Received: 14 May 2021 | Accepted: 7 October 2021

DOI: 10.1002/tpg2.20177

The Plant Genome

INVITED REVIEW

Pangenomics in crop improvement—from coding structural variations to finding regulatory variants with pangenome graphs

Silvia F. Zanini[1] | Philipp E. Bayer[2] | Rachel Wells[3] | Rod J. Snowdon[1] | Jacqueline Batley[2] | Rajeev K. Varshney[4,5] | Henry T. Nguyen[6] | David Edwards[2] | Agnieszka A. Golicz[1]

[1] Dep. of Plant Breeding, IFZ Research Centre for Biosystems, Land Use and Nutrition, Justus Liebig Univ. Giessen, Giessen 35392, Germany

[2] School of Biological Sciences and Institute of Agriculture, Univ. of Western Australia, Perth, Western Australia, Australia

[3] Dep. of Crop Genetics, John Innes Centre, Norwich Research Park, Norwich, NR47UH, UK

[4] Center of Excellence in Genomics & Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India

[5] State Agricultural Biotechnology Centre, Centre for Crop Food Innovation, Food Futures Institute, Murdoch Univ., Murdoch, WA, Australia

[6] Division of Plant Sciences, Univ. of Missouri, Columbia, MO, USA

Correspondence
Agnieszka Golicz, Dep. of Plant Breeding, Justus Liebig Univ. Giessen, 35392, Giessen, Germany.
Email:
Agnieszka.Golicz@agrar.uni-giessen.de

Assigned to Associate Editor Hon-Ming Lam.

Abstract

Since the first reported crop pangenome in 2014, advances in high-throughput and cost-effective DNA sequencing technologies facilitated multiple such studies including the pangenomes of oilseed rape (*Brassica napus* L.), soybean [*Glycine max* (L.) Merr.], rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), and barley (*Hordeum vulgare* L.). Compared with single-reference genomes, pangenomes provide a more accurate representation of the genetic variation present in a species. By combining the genomic data of multiple accessions, pangenomes allow for the detection and annotation of complex DNA polymorphisms such as structural variations (SVs), one of the major determinants of genetic diversity within a species. In this review we summarize the current literature on crop pangenomics, focusing on their application to find candidate SVs involved in traits of agronomic interest. We then highlight the potential of pangenomes in the discovery and functional characterization of noncoding regulatory sequences and their variations. We conclude with a summary and outlook on innovative data structures representing the complete content of plant pangenomes including annotations of coding and noncoding elements and outcomes of transcriptomic and epigenomic experiments.

1 | INTRODUCTION

The number of people globally affected by hunger has been rising since 2014 (FAO et al., 2020). Almost 690 million people—8.9% of the world population—are estimated to have been undernourished in 2019. Without change, the world is on track to reach 840 million undernourished people by

Abbreviations: ACR, accessible chromatin region; CNV, copy-number variation; CNV, copy-number variation; CRE, cis-regulatory element; GS, genomic selection; GWAS, genome-wide association studies; MAS, marker-assisted selection; PAV, presence-or-absence variation; PHG, Practical Haplotype Graph; QTL, quantitative trait loci; SNP, single-nucleotide polymorphism; SV, structural variation; SV, structural variation or structural variant; TE, transposable element; VG, Variation Graph

2030 (FAO et al., 2020). The reasons for an increasingly undernourished population are manifold including unequal resource distribution, food waste, and crop loss arising from climate change (Barrera & Hertel, 2021; Hasegawa et al., 2016; Janssens et al., 2020). While an integrated approach is vital to successfully curb this alarming trend, climate-change-resilient crops are needed to counter more frequent and extreme weather events that affect mostly populations with already high rates of undernourishment (FAO et al., 2020).

During recent decades, significant technological progress has been made in plant breeding to develop new cultivars including marker-assisted selection (MAS), which has been in use since the early 1990s (Dudley, 1993; Ribaut & Hoisington, 1998). In MAS, genomic markers are identified in silico and used to select individuals for use in specialized breeding programs (Hillel et al., 1990; Tanksley & Nelson, 1996). Originally, trait-associated genomic markers targeted by these approaches were identified as quantitative trait loci (QTL) (Geldermann, 1975), which in their earliest iteration could contain many genes within the same locus (Beckmann & Soller, 1983; Westman & Kresovich, 1997). The ability to identify candidate genes was accelerated by the construction of reference assemblies representing the DNA sequence of an individual's genome. Reference genomes were used to identify single-nucleotide polymorphisms (SNPs), copy-number variations (CNVs), and insertion–deletions (InDels). These markers became the foundation for genome-wide association studies (GWAS) and genomic selection (GS), where diversity sequencing datasets are compared with reference genomes and the identified variations statistically associated with phenotypes (Crossa et al., 2017; Hayes & Goddard, 2010; Ozaki et al., 2002; Varshney et al., 2005, 2009). The identification of trait-associated alleles and genes has also been used to drive a revolution in green biotechnology. The first genetically modified crop, the tomato (*Solanum lycopersicum* L.) 'FlavrSavr', was released to U.S. markets in the 1990s (Kramer & Redenbaugh, 1994). Since then, commercialization and distribution of genetically modified and, now, gene-edited crops have the potential to increase yield and improve traits such as biotic and abiotic stresses tolerance (Liu et al., 2021; Shi et al., 2017; Singh et al., 2018; Singh et al., 2020; Varshney et al., 2011; Wang et al., 2014; Zeng et al., 2019).

The approaches described above require the identification of candidate genes or sequences for modification; however, they usually rely on a single reference genome, which does not contain the full extent of genetic variation present in the species, especially in polyploidy crops (Bayer et al., 2020; Golicz et al., 2016a). This limitation has led to the rise of pangenomes, which combine the genomic data derived from multiple accessions and cultivars to detail the full extent of sequence variation within a species, finding genes and alleles to accelerate crop breeding. The pangenome concept was first proposed by Tettelin et al. (2005) to describe a *Streptococcus agalactiae*, or group B *Streptococcus*, reference combining datasets derived from eight different bacterial isolates (Tettelin et al., 2005). The genes present in all individuals were defined as 'core' genes, while the variable fraction was termed 'dispensable' (also referred to as 'accessory' or 'variable'). Functional characterization of group B *Streptococcus* core genes highlighted their involvement in essential processes, while a significant portion of disposable genes was found to be the cause of newly acquired traits such as antibiotic resistance (Tettelin et al., 2005). Later pangenomics studies confirmed the critical roles of several variable genes in the organism's adaptation to a certain environment, making them effectively indispensable to the specific strain or cultivar fitness (Scheben et al., 2016). Tettelin et al. (2005) also noted that increasing the sample size would expand the pangenome size indefinitely, making it an 'open' pangenome as opposed to a 'closed' one, whereby after inclusion of a sufficient number of samples, the addition of further datasets would not result in the identification of novel sequences (Tettelin et al., 2005). Since then, the adoption of high-throughput and cost-effective DNA sequencing technologies has resulted in the proliferation of pangenomes, including major crop species such as rice (*Oryza sativa* L.), maize (*Zea mays* L.), soybean [*Glycine max* (L.) Merr.], and rapeseed (*Brassica napus* L.) (Hirsch et al., 2014; Li et al., 2014; Song et al., 2020; Wang et al., 2018; Yao et al., 2015). Compared with single-reference genomes, pangenomes enable more accurate identification and representation of complex DNA polymorphisms within a species including large insertions, deletions, inversions, duplications, translocations, presence-or-absence variations (PAVs,) and copy-number variations (CNVs). Structural variations (or structural variants, SVs), ranging from few base pairs to several megabase pairs are the result of a variety of mechanisms including transposable elements (TEs) insertion, recombination, and double-strand breaks repair (Saxena et al., 2014). Structural variations are one of the major determinants of phenotypic variability within a species

Core Ideas

- Pangenomes allow integration of many types of DNA variants in a single reference.
- Pangenome studies highlight the importance of structural variations affecting agronomical traits.
- Many structural variations occur in regulatory regions and affect gene expression.
- Adopting pangenome graphs will help understand coding and noncoding variation.

including many agronomically important traits (Gao et al., 2020; Ledesma-Ramírez et al., 2019; Song et al., 2020).

Recent studies highlighted the value of analyzing pangenome content beyond the collection of protein coding genes including regulatory regions and repeat content (Gao et al., 2019; Hufford et al., 2021). However, such detailed analysis requires availability of multiple high-quality genome assemblies and advanced methods for their analysis and visualization, for example using the tools of graphical pangenomics (Eizenga et al., 2020). Pangenomic and comparative analyses are especially challenging for many crop plants that have high repeat content, are ancient, or are recent polyploids.

Here we review recent literature on pangenomics in crops with a focus on their application to find candidate SVs involved in traits of agronomic interest such as disease resistance or flowering time. We highlight the potential of using pangenomes for noncoding regulatory element discovery, functional characterization, and evaluation of SV impact on plant function. We finish with a discussion of innovative data structures representing the complete content of plant pangenomes including the complete repertoire of functional sequences.

2 | USING PANGENOMES TO DISSECT AGRONOMIC TRAITS

Consistent with early findings that the dispensable genome is enriched with genes involved in environmental responses, pangenomes are being increasingly used in the detection of sequences associated with agronomically relevant traits, such as yield and stress resistance, thereby leading the transition from so-called genomic-assisted to pangenomic-assisted breeding strategies (Table 1).

