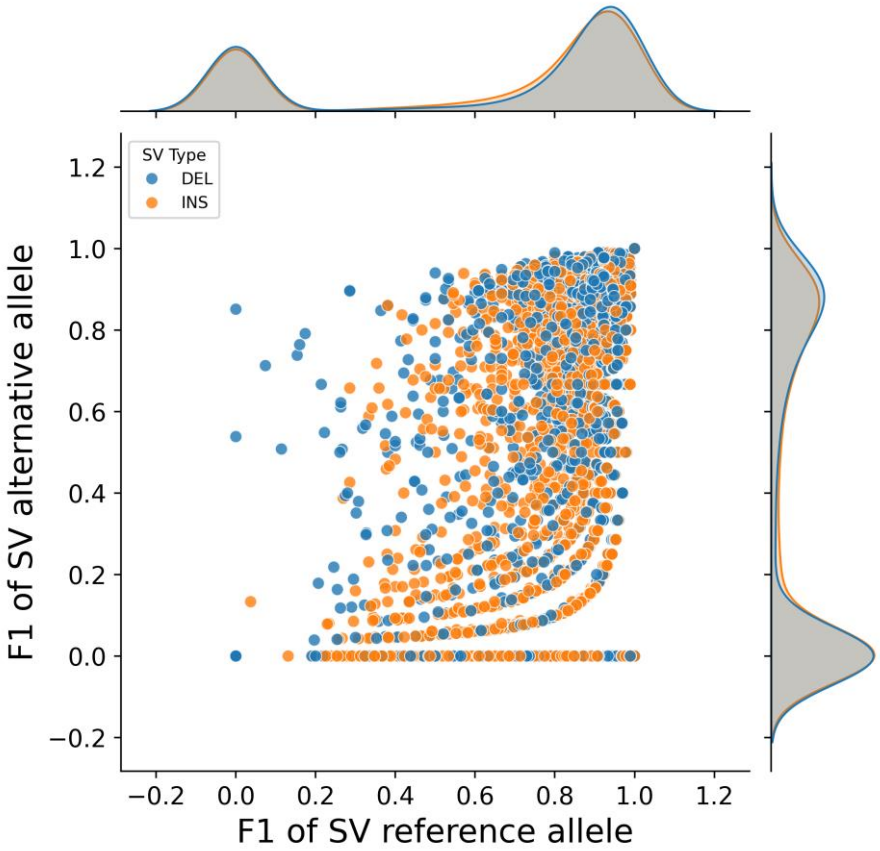


Figure S3C: Distribution of F1-scores of SVs genotyped from short reads for eQTL-SVs unique to analysis with long read-derived genotypes (n:57 long reads).

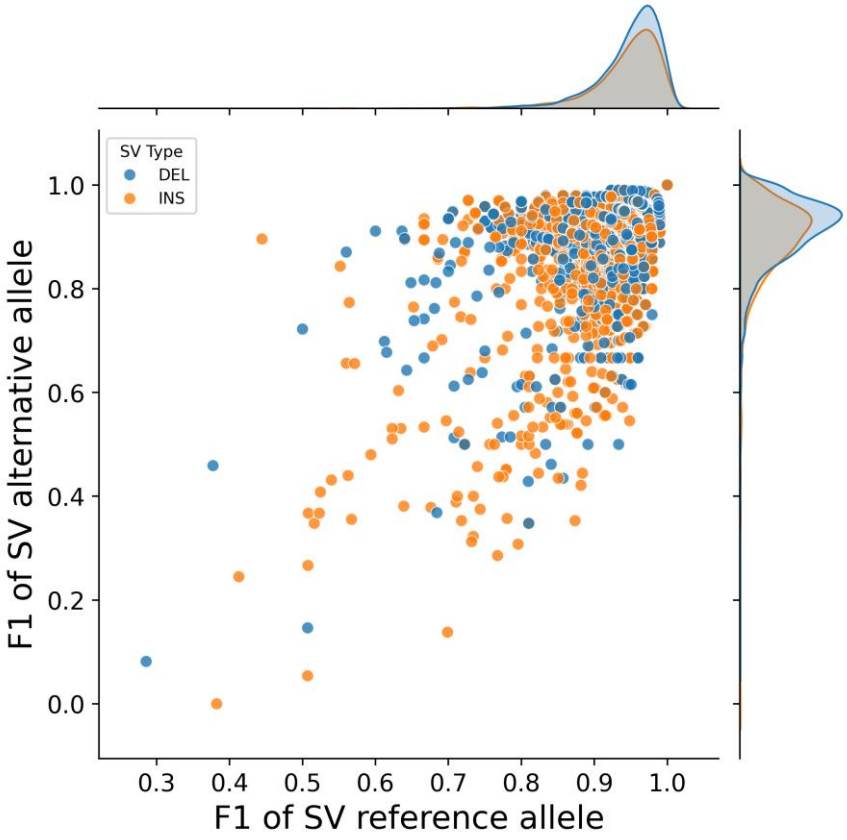


Figure S3D: Distribution of F1-scores of overlapped SVs between eQTL-SVs was discovered using genotyping with short (n:57) and long (n:57) reads.

