Quantitative Plant Biology

cambridge.org/qpb

Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions

Boas Pucker[1,2,*], Iker Irisarri[3,4], Jan de Vries[3,4,5] and Bo Xu[6]

[1]Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom; [2]Institute of Plant Biology & Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig, Germany; [3]Department of Applied Bioinformatics, Institute for Microbiology and Genetics, University of Goettingen, Göttingen, Germany; [4]Campus Institute Data Science (CIDAS), University of Goettingen, Göttingen, Germany; [5]Department of Applied Bioinformatics, Göttingen Center for Molecular Biosciences (GZMB), University of Goettingen, Göttingen, Germany; [6]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China

Review

Cite this article: B. Pucker et al. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions . *Quantitative Plant Biology*, **3**:e5, 1–14
https://dx.doi.org/10.1017/qpb.2021.18

Received: 29 June 2021
Revised: 24 November 2021
Accepted: 21 December 2021

Keywords:
haplophasing; long read sequencing; Oxford Nanopore Technologies (ONT); Pacific Biosciences (PacBio); plant genome assembly; plant genomics.

Author for correspondence:
Boas Pucker
E-mail: b.pucker@tu-braunschweig.de

Abstract

Third-generation long-read sequencing is transforming plant genomics. Oxford Nanopore Technologies and Pacific Biosciences are offering competing long-read sequencing technologies and enable plant scientists to investigate even large and complex plant genomes. Sequencing projects can be conducted by single research groups and sequences of smaller plant genomes can be completed within days. This also resulted in an increased investigation of genomes from multiple species in large scale to address fundamental questions associated with the origin and evolution of land plants. Increased accessibility of sequencing devices and user-friendly software allows more researchers to get involved in genomics. Current challenges are accurately resolving diploid or polyploid genome sequences and better accounting for the intra-specific diversity by switching from the use of single reference genome sequences to a pangenome graph.

1. Introduction

Resolving the genome structure of plants is the key to unlock the complex chassis of genetic factors determining phenotypic traits. As a biochemically homogeneous molecule, DNA can be analysed at high throughput. Enormous progress has been made in the sequencing fields over the last decades. The increase in sequencing capacity is frequently displayed outpacing Moore's law. This technological advancement facilitated major discoveries in numerous fields of life science, such as the discovery of biosynthetic gene clusters in crops (Ma, Vaistij, et al., 2021), insights into the genomic diversity of crops (Jayakodi et al., 2020; Walkowiak et al., 2020; Zhou, Chebotarov, et al., 2020), and generally a better understanding of land plant genome evolution (Carta et al., 2020; Liu et al., 2021). Plant genomics is often applied to unlock the agronomic potential of plants through identification of genetic loci underlying agronomical traits. Loci responsible for a certain trait might involve multiple genes and span hundreds or even thousands of kilobasepairs (kb). Extreme examples are biosynthetic gene clusters that can reach sizes of several hundred kb or even multiple megabases (Mbp) (Nützmann et al., 2016; Zheng et al., 2021). Therefore, it becomes useful to investigate specific allele combinations of neighbouring genes which are forming a haplotype. A sequence representing this combination of neighbouring alleles is called a haplophase. Many application cases require a genome sequence that represents all haplophases of the investigated species. Long-read sequencing is currently the method of choice to generate highly contiguous plant genome assemblies.

Here, we summarise the latest developments in the fast progressing field of plant genome sequencing, identify current challenges, highlight opportunities and postulate future directions. Our objective is to give an introduction to this field so that more plant scientists can benefit from the extensive potential of long read genomics.

John Innes Centre
Unlocking Nature's Diversity
CAMBRIDGE UNIVERSITY PRESS

2. Long-read sequencing technologies

There is no unified definition of "third-generation" or "long-read" sequencing technologies. Therefore, we will use a pragmatic approach and focus on the most important sequencing technologies. Refer to previous reviews about Roche/454 pyrosequencing (Metzker, 2010), Ion Torrent sequencing (Rothberg et al., 2011) or BGI's Single Tube Long Fragment Read method (Wang et al., 2019). Mainly two companies offer technologies which are expected to be the workhorses of genome sequencing projects in the future: Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). The general concept and technical details of the ONT (Branton et al., 2008; Jain et al., 2017) and the PacBio (Eid et al., 2009; Hon et al., 2020; Metzker, 2010) technologies have been described and reviewed before.

Briefly, ONT sequencing is based on measuring changes of an electric signal over a membrane while a DNA strand slides through a nanopore in this membrane (Figure 1a). The recorded changes in the electric signal are characteristic for a certain composition of nucleotides partially blocking the pore and can be translated into a nucleotide sequence. Since this measuring in nanopores is not inherently restricted to DNA, this technology is currently the only method to analyse entire RNA molecules directly at high throughput. Two substantially different types of nanopores are currently distributed by ONT in the R9 and R10 flow cell families, which can be further subclassified. While R9 flow cells tend to have higher output than R10, more bases determine the signal of R10 flow cells. This is due to a longer barrel of the nanopore with a dual reader head in R10 instead of only one reader head in the R9. A reader head measures the electrical signal caused by about six bases that are located in the nanopore. Consequently, R10 flow cells are better suited to resolve homopolymers (ONT, 2021a). Models for the conversion of electric signal to a nucleotide sequence need to be trained individually for each nanopore type. An important feature of the nanopore technology is that there is no limit to the read length—other than the length/integrity of the molecule itself. The raw read accuracy can be increased from 90–95% to over 97% if a species-specific model for basecalling is available (Vereecke et al., 2020). A recent update of flow cells and chemistry enables average

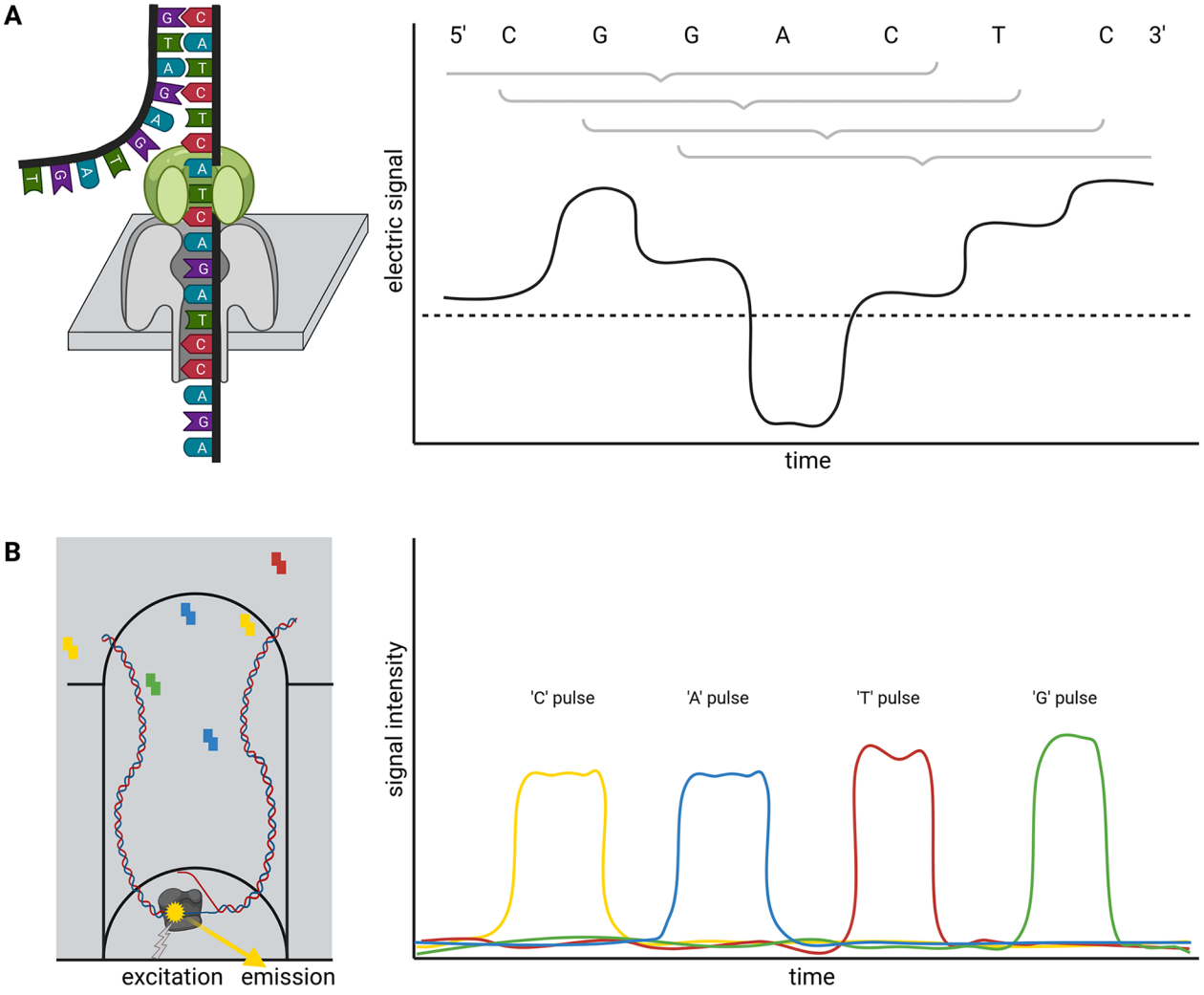

Fig. 1. Schematic illustration of nanopore sequencing (a) and Single-Molecule Real Time (SMRT) sequencing (b). Nanopore sequencing is based on the translocation of a DNA or RNA strand through a nanopore located in an artificial membrane. Multiple nucleotides located in the nanopore determine the flow of ions through this nanopore in a specific way by physically blocking the space. This change in ion flux is recorded as an electric signal and further converted into sequence information. The illustration shows the contribution of six bases to the signal, but the number of bases depends on the pore type. SMRT sequencing detects fluorescent light emitted from nucleotides upon incorporation into a DNA strand. The DNA polymerase is located at the bottom of a well and synthesises a new DNA strand. The integration into the new DNA strand keeps the nucleotide for a sufficiently long time in the well to allow detection.

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

raw read qualities around Q20 (99% accuracy). Various DNA or RNA modifications can be analysed based on ONT sequence reads (Karsten et al., 2017; Parker et al., 2019).

Single molecule real time (SMRT) sequencing offered by PacBio is based on a polymerase located in a well (Zero-Mode Waveguide, ZMW). This polymerase builds a complementary strand to a template DNA strand (Figure 1b). The incorporation of fluorescently labeled nucleotides is detected and reveals the sequence of the analysed DNA strand. PacBio offers Continuous Long Reads (CLR) and Circular Consensus Sequencing (CCS) reads also called High-Fidelity (HiFi) (Wenger et al., 2019). The later read type is the result of sequencing the same circularised DNA molecule multiple times and correcting the reads through alignment. Consequently, there is a tradeoff between the consensus read length and the per-base accuracy. The 99.5% accuracy of HiFi reads exceeds the average accuracy of CLR, but HiFi reads are usually shorter than 25 kb (Hon et al., 2020; Wenger et al., 2019). The combination of long-read length with high per-base accuracy in one technology allows the investigation of highly repetitive genomic regions.

The latest long-read technologies have the capacity to analyse extremely long DNA fragments up to millions of nucleotides in the case of ONT (Payne et al., 2019). While top read lengths of up to 500 kb can be achieved routinely in ONT sequencing runs, the longest observed plant DNA reads reached about 1.5 Mbp (Benjamin Schwessinger, personal communication). Since many sequencing projects are focussed on species without existing reference genome sequence assemblies, it is often not possible to confirm these reads through alignment against a reference genome sequence. However, long-read sequencing technologies allow to generate new assemblies for the species of interest with relative ease. Since there is no technological limit to the read length, the major challenge is the efficient isolation of high molecular weight DNA in order to obtain ultra-long reads that facilitate genome assembly. Due to the stable cell wall and a plethora of specialised metabolites, DNA extraction from plant cells is more complicated than DNA extraction from many animal cells. Challenges increase further when considering the high diversity of plants including algal species. Various DNA extraction protocols and adjustments of existing methods were developed in the last years (Li, Parris, & Saski, 2020; Siadjeu et al., 2020). Additional enrichment methods like the Short Read Eliminator kit (Circulomics) help to exclude short fragments resulting in an increased average read length. In addition to the enrichment of long molecules, reducing the amount of required DNA input is an additional challenge. Limited availability of suitable plant samples combined with large genome sizes can pose a challenge to sequencing projects. Long-read sequencing is still characterised by substantial variation between sequencing runs. This can partially be explained by differences in DNA quality. Improvements in the consumable production process might mitigate issues arising from low output runs by ensuring constant high quality. Warranty of minimal output by the supplier is a solution for the meantime. Users of commercial sequencing services might want to negotiate pricing based on the quality and quantity of sequence reads rather than on the amount of consumed materials.