2.1 | Disease resistance

The sunflower (*Helianthus annuus* L.) pangenome was recently generated from 493 cultivars, landraces, and compatible wild species to investigate the impact of introgression from wild species on the dispensable genome composition of cultivars (Hübner et al., 2019). Approximately 1.5% of genes were identified as exclusively derived from introgression with wild relatives, including two candidate genes associated with resistance to downy mildew: a syntaxin (*SYP132*), known to contribute to bacterial resistance and PR-1 protein secretion (Kalde et al., 2007); and a GDSL-motif lipase involved in plant resistance to fungal infection (Oh et al., 2005). Similarly, multiple pangenomic studies in rapeseed confirmed the role played by dispensable genes in plant defense, with ∼70% of the total predicted resistance genes being found in the variable portion of the pangenome (53 rapeseed accessions) with almost 50% of them absent from the 'Darmor-*bzh*' reference (Hurgobin et al., 2018). Dispensable resistance genes include *BnaA03g43460.1D2*, a potential orthologue of a clubroot resistance gene in *Brassica rapa* L. , *CRa* (Ueno et al., 2012). A later study by Dolatabadian et al. (2020) expanded on this discovery, identifying 753 variable resistance gene analogs (RGA) in the rapeseed pangenome, with 106 resistance gene analog candidates predicted to contribute to blackleg resistance, one of the major diseases affecting *Brassica* species (Dolatabadian et al., 2020; Howlett et al., 2001). Among the dispensable sequences in the *Brachypodium distachyon* (L.) Beauv. pangenome assembly (54 *Brachypodium* lines) (Gordon et al., 2017), *Brdisv1Bd1-11011965m* was identified as a dispensable gene absent from the reference line (Bd21) and induced during infection of wheat stem rust in Bd1-1, a line resistant to *Puccinia graminis* f. sp. *tritici* (Figueroa et al., 2013). While the gene encodes for an uncharacterized protein, the region on which this gene is located is syntenic to the stem rust resistance gene *Sr2* locus in wheat (McFadden, 1930), supporting its potential role in *P. graminis* f. sp. *tritici* resistance in *Brachypodium*.

2.2 | Vernalization and flowering time

As the *Brachypodium distachyon* reference (Bd21) has the shortest flowering time recorded in the species, the pangenome was instrumental in investigating PAVs involved in flowering time variation (Gordon et al., 2017). The gene *Brdisv1ABR21022861m* was absent from both rapid and intermediate flowering lines (including Bd21) but present in all delayed or extremely delayed flowering lines. This gene encodes a NF-Y subunit transcription factor, a class of transcription factors known to regulate flowering in *Arabidopsis thaliana* (L.) Heynh., wheat, and rice (Kumimoto et al., 2008; Mayer et al., 2014; Wei et al., 2010), confirming the role *Brdisv1ABR21022861m* plays in determining *Brachypodium* flowering time.

In *Brassica napus*, pangenome-based comparative analysis and PAV GWAS identified variations causing several agronomically relevant traits including silique length, seed weight, and flowering time (Song et al., 2020). Three SVs were found to affect the expression of the same *FLOWERING LOCUS C* (*FLC*) gene, *BnaA10.FLC*, encoding a key transcriptional regulator responsible for delayed flowering and stronger vernalization requirement (Tadege et al., 2001). A LINE (long interspersed nuclear elements) insertion in the first exon of *BnaA10.FLC* detected in spring-type rapeseed led to low or no vernalization requirement for this ecotype. Conversely, a MITE (miniature inverted-repeat transposable elements) insertion in the promoter region of *BnaA10.FLC* was found in winter-type oilseed rape and resulted in a higher expression of *BnaA10.FLC* and stronger vernalization requirement.

TABLE 1 Agronomical traits addressed by selected pangenome studies highlighting whether a coding or noncoding associated variant was identified

Species	No. of accessions	Single reference size	Pangenome size	Traits studied using the pangenome	Variant type	Coding or noncoding	Reference
Brachypodium distachyon (purple false brome)	54	(Bd21) 272 Mb; 34,310 genes	430 Mb; 37,886 HC genes	Disease resistance, flowering time	PAV	Coding	Gordon et al., 2017
Brassica oleracea, *B. macrocarpa* (cultivated and wild cabbage)	9+1	(*Bo* TO1000) 488 Mb; 59,225 genes	587 Mb; 61,379 genes	Disease resistance, flowering time, secondary metabolites	PAV	Coding	Golicz, et al., 2016b
B. napus (rapeseed)	53	(Darmor-bzh v8.1) 850 Mb; 80,382 genes	1,044 Mb; 94,013 genes	Disease resistance	PAV, SNP	Coding	Dolatabadian et al., 2020; Hurgobin et al., 2018
B. napus (rapeseed)	8	(Zs11) 960.8 Mb, 100,919 genes	1.8 Gb; 152,185 genes	Silique length, seed weight, flowering time	PAV	Noncoding	Song et al., 2020
B. rapa (turnip, napa cabbage)	3	(Chiifu v1.2) 41,019 genes	41,858 genes	Flowering time, stress resistance, lignin formation	PAV	Coding	Lin et al., 2014
Cajanus cajan (pigeon pea)	89	(Asha) 606 Mb; 53,612 genes	622 Mb; 55,512 genes	Self-fertilization, disease resistance, seed weight	SNP, PAV	Coding	Zhao et al., 2020
Capsicum annuum, *C. baccatum*, *C. chinense*, *C. frutescens*(pepper)	383	(*Ca* Zunla 1) 35,336 HC genes	4.316 Gb; 51,757 HC genes	Carotenoid and capsaicinoid biosynthetic pathways	PAV	Coding	Ou et al., 2018
Glycine soja (wild soybean)	7	(GsojaD, Shandong) 985 Mb; 57,631 genes	986.3 Mb; 59,080 gene families	Disease resistance, flowering time, oil content, height and lodging, yield	CNV, PAV, SNP, InDel	Coding	Li et al., 2014
G. max, *G. soja* (cultivated and wild soybean)	29	(Zhonghuang 13) 1.025 Gb; 52,051 genes	57,492 orthologues	Iron uptake	PAV	Coding, noncoding	Liu et al., 2020
Gossypium hirsutum (upland cotton)	1581	(TM-1) 2,347 Mb; 70,199 genes	3,388 Mb; 102,768 genes	Flowering time, morphology, yield, fiber traits	PAV, SNP	Coding, noncoding	Li et al., 2021
Gossypium. barbadense (tropical cotton)	226	(3–79) 2,266 Mb; 71,297 genes	2,575 Mb; 80,148 genes	Disease resistance, stress response			
Helianthus annuus (sunflower)	493	(HA412-HO.v.1.1) 3.6 Gb; 52,232 genes	61,205 genes	Disease resistance	SNP, PAV	Coding	Hübner et al., 2019

(Continues)

TABLE 1 (Continued)

Species	No. of accessions	Single reference size	Pangenome size	Traits studied using the pangenome	Variant type	Coding or noncoding	Reference
Hordeum vulgare (barley)	20	(Morex V2) 32,878 genes	40,176 orthologues	Yield	PAV inversions	Coding	Jayakodi et al., 2020
Malus domestica, *M. sieversii*, *M. sylvestris* (cultivated and wild apple)	91	(*Md* Gala, haploid assembly) 45,352 protein-coding genes	69,411 orthologues	Fruit quality	SNP, PAV	Coding	Sun et al., 2020
Medicago truncatula (barrel medic)	15	(HM101, Mt4.0) 412 Mb; 46,523 genes	431 Mb; 74,700 genes	Pathogen detection	SNP, CNV	Coding	Zhou et al., 2017
Oryza sativa (rice)	1483	(Nipponbare) 384 Mb	*indica*: + 52,976 sequences; *japonica*: + 30,349 sequences	Disease resistance, stress resistance, grain width and size	SNP	Coding	Yao et al., 2015
O. sativa, *O. rufipogon* (cultivated and wild rice)	66+1	(*Os* Nipponbare) 384 Mb	42,580 genes	Flowering time, stress tolerance, grain weight, tiller angle and plant height, hull color	SNP, PAV	Coding	Zhao et al., 2018
O. sativa, *O. glaberrima* (cultivated and wild rice)	3010	(*Os* Nipponbare) 384 Mb	+ 268 Mb; + 12,465 genes	Flowering time, grain length and width, disease resistance	SNP, PAV	Coding	Wang et al., 2018
Solanum lycopersicum, *S. cheesmaniae*, *S. galapagense* (cultivated and wild tomato)	725	(*Sl* Heinz 1706, v ITAG3.2) 900 Mb; 35,496 genes	1,179 Mb; 40,369 genes	Disease resistance, fruit flavor	PAV	Coding, noncoding	Gao et al., 2019
Sorghum bicolor (sorghum)	176+ 1(354 for gPAV)	(Moench) 732.2 Mb	883.3 Mb; 35,719 genes	Drought resistance	SNP, PAV	Coding	Ruperao et al., 2021
Triticum aestivum (bread wheat)	18+1	(Chinese Spring) 10.7 Gb	~11 Gb; 140,500 genes	Disease resistance, stress resistance	PAV	Coding	Montenegro et al., 2017
Zea mays (maize) pan-transcriptome	503	(B73) 22,354 RTAs	+ 8,681 RTAs	Flowering time	SNP	Coding	Hirsch et al., 2014
Zea mays (maize)	26	(B73) 2,182 Mb	103,538 genes	Disease resistance, flowering time	TE-insertion; SNP, PAV	Coding, noncoding	Hufford et al., 2021

Note. CNV, copy-number variation; HC, high confidence; InDel, insertion–deletions; PAV, presence or absence variation; RTA, representative transcript assemblies; SNP, single-nucleotide polymorphism; TE, transposable element.

A *hAT* insertion in the promoter region of *BnaA10.FLC* was identified as the cause for the intermediate vernalization needs of the semi-winter oilseed rape ecotype, confirming the role played by *BnaA10.FLC* and the impact of SVs in regulating vernalization in rapeseed (Song et al., 2020). In addition to the four *FLC* paralogues identified in the reference sequence for TO1000, *B. oleracea* (L. var. *alboglabra* (L. H. Bailey) Musil] (Okazaki et al., 2007), a *B. oleracea* pangenome assembly led to the identification of another candidate *FLC* gene, *BoFLC2*, which is missing from the reference genome but present in all other lines (Golicz, et al., 2016b). This absence is presumed to be caused by a deletion event that occurred in TO1000 and to be the determinant of the early flowering phenotype of this rapid-cycling line. Among *B. oleracea* cultivars, curd initiation and flowering time in cauliflower were also shown to be associated with *BoFLC2* in a dose-dependent manner (Ridge et al., 2015). Finally, a recent maize study combined de novo genome, transcriptome, and methylome analyses of 26 inbred lines to further characterize maize genomic diversity and uncover novel variations in both coding and regulatory regions (Hufford et al., 2021). The resulting pangenome and pan-epigenome highlighted several SVs and their impact on phenotypes, including TE insertions upstream of transcription start sites of *GL15, ZCN10*, and *Dof21*, three known flowering-time-related genes. These insertions were found to correlate with the gene expression levels and, most notably, discern between temperate and tropical lines (Hufford et al., 2021).

