Adv Biochem Eng Biotechnol (2018) 164: 11–36
DOI: 10.1007/10_2017_46

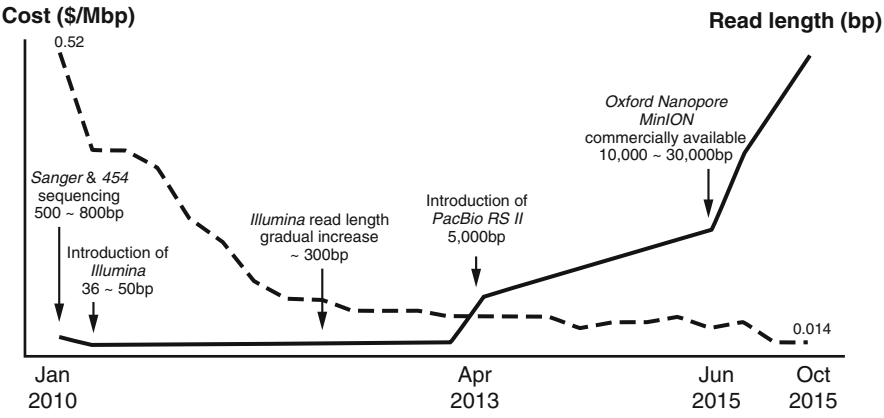
Published online: 8 March 2018

Advances in Sequencing and Resequencing in Crop Plants

Pradeep R. Marri, Liang Ye, Yi Jia, Ke Jiang, and Steven D. Rounsley

Abstract DNA sequencing technologies have changed the face of biological research over the last 20 years. From reference genomes to population level resequencing studies, these technologies have made significant contributions to our understanding of plant biology and evolution. As the technologies have increased in power, the breadth and complexity of the questions that can be asked has increased. Along with this, the challenges of managing unprecedented quantities of sequence data are mounting. This chapter describes a few aspects of the journey so far and looks forward to what may lie ahead.

Graphical Abstract



P. R. Marri, L. Ye, Y. Jia, and K. Jiang
Dow AgroSciences, Indianapolis, IN, USA

S. D. Rounsley (✉)
Genus plc, De Forest, WI, USA
e-mail: steve.rounsley@gmail.com

Keywords Assembly, Crops, NGS, Sequencing

Contents

Abbreviations

ABYSS	Assembly by Short Sequences
AGI	Arabidopsis Genome Initiative
API	Application Programming Interface
BAC	Bacterial Artificial Chromosome
CCD	Charge Coupled Device
CIGAR	Concise Idiosyncratic Gapped Alignment Report
CNV	Copy Number Variation
CRT	Cyclic Reversible Termination
DBG	de Bruijn Graph
ddNTPs	Dideoxynucleotides
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotides
GB	Giga-basepairs
GMOD	Generic Model Organism Database
GWAS	Genome Wide Association Mapping
HapMap	Haplotype Map
IGV	Integrative Genomics Viewer
InDel	Insertion-Deletion
Kb	Kilo-basepairs
LD	Linkage Disequilibrium
MAGIC	Multiparent Advanced Generation InterCross
Mb	Mega-basepairs

MTP	Minimum Tiling Path
NAM	Nested Association Mapping
NGS	Next-Generation Sequencing
OLC	Overlap Layout Consensus
ONT	Oxford Nanopore
PacBio	Pacific Biosciences
PAV	Presence-Absence Variation
PCAP	Parallel Contig Assembly Program
PCR	Polymerase Chain Reaction
PHRAP	Phil's Revised Assembly Program
PHRED	Phil's Read Editor
SBL	Sequencing by Ligation
SBS	Sequencing by Synthesis
SMRT	Single Molecule Real Time
SNA	Single Nucleotide Addition
SNP	Single Nucleotide Polymorphism
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
Tb	Tera-basepairs
TIGR	The Institute for Genomic Research
UCSC	University of California at Santa Cruz
VCF	Variant Call Format
VEP	Variant Effect Predictor
WGS	Whole Genome Shotgun
ZMV	Zero Mode Waveguide

1 Introduction

When History of Science books are written in the future, there seems to be a more-than-reasonable chance that DNA sequencing and the birth of genomics will feature prominently. It is hard to think of a technology that has had a more dramatic effect on the study of biology than DNA sequencing. For those active in research today, with all the data and technology available, it is also hard to remember how little we knew about genomes before the mid 1990s. And despite the huge gulf in technology and knowledge between then and now, the field may still be in its infancy – in the first stages of a journey with a double helix as its guide. This chapter describes a few aspects of the journey so far and looks forward to what may lie ahead.