3. High molecular weight DNA extraction for long-read sequencing

Enormous improvements of the actual sequencing capacity turned high molecular weight DNA extraction from plants into a limiting step. Many protocols for high molecular weight DNA extraction were developed previously (Jones et al., 2021; Li, Parris, & Saski, 2020; Maghini et al., 2021; Murray & Thompson, 1980; Siadjeu et al., 2020; Vilanova et al., 2020).

While the presence of long DNA molecules in the sample is crucial, short fragments can be depleted in a purification step. Moreover, the purity of the DNA is important to avoid interference with the library preparation and sequencing chemistry. Specialised metabolites and proteins might interact with the DNA and reduce the final sequencing output. Long read sequencing projects usually require several micrograms of DNA which is substantially more than needed for short-read sequencing (Siadjeu et al., 2020). This can become a challenge if no suitable plant tissues are available. Young leaves are often a good source of DNA (Pucker et al., 2021), because the number of cells (and nuclei) is high and the amount of specialised metabolites is low. Incubation in the dark for a few days can help reduce starch and sugar concentrations, thereby reducing the sugar contamination in the DNA sample. Extraction protocols should avoid shearing of the DNA molecules and storage of the final elution is recommended at 4°C. As DNA can degrade over time, the extraction should be performed in time for the sequencing experiment to ensure optimal performance.

4. Genome sequencing is accelerated, affordable and accessible

4.1. Accelerated

The 20th anniversary of the *Arabidopsis thaliana* genome sequence (Provart et al., 2020) highlights the enormous progress that has been achieved in plant genomics within two decades. While the sequencing of the first plant genome was an expensive and tedious undertaking performed by a large international consortium, *A. thaliana* genomes are now being sequenced and assembled by many labs within days (Jiao & Schneeberger, 2020; Michael et al., 2018; Pucker et al., 2019). There is also substantial progress when looking at crop genome sequencing projects. Large international genome sequencing consortia were necessary to unravel the first genome sequences of crops like rice (Goff et al., 2002; Yu et al., 2002), poplar (Tuskan et al., 2006), grapevine (Jaillon et al., 2007) and tomato (Sato et al., 2012). Now, enormous genome sequencing projects like the Darwin Tree of Life (Darwin Tree of Life Project, 2021), Earth BioGenome Project (Lewin et al., 2018) or the European Research Genome Atlas (ERGA; https://www.erga-biodiversity.eu/) are starting to sequence the genomes of all eukaryotic species within the next few years. These projects advance an open data policy and might have a positive impact beyond genomics. Therefore, it can be assumed that high-quality reference genome sequences will be available for most species in the near future. The workflow from harvesting plant material in the greenhouse or field to DNA extraction, sequencing, and *de novo* genome assembly can be completed within days (Michael et al., 2018; Pucker et al., 2021). However, current long-read technologies do not allow the construction of gapless telomere-to-telomere genome sequences on a routine basis yet. Regions like the centromere and nucleolus organising regions are not even completely resolved in the latest *A. thaliana* genome assemblies (Michael et al., 2018; Pucker et al., 2021). Consequently, challenges to close the remaining gaps in genome sequences of most species will remain for the foreseeable future. Since the read lengths of both long-read technologies is impressive, the major factor to optimise in the future is per-base accuracy. Rapid increase of the raw read quality during the last years accelerated many genome

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

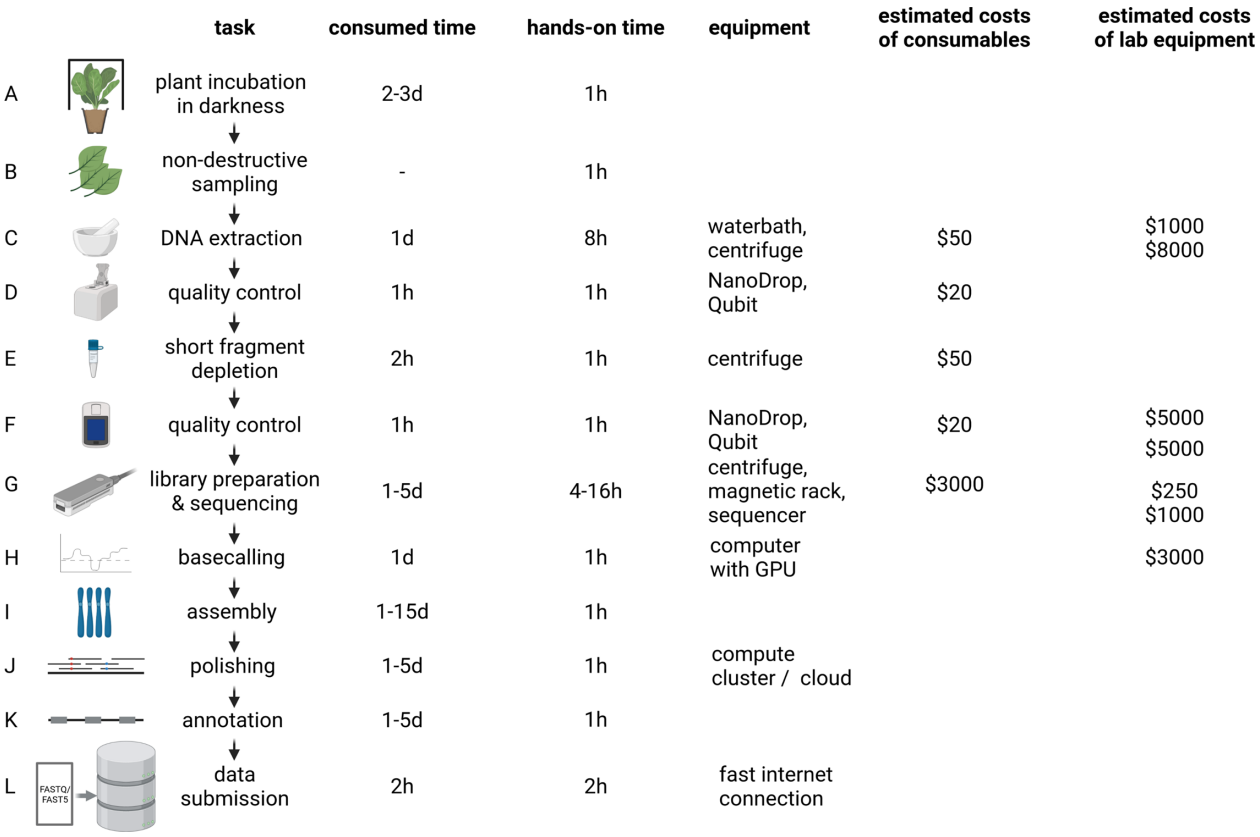

Fig. 2. Plant genome project workflow from DNA extraction over Oxford Nanopore Technologies (ONT) sequencing to data submission. The indicated durations depend on the size and complexity of the investigated plant genome, with larger genomes generally taking longer to analyse. To reduce sugar content, plants are incubated in the dark for a few days prior to DNA extraction (a). Non-destructive sampling is important to allow additional genomic sequencing and also RNA-Seq if required in later stages of a project (b). Mechanical disruption of cell walls is required for the DNA extraction (c). Photometric analysis of the DNA solution (including quantification) is often the first step of quality control (d and f). Removal of short DNA fragments is highly recommended to improve the sequencing output and quality (e). ONT library preparation and sequencing can be repeated several times to increase the output (g). Graphic cards are an efficient resource to convert electric signal into sequence information in real time (h). Multiple tools are available to generate a chromosome-arm level assembly based on long reads (i). Additional polishing in multiple rounds can be necessary due to the noisy character of long reads (j). The value of a genome sequence can be enriched through the identification of relevant genetic elements like genes and transposable elements (k). All data should be shared with the community via submission to a public repository which ensures long-term storage (l). d, day(s); hr, hour(s). The given time estimates for assembly, polishing and annotation are the minimal run time required for the analyses. Manual curation and iterative improvements can take substantially longer. The estimated costs of consumables are based on a haploid 1-Gbp genome and a targeted coverage of 30× which would require six libraries to be sequenced on three MinION/GridION flow cells when assuming an average output of 10 GB per flow cell with two libraries sequenced per flow cell. Investment costs for non-standard lab equipment are independent of the specific sequencing project and only required for high-output experiments in the lab. There is an option to perform rapid sequencing without these instruments in the field, but the lower output does not make that option attractive for large plant genomes.

sequencing projects. PacBio offers HiFi reads which are highly accurate and up to 25 kb long. Since per-base accuracy is based on sequencing the same molecule numerous times, improving the polymerase lifetime could increase raw read accuracy and simultaneously shift the length limit. ONT recently released a 'Q20+' technology together with R10.4 flow cells, which is pushing the raw read accuracy beyond 99% (ONT, 2021b). Since the length of ONT reads is only limited by the length of the DNA molecule, this could become the routine technology to resolve rDNA clusters. The high accuracy of PacBio and ONT long reads accelerates the assembly process and removes the need for short-read polishing, which was previously required to correct errors in non-repetitive regions. As short reads cannot be mapped onto sequences of repetitive regions with reliability, long-read only assemblies could also accelerate the research on transposable elements.

4.2. Affordable

The distribution of affordable ONT MinION sequencers started the democratisation of sequencing (The long view on sequencing, 2018). Increase in read length and output enabled substantial improvements of assembly contiguity and reduced costs associated with genome sequencing projects. Genome sequencing is likely to replace classic polymerase chain reaction-based genotyping methods in certain application cases due to higher cost-effectiveness (Pucker et al., 2021). Plant genome assemblies at chromosome-arm level often cost less than $10,000 and can be completed within days to weeks for many species (Figure 2). However, reaching a telomere-to-telomere assembly is still difficult and expensive. Commercial service centres offer the generation of data at continuously decreasing prices rendering genome sequencing affordable for most research groups. This democratisation might shift the focus of genome sequencing projects from crops with importance in agriculture to neglected crops in developing countries. Improved technologies and substantially reduced sequencing costs have the potential to establish genome sequences as a standard for all plant species. Genetic markers, Hi-C or optical mapping data can be used to arrange contigs into representations of entire chromosomes so-called pseudochromosomes or C-scaffolds (Lewin et al., 2019; Li, Xiang, et al., 2020; Paajanen et al., 2019). Pseudochromosomes

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

contain ordered contigs connected by stretches of ambiguous bases (Ns) to indicate assembly gaps that are only bridged by information about the distance of specific sequences without knowledge about the interleaved sequence. The concept could be considered analogous to paired-end or mate-pair reads, but the distance between the markers is substantially larger. Assemblies generated with the latest long-read technologies can surpass long-standing reference genome sequences with respect to quality and contiguity (Pucker et al., 2021; Rai et al., 2021). Portable sequencers like MinION and Flongle might not be the choice for crop genome sequencing because affordability and throughput are more important than on-site sequencing.

4.3. Accessible

Initial crop genome sequencing projects relied mostly on short reads of second-generation sequencing technologies such as Roche/454 pyrosequencing and Illumina sequencing-by-synthesis which are only accessible to large sequencing centres that can afford the maintenance of expensive instruments. Costs associated with PacBio sequencers still prevent single research groups from buying their own instruments; thus services provided by companies or core facilities are required. However, portable ONT sequencers provide new opportunities for small labs thereby opening an unprecedented opportunity for genome sequencing in low-income countries and for non-model plants such as algae. Substantially, more researchers get involved in genome sequencing and the awareness for opportunities increases. It is also likely that orphan crops, that is species with untapped economic potential, will be made accessible through the publication of their genome sequences (Hunt et al., 2020; Siadjeu et al., 2020; Wang et al., 2021). Huge community engagement inspired the development of more user-friendly and mobile software tools (de Koning et al., 2020; Oliva et al., 2020; Palatnick et al., 2020; Samarakoon et al., 2020), which are paving the way for the democratisation of sequencing data analysis. Both PacBio and ONT come with the opportunity to identify DNA modifications. Even if this opportunity is not used in all sequencing projects, re-use of datasets is possible if all raw data are deposited in public repositories like the Sequence Read Archive and European Nucleotide Archive. Pure bioinformatics groups without experience in genome sequencing can harness these datasets for their analyses. Finally, there is also an educational aspect to portable sequencers. MinION and Flongle can be used to perform plant genomics projects in practical courses at universities and beyond. Persons with basic laboratory skills can operate these sequencers based on instruction videos and manuals without additional training.