2.3 | Fruit, grain, and seed quality

The pepper (*Capsicum* spp.) pangenome was assembled from 383 cultivars and used as a reference for PAV GWAS analysis to detect deletions in genes involved in carotenoid and capsaicinoid biosynthetic pathways (Ou et al., 2018). A 2.5-kb deletion in the predicted *Pungent gene 1* (*Pun1*, *pan02g021380*) was uncovered in 50 cultivars with known low capsaicin content, while a deletion in gene *pan06g005570* (a predicted capsanthin–capsorubin synthase) was found in 26 cultivars with yellow or orange fruits. Similarly, 725 tomato accessions were sequenced and compared with the ‘Heinz 1706’ assembly to identify nonreference sequences and build a pangenome comprising both cultivated and wild tomato varieties (Alonge et al., 2020; Gao et al., 2019). Pangenome analysis led to the identification of a shorter, nonreference allele in the *TomLoxC* promoter, which was abundant in wild relatives but negatively selected during domestication, resulting in lower expression levels of *TomLoxC* and a less desirable flavor in cultivated tomatoes. Several modern elite breeding lines originally selected for stress tolerance also showed an unintended and unexpected recovery of this rare allele, with an increase in apocarotenoid production levels in these varieties (Gao et al., 2019). In addition to the examples reported above, a recent soybean pangenome study analyzed nonredundant SVs identified by comparing 29 whole-genome assemblies to genotype 2,898 accessions. These SVs were used in a pangenome graph-based PAV GWAS, which lead to the identification of a 10-kb PAV. The deleted gene located in this region encodes a hydrophobic protein from soybean (HPS) and is associated with seed luster (Liu et al., 2020).

2.4 | Abiotic stress tolerance

A recently constructed sorghum [*Sorghum bicolor* (L.) Moench] pangenome was used to identify drought-responsive genes in transcriptomic data generated from both resistant and susceptible genotypes (Abdel-Ghany et al., 2020; Ruperao et al., 2021; Varoquaux et al., 2019). A total of 79 genes absent from the reference genome were confirmed as differentially expressed during drought stress, including two resistance specific genes, *Sobic.005G069800* and *Sobic.006G127800*, which were found to co-map with plant height and leaf pigment traits. Analysis of structural variations in a soybean pangenome identified a gene encoding a Fe^{2+}/Zn^{2+} regulated transporter associated with iron deficiency chlorosis (Liu et al., 2020).

3 | PANGENOMES, SVs, AND REGULATORY ELEMENTS—A NEW AVENUE FOR CROP IMPROVEMENT

As outlined above, while variations in coding regions have been the major focus of functional studies thus far, recent reports highlight the importance of SVs affecting *cis*-regulatory elements (CREs) including promoters and enhancers (Figure 1a). Structural variations can potentially affect gene expression via several mechanisms including regulatory element insertion, duplication, deletion, and disruption of the three-dimensional chromatin structure (Figure 1b) (Chiang et al., 2017; Doğan & Liu, 2018). From a plant breeding perspective, regulatory variants are especially valuable because of their potential to provide a spectrum of trait variation by fine tuning gene expression and precise manipulation of quantitative traits (Rodríguez-Leal et al., 2017). Changes in CREs, which are often modular and tissue-specific, are also predicted to be less pleiotropic than those in protein-coding genes (Wittkopp & Kalay, 2011). Alterations of the regulatory regions are therefore expected to result in more subtle phenotypic effects.

The introduction of chromatin accessibility assays, such as ATAC-seq (Buenrostro et al., 2013), allowed for genome-wide identification of accessible chromatin regions (ACRs), which often coincide with regulatory elements. This

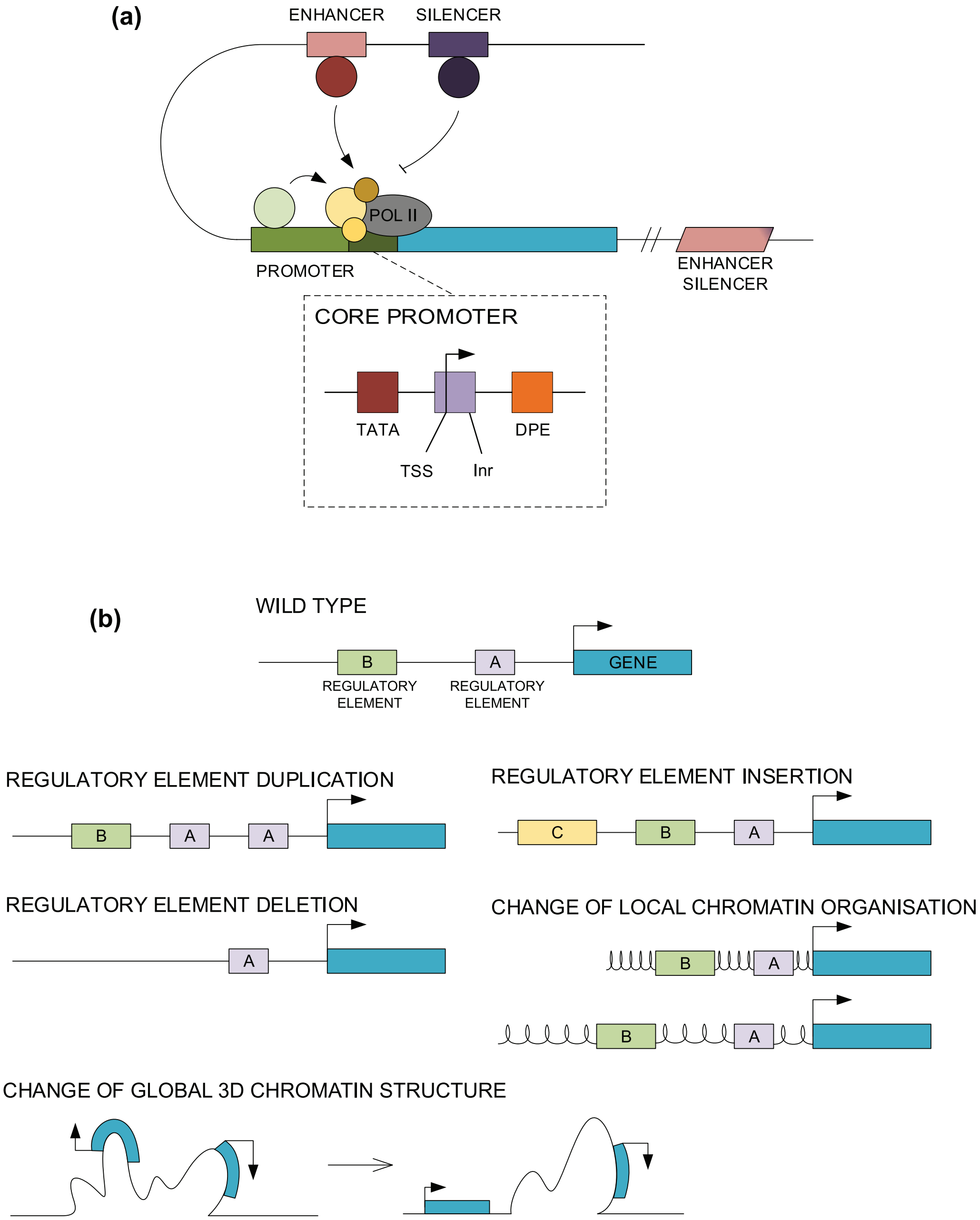
(a)
ENHANCER
SILENCER
POL II
PROMOTER
ENHANCER
SILENCER
CORE PROMOTER
TATA
TSS
Inr
DPE
(b)
WILD TYPE
B
REGULATORY ELEMENT
A
REGULATORY ELEMENT
GENE
REGULATORY ELEMENT DUPLICATION
B
A
A
REGULATORY ELEMENT INSERTION
C
B
A
REGULATORY ELEMENT DELETION
A
CHANGE OF LOCAL CHROMATIN ORGANISATION
B
A
B
A
CHANGE OF GLOBAL 3D CHROMATIN STRUCTURE

FIGURE 1 (a) Overview of *cis*-regulatory elements (CREs) controlling gene expression. The CREs are noncoding DNA sequences capable of recruiting transcription factors and affecting gene expression. The CREs can be broadly subdivided into promoters and enhancers or silencers. Promoters are found directly upstream of the transcription start site, whereas enhancers and silencers can be found megabase pairs (Mbp) away (the promoter configuration shown is not universal). Green box, promoter; pink, enhancer; purple, silencer; blue, gene. (b) Graphical representation of different structural variants potentially affecting gene expression including duplication, deletion, and insertion of entire or segments of regulatory elements, changes in local chromatin organization (or accessibility), and global 3D structure. Purple, green, yellow boxes (A, B, C), regulatory elements; blue, gene

technique was originally developed for animal and human research, though several protocols have now been optimized for plant tissues and applied to study ACR distribution in major crops such as maize, wheat, and barley (*Hordeum vulgare* L.) (Concia et al., 2020; Lu et al., 2019; Ricci et al., 2019; Sun et al., 2020). In combination with ACR profiling, the application of long-read sequencing and pangenomics provides a powerful approach for the assessment of the effects of structural variants on CREs and gene expression.

3.1 | Effect of SVs on gene expression

A recent study on tomato reported that 95% of the 34,075 annotated genes had at least one SV within 5 kb of the coding sequence across the 100 genomes analyzed (Alonge et al., 2020). Approximately 10% of variants occurring in regulatory regions were associated with changes in expression vs. ~50% of variants affecting coding regions. Consistent with predicted subtler effects of regulatory variants, SVs affecting regulatory regions were reported to have less impact on gene expression than those affecting coding sequences, with mean $\log_2$ fold change of 1.36 and 2.47, respectively (Alonge et al., 2020). In maize, expression QTL (eQTL) analysis using kernel transcriptome data from 368 maize inbred lines confirmed that while eQTLs with a greater effect tended to overlap genes, the majority (~80%) of expression-associated SVs were found in intergenic regions (Yang et al., 2019). Furthermore, by proportion, eQTLs were seven times more likely to be detected using SVs rather than SNP markers, suggesting that SVs have stronger impact on gene expression than SNPs.

