the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) **113,** 446–459 doi: 10.1111/tpj.16070

FOCUSED REVIEW

Going broad and deep: sequencing-driven insights into plant physiology, evolution, and crop domestication

Songtao Gui[1], Felix Juan Martinez-Rivas[2], Weiwei Wen[1], Minghui Meng[1], Jianbing Yan[1], Björn Usadel[3,4,*] and Alisdair R. Fernie[2,*]

[1]*National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China,*
[2]*Max-Planck-Institute of Molecular Plant Physiology, Am Mühlenberg 1, Potsdam-Golm 14476, Germany,*
[3]*IBG-4 Bioinformatics, Forschungszentrum Jülich, Wilhelm Johnen Str, BioSc, 52428, Jülich, Germany, and*
[4]*Institute for Biological Data Science, CEPLAS, Heinrich Heine University, 40225, Düsseldorf, Germany*

Received 26 October 2022; revised 12 December 2022; accepted 13 December 2022; published online 19 December 2022.
*For correspondence (e-mail b.usadel@fz-juelich.de and fernie@mpimp-golm.mpg.de)

SUMMARY

Deep sequencing is a term that has become embedded in the plant genomic literature in recent years and with good reason. A torrent of (largely) high-quality genomic and transcriptomic data has been collected and most of this has been publicly released. Indeed, almost 1000 plant genomes have been reported (www.plabipd.de) and the 2000 Plant Transcriptomes Project has long been completed. The EarthBioGenome project will dwarf even these milestones. That said, massive progress in understanding plant physiology, evolution, and crop domestication has been made by sequencing broadly (across a species) as well as deeply (within a single individual). We will outline the current state of the art in genome and transcriptome sequencing before we briefly review the most visible of these broad approaches, namely genome-wide association and transcriptome-wide association studies, as well as the compilation of pangenomes. This will include both (i) the most commonly used methods reliant on single nucleotide polymorphisms and short InDels and (ii) more recent examples which consider structural variants. We will subsequently present case studies exemplifying how their application has brought insight into either plant physiology or evolution and crop domestication. Finally, we will provide conclusions and an outlook as to the perspective for the extension of such approaches to different species, tissues, and biological processes.

Keywords: sequencing, pangenome, single-cell, domestication, physiology.

Bullet-Point Summary

- Sequencing technology has driven the release of complex genomes, the emergence of population-scale and single-cell-scale multi-omics data, and the identification of more genetic variations in plant science.
- These broadly and deeply developed technologies provide opportunities to answer important evolutionary and physiological questions in plant science with respect to phenomena such as gene loss and gain during domestication and the convergent evolution of different plants.

Open Questions

- Efficient ways to construct unbiased reference pangenomes are needed
- Little is known about the roles of genetic variations and gene gain and loss during crop domestication.
- Combining prior knowledge, public databases, and omics data, we can go further into the plant regulatory networks.

SEQUENCING – THE STATE OF THE ART

Alongside CRISPR, next-generation sequencing has undoubtedly been the most prominent technology of 21st-century plant science. Its scope is vast and as we will describe below, it is difficult to think of a sub-discipline of plant science that has not been dramatically impacted by its application. Given that there are several excellent summaries of the underlying technologies (Dumschott et al., 2020; Metzker, 2010; Wenger et al., 2019), we will not detail these aspects here, but rather provide a brief overview of the novel insights that became achievable thanks to the adoption of so called next-generation sequencing as well as long-read sequencing technologies. For this purpose, arguable as good a starting point as any is the review by Weigel and Tautz, who stated 12 years ago in their visionary overview that genomic analysis would soon be applied to large numbers of species to answer important ecological and evolutionary questions (Tautz et al., 2010). The first sequenced plant genome was that of Arabidopsis in 2000 (The Arabidopsis Genome, 2000), rapidly followed by that of the first crop species rice (*Oryza sativa*) in 2002 (Goff et al., 2002), which both spurred basic and applied plant sciences. These genomes were still sequenced using classical Sanger sequencing, as were the few plant genomes that were sequenced in the years following. With the advent of next- or second-generation sequencing driven by Illumina and the long discontinued 454 sequencing technologies, plant genome analyses began to really take off, but were still not as prevalent as those in vertebrates. During evolution, plant genomes were heavily reshaped by polyploidy, several rounds of whole genome duplication, gene family expansion, and transposon amplification. This added complexity made short-read-based analysis of plant genomes more difficult than that of animals (Jiao & Schneeberger, 2017) and has potentially hampered the impact of next-generation sequencing in plants compared to animals. The rise of long-read sequencing technologies brought opportunities to overcome these barriers. Today, 1000 plant species genomes have been sequenced already (see https://plabipd.de/timeline_view.ep), where most new genomes have been analyzed using long sequencing technologies by Oxford Nanopores Technology (ONT) and/or Pacific Biosciences (PacBio) (Marks et al., 2021). This is because long-read sequencing has become much cheaper and at the same time more precise during the last few years, thereby allowing the relatively easy assembly of even complex plant genomes (Jiao & Schneeberger, 2017).

Historically, PacBio technology delivered long reads with error rates above 10% in 'continuous long read' (CLR) mode. Nevertheless, it was taken up rapidly in the field of plant genomics, wherein the longer reads for example allowed the assembly of the genome of the desiccation-tolerant grass *Oropetium thomaeum* with high quality (VanBuren et al., 2015) and provided insights into a new Arabidopsis ecotype besides the Arabidopsis reference genome (Zapata et al., 2016). That said, very recent genome publications still used this technology, including reports describing the genome of a tea tree (*Camellia sinensis* (L.) O. Kuntze) (Zhang et al., 2020), lychee (*Litchi chinensis*) (Hu et al., 2022), or the tomato relative *Solanum lycopersicoides* (Powell et al., 2022). PacBio has continuously evolved since the publication of the *Oropetium* genome, providing continuously increasing output, thus decreasing the cost per sequence.

Since the last few years, most PacBio sequencing data use circular consensus reads (CCS or HIFI), which rely on sequencing the same molecule multiple times and thus achieve very high base accuracy featuring an error rate of well below 1%. However, this technique results in a read length of typically around 20 kb (Hon et al., 2020). It can be argued that the high-quality long-read sequences as provided by the PacBio HIFI technique allow to assemble genomes with very little resources if homozygous plant genomes are to be sequenced, as bioinformatics tools like hifiasm (Cheng et al., 2021) make the whole genome assembly process easy, user-friendly, and computationally tractable without having to rely on high-performance compute clusters.

Similarly, the competing ONT keeps on evolving rapidly and can be accessed without any major capital investment, which is still the case for PacBio (Jain et al., 2018; Schmidt et al., 2017). Improved DNA extraction protocols (e.g., Vaillancourt & Buell, 2020, Vilanova et al., 2020) have helped to extract good-quality DNA more rapidly, and further throughput and accuracy improvements made it possible to obtain a near complete Arabidopsis genome sequence including centromeric regions (Naish et al., 2021).

Usually, the advantage of PacBio HIFI technology still remains the lower error rate, even though ONT has drastically improved recently. An early comparison of barley (*Hordeum vulgare*) assemblies in 2021 (Mascher et al., 2021) concluded that most long-read assemblies are superior to earlier short-read-based assemblies, but that PacBio HIFI performed best. They showed that this is particularly important for the analysis of disease resistance loci, which are difficult to assemble otherwise as they often contain clusters of the same gene families. Since this comparison (van Rengs et al., 2022) could show (i) that both technologies are somewhat complementary and (ii) how they could be combined to generate a gapless tomato (*Solanum lycopersicum*) genome assembly for a cultivar harboring tobacco mosaic virus (TMV) resistance. Rengs et al. used the assembly and the analysis of additional tomato varieties to show that the introgressed TMV resistance was accompanied by a large linkage drag region.

Since then, ONT has released a new 'Q20' chemistry to reduce the error rate. In addition, it is possible to get so-called duplex reads, where both the forward and reverse sequence of a molecule are combined, allowing the consensus sequence to achieve error rates of well below 1%. That said, typically the yield of these duplex reads is currently below 10% of the total (Sanderson et al., 2022). New algorithms have been developed to combine the long-accurate reads with the ultra-long reads and utilize haplotype-specific markers to assemble high-quality genomes (Rautiainen et al., 2022), and these have been added to hifiasm now as well. With the improvement of sequence accuracy and the development of new algorithms, before long will come an era of telomere-to-telomere plant genomes.