5. Pangenomics: From re-sequencing to reference quality genome assemblies of cultivars

The pangenome concept describes all genes or more generally genetic information that is present in a certain group of individuals, for example a population, a species or a higher taxonomic unit. Pangenomes comprise a small set of essential or core genes and numerous genes with different levels of dispensability some of which might be 'accessory' genes (Marroni et al., 2014; Sielemann et al., 2021). A single assembly cannot capture the complete set of genes present in a species and thus a species' pangenome is a better reflection of the diversity. In plants, accessory genes are often enriched in functions related to biotic and abiotic stress response (Bayer et al., 2020). The objective of earlier genome sequencing consortia has been to construct one reference genome sequence that would not just benefit research on one particular species, but would also support research on related species. In such cases, variations in different cultivars or related species were investigated by short read-based re-sequencing and mapping to the reference genome sequence (Figure 3). For example, such studies investigated the pangenome of the model species *A. thaliana* (Alonso-Blanco et al., 2016), tomato (Causse et al., 2013), rice (Lv et al., 2020) and grapevine (Liang et al., 2019). Despite their success, such short-read re-sequencing projects have inherent limitations such

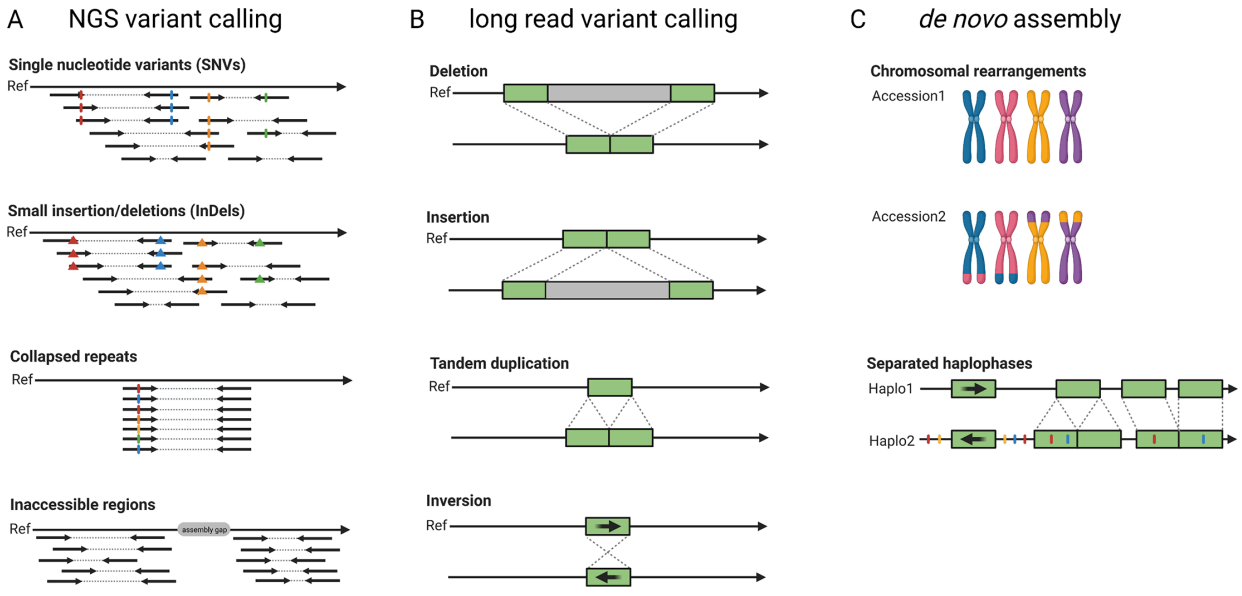

Fig. 3. Development of sequence analysis for exploring genome structure and variability. Read mapping and variant calling was the initial approach to characterise differences between samples based on short-read ('NGS') data (a). Long reads allow an improved variant detection which is especially suited for the detection of structural variants (b). Independent *de novo* genome assemblies allow the identification of all variants and already include an assignment of variants to haplophases (c).

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

as the inabilities to resolve large insertions or to identify variants in repetitive or heterozygous regions (Cameron et al., 2019; Schilbert et al., 2020). Long reads enable the identification of structural sequence variants which have not been identified based on short reads (Chawla et al., 2021). The detection of single nucleotide variants (SNV) requires dedicated tools like DeepVariant (Poplin et al., 2018) and LongShot (Edge & Bansal, 2019), but can outperform the SNV detection based on short reads in difficult-to-map regions (Olson et al., 2021). Nevertheless, *de novo* genome assemblies for multiple different cultivars and comparison of the resulting genome sequences is likely to replace classic variant calling against one reference sequence in most applications (Bayer et al., 2020; Michael & VanBuren, 2020). The feasibility and advantages of constructing a *de novo* genome assembly for the discovery of sequence differences within one species were demonstrated for *A. thaliana* (Michael et al., 2018; Pucker et al., 2021). First crop genome projects generated independent long-read assemblies covering the crop species rice (Choi et al., 2020; Stein et al., 2018; Zhou, Chebotarov, et al., 2020), rapeseed (Song et al., 2020), apple (Sun et al., 2020), wheat (Walkowiak et al., 2020), barley (Jayakodi et al., 2020), soybean (Liu et al., 2020), sorghum (Tao et al., 2021) and maize (Hufford et al., 2021). These studies identified large structural variants including translocations, insertions, deletions, inversions and chromosome fusions. They also found that some 'accessory' genes can have large phenotypic effects including ecotype differentiation, flowering time, stress tolerance or seed weight (Song et al., 2020; Walkowiak et al., 2020). Additional genes and other cultivar specific sequences can be discovered in these projects, but the study of pangenomes should not be limited to crops, because wild relatives might harbour a richer set of 'accessory' genes (Bayer et al., 2020). Some of these 'accessory' genes (e.g., pathogen resistances) could be introduced into crops through breeding. Clearly, long-read sequencing plays a crucial role in the transition towards plant pangenomics.

6. Understanding the deep roots of plant evolution through genomics

Comparative genomic analyses of land plants and their algal relatives provide an unprecedented opportunity to investigate the origin and evolution of embryophytes and their traits. Some agriculturally relevant traits such as tolerance of water scarcity and mutualistic symbioses have deep evolutionary origins predating the origin of land plants. Insights into the evolution of these traits is not only relevant for the understanding of plant terrestrialisation, but can thus also have agronomical implications (Bowles et al., 2021; Fürst-Jansen et al., 2020). Complemented with functional genomic studies, comparative genomics shed light on the innovations of land plant traits such as water conduction systems (Xu et al., 2014), rooting systems (Menand et al., 2007), membrane modifications (Resemann et al., 2021), cuticle (Xu et al., 2021) and stomata (Chater et al., 2017). Deciphering the genomes of species occupying critical phylogenetic positions revealed information on the origin and early evolution of seed-free plants (Szövényi et al., 2021), gymnosperms (Liu et al., 2021), flowering plants (Zhang, Chen, et al., 2020), and grasses (Ma, Liu, et al., 2021), and the genomes of land plants' algal relatives provide a better understanding of genetic changes underpinning the water-to-land transition and associated stress adaptations (Cheng et al., 2019; Jiao et al., 2020; Nishiyama et al., 2018; Wang et al., 2020). In fact, 'alga' is a general term for photosynthetic eukaryotes (and historically also cyanobacteria), which include not only streptophyte algae and land plants, but also an astonishing diversity of green, red and glaucophyte algae—all of which are derived from the singular primary endosymbiotic incorporation of the cyanobacterial progenitor of plastids (de Vries & Archibald, 2017; Keeling, 2013; Sibbald & Archibald, 2020). Additionally, many other eukaryotic groups secondarily acquired plastids by eukaryote–eukaryote endosymbioses, including brown and golden algae or diatoms, among many others (Keeling, 2013; Sibbald & Archibald, 2020; Strassert et al., 2021). This long and convoluted evolutionary history translates into an extraordinary diversity of genomes (Blaby-Haas & Merchant, 2019). Interpretation of these genomes has important biological and biotechnological implications. Over 100 algal genomes have been sequenced to date (Grigoriev et al., 2021) and more are to come.

Until recently, very few algal genome sequences could be considered complete (telomere-to-telomere) and these were on the small range of genome size, with most other assemblies having variable completeness from very short contigs to chromosome-level assemblies (Blaby-Haas & Merchant, 2019). Given the high phylogenetic diversity of algae and the fact that specimens are often sourced from natural populations (most are non-model organisms), high heterozygosity and the presence of many repetitive elements can hamper the assembly of a high-quality algal genome sequence (Michael & VanBuren, 2020). With the exception of a few algal model systems such as *Chlamydomonas reinhardtii* (O'Donnell et al., 2020), *Cyanophora paradoxa* (Price et al., 2019), *Phaeodactylum tricornutum* (Filloramo et al., 2021) and *Thalassiosira pseudonana* (Armbrust et al., 2004), most algal genomes are relatively poorly characterised in comparison to flowering plants. Fortunately, new algal models are flourishing, be it *Ectocarpus siliculosus* (Coelho et al., 2012), *Nannochloropsis* spp. (Radakovits et al., 2012) or *Ulva mutabilis* (De Clerck et al., 2018). A list of available algal and non-seed plant genomes is shown in Table 1. As in other non-model organisms, functional annotation of algal genomes is hampered by the large phylogenetic distance to current model species in which proteins have been functionally characterised (often flowering plants). The likelihood of finding orthologs with the same function across long evolutionary times is low. Currently, about half of the annotated proteins in algal genome sequences, on average, lack functional annotation obtained by searches against Pfam or EggNOG databases (Blaby-Haas & Merchant, 2019). This suggests that algae harbour a vast genetic potential and new gene functions that are yet to be discovered through biochemical characterisation. Gene family analysis using protein similarity networks, co-expression networks and phylogenetic reconstruction are powerful methods to improve functional annotation, providing information on protein domains, condition-specific gene regulation and evolutionary links from knowns to unknowns (de Vries et al., 2021; Gong & Han, 2021; Li et al., 2015; Rhee & Mutwil, 2014; Ruprecht et al., 2017)—especially when novel lineages of algae are involved (Li, Wang, et al., 2020). Reliable genome sequences are the foundation for all these approaches.

Besides nuclear genomes, the plastid (plastome) and mitochondrial (chondrome) counterparts are often of interest in evolutionary biology. The automatic generation of full plastid or mitochondrial genome sequences is now possible as a byproduct of nuclear genome sequencing projects. Long reads also make the more complex chondrome more accessible to genomic studies. Various pipelines have been implemented for the assembly of organellar genomes using exclusively long-read or in combination with short-read data (Soorni et al., 2017; Wick et al., 2017).

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

Table 1. Available streptophyte algae and non-seed plant genomes salient to our understanding of plant diversity and evolution

Species	Assembly size (Mb)	Scaffold N50 (bp)	Lineage	Reference
Chara braunii	1,751.21	2,261,426	Streptophyte algae (Charophyceae)	Nishiyama et al., 2018
Chlorokybus atmophyticus	74.33	752,385	Streptophyte algae (Chlorokybophyceae)	Wang et al., 2020
Klebsormidium nitens	104.21	134,930	Streptophyte algae (Klebsormidiophyceae)	Hori et al., 2014
Mesostigma viride	441.7	2,558,729	Streptophyte algae (Mesostigmatophyceae)	Wang et al., 2020
Mesotaenium endlicherianum	173.75	448,375	Streptophyte algae (Zygnematophyceae)	Cheng et al., 2019
Penium margaritaceum	3,661	116,100	Streptophyte algae (Zygnematophyceae)	Jiao et al., 2020
Spirogloea muscicola	170.82	566,364	Streptophyte algae (Zygnematophyceae)	Cheng et al., 2019
Anthoceros agrestis (Bonn)	116.9	17,300,000	Hornworts	Li, Nishiyama, et al., 2020
Anthoceros angustus	119.35	796,643	Hornworts	Zhang, Fu, et al., 2020
Anthoceros punctatus	132.8	1,700,00	Hornworts	Li, Nishiyama, et al., 2020
Marchantia inflexa	208.75	11,136	Liverworts	Marks et al., 2019
Marchantia paleacea	250.80	2,390,877	Liverworts	Radhakrishnan et al., 2020
Marchantia polymorpha	225.76	1,366,373	Liverworts	Bowman et al., 2017
Ceratodon purpureus	362.51	1,405,213	Mosses	Carey et al., 2021
Fontinatis antipyretica	385.2	45,800	Mosses	Yu et al., 2020
Funaria hygrometrica	340	100,000	Mosses	Kirbis et al., 2020
Pleurozium schreberi	318.34	154,439	Mosses	Pederson et al., 2019
Physcomitrium patens	472.08	17,435,539	Mosses	Lang et al., 2018
Sphagnum fallax	395,1	21,100,000	Mosses	*Sphagnum fallax* v1.1, DOE-JGI, http://phytozome.jgi.doe.gov/
Sphagnum magellanicum	439.0	23,200,000	Mosses	*Sphagnum magellanicum* v1.1, DOE-JGI, http://phytozome.jgi.doe.gov/
Syntrichia caninervis	329.82	21,898,694	Mosses	Silva et al., 2021
Isoetes taiwanensis	1,660	17,400,000	Lycophytes	Wickell et al., 2021
Selaginella lepidophylla	122	163,000	Lycophytes	VanBuren et al., 2018
Selaginella moellendorffii	212.31	119,796	Lycophytes	Banks et al., 2011
Selaginella tamariscina	300.73	407,666	Lycophytes	Xu et al., 2018
Azolla filiculoides	750	964,700	Ferns	Li et al., 2018
Ceratopteris richardii	7,462.46	2,273,607	Ferns	Marchant et al., 2019
Salvinia cucullata	260	719,800	Ferns	Li et al., 2018

Denoted are numbers on the total assembly size, contiguity statistics (N50), taxonomic affiliation and references. Genome statistics were obtained from NCBI's Assembly data base or the corresponding publications.

