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

Genome sequence analysis for structural variation detection in oilseed rape (*Brassica napus* L.)

Inaugural Dissertation for a Doctorate Degree in Agricultural Sciences
in the Faculty of Agricultural Sciences, Nutritional Sciences and
Environmental Management

Examiners
Prof. Dr. Rod Snowdon
Prof. Dr. Matthias Frisch

Submitted by
Harmeet Singh Chawla

Giessen 2020

"Around here, however, we don't look backwards for very long. We keep moving forward, opening up new doors and doing new things, because we're curious ... and curiosity keeps leading us down new paths."

Walt Disney

Table of contents

1 General introduction

1.1 Oilseed rape: Origin, evolution and genetic diversity for breeding

Brassica napus L. (oilseed rape/canola/kale/rutabaga; genome AACC, 2n=38) is the second most important oilseed crop in the world after soybean. Rapeseed oil accounted for 16% of the total oilseed production worldwide in the year 2017. (FAOSTAT 2016) (http://faostat.fao.org). The allopolyploid species *B. napus* originated from interspecific hybridization between *Brassica rapa* ($2n = 2x = 20$, AA) and *Brassica oleracea* ($2n = 2x = 18$, CC) less than 7500 years ago (Chalhoub et al. 2014). *B. napus* is often referred to as a "man-made" species. It has been hypothesized that oilseed rape originated as an unintentional hybrid from human agricultural practices and there have been no reports underlining the existence of a wild type of *B. napus*. Absence of wild germplasm together with rigorous selection for low seed erucic acid and glucosinolate content (00 quality) during the 1970s and 1980s resulted in a major genetic bottleneck. In order to introduce new genetic diversity there have been several efforts to produce synthetic *B. napus* lines by interspecific crossing between *B rapa* and *B oleracea* (Gaeta et al. 2007). These synthetic lines often exhibit a higher frequency of pairing between homeologous chromosomes, thereby leading to increased homeologous exchanges (HE) compared to natural oilseed rape (Hurgobin et al. 2018). Higher rate of HE can often lead to an increase in genome structural variations (SV) in the rapeseed genomes in the form of copy number variations (CNV) or presence-absence variation (PAV).

1.2 The complex *B. napus* genome: Large and small-scale genome restructuring

The rapeseed genome contains a plethora of genome structural variations (SV) in the form of insertion/deletion polymorphisms (InDels), CNV, translocations, inversions, or simple variation in microsatellite repeat number. There is no widely accepted categorization of genome re-arrangements based on their size. Therefore, in this thesis SV are categorized as small scale (30 to 10,000 bp), mid-scale (10,000 bp to 30,000 bp) and large-scale (greater than 30,000 bp). There have been several studies associating agronomically important traits such as disease resistance (Gabur et al. 2018), flowering time (Schiessl et al. 2017b), seed quality (Stein et al. 2017) to SV in oilseed rape. These studies revealed the important role of SV in the creation of *de novo* variation for adaptation and breeding, however all of these studies were only able to detect large-scale SV due to the limited resolution offered by the methods that were used for detection of genomic re-arrangements (SNP genotyping arrays or short read sequencing). Qian et al. (2016) reported a first example of intragenic SV impacting quantitatively inherited traits in *B. napus*

where a deletion of exons 2 and 3 from a *B. napus* orthologue of Mendel's "Green Cotyledon" gene (the Staygreen gene *NON-YELLOWING 1; NYE1*) was associated with quantitative variation for chlorophyll and oil content. Furthermore, the study in chapter 4 revealed a surprisingly high level of widespread, small to mid-scale SV in oilseed rape.

1.3 Homoeologous chromosome exchanges, genome structural variation and gene conversion

The illicit paring between the homeologous chromosome (during meiosis) with high levels of sequence similarity can lead to exchange of DNA fragments between the A and C subgenomes (Gaeta and Pires 2010) of *B. napus*. This type of intra-genome exchange of genetic material is referred to as homoeologous exchange (HE). In oilseed rape HE could shuffle large segments of chromosomes or just a single gene (or even SNP) among the homeologous regions of the genome, gene conversions (Chalhoub et al. 2014). Large HE (more than 80kb in length) could occur both in the form of reciprocal (also known as homeologous reciprocal translocations) and non-reciprocal translocations (also known as homeologous non-reciprocal translocations, HNRT). In *B. napus* HNRT has been shown to create genomic re-arrangements such as duplications (CNV) or deletions (PAV) ranging from a few hundred kilobases (kb) to a few megabases (Mb) (Stein et al. 2017, Samans et al. 2017). Chalhoub et al. (2014) reported a very large HE, where a 9 Mb fragments of the chromosome A01 was replaced by a 13 Mb piece of chromosome C01 in the synthetic *B. napus* accession H165. Such large-scale SV in *B. napus* are associated with eco-geographical adaptability, morphological diversity (Schiessl et al. 2017a) disease resistance (Dolatabadian et al. 2020) and seed quality (Stein et al. 2017) in oilseed rape.

In contrast to large HE, small exchanges (less then 80kb in length) also known as gene conversions (GCs) are necessarily non-reciprocal in nature. GC involves unidirectional transfer of short DNA fragments among homologous non-sister chromatids. GCs can be either allelic, where one allele of the same gene replaces another allele, or ectopic, where one paralogous DNA sequence converts to another (Gardiner et al. 2019). Chalhoub et al. (2014) identified 1.3 times more A to C (56) then C to A (37) converted genes in the rapeseed genome. A similar gene conversion trend was observed in cotton where higher number alleles were converted from the larger repeat-rich genome to the smaller genome (Paterson et al. 2012). Although the genetic factors controlling GC in plant are largely unknown, Gardiner et al. (2019) identified an ATP-dependent RNA helicase gene controlling GC in wheat. The study in chapter 4 of this thesis

describes the presence of widespread intragenic SV in 12 *B. napus* accessions. There is a possibility that some of the SV described in this study are actually GC events. However, additional analysis of the homeologous gene pairs would be required to distinguish GCs from the *de novo* genomic re-arrangements.

1.4 Genome sequencing technologies – short-read vs long-read sequencing for detection of SV underlining complex traits

DNA sequencing has revolutionized the field of plant genomics, and "next-generation" sequencing technologies have made it possible to access highly repetitive or re-arranged regions of the plant genomes which were inaccessible to SNP genotyping arrays. Sequencing technologies can be broadly classified into 3 major generations. Sanger sequencing, the initial (first-generation) DNA sequencing technology, was capable of delivering single DNA reads up to 1000 bp in length. In the mid to late 2000s, Sanger sequencing was superseded by the second generation of massively parallel sequencing technologies (Imelfort and Edwards 2009) capable of tremendously high sequencing throughput at very low cost. These technologies accelerated the process of SV discovery, illuminating the role of SV in important agronomic traits in several plant species such as oilseed rape (Stein et al. 2017, Schiessl et al. 2017a, Schiessl et al. 2017b), cotton (Wang et al. 2015), Arabidopsis (Cao et al. 2011) or barley (Fujii et al. 2012). However, all second-generation sequencing technologies were limited by their read length, with a maximum of 300 bp achievable in a single sequencing read. This severely limited the applicability of these technologies in complex crop genomes. The polyploid genomes of crops such as wheat or oilseed rape presented a challenge, as small sequencing reads often map to more than one locus in the genome due to high levels of homoeology between the subgenomes making up these species. Subsequently, third-generation sequencing or mapping technologies, able to address this major limitation of second-generation sequencing technologies, have recently gained popularity. The key difference distinguishing the third-generation from second-generation sequencing technologies is their enormous read lengths, achieved by the introduction of long-read single-molecule sequencing. The two dominating third-generation sequencing technologies to date are the single-molecule real time (SMRT) sequencing from Pacific Biosciences, PacBio (https://www.pacb.com/) and nanopore-based sequencing from Oxford Nanopore Technologies, ONT (https://nanoporetech.com/)(Sedlazeck et al. 2018). PacBio SMRT technology works via sequencing by synthesis, where a single-stranded circularized DNA molecule is introduced into a

sequencing well and replicated using a DNA polymerase. Unique fluorescence signals emitted by the addition of a specific nucleotide are captured by a high-resolution camera and translated into corresponding DNA base calls. The latest PacBio Sequel II can operate in two modes. The first, circular consensus sequencing (CCS) mode, is optimized for generating high fidelity (Hifi) reads. In CCS mode, a circularized DNA molecule is read in multiple sequencing passes, thereby creating a high accuracy consensus sequence from single reads while making a compromise on the read length and yield. On the contrast, in continuous long read (CLR) mode, each circularized DNA molecule is just read once, providing greater read length and yield but lower accuracy. PacBio has been successfully used to study complex polyploid genomes such as wheat (Zimin et al. 2017, Appels et al. 2018).

On the other hand, ONT sequencing works by capturing current perturbations caused when a DNA molecule passes through a nanopore embedded on a synthetic membrane. The smallest and most popular ONT sequencing device is called a MinION. In comparison to most next-generation sequencing machines, which range in price from around 100,000 EUR to over 1 Million EUR, the MinION is extremely cheap with a price of less than 2000 EUR. A single MinION flowcell contains 2048 nanopores and can produce between 10 to 50 Gb of sequencing data in a single sequencing run. It is well suited for small to medium scale experiments. The larger, more powerful ONT platform is known as PromethION. A PromethION flowcell contains 12,000 nanopores, can generate up to 100 Gb of sequencing data in a single sequencing run and is well suited for large genome centers. Both of the above mentioned ONT sequencing platforms can produce sequencing reads up to 1 Mb (Jain et al. 2018), while an average read length (depending on DNA quality and fragment length) of well over 10,000 kb is achievable under most circumstances. DNA sequence reads of this length can span the entire length of most indels and provide the necessary resolution required for detecting small to mid-scale SV. Since the short-read sequencing technologies have a very high false-positive rate (up to 89%) for SV detection (Mahmoud et al. 2019), small to mid-scale SV were almost invisible using short-read sequencing methods. Song et al. (2020) demonstrated the use of PacBio to detect SV in *B. napus*. They were able to identify 77.2 to 149.6 Mb of sequence affected by PAV in the rapeseed genome. Furthermore, the authors were able to recognize genomic regions associated with silique length, seed weight and flowering time by including PAV into their genome-wide association studies (GWAS) model. These associations were completely overlooked when using GWAS based only on single-nucleotide polymorphism (SNP) markers. This thesis reports a further example where

ONT was used for detecting small to mid-scale SV (Chapter 4). In this case, an astoundingly high proportion of all *B. napus* genes (up to 10%) were found to be affected by small to mid-scale SV events. Nearly half of these SV events ranged between 100 bp to 1000 bp.

1.5 Approaches to detect large-scale and small-scale SV

Several approaches combining various types of genotyping platforms have been used for detecting SV in plant genomes. Most of the studies to date focused on detection of large variants, ranging in size from hundreds of kb to a few Mb. Before the advent of cost-effective next-generation sequencing, inexpensive fluorescence *in situ* hybridization (FISH) and genomic *in situ* hybridization (GISH) methods were widely used for visualizing chromosomal scale SV (Xiong and Pires 2011, Chester et al. 2012, Snowdon 2007). However, these methods are only capable of identifying SV if huge blocks of chromosomes were rearranged. SNP arrays have been another popular choice for detecting genome SV controlling major agronomic traits, for example, flowering time (Schiessl et al. 2017a), seed quality (Stein et al. 2017) and stay-green traits (Qian et al. 2016) in rapeseed, boron toxicity in barley (Sutton et al. 2007), Aluminum tolerance in maize (Maron et al. 2013) and photoperiodicity in wheat (Nishida et al. 2013). SNP arrays rely on the hybridization of DNA fragments to small, 50 nucleotide probe sequences anchored onto a glass surface. The probes are designed to capture approximately 50 nucleotides of unique, non-polymorphic sequence adjacent to a pre-determined SNP site. Several software packages have been developed to quantify and infer CNV for humans (Colella et al. 2007, Korn et al. 2008) and plants (Grandke et al. 2016) from the fluorescence signals generated by a single-base extension on a SNP array. A major drawback of SNP arrays for SV analysis is the limited power of detection, especially for small to mid-scale SV, due to the frequently considerable physical distances between adjacent SNP markers. Furthermore, SNP arrays can introduce an ascertainment bias due to pre-determined design of the arrays, and homoeologous SNP assays are a serious limitation in complex polyploid plants like rapeseed. Next-generation sequencing (NGS) provides a good alternative to avoid ascertainment bias and gain the extra resolution required to identify small SV. Various methodologies have been developed for detection of SV from genome sequencing data, such as read depth (RD), paired read (PR) and split read (SR) analysis. Mace et al. (2013) demonstrated the use of high-coverage short read data to identify SV in sorghum. They were able to detect 1.9 million InDels and gene PAV associated with domestication and breeding in a panel of 44 genetically and geographically diverse *Sorghum*

bicolor accessions. In some cases there have been examples (Schiessl et al. 2017b) where only the protein-coding regions of the genome were sequenced to identify complex SV in genes. Although sequencing-based approaches led to a considerable increase in the resolution of SV detected in comparison to SNP arrays or cytogenetic methods such as FISH and GISH, small to mid-scale SV still remain challenging to detect due to the very high false-positive mapping rate (up to 89%) of short-read sequencing data (Mahmoud et al. 2019, Sedlazeck et al. 2018). In contrast, long-read sequencing platforms such as PacBio and ONT can be effectively used for detection of this size range of SV. Both of these technologies are capable of producing average read lengths of 10 kb or more, thereby spanning the entire length of SV and providing the necessary resolution for detecting small SV. ONT sequencing reads can achieve even longer reads than PacBio, with an N50 of up to 100 kb (Jain et al. 2018). Yang et al. (2019) identified 386,014 genomic rearrangements ranging from 10 to 99,330 bp between the PacBio genome assembly of a tropical small-kernel maize inbred line and two reference cultivars, B73 and Mo17. They also genotyped these SV in a panel of 521 diverse inbred lines using high-depth Illumina data and performed a GWAS. Using the SV data, the authors were able to find a significant new locus affecting oil concentration and long-chain fatty acid composition on chromosome 4, which was invisible using only a SNP-based analysis.

1.6 Pan-genomics vs. resequencing approaches for analysis of genomic variants

Pangenomes are constructed by comparing and identifying core and dispensable genes by *de novo* genome assemblies of several individuals within a species, or by iteratively assembling the sequencing reads that cannot be aligned to a reference genome assembly. Pangenomes from a comparative analysis of whole-genome assemblies are considered to be more robust than read mapping approaches, as *de novo* assemblies provide sufficient resolution to capture the entire size range of genomic rearrangements. However, given the very high read lengths of third-generation sequencing technologies, this raises the question as to whether it is vital to perform *de novo* assembly (in a pangenome type approach) to survey SV diversity in crop species. The long read lengths of third-generation sequencing technologies such as PacBio and ONT now enable identification of SV in the rapeseed genome with an unprecedented resolution (chapter 4). Long ONT and PacBio reads can span the entire length of small to mid-scale SV and enable accurate detection of these re-arrangements. However, large insertions, inversion duplications or translocations would result in the splitting of long reads into two or more alignments (Split-reads,

SR). It is extremely challenging to distinguish the SR due to real SV from the wrongly called genomic re-arrangements because of the errors in reference assembly. Therefore reliable detection of large, complex SV might still require whole-genome assemblies.

1.7 Aims of this study

There is increasing evidence that SV contributes to important agronomic traits in *B. napus*. In the past decade, various methodologies involving short-read sequencing or SNP genotyping arrays have been developed to detect genomic rearrangements in oilseed rape. However, most previous studies focused on the detection of large, chromosomal-scale SV events. This can be mainly attributed to the limited resolution of genotyping arrays or short-read sequencing platforms.

The overall aim of this thesis was to evaluate different methods for genome-wide detection and analysis of small to mid-size SV events in the *B. napus* genome and to elucidate the potential benefits of long-read sequencing for SV detection in a complex polyploid crop. The following major goals were pursued:

i) Exploration and evaluation of technologies and methods for the detection of SV in crop plants (Chapter 2).

ii) Evaluation of a case study for the impact of SV on complex trait expression in oilseed rape, based on an integrated approach combining SNP arrays, Illumina sequencing and optical mapping to identify gene PAV associated with resistance to the fungal pathogen *Verticillium longisporum* (Chapter 3).

iii) Use of long sequencing reads to evaluate the role of small to mid-scale SV in eco-geographical diversification of *B. napus* into the three predominant ecotypes (winter-type, spring-type and semi-winter type rapeseed), and survey the extent and impact of small to mid-size genome-wide SV on genes (Chapter 4).

iv) Provide first insight and ideas about how new long-read sequencing technologies can help to understand complex SV in large plant genomes by providing additional layers of information, such as methylation signatures, chromatin confirmation, or data from target enrichment strategies implementing long-read sequencing, and describe potential cellular mechanisms that might explain the occurrence of small to mid-scale SV in oilseed rape (Chapter 5).

2 Connecting genome structural variation with complex traits in crop plants

Gabur I[#]; **Chawla H.S**[#]; Snowdon R.J.; Parkin, I.A. P
*Theoretical and Applied Genet*ics (2019), vol 132**:** 733–750
doi:10.1007/s00122-018-3233-0
[#] These authors contributed equally to this work.

Theoretical and Applied Genetics (2019) 132:733–750
https://doi.org/10.1007/s00122-018-3233-0

REVIEW ARTICLE

Connecting genome structural variation with complex traits in crop plants

Iulian Gabur[1] · Harmeet Singh Chawla[1] · Rod J. Snowdon[1] · Isobel A. P. Parkin[2]

Received: 15 August 2018 / Accepted: 7 November 2018 / Published online: 17 November 2018

Abstract

Key message **Structural genome variation is a major determinant of useful trait diversity. We describe how genome analysis methods are enabling discovery of trait-associated structural variants and their potential impact on breeding.**

Abstract As our understanding of complex crop genomes continues to grow, there is growing evidence that structural genome variation plays a major role in determining traits important for breeding and agriculture. Identifying the extent and impact of structural variants in crop genomes is becoming increasingly feasible with ongoing advances in the sophistication of genome sequencing technologies, particularly as it becomes easier to generate accurate long sequence reads on a genome-wide scale. In this article, we discuss the origins of structural genome variation in crops from ancient and recent genome duplication and polyploidization events and review high-throughput methods to assay such variants in crop populations in order to find associations with phenotypic traits. There is increasing evidence from such studies that gene presence–absence and copy number variation resulting from segmental chromosome exchanges may be at the heart of adaptive variation of crops to counter abiotic and biotic stress factors. We present examples from major crops that demonstrate the potential of pangenomic diversity as a key resource for future plant breeding for resilience and sustainability.

Introduction: the discovery of structural variation

With rapidly increasing sophistication in genome analysis technologies, there is growing evidence that genome-wide structural variation (SV) is a major factor underlining observed phenotypic variation in eukaryotic organisms. The first report of genic SV affecting a phenotype dates back more than 80 years, when Bridges (1936) discovered that a duplication of the *Bar* gene is associated with small eyes in the fruit fly, *Drosophila*. Genomic rearrangements have been studied extensively in humans due to their association with a range of diseases. Particularly copy number variation (CNV), an important class of structural variation, has been discovered to be causal for various autoimmune disorders (Mamtani et al. 2010), including susceptibility to human immunodeficiency virus (HIV) infection (Gonzalez 2005), Parkinson's disease (Singleton 2003) and Alzheimer's disease (Rovelet-Lecrux et al. 2006; Escaramís et al. 2015). Mounting evidence supporting the importance of SV in human genetics led to the study of the same phenomena in animal species, where numerous examples have been discovered for a role of SV in important traits, for example in mice (Keane et al. 2014), cattle (Fadista et al. 2010), pigs (Esteve-Codina et al. 2013), sheep (Liu et al. 2013) and horses (Ghosh et al. 2014; Wang et al. 2014). Structural variations were initially thought to be rare in plants, but this perspective changed dramatically with the realization that almost all flowering plants derived from multiple rounds of ancient or recent polyploidization (Vialette-Guiraud et al. 2011; Van de Peer et al. 2009; Alix et al. 2017). The ability to generate reference genome sequences even for complex crop plant genomes (Edwards et al. 2013) combined with decreased costs associated with de novo genome assembly and resequencing have accelerated the study of SV (Voss-Fels and Snowdon 2016). Numerous recent reports have clearly demonstrated that both small and large genomic

Communicated by Rajeev K. Varshney.

Iulian Gabur and Harmeet Singh Chawla contributed equally to this work.

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

[1] Department of Plant Breeding, Justus Liebig University, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany

[2] Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S7N OX2, Canada

rearrangements can cause major phenotypic variance affecting an array of important traits in crops (Saxena et al. 2014; Neik et al. 2017; Żmieńko et al. 2014; Schiessl et al. 2017a).

Diversity of structural variants

Genome structural variants occur in diverse forms including translocations, inversions, insertion/deletion polymorphisms (InDels), copy number variation (CNV), or simple variation in microsatellite repeat number. Traditionally, InDels have been defined as short presence/absence nucleotide polymorphism ranging from 1 to 50 bp in length, whereas a variable number of copies for larger DNA segments, ranging from a few hundred bp to several kb, is generally referred to as CNV. Gene CNV represents the most intensively studied class of SV associated directly with trait variation, whereby variants affecting intergenic regions, splicing variants and/or regulatory factors could also infer SV–trait associations. This can be mainly attributed to their ease of detection using simple molecular biology methods. Presence–absence variation (PAV) represents an extreme form of CNV where whole genomic segments are deleted from individuals within a population (Saxena et al. 2014). Different kinds of SV can occur independently or simultaneously, resulting in complex genome alterations. Many important crop genomes arose from multiple polyploidy events, in some cases involving widespread recombination among homoeologous (related but non-homologous) chromosomes. Such exchanges can result in both reciprocal or non-reciprocal exchanges. The latter, often referred to as homoeologous non-reciprocal transpositions, or HNRT (Parkin et al. 1995; Pires et al. 2004; Gaeta and Chris Pires 2010), can lead to loss or gain of DNA fragments on related chromosome homoeologues and consequently to PAV and CNV. As described in more detail later in this review, examples in recent allopolyploids like *Brassica napus* have demonstrated that this kind of exchange during early rounds of polyploidization can be a key driver of modern crop genome diversity and phenotypic plasticity (Chalhoub et al. 2014; Samans et al. 2017; Hurgobin et al. 2017).

Origins of SV

Various cellular mechanisms can trigger generation of SV during meiotic or mitotic cell division. SV events are caused by recombination errors, like non-allelic homologous recombination (NAHR) (Lupski 1998), DNA break repair errors, such as non-homologous end joining (NHEJ) (Moore and Haber 1996), or replication errors, including fork stalling and template switching (FoSTeS) (Lee et al. 2007) and microhomology-mediated break-induced replication (MMBIR) (Hastings et al. 2009). NHEJ can be triggered by misguided fusion of double-strand breaks in DNA, often resulting in insertions and/or deletions; however, in rare cases NHEJ might also generate translocations (McVey and Lee 2008). FoSTeS/MMBIR is another cellular mechanism causing major structural variations (for example large rearrangements, inversions, duplications and translocations) ranging in size from a few kb to several Mb and involves fork stalling and polymerase switching at a nearby single-stranded DNA (Stankiewicz and Lupski 2010). The most likely cause of much of the CNVs observed in plants is NAHR, which is largely the result of misalignment in genomic regions housing highly identical sequences, such as repetitive DNA, leading to duplication or deletion of genomic segments and thus copy number variants. Segmental duplications appear when highly homologous genomic regions (more than 95%) are physically positioned at distances from a few kb to some Mb from one another. Furthermore, depending on the orientation of the homology, NAHR could also cause deletions (upstream orientation on the same chromosome), inversion (downstream orientation on the same chromosome) and translocation (located on different chromosomes) (Sharp et al. 2006). The abundance of repetitive sequences in plant genomes varies widely, with published frequencies ranging from around 10% in Arabidopsis (The Arabidopsis Genome Initiative 2000) to more than 85% (in wheat) (Appels et al. 2018). The prevalence of repetitive DNA, in particular in larger crop genomes, could promote the generation of dosage effects for particular sets of genes, partly explaining the large adaptive phenotypic variation existing within the plant kingdom.

Changes in ploidy can also lead to generation of SV in plants. The majority of angiosperms studied to date show evidence of polyploidization and/or whole-genome duplication in their evolutionary history, and most modern crop species have undergone recent genome duplication events that are now known to have played a significant role in dictating their path to adaptation (Fig. 1). Some major crops contain multiple copies of entire chromosomes from spontaneous genome duplication of the same species, for example autopolyploid potato (*Solanum tuberosum*; $2n = 4x = 48$), while others arose from interspecific hybridization of subgenomes among distinct, yet related species, for example allohexaploid wheat (*Triticum aestivum*; $2n = 6x = 42$) or allotetraploid oilseed rape/canola (*Brassica napus*; $2n = 4x = 38$). Many older crop species arose by ancient duplications and paleopolyploidization. For example, the diploid cabbage species *Brassica oleracea* ($2n = 2x = 18$) and *Brassica rapa* ($2n = 2x = 20$) represent paleohexaploids which have returned to a diploid state by genome fractionation (Lagercrantz et al. 1996; Tang et al. 2012; Parkin et al. 2014).

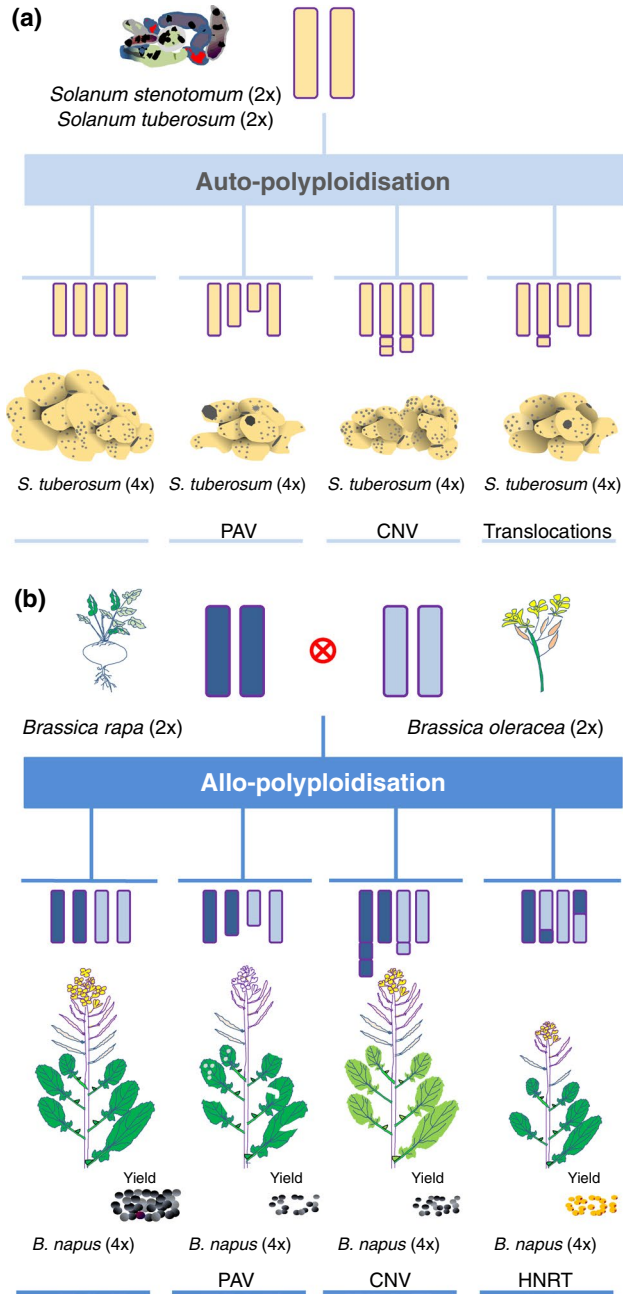


Fig. 1 Origins of different kinds of structural variants in autopolyploid and allopolyploid crops from segmental chromosome rearrangements, illustrated by the coloured bars with examples from **a** autopolyploid potato (*Solanum tuberosum*, 2n = 4x = 48) and **b** allopolyploid rapeseed/canola (*Brassica napus*, 2n = 4x = 38). Autotetraploid *S. tuberosum* arose from a spontaneous genome duplication (auto-polyploidization) of the diploid progenitor *S. stenotomum*/*S. tuberosum* (2n = 2x = 24), while *B. napus* arose from interspecific hybridization between the diploid progenitor species *B. oleracea* (2n = 2x = 18) and *B. rapa* (2n = 2x = 20). SV linked to adaptive and agronomic diversity is represented as presence–absence variation (PAV), copy number variation (CNV), translocations and homoeologous non-reciprocal transpositions (HNRT)

Visualization of large-scale SV

Classical cytology first identified evidence of large-scale chromosomal aberrations in cereals (e.g. Sears 1939), which were later confirmed as translocations using early molecular marker technologies (e.g. Gale and Devos 1998). Comparative genomic hybridization (CGH) was one of the very first methods to visualize large-scale SVs. With CGH it is possible to detect and map relative DNA sequence copy number between genomes, by hybridizing fluorescently labelled DNA from each source genome to metaphase chromosome spreads or genome-wide sequence arrays. An increase or decrease in copy number of genomic DNA (corresponding to segmental SV) can then be detected by measuring the florescence ratios between the two coloured fluorophores (Kallioniemi et al. 1992). The resolution of CGH via in situ hybridization is relatively low, with segmental SV events only visible at megabase scale, whereas array-based CGH (aCGH) can resolve smaller SV events down to a few kilobases in size. For example, Yu et al. (2011) were able to detect 641 CNVs ranging from 1.1 to 180.7 kb between two rice cultivars using aCGH. Large SV events can also be visualized directly at the chromosome level using molecular cytogenetic techniques such as fluorescence in situ hybridization (FISH) and genomic in situ hybridization (GISH) (Xiong et al. 2011; Chester et al. 2012; Horn et al. 2002; Snowdon 2007). These techniques allow physical analysis of chromosomes using chromosome arm ratios, mapping of heterochromatic regions, bacterial artificial chromosome probes (BAC-FISH) containing specific repeat sequences or molecular karyotyping of chromosome-specific probes (Xiong and Pires 2011). FISH has been used successfully in maize to analyse B chromosome non-disjunction due to *r-X1* deficiency (Tseng et al. 2017). Furthermore, the technique has been used to map homoeologous exchanges associated with agronomic traits in polyploid crop genomes. Stein et al. (2017) used BAC-FISH to identify homoeologous exchanges between two *B. napus* chromosomes associated with a QTL for seed fibre content. In potato, FISH was used to identify CNV associated with plant growth and developmental traits (Iovene et al. 2013), while Ali et al. (2016) used FISH in wheat to validate introgression of alien DNA segments that led to mosaic virus resistance. Before the availability of cheap, high-throughput genome sequencing, hybridization methods provided a relatively simple and low cost option for visualizing large SV events at the chromosomal level; however, finer resolution is required for detection of smaller SV events. With sufficient sequence depth, approaches based on next-generation sequencing (NGS) technologies provide an ideal solution.

Sequencing-based SV detection

Next-generation sequencing (NGS) approaches have accelerated the process of assembling plant reference genomes to a speed and accuracy that was unimaginable a decade ago. Furthermore, availability of methods to detect single nucleotide differences between genomes using whole-genome sequencing data, high-coverage exome sequence data or sequence capture data has been a major breakthrough in deciphering complex SV (Chen et al. 2008; Schiessl et al. 2017b). One of the key advantages of NGS-based methods for SV detection is the resolution that can be achieved by using such approaches. To date, sequenced reference genomes of varying quality are available for over 200 plant species, including most major crops (see http://www.plabipd.de/timeline_view.ep for an up-to-date overview of published plant reference genomes). As the quality of reference genome assemblies for more complex genomes continues to improve, for example by utilization of new methods like chromatin conformation associations (e.g. Mascher et al. 2017 or long-read single-molecule sequencing (Jiao et al. 2017), our ability to utilize genome-wide or targeted resequencing techniques for SV analysis in large populations will become even more powerful. Early whole-genome resequencing studies in major crops with relatively simple genomes, like sorghum, used reference-based read-mapping approaches to identify genome-wide SNP and small-scale SV. For example, Mace et al. (2013) identified 1.9 million InDels, including specific gene PAV associated with domestication and breeding, in high-coverage resequencing data from 44 genetically and geographically diverse *Sorghum bicolor* accessions. Different approaches have been developed for characterization of SVs from NGS reads, including combinations of read depth (RD), paired read (PR) and split read (SR) analysis along with de novo sequence assemblies in order to address more complex genomic re-arrangements.

Algorithms for RD analysis rely on the density of sequenced reads aligned to a locus in a reference genome for CNV identification (Alkan et al. 2009; Li and Olivier 2013). In crops, RD approaches focused mainly on calling of large SV, for example in resequencing data from sorghum (Zheng et al. 2011) or rapeseed (Samans et al. 2017). RD-based methods can detect deletions and duplications very effectively. However, limitations of the read length and the quality and coverage of the available reference genomes reduce the efficacy of this approach for detecting insertions or translocations. Furthermore, it should be noted that RD-based approaches are highly sensitive to library preparation methods. For example, PCR amplification during the sequencing library preparation can lead to either over-representation or complete absence of certain genomic regions that are difficult to amplify, which would be interpreted as duplication and deletion events, respectively, using an RD sensing algorithm. Therefore, stringent normalization of read depth is required to remove any kind of bias introduced by the library preparation. These limitations can be somewhat overcome by paired-end sequencing of single DNA fragments (paired reads). Since the sequencing library is enriched for a particular fragment size, the distance between the paired reads is pre-defined. Any insertion or deletion between the paired reads would result in a deviation from the expected mapping distance; hence, abnormally mapped read pairs might indicate the presence of SV (Korbel et al. 2007). Different types of SV can be mapped using paired reads, for example deletions or insertions (when paired reads align further apart or closer than expected), inversions (when the orientation of paired reads is inverted) or translocations (when each of the paired reads maps to different chromosomes). This approach is still highly dependent on the read coverage, size and number of repetitive elements in the genome and the quality of the reference genome, and paired read methods are best suited for detection of medium-sized insertions and deletions. However, they might not be the best choice for identification of small insertion or deletion events, due to the difficulty in distinguishing small deviations in read–pair distance from technical errors. Split read algorithms provide an alternative which also makes use of paired-end sequencing. Split read algorithms depend on accurate mapping of one of the reads from a pair, while the other read maps only partially to the reference genome (Ye et al. 2009). When reads align right across a SV break point, precise calling of breakpoints can be achieved. However, with short read NGS technologies this type of approach is only useful for detecting small-sized SV (Ye et al. 2009; Schröder et al. 2014). New opportunities to overcome these problems using long-read sequencing are described later in this review.

One major bottleneck of the methods described above is the availability of high-quality reference genomes. De novo genome sequence assembly provides the optimal method for fine-scale SV detection, but until now assembly based pangenome approaches have been largely prevented by high cost and time constraints (Hajirasouliha et al. 2010). However, costs can be significantly reduced using reduced-representation sequencing approaches which only address part of the whole genome. Reduced-representation sequencing can be achieved either by selection of restriction fragments for sequencing or by designing baits to capture certain interesting regions of the genome. Whole-exome sequencing is an example of such an approach which reduces computing and sequencing costs by focusing only on protein-coding regions. This reduces the capacity to detect large SV, but can potentially identify causal CNV when sequencing coverage

is sufficient. Exome capture has not yet been used extensively in crops, but recently a capture array was developed for barley to assay species-wide sequence diversity and SV (Mascher et al. 2013). Alternatively, targeted gene sequencing provides opportunities to capture sequence variants for specific panels of target genes, for example for QTL regions (e.g. Clarke et al. 2013) or specific biological pathways (Schiessl et al. 2017a). However, sequence capture does rely on hybridization capture and amplification steps which raise costs of library preparation and can also lead to normalization problems which must be dealt with during data analysis.

Because each method has limitations, a pragmatic approach is to use a combination of SV detection methods (Escaramís et al. 2015; Alkan et al. 2011). However, accurate and unique alignment of short sequence reads to a reference assembly is the foundation of almost every SV detection pipeline. This is extremely challenging in the case of polyploids due to the high homology between their subgenomes. The majority of the crop species reference genomes published to date are themselves based on short read sequencing, containing in some cases thousands of contigs and scaffolds that are not assembled to chromosome level due to the repetitive and complex nature of most crop genomes. The development of third-generation sequencing technologies which generate long-range sequences and enable longer, contiguous scaffolds provide new opportunities for reliable, cost-effective *de novo* assembly at whole-chromosome level (Jiao and Schneeberger 2017). In various research applications, long-read sequencing has become an efficient alternative for SV mapping and phasing. The long-read sequencing platforms from Pacific BioSciences (Menlo Park, CA, USA) and Oxford Nanopore Technologies (Oxford, UK) can provide read lengths ranging from 10 to 150 kb (Schmidt et al. 2017) depending on the DNA library quality. The sequencing error rate for both these platforms is higher compared to short read methods like Illumina sequencing. However, because the sequencing errors are randomly distributed this limitation can be overcome by increasing the depth of sequencing (Schiessl et al. 2018). By spanning rearrangement endpoints and providing more accurate reference assemblies, both of these sequencing platforms enable discovery of complex SV events which were extremely challenging to detect using only short-read methods (English et al. 2015; Chaisson et al. 2014).

Alternative technologies such as optical mapping (Bionano Genomics, San Diego, CA USA) (Lam et al. 2012) or linked-read technologies (10x Genomics, Pleasanton, CA, USA) (Mostovoy et al. 2016), which allow long distances to be effectively spanned in complex genomes, have also contributed substantially to SV detection. Even in challenging polyploid crop genomes, combinations of these different approaches can provide base pair resolution to the study of SV. Unfortunately, these high-resolution techniques are still relatively expensive, meaning that high-throughput analysis in large populations is still prohibited by cost. Until this changes, cheaper high-throughput methods like comparative genomic hybridization (CGH), single nucleotide polymorphism (SNP) arrays or real-time multiplex PCR may be viable alternatives in order to study trait-associated variants in large populations.