2 Current Technologies, Standards, and Strategies

2.1 Sequencing Technologies

2.1.1 Sanger Sequencing

In 1977, Frederick Sanger published a DNA sequencing technique that became the base technology for the field of genomics [1]. Sanger sequencing relies on the chain terminating properties of dideoxynucleotide triphosphates (ddNTPs), which were added to a mix of the four standard deoxynucleotides (dNTPs). When a complementary strand of sequence is synthesized using these reagents (the sequencing reaction), the result is a mixture of DNA fragments each terminated at different lengths. These fragments must then be separated by size (via electrophoresis), detected, and then recorded. Initially, slab polyacrylamide gels, radioactivity, and typing in sequence were integral to the standard (very manual) technique. Automated DNA sequencers were later developed, which automated the detection and capture of the resulting DNA sequence. Improvements such as fluorescently-labeled terminating nucleotides and capillary electrophoresis were incorporated into the ABI line of DNA sequencers. Hundreds of these instruments were sold to large genome centers working on genome projects in the 1990s and early 2000s – including bacteria, yeast, Arabidopsis, mouse, and human genomes [2].

2.1.2 Next-Generation Sequencing (NGS) Technologies

Over the last decade, sequencing technologies have evolved rapidly and led to a significant increase in throughput and reduction in cost, thereby enabling large-scale sequencing of genomes. They have done so by removing a limitation of Sanger sequencing of having to separate DNA fragments by size. In Sanger sequencing, the sequencing reaction occurs outside of the instrument, and the instrument simply separates and detects fragments. For most NGS technologies, the sequencing reaction is occurring on the instrument, and each base addition onto a growing DNA molecule is detected and recorded. The first generation of NGS technologies have relied largely on two approaches for sequencing, sequencing by ligation (SBL) and sequencing by synthesis (SBS) [3]. Both approaches rely on spatially constrained, clonal amplification of DNA and facilitate massive parallelization of sequencing reactions, each with its own clonal DNA template, resulting in the sequencing of millions of sequences in parallel.

SBL involves hybridization and ligation of fluorophore-labelled probes and anchor sequences to a DNA strand and capturing the emission spectrum to identify the DNA base, whereas SBS relies on strand extension using a DNA polymerase and uses changes in color or changes in ionic concentration to identify the incorporated nucleotide [3]. SBL is used in platforms such as SOLiD and Complete Genomics, whereas 454, Ion Torrent and Illumina use the SBS approach.

The SBS technologies can be classified into two approaches: the first, single nucleotide addition (SNA), used in 454 and Ion Torrent sequencers. This approach adds four nucleotides iteratively and scans for a signal after each to record an incorporated nucleotide. In the case of 454, which sold the first NGS instrument (the GS20), template-bound beads are distributed into a PicoTiterPlate and emulsion PCR is performed to clonally amplify a single DNA fragment within a water-in-oil microreactor. The addition of dNTPs triggers an enzymatic reaction that results in a fluorescent signal that is captured by a charge-coupled device (CCD) camera and is indicative of incorporated nucleotide [4]. The SNA method as implemented in Ion Torrent relies on ion sensing rather than fluorescence and detects the H^+ ions that are released after the incorporation of each dNTP and the resulting shift in pH is used to determine the incorporated nucleotide. Both 454 and Ion Torrent methods have limitations in accurately measuring the homopolymer lengths, because all nucleotides in a homopolymer are incorporated at the same time, and the magnitude of the signal must be used to estimate the homopolymer's length.

The other SBS approach is found in the NGS instruments that have come to dominate the market – those manufactured by Illumina. This technology, which was developed by Solexa before they were acquired by Illumina, uses terminating nucleotides similar to Sanger, except the termination is reversible. Cyclic reversible termination (CRT) uses a mixture of four reversible terminators each with a distinct fluorescence. Each template is extended by a single base only using the appropriate terminator and the resulting labeled templates are imaged recording which nucleotide was added to each template. The terminators are then cleaved off, and the cycle continues with the addition and imaging of the next nucleotide. An additional key to Illumina's success is the massive number of templates the technology can sequence in parallel – approaching three billion on a single flow cell in the HiSeq-X instrument. They achieve this through the immobilization of a DNA library onto a glass flow cell coated with adapter oligos. Clonal clusters of each DNA fragment are synthesized using bridge amplification on the flow cell resulting in a very large number of sequence-ready templates. Illumina currently has the largest market share for sequencing instruments and offers a wide variety of sequencing systems, read lengths, and throughput to cater to a wider range of applications (Table 1).

2.2 Assembly Technologies

The developments in automated, higher throughput sequencing technologies have been matched by concomitant development of algorithms and tools to use the resulting data in various applications. For projects where the goal is the generation of a reference genome, assembly algorithms have been a key area of development. The selection of an appropriate algorithm depends on the sequencing strategy being used (see next section), but here we will describe the main classes available.