HOTSPOTS OF REARRANGEMENTS/COMPLEX GENOMES, PHASING, AND SINGLE-CELL SEQUENCING

The advent of these technologies and here especially PacBio HIFI sequencing has enabled to also determine individual haplotypes (Jiang et al., 2022). The aforementioned hifiasm provides this information by default, but necessitates high-quality HIFI reads to do so. An alternative is to use biological information in the form of haploid gametic cells. Sequencing of multiple gametes can be used to reveal crossover events (Li et al., 2015; Luo et al., 2019). Furthermore, by analyzing many different sperm cell 'genomes', Zhang et al. (2021a) and (2021b) were able to reconstruct the full haplotyped genome of the paternal plant (Zhang et al., 2021b). The authors of (Sun et al., 2022b) combined this technique with Hi-C data to reconstruct a tetraploid phased genome for potato (*Solanum tuberosum*) and showed that multiple regions were 'identical by descent' between haplotypes. Such techniques are often necessary for autopolyploid haplotype phasing as typical haplotype-based assemblers such as hifiasm were programmed for diploid cases. In the particular case of potato, given the large 'identical-by-descent' regions it is conceptionally not possible to resolve haplotypes in these regions without information spanning over these blocks. Hence, also longer ONT reads only allowed partial reference-based phasing using whatshap polyphase (Schrinner et al., 2020), but for (near) chromosome phasing, deep Hi-C data (Tang et al., 2022; Wang et al., 2022a) or offspring information (Schrinner et al., 2022; Serra Mari et al., 2022) is often available for agricultural species where necessary.

GWAS AND TWAS

Genome-wide association study (GWAS) has been widely used in human (Visscher et al., 2017), livestock (Gan et al., 2020; Uemoto et al., 2021), and crop populations (Huang et al., 2010; Zhou et al., 2015) to identify associations between genetic and phenotypic variations and analyze the genetic basis of complex traits, especially after the advent of next-generation sequencing. With the increasing number of sequence tags due to the advances in sequencing technology and the expansion of population size, new software and methods based on the mixed linear model (MLM) (Yu et al., 2006) have been developed. The compressed MLM decreases the effective sample size by clustering individuals into groups and reduces the computational effort (Zhang et al., 2010), the multi-locus mixed model (MLMM) method, which takes multi-locus effects into account, showed a better performance with respect to false discovery rate and effect power according to data in human and *Arabidopsis thaliana* (Segura et al., 2012), and the multi-trait mixed model (MTMM) method, considering the variation between and within traits, is suitable for the analysis of multiple traits (Korte et al., 2012). Despite these improvements, more than 90% of variants associated with human disease are located in non-protein-coding regions and far away from annotated genes (Maurano et al., 2012), implying the importance of uncovering the regulatory mechanisms underlying complex traits. Intermediate phenotypes, such as expression, can also integrate signals from changes in multiple components of a network. Recent technologies which permit the quantification of intermediate phenotypes like mRNA, metabolite, or protein abundance now enable mapping and trait dissection to be done between intermediate levels of biological organization, and new pipelines have been implemented for sample preparation and data normalization when performing GWAS with large populations (Bulut et al., 2021).

Transcriptome-wide association study (TWAS) using gene-based association methods such as PrediXcan and FUSION could detect known or novel genes associated with human disease (Gamazon et al., 2015; Gusev et al., 2016). This approach includes three main steps: (i) training a predictive model between variants and gene expression based on a known expression reference panel such as Genotype-Tissue Expression (GTEx) (Battle et al., 2017), (ii) using the trained model to predict the expression in individuals in the GWAS population, (iii) association analysis between expression and phenotypes (Wainberg et al., 2019). This approach has been widely used in the study of human diseases, including schizophrenia (Gandal et al., 2018), Parkinson's disease (Li et al., 2019), breast cancer (Feng et al., 2020), depression (Dall'Aglio et al., 2021), and anxiety (Su et al., 2021), due to the availability of comprehensive expression quantitative trait locus (eQTL) databases. For other species, large-scale transcriptome sequencing is an alternative approach to identify candidate genes. Li et al. (2020) sequenced fiber transcriptomes of 251 cotton (*Gossypium hirsutum*) accessions and identified 15 330 eQTLs, and 13 causal genes for differential fiber quality were prioritized through a TWAS of the local eQTL and GWAS data. Tang et al. (2021) sequenced transcriptomes of *Brassica napus* seed at two developmental stages (309 accessions and 274 accessions,

respectively); as a result, 605 and 148 genes were detected to be related to seed oil content according to the association analysis between gene expression and seed oil content, and the sequenced data were also applied to the association of seed glucosinolate content in *B. napus* (Tan et al., 2022). Although TWAS has its advantages in identifying trait-associated genes compared with GWAS (see also Wainberg et al., 2019) and methods like FOCUS and MR-JTI have been proven to be effective in improving the resolution for causal gene mapping (Mancuso et al., 2019; Zhou et al., 2020), its application in complex traits research of animals and plants is not so extensive as the latter. Further efforts such as building a species genotype–tissue expression database and development of tools for multi-omics analysis would make full use of the rapidly expanding biological sequencing data.

SINGLE-CELL TRANSCRIPTOMICS

Transcriptomics analysis has been the golden standard to investigate how organisms respond to developmental and environmental cues at the global transcriptional level, which contributes to our understanding of spatiotemporally regulated transcriptional programs in organisms (Ozsolak & Milos, 2011; Wang et al., 2009). Early transcriptomics analyses were mainly applied to parts of tissues or organs, or even entire organisms, which can only generate average cell data and lose information about cell heterogeneity (Gutjahr et al., 2015; O'Connell et al., 2012). However, different cell types have biologically distinct roles in development and environmental adaptation (Hong et al., 2017; Libault et al., 2017; Misra et al., 2014), and compared with traditional sequencing, single-cell RNA sequencing (scRNA-seq) methodologies have overcome the problem of averaging gene expression levels across whole tissues, enabling the identification of individual cells at high resolution, the discovery of novel cell types, and more accurate and integrated understanding of their roles in life processes (Kolodziejczyk et al., 2015).

The scRNA-seq workflow includes dissociation of target cells from the tissue, cell isolation, RNA extraction, cDNA synthesis by reverse transcription, and single-cell sequencing, followed by bioinformatics analyses, including expression matrix construction, cell type identification, and cell cluster annotation (Bawa et al., 2022). Early attempts to increase the spatiotemporal solution with techniques such as laser capture microdissection (LCM) (Asano et al., 2002; Cai & Lashbrook, 2006; Tomlins et al., 2007), capturing cells with fluorescence-activated cell sorting (FACS) (Birnbaum et al., 2003; Brady et al., 2007), or isolation of nuclei tagged in individual cell types (INTACT) (Deal & Henikoff, 2011) could only analyze hundreds of cells at very high resolution. Most recently, the droplet-based method is commonly used for the isolation of massive numbers of cells in scRNA-seq analyses (Macosko et al., 2015). Commercial microfluidics technologies such as the 10x Genomics platform can increase the throughput to nearly 10 000 cells per run (Seyfferth et al., 2021).

Compared to animals, the development of plant scRNA-seq has been hampered by the diversification of plant samples and the difficulty of preparing single-cell suspensions. Various factors can influence the efficiency of protoplast isolation, such as cell wall composition, size of cells, and isolation methods; therefore, scRNA-seq datasets often do not truly reflect the original cell populations (Denyer et al., 2019; Shulse et al., 2019; Valihrach et al., 2018). But the development of new technologies such as microwell-based systems and single-nuclei RNA-seq may be able to mitigate these issues to some extent (Ding et al., 2019; Lake et al., 2019). So far, the most profiled tissue by scRNA-seq is the Arabidopsis primary root tip (Apelt et al., 2022; Gala et al., 2021; Jean-Baptiste et al., 2019; Zhang et al., 2019). These studies generated a spatiotemporal gene expression atlas with specific cell types and enabled the reconstruction of a continuous differentiation trajectory of root development. For example, the earliest steps of lateral root development only occur in a small number of cells, and scRNA-seq captured the transcriptomes of the xylem pole pericycle cells where lateral roots originate and discovered many upregulated target genes associated with this process (Gala et al., 2021). Apelt et al. (2022) conducted scRNA-seq analysis and found differences between root and aboveground tissues. Depending on the time of day, alterations in RNA levels occur in distinct tissues to various degrees. MERCY1 was found to be a marker of dividing cells, suggesting its function in meristematic development.

In addition to the applications in Arabidopsis primary root tips, scRNA-seq has been applied to study other dicots and monocots, as well as in leaves, inflorescences, and other tissues (Sun et al., 2022a; Zong et al., 2022). More recently, it was found that CYCLING DOF FACTOR 5 (CDF5) and REPRESSOR OF GA (RGA) have potential roles in the early development and function of leaf veins in cotyledons (Liu et al., 2022b). Sun et al. (2022a) identified cell types in mature and developing stomata of maize (*Zea mays*) epidermis-enriched tissue and found that guard cells (GCs) and subsidiary cells (SCs) displayed differential expression of genes, which, besides those encoding transporters, were involved in the abscisic acid, CO_2, Ca^{2+}, starch metabolism, and blue light signaling pathways.