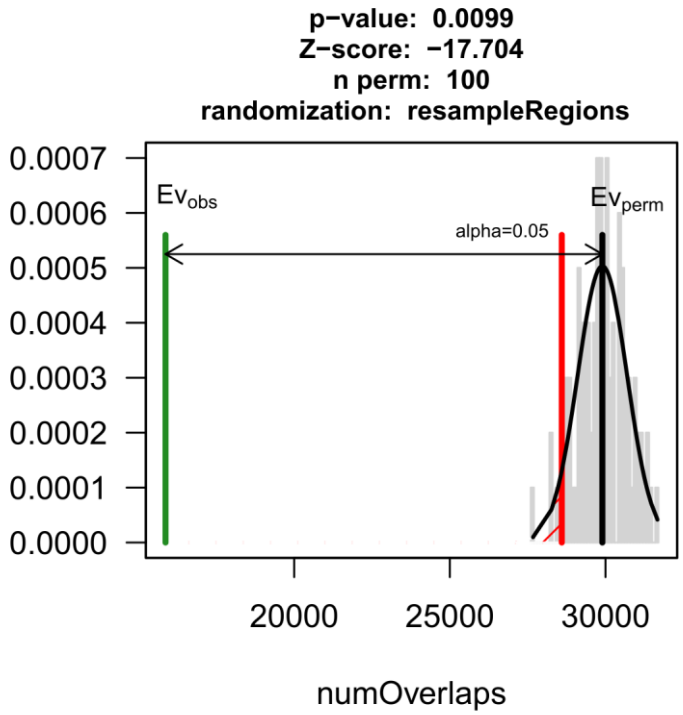


Figure S4A. Transcripts with high concordance of quantification results from Kallisto and RPVG are under-represented in SNPs. Green line – observed value, grey line - mean of permutation results, red line – significance threshold.

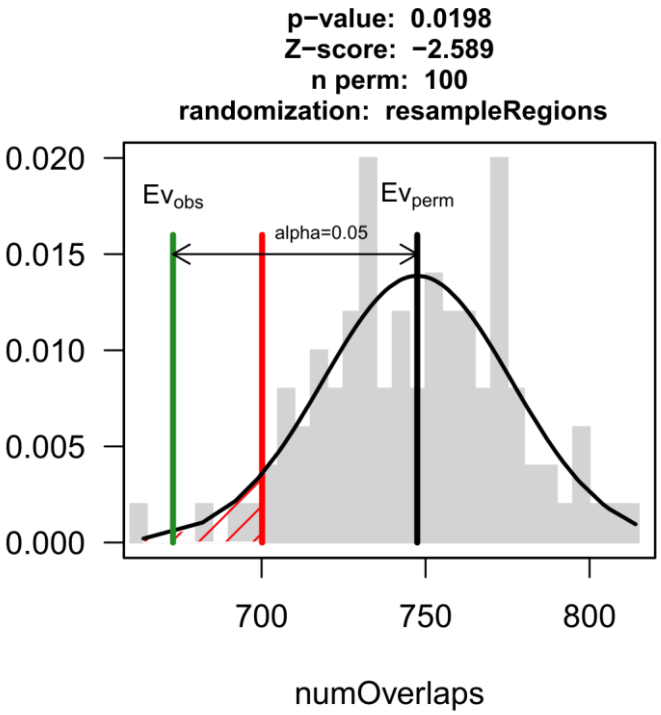


Figure S4B. Transcripts with high concordance of quantification results from Kallisto and RPVG are under-represented in SVs. Green line – observed value, grey line - mean of permutation results, red line – significance threshold.