Analysing SV using SNP genotyping arrays

High-density SNP arrays provide a popular and cost-effective solution to analyse genetic differences among many individuals within a species. Presently the most popular platform for SNP genotyping is the Infinium™ assay from Illumina (San Diego, CA USA), which relies on hybridization of genomic fragments to probe sequences anchored in flowcells on a glass surface, with probes designed to capture approximately 50 nucleotides of unique, non-polymorphic sequence adjacent to a pre-determined SNP site. This is followed by a single-base extension using hapten-labelled nucleotides and generation of fluorescence signal by adding fluorescently labelled antibodies (Mason et al. 2017). Development of algorithms that can detect SV by quantifying the relative light intensities generated during a SNP call has been an area of research for many years. In human genetics, algorithms like QuantiSNP (Colella et al. 2007) and Birdsuite (Korn et al. 2008) use the fluorescent signal intensity of one allele relative to the other to infer a duplication or deletion event. For polyploid crop plants, the R package "gsrc" (for "genome-wide structural rearrangement calling") was developed to call rearrangements using SNP intensity information (Grandke et al. 2016). A wide range of SNP arrays have been developed in multiple crops for use in crop breeding and genetic research (Voss-Fels and Snowdon 2016).

Despite the widespread use of SNP arrays, there are some inherent problems associated with them when it comes to SV detection. The greatest problem is the limited power of detection of small SVs, due to poor resolution and ascertainment bias due to the pre-determined design of the arrays. PCR-based methods provide a simple and cheap alternative to SNP arrays, especially for detection of SVs at ultra-high resolution; however, prior knowledge of regions of interest is required and throughput is limited. Quantitative real-time PCR (qRT-PCR) and digital PCR (dPCR) are further methods capable of efficiently identifying small SV (InDels) but also translocations, inversions and CNV (Schiessl et al. 2017a; Qian et al. 2016; Ma and Chung 2001). Genes for different traits affected by PAV, CNV and InDel polymorphisms have been validated in a number of major crops using PCR, for example flowering time genes (Schiessl et al. 2017a), lignin biosynthesis genes (Stein et al. 2017)

and a stay-green gene (Qian et al. 2016) in rapeseed, copies of the boron toxicity tolerance gene *Bot1* in barley (Sutton et al. 2007), the aluminium tolerance gene *MATE1* in maize genotypes (Maron et al. 2013) and InDels in the wheat photoperiodicity genes *Ppd-A1a Ppd-B1a* (Nishida et al. 2013).

Crop pangenomes as a future reference paradigm

The unprecedented low cost and high throughput of DNA sequencing today makes it possible to generate genome sequence data for hundreds or thousands of individuals within a species. This provides a new wealth of data to discover genomic re-arrangements in crop genomes in the form of CNVs and PAVs. Insights into genomic SV have conclusively established that a single reference assembly cannot reflect the entire diversity within a species. This gave rise to the concept of pangenomes, which ideally represent all structural genome diversity present in a species. Originally coined for analysis of bacterial genomes (Tettelin et al. 2005), the pangenome concept was first adapted to plants after comparative sequencing of grass genomes revealed widespread structural variation on a previously unknown scale (Morgante et al. 2007). Since the first crop genomes became available, the pangenome concept has been investigated at many different levels especially in maize (Morgante et al. 2007; Springer et al. 2009; Lai et al. 2010; Chia et al. 2012; Hirsch et al. 2014). Most pangenome analyses so far have focused primarily on differences in gene content between individuals in a species; however, as de novo genome assemblies become more feasible in more complex organisms there is growing scope for assembly scale pangenome analysis.

A pangenome for any species is considered to comprise a so-called "core" genome, comprising all genes common to all individuals within the species, along with a "dispensable" genome consisting of partially shared genic regions that are present in some individuals but absent from others (Tettelin et al. 2005). To maximize discovery and coverage of the dispensable genome component, a pangenome should ideally include data from a broad range of individuals representative of all diversity present in the species. Pangenomic description of SV is best achieved by assembly based approaches, but due to their prohibitive cost for large and complex genomes the early pangenomes for most crop species have been generated by genomic resequencing (or in some cases skim sequencing) of representative diversity and analysis by techniques to detect and place SV in existing reference assemblies. This type of approach generally comprises three major steps: resequencing reads are mapped to a high-quality reference assembly, unmapped reads are independently assembled into additional contigs and these are inserted at the appropriate positions in the original assembly using end alignments and/or genetic mapping data. Although faster and cheaper than a de novo assembled pangenome, this method relies strongly on a high-quality reference assembly and can only capture SV in contigs that are able to be successfully assembled and placed from unmapped reads. Nevertheless, such approaches can provide cost-effective opportunities to efficiently capture genic CNV and PAV (Golicz et al. 2016a; Montenegro et al. 2017; Zhou et al. 2017).

To date there are only a handful of studies in which crop pangenomes have been created by de novo assembly of diverse individuals. The most comprehensive study so far was a pangenomic analysis of genomic variation in cultivated and wild rice (Zhao et al. 2018) in which whole-genome de novo assemblies were generated for 66 diverse genotypes chosen to represent a panel of 1529 accessions across the *Oryza sativa–Oryza rufipogon* species complex. The resulting rice pangenome identified 26,372 core genes and 16,208 dispensable genes, enabling associations of SV signatures across the pangenome to domestication sweeps and other signals of natural and artificial selection. Interestingly, several important known genes which were not observed in the original Nipponbare reference genome sequence, including the submergence tolerance1 genes *Sub1A* (Xu et al. 2006) *SNORKEL1* and *SNORKEL2* (Hattori et al. 2009), and the phosphorus-deficiency tolerance gene *Pstol* (Gamuyao et al. 2012) were discovered in the pangenome sequence (Zhao et al. 2018).

These findings reflect observations from Samans et al. (2017) in allotetraploid *B. napus* that genes involved in stress adaptation responses are particularly prevalent among genome structural variants resulting in CNV and PAV, underlining the key role of SV in crop adaption and breeding selection. Similarly, in hexaploid wheat, for which the first high-quality whole-genome reference assembly was recently published (The International Wheat Genome Sequencing Consortium 2014), a resequencing-based pangenome study including 18 wheat cultivars revealed an average of 128,656 genes per cultivar, of which 64% were found to be present in all cultivars and 49,952 genes were dispensable (Appels et al. 2018). Again, annotation of the variable set of genes revealed an enrichment for genes involved in environmental stress and defence response. Assembly based approaches to pangenome analysis will further refine these initial studies as they become more feasible with new assembly strategies and long-read sequencing capabilities. In the foreseeable future, assembly based pangenome analysis is likely to become the method of choice for generating and analysing reference genome data, even in crops with large, complex genomes like those of barley (Stein and Mascher 2019) or wheat. In a pangenome analysis based on *de novo* assemblies for wild relatives of soybean (*Glycine soja*), Li et al. (2014) found high variation in a dispensable genome

comprising around 20% of all assembled sequences, with CNV and mutations in dispensable genes showing evidence for positive selection and a strong influence on important agronomical traits. McHale et al. (2012) found previously that CNV and PAV between wild and domesticated soybean affect over 800 genes involved in biotic stress resistance, and detailed assemblies of wild vs. cultivated forms can deliver important sequence information with regard to potentially important genes that may be absent from reference cultivars.

In contrast to pangenome assembly approaches, which can miss genes not picked up by algorithms for prediction of open reading frames (ORF), some authors advocate the use of transcriptomics-based approaches as a cost-effective way to circumvent this problem. For example, He et al. (2015) introduced the concept of an ordered transcriptome for the allopolyploid *B. napus* based on gene models from its diploid progenitors *B. rapa* and *B. oleracea*, and the homoeologous diploid pan-transcriptomes as a reference to visualize SV in genetically diverse *B. napus* accessions using mRNAseq data (He et al. 2016). Such approaches provide a clear visual impression of the high degree of SV in recent polyploid crop genomes (Fig. 2). Hirsch et al. (2014) took a transcriptomics-based approach to assemble a pangenome for maize. Using this approach, they succeeded in identifying 8681 representative transcript assemblies (longest transcripts within the respective loci) which did not map to the B73 reference, 83% of which mapped only in subsets of 503 investigated lines and can be considered as dispensable genes.

Lu et al. (2015) used an alternative approach for an assembly independent pangenome analysis in maize, using linkage information to map 26 million sequence tags generated by reduced-representation sequencing of 14,129 maize inbred lines. A total of 4.4 million tags with high-confidence map

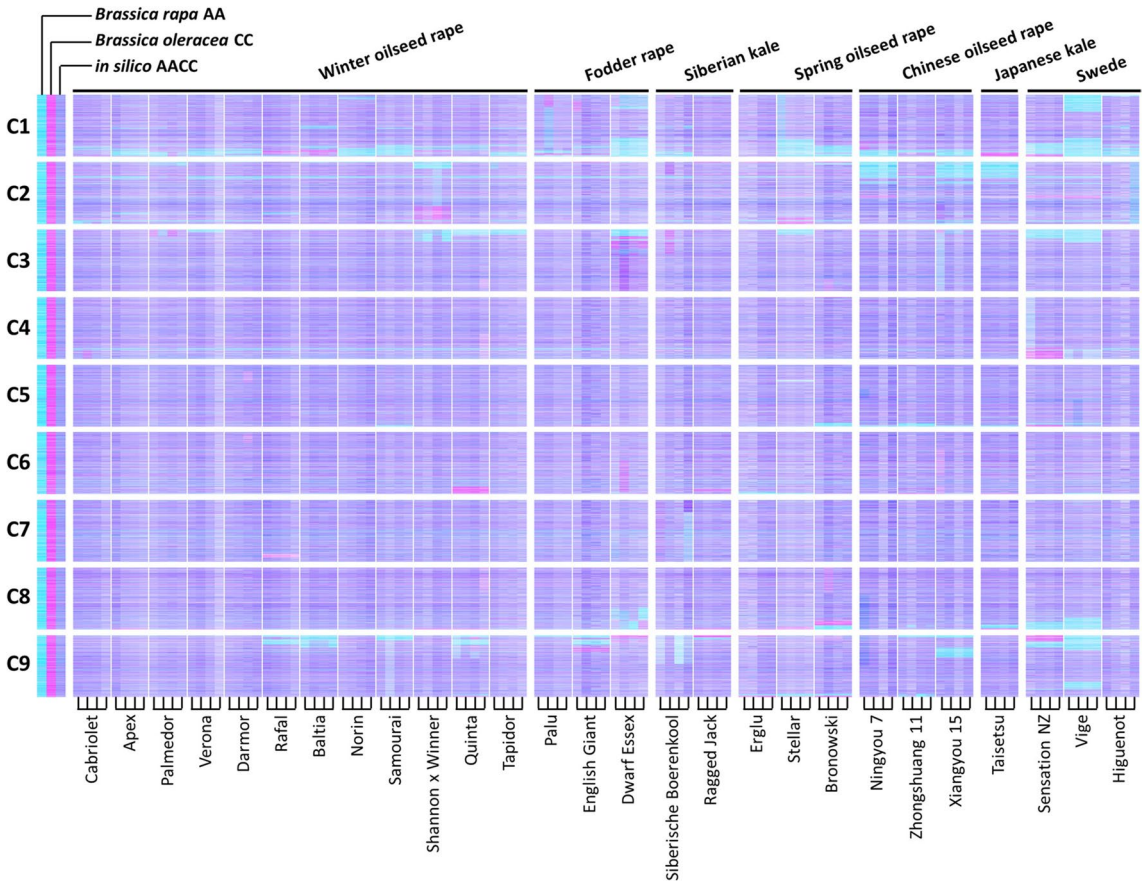


Fig. 2 Visualization of extensive structural variation (SV) caused by homoeologous genome exchanges between the A and C subgenomes of the allopolyploid crop species *Brassica napus* (oilseed rape), using Transcriptome Display Tile Plots derived from mRNAseq reads mapped to an ordered pan-transcriptome. The relative transcript abundance of homoeologous gene pairs is represented in CMYK colour space, with cyan component representing transcript abundance of the A-subgenome copy and magenta component representing transcript abundance of the C-subgenome copy. The pairs are plotted in *Brassica* C genome order (chromosomes denoted C1 to C9) for four biological replicates of each of 27 accessions of *B. napus* and controls comprising parental species and their in silico combination. Image reproduced from He et al. (2015; https://doi.org/10.1111/pbi.12657) under the terms of the Creative Commons Attribution licence 2.0

positions were selected as anchors for a high-density pangenome map. One quarter of these anchors represented PAV and showed enriched associations with phenotypic traits, providing a basis to discover genes where SV is involved in maize adaptation and agronomy. This example shows the power of combinatory approaches involving low-cost, high-throughput sequencing and population genetic analysis to define and analyse SV. Such techniques can potentially also be applied in species without extensive genomics resources.

The Brassicaceae (Cruciferae) family represents an important crop model for studying polyploidy and genomic structural re-arrangements (Mason and Snowdon 2016). Present-day allopolyploid *Brassica* crops originated by interspecific hybridization between different diploid progenitors, for example *B. napus* was formed by hybridization between *B. oleracea* and *B. rapa*. Because the diploids are themselves closely related paleopolyploids with high homoeology between their genomes, synthetic hybrids among them undergo extensive genome restructuring due to inter-homoeologue pairing during the early rounds of meiosis, leading to extensive SV (e.g. Samans et al. 2017; Zou et al. 2018). It might be reasonable to hypothesize that corresponding processes during ancient polyploidization had a similar influence on genome-wide SV and adaptive diversity in *Brassica* spp., giving rise to substantial PAV and CNV observed in present-day diploid cabbage species: such events have been found to have particularly profound effects on genes involved in biotic stress responses in *B. oleracea* (Golicz et al. 2016b) or phenylpropanoid biosynthesis in *B. rapa* (Lin et al. 2014).

Structural variation and trait diversity in major crops: key examples

Table 1 provides examples for demonstrated associations of SV to plant phenotypes in crop species. Wheat is one of the most complex plant genomes due to its large size and polyploid nature. As a result, there has been considerable delay in the detailed genomic analysis of wheat. However, early genetic mapping studies already showed that rearrangements on a number of chromosomes impact numerous important genes for resistance and adaptation traits (Nelson et al. 1995). It was also known for some time that some genes duplicated via polyploidy have remained unaltered, whereas others have been deleted or rendered non-functional by transposon insertions or premature stop codons (Gu et al. 2004). Major translocations in wheat have been associated with specific geographical regions (Riley et al. 1967; Belay and Merker 2004, 2006; Ma et al. 2015) and associated with adaptive and biotic resistance traits (Liu et al. 2016; Law and Worland 2006). With growing access to gene and sequence data, the influence of

Table 1 Examples for structural variations with demonstrated effects on agronomic traits in different crop species

Species	Type of variant	Traits associated	Reference
Barley (*Hordeum vulgare*)	CNV	Boron toxicity tolerance	Sutton et al. (2007)
	CNV	Disease resistance	Muñoz-Amatriaín et al. (2013)
Maize (*Zea mays*)	PAV, CNV	Domestication	Springer et al. (2009)
	CNV	Disease response, heterosis	Beló et al. (2010)
	CNV	–	Swanson-Wagner et al. (2010)
	CNV	Breeding selection	Jiao et al. (2012)
	CNV	Aluminium tolerance	Maron et al. (2013)
Rice (*Oryza sativa*)	PAV, CNV	Grain size, disease resistance	Xu et al. (2012)
	CNV	Disease resistance	Yang et al. (2013), Yu et al. (2013)
	InDel	Root system architecture	Uga et al. (2013)
Soybean (*Glycine max*)	PAV, CNV	Stress responses	Haun et al. (2010), McHale et al. (2012)
	CNV	Disease resistance	Lee et al. (2015)
Sorghum (*Sorghum bicolor*)	PAV, CNV	Disease resistance	Zheng et al. (2011), Mace et al. (2014)
Wheat (*Triticum aestivum*)	CNV	Vernalization, flowering time	Díaz et al. (2012), Würschum et al. (2015)
	CNV	Plant height	Li et al. (2012)
	PAV	Heading date	Nishida et al. (2013)
	CNV	Frost tolerance	Sieber et al. (2016)
	CNV	Winter hardiness	Würschum et al. (2016)
Oilseed rape (*Brassica napus*)	PAV, CNV	Flowering time	Schiessl et al. (2017b)
	HE	Seed fibre	Stein et al. (2017)
	PAV	Stay-green	Qian et al. (2016)
	PAV	Disease resistance	Gabur et al. (2018)

CNV due to polyploidization on adaptive traits like flowering time has been elucidated in more detail. For example, Díaz et al. (2012) found that variation in flowering behaviour in commercial wheat cultivars resulted from CNV for the photoperiodicity gene *Ppd-B1* and the vernalization gene *Vrn-A1*, rather than direct DNA mutations. An increase in the copy number of *Ppd-B1* was found to be associated with an early-flowering, day-length neutral phenotype, whereas plants with a higher *Vrn-A1* copy number exhibited an increased vernalization requirement. In another example, Würschum et al. (2016) reported that copy number variation of *C-repeat Binding Factor* (CBF) genes at the *Fr-A2* locus was the pivotal component for winter hardiness in a panel of 407 European winter wheat cultivars.

In addition to inter-homoeologue chromosome exchanges, interspecific hybrids of wheat with related grasses have led to rich cytogenetic stocks with segmental chromosome insertions or translocations, with a particular focus on resistance traits (Friebe et al. 1996; Wulff and Moscou 2014). One of the most well-known events is the 1BS/1RS translocation from rye, which increased drought adaptation and promoted yield performance of spring wheat in dryland production systems (Villareal et al. 1995; Reeves et al. 1999). However, the exact molecular basis of this improvement is still elusive. On the other hand, gene CNV has also been shown to shape other important phenotypic traits such as plant height in wheat. In cultivars carrying the semi-dwarfing genes *Reduced height* (*Rht*)*-B1b* and *Rht-D1b*, previously uncharacterized CNV polymorphisms of *Rht-D1* were reported to be causal for extreme dwarf phenotype, while a 90 bp insertion in *Rht-B*1 also contributed to severe dwarfism (Pearce et al. 2011). Another critical factor affecting wheat yield is photosynthetic activity determined by chlorophyll content. CNV in the wheat cytokinin oxidase gene *Tackx4*, which influences chlorophyll content and chloroplast stability via modulation of cytokinin concentration, was found to be associated with the chlorophyll content after anthesis as well as grain weight in 102 wheat varieties (Chang et al. 2016).

In barley, several studies have described gene copy number polymorphisms associated with environmental adaptation. As in wheat, CNV in the *H. vulgare CBF* orthologue at the *Frost Resistant-2 locus* (*FR-2*) was found to confer frost tolerance, with an increase in *CBF* coding sequences in winter barley compared to spring forms (Knox et al. 2010; Francia et al. 2016). Similarly, dosage effects from an increase in the number of *H. vulgare* boron transporter (*Bot1*) gene copies were found to confer boron toxicity tolerance (Sutton et al. 2007). Muñoz-Amatriaín et al. (2013) found that CNV between the barley cultivars Barke and Morex was particularly prominent for disease resistance proteins and protein kinases, while increased levels of CNVs were observed for wild accessions in comparison with cultivated barley. As for the examples mentioned above, these studies suggest a key role of SV in conferring the genome plasticity needed for adaption of barley to diverse environmental conditions.

In oilseed rape/canola, anomalies in marker segregation in mapping populations displaying otherwise normal patterns of inheritance (Parkin et al. 1995; Sharpe et al. 1995; Udall et al. 2005) provided the first evidence for exchange of genetic material between homoeologous chromosomes. Detailed elucidation revealed that in the most extreme cases such chromosomal rearrangements can range up to ~ 40 Mb in length, effectively involving whole chromosomes (Higgins et al. 2018). Interestingly, all evidence thus far shows subgenomic bias in direction of exchanges, with loss of the C genome and concomitant gain of the A genome being far more prevalent (Samans et al. 2017; Higgins et al. 2018). Early studies already suggested an important adaptive role, with a well-documented exchange between *B. napus* chromosomes A07 and C06 being associated with higher seed yields (Osborn et al. 2003). The ubiquity of such events, which have shaped the modern *B. napus* genome, was confirmed with the sequencing of the genome reference. Fixed homoeologous exchange events were found to be shared among cultivars due to intentional or inadvertent selection during allopolyploidization and/or breeding, and they thus underlie loci for a number of important traits (Chalhoub et al. 2014). One event led to loss of the C genome copy of a MYB28 transcription factor on chromosome C02 that was replaced with a non-functional A-genome copy of the same gene; this SV defined one of the strongest loci controlling the low glucosinolate phenotype that has underpinned the global success of canola as a major crop. A similar event involving the same *B. napus* chromosomes (A02/C02) created CNV for the flowering time gene *FLC* which leads to accelerated flowering in annual types. Methods to catalogue homoeologous exchanges in *B. napus* from read depth data (Samans et al. 2017) also led to the discovery of PAV underlying a QTL for seed fibre in *B. napus* (Stein et al. 2017).

A comprehensive study of natural genetic variation in homologues of 35 flowering time regulation genes in diverse *B. napus* morphotypes identified an extensive range of structural variation and potential associations to phenotypes related to flowering and secondary processes (Schiessl et al. 2017b). Different homoeologues of the vernalization response gene *Flowering Locus C* (*FLC*), the photosynthetic regulator *Phytochrome A* (*PHYA*) and the hormone *Gibberellic Acid 3-oxidase 1* (*GA3ox1*) all showed CNV and PAV associated with the derivation of *B. napus* morphotypes, again demonstrating the importance of SV on genes involved in human agricultural selection. The use of genome-wide SNP arrays to catalogue SV in multiparental segregating *B. napus* populations enables the inclusion of SV polymorphisms in genome-wide association studies. Gabur et al. (2018) utilized segregating PAV identified by

"single nucleotide absence polymorphism" (SNAP) markers (or "missing" SNP data) to discover a strong involvement of SV in the quantitative control of disease resistance in *B. napus*. Further, in combination with genome sequencing data from mapping parents the QTL could be delineated to small PAV spanning just one or a few potential candidate genes. The success of this study and the continued discovery of SVs as determinants in the control of key agronomic traits suggests that the discovery of SV should become a standard tool in future genetic analyses of crop traits.

In soybean, the world's primary pulse crop, self-pollination together with genetic bottlenecks during domestication have eroded the genetic diversity within the species (Hyten et al. 2006), with sequence polymorphism among soybean accessions typically as low as one SNP per 1000 bases. Therefore, it might also be reasonable to assume a low level of genomic SV. However, this assumption was shattered by Anderson et al. (2014) who found that a panel of 41 soybean accessions contained almost 1528 genes affected by SV. Interestingly, genes exhibiting CNVs were enriched for resistance genes with nucleotide-binding site (NBS) or NBS-leucine-rich repeat (LRR) domains, suggesting involvement of CNVs in interactions with plant pathogens. A well-known example was described by Cook et al. (2012), who found that a 31 kb sequence fragment containing an amino acid transporter, an α-SNAP protein and a WI12 (wound-inducible domain) protein that each contribute to soybean cyst nematode (SCN), one of the most devastating pathogens of soybean, was present in ten tandem copies in resistant cultivars but only a single copy in susceptible cultivars. A similar gene CNV was also reported by Liu et al. (2017) who also showed that CNV of multiple genes present in a single DNA fragment contribute towards SCN resistance.

There have been numerous examples of genome structural variations underlying commercially important traits in many vegetable species. Hardigan et al. (2016) studied genome-wide SV in homozygous clones of diploid potato (*S. tuberosum*), finding that almost 30% of the genes were tolerant to deletion or duplication, with an impact of SV on performance. As in other crop species, there was evidence that PAV and CNV impacted gene clusters in potato involved in environmental stress responses. The authors concluded that CNV may drive adaptation of potato through evolution of important pathways involved in stress responses. SVs have also been reported to play an important role in controlling several traits in tomato. Tranchida-Lombardo et al. (2018) reported over 200 deletions by resequencing and assembling of two tomato landraces. Many of these deletions were found to be localized in the genes annotated for ripening, shelf life and quality of the fruit. In cucumber (*Cucumis sativus*), a model system for sex determination studies in plants, Zhang et al. (2015) constructed a nucleotide-resolution SV map which revealed SVs in their coding regions of over 1600 genes. Using this SV map, they were able to prove that the sex determination in cucumber is controlled by CNV in four genes at the *Female* (*F*) locus.

Approximately 85% of the maize genome is composed of transposable elements (Schnable et al. 2009), which contribute significantly to genomic re-arrangements and gene PAV. In a recent effort to create a newer reference genome assembly for maize, more than 3000 SVs were detected by comparing optical maps for two inbred lines Ki11 and W22 to the B73 reference. The individual SV events ranged from 1 kb to over 1 Mb in length, with an average length of about 20 kb (Jiao et al. 2017). Because this phenomenon has been studied extensively in maize, it is not surprising that a plethora of agronomically interesting traits have been found be to be controlled by PAV in maize, ranging from abiotic and biotic stress responses to plant architecture and heterosis. For example, Wang et al. (2016) reported that an insertion in the promoter region of the *ZmVPP1* gene induces drought-dependent expression of *ZmVPP1* in drought-tolerant genotypes. The PAV in *ZmVPP1* was also associated with enhancement of photosynthetic efficiency and root development under both stress and non-stress conditions. CNV events also play a role in stress resistance responses in maize. For example, the *rp1* locus, responsible for race-specific resistance to the common rust fungus, is a hotspot for unequal crossovers leading to gain, loss or duplications in this NBS-LRR gene family. This creates a diverse haplotype makeup at the *rp1* locus, translating to variable resistance responses to various rust races (Richter et al. 1995). The same authors found a similar example for a CNV in a wall-associated kinase (*ZmWAK*) gene, which was found to confer resistance to head smut in maize. Interestingly, the responsible WAK gene was absent in many modern maize lines but present in wild relatives.

Although there is considerable indirect evidence for a role of PAV in stress responses, there is still only a handful of examples for their importance in adaptive traits in maize. Maron et al. (2013) reported association of a rare CNV in the multidrug and toxin exudation gene *MATE1* in maize to aluminium tolerance. Strikingly, the geographical origin of maize lines carrying three copies of *MATE1* coincided with highly acidic soil, implicating this CNV as an important SV conferring maize adaptation to a specific environment.

A DNA segment present or absent exclusively in germplasm adapted to a particular type of environmental cue might be indicative of the fact that genic PAV in this segment play a crucial role in adaptation. An array of INDELs in the 5′ regulatory region of the *FLOWERING LOCUS T* homologue, *LanFTc1* was found to be associated with differential vernalization response, flowering time, in narrow-leafed lupin (Taylor et al. 2018). Genotypes with no deletion exhibited an early flowering behaviour and a reduced or no response to vernalization. Such a catalogue

of structural variations could serve as the basis for the necessary environmental plasticity needed for designing the future crops adapted to wide range of environments. Darracq et al. (2018) identified in total 88 Mb of DNA in a French maize inbred line that was absent in an American inbred line, and contained 395 putative coding genes. Evidence was also observed for de novo SV in European maize alongside ancient SV, demonstrating ongoing adaptive genome evolutionary dynamics. Annotation of the novel genes revealed putative roles in biotic and biotic stress responses, in biosynthetic processes, in development, in protein synthesis and in chromatin remodelling. Intriguingly, expression of most of the novel genes was restricted to particular conditions or tissues, leading to a conclusion that at least some of the genes from the dispensable part of the genome might be involved in environmental adaptation.

The realization that heterotic pools in maize breeding programs can have vastly different gene content gave new insight into the impact of SV on heterosis. Springer et al. (2009) compared the genome structures of two maize inbred lines by comparative genomic hybridization and found that a copious amount of genomic sequences exhibited copy number differences between the two genomes. Sun et al. (2018) underlined the extent of genome-wide SV in maize by assembling the genome of Mo17 and comparing it to the B73 reference assembly. This confirmed that almost 10% of the annotated genes were exclusive to one or the other accession, while more than 20% were found to show substantial structural variation. It was also hypothesized that these SVs might be involved in heterosis and genome evolution. Furthermore, many sequences annotated as single-copy genes were found to be present in one genome but completely absent from the other genome. Although the contributing mechanisms for heterosis are still not completely elucidated and may differ from crop to crop, there is good reason to believe that fixation of complementary PAV in different heterotic pools can play an important role in exploitation of additive heterosis in hybrid breeding.

The huge diversity in rice, combined with well-defined phylogeny available for the genus *Oryza*, the small size of the genome and vast genome sequence resources make it an ideal candidate for studying effects of genomic SV on traits. Bai et al. (2016) generated a CNV map, at single nucleotide resolution for 50 rice accessions, comprising 9196 deletions compared to the Nipponbare reference genome. More than 2000 annotated genes were reported to be affected by CNV. Uga et al. (2013) identified the gene *Deep Rooting 1* (*DRO1*) as a key regulator of root system architecture with a profound effect on yield under different water regimes in rice. A japonica upland rice (Kinandang Patong) containing a full-length *DRO1* copy was found to have a deeper root system architecture, whereas the indica lowland rice cultivar IR64 carried a truncated copy due to a 1 bp InDel. This results in shallower roots due to the introduction of a premature stop codon in *DRO1* (Uga et al. 2013). Yu et al. (2011) and Yao et al. (2015) both found enrichment for disease resistance or defence response genes among dispensable genes in large rice resequencing studies. Wang et al. (2015) reported CNV at the *Grain Length on Chromosome 7* (GL7) locus associated with regulation of grain dimensions in rice. A tandem duplication of a DNA fragment within the *GL7* locus lead was found to cause upregulation of *GL7* expression and suppression of its negative regulator, thereby resulting in an increased grain length and improved grain appearance.

Outlook

As more and more genome data become available for major crops, our insight into the profound importance of SV on trait diversity continues to grow. Understanding the contribution of gene copy number and presence–absence variation to important traits will be an important factor in improving the accuracy and efficacy of many new genetic technologies in plant breeding, from genomic selection to genome editing strategies. Just a decade ago, the notion that full, high-quality reference genome assemblies for any major crop could be generated reasonably simply and quickly, at low cost, was barely conceivable. Today this is (almost) a reality, and a new era of high-throughput pangenomic analyses is set to dominate crop genetic studies in the immediate future. Although current costs of third-generation DNA sequencing technologies and chromatin conformation technologies for scaffold improvement are still high, and computational bottlenecks associated with creation of reference assemblies need to be overcome, the plummeting cost of long-read sequencing and improvement in computational algorithms and hardware could make de novo genome assembly more routine in the foreseeable future, even for complex polyploid crop genomes. One aspect of interest for breeders in a changing world is associations of SV with ecogeographical adaptations, abiotic stress adaptation and biotic stress responses. This knowledge opens fascinating new opportunities to learn from adaptive evolution of polyploid crop species in order to improve crop resilience against biotic and abiotic stress constraints in the face of climate change. From a broader perspective, studies of SV in model and crop plants derived from recent polyploidy have revealed an involvement of gene dosage and/or PAV in a wide number of different traits under natural and human (breeding) selection and showed how genome rearrangements resulting from de novo polyploidization might even be used to generate new variation for breeding. Such examples underline the role of SV as a key driver of genetic diversity for future breeding of sustainable, resilient and healthy crops. Novel methods to detect,

assay, harness and select for useful SV events will therefore be a valuable future resource for crop breeding.

Author contribution statement IG and HSC drafted the manuscript. IAP and RJS contributed text sections and edited the content.

Funding Deutscher Akademischer Austauschdienst and Deutsche Forschungsgemeinschaft (Grant No. SN14/17-1).

Compliance with ethical standards

Conflict of interest All authors jointly state that there is no conflict of interest. IAPP and RJS serve on the editorial board for this journal, but this is not considered to constitute a conflict of interest.

References

Alix K, Gérard PR, Schwarzacher T, Heslop-Harrison JSP (2017) Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. Ann Bot 120(2):183–194. https://doi.org/10.1093/aob/mcx079

Ali N, Heslop-Harrison JP, Ahmad H, Graybosch RA, Hein GL, Schwarzacher T (2016) Introgression of chromosome segments from multiple alien species in wheat breeding lines with wheat streak mosaic virus resistance. Heredity 117:114–123. https://doi.org/10.1038/hdy.2016.36

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41(10):1061–1067. https://doi.org/10.1038/ng.437

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12(5):363–376. https://doi.org/10.1038/nrg2958

Anderson JE, Kantar MB, Kono TY, Fu F, Stec AO, Song Q, Cregan PB, Specht JE, Diers BW, Cannon SB, McHale LK, Stupar RM (2014) A roadmap for functional structural variants in the soybean genome. G3 (Bethesda) 4(7):1307–1318. https://doi.org/10.1534/g3.114.011551

Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, Pozniak CJ, Choulet F, Distelfeld A, Poland J, Ronen G, Sharpe AG, Pozniak C, Barad O, Baruch K, Keeble-Gagnère G, Mascher M, Ben-Zvi G, Josselin A-A, Himmelbach A, Balfourier F, Gutierrez-Gonzalez J, Hayden M, Koh C, Muehlbauer G, Pasam RK, Paux E, Rigault P, Tibbits J, Tiwari V, Spannagl M, Lang D, Gundlach H, Haberer G, Mayer KFX, Ormanbekova D, Prade V, Šimková H, Wicker T, Swarbreck D, Rimbert H, Felder M, Guilhot N, Kaithakottil G, Keilwagen J, Leroy P, Lux T, Twardziok S, Venturini L, Juhász A, Abrouk M, Fischer I, Uauy C, Borrill P, Ramirez-Gonzalez RH, Arnaud D, Chalabi S, Chalhoub B, Cory A, Datla R, Davey MW, Jacobs J, Robinson SJ, Steuernagel B, van Ex F, Wulff BBH, Benhamed M, Bendahmane A, Concia L, Latrasse D, Alaux M, Bartoš J, Bellec A, Berges H, Doležel J, Frenkel Z, Gill B, Korol A, Letellier T, Olsen O-A, Singh K, Valárik M, van der Vossen E, Vautrin S, Weining S, Fahima T, Glikson V, Raats D, Číhalíková J, Toegelová H, Vrána J, Sourdille P, Darrier B, Barabaschi D, Cattivelli L, Hernandez P, Galvez S, Budak H, Jones JDG, Witek K, Yu G, Small I, Melonek J, Zhou R, Belova T, Kanyuka K, King R, Nilsen K, Walkowiak S, Cuthbert R, Knox R, Wiebe K, Xiang D, Rohde A, Golds T, Čížková J, Akpinar BA, Biyiklioglu S, Gao L, N'Daiye A, Kubaláková M, Šafář J, Alfama F, Adam-Blondon A-F, Flores R, Guerche C, Loaec M, Quesneville H, Condie J, Ens J, Maclachlan R, Tan Y, Alberti A, Aury J-M, Barbe V, Couloux A, Cruaud C, Labadie K, Mangenot S, Wincker P, Kaur G, Luo M, Sehgal S, Chhuneja P, Gupta OP, Jindal S, Kaur P, Malik P, Sharma P, Yadav B, Singh NK, Khurana J, Chaudhary C, Khurana P, Kumar V, Mahato A, Mathur S, Sevanthi A, Sharma N, Tomar RS, Holušová K, Plíhal O, Clark MD, Heavens D, Kettleborough G, Wright J, Balcárková B, Hu Y, Salina E, Ravin N, Skryabin K, Beletsky A, Kadnikov V, Mardanov A, Nesterov M, Rakitin A, Sergeeva E, Handa H, Kanamori H, Katagiri S, Kobayashi F, Nasuda S, Tanaka T, Wu J, Cattonaro F, Jiumeng M, Kugler K, Pfeifer M, Sandve S, Xun X, Zhan B, Batley J, Bayer PE, Edwards D, Hayashi S, Tulpová Z, Visendi P, Cui L, Du X, Feng K, Nie X, Tong W, Wang L (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. https://doi.org/10.1126/science.aar7191

Bai Z, Chen J, Liao Y, Wang M, Liu R, Ge S, Wing RA, Chen M (2016) The impact and origin of copy number variations in the *Oryza* species. BMC Genomics 17:261. https://doi.org/10.1186/s12864-016-2589-2

Belay G, Merker A (2004) Cytogenetic studies in Ethiopian landraces of tetraploid wheat (*Tviticum Turgidum* L.). II. Spontaneous chromosome translocations and fertility. Hereditas 126(1):35–43. https://doi.org/10.1111/j.1601-5223.1997.00035.x

Belay G, Merker A (2006) Cytogenetic analysis of a spontaneous 5B/6B translocation in tetraploid wheat landraces from Ethiopia, and implications for breeding. Plant Breed 117(6):537–542. https://doi.org/10.1111/j.1439-0523.1998.tb02203.x

Beló A, Beatty MK, Hondred D, Fengler KA, Li B, Rafalski A (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor Appl Genet 120(2):355–367. https://doi.org/10.1007/s00122-009-1128-9

Bridges CB (1936) The BAR "gene" a duplication. Science 83(2148):210–211. https://doi.org/10.1126/science.83.2148.210

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE (2014) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517(7536):608–611. https://doi.org/10.1038/nature13907

Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, Correa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, Arnaud D, Guo H, Daviaud C, Alamery S, Jabbari K, Zhao M, Edger PP, Chelaifa H, Tack D, Lassalle G, Mestiri I, Schnel N, Le Paslier M-C, Fan G, Renault V, Bayer PE, Golicz AA, Manoli S, Lee T-H, Thi VHD, Chalabi S, Hu Q, Fan C, Tollenaere R, Lu Y, Battail C, Shen J, Sidebottom CHD, Canaguier A, Chauveau A, Berard A, Deniot G, Guan M, Liu Z, Sun F, Lim YP, Lyons E, Town CD, Bancroft I, Meng J, Ma J, Pires JC, King GJ, Brunel D, Delourme R, Renard M, Aury J-M, Adams KL, Batley J, Snowdon RJ, Tost J, Edwards D, Zhou Y, Hua W, Sharpe AG, Paterson AH, Guan C, Wincker P (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345(6199):950–953. https://doi.org/10.1126/science.1253435