Many studies on stress resistance and adaptation to the environment also rely on scRNA-seq. The comprehensive definition of cell types by scRNA-seq helps us understand how root cells differentiate and how root system architecture is shaped in response to environmental cues, such as drought or flooding (Zhang et al., 2021a). By generating and exploiting a high-resolution single-cell gene expression atlas of Arabidopsis roots, Wendrich

et al. (2020) found that the TMO5/LHW heterodimer triggers biosynthesis of mobile cytokinin in vascular cells and increases the root hair density under low-phosphate conditions by modifying both the length and the cell fate of epidermal cells. Bai et al. (2022) using scRNA-seq obtained a pseudotime trajectory revealing mechanisms underlying the transition from normal functioning to the defense response in epidermal and mesophyll cells upon *Botrytis cinerea* infection.

In addition to scRNA-seq, other single-cell omics technologies, such as scATAC-seq, have been developed. Recently, Marand et al. (2021) built a single-cell *cis*-regulatory atlas in maize leveraging scATAC-seq data from six maize organs and provided a software program, Socrates, to help better understand the single-cell *cis*-regulatory variation.

PLANT PANGENOMES

Pangenomics, as first proposed in microbiology (Vernikos et al., 2015), has now been increasingly seen application to eukaryotic genomes for its abilities to reduce the reference bias and provide additional information of the species (Eizenga et al., 2020). With the ability to acquire high-quality genome sequences for individuals of a given plant species or clade, the past few years have witnessed a rapid growth of plant pangenomics-related works. Since the detailed methodologies for linear pangenome construction (Golicz et al., 2016a), pangenome graphs (Eizenga et al., 2020), and the advances of plant pangenomics until 2021 (Lei et al., 2021) have already been well reviewed, here we will focus on the newly published plant pangenomes (Table 1), updating the work previously reported by (Lei et al., 2021).

With the explosion of sequence data, public resources available online allow researchers to construct pangenomes with a reanalysis of previously published genomes, which enhances the possibilities of generating new pangenomes in the future. A typical example is the sorghum (*Sorghum bicolor*) pangenome, which was constructed using 176 genomic sequences (with a minimum coverage of 10×) available in the databases (Ruperao et al., 2021). This pangenome was constructed following the iterative assembly strategy used for *Brassica* (Golicz et al., 2016b). Each iteration added on average 1.9 Mb of new sequence to the genome, resulting in a 174.5 Mb (approximately 20%) increase compared to the reference genome. Of the genes identified, 47% were 'core genes', or genes found in every accession sequenced. Of the genes that were newly identified, 79 were drought-related, as their expression was increased under drought conditions, making them potential subjects for further characterization. Up to 91 000 new single nucleotide polymorphisms (SNPs) were identified; some of them were identified as trait-related SNPs, which enhances the use of the pangenome for further identification of candidate genes in sorghum improvement.

Two different pangenomes for soybean (*Glycine max*) have been published recently (Bayer et al., 2022; Torkamaneh et al., 2021). Torkamaneh et al., 2021 used previously reported sequencing data, selecting 204 different accessions with a sequencing depth of ≥15×. Their study added 108 Mb to the reference genome used (cv. Williams 82), with 1659 new genes. Of the identified genes, 90% were found in all the accessions sequenced, a number much higher than for the sorghum pangenome. Nonetheless, many non-common genes were related to defense responses and plant development, according to Gene Ontology (GO) enrichment analysis, which makes them potential subjects for further study. Bayer et al., 2022 reported the sequencing data of 1110 different accessions (886 newly sequenced) with a minimum depth of 8.5× of the genome. Using the reference genome of cv. Lee, 198.4 Mb was added to the pangenome, with 3765 new genes (Bayer et al., 2022). Of the identified genes, 86.8% were found in all the sequenced accessions, a similar number previously reported by Torkamaneh et al., 2021. The GO terms associated with the variable genes are also similar to the ones reported in previous work, which supports the idea that different accessions have evolved to cope with certain environmental factors and to survive in these environments. Among the sequenced accessions, Bayer et al., 2022 included wild *Glycine soja* plants, different landraces, and commercial cultivars; they found that during domestication, up to 1.5% of genes were lost. Moreover, different gene frequencies were reported among the population screened. The frequency of genes related to defense and salt responses was decreased, while the frequency of genes related to flowering time and seed composition was increased. This reveals that during domestication, growers selected those lines with shorter flowering time and enhanced pod traits rather than disease-resistant plants.

Using PacBio sequencing, Li et al., 2022 constructed the cucumber (*Cucumis sativus*) pangenome. As discussed above, PacBio allows to obtain longer sequences, resulting in more reliable sequence results. The 11 cucumber accessions sequenced yielded a genome coverage higher than 45×. In the pangenome constructed, around 80% of the identified genes were core genes, and the variable genes were shorter and their expression was lower than that of the core genes (Li et al., 2022). The GO terms of the variable genes were related to the auxin response and cell proliferation. Over 2.5 million SNPs and 1.3 million small insertions/deletions (InDels) were also identified, of which 2.5% and 1.5%, respectively, induced changes in protein sequences as amino acid changes or premature stop codons. A deep study of these InDels identified a 51-nt deletion in the *CsTu* gene, resulting in warted fruits. This deletion leads to a loss of DNA binding ability of the

Table 1 Plant pangenomes reported. This report continues the work previously reported in Lei et al. (2021)

Species	Year	Individual genomes	Reference genome size (Mb)	Technology	Pangenome construction approach	Number of genes in reference[a]	Genes identified in pangenome[b]	Non-reference genes identified	Pangenome/reference	Database website	Reference
Cajanus cajan	2020	89	606	Illumina	Iterative mapping and assembly	53 612	55 512	1900	1.03×	https://research-repository.uwa.edu.au/en/datasets/pigeon-pea-pangenome-contig-assembly-annotation-snps-pav	(Zhao et al., 2020)
Hordeum vulgare	2020	20	478	Illumina	Single copy pangenome	37 848	37 515			https://barley-pangenome.ipk-gatersleben.de	(Jayakodi et al., 2020)
Brassica napus	2020	1688	850	PacBio-Illumina	PVs + map to pan	80 382	101 402	21 020	1.26×	http://cbi.hzau.edu.cn/bnapus	(Song et al., 2021)
Gossypium hirsutum	2021	1581	2347	PacBio-Illumina	Reference-guided assembly approach	70 199	102 768	32 569	1.46×		(Li et al., 2021)
Gossypium barbadense	2021	226	2266	PacBio-Illumina	Reference-guided assembly approach	71 297	80 148	8851	1.12×		(Li et al., 2021)
Brassica genomes *A*	2021	[c]	284	Illumina		45 819	57 558	11 739	1.25×		(He et al., 2021)
B			570	Illumina		59 422	63 630	4208	1.07×		
C			488	Illumina		60 457	76 277	15 820	1.26×		
Oryza sativa, *O. glaberrima*	2021	33 (32 *O. sativa*, 1 *O. glaberrima*)	336	PacBio	Gene base pangenome	35 596	66 636	31 040	1.87×	https://www.ricerc.com	(Qin et al., 2021)
Musaceae species	2021	15 (12 *Musa* species, 3 *Ensete*).	451 (*M. acuminata*)	Illumina	Iterative mapping and assembly	35 276	47 586	12 310	1.35×		(Rijzaani et al., 2022)
Glycine max	2021	204	978	Illumina	Map to pan	52 872	54 531	1659	1.03×		(Torkamaneh et al., 2021)
Cicer arietinum	2021	3366	738	Illumina	Targeted *de novo* assembly	28 269	29 870	1601	1.05×		(Varshney et al., 2021)
Lupinus albus	2021	39	451	Illumina	Map to pan	38 258	38 466	178	1.01×	www.whitelupin.fr	(Hufnagel et al., 2021)
Raphanus spp.	2021	11	513[d]	PacBio-Illumina	Graph-based *de novo* assembly	47 305	45 635				(Cheng et al., 2021)

(continued)

The Plant Journal, (2023), **113**, 446–459

Table 1. (continued)

Species	Year	Individual genomes	Reference genome size (Mb)	Technology	Pangenome construction approach	Number of genes in reference[a]	Genes identified in pangenome[b]	Non-reference genes identified	Pangenome/reference	Database website	Reference
Sorghum bicolor	2021	354	729		Iterative mapping and assembly	35 467	35 719				(Ruperao et al., 2021)
Fragaria spp.	2021	5	240	PacBio-Nanopore		28 588	25 295				(Qiao et al., 2021)
Solanum tuberosum (tetraploid)	2022	6	741[e]	Illumina-Nanopore		32 917[e]	89 187[f]	56 270	2.70×		(Hoopes et al., 2022)
Cucumis sativus	2022	11	225	PacBio	Graph-based *de novo* assembly	24 714	25 128	414	1.02×		(Li et al., 2022)
Oryza	2022	251	336	Nanopore	Graph-based	37 864	51 359	13 495	4.10×	http://www.ricesuperpir.com/	(Shang et al., 2022)
Zea spp.	2022	721	2191	Illumina	Graph-based	39 591	58 944	19 353	1.48×		(Gui et al., 2022)
Lupin angustifolius	2022	55	653	Illumina	Iterative mapping and assembly	38 545	39 339	794	1.02×	http://lupinexpress.org/	(Garg et al., 2022)
Vigna radiata	2022	217	475	Illumina	Map to pan	40 125	43 462	3337	1.08×		(Liu et al., 2022a)
Cucumis melo	2022	297	375	Illumina	PVs + Map to pan	29 980	34 305	4325	1.14×		(Sun et al., 2022c)
Solanum tuberosum (*petota*)	2022	44	741	PacBio	Graph-based	32 917[e]	71 525	38 608	2.17×	http://solomics.agis.org.cn/potato/	(Tang et al., 2022)
Solanum lycopersicum	2022	32	782	PacBio	Graph-based	36 648	51 155	14 507	1.39×	http://solomics.agis.org.cn/tomato/	(Zhou et al., 2022)