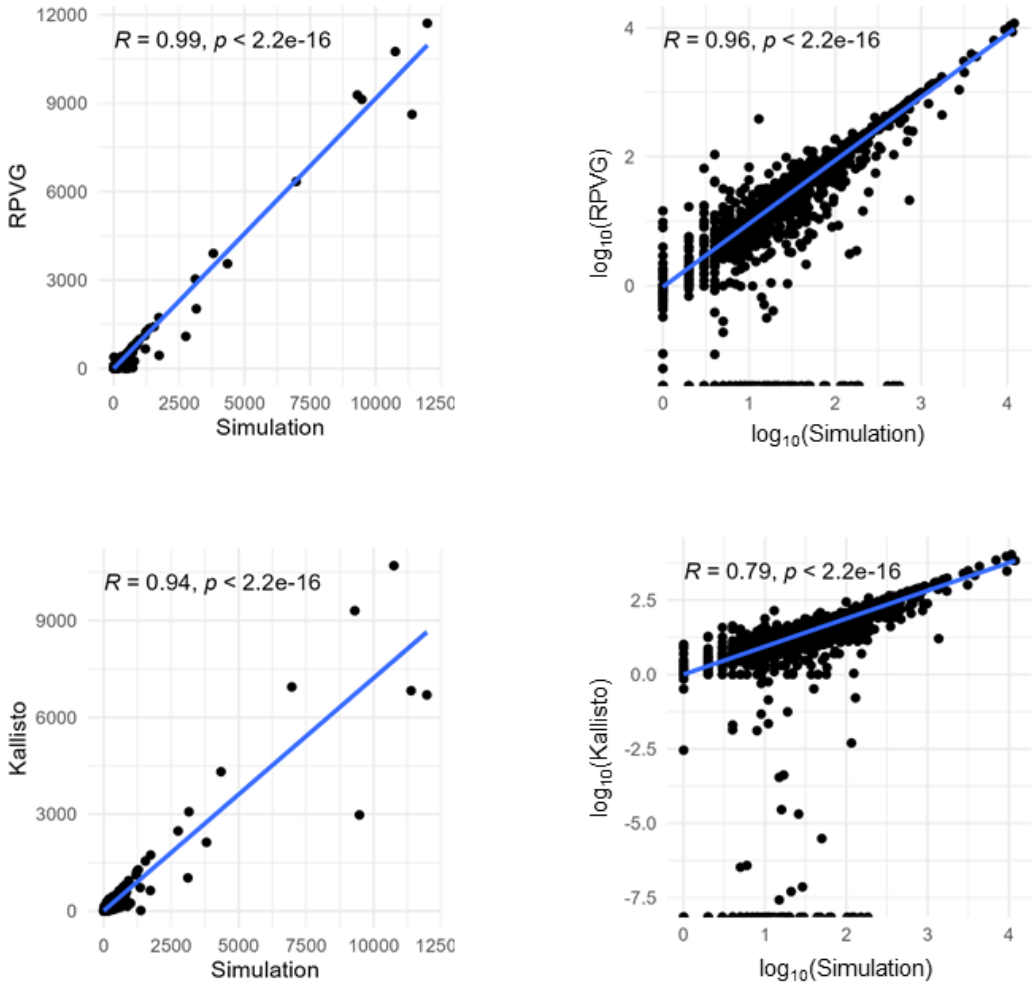


Figure S5. Correlation between simulated read counts and quantification results from rpvg and Kallisto.

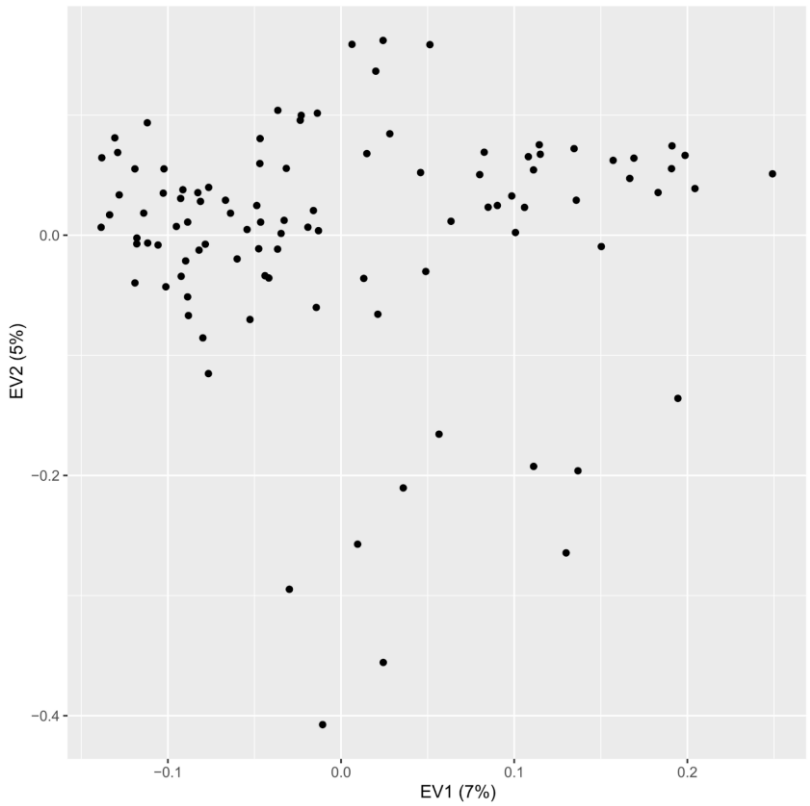


Figure S6A: Principal Component Analysis (PCA) plot of SNPs (n: 100) discovered based on linear reference.

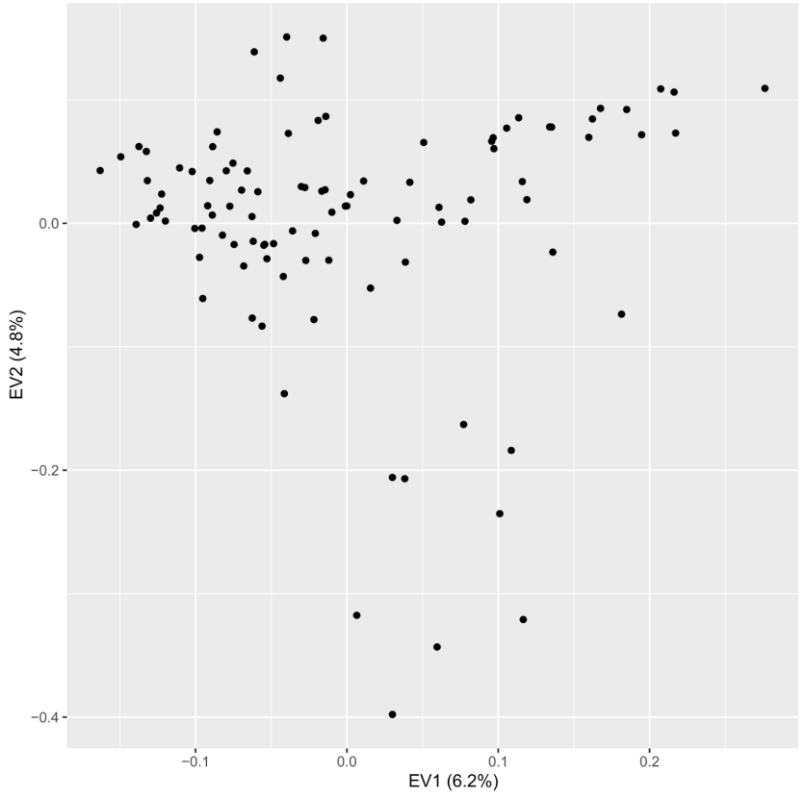


Figure S6B: Principal Component Analysis (PCA) plot of SVs (n: 57) genotyped from graph.