Chang C, Lu J, Zhang H-P, Ma C-X, Sun G (2016) Copy number variation of cytokinin oxidase gene *Tackx4* associated with grain weight and chlorophyll content of flag leaf in common wheat. PLoS ONE 10(12):e0145970. https://doi.org/10.1371/journal.pone.0145970

Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, Schulz MH, Erdogan F, Li N, Kijas Z, Arkesteijn G, Pajares IL, Goetz-Sothmann M, Heinrich U, Rost I, Dufke A, Grasshoff U, Glaeser B, Vingron M, Ropers HH (2008) Mapping translocation breakpoints by next-generation sequencing. Genome Res 18(7):1143–1149. https://doi.org/10.1101/gr.076166.108

Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE (2012) Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). Proc Natl Acad Sci USA 109(4):1176–1181. https://doi.org/10.1073/pnas.1112041109

Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC, Gore M, Guill KE, Holland J, Hufford MB, Lai J, Li M, Liu X, Lu Y, McCombie R, Nelson R, Poland J, Prasanna BM, Pyhäjärvi T, Rong T, Sekhon RS, Sun Q, Tenaillon MI, Tian F, Wang J, Xu X, Zhang Z, Kaeppler SM, Ross-Ibarra J, McMullen MD, Buckler ES, Zhang G, Xu Y, Ware D (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 44:803. https://doi.org/10.1038/ng.2313

Clarke WE, Parkin IA, Gajardo HA, Gerhardt DJ, Higgins E, Sidebottom C, Sharpe AG, Snowdon RJ, Federico ML, Iniguez-Luy FL (2013) Genomic DNA enrichment using sequence capture microarrays: a novel approach to discover sequence nucleotide polymorphisms (SNP) in *Brassica napus* L. PLoS ONE 8(12):e81992. https://doi.org/10.1371/journal.pone.0081992

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35(6):2013–2025. https://doi.org/10.1093/nar/gkm076

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, Diers BW, Jiang J, Hudson ME, Bent AF (2012) Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. Science 338(6111):1206–1209. https://doi.org/10.1126/science.1228746

Darracq A, Vitte C, Nicolas S, Duarte J, Pichon J-P, Mary-Huard T, Chevalier C, Bérard A, Le Paslier M-C, Rogowsky P, Charcosset A, Joets J (2018) Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. BMC Genomics 19(1):119. https://doi.org/10.1186/s12864-018-4490-7

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the *Photoperiod-B1* and *Vernalization-A1* genes Is associated with altered flowering time in wheat (*Triticum aestivum*). PLoS ONE 7(3):e33234. https://doi.org/10.1371/journal.pone.0033234

Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet 126(1):1–11. https://doi.org/10.1007/s00122-012-1964-x

English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, Beck CR, Davis CF, Dahdouli M, Ma S, Carroll A, Veeraraghavan N, Bruestle J, Drees B, Hastie A, Lam ET, White S, Mishra P, Wang M, Han Y, Zhang F, Stankiewicz P, Wheeler DA, Reid JG, Muzny DM, Rogers J, Sabo A, Worley KC, Lupski JR, Boerwinkle E, Gibbs RA (2015) Assessing structural variation in a personal genome—towards a human reference diploid genome. BMC Genomics 16(1):332. https://doi.org/10.1186/s12864-015-1479-3

Escaramís G, Docampo E, Rabionet R (2015) A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics 14(5):305–314. https://doi.org/10.1093/bfgp/elv014

Esteve-Codina A, Paudel Y, Ferretti L, Raineri E, Megens H-J, Silió L, Rodríguez MC, Am Groenen M, Ramos-Onsins SE, Pérez-Enciso M (2013) Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. BMC Genomics 14(1):148. https://doi.org/10.1186/1471-2164-14-148

Fadista J, Thomsen B, Holm L-E, Bendixen C (2010) Copy number variation in the bovine genome. BMC Genomics 11(1):284. https://doi.org/10.1186/1471-2164-11-284

Francia E, Morcia C, Pasquariello M, Mazzamurro V, Milc JA, Rizza F, Terzi V, Pecchioni N (2016) Copy number variation at the *HvCBF4–HvCBF2* genomic segment is a major component of frost resistance in barley. Plant Mol Biol 92(1):161–175. https://doi.org/10.1007/s11103-016-0505-4

Friebe B, Jiang J, Raupp WJ, McIntosh RA, Gill BS (1996) Characterization of wheat-alien translocations conferring resistance to diseases and pests: current status. Euphytica 91(1):59–87. https://doi.org/10.1007/BF00035277

Gabur I, Chawla HS, Liu X, Kumar V, Faure S, von Tiedemann A, Jestin C, Dryzska E, Volkmann S, Breuer F, Delourme R, Snowdon R, Obermeier C (2018) Finding invisible quantitative trait loci with missing data. Plant Biotechnol J. https://doi.org/10.1111/pbi.12942

Gaeta RT, Chris Pires J (2010) Homoeologous recombination in allopolyploids: the polyploid ratchet. New Phytol 186(1):18–28. https://doi.org/10.1111/j.1469-8137.2009.03089.x

Gale MD, Devos KM (1998) Comparative genetics in the grasses. Proc Natl Acad Sci USA 95(5):1971. https://doi.org/10.1073/pnas.95.5.1971

Gamuyao R, Chin JH, Pariasca-Tanaka J, Pesaresi P, Catausan S, Dalid C, Slamet-Loedin I, Tecson-Mendoza EM, Wissuwa M, Heuer S (2012) The protein kinase *Pstol1* from traditional rice confers tolerance of phosphorus deficiency. Nature 488(7412):535–539. https://doi.org/10.1038/nature11346

Ghosh S, Qu Z, Das PJ, Fang E, Juras R, Cothran EG, McDonell S, Kenney DG, Lear TL, Adelson DL, Chowdhary BP, Raudsepp T (2014) Copy number variation in the horse genome. PLoS Genet 10(10):e1004712. https://doi.org/10.1371/journal.pgen.1004712

Golicz AA, Batley J, Edwards D (2016a) Towards plant pangenomics. Plant Biotechnol J 14(4):1099–1105. https://doi.org/10.1111/pbi.12499

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 (2016b) The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun 7:13390. https://doi.org/10.1038/ncomms13390

Gonzalez E (2005) The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307(5714):1434–1440. https://doi.org/10.1126/science.1101160

Grandke F, Snowdon R, Samans B (2016) gsrc: an R package for genome structure rearrangement calling. Bioinformatics 3:545–546. https://doi.org/10.1093/bioinformatics/btw648

Gu YQ, Coleman-Derr D, Kong X, Anderson OD (2004) Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four *Triticeae* Genomes. Plant Physiol 135(1):459. https://doi.org/10.1104/pp.103.038083

Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC (2010) Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 26(10):1277–1283. https://doi.org/10.1093/bioinformatics/btq152

Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, Leisner CP, Manrique-Carpintero NC, Newton L, Pham GM, Vaillancourt B, Yang X, Zeng Z, Douches DS, Jiang J, Veilleux RE, Buell CR (2016) Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. Plant Cell 28(2):388–405. https://doi.org/10.1105/tpc.15.00538

Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. Nat Rev Genet 10(8):551–564. https://doi.org/10.1038/nrg2593

Hattori Y, Nagai K, Furukawa S, Song X-J, Kawano R, Sakakibara H, Wu J, Matsumoto T, Yoshimura A, Kitano H, Matsuoka M, Mori H, Ashikari M (2009) The ethylene response factors *SNORKEL1* and *SNORKEL2* allow rice to adapt to deep water. Nature 460(7258):1026–1030. https://doi.org/10.1038/nature08258

Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ, Richmond T, Jeddeloh JA, Jia G, Springer NM, Vance CP, Stupar RM (2010) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol 155(2):645–655. https://doi.org/10.1104/pp.110.166736

He Z, Cheng F, Li Y, Wang X, Parkin IAP, Chalhoub B, Liu S, Bancroft I (2015) Construction of *Brassica* A and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research. Data Brief 4:357–362. https://doi.org/10.1016/j.dib.2015.06.016

He Z, Wang L, Harper AL, Havlickova L, Pradhan AK, Parkin IAP, Bancroft I (2016) Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. Plant Biotech J 15:594–604. https://doi.org/10.1111/pbi.12657

Higgins EE, Clarke WE, Howell EC, Armstrong SJ, Parkin IAP (2018) Detecting *de Novo* homoeologous recombination events in cultivated *Brassica napus* using a genome-wide SNP array. G3 (Bethesda) 8(8):2673–2683. https://doi.org/10.1534/g3.118.200118

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Penagaricano F, Lindquist E, Pedraza MA, Barry K, de Leon N, Kaeppler SM, Buell CR (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26(1):121–135. https://doi.org/10.1105/tpc.113.119982

Horn R, Snowdon R, Kusterer B (2002) Structural genome analysis using molecular cytogenetic techniques. In: Esser K, Lüttge U, Beyschlag W, Hellwig F (eds) Progress in botany: genetics, physiology, ecology. Springer, Berlin, pp 55–79

Hurgobin B, Golicz AA, Bayer PE, Chan C-KK, Tirnaz S, Dolatabadian A, Schiessl SV, Samans B, Montenegro JD, Parkin IAP, Pires JC, Chalhoub B, King GJ, Snowdon R, Batley J, Edwards D (2017) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. Plant Biotechnol J 16(7):1265–1274. https://doi.org/10.1111/pbi.12867

Hyten DL, Song Q, Zhu Y, Choi I-Y, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci USA 103(45):16666. https://doi.org/10.1073/pnas.0604379103

Iovene M, Zhang T, Lou Q, Buell CR, Jiang J (2013) Copy number variation in potato—an asexually propagated autotetraploid species. Plant J 75(1):80–89. https://doi.org/10.1111/tpj.12200

Jiao W-B, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 36:64–70. https://doi.org/10.1016/j.pbi.2017.02.002

Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, Wang B, Liu Z, Chen J, Li W, Zhang M, Xie S, Lai J (2012) Genome-wide genetic changes during modern breeding of maize. Nat Genet 44(7):812–815. https://doi.org/10.1038/ng.2312

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S, Guill K, Regulski M, Kumari S, Olson A, Gent J, Schneider KL, Wolfgruber TK, May MR, Springer NM, Antoniou E, McCombie WR, Presting GG, McMullen M, Ross-Ibarra J, Dawe RK, Hastie A, Rank DR, Ware D (2017) Improved maize reference genome with single-molecule technologies. Nature 546:524. https://doi.org/10.1038/nature22971

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258(5083):818. https://doi.org/10.1126/science.1359641

Keane TM, Wong K, Adams DJ, Flint J, Reymond A, Yalcin B (2014) Identification of structural variation in mouse genomes. Front Genet 5:1061. https://doi.org/10.3389/fgene.2014.00192

Knox AK, Dhillon T, Cheng H, Tondelli A, Pecchioni N, Stockinger EJ (2010) CBF gene copy number variation at *Frost Resistance-2* is associated with levels of freezing tolerance in temperate-climate cereals. Theor Appl Genet 121(1):21–35. https://doi.org/10.1007/s00122-010-1288-7

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318(5849):420–426. https://doi.org/10.1126/science.1149504

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40(10):1253–1260. https://doi.org/10.1038/ng.237

Lagercrantz U, Putterill J, Coupland G, Lydiate D (1996) Comparative mapping in *Arabidopsis* and *Brassica*, fine scale genome collinearity and congruence of genes controlling flowering time. Plant J 9(1):13–20

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, Jiao Y, Ni P, Zhang J, Li D, Guo X, Ye K, Jian M, Wang B, Zheng H, Liang H, Zhang X, Wang S, Chen S, Li J, Fu Y, Springer NM, Yang H, Wang J, Dai J, Schnable PS, Wang J (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42(11):1027–1030. https://doi.org/10.1038/ng.684

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, Kwok P-Y (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol 30(8):771–776. https://doi.org/10.1038/nbt.2303

Law CN, Worland AJ (2006) The control of adult-plant resistance to yellow rust by the translocated chromosome 5BS–7BS of bread wheat. Plant Breed 116(1):59–63. https://doi.org/10.1111/j.1439-0523.1997.tb00975.x

Lee JA, Carvalho CMB, Lupski JR (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131(7):1235–1247. https://doi.org/10.1016/j.cell.2007.11.037

Lee TG, Kumar I, Diers BW, Hudson ME (2015) Evolution and selection of *Rhg1*, a copy-number variant nematode-resistance locus. Mol Ecol 24(8):1774–1791. https://doi.org/10.1111/mec.13138

Li W, Olivier M (2013) Current analysis platforms and methods for detecting copy number variation. Physiol Genom 45(1):1–16. https://doi.org/10.1152/physiolgenomics.00082.2012

Li Y, Xiao J, Wu J, Duan J, Liu Y, Ye X, Zhang X, Guo X, Gu Y, Zhang L, Jia J, Kong X (2012) A tandem segmental duplication (TSD) in green revolution gene *Rht-D1b* region underlies plant height variation. New Phytol 196(1):282–291. https://doi.org/10.1111/j.1469-8137.2012.04243.x

Li Y, Zhou G, Ma J, Jiang W, Jin L, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang S, Zuo Q, Shi X, Li Y, Zhang W, Hu Y, Kong G, H-l H, Tan B, Song J, Liu Z, Wang Y, Ruan H, Yeung CKL, Liu J, Wang H, Zhang L, Guan R, Wang K, Li W, Chen S, Chang R, Jiang Z, Jackson SA, Li R, Qiu L (2014) *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and

agronomic traits. Nat Biotechnol 32(10):1045–1052. https://doi.org/10.1038/nbt.2979

Lin K, Zhang N, Severing EI, Nijveen H, Cheng F, Visser RGF, Wang X, de Ridder D, Bonnema G (2014) Beyond genomic variation—comparison and functional annotation of three *Brassica* rapagenomes: a turnip, a rapid cycling and a Chinese cabbage. BMC Genomics 15(1):250. https://doi.org/10.1186/1471-2164-15-250

Liu J, Zhang L, Xu L, Ren H, Lu J, Zhang X, Zhang S, Zhou X, Wei C, Zhao F, Du L (2013) Analysis of copy number variations in the sheep genome using 50 K SNP BeadChip array. BMC Genomics 14(1):229. https://doi.org/10.1186/1471-2164-14-229

Liu M, Stiller J, Holušová K, Vrána J, Liu D, Doležel J, Liu C (2016) Chromosome-specific sequencing reveals an extensive dispensable genome component in wheat. Sci Rep. https://doi.org/10.1038/srep36398

Liu S, Kandoth PK, Lakhssassi N, Kang J, Colantonio V, Heinz R, Yeckel G, Zhou Z, Bekal S, Dapprich J, Rotter B, Cianzio S, Mitchum MG, Meksem K (2017) The soybean *GmSNAP18* gene underlies two types of resistance to soybean cyst nematode. Nat Commun 8:14822. https://doi.org/10.1038/ncomms14822

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X, Hernandez AG, Mikel MA, Soifer I, Barad O, Buckler ES (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun. https://doi.org/10.1038/ncomms7914

Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14(10):417–422. https://doi.org/10.1016/S0168-9525(98)01555-8

Ma L, Chung WK (2001) Quantitative analysis of copy number variants based on real-time lightcycler PCR. In: Haines JL, Korf BR, Morton CC, Seidman CE, Seidman JG, Smith DR (eds) Current protocols in human genetics, vol 107. Wiley, Hoboken, pp 7.21.1–7.21.8

Ma J, Stiller J, Zheng Z, Wei Y, Zheng Y-L, Yan G, Doležel J, Liu C (2015) Putative interchromosomal rearrangements in the hexaploid wheat (*Triticum aestivum* L.) genotype 'Chinese Spring' revealed by gene locations on homoeologous chromosomes. BMC Evol Biol 15:37. https://doi.org/10.1186/s12862-015-0313-5

Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, Campbell BC, Hu W, Innes DJ, Han X, Cruickshank A, Dai C, Frère C, Zhang H, Hunt CH, Wang X, Shatte T, Wang M, Su Z, Li J, Lin X, Godwin ID, Jordan DR, Wang J (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun 4:2320. https://doi.org/10.1038/ncomms3320

Mace E, Tai S, Innes D, Godwin I, Hu W, Campbell B, Gilding E, Cruickshank A, Prentis P, Wang J, Jordan D (2014) The plasticity of NBS resistance genes in sorghum is driven by multiple evolutionary processes. BMC Plant Biol. https://doi.org/10.1186/s12870-014-0253-z

Mamtani M, Anaya J-M, He W, Ahuja SK (2010) Association of copy number variation in the *FCGR3B* gene with risk of autoimmune diseases. Genes Immun 11(2):155–160. https://doi.org/10.1038/gene.2009.71

Maron LG, Guimaraes CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D, Magalhaes JV, Pineros MA, Schatz MC, Wing RA, Kochian LV (2013) Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. Proc Natl Acad Sci USA 110(13):5241–5246. https://doi.org/10.1073/pnas.1220766110

Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, Sampath D, Ayling S, Steuernagel B, Pfeifer M, D'Ascenzo M, Akhunov ED, Hedley PE, Gonzales AM, Morrell PL, Kilian B, Blattner FR, Scholz U, Mayer KFX, Flavell AJ, Muehlbauer GJ, Waugh R, Jeddeloh JA, Stein N (2013) Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. Plant J 76(3):494–505. https://doi.org/10.1111/tpj.12294

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang X-Q, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544(7651):427–433. https://doi.org/10.1038/nature22043

Mason AS, Snowdon RJ (2016) Oilseed rape: learning about ancient and recent polyploid evolution from a recent crop species. Plant Biol (Stuttg) 18(6):883–892. https://doi.org/10.1111/plb.12462

Mason AS, Higgins EE, Snowdon RJ, Batley J, Stein A, Werner C, Parkin IAP (2017) A user guide to the Brassica 60 K Illumina Infinium™ SNP genotyping array. Theor Appl Genet 130(4):621–633. https://doi.org/10.1007/s00122-016-2849-1

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol 159(4):1295–1308. https://doi.org/10.1104/pp.112.194605

McVey M, Lee SE (2008) MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet 24(11):529–538. https://doi.org/10.1016/j.tig.2008.08.007

Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan C-KK, Visendi P, Lai K, Doležel J, Batley J, Edwards D (2017) The pangenome of hexaploid bread wheat. Plant J 90(5):1007–1013. https://doi.org/10.1111/tpj.13515

Moore JK, Haber JE (1996) Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in *Saccharomyces cerevisiae*. Mol Cell Biol 16(5):2164–2173. https://doi.org/10.1128/MCB.16.5.2164

Morgante M, de Paoli E, Radovic S (2007) Transposable elements and the plant pan-genomes. Curr Opin Plant Biol 10(2):149–155. https://doi.org/10.1016/j.pbi.2007.02.001

Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, Cao H, Schlebusch SA, Giorda K, Schnall-Levin M, Wall JD, Kwok P-Y (2016) A hybrid approach for *de novo* human genome sequence assembly and phasing. Nat Methods 13(7):587–590. https://doi.org/10.1038/nmeth.3865

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, Mayer KFX, Taudien S, Platzer M, Jeddeloh JA, Springer NM, Muehlbauer GJ, Stein N (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol 14(6):203. https://doi.org/10.1186/gb-2013-14-6-r58

Neik TX, Barbetti MJ, Batley J (2017) Current status and challenges in identifying disease resistance genes in *Brassica napus*. Front Plant Sci 8:1788. https://doi.org/10.3389/fpls.2017.01788

Nelson JC, Sorrells ME, Van-Deynze AE, Lu YH, Atkinson M, Bernard M, Leroy P, Faris JD, Anderson JA (1995) Molecular mapping of wheat: major genes and rearrangements in homoeologous groups 4, 5, and 7. Genetics 141(2):721–731

Nishida H, Yoshida T, Kawakami K, Fujita M, Long B, Akashi Y, Laurie DA, Kato K (2013) Structural variation in the 5′ upstream region of photoperiod-insensitive alleles *Ppd-A1a* and *Ppd-B1a* identified in hexaploid wheat (*Triticum aestivum* L.), and their effect on heading time. Mol Breeding 31(1):27–37. https://doi.org/10.1007/s11032-012-9765-0

Osborn TC, Butrulle DV, Sharpe AG, Pickering KJ, Parkin IAP, Parker JS, Lydiate DJ (2003) Detection and effects of a homeologous reciprocal transposition in *Brassica napus*. Genetics 165(3):1569–1577

Parkin IAP, Sharpe AG, Keith DJ, Lydiate DJ (1995) Identification of the A and C genomes of amphidiploid *Brassica napus* (oilseed rape). Genome 38(6):1122–1131. https://doi.org/10.1139/g95-149

Parkin IAP, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL, Denoeud F, Belcram H, Links MG, Just J, Clarke C, Bender T, Huebert T, Mason AS, Pires JC, Barker G, Moore J, Walley PG, Manoli S, Batley J, Edwards D, Nelson MN, Wang X, Paterson AH, King G, Bancroft I, Chalhoub B, Sharpe AG (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biol 15(6):R77. https://doi.org/10.1186/gb-2014-15-6-r77

Pearce S, Saville R, Vaughan SP, Chandler PM, Wilhelm EP, Sparks CA, Al-Kaff N, Korolev A, Boulton MI, Phillips AL, Hedden P, Nicholson P, Thomas SG (2011) Molecular characterization of *Rht-1 d*warfing genes in hexaploid wheat. Plant Physiol 157(4):1820. https://doi.org/10.1104/pp.111.183657

Pires JC, Zhao JW, Schranz ME, Leon EJ, Quijada PA, Lukens LN, Osborn TC (2004) Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (*Brassicaceae*). Biol J Linn Soc Lond 82(4):675–688. https://doi.org/10.1111/j.1095-8312.2004.00350.x

Qian L, Voss-Fels K, Cui Y, Jan HU, Samans B, Obermeier C, Qian W, Snowdon RJ (2016) Deletion of a stay-green gene associates with adaptive selection in *Brassica napus*. Mol Plant 9(12):1559–1569. https://doi.org/10.1016/j.molp.2016.10.017

Reeves TG, Rajaram S, van Ginkel M, Trethowan R, Braun H, Cassaday K (1999) New wheats for a secure, sustainable future. CIMMYT, Mexico

Richter TE, Pryor TJ, Bennetzen JL, Hulbert SH (1995) New rust resistance specificities associated with recombination in the *Rp1* complex in maize. Genetics 141(1):373–381

Riley R, Coucoli H, Chapman V (1967) Chromosomal interchanges and the phylogeny of wheat. Heredity 22:233. https://doi.org/10.1038/hdy.1967.29

Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerrière A, Vital A, Dumanchin C, Feuillette S, Brice A, Vercelletto M, Dubas F, Frebourg T, Campion D (2006) *APP* locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat Genet 38(1):24–26. https://doi.org/10.1038/ng1718

Samans B, Chalhoub B, Snowdon RJ (2017) Surviving a genome collision: genomic signatures of allopolyploidization in the recent crop species. Plant Genome. https://doi.org/10.3835/plantgenome2017.02.0013

Saxena RK, Edwards D, Varshney RK (2014) Structural variations in plant genomes. Brief Funct Genomics 13(4):296–307. https://doi.org/10.1093/bfgp/elu016

Schiessl S, Huettel B, Kuehn D, Reinhardt R, Snowdon R (2017a) Post-polyploidisation morphotype diversification associates with gene copy number variation. Sci Rep 7:41845. https://doi.org/10.1038/srep41845

Schiessl S, Huettel B, Kuehn D, Reinhardt R, Snowdon RJ (2017b) Targeted deep sequencing of flowering regulators in *Brassica napus* reveals extensive copy number variation. Sci Data. https://doi.org/10.1038/sdata.2017.13

Schiessl S-V, Katche E, Ihien E, Chawla HS, Mason AS (2018) The role of genomic structural variation in the genetic improvement of polyploid crops. Crop J. https://doi.org/10.1016/j.cj.2018.07.006

Schmidt MH-W, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Maß J, Pfaff C, Schurr U, Chetelat R, Maumus F, Aury J-M, Koren S, Fernie AR, Zamir D, Bolger AM, Usadel B (2017) De novo Assembly of a new *Solanum pennellii* accession using nanopore sequencing. Plant Cell 29(10):2336–2348. https://doi.org/10.1105/tpc.17.00521

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, van Buren P, Vaughn MW, Ying K, Yeh C-T, Emrich SJ, Jia Y, Kalyanaraman A, Hsia A-P, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia J-M, Deragon J-M, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956):1112–1115. https://doi.org/10.1126/science.1178534

Schröder J, Hsu A, Boyle SE, Macintyre G, Cmero M, Tothill RW, Johnstone RW, Shackleton M, Papenfuss AT (2014) Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. Bioinformatics 30(8):1064–1072. https://doi.org/10.1093/bioinformatics/btt767

Sears ER (1939) Cytogenetic studies with polyploid species of wheat. I. chromosomal aberrations in the progeny of a haploid of *Triticum Vulgare*. Genetics 24(4):509–523

Sharp AJ, Cheng Z, Eichler EE (2006) Structural variation of the human genome. Annu Rev Genomics Hum Genet 7:407–442. https://doi.org/10.1146/annurev.genom.7.080505.115618

Sharpe AG, Parkin IA, Keith DJ, Lydiate DJ (1995) Frequent nonreciprocal translocations in the amphidiploid genome of oilseed rape (*Brassica napus*). Genome 38(6):1112–1121

Sieber A-N, Longin CFH, Leiser WL, Würschum T (2016) Copy number variation of *CBF-A14* at the *Fr-A2* locus determines frost tolerance in winter durum wheat. Theor Appl Genet 129(6):1087–1097. https://doi.org/10.1007/s00122-016-2685-3

Singleton AB (2003) Synuclein locus triplication causes Parkinson's disease. Science 302(5646):841. https://doi.org/10.1126/science.1090278

Snowdon RJ (2007) Cytogenetics and genome analysis in *Brassica* crops. Chromosome Res 15(1):85–95. https://doi.org/10.1007/s10577-006-1105-y

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS, Ecker JR (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5(11):e1000734. https://doi.org/10.1371/journal.pgen.1000734

Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61(1):437–455. https://doi.org/10.1146/annurev-med-100708-204735

Stein N, Mascher M (2019) Capturing pangenome diversity for breeding: a case study in barley. Theor Appl Genet **(to be published in same special issue)**

Stein A, Coriton O, Rousseau-Gueutin M, Samans B, Schiessl SV, Obermeier C, Parkin IAP, Chèvre A-M, Snowdon RJ (2017) Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in *Brassica napus*. Plant Biotechnol J 15(11):1478–1489. https://doi.org/10.1111/pbi.12732

Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X, Liu H, Ma X, Jiao Y, Wang B, Wei X, Stein JC, Glaubitz JC, Lu F, Yu G, Liang C, Fengler K, Li B, Rafalski A, Schnable PS, Ware DH, Buckler ES, Lai J (2018) Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet 50(9):1289–1295. https://doi.org/10.1038/s41588-018-0182-0

Sutton T, Baumann U, Hayes J, Collins NC, Shi B-J, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, Langridge P (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318(5855):1446–1449. https://doi.org/10.1126/science.1146853

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20(12):1689–1699. https://doi.org/10.1101/gr.109165.110

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC (2012) Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics 190(4):1563–1574. https://doi.org/10.1534/genetics.111.137349

Taylor CM, Kamphuis LG, Zhang W, Garg G, Berger JD, Mousavi-Derazmahalleh M, Bayer PE, Edwards D, Singh KB, Cowling WA, Nelson MN (2018) INDEL variation in the regulatory region of the major flowering time gene *LanFTc1* is associated with vernalization response and flowering time in narrow-leafed lupin (*Lupinus angustifolius* L.). Plant Cell Environ. https://doi.org/10.1111/pce.13320

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102(39):13950. https://doi.org/10.1073/pnas.0506758102

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796. https://doi.org/10.1038/35048692

The International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345(6194):1251788. https://doi.org/10.1126/science.1251788

Tranchida-Lombardo V, Aiese Cigliano R, Anzar I, Landi S, Palombieri S, Colantuono C, Bostan H, Termolino P, Aversano R, Batelli G, Cammareri M, Carputo D, Chiusano ML, Conicella C, Consiglio F, D'Agostino N, de Palma M, Di Matteo A, Grandillo S, Sanseverino W, Tucci M, Grillo S (2018) Whole-genome re-sequencing of two Italian tomato landraces reveals sequence variations in genes associated with stress tolerance, fruit quality and long shelf-life traits. DNA Res 25(2):149–160. https://doi.org/10.1093/dnares/dsx045

Tseng S-H, Peng S-F, Cheng Y-M (2017) Analysis of B chromosome nondisjunction induced by the *r-X1* deficiency in maize. Chromosome Res 42(2):223. https://doi.org/10.1007/s10577-017-9567-7

Udall JA, Quijada PA, Osborn TC (2005) Detection of chromosomal rearrangements derived from homologous recombination in four mapping populations of *Brassica napus* L. Genetics 169(2):967–979. https://doi.org/10.1534/genetics.104.033209

Uga Y, Sugimoto K, Ogawa S, Rane J, Ishitani M, Hara N, Kitomi Y, Inukai Y, Ono K, Kanno N, Inoue H, Takehisa H, Motoyama R, Nagamura Y, Wu J, Matsumoto T, Takai T, Okuno K, Yano M (2013) Control of root system architecture by *DEEPER ROOTING 1* increases rice yield under drought conditions. Nat Genet 45(9):1097–1102. https://doi.org/10.1038/ng.2725

Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K (2009) The flowering world: a tale of duplications. Trends Plant Sci 14(12):680–688. https://doi.org/10.1016/j.tplants.2009.09.001

Vialette-Guiraud ACM, Adam H, Finet C, Jasinski S, Jouannic S, Scutt CP (2011) Insights from ANA-grade angiosperms into the early evolution of *CUP-SHAPED COTYLEDON* genes. Ann Bot 107(9):1511–1519. https://doi.org/10.1093/aob/mcr024

Villareal RL, Toro E, Mujeeb-Kazi A, Rajaram S (1995) The 1BL/1RS chromosome translocation effect on yield characteristics in a *Triticum aestivum* L. cross. Plant Breed 114(6):497–500. https://doi.org/10.1111/j.1439-0523.1995.tb00843.x

Voss-Fels K, Snowdon RJ (2016) Understanding and utilizing crop genome diversity via high-resolution genotyping. Plant Biotechnol J 14(4):1086–1094. https://doi.org/10.1111/pbi.12456

Wang W, Wang S, Hou C, Xing Y, Cao J, Wu K, Liu C, Zhang D, Zhang L, Zhang Y, Zhou H (2014) Genome-wide detection of copy number variations among diverse horse breeds by array CGH. PLoS ONE 9(1):e86860. https://doi.org/10.1371/journal.pone.0086860

Wang Y, Xiong G, Hu J, Jiang L, Yu H, Xu J, Fang Y, Zeng L, Xu E, Xu J, Ye W, Meng X, Liu R, Chen H, Jing Y, Wang Y, Zhu X, Li J, Qian Q (2015) Copy number variation at the *GL7* locus contributes to grain size diversity in rice. Nat Genet 47:944. https://doi.org/10.1038/ng.3346

Wang X, Wang H, Liu S, Ferjani A, Li J, Yan J, Yang X, Qin F (2016) Genetic variation in *ZmVPP1* contributes to drought tolerance in maize seedlings. Nat Genet 48(10):1233–1241. https://doi.org/10.1038/ng.3636

Wulff BBH, Moscou MJ (2014) Strategies for transferring resistance into wheat: from wide crosses to GM cassettes. Front Plant Sci 5:692. https://doi.org/10.3389/fpls.2014.00692

Würschum T, Boeven PHG, Langer SM, Longin CFH, Leiser WL (2015) Multiply to conquer: copy number variations at *Ppd-B1* and *Vrn-A1* facilitate global adaptation in wheat. BMC Genet 16(1):949. https://doi.org/10.1186/s12863-015-0258-0

Würschum T, Longin CFH, Hahn V, Tucker MR, Leiser WL (2016) Copy number variations of CBF genes at the *Fr-A2* locus are essential components of winter hardiness in wheat. Plant J 89(4):764–773. https://doi.org/10.1111/tpj.13424

Xiong Z, Pires JC (2011) Karyotype and identification of all homoeologous chromosomes of allopolyploid *Brassica napus* and its diploid progenitors. Genetics 187(1):37–49. https://doi.org/10.1534/genetics.110.122473

Xiong Z, Gaeta RT, Pires JC (2011) Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. Proc Natl Acad Sci USA 108(19):7908–7913. https://doi.org/10.1073/pnas.1014138108

Xu K, Xu X, Fukao T, Canlas P, Maghirang-Rodriguez R, Heuer S, Ismail AM, Bailey-Serres J, Ronald PC, Mackill DJ (2006) *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442(7103):705–708. https://doi.org/10.1038/nature04920

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30(1):105–111. https://doi.org/10.1038/nbt.2050

Yao W, Li G, Zhao H, Wang G, Lian X, Xie W (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol 16:187. https://doi.org/10.1186/s13059-015-0757-3

Yang S, Li J, Zhang X, Zhang Q, Huang J, Chen J-Q, Hartl DL, Tian D (2013) Rapidly evolving R genes in diverse grass species confer resistance to rice blast disease. Proc Natl Acad Sci USA 110(46):18572–18577. https://doi.org/10.1073/pnas.1318211110

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25(21):2865–2871. https://doi.org/10.1093/bioinformatics/btp394

Yu P, Wang C, Xu Q, Feng Y, Yuan X, Yu H, Wang Y, Tang S, Wei X (2011) Detection of copy number variations in rice using array-based comparative genomic hybridization. BMC Genomics 12(1):372. https://doi.org/10.1186/1471-2164-12-372

Yu P, Wang C-H, Xu Q, Feng Y, Yuan X-P, Yu H-Y, Wang Y-P, Tang S-X, Wei X-H (2013) Genome-wide copy number variations in *Oryza sativa* L. BMC Genomics 14(1):649. https://doi.org/10.1186/1471-2164-14-649

Zhang Z, Mao L, Chen H, Bu F, Li G, Sun J, Li S, Sun H, Jiao C, Blakely R, Pan J, Cai R, Luo R, Van de Peer Y, Jacobsen E, Fei Z, Huang S (2015) Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. Plant Cell 27(6):1595–1604. https://doi.org/10.1105/tpc.114.135848

Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T, Wang Y, Fan D, Zhao Y, Wang Z, Zhou C, Chen J, Zhu C, Li W, Weng Q, Xu Q, Wang Z-X, Wei X, Han B, Huang X (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50(2):278–284. https://doi.org/10.1038/s41588-018-0041-z

Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, Liu T-F, Jiang S, Ramachandran S, Liu C-M, Jing H-C (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). Genome Biol 12(11):R114. https://doi.org/10.1186/gb-2011-12-11-r114

Zhou P, Silverstein KAT, Ramaraj T, Guhlin J, Denny R, Liu J, Farmer AD, Steele KP, Stupar RM, Miller JR, Tiffin P, Mudge J, Young ND (2017) Exploring structural variation and gene family architecture with *de novo* assemblies of 15 *Medicago* genomes. BMC Genomics 18(1):261. https://doi.org/10.1186/s12864-017-3654-1

Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M (2014) Copy number polymorphism in plant genomes. Theor Appl Genet 127(1):1–18. https://doi.org/10.1007/s00122-013-2177-7

Zou J, Hu D, Mason AS, Shen X, Wang X, Wang N, Grandke F, Wang M, Chang S, Snowdon RJ, Meng J (2018) Genetic changes in a novel breeding population of *Brassica napus* synthesized from hundreds of crosses between *B. rapa* and *B. carinata*. Plant Biotechnol J 16(2):507–519. https://doi.org/10.1111/pbi.12791

3 Gene presence-absence variation associates with quantitative *Verticillium longisporum* disease resistance in *Brassica napus*

Gabur I[#]; **Chawla H.S**[#]; Lopisso, D.T; Tiedemann A.v; Snowdon R.J; Obermeier C
Scientific Reports (2020), vol: 10:4131
doi:.1038/s41598-020-61228-3
[#] These authors contributed equally to this work.

SCIENTIFIC REPORTS
nature research

OPEN

Gene presence-absence variation associates with quantitative *Verticillium longisporum* disease resistance in *Brassica napus*

Iulian Gabur[1,4], Harmeet Singh Chawla[1,4], Daniel Teshome Lopisso[2,3], Andreas von Tiedemann[2], Rod J. Snowdon[1] & Christian Obermeier[1*]

Although copy number variation (CNV) and presence-absence variation (PAV) have been discovered in selected gene families in most crop species, the global prevalence of these polymorphisms in most complex genomes is still unclear and their influence on quantitatively inherited agronomic traits is still largely unknown. Here we analyze the association of gene PAV with resistance of oilseed rape (*Brassica napus*) against the important fungal pathogen *Verticillium longisporum*, as an example for a complex, quantitative disease resistance in the strongly rearranged genome of a recent allopolyploid crop species. Using Single Nucleotide absence Polymorphism (SNaP) markers to efficiently trace PAV in breeding populations, we significantly increased the resolution of loci influencing *V. longisporum* resistance in biparental and multi-parental mapping populations. Gene PAV, assayed by resequencing mapping parents, was observed in 23–51% of the genes within confidence intervals of quantitative trait loci (QTL) for *V. longisporum* resistance, and high-priority candidate genes identified within QTL were all affected by PAV. The results demonstrate the prominent role of gene PAV in determining agronomic traits, suggesting that this important class of polymorphism should be exploited more systematically in future plant breeding.

Duplication of genes followed by diversification is a common process shaping the evolution of plant species by natural and artificial (breeding) selection[1]. Genes can be duplicated by different mechanisms, including tandem duplication, transposon-mediated duplication, segmental duplication, or in the most extreme form by whole-genome duplication (WGD) or polyploidization. WGD is common in the evolutionary history of many wild and cultivated plant species. Different terms have been used frequently to describe short- and long-range genomic duplication and genome structural variation (SV), a term originally defined in reference to insertions, deletions and inversions greater than 1 kb in size[2–4]. In contrast to small-scale insertion-deletion (InDel) polymorphisms, which are generally defined as small insertions or deletions of a few nucleotides (up to 50 bp), SV in the size range of genes (up to a few kb) can give rise to copy number variation (CNV) or presence/absence variation (PAV). The latter is an extreme form of CNV where fragments in the size range of genes are missing from the genomes of some investigated genotypes.