Shaded rows refer to work without newly sequenced data, but using previously sequenced reports.
[a]Number of genes found in the reference genome of the species sequenced.
[b]Average number of genes identified in the pangenome.
[c]Used different published pangenomes and polyploid sequence genomes to extract each subgenome.
[d]As a mean of the previously available *Raphanus* genomes.
[e]From diploid species.
[f]From tetraploid species.

transcription factor, thus resulting in a non-active form. Newly identified InDels related to flowering locus T were associated with another important trait, flowering time. This information may be useful in cucumber domestication, as early flowering rather than late flowering accessions were selected, similar to the situation in soybean.

Recently, the pangenomes of two species with high economical value, tomato and potato, have been published (Tang et al., 2022; Zhou et al., 2022). For tomato, the authors constructed a new reference genome, adding almost 20 Mb of new sequence and eliminating almost 90% of the sequencing gaps from the previous version. This, added to the higher quality obtained by PacBio, allowed the identification of new polymorphisms with a dataset containing over 17 million SNPs. In addition, long-read sequencing of additional genomes and the incorporation of data of earlier ONT long-read genomes (Alonge et al., 2020) allowed in-depth identification of structural variants (SVs) whose incorporation into the genetic framework markedly increased the estimated heritability. The authors also identified genes related to the soluble solid contents and to flavor traits, which confirms that the identification of new genome variants is a helpful tool for tomato breeding (Zhou et al., 2022). The potato pangenome permitted researchers to identify genes related to tuber development and to understand its appearance during potato domestication. Also, they determined the level of homozygosity, which will help to select lines for diploid hybrid breeding to genetically improve this crop (Tang et al., 2022).

Most of the pangenomic studies focused on the pangenome of a species. Recently, the 'super-pangenome' (Khan et al., 2020), a broader level of pangenome that represents the genome of a genus, has stood out in crops for better leveraging the wild relatives in genomics-assisted breeding. The super-pangenomes of two main crops, rice and maize, have been released in 2022 (Gui et al., 2022; Shang et al., 2022). For the rice super-pangenome, the authors have selected 251 rice accessions representing the genetic and phenotypic diversity of cultivated and wild rice germplasm, and the super-pangenome was constructed based on the ONT sequences of these 251 accessions. The resulting super-pangenome consisted of 1.52 Gb non-redundant sequences, including 1.15 Gb non-reference sequences. The super-pangenome harbors 51 359 genes, with 42.62% core genes and 57.38% variable genes. Based on the super-pangenome, a total of 193 880 SVs were identified. Further analyses have highlighted the important effects of SVs on important agronomic traits (thousand-grain weight and grain length) through altering gene expression levels. For the *Zea* super-pangenome, the authors have constructed a pan-*Zea* genome of approximately 6.71 Gb using publicly available maize genome assemblies and *de novo* fragmental assemblies of 721 pan-*Zea* individuals (Gui et al., 2022). The authors have highlighted the potential value of this pan-*Zea* genome in maize breeding (introduced in more detail in the case studies section below). Moreover, the sequencing of 744 genomes from maize and all wild taxa of the *Zea* genus revealed over 70 million SNPs, underlining the importance of studying variation in wild relatives to identify genes that are important in crops (Chen et al., 2022a).

The construction of plant pangenomes allows not only to identify new genes that might serve as targets for breeding or functional studies, but also to reveal SVs in genes already selected in breeding that help to understand the domestication of plants. However, most of the current pangenomic studies focused on gene presence/absence variants and canonical SVs. Besides, while the pangenomes of most major crops have been studied, less effort has been made to construct the pangenomes of orphan crops despite their importance for dryland agriculture. Recently, several African orphan crops have had their genome sequenced (reviewed in Ghazal et al., 2021), which provides opportunities to perform pangenomic analyses to further understand the domestication history and support genome-assisted breeding of these African orphan crops. With the development of more whole genome comparative algorithms (Kille et al., 2022), graph-genome-based mapping (Siren et al., 2021), and downstream analysis tools (Liao et al., 2022), we could soon be able to get accurate multi-alignment results of whole genomes, identify more complex sequence variations such as transposable elements (TEs) and nested SVs, and use pan-reference genomes to update traditional bioinformatics pipelines for different omics (for the further application of graph pangenomes in crop improvement, please also refer to Hameed et al., 2022, Wang et al., 2022b, Zanini et al., 2022). By then, pangenomic analysis could help us better understand the evolution of different plants, the origin of new genes, and the molecular basis of complex phenotype variations.

INSIGHT INTO STRUCTURAL VARIANTS

Various types of SVs, including transposons, constitute the majority of the genomic differences among plant species. Two breakthrough studies in tomato recently explored the effects of retrotransposons and other SVs on tomato phenotypes of agricultural importance (Alonge et al., 2020; Domínguez et al., 2020). Despite its narrow genetic basis, tomato (*S. lycopersicum*) exhibits wide phenotypic diversity in metabolic and developmental traits with much of this diversity being historically ascribed to the selection of rare alleles with large effects (Gao et al., 2019). However, this viewpoint is challenged by the fact that GWAS tends to focus only on SNPs and short InDels (for a notable exception to this statement, see Akakpo et al., 2020), although pangenome analyses have revealed that SVs

account for a larger proportion of sequence differences in this species. The approach taken in both papers was essentially the same that of Dominguez et al. (2020), who relied on available resequencing data from over 600 cultivated and wild accessions (Zhu et al., 2018), whereas Alonge et al. additionally generated long-read data for 100 tomato accessions. These data allowed the establishment of a pan-SV genome with any individual accession harboring between 1928 and 45 840 SVs (with the wild species being the most highly divergent). These SVs were largely composed of, or generated by, transposons (Lisch, 2013), the function of which, whilst well characterized at the molecular level (Chuong et al., 2017), remains somewhat controversial on a wider genomic scale. To address this, Dominguez et al. (2020) assessed the set of TE families with recent mobilization activity and revealed that the majority of transposon insertion polymorphisms resulted from the mobilization of COPIA-like retrotransposons. Intriguingly, the mobilome fraction was much greater in wild tomato relatives and early domesticates than in established cultivars as a result of significant gene flow between these groups. Further analysis revealed that COPIA and many other TE families were found preferentially in or near genes, whilst GYPSY TEs mainly cluster in the pericentromeric regions. Moreover, COPIA TEs integrate within environmentally responsive genes (Quadrana et al., 2019), as was identified as being conspicuous in the *Solanum pennellii* genome (Bolger et al., 2014).

CASE STUDIES ON EVOLUTION AND DOMESTICATION

Similarly, deep and broad sequencing was instrumental in demonstrating the convergent selection of a WD40 transcriptional regulator which enhances grain yield in both maize and rice (Chen et al., 2022b). Indeed, knockout of *KRN2* in maize or *OsKRN2* in rice increased grain yield by approximately 10% and approximately 8%, respectively, with no apparent trade-offs in other agronomic traits (Figure 1). Beyond this, genome-wide scans relying on the sequencing data of approximately 600 maize and 170 rice genotypes identified a total of 490 pairs of orthologous genes that underwent convergent selection during maize and rice evolution. These genes were considerably enriched for two shared molecular pathways (starch and sucrose metabolism, biosynthesis of cofactors), suggesting these pathways to be an excellent target for future cereal crop improvement. That said, the recent finding of a role for a WD40 transcription factor in the coordination of tomato fruit ripening (Zhu et al., 2022) suggests that these results may have even broader importance in this respect. Another example of the application of broad sequencing in maize is that carried out to better understand the unilateral cross-incompatibility (UCI) that occurs between popcorn (*Z. mays* var. *everta*) and dent corn (*Z. mays* var. *indentata*) and is associated with the Gametophyte factor1 (Ga1) locus. However, the underlying genetic basis has remained unclear for decades (Wang et al., 2022c). In this study a Z58 × SK recombinant inbred line (RIL) population was developed and genotyped using high-density markers. A segregation distortion locus where the genotypes of the two parents significantly deviated from the expected 1:1 ratio was detected on chromosome 4, which overlapped with the defined Ga1 locus in previous reports (Zhang et al., 2012; Zhang et al., 2018). Approximately 5000 individuals from the Z58 × SK F2 population were subjected to sequencing in order to fine-map the Ga1 locus using genotypic segregation distortion as a phenotype. Two components which influenced segregation distortion were narrowed to a small region. Following this approach, seven linked genes were identified, encoding three types of proteins that affect UCI. These include five pollen-expressed PECTIN METHYLESTERASE (PME) genes (ZmPMEs-m), one silk-expressed PME gene (ZmPME3), and one silk-expressed gene encoding a cysteine-rich protein (ZmPRP3). Whereas ZmPMEs-m confer pollen compatibility, ZmPME3 causes silk reject incompatibility, while ZmAPG1 promotes pollen tube growth and thereby breaks the inhibitory effect of ZmPME3. The three types of genes independently regulate the growth of pollen tubes but with both antagonistic and synergistic effects. This deepens our understanding of the complex regulation of cross-incompatibility. The near isogenic lines (NILs) presented variation in *ZmBAM1d*, a gene associated to a yield QTL, identified due to the high sequencing quality of the SK line (Yang et al., 2019). This thus represents an excellent case study exemplifying the power of broad sequencing to dissect hitherto recalcitrant genomic regions.