Assembly algorithms can be broadly divided into two classes: overlap-layout-consensus (OLC) and De-Bruijn-graph (DBG) [5]. The OLC approach identifies

Table 1 Illumina sequencing systems

Metrics	MiniSeq	MiSeq v3	NextSeq	HiSeq2500 v4	HiSeq3000/4000	HiSeq X
Maximum output	7.5 Gb	15 Gb	120 Gb	1,000 Gb	1,500 Gb	1,800 Gb
Cluster number (millions)	25	25	400	4,000	5,000	6,000
Read length	1 × 75 bp	2 × 75 bp	1 × 75 bp	1 × 36 bp	1 × 50 bp	2 × 150 bp
	2 × 75 bp	2 × 300 bp	2 × 75 bp	2 × 50 bp	2 × 75 bp	
	2 × 150 bp		2 × 150 bp	2 × 100 bp	2 × 150 bp	
				2 × 125 bp		
Run time	7–24 h	21–56 h	11–29 h	1–6 days	<1–3.5 days	<3 days

bp basepairs, *Gb* gigabase pairs, *PE* paired-end sequencing, *SE* single-end sequencing

overlaps between all reads, and the reads and overlap information are laid out on a graph and consensus sequences are then inferred. This algorithm, often used with Sanger-generated data, has been widely incorporated into assembly programs such as Arachne [6], Celera Assembler [7], PCAP [8], and PHRAP [9]. Although this approach provides a cheaper and faster way of utilizing Sanger sequencing for reference genome development, with larger datasets the assemblies usually have gaps and result in unplaced scaffolds that require more effort to verify and finish. This heralded the era of draft genome assemblies and a subsequent change in standards for the quality of a reference genome.

The significantly higher data volume, shorter read lengths, and platform-specific error profiles of NGS data present challenges for algorithm developers. The higher amounts of short-read data from the next generation sequencers furthered new developments in assembly algorithms and a few overlap-layout-consensus assemblers such as Celera Assembler [7], PCAP [8], and Newbler [4] were extended from their original versions to handle both Sanger and NGS data from 454 sequencers. However, the increased usage of short read Illumina sequences for assembling large complex genomes spurred the development of the second class of assembly algorithms – those using the more efficient DBG-based approaches. The DBG approach works by first chopping reads into shorter k-mers, using those k-mers to build a graph and using the graph to infer the genome sequence. Assemblers such as ABySS [10], ALLPATHS-LG [11], and SOAPdenovo [12, 13] rely on the DBG approach for increased efficiency.

2.3 Reference Genome Project Strategies

2.3.1 Sanger-only Assemblies

Sequencing technologies have enabled the study of genomes across all spheres of life. The first genomes to be sequenced were bacterial [14, 15] and employed a whole genome shotgun approach. However, at the time, larger genomes were not considered good candidates for this approach. Consequently, a hierarchical shotgun strategy was developed for the first large genomes, including the generation of the first plant reference genome for the model plant *Arabidopsis thaliana*. The Arabidopsis Genome Initiative (AGI), an international consortium, generated comprehensive BAC libraries and used the BAC end-sequences and fingerprints of individual BAC clones to create a physical map. A minimum tiling path of BAC clones along each chromosome was identified and the selected BACs were then individually shotgun-sequenced by consortium members and assembled using assemblers such as the TIGR Assembler [16] to produce assembled contigs. The BAC ends were later used to link contigs into scaffolds and the genetic map served as a foundation for integrating assembled scaffolds into chromosomes [17].

The initial strategies for reference genomes relied predominantly on Sanger sequencing and continued to make advancements through automation or

incorporating improved methodologies. For instance, the rice genome sequences were assembled using PHRED and PHRAP software packages or the TIGR Assembler with the finishing step incorporating some automated and manual improvements and sequence gaps resolved by full sequencing of gap-bridge clones, PCR fragments, or direct sequencing of BACs [18]. The maize genome also relied on the hierarchical approach and Sanger sequencing while utilizing optical mapping to order and orient contigs into chromosomes [19].

The generation of the soybean reference genome [20] used the whole genome shotgun strategy – first used in the early bacterial genomes in 1995, and later adapted for the Celera human genome and many other mammalian genomes. The basic WGS strategy involves randomly shearing the genome and sequencing the fragments from this WGS library. The modified approach for larger genomes generates sequence libraries from multiple-sized fragments. For soybean, an initial WGS library of ~1,000 bp inserts was combined with 3, 8 kb, Fosmid and BAC libraries. The soybean sequence data were assembled using Arachne [6], where an initial assembly generated from the WGS library was combined with paired end data from multiple libraries for scaffolding the contigs [20]. Subsequently, many other plant genomes have been sequenced with this approach [21–25].