Genes affected by duplications, InDels, CNVs and PAVs in diploid and polyploid plant species have been linked to local adaptation of wild populations[5] and to important agronomical traits in crops[1,3,6–8] for example flowering time and vernalization requirement in oilseed rape[9,10] and wheat[11], abiotic stress tolerance in wheat[12,13] and biotic stress tolerance in tobacco[14]. However, the strong impact of CNV and other forms of SV in polyploid crop genomes on evolution and trait selection was not recognized until the last few years, when recognition of their relevance was facilitated by large-scale genotyping including long-range sequencing technologies in large breeding populations of numerous crops[15]. One recent genotyping technology which is particularly suitable for detection of long-range SV is Bionano optical genome mapping using nano-channel arrays[16]. This method

[1]Department of Plant Breeding, IFZ Research Centre for Biosystems, Land Use and Nutrition, Justus Liebig University Giessen, 35392, Giessen, Germany. [2]Section of General Plant Pathology and Crop Protection, Georg August University Göttingen, 37077, Göttingen, Germany. [3]College of Agriculture and Veterinary Medicine, Jimma University, Jimma, Ethiopia. [4]These authors contributed equally: Iulian Gabur and Harmeet Singh Chawla. *email: christian.obermeier@agrar.uni-giessen.de

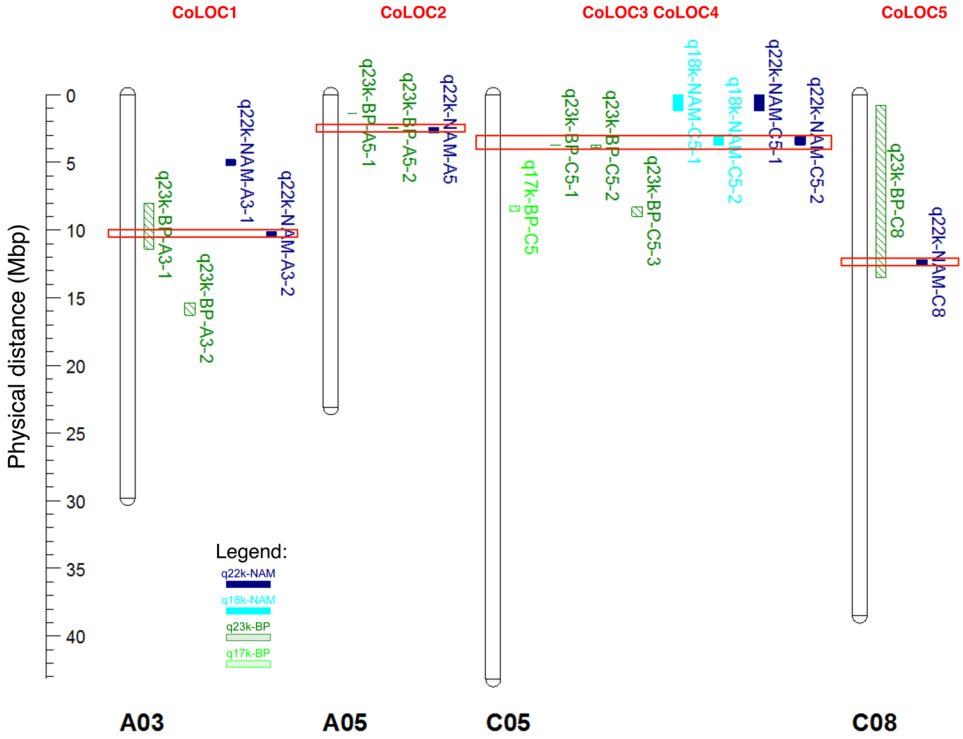


Figure 1. Comparison of co-localizing (CoLOC) QTL positions (in Mbp anchored to Darmor-*bzh*) obtained by QTL mapping in the biparental ExR53-DH population using maps produced with SNP markers only (light green, hatched) or SNP plus SNaP markers (dark green, hatched), and by GWAS in a NAM panel with 5 subpopulations using maps produced with SNP markers only (light blue, solid) or SNP plus SNaP markers (dark blue, solid). Only QTL above a threshold or LOD > 3 and −log (p-value) > 3 were included in the figure. Red boxes indicate QTL regions overlapping in biparental QTL and GWAS.

involves imaging of high-molecular weight, fluorescently-labeled DNA molecules and creation of large restriction maps represented as stretches of light and dark regions (resembling a barcode), which then can be aligned to an *in silico* generated optical map of a reference genome assembly. A key factor distinguishing this approach from previous technologies for SV analysis is that the DNA molecules are not sheared, thus enabling the capture of long-range genomic information stretching up to several hundred kilobases. In combination with accurate genome assemblies even for strongly complex polyploid crop genomes, optical mapping opens new avenues for dissection of genomic rearrangements associated with traits relevant for commercial plant breeding.

Oilseed rape (*Brassica napus*) is a recent polyploid crop species originating from the inter-specific hybridization between the two diploid progenitor species, *B. oleracea* and *B. rapa*. Due to high levels of homoeology between the two progenitor subgenomes, widespread structural rearrangements are a common phenomenon within the rapeseed genome[17–19], while its ancestral hexaploid progenitor genomes already carried intensive structural and functional modifications through long-term genome fractionation and evolution[20–23].

This study focuses exclusively on gene PAV in oilseed rape breeding populations and evaluates the biological relevance and prevalence of gene PAV associated with resistance to a common fungal disease of oilseed rape, Verticillium stem striping caused by *Verticillium longisporum*. By genetic mapping of genome-wide SNP/SNaP markers[24], along with short-read Illumina sequencing in combination with long-range optical mapping, we demonstrate that inclusion of presence-absence polymorphisms in quantitative trait locus (QTL) mapping strategies enables also reliable identification of gene PAV with a putative role in *V. longisporum* resistance. The results provide a valuable example for the importance of pangenomic gene variation for breeding of a key trait in a major polyploid crop.

Results

QTL detection for *V. longisporum* resistance. Analysis of raw genotype data from the Brassica 60k SNP array[25,26] for a doubled haploid population ExR53-DH was performed for 244 genotypes. Two genetic maps were produced, one using only SNP markers (the "SNP map") and one using SNP plus SNaP markers[24] (the "SNaP map"), respectively. Comparison of the SNP and SNaP maps revealed that large chromosomal regions were not covered in the SNP map (e.g. for chromosome A03 compare Fig. 1a,b). Surprisingly, QTL mapping using the SNaP map increased the number of detectable QTL from 5 to 17 (Supplementary Table S1), with substantially increased LOD scores also indicating a dramatic increase in QTL detection power when including SNaP marker data. Furthermore, the map resolution and precision across QTL intervals were considerably increased by inclusion of SNaP markers. Interestingly, some QTL detected only in the SNaP map contained only SNP markers within the QTL confidence interval (e.g. q23k-BP-A1-1, q23k-BP-A3-2, q23k-BP-A3-2, q23k-BP-A5-1), while

chromosome	QTL ID in biparental population	Start position of QTL interval in ExR53-DH	Stop position of QTL in ExR53-DH	Size of QTL interval (bp)	QTL ID in NAM population	Start position of QTL LD block in NAM population	Stop position of QTL LD block in NAM population	Size of LD block (bp)
chrA03	q23k-BP-A3-1	7,963,059	11,419,476	3,456,417	q22k-NAM-A3-2	10,075,388	10,458,202	382,814
chrA05	q23k-BP-A5-2	2,357,535	2,473,365	115,830	q22k-NAM-A5	2,384,153	2,808,636	424,483
chrC05	q23k-BP-C5-1 and q23k-BP-C5-2	3,670,200 and 3,688,115	3,672,842 and 3,949,617	2,642 and 261,502	q22k-NAM-C5-2	3,089,132	3,698,279	609,147
chrC08	q23k-BP-C8	801,925	13,488,675	12,686,750	q22k-NAM-C8	12,201,749	12,596,542	394,793

Table 1. Comparison of QTL locations detected by biparental QTL mapping in ExR53-DH and by GWAS in a NAM panel using SNP and SNaP markers.

other QTL spanned intervals containing only SNaP markers (e.g. q23k-BP-C2-1). Many SNaP markers clustered in groups, spanning large regions up to chromosome scale, while other SNaP markers were located within blocks of SNP markers (Supplementary Table S1).

Additionally, a subset of subpopulations from crosses of a common elite oilseed rape parent with five synthetic *B. napus* parents was selected based on segregation of parental lines for *V. longisporum* resistance from the *B. napus* nested association mapping (NAM) panel described by Snowdon *et al.*[27]. GWAS including both SNP and SNaP markers increased QTL detection power and detected a total of 41 significant marker-trait associations (Supplementary Fig. S2, Supplementary Table S2). Most of the additional detected QTL harboured only SNaP markers.

Co-localizing resistance QTL in diverse genetic backgrounds. Using only SNP markers, comparison of QTL detected by biparental QTL mapping and by GWAS revealed no co-localizing QTL (Fig. 1 in light green and light blue, Supplementary Tables S1 and S2). In contrast, adding SNaP markers to the QTL analyses revealed a strong increase to 5 co-localizing QTL (Fig. 1 in dark green and dark blue, Table 1) harbouring 2 to 90 genes per confidence interval (Supplementary Table S3). Sizes of QTL intervals for *V. longisporum* resistance showed differences between the biparental QTL mapping and the multi-parental GWAS (Supplementary Tables S1 and S2). Generally, smaller QTL intervals are expected in a NAM-GWAS approach, as higher numbers of recombinations are expected from crosses involving multiple non-related parents. On the other hand, QTL intervals in a GWAS mapping approach can only be measured for markers which can be positionally anchored, whereas biparental mapping can also consider marker loci which can be genetically, but not physically anchored. This led to some co-localizing resistance QTL with smaller confidence intervals observed in the biparental mapping compared to the NAM-GWAS (e.g. QTL on chromosomes A05 and C05, Table 1).

Gene ontology and enrichment analysis for genes underlying resistance QTL. Gene ontology enrichment analysis of the biparental population revealed five enriched GO terms in 17 QTL regions, with highest significance (topgoFisher score) attributable to the terms 'chitin catabolic process' and 'response to biotic stimulus' (Supplementary Table S4). In the NAM population eight enriched terms for genes harboured in 28 QTL regions were mainly related to cell growth. In the co-localizing QTL sections six enriched GO terms were identified. A total of 144 genes are harboured within the confidence intervals of the 5 co-localizing resistance QTL (CoLOC1 to CoLOC5, Supplementary Table S5). The GO terms for the 144 genes harboured in the 5 co-localizing QTL intervals reflect annotations from just 8 genes (labelled in green in Supplementary Table S5) related to cell-wall modification (expansins) and pathogen defence (defensins) on chromosome A05 (CoLOC2) and selenium binding on chromosome C08 (CoLOC5). One gene each returned annotation terms containing 'response to stress' and 'systemic acquired resistance', respectively. For four genes an annotation term 'defense response to fungus' or 'response to symbiotic fungus' was assigned, mostly based on plant defensin genes, whereas seven genes returned annotations containing the term 'cell wall', mostly based on expansin genes and pectin esterase-like protein genes.

Both long and short-range PAV associate with *V. longisporum* resistance. To validate if consecutively mapped SNaP markers can help to reliably detect deletions within gene-range size associated with otherwise invisible QTL, resistance-associated QTL detected in the ExR53-DH population were physically located in the *B. napus* Darmor-*bzh* reference genome and the corresponding sequences were compared with optical genome maps. *De novo* assembly for parental lines Express 617 and R53 was performed using 300 Gb (~250x coverage) from Express 617 and 140 Gb (~116x coverage) of Bionano data, respectively. DNA molecules from both genotypes exhibited a very high N50 > 180 kbp, ensuring that long-range genomic information was covered (Supplementary Table S6). One nick label was detected for every 10,000 bp of the molecules, also indicating a uniform coverage and a high SV detection power. The final assembly comprised 1,368 and 1,331 optical maps with N50 values of 234 kb and 235 kb, respectively. The total lengths of the optical mapping assemblies were 978 Mb for Express617 and 874 Mb for R53. The high molecule sizes, along with the total size of the assemblies close to the predicted genome size for *B. napus*, indicate a good assembly quality suitable for reliable detection of long-range SV. Large-scale deletions were consistently detected by consecutively anchored SNaP markers in the ExR53-DH genetic map as well as by optical mapping data (Supplementary Fig. S1). Optical maps enabled accurate detection of small to medium-size deletions and insertions in the size range of genes (from 3 to 5 kb).

The 17 detected QTL regions harboured between 3 and 21 SNP and/or SNaP markers. In total, 122 markers were contained within QTL regions and 72 were anchored to the Darmor-*bzh* reference genome (Supplementary

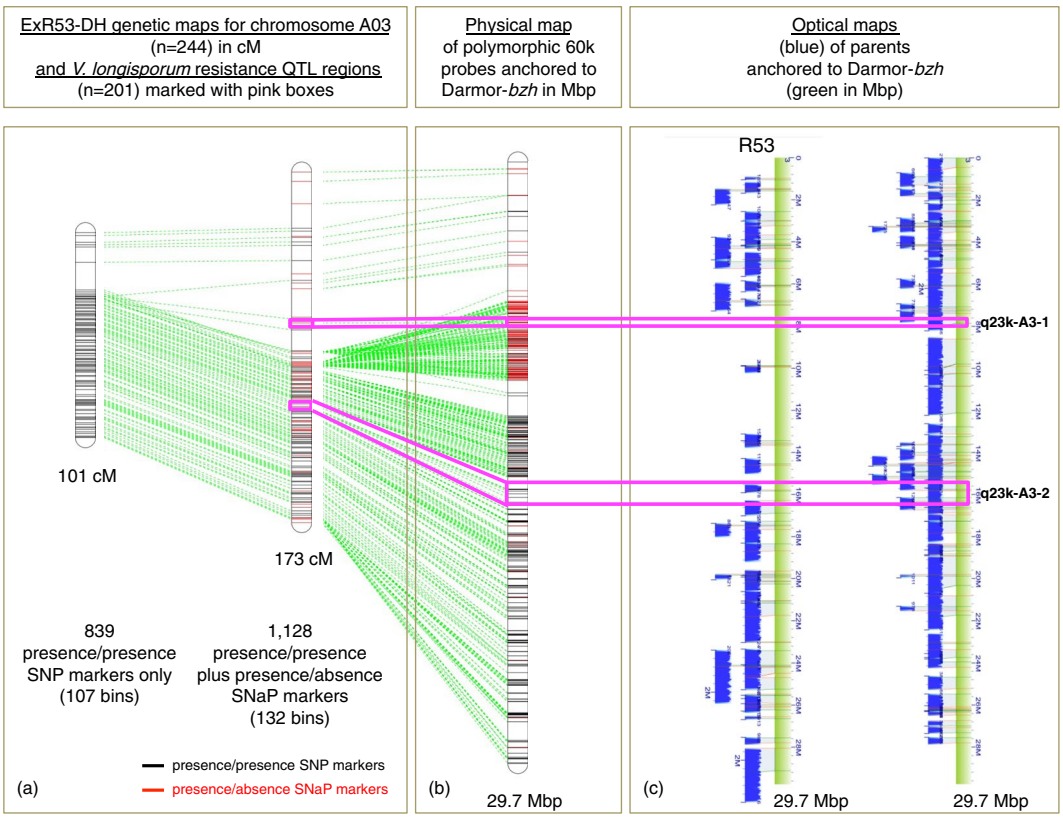


Figure 2. Genetic and physical localisation of biallelic SNP and presence/absence SNaP markers in two *V. longisporum* resistance QTL on chromosome A03 in the ExR53-DH population and its parents. (**a**) Genetic linkage maps showing positions of biallelic SNP (black) and SNaP (red) markers. Green lines connect consensus markers between the different map versions. (**b**) Positions of SNP probe sequences anchored by BLASTn to the Darmor-*bzh* reference sequence. (**c**) Optical Bionano genome maps (blue) of the two parental lines Express617 and R53 aligned to the Darmor-*bzh* reference sequence (green). The pink lines connect marker positions flanking QTL regions in the SNaP map, the physical map and the optical maps (**c**). No resistance QTL were detected using the genetic SNP map.

Table S7). All regions were investigated for long-range structural variants within QTL regions in parental genomes by analyzing the optical map data and comparing the data with SNP/SNaP marker patterns in the segregating population. Short to medium -range structural variation (deletion <5 kb) were confirmed for 16 out of 18 (89%) SNaP marker positions in the two parents (Supplementary Table S7), suggesting that genetically anchored consecutive SNaP markers can reliably detect short to medium-range presence/absence polymorphism associated with *V. longisporum* resistance.

Figure 2 shows an example for comparison of genetic mapping from the DH population, reference anchoring of SNP and SNaP markers and optical mapping data of two parents for chromosome A03, which harbours two additional QTL not detectable using the SNP map. One of the QTL detectable only with the SNaP map, q23k-BP-A3-1, harboured 4 SNaP markers, whereas another, QTL q23k-BP-A3-2, harboured 14 SNP markers within the confidence interval (Supplementary Table S7). The QTL q23-BP-A3-1 is localized in a region with long consecutively ordered stretches of SNaP markers. For the parental line Express 617, both QTL regions were covered in the optical maps (Fig. 2c), whereas for the synthetic *B. napus* parent R53 about 6 Mb overlapping the QTL region q-23k-BP-A3-1 was not covered. The region not covered in R53 corresponds to a 13.8 cM interval on the genetic SNaP map (3.456 Mb), with the resistance allele contributed by Express 617 (Fig. 2a,b). This lack of optical map alignments in QTL q23k-BP-A3-1, combined with segregating, consecutively anchored SNaP markers in the segregating DH population, confirms the deletion of this large chromosomal region in parental line R53 and suggests that this deletion is involved in resistance expression. In contrast, the second QTL, q23k-BP-A3-2, is located in a region on chromosome A03 with long consecutively mapped stretches of SNP markers. Flanking the QTL region, only isolated SNaP markers were mapped. The isolated SNaP markers detected close to the QTL thus probably represent short-range PAV, potentially down to even single-nucleotide level. Nevertheless, saturating the genetic map on chromosome A03 by adding SNaP markers facilitated the detection of QTL q23k-BP-A3-2, whereas no QTL could be mapped in this region using only the SNP map (Fig. 2a, Supplementary Table S1). This example demonstrates that the addition of SNaP markers for genetic mapping was causal for a higher detection power of *V. longisporum* resistance QTL associated with both presence/presence as well as presence/absence polymorphisms.

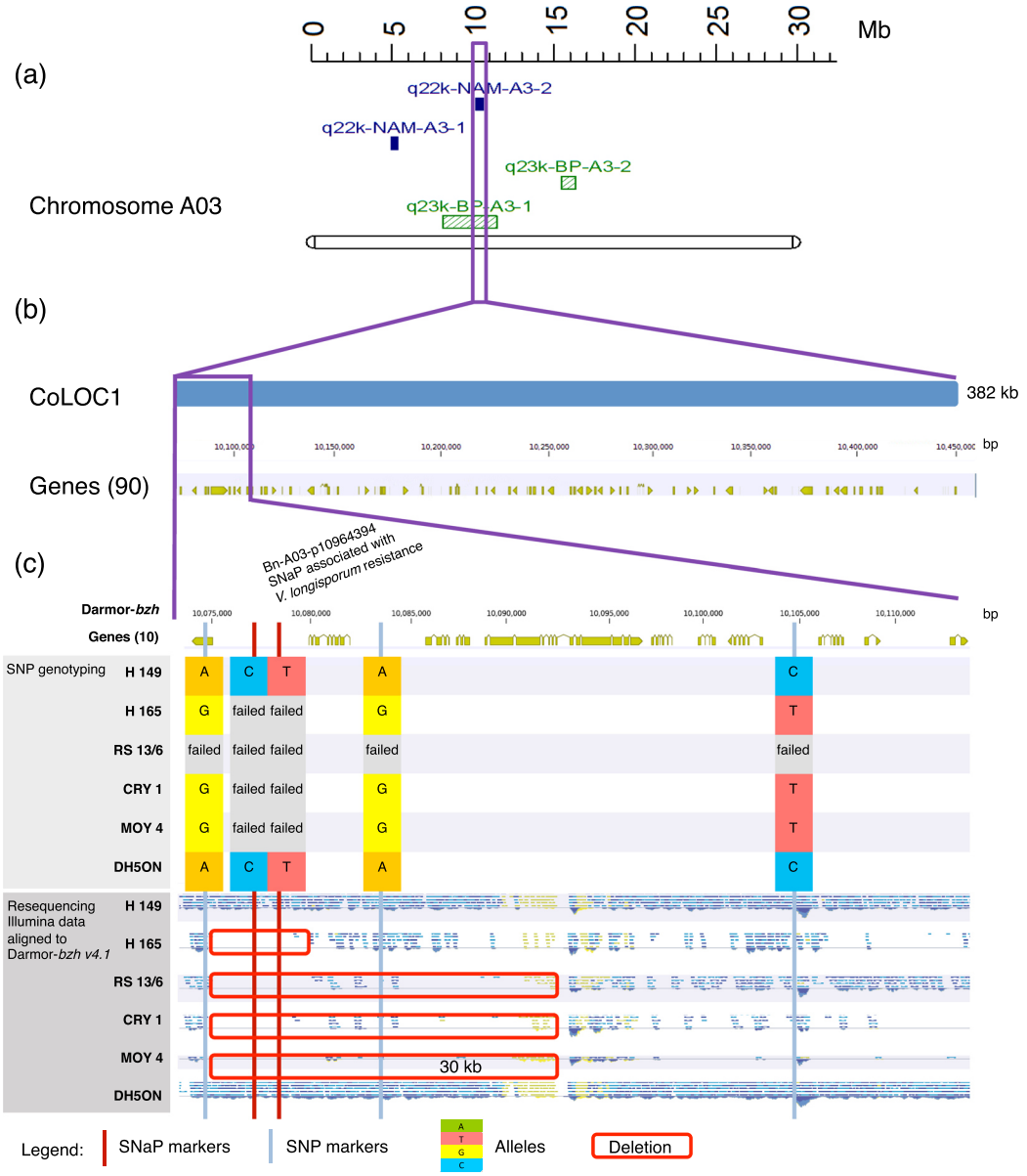


Figure 3. Comparison of SNP/SNaP marker polymorphism, sequence coverage and gene content at the co-localizing QTL CoLOC1 on *B. napus* chromosome A03. (**a**) Positions of QTL for *V. longisporum* resistance in the NAM panel (purple) and in the biparental mapping population ExR53-DH (green). (**b**) Chromosome interval and annotated genes in the QTL region in the *B. napus* Darmor-*bzh* v4.1 reference genome. (**c**) Allele patterns of reference-anchored markers in parents of six NAM subpopulations along with resequencing coverage data in the six parents. Red boxes indicate confirmed segmental deletions involving genes. Failed alleles represent SNaP (absence) alleles in susceptible parents.

Gene PAV is a key determinant of *V. longisporum* resistance. In order to identify putatively absent genes from regions associated with *V. longisporum* disease resistance in the NAM panel, we combined Illumina 60 K SNP chip array genotyping and Illumina resequencing data for the six NAM parental lines with GWAS data from the segregating NAM population. Illumina resequencing confirmed medium to long-range presence/absence variation in the respective parents for 17 QTL regions (Supplementary Table S6). 60% of the 28 detected resistance QTL were affected by medium to long-range PAV in the NAM population.

Figure 3 shows an example of the analyses for the SNaP marker Bn-A03-p10964394, which is associated with *V. longisporum* resistance within the QTL interval q22k-NAM-A3-2 (Supplementary Table S8). This region CoLOC1 (Supplementary Table S3) also overlaps with QTL q23k-BP-A3-1 detected in the biparental population (Fig. 3a). Comparison of the marker-trait segregation in the NAM panel with marker and resequencing data for the six parental genotypes confirmed the expected pattern and revealed a putatively deleted interval within a part (8%) of the QTL region of the susceptible parent. This deletion corresponds to a 30 kb region containing 4 protein-coding genes (Fig. 3c, red box) within the entire QTL interval of 382 kb containing a total of 90 genes

(Fig. 3b). However, from 5 markers (3 SNaPs, 2 SNPs) within the LD block of the co-localizing QTL interval CoLOC1, only the SNaP marker Bn-A03-p10964394 within the deleted region in R53 is significantly associated with *V. longisporum* resistance, suggesting that presence/absence polymorphism is involved in resistance. Based on the resequencing data of the parental lines, two to four genes are affected by deletions within the QTL interval (Fig. 3c), namely BnaA03g21190D (coding for an uncharacterized protein), BnaA03g21200D (coding for a skp1-like protein involved in ubiquitin-dependent protein catabolic processes), BnaA03g21210D (coding for aquaporin pip1-2 protein involved in transporter activities), BnaA03g21220D (coding for an uncharacterized protein) and BnaA03g21230D (coding for an ATP-dependent helicase brm-like protein involved in DNA binding).

SNP/SNaP marker patterns within *V. longisporum* QTL regions were also compared with parental whole-genome sequence coverage in the two parents of the biparental population ExR53-DH. Complete or partial presence/absence events were reconfirmed by resequencing coverage analysis of the two parents for 12 out of 18 reference-anchored SNaP markers (67%) and for 5 out of 6 QTL regions harbouring one or several reference-anchored SNaP markers (83%, Supplementary Table S9). The 2.5 cM QTL interval q23k-BP-A3-1 spans 3,45 Mb on the Darmor-*bzh* reference, containing 683 genes (Supplementary Table S9). Part of this region is shown in detail in Fig. 2 as co-localizing QTL region CoLOC1. Within CoLOC1 a total of 382 kb, containing 90 genes, overlaps in the biparental and multi-parental QTL analyses (see above, Supplementary Table S3 and Fig. 3).

Using genomic resequencing data from 52 accessions[28], including the six NAM parents investigated in our study, we tested the software package SGSGeneLoss[29] for gene loss calling and classification of genes as PAV genes or non-PAV genes. Parameters of 5% minimum gene size coverage (lostCutoff = 0.05) and minimum coverage of two reads for a gene (minCov = 2) was used by Hurgobin *et al.*[19] to call a gene as present in a genotype. However, using these parameters for the 6 NAM parents used in our study, and comparing gene loss calling with PCR amplification data for 11 gene fragments from two genomic regions on chromosome A03 and C04 (Gabur *et al.*[24], Fig. 3 and Supplementary Fig. S3) revealed an inaccurate gene fragment loss classification in 9 of 66 cases (86% accuracy) and no PAV was called for the chromosome C04 regions. This suggests that the chosen parameters were not stringent enough in the six genotypes to distinguish genuine gene loss from data processing noise due to misalignments. Hence, gene loss was accordingly underestimated using these SGSGeneLoss program parameters (and using Darmor-*bzh* v.4.1 as a reference). We therefore developed a customized pipeline which calibrated gene loss calling parameters based on PCR wet lab data and also included the gene size in the calculations to reduce further alignment biases. This procedure increased accuracy of gene loss calling to 96%. Both methods were then used for gene loss calling in the six NAM parents, Express617 and R53, firstly across the whole genome and then within identified QTL regions. Using our modified approach, 49% of total genes (49,701 of 101,039) were estimated to be affected by gene loss and 51% of genes within the 28 QTL regions (1,334 of 2,601) were estimated to be affected by gene loss (PAV genes) between the 6 NAM parents. Between Express 617 and R53, 23% of total genes (23,772 of 101,039) and 23% of genes within 17 QTL regions (590 of 2,646) were estimated to be affected by gene loss. A total of 144 genes were harboured in the 5 QTL intervals which co-localized between the biparental and multi-parental QTL mapping. Of these, 74 (52%) in the NAM panel and 37 (26%) in the ExR53-DH panel were estimated to be affected by PAVs. The genes estimated to be affected by PAV overlap in 20% of cases within the 5 co-localizing QTL regions between the two populations (Supplementary Table S10).

For all genes including PAV genes (144) within the co-localizing QTL regions, 30% of genes have no annotation (43 out of 144). This could suggest that some of the gene models are bioinformatics artefacts. However, from the 43 not annotated genes, 38 (92%) had expression in pan-transcriptome data from He *et al.*[30] or in RNASeq data from the ExR53-DH population (mock or *V. longisporum*-infected, Supplementary Table S11). This suggests that the lack of gene annotation does not necessarily mean that the gene models are incorrect. 12% of all gene models in Darmor-*bzh* could not be annotated using Blast2GO. This is in contrast to the 30% of genes we found in the co-localizing QTL regions which could not be annotated using Blast2GO. Thus, the prevalence of these transcript-validated genes in QTL regions which cannot be annotated from public databases seems to be not a random phenomenon.

Discussion

The reanalysis and integration of SNP array data, short-range Illumina sequencing data and long-range Bionano optical mapping data with QTL data provided new insights into the importance of gene PAV for disease resistance expression against Verticillium stem striping in oilseed rape. QTL analyses including presence/absence markers increased the power of detection for *V. longisporum* resistance in two *B. napus* breeding population. This type of variation (genic or nongenic) was previously reported to be associated with quantitative resistance against two other major fungal disease of oilseed rape 24 suggesting that it is a common and prevalent phenomenon in *B. napus*.

The recent allopolyploid crop species *B. napus* shows a strong abundance of SV and genomic rearrangements[18,31]. In maize, Beló *et al.*[32] used comparative genomic hybridization arrays for detailed genome-wide analysis of SV, and a more recent study discovered some degree of CNV between 100 analyzed lines across more than 90% of the maize genome[33]. However, to date it is unknown how widespread this phenomenon is in other, older polyploid crop species and it is unclear for *B. napus* and most crops how many genes, what kind of genes and what traits are predominantly affected. Particularly for complex polyploid crops, considerably higher numbers of diverse, high-quality reference assemblies are required before reliable pangenomic analysis of genome-wide gene PAV becomes possible. Hurgobin *et al.*[19] classified 38% of all *B. napus* genes in a diversity set of 53 genotypes as PAV-affected, including genes involved in important agronomical traits. In the NAM parental lines, Express617 an R53 we classified 49% and 23% of all genes as PAV-affected by using a PCR-calibrated gene loss bioinformatics pipeline. As five of the 6 NAM parents are resynthesized from exotic *Brassica* species a high gene loss could be expected which are known to exhibit an elevated ratio of rearrangements and SVs[17,18]. Surprisingly the percentage

of genes affected by PAV was similar within the whole genome, the *V. longisporum* QTL for both populations suggesting that the prevalence of gene PAV is a general phenomenon randomly distributed throughout the genome and not restricted to *V. longisporum* resistance or other traits. For gene cluster, gene families and LRR genes involved in qualitative disease resistance it has documented that gene PAV is multiple times higher than on average within the entire genome[19]. However, this seems not to apply for quantitative resistance where diverse genes involved in complex pathways influence resistance responses.

Different mixtures of *V. longisporum* lineages/pathotypes were used for disease resistance screening in the two different *B. napus* mapping populations. Comparison of QTL regions detected by including SNaP markers between the biparental and the multi-parental populations revealed a considerable number of common QTL. This indicates broad-spectrum, lineage/pathotype-independent resistance reactions in genetically diverse germplasm which are of great interest for commercial resistance breeding. Thus we focused for a more detailed analysis on these 5 co-localizing QTL regions, which harbour 2 to 90 genes each within the co-localizing section of the QTL intervals. Based on the quantitative genetic nature of the disease resistance, we assumed that single specialized genes involved in very different biological functions from each QTL interval will not contribute to *V. longisporum* resistance, but rather a number of genes from common biosynthesis pathways or with common biological functions from one or all QTL intervals. To prioritize candidate PAV genes putatively involved in broad-range resistance expression we performed a gene ontology enrichment analysis for all QTL-associated genes, separately for both populations, and also for the genes from the 5 colocalizing QTL intervals. The enriched GO terms for the genes from the QTL regions were quite different between the two populations, being mainly related to chitin and cell-wall metabolism for the biparental ExR53-DH population and mainly related to cell growth for the multi-parental NAM population. This difference is not unexpected, as the disease resistance screenings are known to be highly susceptible to the environment and slightly different mixes of *V. longisporum* lineages/pathotypes were used for inoculation of the two populations. However, the result might also indicate that common as well as different resistance mechanisms are activated by different pathotypes.

Within the 144 genes from the co-localizing QTL putatively involved in pathotype-independent resistance, we found a number of genes coding for selenium-binding proteins, for plant defensin proteins and for expansin proteins have been found to be affected by PAV. In *A. thaliana*, expression of selenium binding proteins is tightly linked to detoxification processes related to oxidative stress[34]. Plant defensins are major components of the innate immune system of plants, are involved in the cell wall integrity signaling pathway and often show a potent, broad-spectrum antifungal activity[35]. The antifungal protein RsAFP2 from *Raphanus sativus*, a close relative of *B. napus*, has been described to exhibit antifungal activitiy against the fungus *V. dahliae*, which is closely related to *V. longisporum*[36]. Furthermore, a synthetic defensin expressed in *A. thaliana* has also been shown to exhibit antifungal activity against *V. dahliae*[37]. Expansins mediate cell wall-loosening and down-regulation of an expansin-like protein in *A. thaliana* that has been shown to increase resistance against nectotrophic fungi[38]. This suggest that genes affected by PAV and involved in cell wall integrity and signaling at the cell wall surface are key components of *V. longisporum* broad-spectrum resistance in *B. napus*.

We found that in the majority of cases gene presence was associated with resistance against *V. longisporum*. However, interestingly one of the strongest QTL, q23k-BP-C2-3 ($R^2 = 23\%$) was mapped using SNaP markers only. In this case, the absence alleles were inherited by the parent R53 and associated with resistance. Bionano Optical Mapping confirmed the deletion of a region containing 70 genes in the parent R53. This phenomenon has been rarely reported for quantitative disease resistance, but has previously been shown for Sclerotinia stem rot in *B. napus*[24] and for three other fungal pathogens in *Medicago truncatula*[39].

Surprisingly, we found another QTL region containing a total of 7 nucleotide binding site-leucine-rich repeat (NLR) resistance genes (TIR-NBS-LRR) to be affected by PAV and to harbour a QTL on chromosome C09 for *V. longisporum* resistance in the NAM population in this study. This region was also described by Samans *et al.*[18] to be part of two gene clusters of 14 and 8 TIR-NBS-LRR genes on *B. napus* chromosome C09, which are frequently deleted in natural *B. napus* compared to synthetic *B. napus* accessions. NBS-LRR genes are frequently described to be involved in monogenic disease resistance. For NLR resistance genes it is well known that they are often organized in clusters or tandem repeats in a number of plant species and crops and numerous studies have shown that fitness costs can lead to multiplication and deletion of gene family members, such as in *A. thaliana*[40,41] and in *B. napus*[17]. *V. longisporum* resistance in the NAM and ExR53-DH population is quantitatively inherited. However, similar co-segregation of TIR-NBS-LRR genes with QTL have been described for other crops, e.g. for soybean[42], barley[43], potato[44] which might result from evolution and local genome diversification of genes involved in qualitative and quantitative disease resistance mechanisms.

The prioritization of these candidate PAV genes mainly involved in cell wall integrity, growth and modification is consistent with our earlier findings that QTL for the concentration of soluble simple phenylpropanoids, which are putative precursors and degradation products of cell-wall modifications and cross-linking, are co-localizing with major resistance QTL in the biparental ExR53-DH population. In addition, the concentrations of some of these cell wall-associated compounds are significantly correlated with *V. longisporum* resistance[45]. Further functional characterization of PAV genes from QTL regions may help to improve our understanding of disease resistance mechanisms and to improve fungal resistance in *B. napus* by exploitation for in future plant breeding programs.

Materials and Methods

Phenotyping for *Verticillium longisporum* resistance in *B. napus*. Resistance phenotyping was conducted in the greenhouse at Georg August University Göttingen, Germany. In order to represent a broad range of pathogenicity traits occurring in oilseed rape fields, a spore suspension mixture of *V. longisporum* isolates VL43 (lineage A1/D1, North Germany), VLS3 (lineage A1/D1, Sweden) and PD589 (lineage A1/D3, Japan)[46,47] with a density of 1×10^6 spores/ml concentrations was used to inoculate the 200 NAM lines applying the root-dipping

method[48] in four experiments. Each experiment included 20 inoculated and 20 control plants for each tested genotype. Rating of symptoms was done weekly over a 4-week period, using the 1–9 disease scoring scale described by Eynck *et al.*[48]. Resistance screening for the biparental population ExR53-DH was performed similarly using a mixture of isolates VL40 and VL43 (both lineage A1/D1) with 202 DH lines as described in detail by Obermeier *et al.*[45].

High molecular weight DNA isolation for optical mapping. High molecular weight (HMW) DNA isolation was carried out for Express617 and R53 according to the IrysPrep™ Plant Tissue-Nuclei protocol provided by Bionano Genomics. Young leaves (approximately 2 grams) were harvested from dark-treated rapeseed plants. The harvested leaves were immediately fixed with 2% formaldehyde followed by homogenization in isolation buffer containing PVP-10, BME and Triton X-100. The isolated nuclei were then purified on Percoll cushions. Purified nuclei were further embedded in an agarose matrix. Agarose plugs were further subjected to proteinase K treatment followed by rigorous washings steps. Finally, HMW DNA was recovered by melting the plugs using GELase™ (Epicentre) treatment. An additional drop dialysis step was performed to ensure ultra-clean DNA. High molecular weight DNA was further subjected to sequence-specific nick-labeling using the IrysPrep™ Labeling-NLRS protocol provided by Bionano Genomics. HMW DNA was subjected to digestion by the single-stranded nicking endonuclease *Nt.BspQI* (recognition site GCTCTTC). The nicks created by *Nt.BspQI* were then repaired using fluorophore-labeled nucleotides. Nicked and labeled single DNA molecules were subsequently loaded onto an IrysChip for imaging on the Bionano Genomics Irys system.