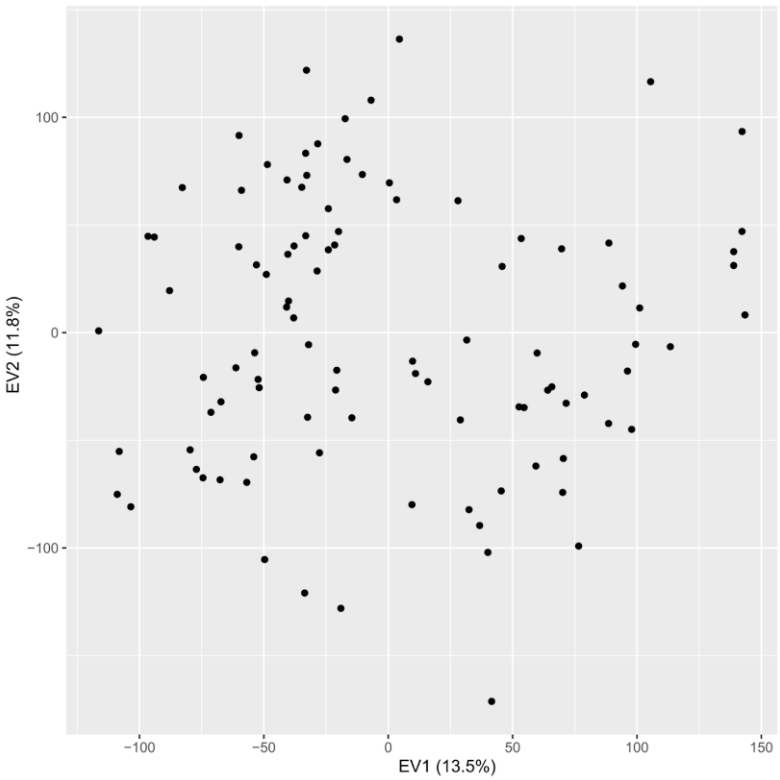


Figure S6C: Principal Component Analysis (PCA) plot of RNASeq (n: 100) quantified from graph.

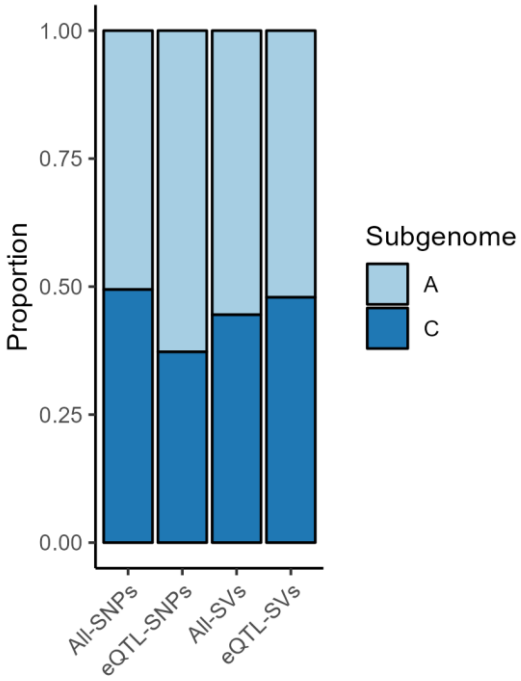


Figure S7: Distribution of eQTLs relative to genomic features. eQTL-SNPs and eQTL-SVs are differently distributed across the two sub-genomes of *Brassica napus*.