The study of regulatory regions will be of particular interest with regard to pangenomes of polyploid crops, where multiple copies of genes exist and alterations in CREs can result in neofunctionalization by promoting changes in timing and location of gene expression. A study in rapeseed, for example, showed that flowering time genes are preferentially retained post-polyploidization, resulting in diverged expression patterns between homologues, most likely resulting from CRE divergence (Jones et al., 2018).

3.2 | TEs as novel regulatory elements

Transposable elements are a known driver of SVs, and they have been shown to shape the regulatory landscape of crop plant genomes either by the disruption of existing or the generation of new regulatory elements (Chuong et al., 2017; Feschotte, 2008; Gill et al., 2021). In maize, the comparison of four lines identified almost 3,000 ACRs contained within TEs and several hundred that overlapped a TE insertion in at least one of the lines. The TE insertions were shown to be associated with changes in methylation, chromatin accessibility, and, potentially, regulatory functions. TEs carrying ACRs were enriched for association with higher expression of nearby genes, suggesting they contribute novel regulatory elements (Noshay et al., 2021). Beyond genome-wide analyses, individual examples also highlight the role of TEs in gene expression regulation. A TE (*Hopscotch*) inserted in a regulatory element of maize *teosinte brached1* (*tb1*) was shown to act as an enhancer and partially explained apical dominance in maize (Studer et al., 2011). Also in maize, *Ac*/*fAc*, a *hAT* family element that can undergo transposition using termini of two adjacent elements, was shown to induce expression of *pericarp color 2* gene (*p2*) by capturing the enhancer sequence of another gene (Su et al., 2020). In *Brassica napus*, a CACTA-like TE inserted upstream of a P450 monooxygenase (*BnaA9.CYP78A9*) led to an increase in silique length and seed weight (Shi et al., 2019). In addition, insertion of a highly methylated COPIA-like long terminal repeat retrotransposon in the promoter region of *SHATTERPROOF1* homologue resulted in the repression of its expression and shattering resistance in rapeseed (Liu et al., 2020). Considering the key role played by TEs in generating SVs and modulating plant yield and fitness, their precise identification and annotation are among the outstanding challenges to generating high-quality pangenomes (Coletta et al., 2021).

3.3 | Editing regulatory elements for crop improvement

Mounting evidence supports the importance of regulatory region variation for crop improvement. In maize, a strong relationship was observed between the genetic variation in putative regulatory elements and complex trait variation (Rodgers-Melnick et al., 2016). The divergence of *cis*-regulatory regions associated with domestication also underscores their important roles in the control of traits targeted by artificial selection (Lemmon et al., 2014; Wang et al., 2017). Considering the lower potential for pleiotropic effects, CREs constitute attractive genome editing targets.

In tomato, inspired by the natural variation observed between wild and domesticated relatives, researchers used CRISPR-Cas9 mutagenesis to generate novel alleles of the promoter of *SlCLV3*, a gene affecting fruit size and engineered a continuum of phenotypic variation. It was noted that although patterns were observed, with larger deletions having the most impact, the magnitude of the phenotypic effect could not be easily predicted from mutations alone, suggesting a complex interplay between CREs controlling gene expression (Rodríguez-Leal et al., 2017). Similarly, in maize, it was possible to engineer a spectrum of variation for yield-related traits by targeting homologues of *CLV3* (Liu et al., 2021). It is conceivable that facilitated by pangenomics-based CRE identification and functional characterization, similar approaches can be applied to other species and candidate genes (Figure 2).

4 | PANGENOME GRAPH—UNIFYING FRAMEWORK FOR PANGENOME ANALYSIS

Construction of high-quality reference genomes of multiple individuals of the same species has become the norm. The availability of corresponding epigenomic and transcriptomic data has enabled the functional annotation of these references—from coding and noncoding genes to regulatory elements. The challenge is the accurate representation of the wealth of available information (Jayakodi et al., 2021). Several pangenomic models have been established thus far—from

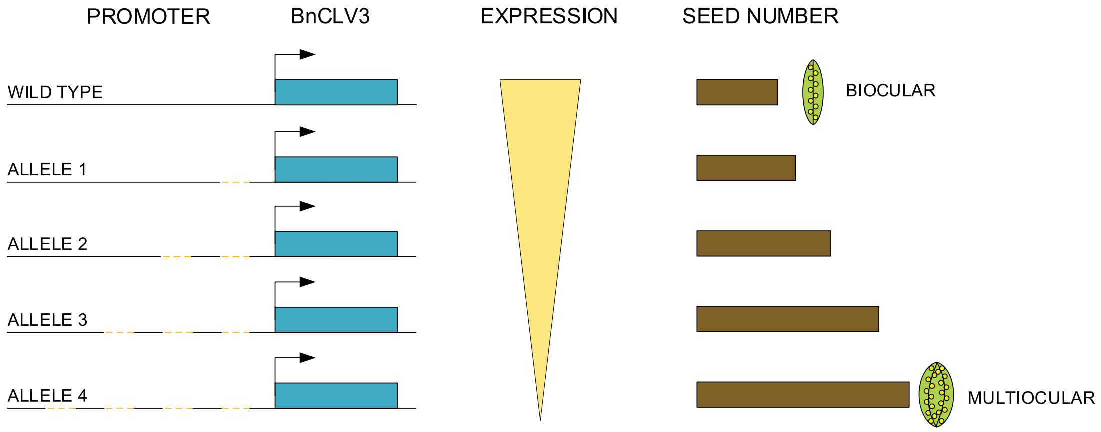

FIGURE 2 Genome editing of *cis*-regulatory elements: a hypothetical scenario of editing of *Brassica napus CLV3* homologues' *cis*-regulatory elements to generate multiocular siliques and range of variation in seed number. *Brassica napus* has two, mostly redundant, copies of *BnCLV3*, so editing of both would be necessary (Xu et al., 2021; Yang et al., 2018)

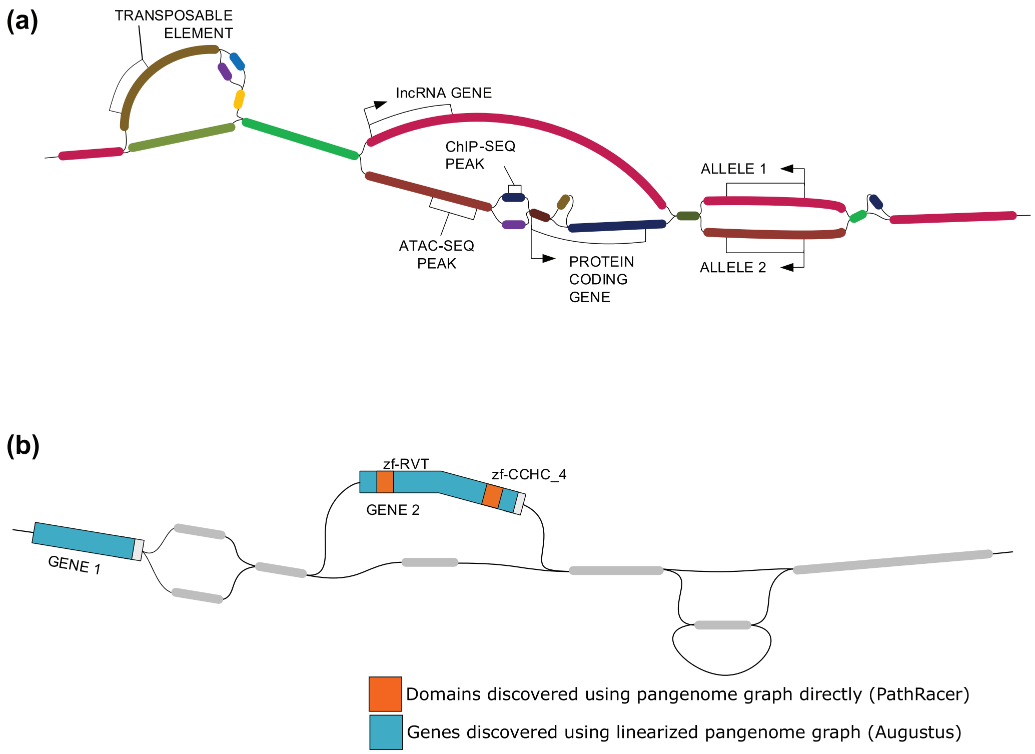
(a)
TRANSPOSABLE ELEMENT
lncRNA GENE
ChIP-SEQ PEAK
ALLELE 1
ATAC-SEQ PEAK
PROTEIN CODING GENE
ALLELE 2
(b)
zf-RVT
zf-CCHC_4
GENE 2
GENE 1
🟧 Domains discovered using pangenome graph directly (PathRacer)
🟦 Genes discovered using linearized pangenome graph (Augustus)

FIGURE 3 Pangenome graph annotation. (a) A hypothetical example of pangenome graph annotation and visualization integrating multiple layers of information including annotation with coding and noncoding loci, regulatory elements, results of transcriptomic, and ChIP-seq and ATAC-seq experiments. Annotating pangenome graph directly allows discovery of features spanning multiple segments (nodes). (b) Actual functional annotation of a segment of pangenome graph representing *Brassica napus* chromosome A01. The graph was built using minigraph with Darmor-bzh (Rousseau-Gueutin et al., 2020) and Express 617 (Lee et al., 2020) assemblies and annotated using PathRacer (applied directly to the pangenome graph (only selected domains were used in the search); Shlemov & Korobeynikov, 2019) and Augustus (applied to a linearized version of the pangenome graph; Stanke et al., 2006; linearized pangenome was obtained with gfatools gfa2fa)

TABLE 2 Overview of key bioinformatics analytical tasks and corresponding tools that can be used for liner genome and pangenome graph analysis