7. From haploid to diploid genome assembly

Crop genome sequencing projects were focussed on almost homozygous cultivars (Jaillon et al., 2007) or even doubled haploid lines when possible (Dohm et al., 2014). Even human genome initiatives that are usually a few years ahead of plant sciences, have only recently managed to produce a complete haploid genome assembly (Nurk et al., 2021). This implies that two separate genome sequences need to be assembled to represent the two haplotypes of heterozygous genotypes. Haplotypes are the biological molecules i. e. a group of alleles that are inherited together. Haplotypes are represented by haplophases in the assembly. The need to distinguish between these two haplophases when targeting heterozygous genotypes adds an additional overhead that makes the situation more complicated. When possible, genome sequencing projects avoided the challenge of separating haplophases by focussing on homozygous or haploid genotypes. The genomes of polyploidy species are an even bigger challenge, because more than two haplotypes need to be represented in the assembly. Polyploid

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

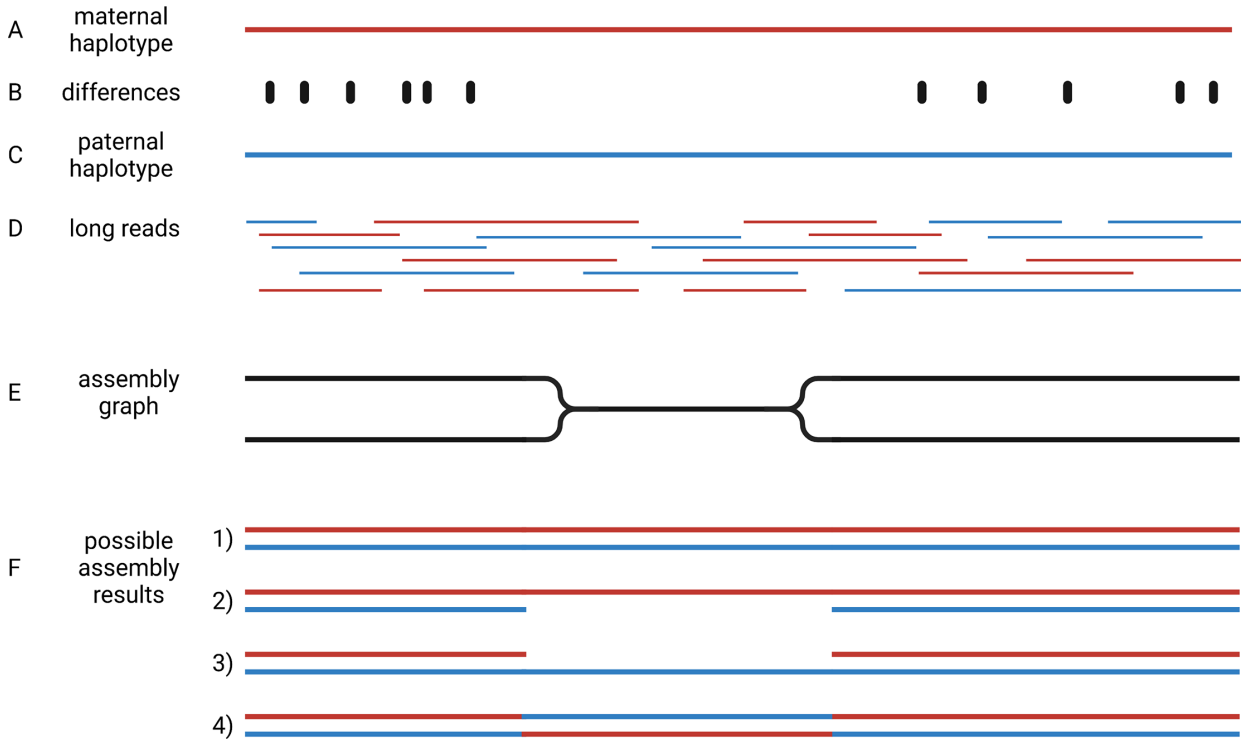

Fig. 4. Assembly of haplophases. Diploid plant genomes have a maternal (a) and a paternal haplotype (c), which differ at specific positions (b). Long reads belong to one or the other haplotype (d). The assembly graph separates haplophases in regions with sufficient differences between both parental haplotypes, but collapses them in identical (homozygous) regions (e). Resolving the assembly graph into final sequences is possible in four different ways (f): It is possible that both haplophases are resolved by connecting the two divergent blocks correctly (1), identical regions could be assigned to one haplophase leading to a less continuous second haplophase (2 and 3), or the identical region can cause an erroneous connection of the flanking distinct sequences (4). This illustration shows the analysis of a diploid genome, but the concept is generalisable to polyploids.

species were investigated by sequencing relatives with a lower ploidy (Kyriakidou et al., 2018; Schmutz et al., 2010; Zimin et al., 2017). Substantially increased read length and sequence accuracy of ONT and PacBio reads enabled the investigation of more challenging genomes. Objectives of current genome sequencing projects are the construction of phased genome sequences that represent both haplotypes by accurate haplophases (Girollet et al., 2019; Siadjeu et al., 2020; Sun et al., 2020). Accurate separation of haplophases is particularly important in highly heterozygous species like grapevine, in which alleles can differ by numerous presence/absence variations. While several assemblies of heterozygous species contain contigs representing two haplophases, it is not clear if contigs accurately represent a single haplotype. One major challenge to the accurate assembly of haplophases is the heterogeneous distribution of differences between the haplotypes. Regions rich in differences are easily separated into phases, but such regions are interleaved with homozygous regions that are more difficult to separate (Figure 4). The major challenge is to avoid switches between the haplophases in these homozygous regions.

Incorporation of external information, for example parental sequencing data are well-established approach to separate haplophases. TrioBinning identifies unique k-mers in each of the parental sequencing datasets and bins the reads of their offspring accordingly (Koren et al., 2018). This approach allows the separate assembly of both haplophases, avoiding phase-switching issues. Each assembly is resolving the structure of one haploid genome. Other approaches subject gametes to single cell sequencing (Campoy et al., 2020; Shi et al., 2019) because these cells contain only DNA of one of the haplotypes. The availability of HiFi reads enables the accurate assembly of haplophases (Zhou, Tang, et al., 2020). Another approach is based on high-throughput chromosome conformation capture (Hi-C) or Omni-C (Dovetail Genomics) data, which provide information about the physical proximity of different parts of the chromatin. Briefly, DNA strands are cross-linked with formaldehyde and digested by endonucleases. Cross-linked DNA fragments are ligated with an adapter in between and subjected to sequencing. It is similar to a mate pair library with huge insert sizes. Tools like ALLHiC (Zhang et al., 2019), hifiasm (Cheng et al., 2021) and FALCON-phase (Kronenberg et al., 2021) allow the integration of these data for a high-level scaffolding of large contigs in an allele-specific manner, thus paving the way for phased assemblies of heterozygous and polyploidy species.

8. Computational future of plant genomics

While sequencing costs drop and the amounts of data increase, the computational data analysis has become the major challenge. Higher raw-read accuracy is likely going to change this again, but the conversion of physical signals into sequence information (basecalling) during the actual sequencing process remains computationally intensive. For example, ONT's GridION is relying on the processors of graphic cards to perform basecalling in real time. Performing the basecalling after completion of a sequencing run on CPUs is an alternative, for example when using a MinION. Miles Benton maintains an excellent blog about technical details and gives advices about the best graphic cards for basecalling of ONT data (Benton, 2021). The primary analysis of PacBio sequencing

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

data involves multiple steps resulting in trace, pulse, base and FASTQ or BAM files. The base file is usually stored as it provides the basis for all secondary analyses. In contrast to second generation short-read sequencing technologies, it is important to store raw data (fast5 for ONT and base files for PacBio) of long-read sequencing runs. Rapid improvements in the basecalling algorithms (Amarasinghe et al., 2020) will allow drawing substantially more accurate information from the same raw reads in the future. The rapid development of new basecalling tools also poses a challenge to users looking for the best solution.

Genome assemblies based on noisy long reads often require a first correction step, which involves the computationally challenging all-vs-all alignment of reads. This step involves the generation of temporary files which are several times the size of the initial sequence data (FASTQ files). More stringent settings in the detection of matches between the reads can help reduce the disk space requirements in this step. The >99.5% accuracy of HiFi reads is a first step to reduce the computational costs of plant genome assemblies by an order of magnitude (Cheng et al., 2021; Mascher et al., 2021; Nurk et al., 2020) because alignments between reads can be restricted to almost perfect matches or the correction step can be skipped altogether.

Genome assemblies require high-performance hardware. However, their usage is characterised by peaks in memory and CPU consumption for assemblies and idle time while no assemblies are computed. Institutional compute clusters can make the necessary resources available to users for the assembly process, but not all institutions can offer this support. Commercial cloud computing offering large resources temporarily could be a good solution for groups that do not have access to high-performance hardware. However, data storage and transfer remains expensive. Several organisations already recognised this issue and offer computational resources and support for researchers, for example de.NBI (Belmann et al., 2019) and CYVERSE (https://cyverse.org/). As described for basecalling and read correction, the settings of the assembly process influence the required computational resources. There is a trade-off between the quality of a genome representation and the associated computational costs (Kaye & Wasserman, 2021). (Hi)Canu (Nurk et al., 2020; Zimin et al., 2017) produced the plant genome assembly of choice in many projects, but other assemblers like Flye (Kolmogorov et al., 2019) might be better if repetitive sequences are the focus of a study (Naish et al., 2021).

While genome assemblies are 'only' computationally challenging, the prediction of gene models and the functional annotation of predicted gene models will remain a challenge for the foreseeable future. The prediction of gene models is usually supported by RNA-Seq. The direct RNA sequencing offered by ONT or full length cDNA sequencing by PacBio or ONT is a good way to improve the annotation and detection of splicing isoforms. Given that multiple genome sequences of closely related plants are generated, the identification of gene models should be performed simultaneously on all sequences as implemented in the Comparative Annotation Toolkit (Fiddes et al., 2018). However, there are many other tools or pipelines including BRAKER2 (Brůna, Hoff, et al., 2020; Hoff et al., 2019), SNAP (Korf, 2004), GeneMark-EP+ (Brůna, Lomsadze, & Borodovsky, 2020) and Gnomon (Souvorov et al., 2018).

Many different tools for the analysis of long-read data are available and new ones are continuously developed. Every tool has its specific strengths and weaknesses with respect to applications, but this also depends on the nature of the data at hand. Therefore, there is a need for benchmarking studies to provide guidance to potential users. Benchmarking studies on short-read assemblers like the Assemblathons (Bradnam et al., 2013; Earl et al., 2011) were informative for many years until long-read sequencing technologies became the *de facto* standard for plant genome assemblies. However, a mechanism to continuously update the benchmarking results would be important for modern long-read assemblers. New software and technology versions are frequently released, thus making comparisons obsolete within months. There are efforts to optimise assemblers towards speed and reduced memory usage (Gatter et al., 2021; Haghshenas et al., 2020; Shafin et al., 2020). While this is important to complete extremely large plant genome assemblies and to reduce the environmental impact of bioinformatics, quality improvements are still of interest and would be beneficial for smaller genomes. Projects aiming for better assembly quality are often trying to achieve this through accurate separation of the haplophases (Chin et al., 2016; Koren et al., 2018; Nurk et al., 2020).

9. Conclusion

Genome sequencing is a rapidly developing field with an exponential growth in the amount of produced data and biological insights gained from them. Technological developments solve the long-standing assembly contiguity issue and enable novel analyses like the study of DNA modifications at a genome-wide scale. As a consequence, we as genomicists gain not only quantity, but also quality. The accurate separation of haplophases remains a challenge. Open science principles including an effective data sharing have been important in the past and will open even more opportunities in the future. Dropping sequencing costs and technological improvements will help to move from single reference genome sequences to pangenomics in order to better understand the genomic diversity within every species.