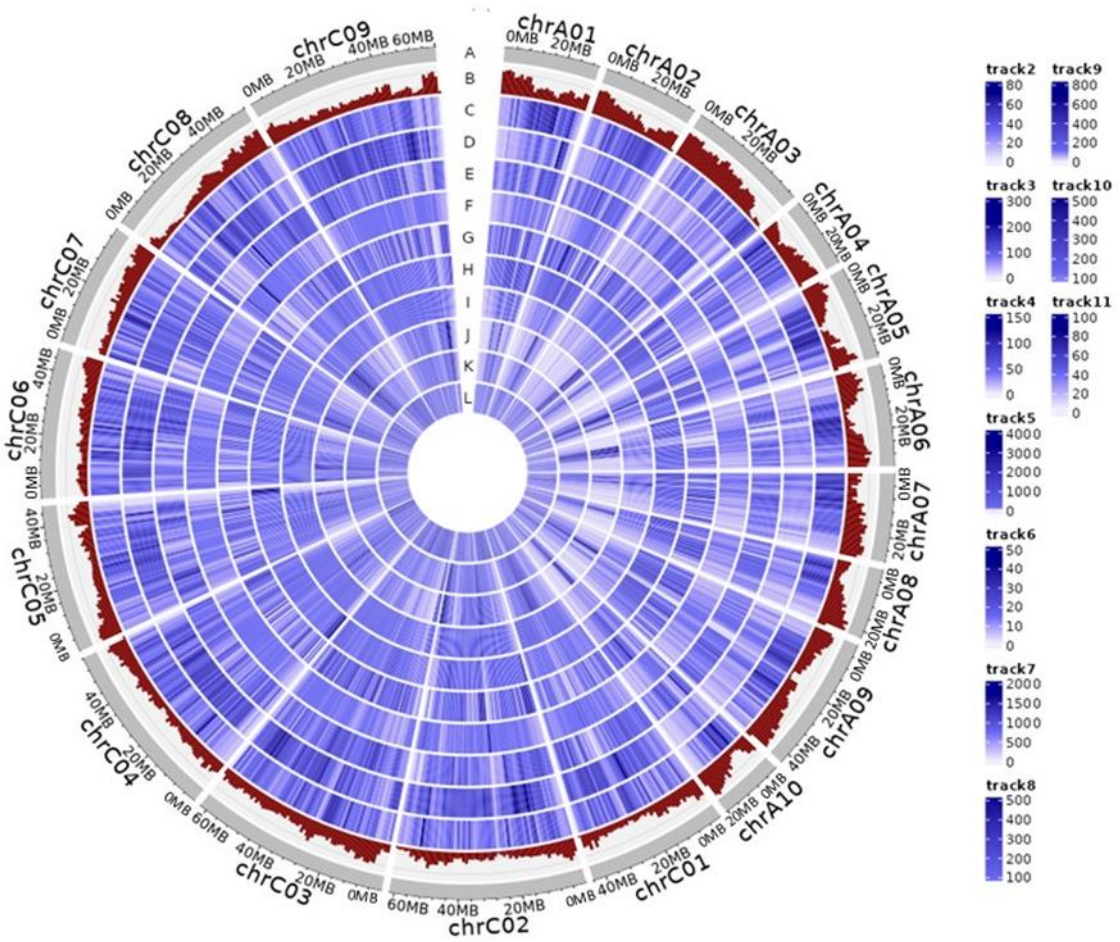


Figure S8. Distribution of transposable elements (TEs) abundance of the *B. napus* genome. Chromosomes (A), Gene density (B), DNA/DTA TEs density (track 2) (C), DNA/DTC TEs density (track 3) (D), DNA/DTH TEs density (track 4) (E), DNA/DTM TEs density (track 5) (F), DNA/DTT TEs density (track 6) (G), DNA/Helitron TEs density (track 7) (H), DNA/Copia TEs density (track 8) (I), LTR/Gypsy TEs density (track 9) (J), LTR/unknown TEs density (track 10) (K), MITE TEs density (track 11) (L). Densities were calculated using 1 Mb window size.

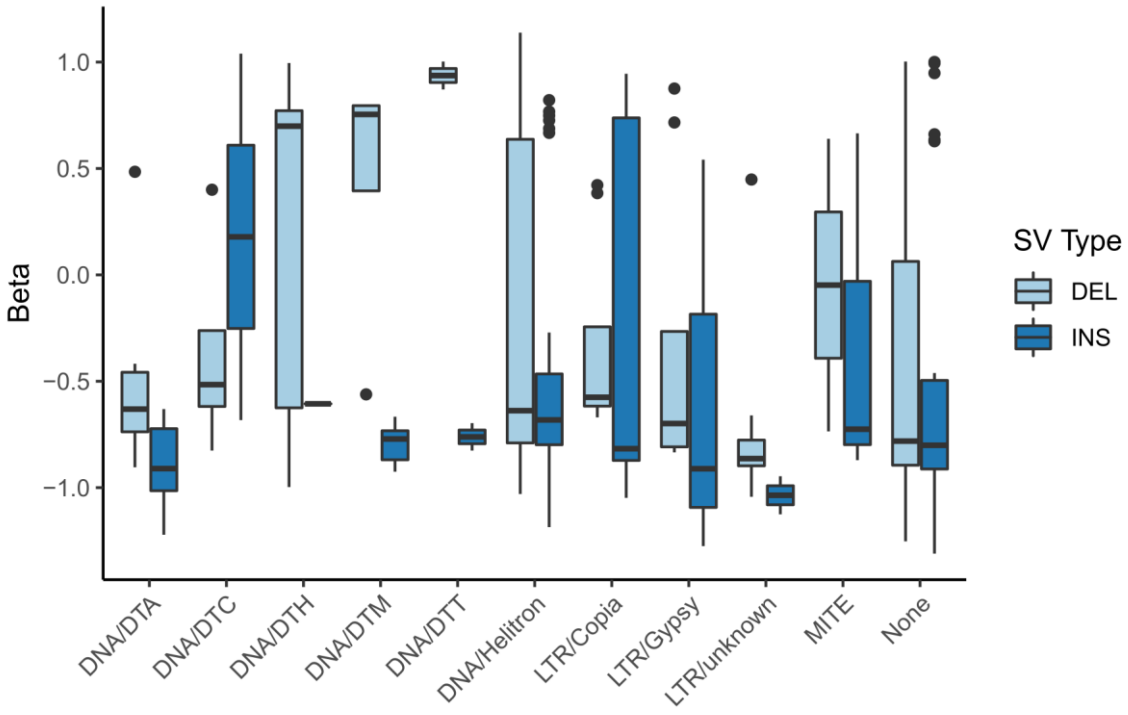


Figure S9. Effect of sequence insertions and deletions on gene expression.