Bio-informatics analysis for Bionano optical mapping data. DNA molecule images generated from the Irys system were computationally translated into single-molecule optical maps. These single molecules were then assembled into consensus maps using the dedicated IrysSolve pipeline (v5134) provided by Bionano Genomics. An *in silico* optical map was generated for the Darmor-*bzh* v. 4.1 ref. [17] using Knickers v1.5.5 and was used to calculate noise parameters for the final assembly. Optical map assemblies from Express617 and R53 were finally aligned to the Darmor-*bzh* reference using the runCharacterize script provided by Bionano Genomics, with the settings published by Kawakatsu *et al.*[49]. The alignment was imported into Bionano IrysView (v2.5.1.29842) software for visualizing and detecting structural variations between Express617, R53 and the Darmor-*bzh B. napus* reference genome.

Global RNA-Seq analysis for confirmation of gene expression. Two contrasting lines from the mapping population ExR53-DH, DH41 (partially resistant genotype) and DH94 (susceptible genotype), were grown for four weeks after inoculation with *V. longisporum*- (isolate VL43) and mock-inoculation in two independent experiments under different environmental conditions (optimal, drought stress and heat stress). For each mock and *V. longisporum*-inoculated plants, pooled hypocotyl samples from 20 plants/line were harvested 28 days after inoculation. In total, 16 samples were analyzed and total RNA was extracted. Sequencing libraries were produced by service provider LGC Genomics GmbH (Berlin, Germany) and in total 1,022 Million 100 bp pair-end raw reads were obtained by Illumina HiSeq 2000 3′end sequencing. Alignment of reads to the *B. napu*s reference genome Darmor-*bzh* v4.1 and all statistical analysis were performed using CLC Genomics Workbench version 9.0 (QIAGEN Bioinformatics CLC bio, Aarhus, Denmark).

Genetic mapping and QTL analysis. 244 DH lines from the F1 of the cross Express617 × R53 and the two parents were analyzed with the Brassica 60 k Illumina Infinium array[26]. DNA was extracted from leaves using the CTAB method[50] and array genotyping assays were outsourced to TraitGenetics (Seeland, Germany). SNP calls with >85% failed calls across all 244 genotypes were removed from further analyses. Also, SNP probe calls with >90% or <10% of a single allele across the population were removed, leaving a total of 22,064 SNP markers. From these, 4,654 SNP probes (21.1%) showed a segregation pattern with one allele displaying a failed call and 17,410 SNP probes showed a normal biallelic segregation. A "SNaP map" was created from all quality-filtered 22,064 markers (2,714 marker bins), including biallelic and presence/absence polymorphisms, while a "SNP map" was created using only the 17,410 markers showing presence/presence polymorphism (2,176 marker bins). Genetic maps were created using the software MSTMap[51] applying the kosambi distance function and a cut-off *p* value of 10^{-30}. QTL analyses were performed using the software QGene 4.3.9 and 4.4.0[52] applying composite interval mapping with a scan interval of 1 milliMorgan and automatic forward cofactor selection. Mean normalized AUDPC values from the four *V. longisporum* greenhouse resistance screenings were used as trait input data[45].

Genome-wide association studies. The NAM panel was also genotyped using the 60 K Illumina Infinium Brassica SNP array as described above and data was filtered according to Gabur *et al.*[24]. Using the Darmor-*bzh* reference v4.1 we anchored 28,073 SNP markers by BLASTn using CLC Genomics Workbench v. 9.0 (Qiagen Bioinformatics). The SNP map contained 18,068 markers, the SNaP map contained 21,695 markers. Association analyses were conducted using the R package GenABEL[53]. A mixed linear model approach that increases detection power[54] was adjusted for population stratification by including the kinship matrix and the first two principal components as covariates[55]. Stringent significance cutoff values were set at a false discovery rate (FDR) correction of 10%[56]. To reduce the type II error rate, we also captured the SNP-trait associations for disease resistance using an arbitrary threshold of $-\log10(p) \geq 3$.

Linkage disequilibrium (LD) analysis and haplotype construction. Whole genome linkage disequilibrium (LD) was calculated using the squared allele-frequency correlations (r^2) between pairs of SNPs. Only markers with a maximum of 10% missing data and MAF ≥ 0.05 were included in the analysis. Haplotype patterns were assessed for SNP and SNaP markers that showed significant marker trait association at the adjusted

Bonferroni threshold of $-\log10(\text{P-value}) \geq 4.33$. Haplotype blocks were defined with the confidence interval method described by Gabriel *et al.*[57] in Haploview version 4.2[58] and visualized with the R package LDheatmap[59].

Resequencing and coverage analysis. The Illumina 250 bp paired-end raw sequencing data for the two parents of the biparental population, Express617 and R53, was described previously by Stein *et al.*[60], while Illumina 100 bp paired-end raw sequencing data for the 6 NAM parents was described by Schmutzer *et al.*[29]. Sequences were aligned to the Darmor-*bzh* v. 4.1 reference using CLC Genomics Workbench v. 9.0 (Qiagen Bioinformatics). Genes from QTL regions were classified as affected or not affected by presence/absence variation based on coverage analysis of the WGS data from the parental lines. Coverage differences were calculated using the bedtools software package v. 2.27.0 with multiBamCov. A minimum cutout threshold of 1.5 aligned reads was used to differentiate between gene presence and absence. The threshold was selected using PCR data for calibration available for the six NAM parental lines from Gabur *et al.*[24]. Additionally, we used the SGSGeneLoss v0.1 described by Golicz *et al.*[30]. The visualization in Fig. 1 was performed using the software MapChart v.2.3, while Fig. 3 was generated using CLC Genomics Workbench 9 track lists. Identification of homoeologous exchanges (HE) was performed using the method described by Samans *et al.*[18]. Further analysis was partly performed by using the R package 'gsrc'[61].

Gene ontology enrichment analyses. To produce gene ontology information for all *B. napus* genes, 101,039 Darmor-*bzh* peptide sequences (Brassica_napus.annotation_v5.pep.fa.gz) were downloaded from the website: http://www.genoscope.cns.fr/brassicanapus/data/ and were used as input for Blast2Go v. 4.1.9. The R package topGO v1.0[62,63] was used for gene ontology enrichment analysis.

Received: 10 October 2018; Accepted: 7 February 2020;
Published online: 05 March 2020

References

1. Panchy, N., Lehti-Shiu, M. & Shiu, S.-H. Evolution of Gene Duplication in Plants. *Plant. Physiol.* **171**, 2294–2316 (2016).
2. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
3. Saxena, R. K., Edwards, D. & Varshney, R. K. Structural variations in plant genomes. *Brief. Funct. Genomics* **13**, 296–307 (2014).
4. Żmieńko, A., Samelak, A., Kozłowski, P. & Figlerowicz, M. Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* **127**, 1–18 (2014).
5. Xu, X. *et al*. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111 (2012).
6. Yang, S. *et al*. Rapidly evolving R genes in diverse grass species confer resistance to rice blast disease. *Proc. Natl Acad. Sci.* **110**, 18572–18577 (2013).
7. Renny-Byfield, S. & Wendel, J. F. Doubling down on genomes: polyploidy and crop plants. *Am. J. Botany* **101**, 1711–1725 (2014).
8. Lee, T. G., Kumar, I., Diers, B. W. & Hudson, M. E. Evolution and selection of *Rhg1*, a copy-number variant nematode-resistance locus. *Mol. Ecol.* **24**, 1774–1791 (2015).
9. Schiessl, S., Hüttel, B., Kühn, D., Reinhardt, R. & Snowdon, R. J. Post-polyploidisation morphotype diversification associates with gene copy-number variation. *Sci. Rep.* **7**, 41845 (2017a).
10. Schiessl, S., Hüttel, B., Kühn, D., Reinhardt, R. & Snowdon, R. J. Targeted deep sequencing of flowering regulators in *Brassica napus* reveals extensive copy number variation. *Sci. Data* **4**, 170013 (2017b).
11. Würschum, T., Boeven, P. H. G., Langer, S. M., Longin, C. F. H. & Leiser, W. L. Multiply to conquer: copy number variations at *Ppd-B1* and *Vrn-A1* facilitate global adaptation in wheat. *BMC Genet.* **16**, 96 (2015).
12. Sieber, A.-N., Longin, C. F. H., Leiser, W. L. & Würschum, T. Copy number variation of CBF-A14 at the Fr-A2 locus determines frost tolerance in winter durum wheat. *Theor. Appl. Genet.* **129**, 1087–1097 (2016).
13. Würschum, T. *et al*. Copy number variations of *CBF* genes at the *Fr-A2* locus are essential components of winter hardiness in wheat. *Plant. J.* **89**, 764–773 (2017).
14. Chen, S. *et al*. Unstable allotetraploid tobacco genome due to frequent homeologous recombination, segmental deletion, and chromosome loss. *Mol. Plant.* **11**, 914–927 (2018).
15. Bevan, M. W. *et al*. Genomic innovation for crop improvement. *Nat.* **543**, 346–354 (2017).
16. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
17. Chalhoub, B. *et al*. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Sci.* **345**, 950–953 (2014).
18. Samans, B., Chalhoub, B. & Snowdon, R. J. Surviving a genome collision: Genomic signatures of allopolyploidiation in the recent crop species *Brassica napus*. Plant. *Genome* **10**, 1–15 (2017).
19. Hurgobin, B. *et al*. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant. Biotechnol. J.* **16**, 1265–1274 (2018).
20. Wang, X. *et al*. *Brassica rapa* Genome Sequencing Project Consortium. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039 (2011).
21. Cheng., F. *et al*. Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. *Plant. Cell* **25**, 1541–1554 (2013).
22. Liu, S. *et al*. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.* **5**, 3930 (2014).
23. Parkin, I. A. *et al*. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biol.* **15**, R77 (2014).
24. Gabur, I. *et al*. Finding invisible quantitative trait loci with missing data. *Plant. Biotech. J.* **16**, 2102–2112 (2018).
25. Clarke, W. E. *et al*. A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theor. Appl. Genet.* **129**, 1887–1899 (2016).
26. Mason, A. S. *et al*. A user guide to the Brassica 60K Illumina Infinium™ SNP genotyping array. *Theor. Appl. Genet.* **130**, 621–633 (2017).
27. Snowdon, R. J., Abbadi, A., Kox, T., Schmutzer, T. & Leckband, G. Heterotic Haplotype Capture: precision breeding for hybrid performance. *Trends Plant. Sci.* **20**, 410–413 (2015).
28. Schmutzer, T. *et al*. Species-wide genome sequence and nucleotide polymorphisms from the model allopolyploid plant *Brassica napus*. *Sci. Data* **2**, 150072 (2015).
29. Golicz, A. A. *et al*. Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Funct. Integr. Genomics* **15**, 189–196 (2015).
30. He, Z. *et al*. Construction of Brassica A and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research. *Data Brief.* **4**, 357–362 (2015).

31. Higgins, E. E., Clarke, W. E., Howell, E. C., Armstrong, S. J. & Parkin, I. A. P. Detecting de novo homoeologous recombination events in cultivated *Brassica napus* using a genome-wide SNP array. G3: Genes|Genomes|Genetics, https://doi.org/10.1534/g3.118.200118 (2018).
32. Beló, A. *et al*. Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor. Appl. Genet.* **120**, 355–367 (2010).
33. Darracq, A. *et al*. Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. *BMC Genomics* **19**, 119 (2018).
34. Valassakis, C., Livanos, P., Minopetrou, M., Haralampidis, K. & Roussis, A. Promoter analysis and functional implications of the selenium binding protein (SBP) gene family in *Arabidopsis thaliana*. *J. Plant. Physiol.* **224–225**, 19–29 (2018).
35. Parisi, K. *et al*. The evolution, function and mechanisms of action for plant defensins. *Semin. Cell Dev. Biol*, https://doi.org/10.1016/j.semcdb.2018.02.004 (2018).
36. De Samblanx, G. W. *et al*. Mutational analysis of a plant defensin from radish (*Raphanus sativus* L.) reveals two adjacent sites important for antifungal activity. *J. Biol. Chem.* **272**, 1171–1179 (1997).
37. Li, F. *et al*. A synthetic antimicrobial peptide BTD-S expressed in *Arabidopsis thaliana* confers enhanced resistance to *Verticillium dahliae*. *Mol. Genet. Genomics* **291**, 1647–1661 (2016).
38. Abuqamar, S., Ajeb, S., Sham, A., Enan, M. R. & Iratni, R. A mutation in the expansin-like A2 gene enhances resistance to necrotrophic fungi and hypersensitivity to abiotic stress in *Arabidopsis thaliana*. *Mol. Plant. Pathol.* **14**, 813–27 (2013).
39. Uppalapati, S. R. *et al*. Loss of abaxial leaf epicuticular wax in *Medicago truncatula* irg1/palm1 mutants results in reduced spore differentiation of anthracnose and nonhost rust pathogens. *Plant. Cell* **24**, 353–370 (2012).
40. Shen, J. D., Araki, H., Chen, L. L., Chen, J. Q. & Tian, D. C. Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in *Arabidopsis thaliana*. *Genet.* **172**, 1243–1250 (2006).
41. Tan, S., Zhong, Y., Hou, H., Yang, S. & Tian, D. Variation of presence/absence genes among Arabidopsis populations. *BMC Evol. Biol.* **12**, 86 (2012).
42. Borrelli, G. M. *et al*. Regulation and Evolution of NLR Genes: A Close Interconnection for Plant Immunity. *Int. J. Mol. Sci.* **19**, 6 (2018).
43. Kang, Y. J. *et al*. Genome-wide mapping of NBS-LRR genes and their association with disease resistance in soybean. *BMC plant. Biol.* **12**, 139 (2012).
44. Madsen, L. H. *et al*. Barley disease resistance gene analogs of the NBS-LRR class: identification and mapping. *Mol. Gen. Genomics* **269**, 150 (2003).
45. Bakker, E. *et al*. A genome-wide genetic map of NB-LRR disease resistance loci in potato. *Theor. Appl. Genet.* **123**, 493–508 (2011).
46. Obermeier, C. *et al*. Genetic analysis of phenylpropanoid metabolites associated with resistance against *Verticillium longisporum* in *Brassica napus*. *Mol. Breed.* **31**, 347–361 (2013).
47. Zeise, K. & von Tiedemann, A. Host specialization among vegetative compatibility groups of *Verticillium dahliae* in relation to *Verticillium longisporum*. *J. Phytopathology* **150**, 112–119 (2002).
48. Novakazi, F. *et al*. The three lineages of the diploid hybrid *Verticillium longisporum* differ in virulence and pathogenicity. *Phytopathology* **105**, 662–673 (2015).
49. Eynck, C., Koopmann, B., Karlovsky, P. & von Tiedemann, A. Internal resistance in winter oilseed rape inhibits systemic spread of the vascular pathogen *Verticillium longisporum*. *Phytopathology* **99**, 802–811 (2009).
50. Kawakatsu, T. *et al*. Epigenomic Diversity in a Global Collection of *Arabidopsis thaliana* Accessions. *Cell* **166**, 492–505 (2016).
51. Doyle, J. J. & Doyle, J. L. Isolation of plant DNA from fresh tissue. *Focus.* **12**, 13–15 (1990).
52. Wu, Y., Bhat, P. R., Close, T. J. & Lonardi, S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.* **4**, e1000212 (2008).
53. Joehanes, R. & Nelson, J. C. QGene 4.0, an extensible Java QTL-analysis platform. *Bioinforma.* **24**, 2788–2789 (2008).
54. Aulchenko, Y. S., Koning, D. J. & Haley, C. Genome wide rapid association using mixed model and regression: a fast and simple method for genome wide pedigree-based quantitative trait loci association analysis. *Genet.* **177**, 577–585 (2007).
55. Stich, B. *et al*. Comparison of mixed-model approaches for association mapping. *Genet.* **178**, 1745–1754 (2008).
56. Price, A. L. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
57. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300 (1995).
58. Gabriel, S. B. *et al*. The structure of haplotype blocks in the human genome. *Sci.* **296**, 2225–2229 (2002).
59. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinforma.* **21**, 263–265 (2005).
60. Shin, J. H., Blay, S., McNeney, B. & Graham, J. LDheatmap: An R function for graphical display of pairwise linkage disequilibria between Single Nucleotide Polymorphisms. *J. Stat. Softw.* **16**, 1–9 (2006).
61. Stein, A. *et al*. Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in *Brassica napus*. *Plant. Biotechnol. J.* **15**, 1478–1489 (2017).
62. Grandke, F., Snowdon, R. & Samans, B. gsrc: an R package for genome structure rearrangement calling. *Bioinforma.* **33**, 545–546 (2017).
63. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinforma.* **22**, 1600–1607 (2006).

Acknowledgements

We thank Jutta Schaper (Georg August University Goettingen) for excellent technical assistance in resistance phenotyping, Matthias Frisch (Justus Liebig University Giessen) for advice in statistical analyses, Philip Howard and Andrew Leitch (Queen Mary University of London) for processing HMW DNA samples and analysis on a Bionano Genomics Irys System for optical mapping. Iulian Gabur acknowledges funding through a scholarship from the DAAD. Funding was provided by the German Federal Ministry of Food and Agriculture (BMEL grant 28-1-45.051-10) and from the German Federal Ministry of Education and Research (BMBF grant 031A325 for the German-French Plant-KBBE consortium GEWIDIS and grant 0315964 for the national consortium PreBreedYield).

Author contributions

C.O. and R.S. designed the research. A.v.T. and D.T.L. produced the phenotyping data. H.S.C. performed the Optical mapping and bioinformatics analysis. D.T.L. performed the RNASeq experiment. I.G. performed genetic and other bioinformatics analyses. I.G., H.S.C. and C.O. analyzed the data. I.G. and C.O. wrote the original draft; all authors discussed and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-61228-3.

Correspondence and requests for materials should be addressed to C.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4 Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant

Chawla H.S; Lee H.T; Gabur I; Amutha S; Obermeier C; Schiessl S.V; Song J; Liu K; Guo L; Parkin I.A.P; Snowdon R.J

Plant Biotechnology Journal (2020), pp. 1–11
doi: 10.1111/pbi.13456

Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2020), pp. 1–11 doi: 10.1111/pbi.13456

Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant

Harmeet Singh Chawla[1], HueyTyng Lee[1], Iulian Gabur[1], Paul Vollrath[1], Suriya Tamilselvan-Nattar-Amutha[1], Christian Obermeier[1], Sarah V. Schiessl[1,2], Jia-Ming Song[3], Kede Liu[3], Liang Guo[3], Isobel A. P. Parkin[4] and Rod J. Snowdon[1,*]

[1]*Department of Plant Breeding, Justus Liebig University, Giessen, Germany*
[2]*Department of Botany and Molecular Evolution, Senckenberg Research Institute and Natural History Museum Frankfurt, Frankfurt am Main, Germany*
[3]*National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China*
[4]*Agriculture and Agri-Food Canada, Saskatoon, SK, Canada*

Received 10 February 2020;
revised 12 July 2020;
accepted 21 July 2020.
**Correspondence* (Tel +49 641 9937420; fax +49 641 9937429; email rod.snowdon@agrar.uni-giessen.de)

Keywords: *Brassica napus*, polyploidy, genome rearrangement, presence–absence variants, PAV.

Summary

Genome structural variation (SV) contributes strongly to trait variation in eukaryotic species and may have an even higher functional significance than single-nucleotide polymorphism (SNP). In recent years, there have been a number of studies associating large chromosomal scale SV ranging from hundreds of kilobases all the way up to a few megabases to key agronomic traits in plant genomes. However, there have been little or no efforts towards cataloguing small- (30–10 000 bp) to mid-scale (10 000–30 000 bp) SV and their impact on evolution and adaptation-related traits in plants. This might be attributed to complex and highly duplicated nature of plant genomes, which makes them difficult to assess using high-throughput genome screening methods. Here, we describe how long-read sequencing technologies can overcome this problem, revealing a surprisingly high level of widespread, small- to mid-scale SV in a major allopolyploid crop species, *Brassica napus*. We found that up to 10% of all genes were affected by small- to mid-scale SV events. Nearly half of these SV events ranged between 100 bp and 1000 bp, which makes them challenging to detect using short-read Illumina sequencing. Examples demonstrating the contribution of such SV towards eco-geographical adaptation and disease resistance in oilseed rape suggest that revisiting complex plant genomes using medium-coverage long-read sequencing might reveal unexpected levels of functional gene variation, with major implications for trait regulation and crop improvement.

Introduction

The recent allopolyploid species *Brassica napus* L. (oilseed rape/canola/kale/rutabaga; genome AACC, 2n = 38) rapidly emerged as a globally important crop. Genome assembly and resequencing of *B. napus* (Chalhoub *et al.*, 2014) revealed a highly complex and strongly duplicated genome with an unexpected extent of segmental exchanges among homoeologous chromosomes. In synthetic *B. napus* accessions, genome structural variants frequently span whole chromosomes or chromosome arms (Chalhoub *et al.*, 2014; Samans *et al.*, 2017). Naturally formed *B. napus* also shows widespread homoeologous exchanges, with similar distribution patterns (Hurgobin *et al.*, 2018; Samans *et al.*, 2017), that apparently arose during the allopolyploidization process (Leflon *et al.*, 2006; Nicolas *et al.*, 2007; Szadkowski *et al.*, 2010). The wide extent of segmental deletion/duplication events in both synthetic and natural *B. napus* has been confirmed using other genome-wide analysis methods, for example visualization based on mRNAseq data (He *et al.*, 2017) or deletion calling from SNP array data (Gabur *et al.*, 2018; Grandke *et al.*, 2016). Critically, numerous examples have connected genome SV in *B. napus* to important agronomic traits (Gabur *et al.*, 2018; Gabur *et al.*, 2019; Liu *et al.*, 2012; Stein *et al.*, 2017). These studies revealed the important role of SV in the creation of *de novo* variation for adaptation and breeding; however, the methods used were not yet capable of resolving SV at gene scale.

A first example of intragenic SV impacting quantitatively inherited traits in *B. napus* was reported by Qian *et al.* (2016), who demonstrated that deletion of exons 2 and 3 from a *B. napus* orthologue of Mendel's 'Green Cotyledon' gene (the Staygreen gene *NON-YELLOWING 1*; *NYE1*) associated with quantitative variation for chlorophyll and oil content. Unfortunately, such small deletions are challenging to reliably detect using short-read sequencing or low-cost marker arrays, so that their genome-wide extent could not yet be investigated in detail. In this study, using *B. napus* as an example for a plant genome with widespread structural variation, we demonstrate the power of whole-genome long-read sequencing for high-resolution detection of intragenic SV. The results reveal widespread functional variation on a completely unexpected scale, suggesting that small- to mid-scale SV may be a major driver of functional gene diversity in this recent polyploid crop. With the growing accessibility, accuracy and cost-effectiveness of long-read sequencing,

Please cite this article as: 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. and Snowdon, R. J. (2020) Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. *Plant Biotechnol J.*, https://doi.org/10.1111/pbi.13456

our results suggest that there could be enormous promise in revisiting complex crop genomes to discover potentially novel functional SV which has previously been overlooked.

Results and discussions

Long-read sequencing reveals novel SV diversity in *B. napus*

We sequenced 4 *B. napus* accessions with long reads using the Oxford Nanopore Technology (ONT) and 8 additional accessions using the Pacific Biosciences (PacBio) platform (data from Song *et al*. (2020)). The genotype panel included three vernalization-dependent winter-type accessions, 3 vernalization-independent spring-type accessions, 4 semi-winter accessions and 2 synthetic *B. napus* accessions (a winter-type and a spring-type). All accessions were sequenced to between ~ 30x and ~ 50x whole-genome coverage (between 30 and 50 GB of data). Reads were aligned to the *B. napus* Darmor-*bzh* version 4.1 reference genome (Chalhoub *et al*., 2014) using the long-read aligner NGMLR (https://github.com/philres/ngmlr) (Sedlazeck *et al*., 2018) and called for genome-wide SV using the SV-calling algorithm Sniffles (Sedlazeck *et al*., 2018). N50 values ranging from 10 552 to 15 369 bp were obtained for the 8 PacBio datasets, while in the 4 ONT datasets the N50 ranged from 10 756 to 28 916 bp (Table 1, Table S1). After aligning to the Darmor-*bzh* v4.1 reference genome, the total number of SV events called by Sniffles ranged from 51 463 to 108 335. PacBio and ONT sequencing can potentially result in systemic differences in SV calling because of their different error profiles. PacBio sequencing is known to be enriched for small insertion errors, whereas ONT suffers from deletions especially in homopolymer regions (Sedlazeck *et al*., 2018). To neutralize systematic bias due to different error profiles, we used different noise models for aligning ONT and PacBio datasets. For ONT datasets, we used the '-x ont' flag for NGMLR, whereas this was omitted for PacBio datasets because NGMLR expects a PacBio dataset by default. To demonstrate that this flag was effective in expunging the majority of spurious SV calls, we compared variant calls between ONT and PacBio datasets from the same winter oilseed rape genotype, Express 617, using data used by Lee *et al*. (2020) to assemble the Express 617 genome. Overall, 27 106 and 33 424 quality-filtered SV calls from respective ONT and PacBio libraries of Express 617 were merged using SURVIVOR, resulting in a combined set of 34 885 SV. We found 82.7 per cent (28 857) of the total SV to be supported by both ONT and PacBio reads, indicating that both long-read technologies generate highly suitable data to accurately capture a majority of small- to mid-scale genome-wide SV events.

To minimize false-positive calls derived from reference misassemblies, we followed a highly stringent quality-filtering approach that removed 54.4–59.4% of the total predicted SV. This procedure resulted in a final set of 27 106 to 44 516 high-quality SV events (Table 1). To evaluate the impact of assembly errors on SV-calling rates, we compared results after aligning (using the same procedure) to a pseudo-reference constructed by combining the high-quality long-read reference assemblies of *Brassica rapa* (A subgenome) and *Brassica oleracea* (C subgenome) published recently by Belser *et al*. (2018). Using this pseudo-reference assembly, we detected between 41 436 and 50 907 quality-filtered SV across the 12 *B. napus* genotypes. There are two possible explanations for the higher number of SV. Firstly, the pseudo-reference assembly (957 Mbp) is nearly 10 per cent larger than the *B. napus* Darmor-*bzh* v4.1 reference (849.7 Mbp). Secondly, SV detected using the pseudo-reference assembly will also reflect genomic differences between the unknown diploid progenitors of *B. napus* and the two diploid genotypes from which this pseudo-assembly was generated. To further validate our SV detection approach, we therefore compared the number of SV per megabase, detected using the two different genome assemblies for each of the 19 chromosomes across 12 genotypes. This showed a correspondence of 77.08 per cent, suggesting that the latter may be the predominant cause.

After alignment to the Darmor-*bzh* v4.1 reference genome, the median detected SV size across the 12 accessions ranged from 296 bp to 584 bp. The spring-type accessions N99 and PAK85912 had the largest median SV size (509 and 584 bp, respectively), which might be attributable to the longer read lengths for these two genotypes (N50 = 27 139 bp and 28 916 bp, respectively) (Figure 1a). The largest SV event (28 777 bp) was also detected in the spring-type accession PAK85912, suggesting that read length plays a critical role in the ability to detect large and complex SV events. A complete lack of SV calls on chromosome C02 of the winter oilseed rape accession R53, using both the pseudo-reference and the Darmor-*bzh* assembly. No SV calls were observed on chromosome C02 of the synthetic *B. napus* accession R53. This is because C02 is deleted

Table 1 Number and size distributions of SV detected in 12 *B. napus* genotypes

Genotype	Data type	Ecotype	N50 for raw reads	Quality-filtered SV	Intragenic SV	Maximum size of SV	Median size SV
Express 617	ONT	Winter	10 756	27 106	5898	16 931	341
Quinta	PacBio	Winter	14 192	32 349	7085	15 869	353
Tapidor	PacBio	Winter	14 448	32 757	7291	15 289	344
ZS11	PacBio	Semi-winter	10 552	37 496	9004	11 312	281
Zheyou7	PacBio	Semi-winter	12 370	38 590	9042	17 001	305
Gangan	PacBio	Semi-winter	14 064	35 560	8366	14 264	335
Shengli	PacBio	Semi-winter	13 828	39 622	9501	12 207	321
PAK85912	ONT	Spring	28 916	23 177	5011	28 777	584
N99	ONT	Spring	27 139	34 848	7482	26 183	509
Westar	PacBio	Spring	13 810	37 138	8575	17 615	332
R53	ONT	Winter synthetic	11 253	33 851	8647	12 635	296
No2127	PacBio	Spring synthetic	15 369	44 516	10 675	15 565	304

ONT, Oxford Nanopore Technologies; PacBio, Pacific Biosciences; SV, structural variant.

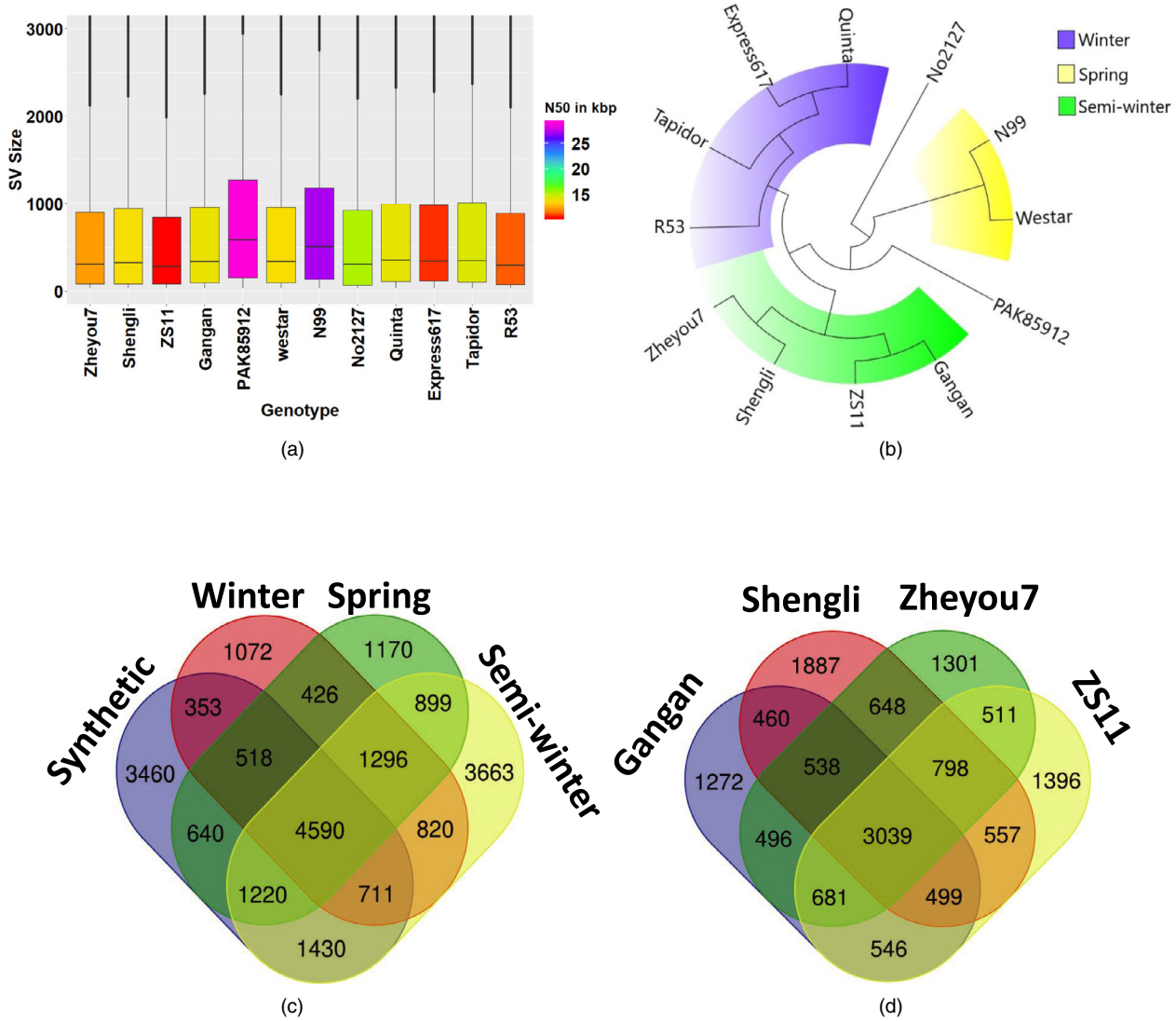


Figure 1 Gene scale SV in oilseed rape. (a) Box plots showing size distributions of SV events detected in 12 *B.napus* genotypes. (b) Maximum-likelihood tree showing genetic relationships among 12 *B. napus* genotypes based solely on genome-wide SV events, revealing clear clustering into the appropriate eco-geographical morphotype groups. (c) Venn diagram showing the numbers of common or unique genes carrying intragenic SV events across three divergent ecotypes and synthetic *B. napus*, respectively. (d) Venn diagram representing the numbers of common or unique genes carrying intragenic SV events across four semi-winter *B. napus* accessions.

in R53 and replaced by its homeolog A02, as reported previously by Stein *et al*. (2017). Around half of all detected, high-confidence SV events (46.8 to 53.2 % across the 12 genotypes) ranged in size from ~ 100–1000 bp (Tables S2-S3). These small SV represent a novel genetic diversity resource that was previously unnoticed due to the insufficient resolution of high-throughput genotyping platforms such as SNP genotyping arrays and a very high false-positive rates (up to 89%) of short-read sequencing data (Mahmoud *et al*., 2019; Sedlazeck *et al*., 2018).