Acknowledgements

We thank Quantitative Plant Biology for the invitation to submit this review article. Some figures were generated using bioRender.com.

Financial support. B.P. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—436841671. I.I. and J.d.V. are part of the framework of MAdLand (http://madland.science, DFG priority programme 2237); J.d.V. is grateful for funding by the DFG (VR132/4-1). Work in the lab of J.d.V. is further supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 852725; ERC Starting Grant 'TerreStriAL'). B.X. is supported by the National Natural Science Foundation of China (32070249), and the Strategic Priority Research Programme of the Chinese Academy of Sciences (XDA26030104).

Conflicts of interest. B.P. was an invited speaker without financial compensation at a virtual conference (London Calling 2021) organised by Oxford Nanopore Technologies. J.V., I.I. and B.X. declare no conflicts of interest.

Authorship contributions. B.P. initiated and coordinated the project. All authors contributed to the manuscript and have approved the final version.

Data and availability statement. Data availability is not applicable to this article as no new data were created or analysed in this study.

References

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., ... **Zhou, X.** (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, **166**, 481–491.

Amarasinghe, S. L., **Su, S.**, **Dong, X.**, **Zappia, L.**, **Ritchie, M. E.**, & **Gouil, Q.** (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, **21**, 30.

Armbrust, E. V., **Berges, J. A.**, **Bowler, C.**, **Green, B. R.**, **Martinez, D.**, **Putnam, N. H.**, **Zhou, S.**, **Allen, A. E.**, **Apt, K. E.**, **Bechner, M.**, **Brzezinski, M. A.**, **Chaal, B. K.**, **Chiovitti, A.**, **Davis, A. K.**, **Demarest, M. S.**, **Detter, J. C.**, **Glavina, T.**, **Goodstein, D.**, **Hadi, M. Z.**, … **Rokhsar, D. S.** (2004). The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science*, **306**, 79–86.

Banks, J. A., **Nishiyama, T.**, **Hasebe, M.**, **Bowman, J. L.**, **Gribskov, M.**, **dePamphilis, C.**, **Albert, V. A.**, **Aono, N.**, **Aoyama, T.**, **Ambrose, B. A.**, et al. (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*, **332**, 960–963.

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.

Belmann, P., **Fischer, B.**, **Krüger, J.**, **Procházka, M.**, **Rasche, H.**, **Prinz, M.**, **Hanussek, M.**, **Lang, M.**, **Bartusch, F.**, **Gläßle, B.**, **Krüger, J.**, **Pühler, A.**, & **Sczyrba, A.** (2019). de.NBI cloud federation through ELIXIR AAI. *F1000Research*, **8**, 842.

Benton, M. C. (2021). GPU-musings. https://zenodo.org/record/5005787

Blaby-Haas, C. E., & **Merchant, S. S.** (2019). Comparative and functional algal genomics. *Annual Review of Plant Biology*, **70**, 605–638.

Bowles, A. M. C., **Paps, J.**, & **Bechtold, U.** (2021). Evolutionary origins of drought tolerance in spermatophytes. *Frontiers in Plant Science*, **12**, 655924.

Bowman, J. L., **Kohchi, T.**, **Yamato, K. T.**, **Jenkins, J.**, **Shu, S.**, **Ishizaki, K.**, **Yamaoka, S.**, **Nishihama, R.**, **Nakamura, Y.**, **Berger, F.**, et al. (2017). Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell*, **171**, 287–304.e15.

Bradnam, K. R., **Fass, J. N.**, **Alexandrov, A.**, **Baranay, P.**, **Bechner, M.**, **Birol, I.**, **Boisvert, S.**, **Chapman, J. A.**, **Chapuis, G.**, **Chikhi, R.**, **Chitsaz, H.**, **Chou, W.-C.**, **Corbeil, J.**, **Del Fabbro, C.**, **Docking, T. R.**, **Durbin, R.**, **Earl, D.**, **Emrich, S.**, **Fedotov, P.**, … **Korf, I. F.** (2013). Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, **2**, 2047-217X-2-10.

Branton, D., **Deamer, D. W.**, **Marziali, A.**, **Bayley, H.**, **Benner, S. A.**, **Butler, T.**, **Di Ventra, M.**, **Garaj, S.**, **Hibbs, A.**, **Huang, X.**, **Jovanovich, S. B.**, **Krstic, P. S.**, **Lindsay, S.**, **Ling, X. S.**, **Mastrangelo, C. H.**, **Meller, A.**, **Oliver, J. S.**, **Pershin, Y. V.**, **Ramsey, J. M.**, … **Schloss, J. A.** (2008). The potential and challenges of nanopore sequencing. *Nature Biotechnology*, **26**, 1146–1153.

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

Brůna, T., **Lomsadze, A.**, & **Borodovsky, M.** (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, **2**, lqaa026.

Cameron, D. L., **Di Stefano, L.**, & **Papenfuss, A. T.** (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, **10**, 3240.

Campoy, J. A., **Sun, H.**, **Goel, M.**, **Jiao, W.-B.**, **Folz-Donahue, K.**, **Wang, N.**, **Rubio, M.**, **Liu, C.**, **Kukat, C.**, **Ruiz, D.**, **Huettel, B.**, & **Schneeberger, K.** (2020). Gamete binning: Chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biology*, **21**, 306.

Carey, S. B., **Jenkins, J.**, **Lovell, J. T.**, **Maumus, F.**, **Sreedasyam, A.**, **Payton, A. C.**, **Shu, S.**, **Tiley, G. P.**, **Fernandez-Pozo, N.**, **Healey, A.**, et al. (2021). Gene-rich UV sex chromosomes harbor conserved regulators of sexual development. *Science Advances*, **7**, eabh2488.

Carta, A., **Bedini, G.**, & **Peruzzi, L.** (2020). A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytologist*, **228**, 1097–1106.

Causse, M., **Desplat, N.**, **Pascual, L.**, **Le Paslier, M.-C.**, **Sauvage, C.**, **Bauchet, G.**, **Bérard, A.**, **Bounon, R.**, **Tchoumakov, M.**, **Brunel, D.**, & **Bouchet, J.-P.** (2013). Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics*, **14**, 791.

Chater, C. C. C., **Caine, R. S.**, **Fleming, A. J.**, & **Gray, J. E.** (2017). Origins and evolution of stomatal development. *Plant Physiology*, **174**, 624–638.

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

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.

Cheng, S., **Xian, W.**, **Fu, Y.**, **Marin, B.**, **Keller, J.**, **Wu, T.**, **Sun, W.**, **Li, X.**, **Xu, Y.**, **Zhang, Y.**, **Wittek, S.**, **Reder, T.**, **Günther, G.**, **Gontcharov, A.**, **Wang, S.**, **Li, L.**, **Liu, X.**, **Wang, J.**, **Yang, H.**, … **Melkonian, M.** (2019). Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell*, **179**, 1057–1067.e14.

Chin, C.-S., **Peluso, P.**, **Sedlazeck, F. J.**, **Nattestad, M.**, **Concepcion, G. T.**, **Clum, A.**, **Dunn, C.**, **O'Malley, R.**, **Figueroa-Balderas, R.**, **Morales-Cruz, A.**, **Cramer, G. R.**, **Delledonne, M.**, **Luo, C.**, **Ecker, J. R.**, **Cantu, D.**, **Rank, D. R.**, & **Schatz, M. C.** (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, **13**, 1050–1054.

Choi, J. Y., **Lye, Z. N.**, **Groen, S. C.**, **Dai, X.**, **Rughani, P.**, **Zaaijer, S.**, **Harrington, E. D.**, **Juul, S.**, & **Purugganan, M. D.** (2020). Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. *Genome Biology*, **21**, 21.

Coelho, S. M., **Scornet, D.**, **Rousvoal, S.**, **Peters, N. T.**, **Dartevelle, L.**, **Peters, A. F.**, & **Cock, J. M.** (2012). *Ectocarpus*: A model organism for the brown algae. *Cold Spring Harbor Protocols*, **2012**, pdb.emo065821.

Darwin Tree of Life Project. (2021). Darwin Tree of Life Project—Wellcome Sanger Institute. https://www.sanger.ac.uk/collaboration/darwin-tree-of-life-project/

De Clerck, O., **Kao, S.-M.**, **Bogaert, K. A.**, **Blomme, J.**, **Foflonker, F.**, **Kwantes, M.**, **Vancaester, E.**, **Vanderstraeten, L.**, **Aydogdu, E.**, **Boesger, J.**, **Califano, G.**, **Charrier, B.**, **Clewes, R.**, **Del Cortona, A.**, **D'Hondt, S.**, **Fernandez-Pozo, N.**, **Gachon, C. M.**, **Hanikenne, M.**, **Lattermann, L.**, … **Bothwell, J. H.** (2018). Insights into the evolution of multicellularity from the sea lettuce genome. *Current Biology*, **28**, 2921–2933.e5.

de Koning, W., **Miladi, M.**, **Hiltemann, S.**, **Heikema, A.**, **Hays, J. P.**, **Flemming, S.**, **van den Beek, M.**, **Mustafa, D. A.**, **Backofen, R.**, **Grüning, B.**, & **Stubbs, A. P.** (2020). NanoGalaxy: Nanopore long-read sequencing data analysis in galaxy. *GigaScience*, **9**.

de Vries, J., & **Archibald, J. M.** (2017). Endosymbiosis: Did plastids evolve from a freshwater cyanobacterium? *Current Biology*, **27**, R103–R105.

de Vries, S., **Fürst-Jansen, J. M.**, **Irisarri, I.**, **Ashok, A. D.**, **Ischebeck, T.**, **Feussner, K.**, **Abreu, I. N.**, **Petersen, M.**, **Feussner, I.**, & **de Vries, J.** (2021). The evolution of the phenylpropanoid pathway entailed pronounced radiations and divergences of enzyme families. *The Plant Journal*, **107**, 975–1002.

Dohm, J. C., **Minoche, A. E.**, **Holtgräwe, D.**, **Capella-Gutiérrez, S.**, **Zakrzewski, F.**, **Tafer, H.**, **Rupp, O.**, **Sörensen, T. R.**, **Stracke, R.**, **Reinhardt, R.**, **Goesmann, A.**, **Kraft, T.**, **Schulz, B.**, **Stadler, P. F.**, **Schmidt, T.**, **Gabaldón, T.**, **Lehrach, H.**, **Weisshaar, B.**, & **Himmelbauer, H.** (2014). The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, **505**, 546–549.

Earl, D., **Bradnam, K.**, **St. John, J.**, **Darling, A.**, **Lin, D.**, **Fass, J.**, **Yu, H. O. K.**, **Buffalo, V.**, **Zerbino, D. R.**, **Diekhans, M.**, **Nguyen, N.**, **Ariyaratne, P. N.**, **Sung, W.-K.**, **Ning, Z.**, **Haimel, M.**, **Simpson, J. T.**, **Fonseca, N. A.**, **Birol, İ .**, **Docking, T. R.**, … **Paten, B.** (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, **21**, 2224–2241.

Edge, P., & **Bansal, V.** (2019). Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nature Communications*, **10**, 4660.

Eid, J., **Fehr, A.**, **Gray, J.**, **Luong, K.**, **Lyle, J.**, **Otto, G.**, **Peluso, P.**, **Rank, D.**, **Baybayan, P.**, **Bettman, B.**, **Bibillo, A.**, **Bjornson, K.**, **Chaudhuri, B.**, **Christians, F.**, **Cicero, R.**, **Clark, S.**, **Dalal, R.**, **deWinter, A.**, **Dixon, J.**, … **Turner, S.** (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Fiddes, I. T., **Armstrong, J.**, **Diekhans, M.**, **Nachtweide, S.**, **Kronenberg, Z. N.**, **Underwood, J. G.**, **Gordon, D.**, **Earl, D.**, **Keane, T.**, **Eichler, E. E.**, **Haussler, D.**, **Stanke, M.**, & **Paten, B.** (2018). Comparative annotation toolkit (CAT)—

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

Simultaneous clade and personal genome annotation. *Genome Research*, **28**, 1029–1038.

Filloramo, G. V., **Curtis, B. A.**, **Blanche, E.**, & **Archibald, J. M.** (2021). Re-examination of two diatom reference genomes using long-read sequencing. *BMC Genomics*, **22**, 379.

Fürst-Jansen, J. M. R., **de Vries, S.**, & **de Vries, J.** (2020). Evo-physio: On stress responses and the earliest land plants. *Journal of Experimental Botany*, **71**, 3254–3269.

Gatter, T., **von Löhneysen, S.**, **Fallmann, J.**, **Drozdova, P.**, **Hartmann, T.**, & **Stadler, P. F.** (2021). LazyB: Fast and cheap genome assembly. *Algorithms for Molecular Biology*, **16**, 8.