Whilst the above two examples clearly show the role of broad sequencing in addressing targeted questions, two further studies that look at a much more global level have also recently been published that are worth discussing. In the first of these, Chen and co-workers characterized a high-density genomic variation map from approximately 700 genomes encompassing maize and all wild taxa of the genus *Zea*, identifying over 65 million SNPs, 8 million InDel polymorphisms, and over 1000 novel inversions (Figure 1). This variation map revealed evidence of selection within taxa displaying novel adaptations such as perenniality and regrowth. However, most striking in this dataset was the evidence of convergent adaptation in highland teosinte and temperate maize. Indeed, this study not only indicated the key role of hormone-related pathways in highland adaptation and flowering time-related pathways in high-latitude adaptation, but also identified significant overlap in the genes underlying adaptation to both environments. In order to illustrate how these data can be used to identify useful genetic variants, the authors subsequently generated and characterized novel mutant alleles for two flowering time

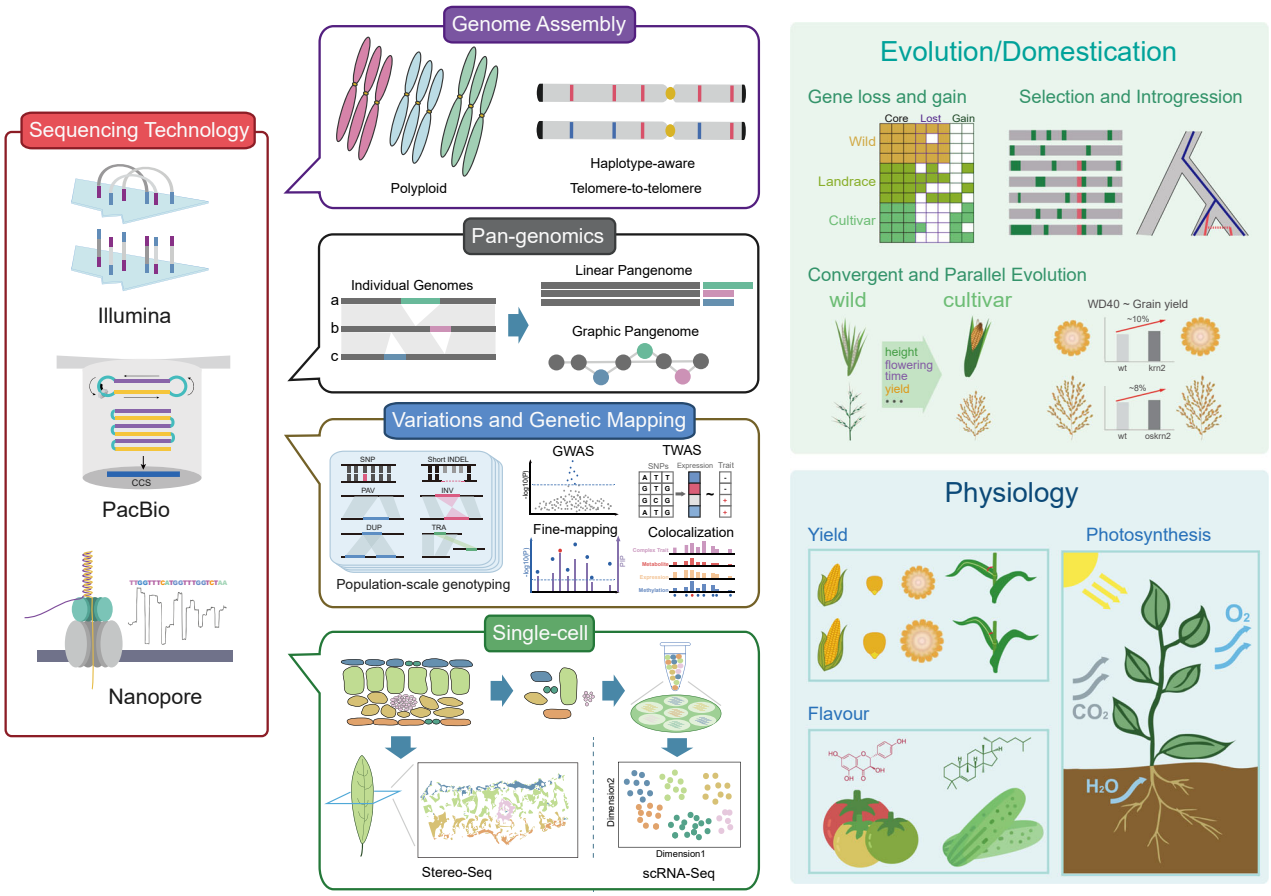

Figure 1. Schematic representation of how sequencing technology has driven the study of plant physiology and evolution. The development of sequencing technologies has promoted the generation of different omics data in plants, including high-quality genome assemblies, pangenomes, population-scale genetic variations, and single-cell transcriptomes. These omics data have helped in studying variation during plant evolution and domestication and in uncovering the molecular mechanisms underlying important plant physiological features.

candidate genes. To summarize, this work provides the most extensive sampling to date of the genetic diversity inherent in the genus *Zea*, resolving questions on evolution and identifying adaptive variants for direct use in modern breeding. In a sister paper, Gui and co-workers constructed both a super-pangenome for the *Zea* genus and a comprehensive SV map for maize breeding. As discussed above, this generated an approximately 6.71-Gb pan-*Zea* genome containing approximately 4.57 Gb of non-B73 reference sequences from fragmented *de novo* assemblies of 721 pan-*Zea* individuals. A total of 58 944 pan-*Zea* genes were annotated, and approximately 44.34% of them were dispensable in the pan-*Zea* population. Moreover, 255 821 common SVs were identified and genotyped in a maize association mapping panel. Further analyses revealed the gene presence/absence variants and their potential roles during domestication of maize. Similar to earlier observations in tomato (Alonge et al., 2020; Alseekh et al., 2020; Domínguez et al., 2020), combining genetic analyses with multi-omics data, Gui and co-authors demonstrated how SVs are associated with complex agronomic traits. As such, their results highlight the underexplored role of the pan-*Zea* genome and SVs to further understand domestication of maize and explore their potential utilization in crop improvement.

CONCLUSIONS AND PERSPECTIVES

In summary, improvements in sequencing technology have made it cost-effective to generate high-quality genome assemblies and population-scale multi-omics data in plants. These new resources, along with the development of statistical and bioinformatics methods, have not only brought great opportunities to answer broad evolutionary questions with respect to, for example, the evolution and domestication history of plants, but also help in digging deeper into the molecular regulatory mechanisms of important plant physiological features and promote the application of genomics-assisted breeding.

With great opportunities come great challenges. One technical issue the community must face is how to make

sense of these multi-omics data efficiently. For pangenomics, the unbiased representation of graph pangenomes is still under rapid development (Garrison & Guarracino, 2023); pangenome-based bioinformatics analyses are still not as effective as the canonical linear reference genome-based ones. For population genetics, we still need new bioinformatics algorithms, new causal inference strategies, and larger-scale omics data to reveal the potential roles of rare and complex genetic variations. For single-cell technologies, solutions to reduce the sequencing noise, to achieve higher throughputs, and to make improvements to algorithms to deal with missing values are needed, as well as methods to isolate single cells from more plant tissues. Additionally, we still need efficient ways to integrate multi-omics data to draw more comprehensive pictures of the regulatory networks of plant physiology.

The rapid growth of biological Big Data has triggered a trend of combining the data-driven 'Kepler paradigm' with the rationale-driven 'Newtonian paradigm' in biology studies. By combining sequencing technology with genetic engineering technology, we are able to go broader in the study of patterns of large evolutionary problems and deeper in the study of the detailed underlying molecular mechanisms. With the further application of technologies such as complex network analysis and deep learning, with the increasing availability of data at different scales such as population-scale and single-cell sequencing data (Yazar et al., 2022) in plant sciences, it will not be too long before we go broad and deep simultaneously to uncover the detailed genetic regulatory mechanisms underlying variation in plant physiology during domestication and more precisely design, breed, and improve crops.