Subgenomic differences in SV frequency

Comparison of subgenomic SV frequency revealed significantly higher numbers of small- to mid-scale SV per megabase in the *B. napus* A subgenome than the C subgenome in all twelve analysed genotypes (Figure 2a and b, Tables S4-S5). This reflects a corresponding subgenomic bias also observed for large-scale SV in *B. napus* (Samans *et al*., 2017), and this could also be attributable to repeated introgressions from the A genome of *B. rapa* during the breeding history of *B. napus* (Lu *et al*., 2019). Samans *et al*. (2017) reported a significant enrichment for large-scale segmental deletions in the C-subgenome of *B. napus* resulting from homoeologous exchanges. In contrast, we observed no bias for small to mid-scale deletions in the C-subgenome of the 12 sequenced *B. napus* accessions (Tables S6-S7). This indicates that a different molecular mechanism may be responsible for the generation of large and small to mid-scale SV events in the rapeseed genome. Unexpectedly, we found that between 5% (Express 617) and 10% (No2127) of all genes detected in the twelve accessions were affected by small to mid-scale SV events. This represents a previously completely unknown extent of functional gene modification as a result of post-polyploidization genome restructuring. It also underlines the massive selection potential arising from intergenomic disruption during the act of allopolyploidization (Nicolas *et al*., 2008; Nicolas

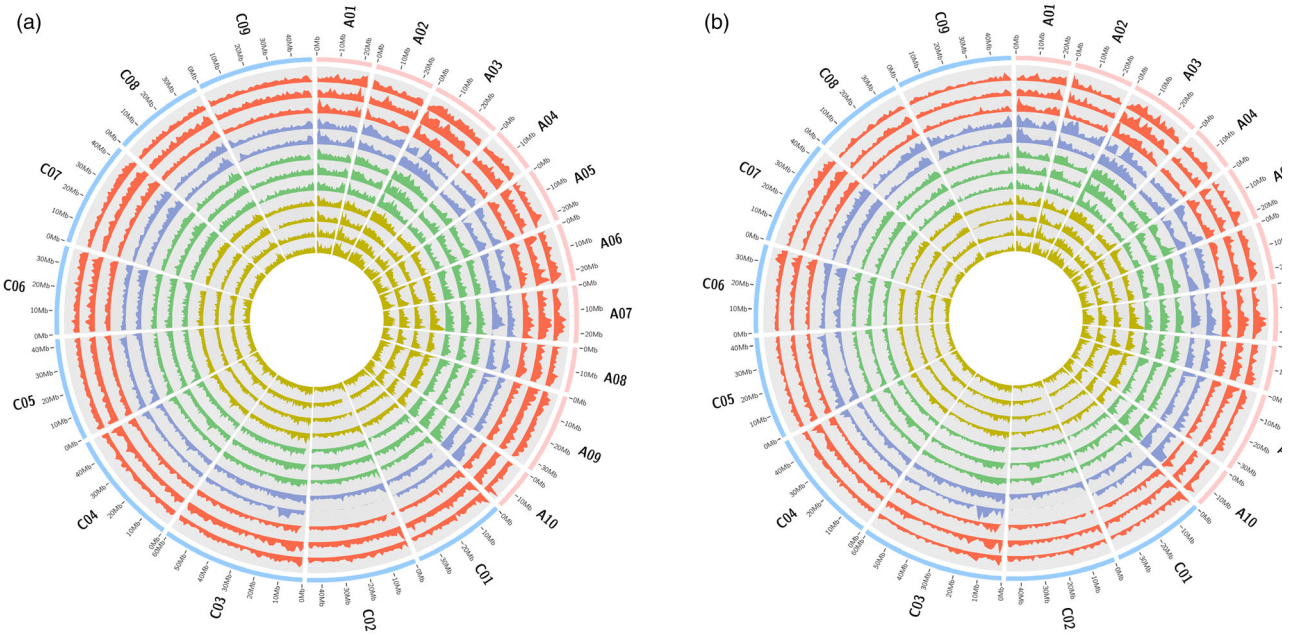


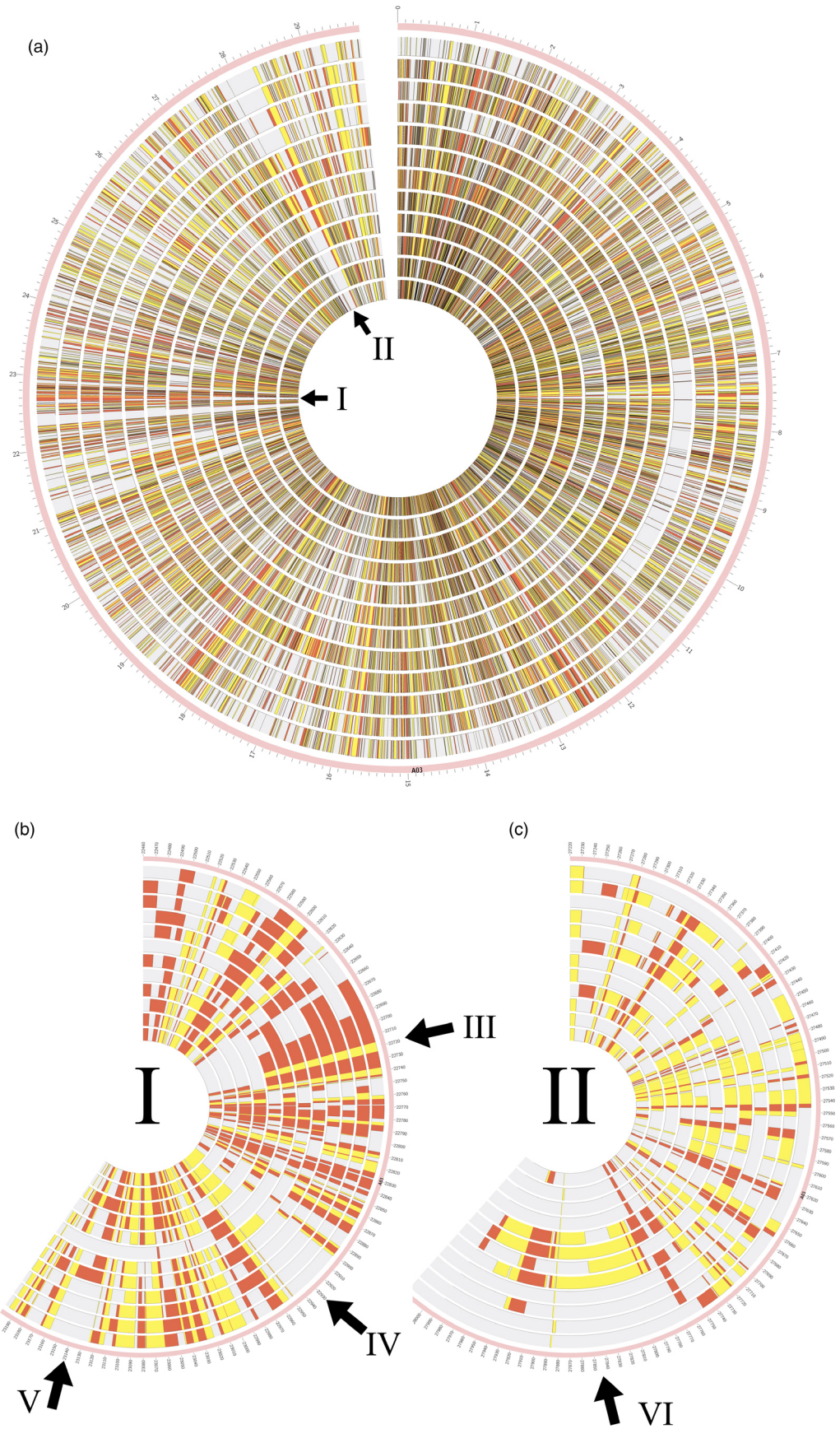
Figure 2 Genome-wide distribution of small- to mid-scale insertions and deletions in *B. napus*. (a) Circos plot depicting number of small to mid-scale deletion events calculated in 1 Mbp windows across 19 chromosomes of 12 *B. napus* accessions. Each track represents a single genotype in the following order from outside to inside: Express 617, Quinta, Tapidor, R53, No2127, N99, Westar, PAK85912, Gangan, Shengli, Zheyou7 and ZS11. Colours of tracks represent different types of *B. napus*. The red, blue, green and yellow track colours represent winter-type, synthetic, spring-type and semi-winter accessions, respectively. (b) Circos plot depicting the frequency of small to mid-scale insertion events in 1 Mbp windows across 19 chromosomes of 12 *B. napus* genotypes. Each track represents a single genotype in the following order from outside to inside: Express 617, Quinta, Tapidor, R53, No2127, N99, Westar, PAK85912, Gangan, Shengli, Zheyou7 and ZS11. Colours of tracks represent different types of *B. napus*. The red, blue, green and yellow track colours represent winter-type, synthetic, spring-type and semi-winter accessions, respectively.

et al., 2007; Szadkowski *et al*., 2010), and the great significance of post-polyploidization intergenomic restructuring for polyploid crop evolution (Samans *et al*., 2017).

Small- to mid-scale SV underlining eco-geographical differentiation in *B. napus*

As expected, strong SV differentiation from the winter-type oilseed reference genotype Darmor-*bzh* was found in the divergent semi-winter and spring ecotypes, and in genetically distant synthetic *B. napus* accessions R53 and No2127 (Figure 3). Unexpectedly, however, the winter-type accessions Express 617, Tapidor and Quinta also showed high levels of SV compared to Darmor-*bzh*, despite a related breeding history and partially shared pedigree (e.g. Express 617). According to (Lu *et al*., 2019), who used whole-genome resequencing data to investigate the species origin and evolution of *B. napus*, spring and semi-winter types arose only very recently (<500 years) from winter types. Our data concur with this assumption, with fewer genes carrying SV in winter-type accessions (1072) than in spring (1170) or semi-winter (3663) ecotypes (Figure 1c). Furthermore, we also detected small- to mid-scale SV within each ecotype; for example, 1272–1887 genes carrying unique SV events were found among the four semi-winter accessions (Figure 1d). The unexpectedly high structural gene diversification both between and within ecotypes suggests that *de novo* generation of small- to mid-scale SV may also be ongoing in recent breeding history. Overall, 4590 of the called intragenic SV were common among the four *B. napus* forms, indicating putative SV events specific to Darmor-*bzh*. These could possibly be attributed to errors in the Darmor-*bzh* reference assembly; however, the similar number of unique intragenic SV detected only in semi-winter types (3663) suggests that this frequency is not unexpected in the context of the other results. Repeating the analysis with the concatenated pseudo-reference from *B. rapa* plus *B. oleracea* gave comparable results

Figure 3 Small- to mid-scale genomic rearrangements on chromosome A03 of *B.napus*. (a) Circos plot showing small- to mid-scale insertion and deletion events in 12 *B. napus* accessions, using chromosome A03 as an example. Each track represents a single accession in the following order from outside to inside: Express 617, Quinta, Tapidor, R53 (all winter type), No2127, N99, Westar, PAK85912 (spring type), Gangan, Shengli, Zheyou7 and ZS11 (semi-winter type). Deletions are represented by yellow blocks, whereas insertions are shown by red blocks. Darker blocks in (A) represent regions containing both deletions and insertions in different genotypes. Arrows I and II mark selected segmental SV events specific for a particular ecotype. (b) Expanded view of the chromosome segment depicted by arrow I in A. Arrow III represents a 50 kbp region containing segmental deletion and insertion events detected in all winter and spring ecotypes but not in the semi-winter types. Arrow IV indicates a 40 kbp region containing segmental deletions detected only in the four semi-winter types and three of spring types. Arrow V indicates a 40 kbp region containing segmental insertions detected only in the four semi-winter types and one of the spring types. (c) Expanded view of the chromosome segment depicted by arrow II in A. Arrow VI indicates a 120 kbp region containing segmental insertions only in the four spring types.



(6248 common among all sequenced *B. napus* forms, 2919 unique to semi-winter ecotypes).

To evaluate the influence of SV on eco-geographical adaptation and potential species diversification, we constructed a maximum-likelihood (ML) tree for the 12 *B. napus* lines based solely on SV detected using long-read sequencing data. The resulting tree (Figure 1b) comprised 3 divergent clades representing 3 ecotypes of *B. napus* (winter, semi-winter and spring). In contrast with genetic clustering based on genome-wide SNP data, which reveals high sequence diversification between synthetic and natural *B. napus* (Bus *et al*., 2011), the two synthetic accessions R53 and No2127 did not fall into separate clades. Instead, the winter-type R53 clustered closest together with the natural winter-type accessions and the spring-type No2127 clustered with the natural spring-type accessions. This suggests that small- to mid-size SV events originating during or immediately after allopolyploidization might rapidly confer eco-geographical adaptation. Although hundreds to thousands of genes carrying unique SV events were detected in each individual accession, the intriguing observation that their cumulative clustering reflects eco-geographical adaptation forms suggests a possible key role of SV in rapid functional adaptation. Overall, the distribution and frequency of SV events in all investigated accessions suggest that small- to mid-scale SV may be a major, previously unknown source of functional genetic variation in *B. napus*.

Unfortunately, a catalogued and validated 'truth set' of genomic SV is not yet established for *B. napus* or other complex plant genomes. This makes it crucial to validate SV predicted from long reads using independent validation methods. On the other hand, manual verification of thousands of SV events (for example using PCR) is not realistic. To obtain first insight into the validity of the SV called using our pipeline, we selected relevant potentially functional examples representing possible functional mutations in flowering-time and disease resistance-related genes. We validated the detected SV events using different independent methods in a total of 4 *B. napus* genotypes including two springs, one winter and a synthetic.

Small- to mid-scale SV events impact *B. napus* flowering-time pathway genes

In order to understand the impact of gene scale rearrangements on eco-geographical adaptations in *B. napus*, we examined the abundance of SV in the known *B. napus* orthologs of all known genes from the *Arabidopsis* flowering-time pathway. Whereas most of these genes are present in only a single copy in *Arabidopsis*, all have multiple duplicates in *B. napus* (Schiessl *et al*., 2014). Although many *B. napus* flowering-time gene orthologues are known to be affected by copy-number variation (CNV), the exact positions of copy-number variants and other small- to mid-scale forms of SV could not be determined from previous short-read resequencing data (Schiessl *et al*., 2017). In this study, we used the long-read data from two winter and two spring oilseed rape accessions to identify CNV in the form of duplications within flowering-time genes (Tables S8 and S9). Only 3 to 4 per cent of all genes in the flowering-time pathway showed CNV in these genotypes, whereas 24.7 % (44 of 178), including numerous key regulatory genes, contained one or more small to mid-scale insertions or deletions. Therefore, we did not perform CNV analysis in the remaining 8 genotypes and focused instead on small- to mid-scale SV in the full dataset. For example, we detected a 90 bp insertion in an orthologue of *Vernalization Insensitive 3* on chromosome C03 (*BnVIN3.C03*, *BnaC03g12980D*) in 3 out of 12 total genotypes, Express 617, No2127 and Zheyou7 (Figure 4a). Successful validation of this insertion via PCR, using primers designed from the SV-flanking sequences, is shown in Figure 4b. The same insertion was undetectable using only the short-read sequence-capture data of Schiessl *et al*. (2017). In two out of three spring accessions, N99 and PAK85912, we detected a 2.8 kbp insertion in a *B. napus* orthologue of the key vernalization regulator *Flowering Locus C* (*BnFLC.A02, BnaA02g00370D*), a variant previously

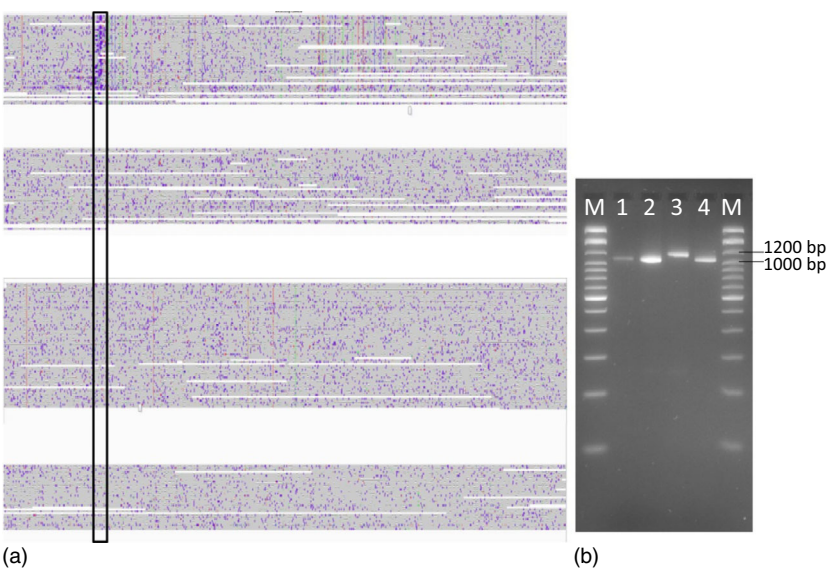


Figure 4 Small-scale SV in a key flowering-time gene *Vernalization Insensitive 3* on chromosome C03 (*BnVIN3.C03*). (a) 90 bp insertion detected only in accession Express 617 (highlighted in the black box) in an orthologue *BnVIN3.C03* was revealed by aligning ONT reads from 4 different genotypes to the Darmor-*bzh* reference version 4.1. (b) Agarose gel image of PCR products spanning the insertion polymorphism. M:100 bp ladder; 1: N99; 2: PAK85912; 3: Express 617; 4: R53. As expected, Express 617 exhibits 1170 bp PCR product, whereas the other three genotypes all show a 1060 bp amplicon.

reported by Chen *et al*. (2018) to be causal for early flowering (Appendix S1).

In a second case study, we analysed SV events in key vernalization genes that differentiate between the vernalization-dependent and vernalization-independent *B. napus* accessions in our panel. A number of interesting, putative functional variants were detected. For example, we detected a 1.3 kbp deletion (Figure 5**)** in the putative promoter of *BnFT.A02* (*BnaA02g12130D*), located between 6 365 143 and 6 366 504 bp on chromosome A02. This deletion was exclusively detected in all 4 spring accessions. *BnFT.A02* has been reported to be differentially expressed among winter, spring and semi-winter type *B. napus* by Wu *et al*. (2019) and the 1.3 kbp deletion in its promoter region might explain the cause for this differential expression. To further validate our hypothesis that this 1.3 kbp deletion in *BnFT.A02* associates with vernalization behaviour of oilseed rape, we genotyped it using a locus-specific PCR assay in 25 vernalization-dependent and 25 vernalization-independent accessions from the ERANET-ASSYST *B. napus* diversity set (Bus *et al.*, 2011) (Table S10). Eighty per cent of the vernalization-independent oilseed rape accessions were found to contain the 1.3 kbp deletion in *BnFT.A02*, whereas a majority of vernalization-dependent winter types (79 per cent) showed no deletion. The strong co-segregation of this SV with vernalization behaviour might indicate a potentially crucial role for this genomic rearrangement in eco-geographical adaptation.

Intragenic SV events associated with disease resistance in oilseed rape

Samans *et al*. (2017) and Hurgobin *et al*. (2018) revealed that defence-related R-genes involved in monogenic resistance are particularly enriched in genome regions affected by large-scale SV in *B. napus*. Gabur *et al*. (2020) reported gene presence–absence variations (PAV) in 23 to 51% genes within confidence intervals of QTL for *V. longisporum* resistance in *B. napus*. In a third case study related to a prominent disease resistance in oilseed rape, we investigated the impact of SV in resistance-related genes co-localizing with QTL for quantitative disease resistance in a bi-parental cross between the sequenced accessions Express 617 and R53. These two accessions differ strongly in their resistance reaction to the important fungal pathogen *V longisporum* (Obermeier *et al*., 2013), and SV detected between the two parental lines were selected for validation based on their co-localization to resistance-related genes in corresponding resistance QTL (see Methods for selection criteria for PCR validation of SV events). Most interestingly, we identified a 700 bp deletion in R53 that caused the loss of three exons of a *4-Coumarate:CoA Ligase* (*4CL*) gene (*BnaC05g15830D*). In the genetic map from the Express 617 × R53 mapping population, this gene is located within a major QTL for *V. longisporum* resistance on *B. napus* chromosome C05 (Obermeier *et al*., 2013). 4CL is a critical enzyme involved in the phenylpropanoid pathway (Li *et al*., 2015) and Obermeier *et al*. (2013) reported that major QTL for phenylpropanoid compounds co-localized with the QTL for *V. longisporum* resistance in the Express 617 × R53 mapping population. Locus-specific PCR primers, spanning the putative SV predicted by the long sequence reads, amplified 900 bp and 200 bp fragments for Express 617 and R53, respectively (Figure 6a,b), confirming the expected 700 bp deletion. Re-screening of the PCR markers for the 700 bp deletion in the doubled haploid mapping population from Express 617 × R53 confirmed their co-localization with the QTL and a strong effect on resistance of up to R^2 = 19.4%.

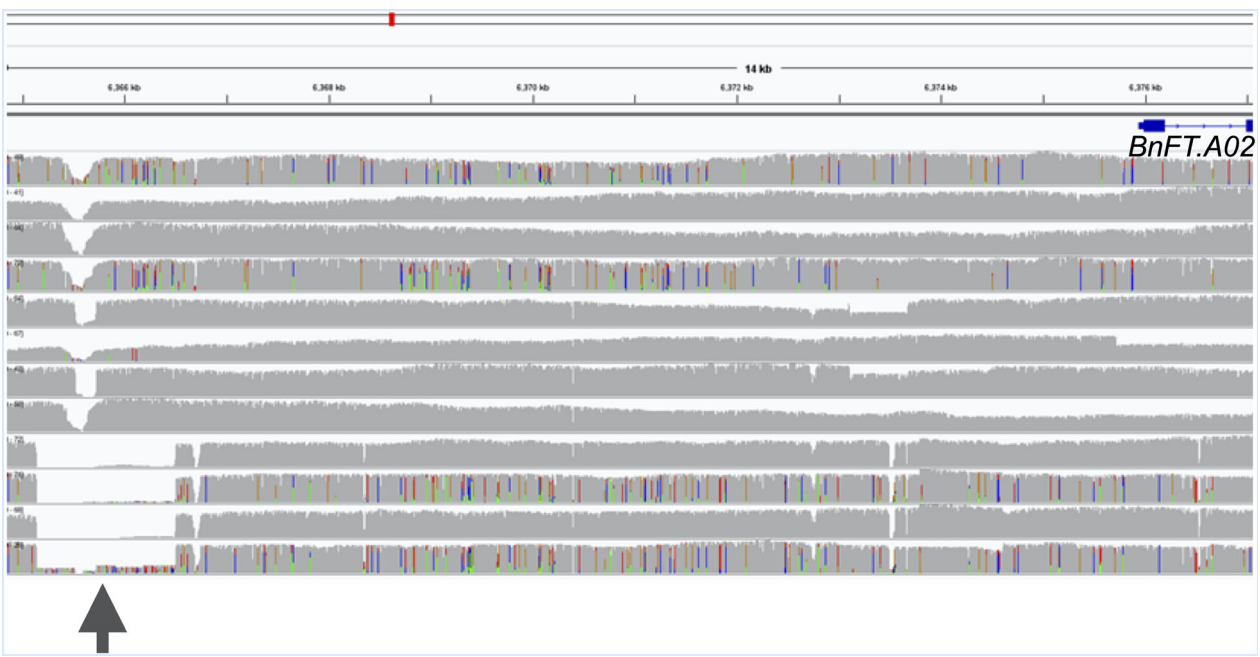


Figure 5 1.3 kbp deletion in the putative promoter of *BnFT.A02* (*BnaA02g12130D*). Each track represents a single genotype in the following order from top to bottom: Express 617, Tapidor, Quinta, R53, Shengli, ZS11, Gangan, Zheyou7, No2127, N99, Westar and PAK85912. The arrow indicates a 1.3 kbp deletion in putative promoter region in *BnFT.A02* (*BnaA02g12130D*) for all four spring accessions (No2127, N99, Westar and PAK85912).

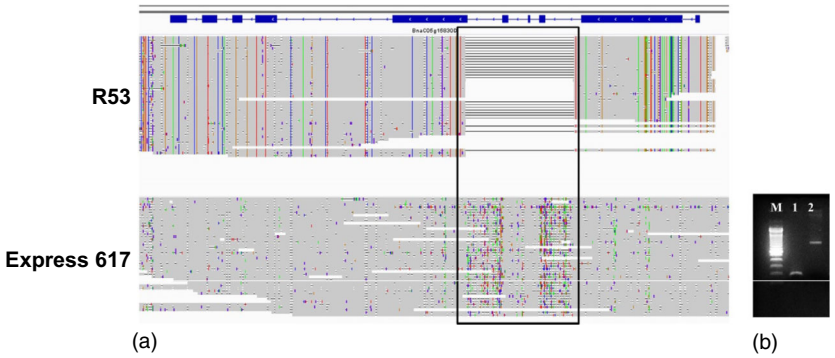


Figure 6 SV contributing to disease resistance. (a) 700 bp deletion (highlighted in the black box) was detected in accession R53 that caused the loss of three exons of a *4-Coumarate:CoA Ligase* (*4CL*) gene *(BnaC05g15830D)*. (b) Agarose gel image of PCR product from the same deletion. M: 100 bp ladder; 1: PCR product originating from R53; 2: product originating from Express 617. As expected, Express 617 exhibits a 900 bp PCR product, whereas R53 shows a 200bp amplicon.

Implications of long-read sequencing technologies for discovery of functional diversity

Of nine additional SV events we evaluated using PCR, all showed the expected PCR products corresponding to the deletions or insertions predicted by the long-read SV calling. These results underline the apparent effectiveness of long sequence reads for accurately detecting and anchoring insertions/deletions in a broad size range from under 100 bp up to multiple kbp. In contrast, Illumina short reads from regions corresponding to insertions not present in available reference genomes remain un-aligned in alignment-based resequencing approaches, meaning that their genomic localization using short-read data can be achieved only by whole-genome *de novo* assembly. Our results in *B. napus* showed that *de novo* SV events appear to occur at an unexpectedly high rate. Depending on the genotype, 49–75% of intragenic SV events were located in exons in relation to introns (Table S11), indicating high potential for functional implications. Given the very high rate and the differences among genotypes, it remains unclear how many high-quality reference genomes will be necessary to construct a representative pangenome that captures the majority of the genome-wide functional SV landscape.

This study provides one of the very first insights into genome-wide, gene scale SV linked to important agronomic traits in a major crop species. Recently, Yang et al. (2019) revealed a similar scale of widespread SV by comparing whole-genome assemblies of two diverse maize accessions. However, the cost of genome assembly is still much too high to capture the full extent of species-wide SV in large numbers of genotypes, particularly in species like *B. napus* with dynamic polyploid genomes in which genome rearrangement may even still be ongoing. Our successful verification of 10 out of 10 SV selected events via PCR (Table S12) gives us high confidence that SV predicted using medium-coverage long-read data with our calling strategy are genuine. This provides a relatively cost-effective method to assay larger germplasm collections without ascertainment bias.

The occurrence of SV events in a size range corresponding to intragenic rearrangements (~100–1000 nt) has been ignored in most crop species in the past, due to the limited resolution of short-read resequencing. Although presence–absence calling from genome-wide SNP array data has been successful in isolated cases in establishing QTL associations (e.g. Gabur et al., 2018a), SNP-based genome-wide association (GWAS) studies are unable to tag causative SV in crops and genome regions in which high levels of LD decay surround the SV events (Zhou *et al*., 2019). Array-based approaches to call PAV or homoeologous exchanges (e.g. Grandke *et al*., 2016) are therefore likely to ignore potentially functional SV events. Reduced costs, considerably improved read accuracy and significantly increased average read lengths today make long-read sequencing technologies a viable option not only for accurate assemblies of complex plant genomes (Belser *et al*., 2018), but increasingly also for genome-wide resequencing. Our results suggest that simple reference-based resequencing and alignment with long reads can uncover a new dimension of genetic and genomic diversity associated with important traits in crop plants. Particularly in polyploid plants (Schiessl *et al*., 2019), this may lead to discovery of previously unknown levels of functional diversity of major interest for breeding and crop adaptation.

Experimental procedures

Plant material

We chose 12 *B. napus* genotypes (Table 1) comprising of three winter, four semi-winter, three spring and two synthetics (one each of winter and spring).

DNA isolation for Oxford Nanopore Technology (ONT) sequencing

High-molecular-weight DNA was isolated using DNA isolation protocol modified from Mayjonade *et al*. (2016). Young leaves were harvested from rapeseed plants at 4–6 leaf stage and flash frozen using liquid nitrogen. Frozen leaf material was ground to fine powder using a mortar and pestle and transferred to 15 mL Falcon tube. A total of 4–5 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) was added in order to disrupt the cell wall. The lysate was incubated for 30 min at 37°C in a thermomixer. A total of 0.3 volumes of 5M potassium acetate was added to the lysate and spun at 8000g for 12 min at 4°C to precipitates sodium dodecyl sulphate (SDS) and SDS-bound proteins in order to obtain clean DNA. Finally, magnetic beads were used to recover cleaned DNA.

Library preparation for ONT sequencing

Between 1–3 ug of DNA was used to prepare the sequencing library, using the ligation sequencing kit SQK-LSK108 or SQK-LSK109 according to the manufacturer's recommendations. Genomic DNA was subjected to end repair followed by a bead clean-up. Sequencing adaptors were then ligated to the end-repaired DNA. Finally, the adaptor ligated DNA was once again subjected to bead cleaning. DNA was finally loaded onto an Oxford Nanopore MinION flow cell for sequencing.

Pacific biosciences (PacBio) sequencing

Raw PacBio reads originating from 8 genotypes (Quinta, Tapidor, No2127, Westar, Gangan, Shengli, Zheyou7 and ZS11) were downloaded from NCBI short-read archive (Accession number PRJNA546246) with the permission from the authors.

Bioinformatics analysis

Alignment and SV calling for ONT data

Raw fast5 files obtained by the MinION device were base-called using ONT provided base-caller, Albacore. Raw uncorrected reads from various flow cells were combined into single fastq file for each genotype. This fastq file was used to align the Nanopore reads to the publically available *B. napus* reference genome assembly Darmor-bzh v4.1 (Chalhoub *et al*., 2014) and also to a concatenated pseudo-reference assembly comprising the *B. rapa* and *B. oleracea* reference assemblies recently published by Belser *et al*. (2018), using NGMLR version 0.2.7 (Sedlazeck *et al*., 2018) with default settings except for '-x ont' flag, representing parameter presets for ONT. NGMLR produced an un-sorted SAM file as an output, which was converted to a sorted BAM file using Samtools version 1.9 (Li *et al*., 2009). Genomic variants were called using Sniffles version 1.0.10 (Sedlazeck *et al*., 2018) using the preset parameters.

Alignment and SV calling for PacBio data

Since 8 PacBio libraries contained nearly 70–80 Gbp of sequencing data, we randomly selected 50 Gbp of data for further analysis in order to obtain quantitatively comparable data to the Nanopore sequencing. This 50 Gbp of data was then aligned as per section 1.4.1 to the publicly available *B. napus* reference and also to the concatenated pseudo-reference assembly, using NGMLR version 0.2.7 with default settings. NGMLR produced an un-sorted SAM file as an output, which was converted to a sorted BAM file using Samtools version 1.9. Genomic variants were called using Sniffles (version 1.0.10) using the preset parameters.

Quality filtering of the predicted SV events for both ONT and PacBio datasets

We performed a very stringent quality filtering on the sniffles predicted SV events. Since the study was focused on small-scale insertions or deletions, we removed all predicted translocations and duplications. Furthermore, it is nearly impossible to validate the authenticity of such SV events, as many may represent mis-positioning of genomic fragments in the reference assembly, we only considered SV scored as 'PASS' by Sniffles and ignored those scored as 'UNRESOLVED'. Sniffles report SVs with both within-alignment (AL) and split-read (SR) information. AL-type SV are usually small indels that can be spanned within a single alignment, whereas large or complex events lead to SR alignments (Sedlazeck *et al*., 2018). To ensure only the high-confidence SV were selected, all SV which were not supported by a 'within-alignment: AL' flag were discarded. This might lead to an under-estimation and bias in the size distribution of the detectable SV. However, at this point of time the accuracy of publically available genome from *B. napus* is not high enough to distinguish large and complex SV events from assembly errors.

Comparison of SV calls between ONT and PacBio datasets for Express 617

To evaluate the impact of the sequencing technology on the general ability to accurately call small- to mid-scale SV, we used SURVIVOR version 1.0.7 (Jeffares *et al*., 2017) to calculate overlaps between SV calls for ONT and PacBio datasets for the same genotype, Express 617. In a first step, quality-filtered SV calls from the respective PacBio and ONT datasets were merged using 'SURVIVOR merge' using the settings 'Max distance between breakpoints 1000, Minimum number of supporting caller 1, Take the type into account 1, Take the strands of SVs into account −1, Estimate distance based on the size of SV −1, Minimum size of SVs to be taken into account −1'. The merged variant calling file (VCF) was then used to 'force call' the SV across both datasets using Sniffles version 1.0.10. These forced-called SV were again merged into a single VCF using 'SURVIVOR merge' with 'Max distance between breakpoints 1000, Minimum number of supporting caller −1, Take the type into account 1, Take the strands of SVs into account −1, Estimate distance based on the size of SV −1, Minimum size of SVs to be taken into account −1'. Overlap was estimated by counting the number of SV that could be genotyped in both datasets.

CNV calling

CNV detection was performed using the method published by Stein *et al*. (2017). Read depth for every nucleotide in the genome was calculated using the *genomecov* command in bedtools package v.2.20.1, using the alignment file from NGMLR as input. This read depth file was then used for estimating median coverage over 1000 bp blocks using an R script published in Stein *et al*. (2017). We then calculated median read coverage and standard deviation separately for every *B. napus* chromosome. Genomic segments with coverage higher than 1 and standard deviation above the median depth of the chromosome were defined as segmental duplications or CNV.

Calculation of overlap between SV events and the gene models

Quality-filtered SV events were overlapped with the gene models from Darmor-*bzh* and also to the combined *B. rapa* and *B. oleracea* reference assemblies using bedtools intersect (Quinlan and Hall, 2009) using the default parameters. In order to calculate the genome-wide frequency of SV events, we also overlapped the quality-filtered SV with a bed file containing 1 Mbp windows for the entire genome assembly. The intersect file between the SV events and 1 Mbp windows for the entire genome assembly was then used for plotting the SV distribution along 19 *B. napus* chromosomes, using Circos (Krzywinski *et al*., 2009). Statistics including length and distribution of quality-filtered SV from the 12 genotypes were calculated with SURVIVOR (Jeffares *et al*., 2017) and plotted with ggplot2 (Wickham, 2016).

Construction of a maximum-likelihood (ML) tree

SV events predicted for each of the 12 genotypes were merged into a single VCF. This combined VCF was then used to force call

all the SV events across all 12 genotypes using Sniffles, resulting in a multi-sample VCF. The multi-sample VCF was then converted into PHYLIP format using an in house bash script and used as an input for IQ-TREE version 1.6.12 (Nguyen *et al*., 2015). The best-fit substitution model for the data was determined by IQ-TREE ModelFinder (Kalyaanamoorthy *et al*., 2017) and used to construct a phylogenetic tree. The tree was then plotted with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Selection of SV events for PCR validation

We looked at two different agronomically interesting traits in order to prioritize the predicted SV events. Firstly, we analyzed the SV events that might contribute to *Verticillium longisporum* (VL) resistance, using a bi-parental double-haploid population derived from a cross between our sequencing panel genotypes Express 617 and R53. Two QTL were defined for VL resistance on chromosome C01 and C05 by Obermeier *et al*. (2013). We mainly focused on C05 QTL, as this was described to be the major genetic control for VL resistance. The genetic map used for identifying C05 QTL was based on SSR (simple sequence repeats) and AFLP (amplified fragment length polymorphism) markers. Therefore, in order to localize the physical position of the QTL on chromosome C05, we anchored the flanking SSR markers (BRMS030_210 and Na12C01_160) to the Darmor-*bzh* version 4.1 assembly and identified a 4.3 Mbp (6 329 426 bp to 10 659 726 bp) region containing 606 genes. Thirty-seven and 45 out of the 606 genes were found to contain SV in the form of insertions or deletions in Express 617 and R53, respectively. A total of 17 genes were found to be common among both the genotypes, so were dropped from the prioritized gene set. We further prioritized the candidate genes, if they were annotated as defence response or phenolpropanoid pathway genes. Secondly, we analyzed the SV located within the genes described to be involved in flowering-time pathway in *B. napus* as described by Schiessl *et al*. (2017). Top prioritized SV were then visualized in IGV viewer (Robinson *et al*., 2017) and selected for PCR validation.

Acknowledgements

RS was supported by DFG grant SN14/22-1 and BMBF grant 031B0890A. The authors acknowledge Stavros Tzigos and Andreas Welke (Justus Liebig University, Giessen) for technical assistance in the laboratory and greenhouse. Informatics infrastructure was provided by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics (de.NBI).

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

HSC, HTL and RJS conceived the study. HSC, STNA and IAPP generated the Oxford Nanopore long-read sequence data. JS, KL and LG contributed PacBio long-read sequence data. SVS contributed Illumina sequence-capture data. HSC, STNA and HTL conducted the experiments and analysed the data. PV performed SV validation. IG, CO, RJS and HTL provided ideas and suggestions for data analysis. HSC and RS drafted the manuscript.

Data availability statement

Raw data from all 4 ONT libraries have been deposited to the NCBI short-read archive under Bio project number PRJNA642096. The PacBio data from Song *et al*. (2020) are available under PRJNA546246. All variants detected in this study are available as a Supplementary Dataset at https://doi.org/10.5281/zenodo.3931391

References

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., Genete, M. *et al*. (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* **4**, 879–887.

Bus, A., Körber, N., Snowdon, R.J. and Stich, B. (2011) Patterns of molecular variation in a species-wide germplasm set of *Brassica napus*. *Theor. Appl. Genet.* **123**, 1413–1423.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A.P., Tang, H., Wang, X., Chiquet, J. *et al*. (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**, 950–953.

Chen, L., Dong, F., Cai, J., Xin, Q., Fang, C., Liu, L., Wan, L. *et al*. (2018) A 2.833-kb insertion in BnFLC.A2 and its homeologous exchange with BnFLC.C2 during breeding selection generated early-flowering rapeseed. *Mol. Plant* **11**, 222–225.

Gabur, I., Chawla, H.S., Liu, X., Kumar, V., Faure, S., von Tiedemann, A., Jestin, C. *et al*. (2018) Finding invisible quantitative trait loci with missing data. *Plant Biotechnol. J.* **16**, 2102–2112.

Gabur, I., Chawla, H.S., Snowdon, R.J. and Parkin, I.A.P. (2019) Connecting genome structural variation with complex traits in crop plants. *Theor. Appl. Genet.* **132**, 733–750.

Gabur, I., Chawla, H.S., Lopisso, D.T., von Tiedemann, A., Snowdon, R.J. and Obermeier, C. (2020) Gene presence-absence variation associates with quantitative Verticillium longisporum disease resistance in *Brassica napus*. *Sci. Rep.* **10**, 4131.

Grandke, F., Snowdon, R. and Samans, B. (2016) gsrc: an R package for genome structure rearrangement calling. *Bioinformatics* **33**, 545–546.

He, Z., Wang, L., Harper, A.L., Havlickova, L., Pradhan, A.K., Parkin, I.A.P. and Bancroft, I. (2017) Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. *Plant Biotechnol. J.* **15**, 594–604.

Hurgobin, B., Golicz, A.A., Bayer, P.E., Chan, C.-K.K., Tirnaz, S., Dolatabadian, A., Schiessl, S.V. *et al*. (2018) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnol. J.* **16**, 1265–1274.

Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F. *et al*. (2017) Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. and Jermiin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. *et al*. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.

Lee, H., Chawla, H.S., Obermeier, C., Dreyer, F., Abbadi, A. and Snowdon, R. (2020) Chromosome-scale assembly of winter oilseed rape *Brassica napus*. *Front. Plant Sci.* **11**. 496. https://www.frontiersin.org/articles/10.3389/fpls.2020.00496/full

Leflon, M., Eber, F., Letanneur, J.C., Chelysheva, L., Coriton, O., Huteau, V., Ryder, C.D. *et al*. (2006) Pairing and recombination at meiosis of Brassica rapa (AA) x *Brassica napus* (AACC) hybrids. *Theor. Appl. Genet.* **113**, 1467–1480.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Li, Y., Im Kim, J., Pysh, L. and Chapple, C. (2015) Four isoforms of Arabidopsis 4-Coumarate:CoA Ligase have overlapping yet distinct roles in phenylpropanoid metabolism. *Plant Physiol.* **169**, 2409–2421.

Liu, L., Stein, A., Wittkop, B., Sarvari, P., Li, J., Yan, X., Dreyer, F. *et al*. (2012) A knockout mutation in the lignin biosynthesis gene CCR1 explains a major QTL for acid detergent lignin content in *Brassica napus* seeds. *Theor. Appl. Genet.* **124**, 1573–1586.

Lu, K., Wei, L., Li, X., Wang, Y., Wu, J., Liu, M., Zhang, C. *et al*. (2019) Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. *Nat. Commun.* **10**, 1154.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C. and Sedlazeck, F.J. (2019) Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246.

Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., Langlade, N. *et al*. (2016) Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques* **61**, 203–205.

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

Nicolas, S.D., Le Mignon, G., Eber, F., Coriton, O., Monod, H., Clouet, V., Huteau, V. *et al*. (2007) Homeologous recombination plays a major role in chromosome rearrangements that occur during meiosis of *Brassica napus* haploids. *Genetics* **175**, 487–503.

Nicolas, S.D., Leflon, M., Liu, Z., Eber, F., Chelysheva, L., Coriton, O., Chèvre, A.M. *et al*. (2008) Chromosome 'speed dating' during meiosis of polyploid Brassica hybrids and haploids. *Cytogenet. Genome. Res.* **120**, 331–338.

Obermeier, C., Hossain, M.A., Snowdon, R., Knüfer, J., von Tiedemann, A. and Friedt, W. (2013) Genetic analysis of phenylpropanoid metabolites associated with resistance against Verticillium longisporum in *Brassica napus*. *Mol. Breed.* **31**, 347–361.