Application	Genome tool	Pangenome tool	Notes
Read mapping	Bowtie2 (Langmead & Salzberg, 2012), BWA (Li & Durbin, 2009), BWA MEM (Li, 2013), HISAT2 (Kim et al., 2019)	VG map (Garrison et al., 2018), GraphAligner (Rautiainen & Marschall, 2020), Minigraph (H. Li et al., 2020), Giraffe (Sirén et al., 2020)	VG map and Minigraph map both short and long reads. GraphAligner maps long reads only. Giraffe maps only short reads. HISAT2 is spliced read alignment (RNA-seq) enabled and can also account for variants by building acyclic variation graph.
Single-nucleotide variant (SNV) genotyping	GATK (Poplin et al., 2018), FreeBayes (Garrison & Marth, 2012), Bcftools (Li, 2011)	Graphtyper (Eggertsson et al., 2017)	Many linear genome tools perform both SNV discovery and genotyping.
Structural variation (SV) genotyping	SVTyper (Chiang et al., 2015), Nebula (Khorsand & Hormozdiari, 2021), SVJedi (Lecompte et al., 2020), Sniffles (Sedlazeck et al., 2018)	Paragraph (Chen et al., 2019), GraphTyper2 (Eggertsson et al., 2019), VG toolkit (Garrison et al., 2018)	Minigraph can also be used for SV calling.
Gene annotation	MAKER2 (Holt & Yandell, 2011), BRAKER2 (Brůna et al., 2021), FINDER (Banerjee et al., 2021), Funannotate (Palmer & Stajich, 2017)	Not available	Performing independent annotation of individual genomes may lead to artefacts (Bayer et al., 2017; König et al., 2016) Annotations of linear genomes can be projected onto graph using VG annotate/rna.
Homology searches	BLAST (Camacho et al., 2009), DIAMOND (Buchfink et al., 2015), HMMER (hmmer.org)	PLAST (Schulz et al., 2021), PathRacer (Shlemov & Korobeynikov, 2019)	PLAST is pangenome-graph counterpart to BLAST (currently only supports graphs built by Bifrost). PathRacer aligns profile HMM directly to the assembly graph (an experimental version supporting minigraph gfa files is available).
ChIP-seq peak calling	MACS (Zhang et al., 2008)	Graph Peak Caller (Grytten et al., 2019)	Graph Peak Caller is based on MACS and VG.
Transcript mapping and quantification	HISAT2 (Kim et al., 2019), STAR (Dobin et al., 2013), Kallisto (Bray et al., 2016), featureCounts (Liao et al., 2014), RSEM (Li & Dewey, 2011)	VG rna (constructs spliced graph), VG mpmap (maps reads to spliced graph), RPVG (quantifies transcript expression)	Pangenome graph-based workflow improves accuracy over state-of-the-art RNA-seq mapping methods (Sibbesen et al., 2021). It allows quantification of haplotype-specific transcript expression.
Association studies	Plink, TASSEL (Bradbury et al., 2007), GAPIT (Lipka et al., 2012), GenABEL (Aulchenko et al., 2007)	Pangenome-wide association studies (approach is based on frequented regions in pangenome graph; Manuweera et al., 2019)	Structural variations genotyped from a pangenome graph can be projected onto a linear reference and the results can be used for analysis using standard linear genome tools, adjusting for population genomics assumptions.
Visualization	GBrowse (Stein, 2013), JBrowse (Buels et al., 2016), Tablet (Milne et al., 2010), IGV (Robinson et al., 2011), MUMmer (Marçais et al., 2018), SyMAP (Soderlund et al., 2006)	Bandage (Wick et al., 2015), MoMI-G (Yokoyama et al., 2019), Sequence Tube Maps (Beyer et al., 2019), GfaViz (Gonnella et al, 2019), Pantograph (https://graphgenome.org/)	One of the standing challenges of pangenome visualization tools is display of metadata and annotations.

Note. For linear genomes, the most popular or representative tools are shown although many more exist.

simple collections of unaligned sequences to graphical representations. Pangenome graphs can be used to represent the sequence content and the corresponding functional annotation of an entire population, species, or a clade by compressing redundant sequences into smaller data structures while retaining information on genomic diversity and whole-genome relationships (Eizenga et al., 2020) (Figure 3). Graphs are composed of the DNA sequences (nodes), links between them (edges), and information about arrangement of nodes found in each constituent genome (paths). Paths provide a stable coordinate system allowing management of positions, annotations, and alignments across multiple genomes and transitions between graph and linear coordinates. To date, several approaches for pangenome graph construction have been adopted including saturation of the reference sequence with variants and using whole genome alignments. The Variation Graph (VG) toolkit (Garrison et al., 2018) can be employed to transform a reference sequence and variation file in Variant Call Format (VCF) into a pangenome graph and build pangenome from multiple genome assemblies (pggb from VG developers). It also contains a suite of tools for mapping of reads to the pangenome graph, variant genotyping, and projection of linear annotations. Minigraph can be used to generate graphs from assembled genomes, map sequences to graphs, and call structural variants (Li et al., 2020).

The main obstacle to the widespread adoption of pangenome graphs is the lack of suitable bioinformatics tools, as those designed for linear genomes are not readily transferable. These include tools for read mapping, small and large variant calling and genotyping, haplotype inference, gene annotation, homology searches, epigenomic, transcriptomic, association studies, and pangenome visualization (Table 2). Integrative genomics approaches have been shown to be invaluable for sequence functional annotation (Golicz et al., 2018; Hassani-Pak et al., 2021). Ideally, a complete and fully annotated pangenome graph would integrate genomic, epigenomic, and transcriptomic datasets, thus facilitating downstream functional and comparative analyses (Figure 3).

Despite the early stage of pangenome graph-specific tool development, several practical applications have already emerged. Practical Haplotype Graph (PHG) is a graph-based computation framework and associated pipeline for inference of high-density genotype from low coverage (skim) sequencing. A PHG was employed in maize to impute genotypes of recombinant inbred lines of a NAM population with an average accuracy of over 99%. Compared with standard genotype files, PHG increased the efficiency of data storage by four orders of magnitude (30,000-fold) (Franco et al., 2020). Graph Peak Caller is another tool designed specifically for the identification of ChIP-seq peaks using pangenome graph as a reference. Analysis of *Arabidopsis* 1000-genomes data showed that the combination of a VG-constructed graph and Graph Peak Caller identified peaks overlapping sequence not found in the linear reference. The peaks found were generally more motif enriched, suggesting higher-quality calls (Grytten et al., 2019). Recently a VG-based pantranscriptome pipeline became available, which allows construction of spliced pangenome graphs, mapping of RNA sequencing data, and haplotype-aware expression quantification (Sibbesen et al., 2021). The pangenome graph also provides a convenient framework for the genotyping of SVs across a large number of individuals, for example, for use in SV GWAS studies (Liu et al., 2020; Ruperao et al., 2021; Song et al., 2020; Zhao et al., 2020). The availability of comprehensive, well-annotated pangenome graphs including both genes and regulatory elements will become a key stepping-stone for the next generation of genomic analyses.

5 | CONCLUSIONS

While most pangenomics studies to date have focused on SVs affecting coding regions, the importance of their impact on the regulatory elements, including promoters and enhancers, has become apparent. Availability of genome-wide chromatin accessibility profiling technologies, which can be applied to multiple species and tissue types, allows inclusion of CRE profiling in pangenome analyses. This will enable, for example, the assessment of CRE diversity within species, evaluation of the SV impact on gene expression, and estimation of the relative contributions of CRE and coding sequences diversity to the phenotypic differences observed. Inclusion of crop-wild relatives in pangenomes, referred as ‘super pangenomes’ (Khan et al., 2020), will improve the identification of CREs targeted by artificial selection (Lemmon et al., 2014; Wang et al., 2017), providing new genome editing targets. Simultaneously, expansion of pangenome studies to higher taxonomic units will help identify core conserved regulatory modules and species-specific layers of regulation. Widespread adoption of the pangenome graph as a reference will allow for greater integration and varying data types, facilitating functional and comparative analyses to develop the next generation of climate change resilient and high-performance crops.

ACKNOWLEDGMENTS

This work was supported by Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research to Agnieszka A. Golicz.

AUTHOR CONTRIBUTIONS

Silvia F. Zanini: Investigation, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. Philipp E. Bayer: Investigation, Methodology, Resources, Writing – original draft. Rachel Wells: Conceptualization, Writing – review & editing. Rod J. Snowdon:

Conceptualization, Writing – review & editing. Jacqueline Batley: Conceptualization, Writing – review & editing. Rajeev K. Varshney: Conceptualization, Writing – review & editing. Henry T. Nguyen: Conceptualization, Writing – review & editing. David Edwards: Conceptualization, Writing – review & editing. Agnieszka A. Golicz: Conceptualization, Resources, Visualization, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Silvia F. Zanini https://orcid.org/0000-0002-9137-8783
Philipp E. Bayer https://orcid.org/0000-0001-8530-3067
Rachel Wells https://orcid.org/0000-0002-1280-7472
Rod J. Snowdon https://orcid.org/0000-0001-5577-7616
Jacqueline Batley https://orcid.org/0000-0002-5391-5824
Rajeev K. Varshney https://orcid.org/0000-0002-4562-9131
David Edwards https://orcid.org/0000-0001-7599-6760
Agnieszka A. Golicz https://orcid.org/0000-0002-9711-4826

REFERENCES

Abdel-Ghany, S. E., Ullah, F., Ben-Hur, A., & Reddy, A. S. N. (2020). Transcriptome analysis of drought-resistant and drought-sensitive sorghum (*Sorghum bicolor*) genotypes in response to PEG-induced drought stress. *International Journal of Molecular Sciences*, *21*, 772. https://doi.org/10.3390/ijms21030772

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

Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). Genabel: An R library for genome-wide association analysis. *Bioinformatics*, *23*, 1294–1296. https://doi.org/10.1093/bioinformatics/btm108

Banerjee, S., Bhandary, P., Woodhouse, M., Sen, T. Z., Wise, R. P., & Andorf, C. M. (2021). FINDER: An automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinformatics*, *22*, 205. https://doi.org/10.1186/s12859-021-04120-9

Barrera, E. L., & Hertel, T. (2021). Global food waste across the income spectrum: Implications for food prices, production and resource use. *Food Policy*, *98*, 101874. https://doi.org/10.1016/j.foodpol.2020.101874

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, *6*, 914–920. https://doi.org/10.1038/s41477-020-0733-0

Bayer, P. E., Hurgobin, B., Golicz, A. A., Chan, C. K. K., Yuan, Y., Lee, H., Renton, M., Meng, J., Li, R., Long, Y., Zou, J., Bancroft, I., Chalhoub, B., King, G. J., Batley, J., & Edwards, D. (2017). Assembly and comparison of two closely related *Brassica napus* genomes. *Plant Biotechnology Journal*, *15*, 1602–1610. https://doi.org/10.1111/pbi.12742

Beckmann, J. S., & Soller, M. (1983). Restriction fragment length polymorphisms in genetic improvement: Methodologies, mapping and costs. *Theoretical and Applied Genetics*, *67*, 35–43. https://doi.org/10.1007/bf00303919