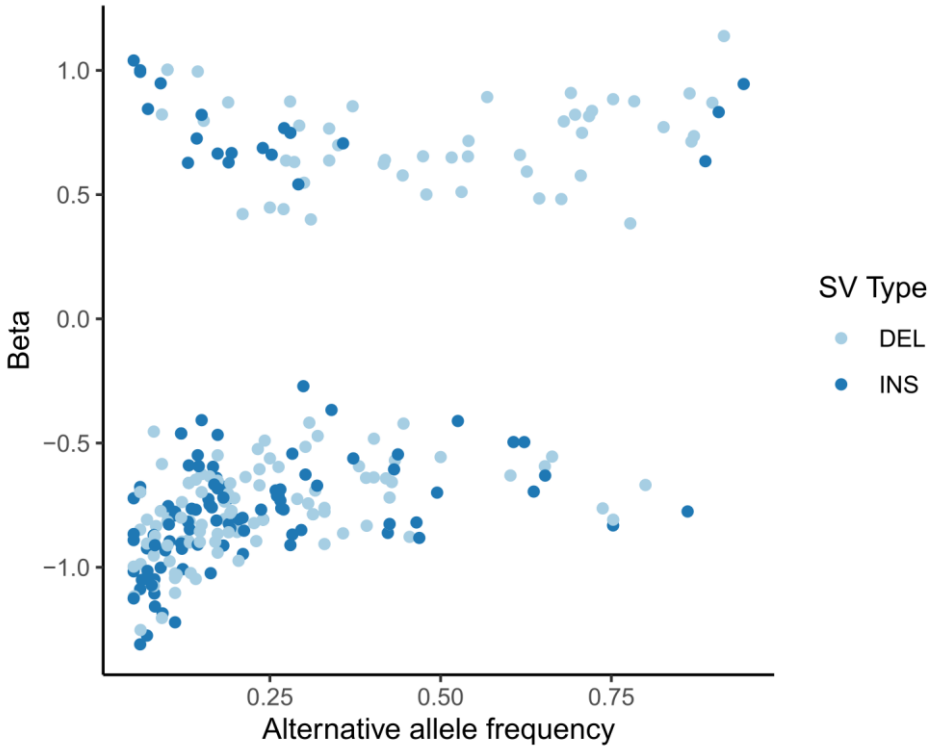


Figure S10A. Relationship between SV alternative allele frequency, effect size (Beta) and variant type.

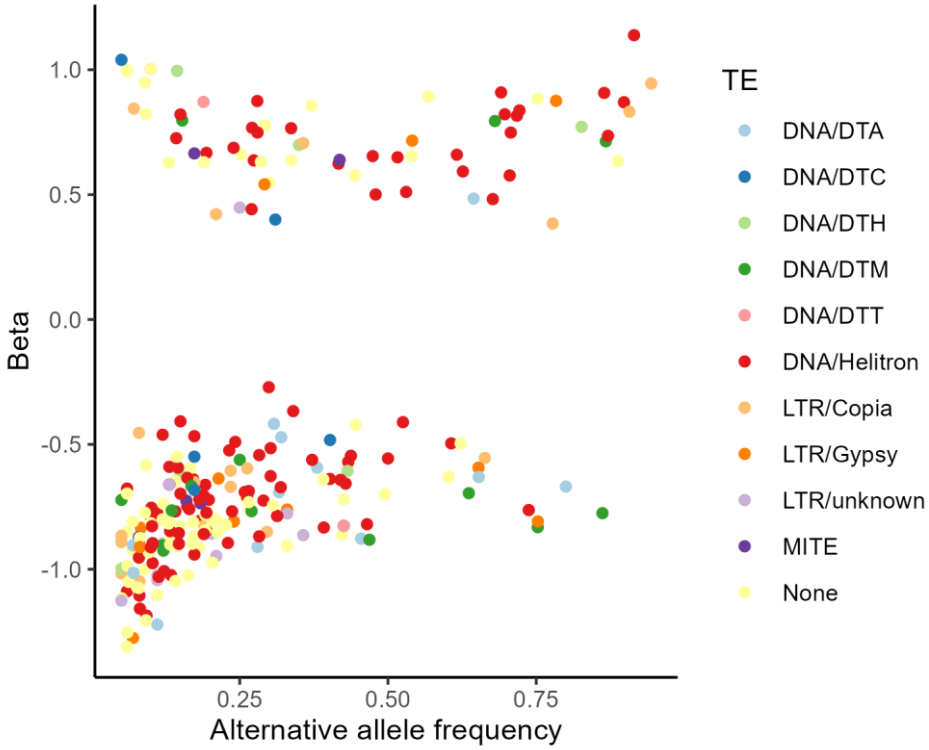


Figure S10B. Relationship between SV alternative allele frequency, effect size (Beta) and TE classification.