Qian, L., Voss-Fels, K., Cui, Y., Jan, H.U., Samans, B., Obermeier, C., Qian, W. *et al*. (2016) Deletion of a stay-green gene associates with adaptive selection in *Brassica napus*. *Mol. Plant* **9**, 1559–1569.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.

Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A. and Mesirov, J.P. (2017) Variant review with the integrative genomics viewer. *Cancer Res.* **77**, e31–e34.

Samans, B., Chalhoub, B. and Snowdon, R.J. (2017) Surviving a Genome Collision: Genomic Signatures of Allopolyploidization in the Recent Crop Species Brassica napus. *The Plant Genome*, **10** (3). http://dx.doi.org/10.3835/plantgenome2017.02.0013

Schiessl, S., Samans, B., Hüttel, B., Reinhard, R. and Snowdon, R.J. (2014) Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species *Brassica napus*. *Front. Plant Sci.* **5**, 404.

Schiessl, S., Huettel, B., Kuehn, D., Reinhardt, R. and Snowdon, R.J. (2017) Targeted deep sequencing of flowering regulators in Brassica napus reveals extensive copy number variation. *Scientific Data*, **4** (1). http://dx.doi.org/10.1038/sdata.2017.13

Schiessl, S.-V., Katche, E., Ihien, E., Chawla, H.S. and Mason, A.S. (2019) The role of genomic structural variation in the genetic improvement of polyploid crops. *Crop J.* **7**, 127–140.

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A. and Schatz, M.C. (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468.

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D. *et al*. (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* **6**, 34–45.

Stein, A., Coriton, O., Rousseau-Gueutin, M., Samans, B., Schiessl, S.V., Obermeier, C., Parkin, I.A.P. *et al*. (2017) Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in *Brassica napus*. *Plant Biotechnol. J.* **15**, 1478–1489.

Szadkowski, E., Eber, F., Huteau, V., Lodé, M., Huneau, C., Belcram, H., Coriton, O. *et al*. (2010) The first meiosis of resynthesized *Brassica napus*, a genome blender. *New Phytol.* **186**, 102–112.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer International Publishing.

Wu, D., Liang, Z., Yan, T., Xu, Y., Xuan, L., Tang, J., Zhou, G. *et al*. (2019) Whole-genome resequencing of a worldwide collection of rapeseed accessions reveals the genetic basis of ecotype divergence. *Mol. Plant.* **12**, 30–43.

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D. *et al*. (2019) The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979.

Supporting information

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

Table S1 Read statistics for raw reads.
Table S2 SV Length distribution compared to *B. napus* 4.1 reference assembly.
Table S3 SV Length distribution compared to *B. rapa* plus *B. oleracea* pseudoassembly.
Table S4 Average number of SV per megabase along each *B. napus* chromosome (compared to *B. napus* 4.1 reference assembly).
Table S5 Average number of SV per megabase along each *B. napus* chromosome (compared to *B. rapa* plus *B. oleracea* pseudo assembly).
Table S6 Average number of deletions per megabase along each *B. napus* chromosomes (compared to *B. napus* 4.1 reference assembly).
Table S7 Average number of insertions per megabase along each *B. napus* chromosome (compared to *B. napus* 4.1 reference assembly).
Table S8 Number of CNV in four representative *B. napus* accessions.
Table S9 CNV calling for Express 617, R53, N99 and PAK85912.
Table S10 Validation of 1.3 Kbp deletion in the promoter region of *BnFT.A02* (BnaA02g12130D) in ERANET-ASSYST consortium diversity set.
Table S11 Distribution of SV in exons and introns.
Table S12 PCR primer sequences used for validation of SV events.

5 Discussion

SNPs have always been regarded as the major source of selectable variation in plants, however it is becoming increasingly evident that SV are a significant source of genetic variation. For example, chapter 3 and 4 of this thesis identified SV underlining disease resistance and eco-geographical adaptation in *B. napus.* Additionally, SV have also been associated with several other agronomically important traits such as plant height (Li et al. 2012), heading date (Nishida et al. 2013), root architecture (Uga et al. 2013). Zhou et al. (2019) found genomic re-arrangements were the causal for at least one-third of all the known domestication alleles in crop species. During the last decade there has been a surge in the number of studies associating SV with key phenotypic traits in plants. This increase can largely be attributed to the advent of next generation sequencing technologies. One such example combining Illumina sequencing with the widely used SNP arrays and ultra-long restriction maps from optical mapping to discover large-scale SV underlining quantitative disease resistance in oilseed rape has been presented in chapter 3 of this thesis. A majority of studies aimed at the genome wide identification of SV in plants have relied on Illumina sequencing. However, Mahmoud et al. (2019) reported a very high false positive rate (up to 89%) for SV detection with Illumina data. Furthermore, it is challenging to unambiguously align short Illumina reads to a reference assembly for a polyploid genome due to the high levels of sequence similarity among the homeologous regions (Schrinner et al. 2020). Although, the large SV might be detected using a high sequencing coverage Illumina data by analysing the density of sequenced reads aligned to a locus in a reference genome (Gabur et al. 2019), such a read depth based approach would not provide the necessary resolution to detect small to mid-scale genome re-arrangements. Therefore, there have been little or no efforts towards cataloguing small to mid-scale SV and their impact on evolution and adaptation related traits. The study in chapter 4 was specifically aimed at detecting the extent of small to mid-scale SV in the rapeseed genome using the long reads from the third generation of sequencing technologies. Both PacBio and ONT have been widely used for genome assembly and SV detection for various plant species such as bread wheat (Zimin et al. 2017), maize (Yang et al. 2019), rice (Zhang et al. 2016), Arabidopsis (Michael et al. 2018), tomato (Schmidt et al. 2017) and banana (Belser et al. 2018). Despite the long read length, there are several challenges associated with the third generation sequencing approaches such as the complexity involved in the isolation of high molecular weight (HMW) DNA, elevated error rates and the high costs associated with handling the huge amounts of data. These challenges need to be addressed before the wide scale adoption of the long-read sequencing for studying the SV landscape for a large number of plant species.

5.1 Challenges involved with long-read sequencing

Read length is the biggest factor separating the third-generation sequencing technologies from their predecessors. Ultra-clean HMW DNA is critical to attain an optimal read length from the third-generation sequencing platforms. Although extracting large quantities of HMW DNA from mammalian cells is relatively simple, it could be extremely challenging for certain plant species. This difficulty to isolate HMW DNA could be possibly attributed to the presence of cell wall and high quantities of secondary metabolites in a plant cell. Secondary metabolites comprise a family of compounds that are not critical for the growth and development of a plant but are crucial in conferring eco-geographical adaptability. These compounds not only interfere with DNA isolation but can also clog the nanopores on an ONT flowcell, thereby decreasing the sequencing yield. Several DNA isolation protocols such as nuclei isolation (Sikorskaite et al. 2013), the standard Cetyl trimethylammonium bromide (CTAB) method (Doyle 1991) and the commercially available DNeasy Plant Mini Kit by Qiagen have been used to obtain HMW DNA for long read sequencing of plant species. While some of the methods, such as CTAB or the DNeasy Plant Mini Kit, offer a good quality of DNA in terms of A260/280 and A260/230 ratios, but can introduce shearing, thereby lowering the read lengths. Other methods based on nuclei isolation offer an excellent read length distribution but are often challenging to implement. Therefore, this thesis establishes a protocol based on Mayjonade et al. (2016) for isolating HMW, ultra-pure DNA from oilseed rape (chapter 4). Several protocols for the removal of short DNA fragments have been described in the literature to improve the read length of ONT sequencing. Long DNA fragments can be enriched in a sequencing library by fine tuning the ratio of NaCl to polyethylene glycol (PEG) in the magnetic bead solution used to remove the residual ligation enzymes and other unbound sequencing adapters during ONT library prep (Dumschott et al. 2020). Pulsed-field electrophoresis based system such as BluePippin (Sage Scientific, Beverly, MA, USA) could also be used to eradicate short DNA fragments from a sequencing library (Dumschott et al. 2020). A BluePippin can enrich DNA fragments between 100bp to 50kb, but the cost for the equipment is around 20,000 EUR along with the running costs for the DNA gel cassettes. Short Read Eliminator (SRE) kits from Circulomics Inc. Baltimore, MD, USA are a cheaper and efficient alternate to BluePippin. These SRE kits can remove small DNA fragments by selectively precipitating HMW DNA for less than 10 EUR per sample. SRE kits can effectively enrich for fragment lengths greater than 5kb or 25 kb depending on the aim of the experiment.

Another major challenge with ONT is the astronomical amounts of raw data produced during a sequencing run. A single run on the ONT MinION system produces between 5 and 15 current pulses per base which are then converted to a 16-bit integer stored in a specific class of the HDF5 format known as fast5. In an uncompressed state the current disruptions from every base recorded by the sequencer requires 18 bytes of hard drive space (Chandak et al. 2020). Therefore, 180 GB of raw data would be written to a storage medium in a standard MinION run (of around 10 Gb) whereas 1.2 TB of data would be generated by a single flowcell on a PromethION (assuming 70 Gb from a flowcell). However, a PromethION system is capable of running 48 flowcells in parallel and at its full capacity it would generate 350 MB of data every second continuously for 48 hours (the duration of a standard sequencing run). To put it in perspective, the Large Hadron Collider (LHC) at CERN produces 25 GB of data per second (https://home.cern/science/computing/processing-what-record). Although the LHC produces 100 times more data than a single PromethION (per second) but there are definitely more than 100 PromethION systems present in the genome sequencing centers around the world. All of the above calculations do not even include the additional storage space required for writing methylation signals associated with each basecalled nucleotide. If methylation signals would be recorded in the raw fast5 files, it would just require 26 PromethION systems to outcompete the LHC in terms of raw the data produced per second. For large scale projects major upgrades to the storage (in terms of vast amounts of solid-state hard drives) and the network capabilities (fast interface network switches) of the IT infrastructure would be indispensable. Compression algorithms such as Picopore (Gigante 2017), VBZ (VBZ Compression.https://github.com/nanoporetech/vbz_compression/) or lossy methods have been developed to reduce the storage resources for ONT data. These algorithms can reduce the size of the raw ONT data by 60 percent.

5.2 Towards understanding gene-scale SV

Although the next generation sequencing technologies have enabled the identification of SV in complex plant genomes not much is known about the cellular mechanisms generating these genomic re-arrangements. Several molecular mechanisms associated with DNA recombination, replication and repair have been put forward to explain the origin of SV in the plant genomes. Genomic re-arrangements can emerge by recombination errors, such as non-allelic homologous recombination (NAHR) (Lupski 1998), DNA break repair errors, like non-homologous end

joining (NHEJ) (Moore and Haber 1996), replication errors, comprising of fork stalling and template switching (FoSTeS) (Lee et al. 2007), microhomology-mediated break-induced replication (MMBIR) (Hastings et al. 2009) or Transposable elements (TE) activity (Schiessl et al. 2019). In polyploids, intra-genomic exchange of genetic material known as homeologous exchange (HE) is one of the major sources of SV (Schiessl et al. 2019). HE are the outcome of various cellular phenomenon such as crossovers (CO) and non-crossovers (NCO or gene conversions). Both CO and NCO originate via double-strand breaks (DSBs) in the DNA. Samans et al. (2017) reported a strong enrichment for HE driven large segmental deletions in the C subgenome of *B. napus*. In contrast, no such enrichment was observed for the small to mid-scale deletions in the C-subgenome of the 12 *B. napus* accessions sequenced in the study in chapter 4. The lack of enrichment for small to mid-scale deletions in the C subgenome indicates that HE might not be the causal for generating small to mid-scale genomic re-arrangements in *B. napus*. A different molecular mechanism such as TE activity might be responsible for the generation of this size range of SV events.

Transposable elements (TE) are a major cause for the genomic rearrangements in plant genomes. (Schiessl et al. 2019). Jiang et al. (2004) found nearly 3000 *Mutator*-like transposable elements (MULEs) in the rice genome. Pack-MULEs, MULEs with genic fragments from the host were found to be responsible for capturing and reshuffling of approximately 1,000 gene fragments in rice. While comparing the sequences from the cellular genes and their Pack-MULE counterparts the authors could identify that fragments of genomic DNA have been captured, re-arranged and amplified over millions of years. Since MULEs are present in many plant genomes such as rice (Jiang et al. 2004), maize (Zhao and Jiang 2014) and Arabidopsis (Yu et al. 2000), they could also generate small to mid-scale SV in the rapeseed genome. *Ac/Ds* (Activator/Dissociation) transposable elements could also explain the creation of small to mid-scale SV in the plant genomes. *Ds* can introduce genomic rearrangements via a chromosomal break at its insertion site and could move in response to another self-mobilizing or autonomous factor, *Ac* (*Activator*) (Du et al. 2011). Xuan et al. (2011) studied the SV inducing impact of *Ac/Ds* transposable elements in the rice genome. The authors found genomic re-arrangements such as insertions, deletions and inversion ranging from 184 bp to 520 kb among the 300 offspring generated from a single T-DNA insertion line housing *Ac* and *Ds* elements. Even after losing the ability to mobilize, TEs can still promote SV by providing localized regions of microhomology, necessary for template switching during the repair of replication errors by fork stalling and template switching

mechanism (Lee et al. 2007). In addition to TEs, high sequence similarity among the homeologous regions in a polyploid genomes could provide these micro-homologous sites and might result in shuffling of DNA fragments among the subgenomes.

Some of the small to mid-scale SV observed in the rapeseed genome (Chapter 4) might actually be GCs. Although the exact mechanism of GC in plants is unknown but in yeast it occurs by synthesis-dependent strand-annealing (SDSA) (Chen et al. 2007). In yeast, a DSB in the DNA is fixed by the SDSA mechanism using homologous region of the genome as a template for the repair. On contrast, in a polyploid genome homeologous regions (instead of homologous) might be used as a template for repairing the breaks in DNA, thereby translocating small genomic fragments between highly similar sequences of the subgenomes.

5.3 Targeted sequencing using long reads

With the current costs of long-read sequencing it is not feasible to sequence whole genomes for a large number of individuals to study the SV landscape of a species. This is especially true for many commercially important plant species with huge genome size such as bread wheat (16 Gb), barley (5.3 Gb), maize (2.5 Gb) or onion (16 Gb). Furthermore, it might not be even necessary to perform whole-genome sequencing in certain scenarios such as for cloning the gene(s) or the genomic region(s) underlining important phenotypic traits in a large number of genotypes. Targeted sequencing provides an economically viable alternative to whole-genome sequencing. Most widely used targeted sequencing strategy involves capturing of genomic regions of interest by hybridizing them to biotin labelled oligonucleotide baits followed by Illumina sequencing. One of the major drawbacks of Illumina based sequence capture is that the sequence of baits is based on a single reference genome assembly. Therefore, the sequencing reads originating from a genetically distant genotype with novel structural variants would not align to the reference genome due to the short read length of Illumina sequencing. Ebbert et al. (2019) identified 527 gene bodies with a few or no mappable Illumina reads in the whole-exome sequencing data from 10 human genomes. The authors detected 36,794 “dark” regions with no or zero aligned reads within the protein-coding exons across 748 genes using the Illumina data. However, the dark protein-coding genes were reduced to 35.6% and 9.6% using the PacBio and ONT reads respectively. Therefore, target enrichment methodologies using the long reads from the PacBio and ONT have been developed to overcome the problem of short read length of Illumina sequencing. Grech-Baran et al. (2020) were able to assemble multiple TIR-NLR-encoding

paralogues co-segregating with *Potato virus Y* resistance in 150 diploid potato lines from a mapping population. The authors were able to enriched for several TIR-NLR-encoding paralogues using oligonucleotide bates followed by long read sequencing with a PacBio RS II system. However, their protocol still involved PCR amplification which not only limits the length of the sequencing reads but also introduces amplification bias. Furthermore, an amplification step would result in the loss of the methylation signatures associated with the DNA bases. Amplification free sequence capture methods eliminate the need for PCR by direct enrichment of the target region using CRISPR-Cas9 system instead of the oligonucleotide baits. Tsai et al. (2017) were able to enrich for the causative loci for Fragile X syndrome, Huntington's disease, frontotemporal dementia, spinocerebellar ataxia type 10 and amyotrophic lateral sclerosis (ALS) by combining CRISPR-Cas9 enrichment with long PacBio reads using a protocol described in Figure 1A. Another amplification free target enrichment method based on ONT known as nCATS (Gilpatrick et al. 2020) involves removing all the free 3 prime ends of the DNA by dephosphorylation. The dephosphorylated DNA is then subjected to a Cas9 cleavage to excise the target region. It is followed by preferential ligation of sequencing adaptors to the newly produced phosphorylated DNA ends at Cas9 cleavage sites (Figure 1B). nCATS was proven to be useful for enriching fragments up to 140 kb in length. A very recent improvement to nCATS method known as Affinity-based Cas9 Mediated Enrichment (Figure 1C) or ACME (Iyer 2019) also involves dephosphorylation of DNA followed by a Cas9 cleavage of the target region. ACME protocol includes an additional cleanup step to eliminate non-target Cas9 bound DNA molecules using His-Tag beads. His-Tag beads selectively bind to the 6 Histidine tag of Cas9 protein (added during its purification) to precipitate out any non-target nucleic acid bound with it. All the above-mentioned target enrichment methods are suitable for capturing a low number of genomic loci and would not be useful for a whole exome sequencing experiment as it would be both time consuming and expensive to design thousands of CRISPR constructs.

Computational approaches such as UNCALLED (Kovaka et al. 2020), Read Until (Payne et al. 2020) have been specifically developed for the applications aimed at the whole exome capture using ONT sequencing. Both the UNCALLED and Read Until approaches rely on the ability of a

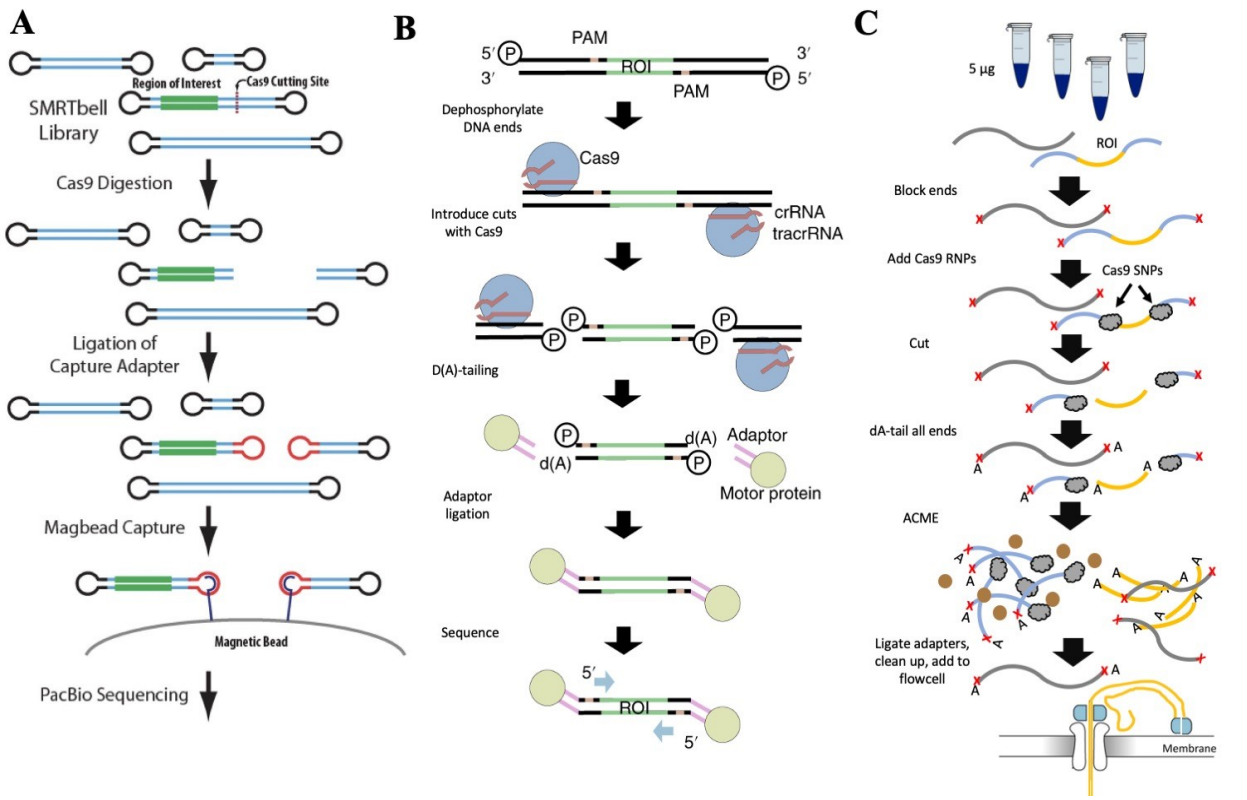


Figure 1: Different target enrichment protocols using third generation long sequencing reads. **A.** Cas9 based target enrichment using PacBio sequencing. A PacBio SMRTbell library is prepared with a CRISPR restriction site adjacent to the region of interest (green). This sequencing library is digested with Cas9 to create free 3' ends to ligate a capture adapter. SMRTbell molecules with a capture adapter are captured using magnetic beads and sequenced on a suitable PacBio platform. Figure adapted from Tsai et al. (2017). **B.** In the nCATS workflow, the DNA is subjected to end dephosphorylation followed by digestion with a Cas9/guideRNA complex. Nanopore sequencing adaptors can then be ligated to newly formed phosphorylated ends from the Cas9 digestion next to the ROI (region of interest), followed by loading of the sample on to a nanopore sequencer. Figure adapted from Gilpatrick et al. (2020) **C.** ACME protocol is similar to nCATS however, after the Cas9 restriction of ROI the background non-target Cas9 bound DNA fragments are further depleted by additional cleaning with His-Tag beads. Figure adapted from Iyer (2019).

nanopore sequencer to recognize and eject a non-target DNA molecule while it is still being sequenced. Selective ejection of non-target molecules ensures that the nanopores on a flowcell are reading the DNA fragments housing the target sequences for the majority of the time during a sequencing run (Figure 2). Therefore increasing the coverage for the target molecules. Payne et al. (2020) enriched 25,600 exon targets from 10,000 human genes using the "Read Until" method. Kovaka et al. (2020) used UNCALLED to enrich 148 human genes associated with hereditary cancers to 29.6x coverage using a single MinION flowcell. Using UNCALLED the authors were able to detect double the number of SV compared to 50x Illumina data. These computational based enrichment approaches do not require any special type of library preparation, which is often the most expensive and time-consuming part of sequence capture with

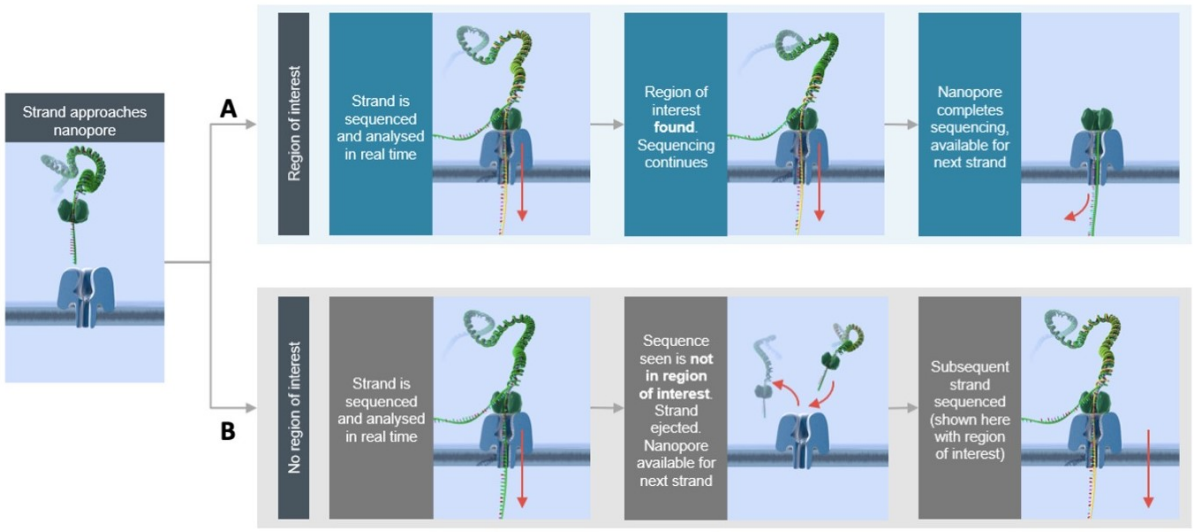


Figure 2: A flow diagram depicting ONT "Read Until" utility. **A.** Scenario when the region of interest is found, and the sequencing is continued until the entire molecule is read. **B.** Scenario when the region of interest is not found and the DNA strand is ejected immediately and the nanopore is available again to sequence another strand.(downloaded from https://nanoporetech.com/about-us/news/nanopore-digest-18th-february-2020)

Illumina reads. Therefore, computational enrichment would be best suited for large plant genomes with huge number of genes. For example, a single MinION run for 600 EUR (including flowcell cost, library prep and DNA isolation) would be sufficient to generate approximately 50x coverage (assuming per flowcell yield of 10 Gb and exome size of ~200 Mb for rapeseed) of an entire *B. napus* exome comprising of 101,040 genes. Furthermore, the length of the sequencing reads from exome capture using "Read Until" method could be theoretically the same as the actual length of the exons in the genome. ONT based enrichment methods would also allow simultaneous capturing of epigenetic signatures on the DNA molecules along with their nucleotide readouts.

5.4 Long read sequencing for identifying epigenetic signatures

All the cells in a plant are genetically similar; however, for performing the vital functions to survive under different types of environmental conditions the plant requires several specialized cell types and tissues. Such a cell specialization is achieved by fluctuations in gene expression patterns in different cell types that are controlled by epigenetic mechanisms. (Cytrynbaum et al. 2019). Epigenetics comprises modifications to DNA (such as DNA methylation), chromatin, and associated molecules that do not change the primary sequence of the DNA (Berger et al. 2009). Illumina based methods such as bisulphite sequencing and Hi-C have been widely used for studying these epigenetic marks in plants and humans (Liu and Weigel 2015). However, using

the third-generation sequencing technologies it is possible to detect these epigenetic signatures with greater resolution and reduced costs.

5.4.1 Chromatin conformation capture with ONT

Genomes of all the eukaryotic organisms are organized into characteristic folding patterns such as loops and domains inside the nucleus of a cell. Regions of high chromatin inter-connectivity known as topologically associated domains; TADs are one of the characteristic features of this 3D arrangement. Contact maps from high-throughput chromosome conformation capture (Hi-C) have aided in the detection of these spatial features. Golicz et al. (2020) showed TADs exhibit an elevated SNP and SV density along with an increased recombination rate compared to inter-TAD regions in rice. However, due to the small read length of Illumina reads the Hi-C approach might only be able to detect simple pairwise interactions between two genomic regions. In contrast, several kb long ONT reads could span multiple contact points within the genome, thereby enabling the identification of multi-way, higher-order information. Pore-C is a recently developed method that couples chromatin conformation capture with ONT sequencing (Ulahannan et al. 2019). Ulahannan et al. (2019) found Pore-C data to be in full agreement with gold-standard Hi-C pairwise contact maps at the compartment, topologically-associated domain, and loop levels. Furthermore, Pore-C does not involve PCR, thereby reducing sequencing bias and enabling dissection of a genomic region with high or low GC-content. Another possible application of Hi-C or Pore-C contact maps could be to identify complex SV events such as translocation or inversion duplications. Low mappability around the breakpoints of such complex genomic re-arrangements limits the applicability of mapping-based approaches for SV detections. Hi-C contact maps have provided insights into these complex genomic rearrangements. However, the short-read length of Illumina limits the resolution of Hi-C for resolving allelic structure of complex SV. In contrast, the long-range information encoded in Pore-C reads could be used to reconstruct complex genomic rearrangements spanning multiple Mb across different chromosomes. Using Pore-C, the Ulahannan et al. (2019) identified a rearrangement junction connecting three genomic regions on chromosomes 9, 12, and 20 in a breast cancer cell line.

5.4.2 DNA modifications detection using third-generation sequencing technologies

In plants, DNA methylation can occur on the 5-prime position of a cytosine residue in various contexts such as CG, CHG and CHH where H represents A, T or C (Lister et al. 2008, Zhang et

al. 2006). There has been no study linking SV to DNA methylation in any plant species. It could be majorly due to the technological and cost constraints for studying methylation signatures for large and complex plant genomes. Illumina bisulphite sequencing has been the most widely used method for methylation detection. However, it requires special library preparation and can deliver a maximum read length of 300 bp. In contrast, the third-generation of sequencing technologies are capable of delivering the long-range epigenetic information at no extra costs together with a regular sequencing run. PacBio platforms can identify methylated nucleotides by recording the shift in the speed of nucleotide incorporation by the DNA polymerase (Kraft and Kurth 2019). In addition to the emission spectra, the PacBio systems also record the duration (pulse width) and interval between successive fluorescence pulses (interpulse duration, IPD) created during incorporation of a specific nucleotide. Different modifications to a nucleotide can introduce shifts in the pulse width and IPD for example Flusberg et al. (2010) found that the IPD for a methylated Adenine was five times larger than its canonical form. However, a very high coverage, 250x per strand is required for a reliable detection of 5mC with PacBio sequencing (Liu et al. 2020). Furthermore, the accuracy of methylation calling using PacBio is primarily influenced by the sequence coverage (Zhu et al. 2018). Therefore, it would not be economically viable to use PacBio for methylation detection. ONT sequencing has also been used widely for studying DNA modifications. The methylated bases would produce a different current perturbation as they pass through a nanopore compared to their canonical form. ONT sequencing can reliably detect various types of DNA methylation such as N6-methyladenine (6 mA) and 5-methylcytosine (5mC) (McIntyre et al. 2019). Perumal et al. (2020) used ONT for CG methylation profiling of *Brassica nigra*. They reported a 93 to 97% correlation between methylation calls from ONT and the bisulphite sequencing. Zhang et al. (2019) suggested that the DSB repair mechanisms responsible for generating genomic re-arrangements can also result alterations of CpG methylation. Sun et al. (2018) provided another example linking SV to DNA methylation. The authors suggested that CNV often affect DNA methylation and gene expression in tumor samples. Therefore, it might be important to study DNA methylation to completely understand the cellular mechanism driving genomic re-arrangements. Recent improvements in ONT basecalling algorithms can enable accurate identification of DNA methylation together with a regular sequencing run at no extra costs. This wealth of knowledge would definitely be instrumental in providing a clear picture about cellular processes creating SV.

5.5 Limitations of the third-generation sequencing technologies

Although, the third-generation sequencing technologies have provided insights into genomic re-arrangements by providing additional layers of information, such as methylation signatures, chromatin confirmation, there are still some limitations associated with these long read sequencing platforms. High error rate (up to 20 percent) and high cost are the two major limitations of the third-generation sequencing technologies. Both PacBio and ONT have a significantly higher error rate compared to Illumina sequencing. Accuracy of a sequencing technology can be calculated at the read level (raw read accuracy) or at a consensus level (consensus accuracy). Raw read accuracy refers to the sequence identity of an individual read when compared to a reliable reference assembly. In contrast, consensus accuracy is the correctness of a consensus sequence assembled by overlapping multiple reads from a single genomic locus. In most cases, the consensus accuracy increases with an increase in read depth for example a consensus sequence assembled from 50 reads is likely to be more accurate than a consensus of 5 reads. However, this might not be the case, especially if the raw reads contain non-random errors such as homopolymer deletions as in the case of ONT. Consensus accuracy is critical for high coverage applications such as *de novo* genome assembly whereas for the low read depths scenarios raw read accuracy could be more meaningful. For example, a clinical metagenomics sample would comprise majorly of non-target sequencing reads from the host and a very few reads from the organisms of interest in the microbiome. Low raw read accuracies would therefore make it very challenging to identify and characterize these organisms in the microbiome (Wick et al. 2019). Both the raw read and consensus accuracy is expressed in terms of Q score. Ewing and Green (1998) defined Q score as a value logarithmically related to the base calling error probabilities ($Q= -10 \log_{10}P$). Sanger sequencing has raw read accuracy of ~99.4%, or ~Q20, whereas Illumina sequencing can produce reads that are 99.99% accurate (~Q30). PacBio reads in their raw form are already a consensus derived from repeated sequencing of a circularized DNA molecule, therefore it would be pointless to calculate raw read accuracies for this type of reads. The consensus accuracy of the Hifi reads from latest PacBio sequel II system has been reported to be at 99.8 % (Wenger et al. 2019). In contrast, the raw read accuracies for ONT is around 85 to 95% depending upon the flowcell chemistry and the basecaller version (Zhang et al. 2020). A consensus accuracy of Q32.2 (99.94% identity) was reported for ONT, making it comparable to Illumina “raw read accuracy”. 99.94% identity would translate to approximately 3000 errors in a 5 Mb genome. (Wick et al. 2019). However, to

achieve a reliable SNP calling from ONT data the consensus accuracy needs to be two folds, i.e. Q70 (one error per 10 Mb) higher than the current levels. ONT data is inherently noisier compared to Illumina sequencing due to the complexity involved in the translation of current signals captured by ONT sequencing device into nucleotide information. The strength of the electrical signal depends on the resistance to flow of current from nucleotides present within the narrowest point of a nanopore. Five nucleotides can reside within the narrowest point of an R9.4 pore (the most widely used nanopore from ONT) at any given point of time (Wick et al. 2019). This 5-mer results in a large number of possible states (4^5 =1024) for a basecalling model aware of just the standard four bases of DNA. Instead, if a methylation aware basecalling model is used the number of possible states can get even higher (5^5 =3125). Due to this complexity, basecalling requires sophisticated machine learning algorithms. These algorithms are computationally demanding and the data from a single flowcell could take weeks to basecall. Performing basecalling on a standard CPU based computer for a MinION (assuming 10 Gb from a flowcell) run would take more than a day to basecall. In contrast, a PromethION (assuming 70 Gb from a flowcell) run would require a week of basecalling. The newer generation of basecalling algorithms such as guppy can leverage graphical processing units, GPUs to accelerate base calling by 10 to 15 times (up to 1,500,000 bp/s) compared to older generation basecallers (120,000 bp/s) (Wick et al. 2019).

Long-read sequencing is still many folds expensive then Illumina sequencing. Moreover, a greater depth of coverage is required to overcome the higher error rate of the third generation sequencing technologies, making them even more expensive. For producing 1 Gb of sequencing data with Illumina would cost around 10 EUR whereas for ONT and PacBio it would be approximately 60 to 80 EUR. Furthermore, specialized computer hardware such as GPU, required for processing ONT data could raise the cost per data point even higher.

5.6 Conclusion

During the last decade, there have been tremendous efforts to detect and catalogues SV in several plant genomes. This could be majorly attributed to the advent of high throughput next-generation sequencing methodologies. These methodologies provided enabled the detection of large-scale SV underlining important agronomic traits in several plant genomes. Many SV detection strategies till date include short-read sequencing technologies to genotype a large number of individuals within a population and mapping of these reads to a reference assembly. However, due to the short read length, these reads cannot be anchored uniquely to a reference assembly, thereby offering a minimal resolution to detect small to mid-scale SV, especially for polyploids species. One of the possible solutions is to construct a pan-genome, but *de novo* assembly of multiple individuals with large genome size such as wheat (genome size of approximately 16 Gb) would require an extremely high depth of coverage which would be time-consuming and expensive. Long reads from ONT and PacBio with can provide a viable alternative for reliably detecting a broad size range of SV. Reads from ONT can reach up to 1 Mb (Jain et al. 2018) in length, thereby spanning even the largest of SV and reducing the need for a pan-genome. Furthermore, both ONT and PacBio can also detect methylated bases, without any special library preparation making them even more suitable for studying SV and the epigenetic signatures. However, read mapping alone might not be sufficient to thoroughly understand the most complex types of SV, such as translocation or inversion duplications. Chromatin conformation capture using Pore-C approach could provide insights into such complex type of SV events.

Despite all the advantages, there are still some challenges that need to be addressed before wide-scale adoption of long-read sequencing technologies. One of the biggest challenges, especially for ONT, is to obtain huge quantities of ultra-pure HMW DNA for certain plant species. Till date, there is no straightforward method or commercial kits applicable to a wide range of plant species for isolating HMW DNA. There have been several attempts to develop HMW DNA isolation methods for a particular species such as Arabidopsis (Michael et al. 2018), Eucalyptus (Schalamun et al. 2019), Banana (Belser et al. 2018) etc. These methods can be used as a baseline for developing a working protocol for sequencing the species of interest.

The error rate for the long-read sequencing is still higher than Illumina data. PacBio has solved the problem of high error rate by introducing the Sequel II system. A PacBio Sequel II system is capable of generating highly accurate (more than 99% single-molecule read accuracy) high-

fidelity (HiFi) reads using the circular consensus sequencing (CCS) mode (Lang et al. 2020). This high accuracy comes at the cost of read length as in the CCS mode the Sequel II is capable of delivering read lengths N50 between 10 to 15 kb. On the other hand ONT reads could be as long as 1 Mb with an N50 of over 100 kb but suffer from systematic errors in homopolymeric regions (Guiglielmoni et al. 2020). To overcome these systematic errors, ONT has constantly pushed newer basecalling algorithms together with more sophisticated type of nanopores containing longer barrel and dual reader head. At the time of writing this document, ONT claims a median raw read accuracy of 96% using the latest version of basecaller, Guppy version 3.6.0 with R9.4.1 flowcell (https://community.nanoporetech.com/posts/guppy-v3-6-0-release).

In conclusion, if *de novo* assemblies are desired for SV detection, PacBio assemblies would be more accurate at a base-pair level whereas the ONT assemblies would be more contiguous (Lang et al. 2020). Therefore, a combined strategy involving both Pacbio and ONT would be best suited for reliable detection of SV, but the long reads are significantly more expensive then Illumina sequencing. However, improved DNA isolation protocols along with newer and better flowcells have improved the yield both for PacBio and ONT sequencing. Therefore, it's a matter of time until the price for long-read sequencing would be comparable to the short-read length technologies.