Beyer, W., Novak, A. M., Hickey, G., Chan, J., Tan, V., Paten, B., & Zerbino, D. R. (2019). Sequence tube maps: Making graph genomes intuitive to commuters. *Bioinformatics*, *35*, 5318. https://doi.org/10.1093/bioinformatics/btz597

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). Tassel: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*, 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*, 525–527. https://doi.org/10.1038/nbt.3519

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*, lqaa108. https://doi.org/10.1093/nargab/lqaa108

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*, 59–60. https://doi.org/10.1038/nmeth.3176

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., Stein, L., & Holmes, I. H. (2016). JBrowse: A dynamic web platform for genome visualization and analysis. *Genome biology*, *17*, 66. https://doi.org/10.1186/s13059-016-0924-1

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*, 1213–1218. https://doi.org/10.1038/nmeth.2688

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). Blast+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., Kirsche, M., Bentley, D. R., Schatz, M. C., Sedlazeck, F. J., & Eberle, M. A. (2019). Paragraph: A graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, *20*, 291. https://doi.org/10.1186/s13059-019-1909-7

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nature Methods*, *12*, 966–968. https://doi.org/10.1038/nmeth.3505

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., Montgomery, S. B., Battle, A., Conrad, D. F., & Hall, I. M., & GTEx Consortium (2017). The impact of structural variation on human gene expression. *Nature Genetics*, *49*, 692–699. https://doi.org/10.1038/ng.3834

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, *18*, 71–86. https://doi.org/10.1038/nrg.2016.139

Coletta R. D., Qiu, Y., Ou, S., Hufford, M. B., & Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biology*, *22*, 3. https://doi.org/10.1186/s13059-020-02224-8

Concia, L., Veluchamy, A., Ramirez-Prado, J. S., Martin-Ramirez, A., Huang, Y., Perez, M., Domenichini, S., Rodriguez Granados N. Y., Kim, S., Blein, T., Duncan, S., Pichot, C., Manza-Mianza, D., Juery, C., Paux, E., Moore, G., Hirt, H., Bergounioux, C., … Benhamed, M. (2020). Wheat chromatin architecture is organized in genome territories and transcription factories. *Genome Biology*, *21*, 104. https://doi.org/10.1186/s13059-020-01998-1

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, *22*, 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Doğan, E. S., & Liu, C. (2018). Three-dimensional chromatin packing and positioning of plant genomes. *Nature Plants*, *4*, 521–529. https://doi.org/10.1038/s41477-018-0199-5

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

Dudley, J. W. (1993). Molecular markers in plant improvement: Manipulation of genes affecting quantitative traits. *Crop Science*, *33*, 660–668. https://doi.org/10.2135/cropsci1993.0011183X003300040003x

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., & Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, *49*, 1654–1660. https://doi.org/10.1038/ng.3964

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., Gudbjartsson, D. F., Stefansson, K., Halldorsson, B. V., & Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, *10*, 5402. https://doi.org/10.1038/s41467-019-13341-9

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J. D., Rounthwaite, R., Ebler, J., Rautiainen, M., Garg, S., Paten, B., Marschall, T., Sirén, J., & Garrison, E. (2020). Pangenome graphs. *Annual Review of Genomics and Human Genetics*, *21*, 139–162. https://doi.org/10.1146/annurev-genom-120219-080406

FAO, IFAD, UNICEF, WFP, & WHO. (2020). *State of food security and nutrition in the world 2020: Transforming food systems for affordable healthy diets*. Food & Agriculture Org. https://doi.org/10.4060/ca9692en

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*, 397–405. https://doi.org/10.1038/nrg2337

Figueroa, M., Alderman, S., Garvin, D. F., & Pfender, W. F. (2013). Infection of *Brachypodium distachyon* by formae speciales of *Puccinia graminis*: Early infection events and host-pathogen incompatibility. *PLoS ONE*, *8*, e56857. https://doi.org/10.1371/journal.pone.0056857

Franco, J. A. V., Gage, J. L., Bradbury, P. J., Johnson, L. C., Miller, Z. R., Buckler, E. S., & Romay, M. C. (2020). A maize practical haplotype graph leverages diverse NAM assemblies. *bioRxiv*, *2020.08.31.268425*. https://doi.org/10.1101/2020.08.31.268425

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., van der Knaap, E., Huang, S., Klee, H. J., …, & Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, *51*, 1044–1051. https://doi.org/10.1038/s41588-019-0410-2

Gao, S., Wu, J., Stiller, J., Zheng, Z., Zhou, M., Wang, Y. G., & Liu, C. (2020). Identifying barley pan-genome sequence anchors using genetic mapping and machine learning. *Theoretical and Applied Genetics*, *133*, 2535–2544. https://doi.org/10.1007/s00122-020-03615-y

Garrison, E., & Marth, G. (July 17). (2012). Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907*. https://arxiv.org/abs/1207.3907v2

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*, 875–879. https://doi.org/10.1038/nbt.4227

Geldermann, H. (1975). Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. *Theoretical and Applied Genetics*, *46*, 319–330. https://doi.org/10.1007/BF00281673

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

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

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016b). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, *7*, 13390. https://doi.org/10.1038/ncomms13390

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

Gonnella, G., Niehus, N., & Kurtz, S. (2019). GfaViz: Flexible and interactive visualization of GFA sequence graphs. *Bioinformatics*, *35*, 2853–2855. https://doi.org/10.1093/bioinformatics/bty1046

Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., Stritt, C., Roulin, A. C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N., Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T. E., Amasino, R., … Vogel, J. P. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications*, *8*, 2184. https://doi.org/10.1038/s41467-017-02292-8

Grytten, I., Rand, K. D., Nederbragt, A. J., Storvik, G. O., Glad, I. K., & Sandve, G. K. (2019). Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes. *PLOS Computational Biology*, *15*, e1006731. https://doi.org/10.1371/journal.pcbi.1006731

Hasegawa, T., Fujimori, S., Takahashi, K., Yokohata, T., & Masui, T. (2016). Economic implications of climate change impacts on human health through undernourishment. *Climatic Change*, *136*, 189–202. https://doi.org/10.1007/s10584-016-1606-4

Hassani-Pak, K., Singh, A., Brandizi, M., Hearnshaw, J., Parsons, J. D., Amberkar, S., Phillips, A. L., Doonan, J. H., , & Rawlings, C. (2021). KnetMiner: A comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnology Journal*, *19*, 1670–1678. https://doi.org/10.1111/pbi.13583

Hayes, B., & Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome*, *53*, 876–883. https://doi.org/10.1139/G10-076

Hillel, J., Schaap, T., Haberfeld, A., Jeffreys, A. J., Plotzky, Y., Cahaner, A., & Lavi, U. (1990). DNA fingerprints applied to gene introgression in breeding programs. *Genetics*, *124*, 783–789. https://www.genetics.org/content/124/3/783.short

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., & Buell, C. R. (2014). Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell*, *26*, 121–135. https://doi.org/10.1105/tpc.113.119982

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491. https://doi.org/10.1186/1471-2105-12-491

Howlett, B. J., Idnurm, A., & Pedras, M. S. C. (2001). *Leptosphaeria maculans*, the causal agent of blackleg disease of Brassicas. *Fungal Genetics and Biology*, *33*, 1–14. https://doi.org/10.1006/fgbi.2001.1274

Hübner, S., Bercovich, N., Todesco, M., Mandel, J. R., Odenheimer, J., Ziegler, E., Lee, J. S., Baute, G. J., Owens, G. L., Grassa, C. J., Ebert, D. P., Ostevik, K. L., Moyers, B. T., Yakimowski, S., Masalia, R. R., Gao, L., Ćalić, I., Bowers, J. E., Kane, N. C., … Rieseberg, L. H. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nature Plants*, *5*, 54–62. https://doi.org/10.1038/s41477-018-0329-0

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tittes, S., Hudson, A. I., Marand, A. P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M. K., Piri, R. D., …, & Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, *373*, 6555. https://doi.org/10.1126/science.abg5289

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C. K. K., Tirnaz, S., Dolatabadian, A., Schiessl, S. V., Samans, B., Montenegro, J. D., Parkin, I. A. P., Pires, J. C., Chalhoub, B., King, G. J., Snowdon, R., Batley, J., & Edwards, D. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal*, *16*, 1265–1274. https://doi.org/10.1111/pbi.12867

Janssens, C., Havlík, P., Krisztin, T., Baker, J., Frank, S., Hasegawa, T., Leclère, D., Ohrel, S., Ragnauth, S., Schmid, E., & Valin, H. (2020). Global hunger and climate change adaptation through international trade. *Nature Climate Change*, *10*, 829–835. https://doi.org/10.1038/s41558-020-0847-4

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V. S., Gundlach, H., Monat, C., Lux, T., Kamal, N., Lang, D., Himmelbach, A., Ens, J., Zhang, X. Q., Angessa, T. T., Zhou, G., Tan, C., Hill, C., Wang, P., Schreiber, M., Boston, L. B., …, & Stein, N. (2020). The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, *588*, 284–289. https://doi.org/10.1038/s41586-020-2947-8

Jayakodi, M., Schreiber, M., Stein, N., & Mascher, M. (2021). Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Research*, *28*, dsaa030. https://doi.org/10.1093/dnares/dsaa030

Jones, D. M., Wells, R., Pullen, N., Trick, M., Irwin, J. A., & Morris, R. J. (2018). Spatio-temporal expression dynamics differ between homologues of flowering time genes in the allopolyploid *Brassica napus*. *The Plant Journal*, *96*, 103–118. https://doi.org/10.1111/tpj.14020

Kalde, M., Nühse, T. S., Findlay, K., & Peck, S. C. (2007). The syntaxin SYP132 contributes to plant resistance against bacteria and secretion of pathogenesis-related protein 1. *Proceedings of the National Academy of Sciences*, *104*, 11850–11855. https://doi.org/10.1073/pnas.0701083104

Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D., & Varshney, R. K. (2020). Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends in Plant Science*, *25*, 148–158. https://doi.org/10.1016/j.tplants.2019.10.012

Khorsand, P., & Hormozdiari, F. (2021). Nebula: Ultra-efficient mapping-free structural variant genotyper. *Nucleic Acids Research*, *49*, e47. https://doi.org/10.1093/nar/gkab025