Girollet, N., **Rubio, B.**, **Lopez-Roques, C.**, **Valière, S.**, **Ollat, N.**, & **Bert, P.-F.** (2019). *De novo* phased assembly of the *Vitis riparia* grape genome. *Scientific Data*, **6**.

Goff, S. A., **Ricke, D.**, **Lan, T.-H.**, **Presting, G.**, **Wang, R.**, **Dunn, M.**, **Glazebrook, J.**, **Sessions, A.**, **Oeller, P.**, **Varma, H.**, **Hadley, D.**, **Hutchison, D.**, **Martin, C.**, **Katagiri, F.**, **Lange, B. M.**, **Moughamer, T.**, **Xia, Y.**, **Budworth, P.**, **Zhong, J.**, … **Briggs, S.** (2002). A draft sequence of the Rice genome (Oryza sativa L. ssp. japonica). *Science*, **296**, 92–100.

Gong, Z., & **Han, G.-Z.** (2021). Flourishing in water: The early evolution and diversification of plant receptor-like kinases. *The Plant Journal*, **106**, 174–184.

Grigoriev, I. V., **Hayes, R. D.**, **Calhoun, S.**, **Kamel, B.**, **Wang, A.**, **Ahrendt, S.**, **Dusheyko, S.**, **Nikitin, R.**, **Mondo, S. J.**, **Salamov, A.**, **Shabalov, I.**, & **Kuo, A.** (2021). PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Research*, **49**, D1004–D1011.

Haghshenas, E., **Asghari, H.**, **Stoye, J.**, **Chauve, C.**, & **Hach, F.** (2020). HASLR: Fast hybrid assembly of long reads. *iScience*, 23.

Hoff, K., **Lomsadze, A.**, **Borodovsky, M.**, & **Stanke, M.** (2019). Whole-genome annotation with BRAKER. *Methods in Molecular Biology (Clifton, N.J.)*, **1962**, 65–95.

Hon, T., **Mars, K.**, **Young, G.**, **Tsai, Y.-C.**, **Karalius, J. W.**, **Landolin, J. M.**, **Maurer, N.**, **Kudrna, D.**, **Hardigan, M. A.**, **Steiner, C. C.**, **Knapp, S. J.**, **Ware, D.**, **Shapiro, B.**, **Peluso, P.**, & **Rank, D. R.** (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, **7**, 399.

Hori, K., **Maruyama, F.**, **Fujisawa, T.**, **Togashi, T.**, **Yamamoto, N.**, **Seo, M.**, **Sato, S.**, **Yamada, T.**, **Mori, H.**, **Tajima, N.**, et al. (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nature Communications*, **5**, 3978.

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.**, **Coletta, R. D.**, **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**, 655–662. https://doi.org/10.1126/science.abg5289.

Hunt, S. P., **Jarvis, D. E.**, **Larsen, D. J.**, **Mosyakin, S. L.**, **Kolano, B. A.**, **Jackson, E. W.**, **Martin, S. L.**, **Jellen, E. N.**, & **Maughan, P. J.** (2020). A chromosome-scale assembly of the garden Orach (*Atriplex hortensis* L.) genome using Oxford Nanopore sequencing. *Frontiers in Plant Science*, **11**.

Jaillon, O., **Aury, J.-M.**, **Noel, B.**, **Policriti, A.**, **Clepet, C.**, **Casagrande, A.**, **Choisne, N.**, **Aubourg, S.**, **Vitulo, N.**, **Jubin, C.**, **Vezzi, A.**, **Legeai, F.**, **Hugueney, P.**, **Dasilva, C.**, **Horner, D.**, **Mica, E.**, **Jublot, D.**, **Poulain, J.**, **Bruyere, C.**, … The French–Italian Public Consortium for Grapevine Genome Characterization . (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Jain, M., **Tyson, J. R.**, **Loose, M.**, **Ip, C. L. C.**, **Eccles, D. A.**, **O'Grady, J.**, **Malla, S.**, **Leggett, R. M.**, **Wallerman, O.**, **Jansen, H. J.**, **Zalunin, V.**, **Birney, E.**, **Brown, B. L.**, **Snutch, T. P.**, **Olsen, H. E.**, & MinION Analysis and Reference Consortium. (2017). MinION analysis and reference consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Research*, **6**, 760.

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.

Jiao, C., **Srensen, I.**, **Sun, X.**, **Sun, H.**, **Behar, H.**, **Alseekh, S.**, **Philippe, G.**, **Palacio Lopez, K.**, **Sun, L.**, **Reed, R.**, **Jeon, S.**, **Kiyonami, R.**, **Zhang, S.**, **Fernie, A. R.**, **Brumer, H.**, **Domozych, D. S.**, **Fei, Z.**, & **Rose, J. K. C.** (2020). The *Penium margaritaceum* genome: Hallmarks of the origins of land plants. *Cell*, **181**, 1097–1111.e12.

Jiao, W.-B., & **Schneeberger, K.** (2020). Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nature Communications*, **11**, 989.

Jones, A., **Torkel, C.**, **Stanley, D.**, **Nasim, J.**, **Borevitz, J.**, & **Schwessinger, B.** (2021). High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. *PLoS One*, **16**, e0253830.

Karsten, L., **Bergen, D.**, **Drake, C.**, **Dymek, S.**, **Edich, M.**, **Haak, M.**, **Kerkhoff, D.**, **Kerkhoff, Y.**, **Liebers, M.**, **März, C.**, **Schlüter, L.**, **Schmidt, O.**, **Vinke, S.**, **Whitford, C.**, **Pucker, B.**, **Droste, J.**, **Rückert, C.**, **Müller, K.**, & **Kalinowski, J.** (2017). Expanding the genetic code. https://doi.org/10.13140/RG.2.2.20342.91203.

Kaye, A. M., & **Wasserman, W. W.** (2021). The genome atlas: Navigating a new era of reference genomes. *Trends in Genetics*, **37**, 807–818.

Keeling, P. J. (2013). The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annual Review of Plant Biology*, **64**, 583–607

Kirbis, A., **Waller, M.**, **Ricca, M.**, **Bont, Z.**, **Neubauer, A.**, **Goffinet, B.**, & **Szövényi, P.** (2020). Transcriptional landscapes of divergent sporophyte development in two mosses, *Physcomitrium (Physcomitrella) patens* and *Funaria hygrometrica*. *Frontiers in Plant Science*, **11**, 747

Kolmogorov, M., **Yuan, J.**, **Lin, Y.**, & **Pevzner, P. A.** (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, **37**, 540–546.

Koren, S., **Rhie, A.**, **Walenz, B. P.**, **Dilthey, A. T.**, **Bickhart, D. M.**, **Kingan, S. B.**, **Hiendleder, S.**, **Williams, J. L.**, **Smith, T. P. L.**, & **Phillippy, A. M.** (2018). *De novo* assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, **36**, 1174–1182.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Kronenberg, Z. N., **Rhie, A.**, **Koren, S.**, **Concepcion, G. T.**, **Peluso, P.**, **Munson, K. M.**, **Porubsky, D.**, **Kuhn, K.**, **Mueller, K. A.**, **Low, W. Y.**, **Hiendleder, S.**, **Fedrigo, O.**, **Liachko, I.**, **Hall, R. J.**, **Phillippy, A. M.**, **Eichler, E. E.**, **Williams, J. L.**, **Smith, T. P. L.**, **Jarvis, E. D.**, … **Kingan, S. B.** (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*, **12**, 1935.

Kyriakidou, M., **Tai, H. H.**, **Anglin, N. L.**, **Ellis, D.**, & **Strömvik, M. V.** (2018). Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science*, **9**.

Lang, D., **Ullrich, K. K.**, **Murat, F.**, **Fuchs, J.**, **Jenkins, J.**, **Haas, F. B.**, **Piednoel, M.**, **Gundlach, H.**, **Van Bel, M.**, **Meyberg, R.**, et al. (2018). The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *The Plant Journal*, **93**, 515–533.

Lewin, H. A., **Graves, J. A. M.**, **Ryder, O. A.**, **Graphodatsky, A. S.**, & **O'Brien, S. J.** (2019). Precision nomenclature for the new genomics. *GigaScience*, **8**.

Lewin, H. A., **Robinson, G. E.**, **Kress, W. J.**, **Baker, W. J.**, **Coddington, J.**, **Crandall, K. A.**, **Durbin, R.**, **Edwards, S. V.**, **Forest, F.**, **Gilbert, M. T. P.**, **Goldstein, M. M.**, **Grigoriev, I. V.**, **Hackett, K. J.**, **Haussler, D.**, **Jarvis, E. D.**, **Johnson, W. E.**, **Patrinos, A.**, **Richards, S.**, **Castilla-Rubio, J. C.**, … **Zhang, G.** (2018). Earth BioGenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, **115**, 4325–4333.

Li, C., **Xiang, X.**, **Huang, Y.**, **Zhou, Y.**, **An, D.**, **Dong, J.**, **Zhao, C.**, **Liu, H.**, **Li, Y.**, **Wang, Q.**, **Du, C.**, **Messing, J.**, **Larkins, B. A.**, **Wu, Y.**, & **Wang, W.** (2020). Long-read sequencing reveals genomic structural variations that underlie creation of quality protein maize. *Nature Communications*, **11**, 17.

Li, F.-W., **Brouwer, P.**, **Carretero-Paulet, L.**, **Cheng, S.**, **de Vries, J.**, **Delaux, P.-M.**, **Eily, A.**, **Koppers, N.**, **Kuo, L.-Y.**, **Li, Z.**, et al. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants*, **4**, 460–472.

Li, F.-W., **Melkonian, M.**, **Rothfels, C. J.**, **Villarreal, J. C.**, **Stevenson, D. W.**, **Graham, S. W.**, **Wong, G. K.-S.**, **Pryer, K. M.**, & **Mathews, S.** (2015). Phytochrome diversity in green plants and the origin of canonical plant phytochromes. *Nature Communications*, **6**, 7852.

Li, F.-W., **Nishiyama, T.**, **Waller, M.**, **Frangedakis, E.**, **Keller, J.**, **Li, Z.**, **Fernandez-Pozo, N.**, **Barker, M. S.**, **Bennett, T.**, **Blázquez, M. A.**, et al. (2020). *Anthoceros* genomes illuminate the origin of land plants and the unique biology of hornworts. *Nature Plants*, **6**, 259–272.

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

Li, L., Wang, S., Wang, H., Sahu, S. K., Marin, B., Li, H., Xu, Y., Liang, H., Li, Z., Cheng, S., Reder, T., Cebi, Z., Wittek, S., Petersen, M., Melkonian, B., Du, H., Yang, H., Wang, J., Wong, G. K.-S., … Liu, H. (2020). The genome of *Prasinoderma coloniale* unveils the existence of a third phylum within green plants. *Nature Ecology & Evolution*, **4**, 1220–1231.

Li, Z., Parris, S., & Saski, C. A. (2020). A simple plant high-molecular-weight DNA extraction method suitable for single-molecule technologies. *Plant Methods*, **16**, 38.

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., Liu, C., Nick, P., Du, F., Fan, P., Mao, R., Zhu, Y., Deng, W., Yang, M., Huang, H., Liu, Y., Ding, Y., Liu, X., Jiang, J., … Dong, Y. (2019). Whole-genome resequencing of 472 *Vitis* accessions for grapevine diversity and demographic history analyses. *Nature Communications*, **10**, 1190.

Liu, H., Wang, X., Wang, G., Cui, P., Wu, S., Ai, C., Hu, N., Li, A., He, B., Shao, X., Wu, Z., Feng, H., Chang, Y., Mu, D., Hou, J., Dai, X., Yin, T., Ruan, J., & Cao, F. (2021). The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. *Nature Plants*, **7**, 748–756.

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.

Lv, Q., Li, W., Sun, Z., Ouyang, N., Jing, X., He, Q., Wu, J., Zheng, J., Zheng, J., Tang, S., Zhu, R., Tian, Y., Duan, M., Tan, Y., Yu, D., Sheng, X., Sun, X., Jia, G., Gao, H., … Yuan, D. (2020). Resequencing of 1,143 indica rice accessions reveals important genetic variations and different heterosis patterns. *Nature Communications*, **11**, 4778.

Ma, P.-F., Liu, Y.-L., Jin, G.-H., Liu, J.-X., Wu, H., He, J., Guo, Z.-H., & Li, D.-Z. (2021). The *Pharus latifolius* genome bridges the gap of early grass evolution. *The Plant Cell*, **33**, 846–864.