A major question that arises here is whether there is any additional genetic gain by investing high amount of money into SV detection for predictive breeding? Weisweiler et al. (2019) included PAV in their genomic selection models to predict leaf angle, plant height and heading date in barley and found a higher prediction accuracy for all the three traits using a PAV aware model. A similar observation was made when PAV was included for prediction of quantitative disease resistance in oilseed rape (unpublished data). Furthermore, the authors found that PAV were not in strong linkage disequilibrium (LD) with the neighboring SNP markers. A similar scenario was also reported for grapevine (Zhou et al. 2019). Therefore, it will not be entirely false to say that including SV into genomic prediction model can lead to higher prediction accuracies which might translate into higher genetic gains. However, whether this genetic gain is significant enough to invest in it, is entirely a different question and remains to be answered. Till the long reads sequencing becomes less expensive, an alternate approach would be to construct an SV atlas by sequencing a few diverse individuals within a species with long-read platforms. Using this SV catalogue, it might be possible to detect SV using existing short-read data in an approach described by Malmberg et al. (2019).

Longer more accurate reads could improve our ability to identify SV which were nearly invisible to short reads. Although, the exact molecular mechanism underlining small to mid-scale SV in *B. napus* is not well understood. We can gain more insights into the mechanism for their emergence by zooming into the regions housing these SV. Highly re-arranged genomic region might reveal an enrichment of genomic feature such as TEs or TADs. Recognition of other cellular features such as methylation signature and chromatin conformation could further aid in a clear understanding of the molecular mechanism responsible for creating genomic rearrangements and their implications on important phenotypic traits.

6 Summary

The allopolyploid species *B. napus* originated from interspecific hybridization between *B. rapa* and *B. oleracea* about 7500 years ago. Due to this recent polyploidization event, the A and the C subgenomes of oilseed rape share high levels of sequence identity. High homeology among the subgenomes of *B. napus* results in a plethora of structural variations (SV) in the form of InDels, copy number variations (CNV), translocations or inversions. There have been several studies associating agronomically important traits such as disease resistance, flowering time and seed quality to SV in oilseed rape. These studies revealed the importance of SV in the creation of the *de novo* genetic variation necessary for adaptation and breeding. In this thesis, I elucidate different approaches for genome-wide detection and analysis of all size ranges of SV in *B. napus*. For the identification of large-scale SV this dissertation describes an integrated approach combining single nucleotide polymorphism (SNP) arrays, Illumina sequencing and optical mapping using resistance to *Verticillium longisporum* as an example for a quantitatively inherited trait in *B. napus*. A significant increase in the resolution of *Verticillium* resistance quantitative trait loci (QTL) was observed by including the SV in the form of single nucleotide absence polymorphism (SNaP) markers in the genetic map or genome-wide association studies (GWAS) model. Furthermore, presence absence variation (PAV) was observed in 23 to 51% of the genes within the *Verticillium* resistance QTL. Moreover, every high-priority candidate gene for *Verticillium* resistance within the QTL was affected by PAV. The widespread PAV in the rapeseed genome suggested that it is an important class of polymorphism and should be exploited more systematically in plant breeding programs.

A majority of studies (including the one mentioned above) aimed at the genome wide identification of SV in plants have relied on Illumina sequencing. However, up to 89% false positive rate has been reported for SV calling with Illumina data. Furthermore, it is challenging to unambiguously align short Illumina reads to a reference assembly for a polyploid genome due to the high levels of sequence similarity among the homeologous regions. Therefore, there have been little or no efforts towards cataloguing small to mid-scale SV. This thesis describes the use of long sequencing reads to evaluate the role of small to mid-scale SV in eco-geographical diversification of *B. napus* into the three predominant ecotypes (winter-type, spring-type and semi-winter type), and survey their extent and impact on genes. Up to 10% of all genes in the rapeseed genome were found to be affected by small to mid-scale SV events. Nearly half of these SV events ranged between 100 bp to 1000 bp, which makes them challenging to detect using short read Illumina sequencing. Furthermore, small SV were also detected in the genes associated

with *Verticillium* resistance in oilseed rape. This thesis also provides first insight and ideas about how new long-read sequencing technologies can help to understand complex SV in large plant genomes by providing additional layers of information, such as methylation signatures, chromatin confirmation, or data from target enrichment strategies implementing long-read sequencing, and describe potential cellular mechanisms that might explain the occurrence of small to mid-scale SV in oilseed rape. Additionally, the dissertation also reviews the challenges and limitations of the third-generation sequencing technologies.

The key finding from this dissertation was the surprisingly high level of widespread, small to mid-scale SV in the rapeseed genome. This size range of SV is almost invisible to Illumina sequencing and was therefore completely ignored by the earlier studies aimed at detecting genomic re-arrangements in *B. napus*. The results from this dissertation suggest that revisiting complex plant genomes using medium-coverage, long-read sequencing might reveal unexpected levels of functional gene variation, with major implications for trait regulation and crop improvement.

7 Zusammenfassung

Die allopolyploide Art *B. napus* entstand vor etwa 7500 Jahren aus einer interspezifischen Kreuzung zwischen *B. rapa* und *B. oleracea*. Aufgrund dieses relativ jungen Polyploidisierungsereignisses sind die Sequenzen der beiden *B. napus* Subgenome A und C sehr ähnlich. Diese Homöologie zwischen den beiden Subgenomen führt zu einer Vielzahl von SV in Form von Indels, Kopienzahlvariationen, Translokationen oder Inversionen. Es gibt zahlreiche Studien, die den Zusammenhang von strukturellen Genomvariationen (SV) mit wichtigen agronomischen Merkmalen wie Krankheitsresistenz, Blühzeitpunkt und Samenqualität von Raps zeigen. Diese Studien offenbaren die Bedeutung von SV zur Schaffung neuer genetischer Variation, die für Anpassung und Züchtung notwendig ist. In dieser Arbeit erläutere ich verschiedene Ansätze zur genomweiten Detektion und Analyse von SV aller Größenbereiche in *B. napus*.

Zur Identifizerung von großen SV beschreibt diese Dissertation einen integrierten Ansatz, welcher Einzelnukleotid-Polymorphismen (SNP) Arrays, Illumina Sequenzierung und Optical Mapping kombiniert und bei dem die Resistenz gegenüber *Verticillium longisporum* als Beispiel für ein quantitativ vererbtes Merkmal in *B. napus* dient. Die Einbeziehung von SV in Form von SNaP (Einzelnukleotid-Absenz-Polymorphismus)-Markern in die genetische Karte oder in das Modell zur genomweiten Assoziationskartierung (GWAS) führte zu einer signifikant gesteigerten Auflösung der *Verticillium*-Resistenz-QTL (Quantitative Merkmalslokus). Überdies wurden in 23 – 51% der Gene innerhalb der Verticillium-Resistenz-QTL PAV (presence absence variation) gefunden. Vielmehr war jedes Kandidatengen für *Verticillium*-Resistenz innerhalb der QTL von PAV betroffen. Die weite Verbreitung von PAV im Rapsgenom deutet darauf hin, dass es sich um eine wichtige Art von Polymophismen handelt, die in Pflanzenzüchtungsprogrammen systematisch genutzt werden sollte.

Die Mehrheit der Studien (einschließlich der oben erwähnten), mit dem Ziel der genomweiten Identifizierung von SV in Pflanzen, basieren auf Illumina-Sequenzierungen. Für SV Calling mittels Illumina Daten wurden jedoch Falsch-Postiv-Raten von bis zu 89% berichtet. Außerdem ist ein eindeutiges Alignment von kurzen Illumina Reads an ein polyploides Referenzgenom, aufgrund der großen Ähnlichkeit der homöologen Regionen, äußerst schwierig. Aus diesem Grund gab es bisher wenig bis gar keine Bemühungen, kleine bis mittelgroße SV zu erfassen. Die vorliegende Arbeit beschreibt die Verwendung von langen Sequenzierungs-Reads, um die Rolle kleiner bis mittelgroßer SV zur Diversifizierung von *B. napus* in die drei vorherrschenden Ökotypen (Wintertyp, Sommertyp und Semi-Wintertyp) zu bewerten und deren Ausmaß und

Einfluss auf Gene zu untersuchen. Es wurde festgestellt, dass bis zu 10% aller Gene im Rapsgenom von kleinen bis mittelgroßen SV-Ereignissen betroffen sind. Fast die Hälfte dieser SV waren in der Größenordnung von 100 bis 1000 bp, was ihren Nachweis mittels Short Read Illumina-Sequenzierung schwierig macht. Zudem wurden kleine SV auch in jenen Genen nachgewiesen, die mit *Verticillium*-Resistenz in Raps assoziiert sind. Darüber hinaus liefert diese Arbeit erste Einblicke und Ideen, wie neue Long Read-Sequenzierungstechnologien helfen können, komplexe SV in großen Pflanzengenomen zu verstehen, indem sie zusätzliche Informationsebenen wie Methylierungsmuster, Chromatinkonformation oder Daten aus Target-Enrichment-Strategien, die die Long-Read-Sequenzierung verwenden, liefern und potenzielle zelluläre Mechanismen beschreiben, die das Auftreten kleiner bis mittelgroßer SV in Raps erklären könnten. Des Weiteren werden auch die Herausforderungen und Grenzen der Sequenziertechnologien der dritten Generation in dieser Dissertation erörtert.

Das zentrale Ergebnis dieser Dissertation war die erstaunlich hohe Anzahl weit verbreiteter, kleiner bis mittelgroßer SV im Rapsgenom. SV in diesem Größenbereich sind mittels Illumina-Sequenzierung nahezu undetektierbar und blieben daher in früheren Studien, die genomische Umlagerungen in *B. napus* untersuchten, völlig unberücksichtigt. Die Ergebnisse dieser Dissertation deuten darauf hin, dass die Analyse komplexer Pflanzengenome mit Hilfe von Long Read-Sequenzierung mit mittlerer Genomabdeckung eine unerwartete Anzahl funktioneller Genvariationen mit großen Auswirkungen auf die Merkmalsausprägung und die Verbesserung der Kulturpflanzen aufdecken könnte.

8 References

Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, Pozniak CJ, Choulet F, Distelfeld A, Poland J, Ronen G, Sharpe AG, Barad O, Baruch K, Keeble-Gagnère G, Mascher M, Ben-Zvi G, Josselin A-A, Himmelbach A, Balfourier F, Gutierrez-Gonzalez J, Hayden M, Koh C, Muehlbauer G, Pasam RK, Paux E, Rigault P, Tibbits J, Tiwari V, Spannagl M, Lang D, Gundlach H, Haberer G, Mayer KFX, Ormanbekova D, Prade V, Šimková H, Wicker T, Swarbreck D, Rimbert H, Felder M, Guilhot N, Kaithakottil G, Keilwagen J, Leroy P, Lux T, Twardziok S, Venturini L, Juhász A, Abrouk M, Fischer I, Uauy C, Borrill P, Ramirez-Gonzalez RH, Arnaud D, Chalabi S, Chalhoub B, Cory A, Datla R, Davey MW, Jacobs J, Robinson SJ, Steuernagel B, van Ex F, Wulff BBH, Benhamed M, Bendahmane A, Concia L, Latrasse D, Bartoš J, Bellec A, Berges H, Doležel J, Frenkel Z, Gill B, Korol A, Letellier T, Olsen O-A, Singh K, Valárik M, van der Vossen E, Vautrin S, Weining S, Fahima T, Glikson V, Raats D, Číhalíková J, Toegelová H, Vrána J, Sourdille P, Darrier B, Barabaschi D, Cattivelli L, Hernandez P, Galvez S, Budak H, Jones JDG, Witek K, Yu G, Small I, Melonek J, Zhou R, Belova T, Kanyuka K, King R, Nilsen K, Walkowiak S, Cuthbert R, Knox R, Wiebe K, Xiang D, Rohde A, Golds T, Čížková J, Akpinar BA, Biyiklioglu S, Gao L, N'Daiye A, Kubaláková M, Šafář J, Alfama F, Adam-Blondon A-F, Flores R, Guerche C, Loaec M, Quesneville H, Condie J, Ens J, Maclachlan R, Tan Y, Alberti A, Aury J-M, Barbe V, Couloux A, Cruaud C, Labadie K, Mangenot S, Wincker P, Kaur G, Luo M, Sehgal S, Chhuneja P, Gupta OP, Jindal S, Kaur P, Malik P, Sharma P, Yadav B, Singh NK, Khurana JP, Chaudhary C, Khurana P, Kumar V, Mahato A, Mathur S, Sevanthi A, Sharma N, Tomar RS, Holušová K, Plíhal O, Clark MD, Heavens D, Kettleborough G, Wright J, Balcárková B, Hu Y, Salina E, Ravin N, Skryabin K, Beletsky A, Kadnikov V, Mardanov A, Nesterov M, Rakitin A, Sergeeva E, Handa H, Kanamori H, Katagiri S, Kobayashi F, Nasuda S, Tanaka T, Wu J, Cattonaro F, Jiumeng M, Kugler K, Pfeifer M, Sandve S, Xun X, Zhan B, Batley J, Bayer PE, Edwards D, Hayashi S, Tulpová Z, Visendi P, Cui L, Du X, Feng K, Nie X, Tong W, Le Wang (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361(6403):eaar7191. doi: 10.1126/science.aar7191

Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, Genete M, Berrabah W, Chèvre A-M, Delourme R, Deniot G, Denoeud F, Duffé P, Engelen S, Lemainque A, Manzanares-Dauleux M, Martin G, Morice J, Noel B, Vekemans X, D'Hont A, Rousseau-Gueutin M, Barbe V, Cruaud C, Wincker P, Aury J-M (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants 4(11):879–887. doi: 10.1038/s41477-018-0289-4

Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A (2009) An operational definition of epigenetics. Genes Dev 23(7):781–783. doi: 10.1101/gad.1787609

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43(10):956–963. doi: 10.1038/ng.911

Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, Correa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, Arnaud D, Guo H, Daviaud C, Alamery S, Jabbari K, Zhao M, Edger PP, Chelaifa H, Tack D, Lassalle G, Mestiri I, Schnel N, Le Paslier M-C, Fan G, Renault V, Bayer PE, Golicz AA, Manoli S, Lee T-H, Thi VHD, Chalabi S, Hu Q, Fan C, Tollenaere R, Lu Y, Battail C, Shen J, Sidebottom CHD, Canaguier A, Chauveau A, Berard A, Deniot G, Guan M, Liu Z, Sun F, Lim YP, Lyons E, Town CD, Bancroft I, Meng J, Ma J, Pires JC, King GJ, Brunel D, Delourme R, Renard M, Aury J-M,

Adams KL, Batley J, Snowdon RJ, Tost J, Edwards D, Zhou Y, Hua W, Sharpe AG, Paterson AH, Guan C, Wincker P (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345(6199):950–953. doi: 10.1126/science.1253435

Chandak S, Tatwawadi K, Sridhar S, Weissman T (2020) Impact of lossy compression of nanopore raw signal data on basecall and consensus accuracy. bioRxivorg:2020.04.19.049262. doi: 10.1101/2020.04.19.049262

Chen J-M, Cooper DN, Chuzhanova N, Férec C, Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 8(10):762–775. doi: 10.1038/nrg2193

Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE (2012) Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus (Asteraceae)*. Proc Natl Acad Sci USA 109(4):1176–1181. doi: 10.1073/pnas.1112041109

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35(6):2013–2025. doi: 10.1093/nar/gkm076

Cytrynbaum C, Choufani S, Weksberg R (2019) Epigenetic signatures in overgrowth syndromes: Translational opportunities. Am J Med Genet 181(4):491–501. doi: 10.1002/ajmg.c.31745

Dolatabadian A, Bayer PE, Tirnaz S, Hurgobin B, Edwards D, Batley J (2020) Characterization of disease resistance genes in the *Brassica napus* pangenome reveals significant structural variation. Plant Biotechnol J 18(4):969–982. doi: 10.1111/pbi.13262

Doyle J (1991) DNA Protocols for Plants. In: Hewitt GM, Johnston AWB, Young JPW (eds) Molecular techniques in taxonomy. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 283–293

Du C, Hoffman A, He L, Caronna J, Dooner HK (2011) The complete *Ac/Ds* transposon family of maize. BMC genomics 12(1):588. doi: 10.1186/1471-2164-12-588

Dumschott K, Schmidt MH-W, Chawla HS, Snowdon R, Usadel B (2020) Oxford Nanopore Sequencing: New opportunities for plant genomics? J Exp Bot. doi: 10.1093/jxb/eraa263

Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, Kauwe JSK, Belzil V, Pregent L, Carrasquillo MM, Keene D, Larson E, Crane P, Asmann YW, Ertekin-Taner N, Younkin SG, Ross OA, Rademakers R, Petrucelli L, Fryer JD (2019) Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol. 20(1):97. doi: 10.1186/s13059-019-1707-2

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Bioinformatics 8(3):186–194

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7(6):461–465. doi: 10.1038/nmeth.1459

Fujii M, Yokosho K, Yamaji N, Saisho D, Yamane M, Takahashi H, Sato K, Nakazono M, Ma JF (2012) Acquisition of aluminium tolerance by modification of a single gene in barley. Nat Commun 3:713. doi: 10.1038/ncomms1726

Gabur I, Chawla HS, Liu X, Kumar V, Faure S, Tiedemann A von, Jestin C, Dryzska E, Volkmann S, Breuer F, Delourme R, Snowdon R, Obermeier C (2018) Finding invisible quantitative trait loci with missing data. Plant Biotechnol J. 16(12): 2102-2112. doi: 10.1111/pbi.12942

Gabur I, Chawla HS, Snowdon RJ, Parkin IAP (2019) Connecting genome structural variation with complex traits in crop plants. Theor. Appl. Genet. 132(3):733–750. doi: 10.1007/s00122-018-3233-0

Gaeta RT, Pires CJ (2010) Homoeologous recombination in allopolyploids: the polyploid ratchet. New Phytol 186(1):18–28. doi: 10.1111/j.1469-8137.2009.03089.x

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC (2007) Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. Plant Cell 19(11):3403–3417. doi: 10.1105/tpc.107.054346

Gardiner L-J, Wingen LU, Bailey P, Joynson R, Brabbs T, Wright J, Higgins JD, Hall N, Griffiths S, Clavijo BJ, Hall A (2019) Analysis of the recombination landscape of hexaploid bread wheat reveals genes controlling recombination and gene conversion frequency. Genome Biol. 20(1):69. doi: 10.1186/s13059-019-1675-6

Gigante S (2017) Picopore: A tool for reducing the storage size of Oxford Nanopore Technologies datasets without loss of functionality. F1000Res 6. doi: 10.12688/f1000research.11022.3

Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, Downs B, Sukumar S, Sedlazeck FJ, Timp W (2020) Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol 38(4):433–438. doi: 10.1038/s41587-020-0407-5

Golicz AA, Bhalla PL, Edwards D, Singh MB (2020) Rice 3D chromatin structure correlates with sequence variation and meiotic recombination rate. Commun Biol 3(1):235. doi: 10.1038/s42003-020-0932-2

Grandke F, Snowdon R, Samans B (2016) gsrc: an R package for genome structure rearrangement calling. Bioinformatics 33(4):545-546. doi: 10.1093/bioinformatics/btw648

Grech-Baran M, Witek K, Szajko K, Witek AI, Morgiewicz K, Wasilewicz-Flis I, Jakuczun H, Marczewski W, Jones JDG, Hennig J (2020) Extreme resistance to *Potato virus Y* in potato carrying the Ry_{sto} gene is mediated by a TIR-NLR immune receptor. Plant Biotechnol J 18(3):655–667. doi: 10.1111/pbi.13230

Guiglielmoni N, Derzelle A, van Doninck K, Flot J-F (2020) Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. bioRxivorg:2020.03.16.993428. doi: 10.1101/2020.03.16.993428

Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. Nat Rev Genet 10(8):551–564. doi: 10.1038/nrg2593

Hurgobin B, Golicz AA, Bayer PE, Chan C-KK, Tirnaz S, Dolatabadian A, Schiessl SV, Samans B, Montenegro JD, Parkin IAP, Pires JC, Chalhoub B, King GJ, Snowdon R, Batley J, Edwards D (2018) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. Plant Biotechnol J 16(7):1265–1274. doi: 10.1111/pbi.12867

Imelfort M, Edwards D (2009) De novo sequencing of plant genomes using second-generation technologies. Brief Bioinform 10(6):609–618. doi: 10.1093/bib/bbp039

Iyer S (2019) Understanding genetic variation in cancer, using targeted nanopore sequencing. Nanopore Community Meeting, NY,USA

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36(4):338–345. doi: 10.1038/nbt.4060

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431(7008):569–573. doi: 10.1038/nature02953

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40(10):1253–1260. doi: 10.1038/ng.237

Kovaka S, Fan Y, Ni B, Timp W, Schatz MC (2020) Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. bioRxivorg:2020.02.03.931923. doi: 10.1101/2020.02.03.931923

Kraft F, Kurth I (2019) Long-read sequencing in human genetics. medizinische genetik 31(2):198–204. doi: 10.1007/s11825-019-0249-z

Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, Tan Y, Hu J, Li X, Lai Q, Han L, Wang D, Hu F, Wang W, Liu S (2020) Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacbio Sequel II system and ultralong reads of Oxford Nanopore. bioRxivorg:2020.02.13.948489. doi: 10.1101/2020.02.13.948489

Lee JA, Carvalho CMB, Lupski JR (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131(7):1235–1247. doi: 10.1016/j.cell.2007.11.037

Li Y, Xiao J, Wu J, Duan J, Liu Y, Ye X, Zhang X, Guo X, Gu Y, Zhang L, Jia J, Kong X (2012) A tandem segmental duplication (TSD) in green revolution gene *Rht-D1b* region underlies plant height variation. New Phytol 196(1):282–291. doi: 10.1111/j.1469-8137.2012.04243.x

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133(3):523–536. doi: 10.1016/j.cell.2008.03.029

Liu C, Weigel D (2015) Chromatin in 3D: progress and prospects for plants. Genome Biol. 16(1):170. doi: 10.1186/s13059-015-0738-6

Liu Y, Cheng J, Siejka-Zielińska P, Weldon C, Roberts H, Lopopolo M, Magri A, D'Arienzo V, Harris JM, McKeating JA, Song C-X (2020) Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. Genome Biol. 21(1):54. doi: 10.1186/s13059-020-01969-6

Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14(10):417–422. doi: 10.1016/S0168-9525(98)01555-8

Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, Campbell BC, Hu W, Innes DJ, Han X, Cruickshank A, Dai C, Frère C, Zhang H, Hunt CH, Wang X, Shatte T, Wang M, Su Z, Li J, Lin X, Godwin ID, Jordan DR, Wang J (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun 4:2320. doi: 10.1038/ncomms3320

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019) Structural variant calling: the long and the short of it. Genome Biol. 20(1):246. doi: 10.1186/s13059-019-1828-7

Malmberg MM, Spangenberg GC, Daetwyler HD, Cogan NOI (2019) Assessment of low-coverage nanopore long read sequencing for SNP genotyping in doubled haploid canola (*Brassica napus* L.). Sci Rep 9(1):8688. doi: 10.1038/s41598-019-45131-0

Maron LG, Guimaraes CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D, Magalhaes JV, Pineros MA, Schatz MC, Wing RA, Kochian LV (2013) Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. Proc Natl Acad Sci USA 110(13):5241–5246. doi: 10.1073/pnas.1220766110

Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, Callot C, Langlade N, Muños S (2016) Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. BioTechniques 61(4):203–205. doi: 10.2144/000114460

McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, Mason CE (2019) Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat Commun 10(1):579. doi: 10.1038/s41467-019-08289-9

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR (2018) High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat Commun 9(1):541. doi: 10.1038/s41467-018-03016-2

Moore JK, Haber JE (1996) Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in *Saccharomyces cerevisiae*. Mol Cell Biol 16(5):2164–2173. doi: 10.1128/MCB.16.5.2164

Nishida H, Yoshida T, Kawakami K, Fujita M, Long B, Akashi Y, Laurie DA, Kato K (2013) Structural variation in the 5′ upstream region of photoperiod-insensitive alleles *Ppd-A1a* and *Ppd-B1a* identified in hexaploid wheat (*Triticum aestivum* L.), and their effect on heading time. Mol Breeding 31(1):27–37. doi: 10.1007/s11032-012-9765-0

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo M-j, Byers R, Chen W, Doron-Faigenboim A, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu G, Lee T-h, Li J, Lin L, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das S, Gingle AR, Haigler CH, Harker D, Hoffmann LV, Hovav R, Jones DC, Lemke C, Mansoor S, ur Rahman M, Rainville LN, Rambani A, Reddy UK, Rong J-k, Saranga Y, Scheffler BE, Scheffler JA, Stelly DM, Triplett BA, van Deynze A, Vaslin MFS, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KFX, Peterson DG, Rokhsar DS, Wang X, Schmutz J (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492(7429):423–427. doi: 10.1038/nature11798

Payne A, Holmes N, Clarke T, Munro R, Debebe B, Loose M (2020) Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels. bioRxivorg:2020.02.03.926956. doi: 10.1101/2020.02.03.926956

Perumal S, Koh CS, Jin L, Buchwaldt M, Higgins E, Zheng C, Sankoff D, Robinson SJ, Kagale S, Navabi Z-K, Tang L, Horner KN, He Z, Bancroft I, Chalhoub B, Sharpe AG, Parkin IAP (2020) High contiguity long read assembly of *Brassica nigra* allows localization of active centromeres and provides insights into the ancestral *Brassica* genome. bioRxivorg:2020.02.03.932665. doi: 10.1101/2020.02.03.932665

Qian L, Voss-Fels K, Cui Y, Jan HU, Samans B, Obermeier C, Qian W, Snowdon RJ (2016) Deletion of a Stay-Green gene associates with adaptive selection in *Brassica napus*. Mol Plant 9(12):1559–1569. doi: 10.1016/j.molp.2016.10.017

Samans B, Chalhoub B, Snowdon RJ (2017) Surviving a Genome Collision: Genomic Signatures of Allopolyploidization in the Recent Crop Species *Brassica napus*. The Plant Genome 10(3). doi: 10.3835/plantgenome2017.02.0013

Schalamun M, Nagar R, Kainer D, Beavan E, Eccles D, Rathjen JP, Lanfear R, Schwessinger B (2019) Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from *Eucalyptus pauciflora*. Mol Ecol Resour 19(1):77–89. doi: 10.1111/1755-0998.12938

Schiessl S, Huettel B, Kuehn D, Reinhardt R, Snowdon R (2017a) Post-polyploidisation morphotype diversification associates with gene copy number variation. Sci Rep 7:41845. doi: 10.1038/srep41845

Schiessl S, Huettel B, Kuehn D, Reinhardt R, Snowdon RJ (2017b) Targeted deep sequencing of flowering regulators in *Brassica napus* reveals extensive copy number variation. Sci Data 4. doi: 10.1038/sdata.2017.13

Schiessl S-V, Katche E, Ihien E, Chawla HS, Mason AS (2019) The role of genomic structural variation in the genetic improvement of polyploid crops. The Crop Journal 7(2):127–140. doi: 10.1016/j.cj.2018.07.006

Schmidt MH-W, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Maß J, Pfaff C, Schurr U, Chetelat R, Maumus F, Aury J-M, Koren S, Fernie AR, Zamir D, Bolger AM, Usadel B (2017) De Novo assembly of a new *Solanum pennellii* accession using nanopore sequencing. Plant Cell 29(10):2336. doi: 10.1105/tpc.17.00521

Schrinner S, Mari RS, Ebler J, Rautiainen M, Seillier L, Reimer J, Usadel B, Marschall T, Klau G (2020) Haplotype Threading: Accurate Polyploid Phasing from Long Reads. bioRxiv 2020.02.04.933523. doi:10.1101/2020.02.04.933523

Sedlazeck FJ, Lee H, Darby CA, Schatz MC (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet 19(6):329–346. doi: 10.1038/s41576-018-0003-4

Sikorskaite S, Rajamäki M-L, Baniulis D, Stanys V, Valkonen JPT (2013) Protocol: Optimised methodology for isolation of nuclei from leaves of species in the *Solanaceae* and *Rosaceae* families. Plant Methods 9(1):31. doi: 10.1186/1746-4811-9-31

Snowdon RJ (2007) Cytogenetics and genome analysis in *Brassica* crops. Chromosome Res 15(1):85–95. doi: 10.1007/s10577-006-1105-y

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(1):34–45. doi: 10.1038/s41477-019-0577-7

Stein A, Coriton O, Rousseau-Gueutin M, Samans B, Schiessl SV, Obermeier C, Parkin IAP, Chèvre A-M, Snowdon RJ (2017) Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in *Brassica napus*. Plant Biotechnol J 15(11):1478–1489. doi: 10.1111/pbi.12732

Sun W, Bunn P, Jin C, Little P, Zhabotynsky V, Perou CM, Hayes DN, Chen M, Lin D-Y (2018) The association between copy number aberration, DNA methylation and gene expression in tumor samples. Nucleic Acids Res 46(6):3009–3018. doi: 10.1093/nar/gky131

Sutton T, Baumann U, Hayes J, Collins NC, Shi B-J, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, Langridge P (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318(5855):1446–1449. doi: 10.1126/science.1146853

Tsai Y-C, Greenberg D, Powell J, Höijer I, Ameur A, Strahl M, Ellis E, Jonasson I, Mouro Pinto R, Wheeler VC, Smith ML, Gyllensten U, Sebra R, Korlach J, Clark TA (2017) Amplification-free, CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions. bioRxivorg:203919. doi: 10.1101/203919

Uga Y, Sugimoto K, Ogawa S, Rane J, Ishitani M, Hara N, Kitomi Y, Inukai Y, Ono K, Kanno N, Inoue H, Takehisa H, Motoyama R, Nagamura Y, Wu J, Matsumoto T, Takai T, Okuno K, Yano M (2013) Control of root system architecture by *DEEPER ROOTING 1* increases rice yield under drought conditions. Nat Genet 45(9):1097–1102. doi: 10.1038/ng.2725

Ulahannan N, Pendleton M, Deshpande A, Schwenk S, Behr JM, Dai X, Tyer C, Rughani P, Kudman S, Adney E, Tian H, Wilkes D, Mosquera JM, Stoddart D, Turner DJ, Juul S, Harrington E, Imielinski M (2019) Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. bioRxivorg:833590. doi: 10.1101/833590

Wang S, Chen J, Zhang W, Hu Y, Chang L, Fang L, Wang Q, Lv F, Wu H, Si Z, Chen S, Cai C, Zhu X, Zhou B, Guo W, Zhang T (2015) Sequence-based ultra-dense genetic and physical maps reveal structural variations of allopolyploid cotton genomes. Genome Biol. 16(1):108. doi: 10.1186/s13059-015-0678-1

Weisweiler M, Montaigu Ad, Ries D, Pfeifer M, Stich B (2019) Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mRNA sequencing and their power to predict phenotypic traits. BMC genomics 20(1):787. doi: 10.1186/s12864-019-6174-3

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37(10):1155–1162. doi: 10.1038/s41587-019-0217-9

Wick RR, Judd LM, Holt KE (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 20(1):129. doi: 10.1186/s13059-019-1727-y

Xiong Z, Pires JC (2011) Karyotype and identification of all homoeologous chromosomes of allopolyploid *Brassica napus* and Its diploid progenitors. Genetics 187(1):37–49. doi: 10.1534/genetics.110.122473

Xuan YH, Piao HL, Je BI, Park SJ, Park SH, Huang J, Zhang JB, Peterson T, Han C-d (2011) Transposon *Ac/Ds*-induced chromosomal rearrangements at the rice *OsRLG5* locus. Nucleic Acids Res 39(22):e149. doi: 10.1093/nar/gkr718

Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, Deng T, Luo J, He L, Wang Y, Xu P, Peng Y, Shi Z, Lan L, Ma Z, Yang X, Zhang Q, Bai M, Li S, Li W, Liu L, Jackson D, Yan J (2019) Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet 51(6):1052–1059. doi: 10.1038/s41588-019-0427-6

Yu Z, Wright SI, Bureau TE (2000) Mutator-like elements in *Arabidopsis thaliana*. Structure, diversity and evolution. Genetics 156(4):2019–2031

Zhang J, Chen L-L, Sun S, Kudrna D, Copetti D, Li W, Mu T, Jiao W-B, Xing F, Lee S, Talag J, Song J-M, Du B, Xie W, Luo M, Maldonado CE, Goicoechea JL, Xiong L, Wu C, Xing Y, Zhou D-x, Yu S, Zhao Y, Wang G, Yu Y, Luo Y, Hurtado BEP, Danowitz A, Wing RA, Zhang Q (2016) Building two *indica* rice reference genomes with PacBio long-read and Illumina paired-end sequencing data. Sci Data 3(1):160076. doi: 10.1038/sdata.2016.76

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126(6):1189–1201. doi: 10.1016/j.cell.2006.08.003

Zhang Y, Yang L, Kucherlapati M, Hadjipanayis A, Pantazi A, Bristow CA, Lee EA, Mahadeshwar HS, Tang J, Zhang J, Seth S, Lee S, Ren X, Song X, Sun H, Seidman J, Luquette LJ, Xi R, Chin L, Protopopov A, Park PJ, Kucherlapati R, Creighton CJ (2019) Global impact of somatic structural variation on the DNA methylome of human cancers. Genome Biol. 20(1):209. doi: 10.1186/s13059-019-1818-9

Zhang Y-z, Akdemir A, Tremmel G, Imoto S, Miyano S, Shibuya T, Yamaguchi R (2020) Nanopore basecalling from a perspective of instance segmentation. BMC Bioinformatics 21(3):136. doi: 10.1186/s12859-020-3459-0

Zhao D, Jiang N (2014) Nested insertions and accumulation of indels are negatively correlated with abundance of *Mutator*-Like Transposable Elements in maize and rice. PLOS ONE 9(1):e87069. doi: 10.1371/journal.pone.0087069

Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS (2019) The population genetics of structural variants in grapevine domestication. Nat Plants 5(9):965–979. doi: 10.1038/s41477-019-0507-8

Zhu S, Beaulaurier J, Deikus G, Wu TP, Strahl M, Hao Z, Luo G, Gregory JA, Chess A, He C, Xiao A, Sebra R, Schadt EE, Fang G (2018) Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. Genome Res 28(7):1067–1078. doi: 10.1101/gr.231068.117

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. GigaScience 6(11):1–7. doi: 10.1093/gigascience/gix097

Declaration

I declare that the dissertation here submitted is entirely my own work, written without any illegitimate help by any third party and solely with materials as indicated in the dissertation. I have indicated in the text where I have used texts from already published sources, either word for word or in substance, and where I have made statements based on oral information given to me. At all times during the investigations carried out by me and described in the dissertation, I have followed the principles of good scientific practice as defined in the "Statutes of the Justus Liebig University Gießen for the Safeguarding of Good Scientific Practice".

Giessen, Tuesday 16th June, 2020 **Harmeet Singh Chawla**

Acknowledgements

I am very grateful to all the people working at the department of plant breeding at the Justus Liebig University in Giessen for their support during my doctoral thesis.

First and foremost, I would like to thank Prof. Dr. Rod Snowdon for providing all the help needed for the successful completion of my PhD. Rod, its entirely because of you I chose plant breeding as a career. Your master's module "Biotechnology and Genomics" inspired me to pursue plant breeding. Since my master thesis I have admired your farsightedness as you have always been encouraging to adopt new and modern methodologies in the field of plant genomics. I am grateful to you for providing me so many wonderful opportunities to travel and present my work at national and international scientific conferences. Finally, I would also like to thank you for writing wonderful letter of recommendations and endorsing me to Curtis for my postdoc position. I would definitely like to be in touch and keep collaborating with you.

I am also grateful to Prof. Dr. Matthias Frisch for agreeing to be the second supervisor for my doctoral thesis.

Another very important person for my scientific career I would like to sincerely thank is Dr. Christian Obermeier. I am thankful to him for providing me the opportunity to work at the department. I have been able to develop a scientific thinking because of the long discussions with him. He has always been supportive and has always been there whenever I needed any kind of help.

I would like to acknowledge the technical support I received during my thesis from our excellent technical assistants. Firstly, I would thank Stavors Tzigos (The Greek Lab officer) for constantly helping me in the lab. Even during his most busy times he took the time out to answer my questions and helped me to the best of his abilities. In addition to the professional help I also want to thank him for providing me with a second home at "Walltorstrasse". I am also thankful to Regina Illgner for her wonderful support for performing qPCR and DNA isolations for me. Lastly, a big thanks to Andreas Welke for all his help in lab and greenhouse work during my thesis.

This thesis would have not been possible without support from my wonderful colleagues and friends at the department of plant breeding. I am extremely thankful to Subhadra Chakrabarty for all her help professionally and also for providing me with best of the memories from Giessen. I am also greatly thankful to Paula Vasquez Teuber for her wonderful company and of course for

the legendary “Friday sessions”. A big thanks to Paul Vollrath for translating for me whenever I was stuck because of my nonexistent German skills and Pablo for a nice “outside work” time. I would like to extend my sincere thanks to Dr. Iulian Gabur for his professional help and motivating discussions. I am also thankful to Jenny Lee for helping me write this thesis and for the awesome discussions at work. I would appreciate Andreas “Hansi” Eckert for maintaining a cheerful environment in the office and for being the best IT support I could imagine and last but not the least I also want to thank Stjepan Vukasovic for the nice discussions.

I would like to express my sincere gratitude to both our wonderful secretaries Miss. Sabine Schomber and Miss Ulla Riedmeier. Thanks for taking care of all administrative stuff for me and most importantly for making me feel welcomed in the department with your warm gestures.

Last but not the least I would like to thank my parents and my little sister. My father Dr. Jasbir Singh Chawla, you are my hero and my inspiration and my mother, Mrs. Kiranjeet, who has been my mental support system. I am able to achieve this milestone because of you both and would always be indebted to you. Ravneet Kaur my sister, you have stood by me in worst of my time. “Thank you” is too small a word to acknowledge your contribution to my success.

I would like to acknowledge deNBI (German Network for Bioinformatics Infrastructure) for providing the compute resources necessary for the completion of this work.