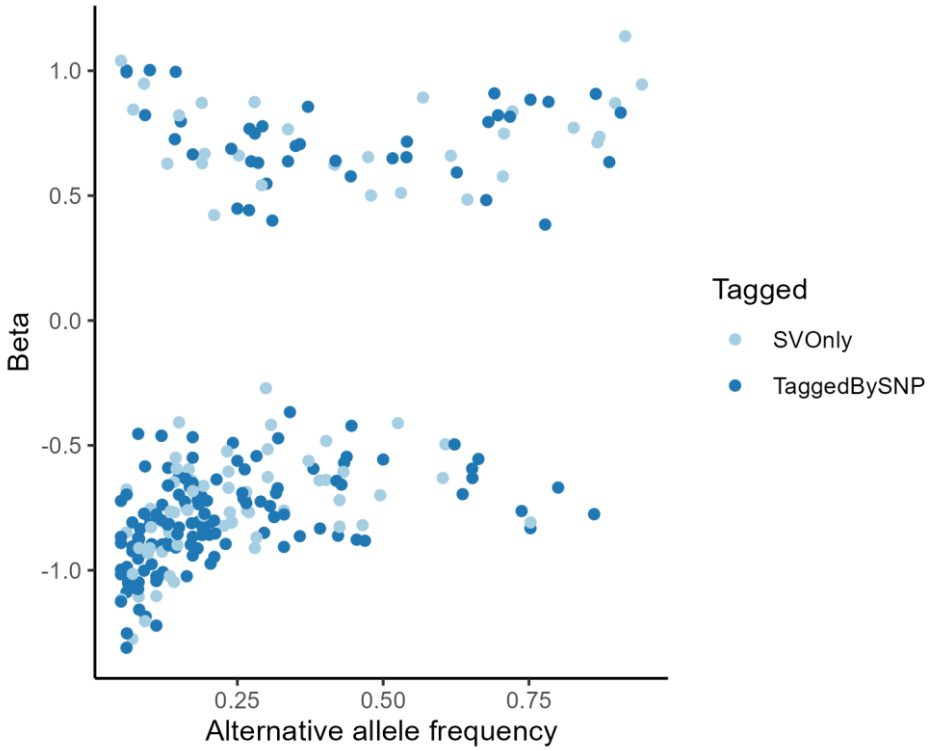


Figure S10C. Relationship between SV alternative allele frequency, effect size (Beta) and being tagged (in LD) with SNP.

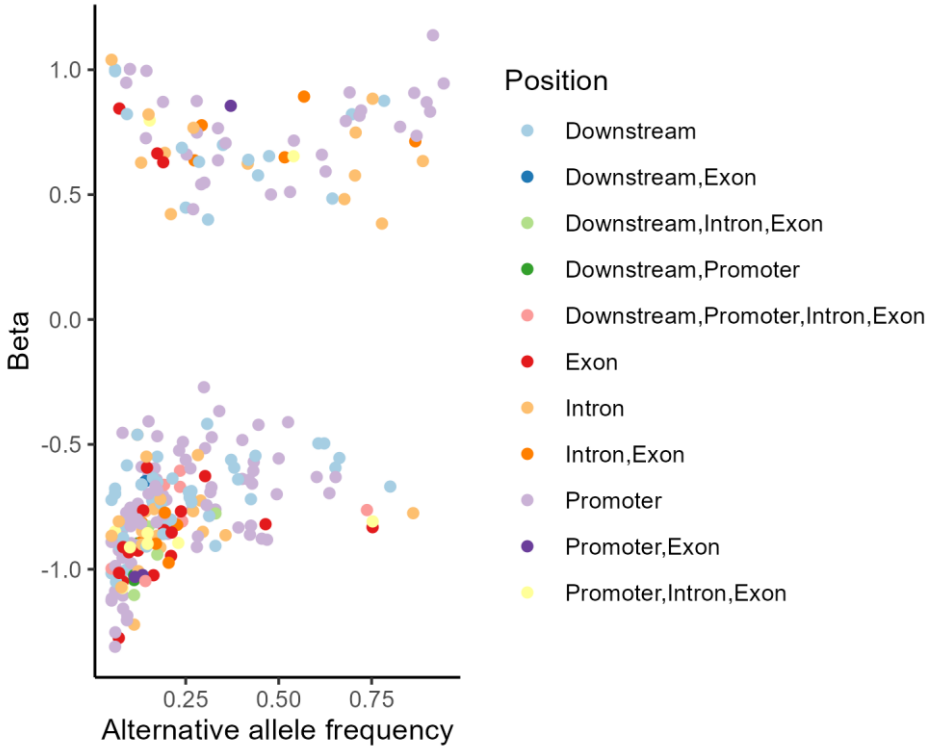


Figure S10D. Relationship between SV alternative allele frequency, effect size (Beta) and variant locations.