ACKNOWLEDGMENTS

FJMR, BU, and ARF acknowledge the Deutsche Forschungsgemeinschaft for projects 452682775, FE 552/39-1, and US 98/23-1; WW, BU, and ARF acknowledge the Deutsche Forschungsgemeinschaft for projects 468870408, US 98/25-1, and FE 552/40-1. Open Access funding enabled and organized by Projekt DEAL.

REFERENCES

Akakpo, R., Carpentier, M.C., Ie Hsing, Y. & Panaud, O. (2020) The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytologist*, **226**, 44–49.

Alonge, M., Wang, X.G., Benoit, M., Soyk, S., Pereira, L., Zhang, L. *et al.* (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, **182**, 145.

Alseekh, S., Scossa, F. & Fernie, A.R. (2020) Mobile transposable elements shape plant Genome diversity. *Trends in Plant Science*, **25**, 1062–1064.

Apelt, F., Mavrothalassiti, E., Gupta, S., Machin, F., Olas, J.J., Annunziata, M.G. *et al.* (2022) Shoot and root single cell sequencing reveals tissue- and daytime-specific transcriptome profiles. *Plant Physiology*, **188**, 861–878.

Asano, T., Masumura, T., Kusano, H., Kikuchi, S., Kurita, A., Shimada, H. *et al.* (2002) Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: toward comprehensive analysis of the genes expressed in the rice phloem. *Plant Journal*, **32**, 401–408.

Bai, Y., Liu, H., Lyu, H., Su, L., Xiong, J. & Cheng, Z.-M.M. (2022) Development of a single-cell atlas for woodland strawberry (Fragaria vesca) leaves during early Botrytis cinerea infection using single-cell RNA-seq. *Horticulture Research*, **9**, uhab055.

Battle, A., Brown, C.D., Engelhardt, B.E. & Montgomery, S.B. (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

Bawa, G., Liu, Z., Yu, X., Qin, A. & Sun, X. (2022) Single-cell RNA sequencing for plant research: insights and possible benefits. *International Journal of Molecular Sciences*, **23**, 4497.

Bayer, P.E., Valliyodan, B., Hu, H., Marsh, J.I., Yuan, Y., Vuong, T.D. *et al.* (2022) Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. *Plant Genome*, **15**, e20109.

Birnbaum, K., Shasha, D.E., Wang, J.Y., Jung, J.W., Lambert, G.M., Galbraith, D.W. *et al.* (2003) A gene expression map of the Arabidopsis root. *Science*, **302**, 1956–1960.

Bolger, A., Scossa, F., Bolger, M.E., Lanz, C., Maumus, F., Tohge, T. *et al.* (2014) The genome of the stress-tolerant wild tomato species Solanum pennellii. *Nature Genetics*, **46**, 1034–1038.

Brady, S.M., Orlando, D.A., Lee, J.-Y., Wang, J.Y., Koch, J., Dinneny, J.R. *et al.* (2007) A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, **318**, 801–806.

Bulut, M., Fernie, A.R. & Alseekh, S. (2021) Large-scale multi-omics genome-wide association studies (Mo-GWAS): guidelines for sample preparation and normalization. *Journal of Visualized Experiments*. https://doi.org/10.1093/plphys/kiac593

Cai, S. & Lashbrook, C.C. (2006) Laser capture microdissection of plant cells from tape-transferred paraffin sections promotes recovery of structurally intact RNA for global gene profiling. *Plant Journal*, **48**, 628–637.

Chen, L., Luo, J., Jin, M., Yang, N., Liu, X., Peng, Y. *et al.* (2022a) Genome sequencing reveals evidence of adaptive variation in the genus Zea. *Nature Genetics*, **54**, 1736–1745.

Chen, W., Chen, L., Zhang, X., Yang, N., Guo, J., Wang, M. *et al.* (2022b) Convergent selection of a WD40 protein that enhances grain yield in maize and rice. *Science*, **375**, eabg7985.

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

Chuong, E.B., Elde, N.C. & Feschotte, C. (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews. Genetics*, **18**, 71–86.

Dall'Aglio, L., Lewis, C.M. & Pain, O. (2021) Delineating the genetic component of gene expression in major depression. *Biological Psychiatry*, **89**, 627–636.

Deal, R.B. & Henikoff, S. (2011) The INTACT method for cell type–specific gene expression and chromatin profiling in Arabidopsis thaliana. *Nature Protocols*, **6**, 56–68.

Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K. & Timmermans, M.C.P. (2019) Spatiotemporal developmental trajectories in the arabidopsis root revealed using high-throughput single-cell RNA sequencing. *Developmental Cell*, **48**, 840–852.e5.

Ding, J., Adiconis, X., Simmons, S.K., Kowalczyk, M.S., Hession, C.C., Marjanovic, N.D., *et al.* (2019) Systematic comparative analysis of single cell RNA-sequencing methods: genomics.

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J.M., Colot, V. *et al.* (2020) The impact of transposable elements on tomato diversity. *Nature Communications*, **11**, 4058.

Dumschott, K., Schmidt, M.H., Chawla, H.S., Snowdon, R. & Usadel, B. (2020) Oxford nanopore sequencing: new opportunities for plant genomics? *Journal of Experimental Botany*, **71**, 5313–5322.

Eizenga, J.M., Novak, A.M., Sibbesen, J.A., Heumos, S., Ghaffaari, A., Hickey, G. *et al.* (2020) Pangenome Graphs. *Annual Review of Genomics and Human Genetics*, **21**, 139–162.

Feng, H., Gusev, A., Pasaniuc, B., Wu, L., Long, J., Abu-Full, Z. *et al.* (2020) Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. *Genetic Epidemiology*, **44**, 442–468.

Gala, H.P., Lanctot, A., Jean-Baptiste, K., Guiziou, S., Chu, J.C., Zemke, J.E. *et al.* (2021) A single-cell view of the transcriptome during lateral root initiation in Arabidopsis thaliana. *The Plant Cell*, **33**, 2197–2220.

Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J. *et al.* (2015) A gene-based association method for

mapping traits using reference transcriptome data. *Nature Genetics*, **47**, 1091–1098.

Gan, Q., Li, Y., Liu, Q., Lund, M., Su, G. & Liang, X. (2020) Genome-wide association studies for the concentrations of insulin, triiodothyronine, and thyroxine in Chinese Holstein cattle. *Tropical Animal Health and Production*, **52**, 1655–1660.

Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S. *et al.* (2018) Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*, **362**, eaat8127. https://doi.org/10.1126/science.aat8127

Gao, L., Gonda, I., Sun, H.H., Ma, Q.Y., Bao, K., Tieman, D.M. *et al.* (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, **51**, 1044.

Garg, G., Kamphuis, L.G., Bayer, P.E., Kaur, P., Dudchenko, O., Taylor, C.M. *et al.* (2022) A pan-genome and chromosome-length reference genome of narrow-leafed lupin (Lupinus angustifolius) reveals genomic diversity and insights into key industry and biological traits. *The Plant Journal*, **111**, 1252–1266.

Garrison, E. & Guarracino, A. (2023) Unbiased pangenome graphs. *Bioinformatics*, **39**(1), btac743. https://doi.org/10.1093/bioinformatics/btac743

Ghazal, H., Adam, Y., Idrissi Azami, A., Sehli, S., Nyarko, H.N., Chaouni, B. *et al.* (2021) Plant genomics in Africa: present and prospects. *The Plant Journal*, **107**, 21–36.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M. *et al.* (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*, **296**, 92–100.

Golicz, A.A., Batley, J. & Edwards, D. (2016a) Towards plant pangenomics. *Plant Biotechnology Journal*, **14**, 1099–1105.

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

Gui, S., Wei, W., Jiang, C., Luo, J., Chen, L., Wu, S. *et al.* (2022) A pan-Zea genome map for enhancing maize improvement. *Genome Biology*, **23**, 178.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W. *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, **48**, 245–252.

Gutjahr, C., Sawers, R.J.H., Marti, G., Andrés-Hernández, L., Yang, S.-Y., Casieri, L. *et al.* (2015) Transcriptome diversity among rice root types during asymbiosis and interaction with arbuscular mycorrhizal fungi. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 6754–6759.

Hameed, A., Poznanski, P., Nadolska-Orczyk, A. & Orczyk, W. (2022) Graph pangenomes track genetic variants for crop improvement. *International Journal of Molecular Sciences*, **23**, 13420. https://doi.org/10.3390/ijms232113420

He, Z., Ji, R., Havlickova, L., Wang, L., Li, Y., Lee, H.T. *et al.* (2021) Genome structural evolution in brassica crops. *Nature Plants*, **7**, 757–765.