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4

König, S., Romoth, L. W., Gerischer, L., & Stanke, M. (2016). Simultaneous gene finding in multiple genomes. *Bioinformatics*, *32*, 3388–3395. https://doi.org/10.1093/bioinformatics/btw494

Kramer, M. G., & Redenbaugh, K. (1994). Commercialization of a tomato with an antisense polygalacturonase gene: The FLAVR SAVR™ tomato story. *Euphytica*, *79*, 293–297. https://doi.org/10.1007/BF00022530

Kumimoto, R. W., Adam, L., Hymus, G. J., Repetti, P. P., Reuber, T. L., Marion, C. M., Hempel, F. D., & Ratcliffe, O. J. (2008). The nuclear factor Y subunits NF-YB2 and NF-YB3 play additive roles in the promotion of flowering by inductive long-day photoperiods in Arabidopsis. *Planta*, *228*, 709–723. https://doi.org/10.1007/s00425-008-0773-6

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.1923

Lecompte, L., Peterlongo, P., Lavenier, D., & Lemaitre, C. (2020). SVJedi: Genotyping structural variations with long reads. *Bioinformatics*, *36*, 4568–4575. https://doi.org/10.1093/bioinformatics/btaa527

Ledesma-Ramírez, L., Solís-Moya, E., Iturriaga, G., Sehgal, D., Reyes-Valdes, M. H., Montero-Tavera, V., Sansaloni, C. P., Burgueño, J., Ortiz, C., Aguirre-Mancilla, C. L., Ramírez-Pimentel, J. G., Vikram, P., & Singh, S. (2019). GWAS to identify genetic loci for resistance to yellow rust in wheat pre-breeding lines derived from diverse exotic crosses. *Frontiers in Plant Science*, *10*, 1390. https://doi.org/10.3389/fpls.2019.01390

Lee, H., Chawla, H. S., Obermeier, C., Dreyer, F., Abbadi, A., & Snowdon, R. (2020). Chromosome-scale assembly of winter oilseed rape Brassica napus. *Frontiers in Plant Science*, *11*, 496. https://doi.org/10.3389/fpls.2020.00496

Lemmon, Z. H., Bukowski, R., Sun, Q., & Doebley, J. F. (2014). The role of *cis* regulatory evolution in maize domestication. *PLoS Genetics*, *10*, e1004745. https://doi.org/10.1371/journal.pgen.1004745

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. https://doi.org/10.1186/1471-2105-12-323

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*. https://arxiv.org/pdf/1303.3997

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, *21*, 265. https://doi.org/10.1186/s13059-020-02168-z

Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M., Liu, Z., …, & Wang, M. (2021). Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biology*, *22*, 119. https://doi.org/10.1186/s13059-021-02351-w

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., Hong, H., Tan, B., …, & Qiu, L. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, *32*, 1045–1052. https://doi.org/10.1038/nbt.2979

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656

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

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., & Zhang, Z. (2012). Gapit: Genome association and prediction integrated tool. *Bioinformatics*, *28*, 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Liu, J., Zhou, R., Wang, W., Wang, H., Qiu, Y., Raman, R., Mei, D., Raman, H., & Hu, Q. (2020). A *copia*-like retrotransposon insertion in the upstream region of the *SHATTERPROOF1* gene, *BnSHP1.A9*, is associated with quantitative variation in pod shattering resistance in oilseed rape. *Journal of Experimental Botany*, *71*, 5402–5413. https://doi.org/10.1093/jxb/eraa281

Liu, L., Gallagher, J., Arevalo, E. D., Chen, R., Skopelitis, T., Wu, Q., Bartlett, M., & Jackson, D. (2021). Enhancing grain-yield-related traits by CRISPR–Cas9 promoter editing of maize *CLE* genes. *Nature Plants*, *7*, 287–294. https://doi.org/10.1038/s41477-021-00858-5

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G. A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., & Tian, Z. (2020). Pan-genome of wild and cultivated soybeans. *Cell*, *182*, 162–176.E13. https://doi.org/10.1016/j.cell.2020.05.023

Loira, N., Zhukova, A., & Sherman, D. J. (2015). Pantograph: A template-based method for genome-scale metabolic model reconstruction. *Journal of Bioinformatics and Computational Biology*, *13*, 1550006. https://doi.org/10.1142/s0219720015500067

Lu, Z., Marand, A. P., Ricci, W. A., Ethridge, C. L., Zhang, X., & Schmitz, R. J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature Plants*, *5*, 1250–1259. https://doi.org/10.1038/s41477-019-0548-z

Manuweera, B., Mudge, J., Kahanda, I., Mumey, B., Ramaraj, T., & Cleary, A. (2019). Pangenome-wide association studies with frequented regions. In BCB '19: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics (pp. 627–632). https://doi.org/10.1145/3307339.3343478

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*, e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Mayer, K. F. X., Rogers, J., Dole el, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A. J., Sourdille, P., Endo, T. R., Kubalakova, M., ihalikova, J., Dubska, Z., Vrana, J., perkova, R., imkova, H., Febrer, M., Clissold, L., …, & Praud, S. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, *345*, 1251788. https://doi.org/10.1126/science.1251788

McFadden, E. S. (1930). A successful transfer of emmer characters to vulgare wheat. *Journal of the American Society of Agronomy*, *22*, 1020–1034.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., & Marshall, D. (2010). Tablet–Next generation sequence assembly visualization. *Bioinformatics*, *26*, 401–402. https://doi.org/10.1093/bioinformatics/btp666

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, *90*, 1007–1013. https://doi.org/10.1111/tpj.13515

Noshay, J. M., Marand, A. P., Anderson, S. N., Zhou, P., Mejia Guerra, M. K., Lu, Z., O'Connor, C. H., Crisp, P. A., Hirsch, C. N., Schmitz, R. J., & Springer, N. M. (2021). Assessing the regulatory potential of transposable elements using chromatin accessibility profiles of maize transposons. *Genetics*, *217*, 1–13. https://doi.org/10.1093/genetics/iyaa003

Oh, I. S., Park, A. R., Bae, M. S., Kwon, S. J., Kim, Y. S., Lee, J. E., Kang, N. Y., Lee, S., Cheong, H., & Park, O. K. (2005). Secretome analysis reveals an *Arabidopsis* lipase involved in defense against *Alternaria brassicicola*. *The Plant Cell*, *17*, 2832–2847. https://doi.org/10.1105/tpc.105.034819

Okazaki, K., Sakamoto, K., Kikuchi, R., Saito, A., Togashi, E., Kuginuki, Y., Matsumoto, S., & Hirai, M. (2007*)*. Mapping and characterization of FLC homologs and QTL analysis of flowering time in *Brassica oleracea*. *Theoretical and Applied Genetics*, *114*, 595–608. https://doi.org/10.1007/s00122-006-0460-6

Ou, L., Li, D., Lv, J., Chen, W., Zhang, Z., Li, X., Yang, B., Zhou, S., Yang, S., Li, W., Gao, H., Zeng, Q., Yu, H., Ouyang, B., Li, F., Liu, F., Zheng, J., Liu, Y., Wang, J., …, & Zou, X. (2018). Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses. *The New Phytologist*, *220*, 360–363. https://doi.org/10.1111/nph.15413

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., & Tanaka, T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, *32*, 650–654. https://doi.org/10.1038/ng1047

Palmer, J., & Stajich, J. E. (2017). Funannotate: Eukaryotic genome annotation pipeline. *Zenodo*. https://doi.org.10.5281/zenodo.3548120

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178. https://doi.org/10.1101/201178

Rautiainen, M., & Marschall, T. (2020). GraphAligner: Rapid and versatile sequence-to-graph alignment. *Genome Biology*, *21*, 253. https://doi.org/10.1186/s13059-020-02157-2

Ribaut, J. M., & Hoisington, D. (1998). Marker-assisted selection: New tools and strategies. *Trends in Plant Science*, *3*, 236–239. https://doi.org/10.1016/S1360-1385(98)01240-0

Ricci, W. A., Lu, Z., Ji, L., Marand, A. P., Ethridge, C. L., Murphy, N. G., Noshay, J. M., Galli, M., Mejía-Guerra, M. K., Colomé-Tatché, M., Johannes, F., Rowley, M. J., Corces, V. G., Zhai, J., Scanlon, M. J., Buckler, E. S., Gallavotti, A., Springer, N. M., Schmitz, R. J., & Zhang, X. (2019). Widespread long-range *cis*-regulatory elements in the maize genome. *Nature Plants*, *5*, 1237–1249. https://doi.org/10.1038/s41477-019-0547-0

Ridge, S., Brown, P. H., Hecht, V., Driessen, R. G., & Weller, J. L. (2015). The role of BoFLC2 in cauliflower (*Brassica oleracea* var. *Botrytis* L.) reproductive development. *Journal of Experimental Botany*, *66*, 125–135. https://doi.org/10.1093/jxb/eru408

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*, 24–26. https://doi.org/10.1038/nbt.1754

Rodgers-Melnick, E., Vera, D. L., Bass, H. W., & Buckler, E. S. (2016). Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences*, *113*, E3177. https://doi.org/10.1073/pnas.1525244113

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., & Lippman, Z. B. (2017). Engineering quantitative trait variation for crop improvement by genome editing. *Cell*, *171*, 470–480.E8. https://doi.org/10.1016/j.cell.2017.08.030

Rousseau-Gueutin, M., Belser, C., Da Silva, C., Richard, G., Istace, B., Cruaud, C., Falentin, C., Boideau, F., Boutte, J., Delourme, R., Deniot, G., Engelen, S., Ferreira de Carvalho, J., Lemainque, A., Maillet, L., Morice, J., Wincker, P., Denoeud, C., Chèvre, A.-M., & Aury, J.-M. (2020). Long-read assembly of the Brassica napus reference genome Darmor-bzh. *GigaScience*, *9*, giaa137. https://doi.org/10.1093/gigascience/giaa137

Ruperao, P., Thirunavukkarasu, N., Gandham, P., Selvanayagam, S., Govindaraj, M., Nebie, B., Manyasa, E., Gupta, R., Das, R. R., Gandhi, H., Edwards, D., Deshpande, S. P., & Rathore, A. (2021). Sorghum pan-genome explores the functional utility to accelerate the genetic gain. *Frontiers in Plant Science*, *12*, 666342. https://doi.org/10.3389/fpls.2021.666342