Ma, X., Vaistij, F. E., Li, Y., Jansen van Rensburg, W. S., Harvey, S., Bairu, M. W., Venter, S. L., Mavengahama, S., Ning, Z., Graham, I. A., Van Deynze, A., Van de Peer, Y., & Denby, K. J. (2021). A chromosome-level *Amaranthus cruentus* genome assembly highlights gene family evolution and biosynthetic gene clusters that may underpin the nutritional value of this traditional crop. *The Plant Journal*, **107**, 613–628.

Maghini, D. G., Moss, E. L., Vance, S. E., & Bhatt, A. S. (2021). Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. *Nature Protocols*, **16**, 458–471.

Marchant, D. B., Sessa, E. B., Wolf, P. G., Heo, K., Barbazuk, W. B., Soltis, P. S., & Soltis, D. E. (2019). The C-Fern (*Ceratopteris richardii*) genome: Insights into plant genome evolution with the first partial homosporous fern genome assembly. *Scientific Reports*, **9**, 18181.

Marks, R. A., Smith, J. J., Cronk, Q., Grassa, C. J., & McLetchie, D. N. (2019). Genome of the tropical plant *Marchantia inflexa*: Implications for sex chromosome evolution and dehydration tolerance. *Scientific Reports*, **9**, 8722.

Marroni, F., Pinosio, S., & Morgante, M. (2014). Structural variation and genome complexity: Is dispensable really dispensable? *Current Opinion in Plant Biology*, **18**, 31–36.

Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C. S., Ens, J., Gundlach, H., Boston, L. B., Tulpová, Z., Holden, S., Hernández-Pinzón, I., Scholz, U., Mayer, K. F. X., Spannagl, M., Pozniak, C. J., Sharpe, A. G., Šimková, H., Moscou, M. J., … Stein, N. (2021). Long-read sequence assembly: A technical evaluation in barley. *The Plant Cell*, **6**, 1888–1906.

Menand, B., Yi, K., Jouannic, S., Hoffmann, L., Ryan, E., Linstead, P., Schaefer, D. G., & Dolan, L. (2007). An ancient mechanism controls the development of cells with a rooting function in land plants. *Science*, **316**, 1477–1480.

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

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., Loudet, O., Weigel, D., & Ecker, J. R. (2018). High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nature Communications*, **9**, 541.

Michael, T. P., & VanBuren, R. (2020). Building near-complete plant genomes. *Current Opinion in Plant Biology*, **54**, 26–33.

Murray, M. G., & Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research*, **8**, 4321–4326.

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Lambing, C., Kuo, P., Yelina, N., Hartwick, N., Colt, K., Smith, L. M., Ton, J., Kakutani, T., Martienssen, R. A., Schneeberger, K., Lysak, M. A., … Henderson, I. R., (2021). The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science,* **374**, eabi7489. https://doi.org/10.1126/science.abi7489.

Nishiyama, T., Sakayama, H., de Vries, J., Buschmann, H., Saint-Marcoux, D., Ullrich, K. K., Haas, F. B., Vanderstraeten, L., Becker, D., Lang, D., Vosolsobě, S., Rombauts, S., Wilhelmsson, P. K. I., Janitza, P., Kern, R., Heyl, A., Rümpler, F., Villalobos, L. I. A. C., Clay, J. M., … Rensing, S. A. (2018). The Chara genome: Secondary complexity and implications for plant terrestrialization. *Cell*, **174**, 448–464.e24.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2021). The complete sequence of a human genome. bioRxiv: 2021.05.26.445798.

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., Miga, K. H., Eichler, E. E., Phillippy, A. M., & Koren, S. (2020). HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research,* **30**, 1291–1305.

Nützmann, H.-W., Huang, A., & Osbourn, A. (2016). Plant metabolic clusters—From genetics to genomics. *New Phytologist*, **211**, 771–789.

O'Donnell, S., Chaux, F., & Fischer, G. (2020). Highly contiguous Nanopore genome assembly of *Chlamydomonas reinhardtii* CC-1690. *Microbiology Resource Announcements*, **9**, e00726-20.

Oliva, M., Milicchio, F., King, K., Benson, G., Boucher, C., & Prosperi, M. (2020). Portable nanopore analytics: Are we there yet? *Bioinformatics*, **36**, 4399–4405.

Olson, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A. G., Johanson, E., Boja, E., Maier, E. J., Serang, O., Jáspez, D., Lorenzo-Salazar, J. M., Muñoz-Barrera, A., Rubio-Rodríguez, L. A., Flores, C., Kyriakidis, K., Malousi, A., Shafin, K., Pesout, T., … Zook, J. M. (2021). precisionFDA truth challenge V2: Calling variants from short- and long-reads in difficult-to-map regions. https://doi.org/10.1101/2020.11.13.380741.

ONT. (2021a). *R10.3: The newest nanopore for high accuracy nanopore sequencing—Now available in store*. Oxford Nanopore Technologies.

ONT. (2021b). *New nanopore sequencing chemistry in developers' hands; set to deliver Q20+ (99%+) "raw read" accuracy*. Oxford Nanopore Technologies.

Paajanen, P., Kettleborough, G., López-Girona, E., Giolai, M., Heavens, D., Baker, D., Lister, A., Cugliandolo, F., Wilde, G., Hein, I., Macaulay, I., Bryan, G. J., & Clark, M. D. (2019). A critical comparison of technologies for a plant genome sequencing project. *GigaScience*, **8**, giy163.

Palatnick, A., Zhou, B., Ghedin, E., & Schatz, M. C. (2020). iGenomics: Comprehensive DNA sequence analysis on your smartphone. *GigaScience*, **9**.

Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., Hall, A., Barton, G. J., & Simpson, G. G. (2019). Nanopore direct RNA sequencing maps an Arabidopsis N6 methyladenosine epitranscriptome. *ELife,* **9**, e49658. https://doi.org/10.7554/eLife.49658.

Payne, A., Holmes, N., Rakyan, V., & Loose, M. (2019). BulkVis: A graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*, **35**, 2193–2198.

Pederson, E. R. A., Warshan, D., & Rasmussen, U. (2019). Genome sequencing of *Pleurozium schreberi*: The assembled and annotated draft genome of a *Pleurocarpous* feather Moss. *G3 Genes|Genomes|Genetics*, **9**, 2791–2797.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, **36**, 983–987.

Price, D. C., Goodenough, U. W., Roth, R., Lee, J.-H., Kariyawasam, T., Mutwil, M., Ferrari, C., Facchinelli, F., Ball, S. G., Cenci, U., Chan, C. X., Wagner, N. E., Yoon, H. S., Weber, A. P. M., & Bhattacharya, D. (2019). Analysis of an improved *Cyanophora paradoxa* genome assembly. *DNA Research*, **26**, 287–299.

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

Provart, N. J., **Brady, S. M.**, **Parry, G.**, **Schmitz, R. J.**, **Queitsch, C.**, **Bonetta, D.**, **Waese, J.**, **Schneeberger, K.**, & **Loraine, A. E.** (2020). Anno genominis XX: 20 years of *Arabidopsis* genomics. *The Plant Cell*, **33**, 832–845.

Pucker, B., **Holtgräwe, D.**, **Stadermann, K. B.**, **Frey, K.**, **Huettel, B.**, **Reinhardt, R.**, & **Weisshaar, B.** (2019). A chromosome-level sequence assembly reveals the structure of the *Arabidopsis thaliana* Nd-1 genome and its gene set. *PLoS One*, **14**, e0216233.

Pucker, B., **Kleinbölting, N.**, & **Weisshaar, B.** (2021). Large scale genomic rearrangements in selected *Arabidopsis thaliana* T-DNA lines are caused by T-DNA insertion mutagenesis. *BMC Genomics*, **22**, 599.

Radakovits, R., **Jinkerson, R. E.**, **Fuerstenberg, S. I.**, **Tae, H.**, **Settlage, R. E.**, **Boore, J. L.**, & **Posewitz, M. C.** (2012). Draft genome sequence and genetic transformation of the oleaginous alga *Nannochloropsis gaditana*. *Nature Communications*, **3**, 686.

Radhakrishnan, G. V., **Keller, J.**, **Rich, M. K.**, **Vernié, T.**, **Mbadinga Mbadinga, D. L.**, **Vigneron, N.**, **Cottret, L.**, **Clemente, H. S.**, **Libourel, C.**, **Cheema, J.**, et al. (2020). An ancestral signalling pathway is conserved in intracellular symbioses-forming plant lineages. *Nature Plants*, **6**, 280–289.

Rai, A., **Hirakawa, H.**, **Nakabayashi, R.**, **Kikuchi, S.**, **Hayashi, K.**, **Rai, M.**, **Tsugawa, H.**, **Nakaya, T.**, **Mori, T.**, **Nagasaki, H.**, **Fukushi, R.**, **Kusuya, Y.**, **Takahashi, H.**, **Uchiyama, H.**, **Toyoda, A.**, **Hikosaka, S.**, **Goto, E.**, **Saito, K.**, … **Yamazaki, M.** (2021). Chromosome-level genome assembly of *Ophiorrhiza pumila* reveals the evolution of camptothecin biosynthesis. *Nature Communications*, **12**, 405.

Resemann, H. C., **Herrfurth, C.**, **Feussner, K.**, **Hornung, E.**, **Ostendorf, A. K.**, **Gömann, J.**, **Mittag, J.**, **van Gessel, N.**, **de, V. J.**, **Ludwig-Müller, J.**, **Markham, J.**, **Reski, R.**, & **Feussner, I.** (2021). Convergence of sphingolipid desaturation across over 500 million years of plant evolution. *Nature Plants*, **7**, 219–232.

Rhee, S. Y., & **Mutwil, M.** (2014). Towards revealing the functions of all genes in plants. *Trends in Plant Science*, **19**, 212–221.

Rothberg, J. M., **Hinz, W.**, **Rearick, T. M.**, **Schultz, J.**, **Mileski, W.**, **Davey, M.**, **Leamon, J. H.**, **Johnson, K.**, **Milgrew, M. J.**, **Edwards, M.**, **Hoon, J.**, **Simons, J. F.**, **Marran, D.**, **Myers, J. W.**, **Davidson, J. F.**, **Branting, A.**, **Nobile, J. R.**, **Puc, B. P.**, **Light, D.**, … **Bustillo, J.** (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Ruprecht, C., **Proost, S.**, **Hernandez-Coronado, M.**, **Ortiz-Ramirez, C.**, **Lang, D.**, **Rensing, S. A.**, **Becker, J. D.**, **Vandepoele, K.**, & **Mutwil, M.** (2017). Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. *The Plant Journal*, **90**, 447–465.

Samarakoon, H., **Punchihewa, S.**, **Senanayake, A.**, **Hammond, J. M.**, **Stevanovski, I.**, **Ferguson, J. M.**, **Ragel, R.**, **Gamaarachchi, H.**, & **Deveson, I. W.** (2020). Genopo : A nanopore sequencing analysis toolkit for portable android devices. *Communications Biology*, **3**, 1–5.

Sato, S., **Tabata, S.**, **Hirakawa, H.**, **Asamizu, E.**, **Shirasawa, K.**, **Isobe, S.**, **Kaneko, T.**, **Nakamura, Y.**, **Shibata, D.**, **Aoki, K.**, **Egholm, M.**, **Knight, J.**, **Bogden, R.**, **Li, C.**, **Shuang, Y.**, **Xu, X.**, **Pan, S.**, **Cheng, S.**, **Liu, X.**, … Universitat Pompeu Fabra. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Schilbert, H. M., **Rempel, A.**, & **Pucker, B.** (2020). Comparison of read mapping and variant calling tools for the analysis of plant NGS data. *Plants*, **9**, 439.

Schmutz, J., **Cannon, S. B.**, **Schlueter, J.**, **Ma, J.**, **Mitros, T.**, **Nelson, W.**, **Hyten, D. L.**, **Song, Q.**, **Thelen, J. J.**, **Cheng, J.**, **Xu, D.**, **Hellsten, U.**, **May, G. D.**, **Yu, Y.**, **Sakurai, T.**, **Umezawa, T.**, **Bhattacharyya, M. K.**, **Sandhu, D.**, **Valliyodan, B.**, … **Jackson, S. A.** (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Shafin, K., **Pesout, T.**, **Lorig-Roach, R.**, **Haukness, M.**, **Olsen, H. E.**, **Bosworth, C.**, **Armstrong, J.**, **Tigyi, K.**, **Maurer, N.**, **Koren, S.**, **Sedlazeck, F. J.**, **Marschall, T.**, **Mayes, S.**, **Costa, V.**, **Zook, J. M.**, **Liu, K. J.**, **Kilburn, D.**, **Sorensen, M.**, **Munson, K. M.**, … **Paten, B.** (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, **38**, 1044–1053.