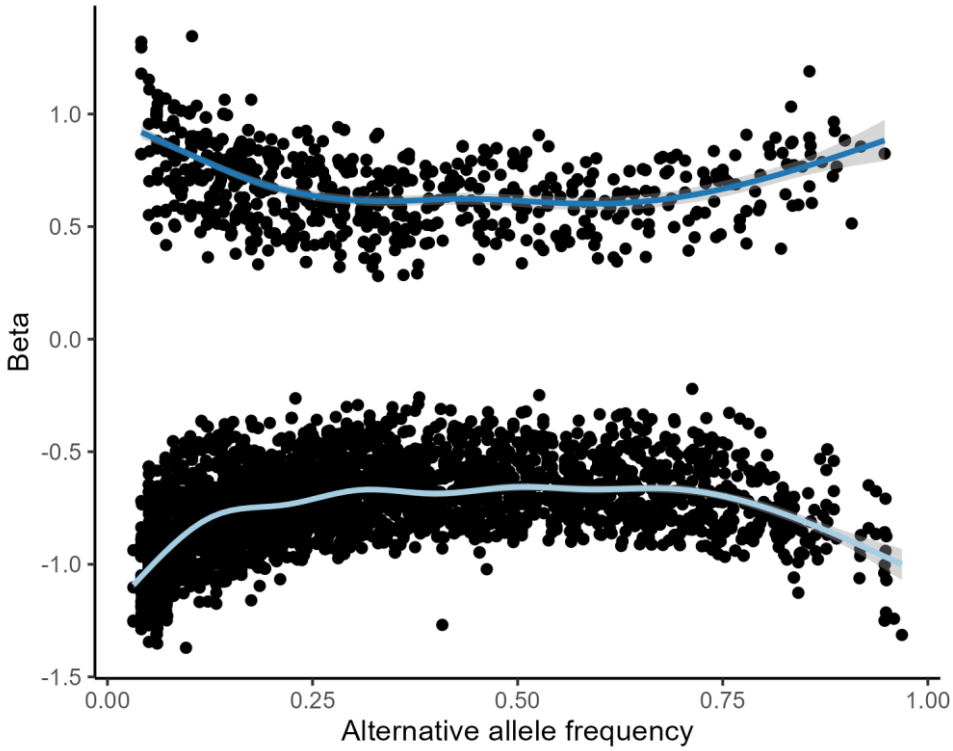


Figure S11. Relationship between SNP alternative allele frequency and effect size (Beta).

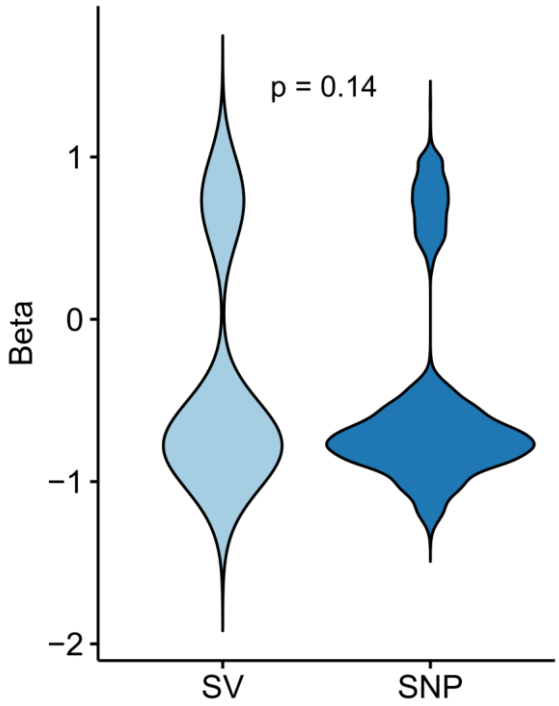


Figure S12. No difference in effect size between eQTL-SNPs and SVs was observed (Wilcoxon test, P=0.14).

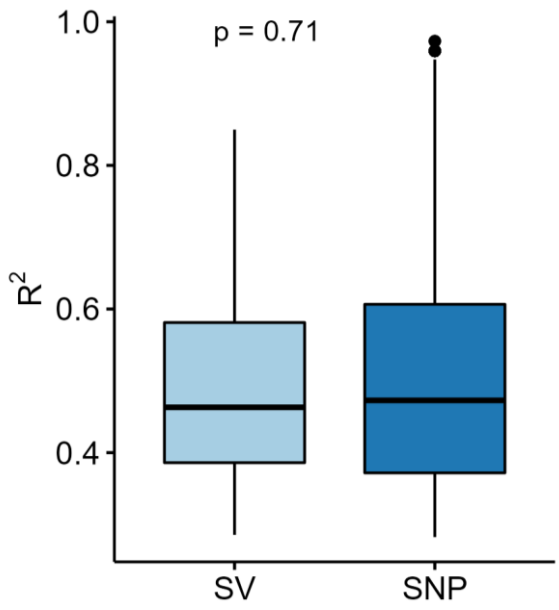


Figure S13. No difference of variance explained between eQTL-SNPs and SVs was observed (Wilcoxon test, P=0.71).

Declaration of Academic Integrity

„Ich erkläre: Ich habe die vorgelegte Dissertation selbständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.“

Giessen, 26.09.2025 Gözde Yildiz

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Agnieszka A. Golicz, for her invaluable guidance, continuous support, and encouragement during my PhD. Her scientific insight, patience, and enthusiasm for research have been an inspiration and shaped me both as a researcher and as a person.

I am also sincerely grateful to Prof. Dr. Rod J. Snowdon for his mentorship, constructive feedback, and encouragement, which were essential for the progress of my work. A very special thanks goes to Dr. Silvia F. Zanini, whose mentoring, discussions, and constant encouragement over the past four years have been a source of both motivation and learning. The countless conversations with Prof. Dr. Agnieszka A. Golicz and Dr. Silvia F. Zanini has deeply enriched my scientific journey and helped me navigate the challenges of my PhD with confidence.

I gratefully acknowledge the Alexander von Humboldt Foundation, which supported this research in the framework. I also wish to acknowledge the bioinformatics resources provided by the de.NBI network and Justus Liebig University (JLU) Bioinformatics Core Facility (BCF), which were essential for the computational aspects of this work.

I would also like to thank all members of the AgroBioinformatics group at JLU for sharing their knowledge, offering constructive feedback, and creating a stimulating research environment. My sincere appreciation also goes to the International Giessen Graduate Centre for Life Sciences (GGL), which provided an interdisciplinary and well-structured graduate programme. Through its seminars, practical courses, research retreats, and transferable skills training, the GGL contributed greatly to my academic development and career preparation. I am also thankful to my thesis jury members for their helpful comments and valuable contributions to this work.

Finally, my deepest thanks go to my family; my mother and father Arife&Veli Yildiz, my sister and brother Hande&Alican Yildiz, for their unconditional love, patience, and encouragement. Their endless support has carried me through every challenge.