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

Scheben, A., Yuan, Y., & Edwards, D. (2016). Advances in genomics for adapting crops to climate change. *Current Plant Biology*, *6*, 2–10. https://doi.org/10.1016/j.cpb.2016.09.001

Schulz, T., Wittler, R., Rahmann, S., Hach, F., & Stoye, J. (2021). Detecting high scoring local alignments in pangenome graphs. *Bioinformatics*, *37*, 2266–2274. https://doi.org/10.1093/bioinformatics/btab077

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

Shi, J., Gao, H., Wang, H., Lafitte, H. R., Archibald, R. L., Yang, M., Hakimi, S. M., Mo, H., & Habben, J. E. (2017). Argos8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. *Plant Biotechnology Journal*, *15*, 207–216. https://doi.org/10.1111/pbi.12603

Shi, L., Song, J., Guo, C., Wang, B., Guan, Z., Yang, P., Chen, X., Zhang, Q., King, G. J., Wang, J., & Liu, K. (2019). A CACTA-like transposable element in the upstream region of *BnaA9.CYP78A9* acts as an enhancer to increase silique length and seed weight in rapeseed. *The Plant Journal*, *98*, 524–539. https://doi.org/10.1111/tpj.14236

Shlemov, A., & Korobeynikov, A. (2019). PathRacer: Racing profile HMM paths on assembly graph. In I. Homes, C. Martín-vide, & m. vega-rodriguez (eds.), *Algorithms for computational biology. alcob 2019. Lecture notes in computer science, Vol. 11488* (pp. 80–94). Springer. https://doi.org/10.1007/978-3-030-18174-1_6

Sibbesen, J. A., Eizenga, J. M., Novak, A. M., Sirén, J., Chang, X., Garrison, E., & Paten, B. (2021). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *bioRxiv*. 2021.03.26.437240. https://doi.org/10.1101/2021.03.26.437240

Singh, M., Kumar, M., Albertsen, M. C., Young, J. K., & Cigan, A. M. (2018). Concurrent modifications in the *three homeologs of Ms45 gene with CRISPR-Cas9 lead to rapid generation of male sterile bread wheat (*Triticum aestivum L.). *Plant Molecular Biology*, *97*, 371–383. https://doi.org/10.1007/s11103-018-0749-2

Singh, R. K., Prasad, A., Muthamilarasan, M., Parida, S. K., & Prasad, M. (2020). Breeding and biotechnological interventions for trait improvement: Status and prospects. *Planta*, *252*, 54. https://doi.org/10.1007/s00425-020-03465-4

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J., Hickey, G., Chang, P. C., Carroll, A., Haussler, D., Garrison, E., & Paten, B. (2020). Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *BioRxiv*, 2020.12.04.412486. https://doi.org/10.1101/2020.12.04.412486

Soderlund, C., Nelson, W., Shoemaker, A., & Paterson, A. (2006). SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Research*, *16*, 1159–1168. https://doi.org/10.1101/gr.5396706

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

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*, W435–W439. https://doi.org/10.1093/nar/gkl200

Stein, L. D. (2013). Using GBrowse 2.0 to visualize and share next-generation sequence data. *Briefings in Bioinformatics*, *14*, 162–171. https://doi.org/10.1093/bib/bbt001

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

Su, W., Zuo, T., & Peterson, T. (2020). Ectopic expression of a maize gene is induced by composite insertions generated through alternative transposition. *Genetics*, *216*, 1039. https://doi.org/10.1534/genetics.120.303592

Sun, X., Jiao, C., Schwaninger, H., Chao, C. T., Ma, Y., Duan, N., Khan, A., Ban, S., Xu, K., Cheng, L., Zhong, G. Y., & Fei, Z. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nature Genetics*, *52*, 1423–1432. https://doi.org/10.1038/s41588-020-00723-9

Sun, Y., Dong, L., Zhang, Y., Lin D., Xu, W., Ke, C., Han, L., Deng, L., Li, G., Jackson, D., Li, X., & Yang, F. (2020). 3D genome architecture coordinates trans and *cis* regulation of differentially expressed ear and tassel genes in maize. *Genome Biology*, *21*, 143. https://doi.org/10.1186/s13059-020-02063-7

Tadege, M., Sheldon, C. C., Helliwell, C. A., Stoutjesdijk, P., Dennis, E. S., & Peacock, W. J. (2001). Control of flowering time by FLC orthologues in *Brassica napus*. *The Plant Journal*, *28*, 545–553. https://doi.org/10.1046/j.1365-313x.2001.01182.x

Tanksley, S. D., & Nelson, J. C. (1996). Advanced backcross QTL analysis: A method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theoretical and Applied Genetics*, *92*, 191–203. https://doi.org/10.1007/s001220050114

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., & Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences*, *102*, 13950–13955. https://doi.org/10.1073/pnas.0506758102

Ueno, H., Matsumoto, E., Aruga, D., Kitagawa, S., Matsumura, H., & Hayashida, N. (2012). Molecular characterization of the *CRa* gene conferring clubroot resistance in *Brassica rapa*. *Plant Molecular Biology*, *80*, 621–629. https://doi.org/10.1007/s11103-012-9971-5

Varoquaux, N., Cole, B., Gao, C., Pierroz, G., Baker, C. R., Patel, D., Madera, M., Jeffers, T., Hollingsworth, J., Sievert, J., Yoshinaga, Y., Owiti, J. A., Singan, V. R., DeGraaf, S., Xu, L., Blow, M. J., Harrison, M. J., Visel, A., Jansson, C., …, Purdom, E. (2019). Transcriptomic analysis of field-droughted sorghum from seedling to maturity reveals biotic and metabolic responses. *Proceedings of the National Academy of Sciences*, *116*, 27124–27132. https://doi.org/10.1073/pnas.1907500116

Varshney, R. K., Bansal, K. C., Aggarwal, P. K., Datta, S. K., & Craufurd, P. Q. (2011). Agricultural biotechnology for crop improvement in a variable climate: Hope or hype? *Trends in Plant Science*, *16*, 363–371. https://doi.org/10.1016/j.tplants.2011.03.004

Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends in Plant Science*, *10*, 621–630. https://doi.org/10.1016/j.tplants.2005.10.004

Varshney, R. K., Nayak, S. N., May, G. D., & Jackson, S. A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, *27*, 522–530. https://doi.org/10.1016/j.tibtech.2009.05.006

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., Ye, Z., Shen, C., Li, J., Zhang, L., Zhou, X., Nie, X., Li, Z., Guo, K., Ma, Y., Huang, C., Jin, S., Zhu, L., Yang, X., … Zhang, X. (2017). Asymmetric subgenome selection and *cis*-regulatory divergence during cotton domestication. *Nature Genetics*, *49*, 579–587. https://doi.org/10.1038/ng.3807

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F., Mansueto, L., Copetti, D., Sanciangco, M., Palis, K. C., Xu, J., Sun, C., Fu, B., Zhang, H., Gao, Y., …, Leung, H. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, *557*, 43–49. https://doi.org/10.1038/s41586-018-0063-9

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., Gao, C., & Qiu, J. L. (2014). Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nature Biotechnology*, *32*, 947–951. https://doi.org/10.1038/nbt.2969

Wei, X., Xu, J., Guo, H., Jiang, L., Chen, S., Yu, C., Zhou, Z., Hu, P., Zhai, H., & Wan, J. (2010). *DTH8* suppresses flowering in rice, influencing plant height and yield potential simultaneously. *Plant Physiology*, *153*, 1747–1758. https://doi.org/10.1104/pp.110.156943

Westman, A. L., & Kresovich, S. (1997). Use of molecular marker techniques for description of plant genetic variation. In J. A. Callow, B. Ford-Lloyd, & H. J. Newbury (Eds.), *Biotechnology and plant genetic resources: Conservation and use* (pp. 9–48). CAB International

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*, 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Wittkopp, P. J., & Kalay, G. (2011). Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews. Genetics*, *13*, 59–69. https://doi.org/10.1038/nrg3095

Xu, P., Wang, X., Dai, S., Cui, X., Cao, X., Liu, Z., & Shen, J. (2021). The multilocular trait of rapeseed is ideal for high-yield breeding. *Plant Breeding*, *140*, 65–73. https://doi.org/10.1111/pbr.12880

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., …, Yan, J. (2019). Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nature Genetics*, *51*, 1052–1059. https://doi.org/10.1038/s41588-019-0427-6

Yang, Y., Zhu, K., Li, H., Han, S., Meng, Q., Khan, S. U., Fan, C., Xie, K., & Zhou, Y. (2018). Precise editing of *CLAVATA* genes in *Brassica napus* L. regulates multilocular silique development. *Plant Biotechnology Journal*, *16*, 1322–1335. https://doi.org/10.1111/pbi.12872

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 Biology*, *16*, 187. https://doi.org/10.1186/s13059-015-0757-3

Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y., & Kasahara, M. (2019). MoMI-G: Modular multi-scale integrated genome graph browser. *BMC Bioinformatics*, *20*, 548. https://doi.org/10.1186/s12859-019-3145-2

Zeng, Y., Wen, J., Zhao, W., Wang, Q., & Huang, W. (2019). Rational improvement of rice yield and cold tolerance by editing the three genes *OsPIN5b*, *GS3*, and *OsMYB30* with the CRISPR-Cas9 system. *Frontiers in Plant Science*, *10*, 1663. https://doi.org/10.3389/fpls.2019.01663

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhao, J., Bayer, P. E., Ruperao, P., Saxena, R. K., Khan, A. W., Golicz, A. A., Nguyen, H. T., Batley, J., Edwards, D., & Varshney, R. K. (2020). Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). *Plant Biotechnology Journal*, *18*, 1946–1954. https://doi.org/10.1111/pbi.13354

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., …, Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, *50*, 278–284. https://doi.org/10.1038/s41588-018-0041-z

Zhou, P., Silverstein, K. A. T., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., Farmer, A. D., Steele, K. P., Stupar, R. M., Miller, J. R., Tiffin, P., Mudge, J., & Young, N. D. (2017). Exploring structural variation and gene family architecture with *De Novo* assemblies of 15 *Medicago* genomes. *BMC Genomics*, *18*, 261. https://doi.org/10.1186/s12864-017-3654-1

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