Hon, T., Mars, K., Young, G., Tsai, Y.C., Karalius, J.W., Landolin, J.M. *et al.* (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, **7**, 399.

Hong, J.H., Savina, M., Du, J., Devendran, A., Ramakanth, K.K., Tian, X. *et al.* (2017) A sacrifice-for-survival mechanism protects root stem cell niche from chilling stress. *Cell*, **170**, 102–113.e114.

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

Hu, G., Feng, J., Xiang, X., Wang, J., Salojarvi, J., Liu, C. *et al.* (2022) Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. *Nature Genetics*, **54**, 73–83.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, **42**, 961–967.

Hufnagel, B., Soriano, A., Taylor, J., Divol, F., Kroc, M., Sanders, H. *et al.* (2021) Pangenome of white lupin provides insights into the diversity of the species. *Plant Biotechnology Journal*, **19**, 2532–2543.

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

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C. *et al.* (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, **588**, 284–289.

Jean-Baptiste, K., McFaline-Figueroa, J.L., Alexandre, C.M., Dorrity, M.W., Saunders, L., Bubb, K.L. *et al.* (2019) Dynamics of gene expression in single root cells of arabidopsis thaliana. *The Plant Cell*, **31**, 993–1011.

Jiang, L., Lin, M., Wang, H., Song, H., Zhang, L., Huang, Q. *et al.* (2022) Haplotype-resolved genome assembly of Bletilla striata (Thunb.) Reichb.f. to elucidate medicinal value. *The Plant Journal*, **111**, 1340–1353.

Jiao, W.B. & Schneeberger, K. (2017) The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, **36**, 64–70.

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.

Kille, B., Balaji, A., Sedlazeck, F.J., Nute, M. & Treangen, T.J. (2022) Multiple genome alignment in the telomere-to-telomere assembly era. *Genome Biology*, **23**, 182.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. (2015) The technology and biology of single-cell RNA sequencing. *Molecular Cell*, **58**, 610–620.

Korte, A., Vilhjálmsson, B.J., Segura, V., Platt, A., Long, Q. & Nordborg, M. (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, **44**, 1066–1071.

Lake, B.B., Chen, S., Hoshi, M., Plongthongkum, N., Salamon, D., Knoten, A. *et al.* (2019) A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. *Nature Communications*, **10**, 2832.

Lei, L., Goltsman, E., Goodstein, D., Wu, G.A., Rokhsar, D.S. & Vogel, J.P. (2021) Plant pan-genomics comes of age. *Annual Review of Plant Biology*, **72**, 411–435.

Li, H., Wang, S., Chai, S., Yang, Z., Zhang, Q., Xin, H. *et al.* (2022) Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nature Communications*, **13**, 682.

Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M., Liu, Z. *et al.* (2021) Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biology*, **22**, 119.

Li, X., Li, L. & Yan, J. (2015) Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nature Communications*, **6**, 6648.

Li, Y.I., Wong, G., Humphrey, J. & Raj, T. (2019) Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nature Communications*, **10**, 994.

Li, Z., Wang, P., You, C., Yu, J., Zhang, X., Yan, F. *et al.* (2020) Combined GWAS and eQTL analysis uncovers a genetic regulatory network orchestrating the initiation of secondary cell wall development in cotton. *The New Phytologist*, **226**, 1738–1752.

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G. *et al.* (2022) A draft human pangenome reference. *bioRxiv*, 2022.2007.2009.499321.

Libault, M., Pingault, L., Zogli, P. & Schiefelbein, J. (2017) Plant systems biology at the single-cell level. *Trends in Plant Science*, **22**, 949–960.

Lisch, D. (2013) How important are transposons for plant evolution? *Nature Reviews Genetics*, **14**, 49–61.

Liu, C., Wang, Y., Peng, J., Fan, B., Xu, D., Wu, J. *et al.* (2022a) High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. *Plant Communications*, **3**(6), 100352. https://doi.org/10.1016/j.xplc.2022.100352

Liu, Z., Wang, J., Zhou, Y., Zhang, Y., Qin, A., Yu, X. *et al.* (2022b) Identification of novel regulators required for early development of vein pattern in the cotyledons by single-cell RNA-sequencing. *The Plant Journal: For Cell and Molecular Biology*, **110**, 7–22.

Luo, C., Li, X., Zhang, Q. & Yan, J. (2019) Single gametophyte sequencing reveals that crossover events differ between sexes in maize. *Nature Communications*, **10**, 785.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Mancuso, N., Freund, M.K., Johnson, R., Shi, H., Kichaev, G., Gusev, A. *et al.* (2019) Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics*, **51**, 675–682.

Marand, A.P., Chen, Z., Gallavotti, A. & Schmitz, R.J. (2021) A cis-regulatory atlas in maize at single-cell resolution. *Cell*, **184**, 3041–3055.e21.

Marks, R.A., Hotaling, S., Frandsen, P.B. & VanBuren, R. (2021) Representation and participation across 20 years of plant genome sequencing. *Nature Plants*, **7**, 1571–1578.

Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C.S. *et al.* (2021) Long-read sequence assembly: a technical evaluation in barley. *Plant Cell*, **33**, 1888–1906.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.

Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nature Reviews Genetics*, **11**, 31–46.

Misra, B.B., Assmann, S.M. & Chen, S. (2014) Plant single-cell and single-cell-type metabolomics. *Trends in Plant Science*, **19**, 637–646.

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A.J., Abramson, B.W., Schmucker, A. *et al.* (2021) The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science*, **374**, eabi7489.

O'Connell, R.J., Thon, M.R., Hacquard, S., Amyotte, S.G., Kleemann, J., Torres, M.F. *et al.* (2012) Lifestyle transitions in plant pathogenic Colletotrichum fungi deciphered by genome and transcriptome analyses. *Nature Genetics*, **44**, 1060–1065.

Ozsolak, F. & Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, **12**, 87–98.

Powell, A.F., Feder, A., Li, J., Schmidt, M.H., Courtney, L., Alseekh, S. *et al.* (2022) A Solanum lycopersicoides reference genome facilitates insights into tomato specialized metabolism and immunity. *The Plant Journal*, **110**, 1791–1810.

Qiao, Q., Edger, P.P., Xue, L., Qiong, L., Lu, J., Zhang, Y. *et al.* (2021) Evolutionary history and pan-genome dynamics of strawberry (*Fragaria* spp.). *Proceedings of the National Academy of Sciences of the United States of America*, **118**, e2105431118.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z. *et al.* (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**, 3542–3558.e3516.

Quadrana, L., Etcheverry, M., Gilly, A., Caillieux, E., Madoui, M.A., Guy, J. *et al.* (2019) Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nature Communications*, **10**, 3421.

Rautiainen, M., Nurk, S., Walenz, B.P., Logsdon, G.A., Porubsky, D., Rhie, A. *et al.* (2022) Verkko: telomere-to-telomere assembly of diploid chromosomes. *bioRxiv*, 2022.2006.2024.497523.

Rijzaani, H., Bayer, P.E., Rouard, M., Doležel, J., Batley, J. & Edwards, D. (2022) The pangenome of banana highlights differences between genera and genomes. *The Plant Genome*, **15**, e20100.

Ruperao, P., Thirunavukkarasu, N., Gandham, P., Selvanayagam, S., Govindaraj, M., Nebie, B. *et al.* (2021) Sorghum pan-Genome explores the functional utility for genomic-assisted breeding to accelerate the genetic gain. *Frontiers. Plant Science*, **12**, 666342.

Sanderson, N., Kapel, N., Rodger, G., Webster, H., Lipworth, S., street, T. *et al.* (2022) Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford nanopore flowcells and chemistries in bacterial genome reconstruction. *bioRxiv*, 2022.2004.2029.490057.

Schmidt, M.H., Vogel, A., Denton, A.K., Istace, B., Wormit, A., van de Geest, H. *et al.* (2017) De novo assembly of a new Solanum pennellii accession using nanopore sequencing. *Plant Cell*, **29**, 2336–2348.

Schrinner, S., Serra Mari, R., Finkers, R., Arens, P., Usadel, B., Marschall, T. *et al.* (2022) Genetic polyploid phasing from low-depth progeny samples. *iScience*, **25**, 104461.

Schrinner, S.D., Mari, R.S., Ebler, J., Rautiainen, M., Seillier, L., Reimer, J.J. *et al.* (2020) Haplotype threading: accurate polyploid phasing from long reads. *Genome Biology*, **21**, 252.

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q. *et al.* (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, **44**, 825–830.

Serra Mari, R., Schrinner, S., Finkers, R., Arens, P., Schmidt, M.H.-W., Usadel, B. *et al.* (2022) Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data. *bioRxiv*, 2022.2005.2010.491293.

Seyfferth, C., Renema, J., Wendrich, J.R., Eekhout, T., Seurinck, R., Vandamme, N. *et al.* (2021) Advances and opportunities in single-cell transcriptomics for plant research. *Annual Review of Plant Biology*, **72**, 847–866.

Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z. *et al.* (2022) A super pan-genomic landscape of rice. *Cell Research*, **32**, 878–896.

Shulse, C.N., Cole, B.J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M. *et al.* (2019) High-throughput single-cell transcriptome profiling of plant cell types. *Cell Reports*, **27**, 2241+–2247.e4.

Siren, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C. *et al.* (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, **374**, abg8871.

Song, J.-M., Liu, D.-X., Xie, W.-Z., Yang, Z., Guo, L., Liu, K. *et al.* (2021) BnPIR: Brassica napus pan-genome information resource for 1689 accessions. *Plant Biotechnology Journal*, **19**, 412–414.

Su, X., Li, W., Lv, L., Li, X., Yang, J., Luo, X.J. *et al.* (2021) Transcriptome-wide association study provides insights into the genetic component of gene expression in anxiety. *Frontiers in Genetics*, **12**, 740134.

Sun, G., Xia, M., Li, J., Ma, W., Li, Q., Xie, J. *et al.* (2022a) The maize single-nucleus transcriptome comprehensively describes signaling networks governing movement and development of grass stomata. *The Plant Cell*, **34**, 1890–1911.

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

Sun, Y., Wang, J., Li, Y., Jiang, B., Wang, X., Xu, W.H. *et al.* (2022c) Pan-Genome analysis reveals the abundant gene presence/absence variations among different varieties of melon and their influence on traits. *Frontiers in Plant Science*, **13**, 835496.

Tan, Z., Xie, Z., Dai, L., Zhang, Y., Zhao, H., Tang, S. *et al.* (2022) Genome- and transcriptome-wide association studies reveal the genetic basis and the breeding history of seed glucosinolate content in Brassica napus. *Plant Biotechnology Journal*, **20**, 211–225.

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

Tang, S., Zhao, H., Lu, S., Yu, L., Zhang, G., Zhang, Y. *et al.* (2021) Genome- and transcriptome-wide association studies provide insights into the genetic basis of natural variation of seed oil content in Brassica napus. *Molecular Plant*, **14**, 470–487.

Tautz, D., Ellegren, H. & Weigel, D. (2010) Next generation molecular ecology. *Molecular Ecology*, **19**(Suppl 1), 1–3.

The Arabidopsis Genome, I. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.

Tomlins, S.A., Mehra, R., Rhodes, D.R., Cao, X., Wang, L., Dhanasekaran, S.M. *et al.* (2007) Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics*, **39**, 41–51.

Torkamaneh, D., Lemay, M.-A. & Belzile, F. (2021) The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. *Plant Biotechnology Journal*, **19**, 1852–1862.

Uemoto, Y., Ichinoseki, K., Matsumoto, T., Oka, N., Takamori, H., Kadowaki, H. *et al.* (2021) Genome-wide association studies for production, respiratory disease, and immune-related traits in landrace pigs. *Scientific Reports*, **11**, 15823.

Vaillancourt, B. & Buell, R.C. (2020) High molecular weight DNA isolation method from diverse plant species for use with Oxford nanopore sequencing. *bioRxiv*, **783159**.

Valihrach, L., Androvic, P. & Kubista, M. (2018) Platforms for single-cell collection and analysis. *International Journal of Molecular Sciences*, **19**, 807. https://doi.org/10.3390/ijms19030807

van Rengs, W.M.J., Schmidt, M.H., Effgen, S., Le, D.B., Wang, Y., Zaidan, M. *et al.* (2022) A chromosome scale tomato genome built from complementary PacBio and nanopore sequences alone reveals extensive linkage drag during breeding. *The Plant Journal*, **110**, 572–588.

VanBuren, R., Bryant, D., Edger, P.P., Tang, H., Burgess, D., Challabathula, D. *et al.* (2015) Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature*, **527**, 508–511.

Varshney, R.K., Roorkiwal, M., Sun, S., Bajaj, P., Chitikineni, A., Thudi, M. *et al.* (2021) A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature*, **599**, 622–627.

Vernikos, G., Medini, D., Riley, D.R. & Tettelin, H. (2015) Ten years of pan-genome analyses. *Current Opinion in Microbiology*, **23**, 148–154.

Vilanova, S., Alonso, D., Gramazio, P., Plazas, M., Garcia-Fortea, E., Ferrante, P. *et al.* (2020) SILEX: a fast and inexpensive high-quality DNA extraction method suitable for multiple sequencing platforms and recalcitrant plant species. *Plant Methods*, **16**, 110.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics*, **101**, 5–22.

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D. *et al.* (2019) Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, **51**, 592–599.

Wang, F., Xia, Z., Zou, M., Zhao, L., Jiang, S., Zhou, Y. *et al.* (2022a) The autotetraploid potato genome provides insights into highly heterozygous species. *Plant Biotechnology Journal*, **20**, 1996–2005.

Wang, S., Qian, Y.Q., Zhao, R.P., Chen, L.L. & Song, J.M. (2022b) Graph-based pan-genome: increased opportunities in plant genomics. *Journal of Experimental Botany*, **74**(1), 24–39. https://doi.org/10.1093/jxb/erac412

Wang, Y., Li, W., Wang, L., Yan, J., Lu, G., Yang, N. *et al.* (2022c) Three types of genes underlying the gametophyte factor1 locus cause unilateral cross incompatibility in maize. *Nature Communications*, **13**, 4498.

Wang, Z., Gerstein, M. & Snyder, M. (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.

Wendrich, J.R., Yang, B., Vandamme, N., Verstaen, K., Smet, W., Van de Velde, C. *et al.* (2020) Vascular transcription factors guide plant epidermal responses to limiting phosphate conditions. *Science*, **370**, 810.

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

Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L. *et al.* (2019) Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nature Genetics*, **51**, 1052–1059.

Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M.G., Andersen, S. *et al.* (2022) Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, **376**, eabf3041.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203–208.

Zanini, S.F., Bayer, P.E., Wells, R., Snowdon, R.J., Batley, J., Varshney, R.K. *et al.* (2022) Pangenomics in crop improvement-from coding structural variations to finding regulatory variants with pangenome graphs. *Plant Genome*, **15**, e20177.

Zapata, L., Ding, J., Willing, E.M., Hartwig, B., Bezdan, D., Jiao, W.B. *et al.* (2016) Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, E4052–E4060.

Zhang, H., Liu, X., Zhang, Y., Jiang, C., Cui, D., Liu, H. *et al.* (2012) Genetic analysis and fine mapping of the Ga1-S gene region conferring cross-incompatibility in maize. *Theoretical and Applied Genetics*, **124**, 459–465.

Zhang, T.-Q., Chen, Y., Liu, Y., Lin, W.-H. & Wang, J.-W. (2021a) Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. *Nature Communications*, **12**, 2053.

Zhang, T.-Q., Xu, Z.-G., Shang, G.-D. & Wang, J.-W. (2019) A single-cell RNA sequencing profiles the developmental landscape of arabidopsis root. *Molecular Plant*, **12**, 648–660.

Zhang, W., Luo, C., Scossa, F., Zhang, Q., Usadel, B., Fernie, A.R. *et al.* (2021b) A phased genome based on single sperm sequencing reveals crossover pattern and complex relatedness in tea plants. *The Plant Journal*, **105**, 197–208.

Zhang, W., Zhang, Y., Qiu, H., Guo, Y., Wan, H., Zhang, X. *et al.* (2020) Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nature Communications*, **11**, 3719.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A. *et al.* (2010) Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, **42**, 355–360.

Zhang, Z., Zhang, B., Chen, Z., Zhang, D., Zhang, H., Wang, H. *et al.* (2018) A PECTIN METHYLESTERASE gene at the maize Ga1 locus confers male function in unilateral cross-incompatibility. *Nature Communications*, **9**, 3678.

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

Zhou, D., Jiang, Y., Zhong, X., Cox, N.J., Liu, C. & Gamazon, E.R. (2020) A unified framework for joint-tissue transcriptome-wide association and mendelian randomization analysis. *Nature Genetics*, **52**, 1239–1246.

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y. *et al.* (2022) Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, **606**, 527–534.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W. *et al.* (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, **33**, 408–414.

Zhu, F., Jadhav, S.S., Tohge, T., Salem, M.A., Lee, J.M., Giovannoni, J.J. *et al.* (2022) A comparative transcriptomics and eQTL approach identifies SlWD40 as a tomato fruit ripening regulator. *Plant Physiology*, **190**, 250–266.

Zhu, G.T., Wang, S.C., Huang, Z.J., Zhang, S.B., Liao, Q.G., Zhang, C.Z. *et al.* (2018) Rewiring of the fruit metabolome in tomato breeding. *Cell*, **172**, 249-+.

Zong, J., Wang, L., Zhu, L., Bian, L., Zhang, B., Chen, X. *et al.* (2022) A rice single cell transcriptomic atlas defines the developmental trajectories of rice floret and inflorescence meristems. *New Phytologist*, **234**, 494–512.