Shi, D., **Wu, J.**, **Tang, H.**, **Yin, H.**, **Wang, H.**, **Wang, R.**, **Wang, R.**, **Qian, M.**, **Wu, J.**, **Qi, K.**, **Xie, Z.**, **Wang, Z.**, **Zhao, X.**, & **Zhang, S.** (2019). Single-pollen-cell sequencing for gamete-based phased diploid genome assembly in plants. *Genome Research*, **29**, 1889–1899.

Siadjeu, C., **Pucker, B.**, **Viehöver, P.**, **Albach, D. C.**, & **Weisshaar, B.** (2020). High contiguity de novo genome sequence assembly of trifoliate yam (*Dioscorea dumetorum*) using long read sequencing. *Genes*, **11**, 274.

Sibbald, S. J., & **Archibald, J. M.** (2020). Genomic insights into plastid evolution. *Genome Biology and Evolution*, **12**, 978–990.

Sielemann, K., **Weisshaar, B.**, & **Pucker, B.** (2021). Reference-based QUantification of gene dispensability (QUOD). *Plant Methods*, **17**, 18.

Silva, A. T., **Gao, B.**, **Fisher, K. M.**, **Mishler, B. D.**, **Ekwealor, J. T. B.**, **Stark, L. R.**, **Li, X.**, **Zhang, D.**, **Bowker, M. A.**, **Brinda, J. C.**, et al. (2021). To dry perchance to live: Insights from the genome of the desiccation-tolerant biocrust moss *Syntrichia caninervis*. *The Plant Journal*, **105**, 1339–1356.

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.

Soorni, A., **Haak, D.**, **Zaitlin, D.**, & **Bombarely, A.** (2017). Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC Genomics*, **18**, 49.

Souvorov, A., **Kapustin, Y.**, **Kiryutin, B.**, **Chetvernin, V.**, **Tatusova, T.**, & **Lipman, D.** (2018). Gnomon—The NCBI eukaryotic gene prediction tool. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/

Stein, J. C., **Yu, Y.**, **Copetti, D.**, **Zwickl, D. J.**, **Zhang, L.**, **Zhang, C.**, **Chougule, K.**, **Gao, D.**, **Iwata, A.**, **Goicoechea, J. L.**, **Wei, S.**, **Wang, J.**, **Liao, Y.**, **Wang, M.**, **Jacquemin, J.**, **Becker, C.**, **Kudrna, D.**, **Zhang, J.**, **Londono, C. E. M.**, … **Wing, R. A.** (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nature Genetics*, **50**, 285–296.

Strassert, J. F. H., **Irisarri, I.**, **Williams, T. A.**, & **Burki, F.** (2021). A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nature Communications*, **12**, 1879.

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.

Szövényi, P., **Gunadi, A.**, & **Li, F.-W.** (2021). Charting the genomic landscape of seed-free plants. *Nature Plants*, **7**, 554–565.

Tao, Y., **Luo, H.**, **Xu, J.**, **Cruickshank, A.**, **Zhao, X.**, **Teng, F.**, **Hathorn, A.**, **Wu, X.**, **Liu, Y.**, **Shatte, T.**, **Jordan, D.**, **Jing, H.**, & **Mace, E.** (2021). Extensive variation within the pan-genome of cultivated and wild sorghum. *Nature Plants*, **7**, 766–773.

The long view on sequencing (2018) *Nature Biotechnology*, **36**, 287–287.

Tuskan, G. A., **DiFazio, S.**, **Jansson, S.**, **Bohlmann, J.**, **Grigoriev, I.**, **Hellsten, U.**, **Putnam, N.**, **Ralph, S.**, **Rombauts, S.**, **Salamov, A.**, **Schein, J.**, **Sterck, L.**, **Aerts, A.**, **Bhalerao, R. R.**, **Bhalerao, R. P.**, **Blaudez, D.**, **Boerjan, W.**, **Brun, A.**, **Brunner, A.**, … **Rokhsar, D.** (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

VanBuren, R., **Wai, C. M.**, **Ou, S.**, **Pardo, J.**, **Bryant, D.**, **Jiang, N.**, **Mockler, T. C.**, **Edger, P.**, & **Michael, T. P.** (2018). Extreme haplotype variation in the desiccation-tolerant clubmoss *Selaginella lepidophylla*. *Nature Communications*, **9**, 13.

Vereecke, N., **Bokma, J.**, **Haesebrouck, F.**, **Nauwynck, H.**, **Boyen, F.**, **Pardon, B.**, & **Theuns, S.** (2020). High quality genome assemblies of *Mycoplasma bovis* using a taxon-specific Bonito basecaller for MinION and Flongle long-read nanopore sequencing. *BMC Bioinformatics*, **21**, 517.

Vilanova, S., **Alonso, D.**, **Gramazio, P.**, **Plazas, M.**, **García-Fortea, E.**, **Ferrante, P.**, **Schmidt, M.**, **Díez, M. J.**, **Usadel, B.**, **Giuliano, G.**, & **Prohens, J.** (2020). SILEX: A fast and inexpensive high-quality DNA extraction method suitable for multiple sequencing platforms and recalcitrant plant species. *Plant Methods*, **16**, 110.

Walkowiak, S., **Gao, L.**, **Monat, C.**, **Haberer, G.**, **Kassa, M. T.**, **Brinton, J.**, **Ramirez-Gonzalez, R. H.**, **Kolodziej, M. C.**, **Delorean, E.**, **Thambugala, D.**, **Klymiuk, V.**, **Byrns, B.**, **Gundlach, H.**, **Bandi, V.**, **Siri, J. N.**, **Nilsen, K.**, **Aquino, C.**, **Himmelbach, A.**, **Copetti, D.**, … **Pozniak, C. J.** (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature*, **588**, 277–283.

Wang, O., **Chin, R.**, **Cheng, X.**, **Wu, M. K. Y.**, **Mao, Q.**, **Tang, J.**, **Sun, Y.**, **Anderson, E.**, **Lam, H. K.**, **Chen, D.**, **Zhou, Y.**, **Wang, L.**, **Fan, F.**, **Zou, Y.**, **Xie, Y.**,

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press

Zhang, R. Y., **Drmanac, S.**, **Nguyen, D.**, **Xu, C.**, … **Peters, B. A.** (2019). Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Research*, **29**, 798–808.

Wang, S., **Li, L.**, **Li, H.**, **Sahu, S. K.**, **Wang, H.**, **Xu, Y.**, **Xian, W.**, **Song, B.**, **Liang, H.**, **Cheng, S.**, **Chang, Y.**, **Song, Y.**, **Cebi, Z.**, **Wittek, S.**, **Reder, T.**, **Peterson, M.**, **Yang, H.**, **Wang, J.**, **Melkonian, B.**, … **Liu, X.** (2020). Genomes of early-diverging streptophyte algae shed light on plant terrestrialization. *Nature Plants*, **6**, 95–106.

Wang, X., **Chen, S.**, **Ma, X.**, **AEJ, Y.**, **Chaluvadi, S. R.**, **Johnson, M. S.**, **Gangashetty, P.**, **Hamidou, F.**, **Sanogo, M. D.**, **Zwaenepoel, A.**, **Wallace, J.**, **Van de Peer, Y.**, **Bennetzen, J. L.**, & **Van Deynze, A.** (2021). Genome sequence and genetic diversity analysis of an under-domesticated orphan crop, white fonio (*Digitaria exilis*). *GigaScience*, **10**.

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

Wick, R. R., **Judd, L. M.**, **Gorrie, C. L.**, & **Holt, K. E.** (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, **13**, e1005595.

Wickell, D., **Kuo, L.-Y.**, **Yang, H.-P.**, **Dhabalia Ashok, A.**, **Irisarri, I.**, **Dadras, A.**, **de Vries, S.**, **de Vries, J.**, **Huang, Y.-M.**, **Li, Z.**, et al. (2021). Underwater CAM photosynthesis elucidated by *Isoetes* genome. *Nature Communications*, **12**, 6348.

Xu, B., **Ohtani, M.**, **Yamaguchi, M.**, **Toyooka, K.**, **Wakazaki, M.**, **Sato, M.**, **Kubo, M.**, **Nakano, Y.**, **Sano, R.**, **Hiwatashi, Y.**, **Murata, T.**, **Kurata, T.**, **Yoneda, A.**, **Kato, K.**, **Hasebe, M.**, & **Demura, T.** (2014). Contribution of NAC transcription factors to plant adaptation to land. *Science (New York, N.Y.)*, **343**, 1505–1508.

Xu, B., **Taylor, L.**, **Pucker, B.**, **Feng, T.**, **Glover, B. J.**, & **Brockington, S. F.** (2021). The land plant-specific MIXTA-MYB lineage is implicated in the early evolution of the plant cuticle and the colonization of land. *New Phytologist*, **229**, 2324–2338.

Xu, Z., **Xin, T.**, **Bartels, D.**, **Li, Y.**, **Gu, W.**, **Yao, H.**, **Liu, S.**, **Yu, H.**, **Pu, X.**, **Zhou, J.**, et al. (2018). Genome analysis of the ancient Tracheophyte *Selaginella tamariscina* reveals evolutionary features relevant to the acquisition of desiccation tolerance. *Molecular Plant*, **11**, 983–994.

Yu, J., **Hu, S.**, **Wang, J.**, **Wong, G. K.-S.**, **Li, S.**, **Liu, B.**, **Deng, Y.**, **Dai, L.**, **Zhou, Y.**, **Zhang, X.**, **Cao, M.**, **Liu, J.**, **Sun, J.**, **Tang, J.**, **Chen, Y.**, **Huang, X.**, **Lin, W.**, **Ye, C.**, **Tong, W.**, … **Yang, H.** (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.

Yu, J., **Li, L.**, **Wang, S.**, **Dong, S.**, **Chen, Z.**, **Patel, N.**, **Goffinet, B.**, **Chen, H.**, **Liu, H.**, **Liu, Y.**, et al. (2020). Draft genome of the aquatic moss *Fontinalis antipyretica* (Fontinalaceae, Bryophyta). *Gigabyte*, **2020**, 1–9.

Zhang, J., **Fu, X.-X.**, **Li, R.-Q.**, **Zhao, X.**, **Liu, Y.**, **Li, M.-H.**, **Zwaenepoel, A.**, **Ma, H.**, **Goffinet, B.**, **Guan, Y.-L.**, et al. (2020). The hornwort genome and early land plant evolution. *Nature Plants*, **6**, 107–118.

Zhang, L., **Chen, F.**, **Zhang, X.**, **Li, Z.**, **Zhao, Y.**, **Lohaus, R.**, **Chang, X.**, **Dong, W.**, **Ho, S. Y. W.**, **Liu, X.**, **Song, A.**, **Chen, J.**, **Guo, W.**, **Wang, Z.**, **Zhuang, Y.**, **Wang, H.**, **Chen, X.**, **Hu, J.**, **Liu, Y.**, … **Tang, H.** (2020). The water lily genome and the early evolution of flowering plants. *Nature*, **577**, 79–84.

Zhang, X., **Zhang, S.**, **Zhao, Q.**, **Ming, R.**, & **Tang, H.** (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants*, **5**, 833–845.

Zheng, J., **Meinhardt, L. W.**, **Goenaga, R.**, **Zhang, D.**, & **Yin, Y.** (2021). The chromosome-level genome of dragon fruit reveals whole-genome duplication and chromosomal co-localization of betacyanin biosynthetic genes. *Horticulture Research*, **8**, 1–16.

Zhou, Q., **Tang, D.**, **Huang, W.**, **Yang, Z.**, **Zhang, Y.**, **Hamilton, J. P.**, **Visser, R. G. F.**, **Bachem, C. W. B.**, **Robin Buell, C.**, **Zhang, Z.**, **Zhang, C.**, & **Huang, S.** (2020). Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nature Genetics*, **52**, 1018–1023.

Zhou, Y., **Chebotarov, D.**, **Kudrna, D.**, **Llaca, V.**, **Lee, S.**, **Rajasekar, S.**, **Mohammed, N.**, **Al-Bader, N.**, **Sobel-Sorenson, C.**, **Parakkal, P.**, **Arbelaez, L. J.**, **Franco, N.**, **Alexandrov, N.**, **Hamilton, N. R. S.**, **Leung, H.**, **Mauleon, R.**, **Lorieux, M.**, **Zuccolo, A.**, **McNally, K.**, … **Wing, R. A.** (2020). A platinum standard pan-genome resource that represents the population structure of Asian rice. *Scientific Data*, **7**, 113.

Zimin, A. V., **Puiu, D.**, **Luo, M.-C.**, **Zhu, T.**, **Koren, S.**, **Marçais, G.**, **Yorke, J. A.**, **Dvořák, J.**, & **Salzberg, S. L.** (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research*, **27**, 787–792.

https://doi.org/10.1017/qpb.2021.18 Published online by Cambridge University Press