2.3.2 NGS Technologies for Reference Genome Generation

With the advent of cheaper and high-throughput NGS technologies, Sanger sequencing was soon relegated to the back seat for sequencing needs. 454 and Illumina platforms that could generate several megabases of sequence data in a short time, opened up genome projects to researchers outside of the large genome centers. Although the newer technologies produced shorter read lengths (32–500 bp), and thus presented assembly challenges, the higher throughputs, lower costs, and faster data turnaround made them hard to resist, and soon there was a surge in reference genomes from plant species, albeit with lower quality than Sanger genomes. NGS has been applied to more genomes as the cost of NGS dropped quickly (Fig. 1). About 73% of first 50 plant genomes published are on crop species and most of them include NGS as part of sequencing [26].

2.3.3 Hybrid Sanger-NGS Assemblies

Although many genome projects started to rely on NGS for generating assembled reference genomes, the contiguity from NGS-only assemblies was far shorter than those from Sanger sequencing. Thus, strategies to sequence large complex crop genomes began to rely on a combination of Illumina, Roche 454 and Sanger platforms to balance the cost and contiguity of assemblies. For example, the genome of oil seed rape, *Brassica napus*, was sequenced using a combination of multiple platforms: 21.2× coverage from GS FLX Titanium sequencing (reads of 450 bp average size), 0.1× Sanger BAC ends (reads of 650 bp average size), and 53.9× Illumina HiSeq sequencing (reads of 100 bp) [27]. The 454 sequencing included

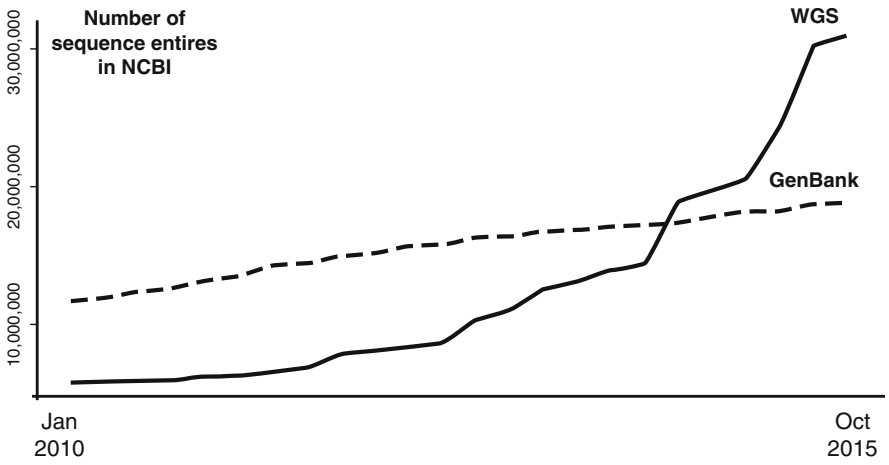


Fig. 1 Decreasing sequencing cost and increasing read length driven by introductions of new technologies. Data sources: Illumina (www.illumina.com), PacBio (www.pacb.com), Oxford Nanopore (www.nanoporetech.com)

regular 8 and 20 kb libraries and Sanger-based BAC ends were from a BAC library of 139 kb average insert size. The longer reads were assembled using Newbler to generate an initial assembly and Illumina reads were used for final error correction and gap filling with the construction of final pseudomolecules facilitated with genetic maps. A similar strategy of combining benefits from multiple technologies was used to generate the reference genomes of tomato, cassava, and African rice [28–30]. As more NGS sequences were used, a drop in assembly quality is generally seen compared to the genomes sequenced using Sanger method.

2.3.4 NGS-only Assemblies

With continuous improvements in the Illumina platform and assembly algorithms, NGS-only genomes have increased in number. The Illumina platform was used to generate a chromosome-based draft sequence of the hexaploid bread wheat [31]. High depth of Illumina sequences was also added to the *B. rapa* genome [32], the diploid [33], and allopolyploid cultivated [34, 35] cotton. Due to the repetitive nature of crop genomes, the contiguity is much lower than that from Sanger sequencing. Although the hierarchical approach algorithmically has advantages over WGS approaches, the overall process of generating a BAC library, physical map, and MTP are very labor and time intensive, making these projects very expensive and time consuming.

2.4 Resequencing Strategies

The availability of high-quality reference genome sequences combined with higher throughput and lower cost of sequencing is making it possible to comprehensively understand diversity within a species by generating sequence from many accessions. Whole genome resequencing is being effectively utilized to understand crop diversity and create genomic resources to enable crop improvement across a wide range of crops. This approach generates low coverage (usually 2× to 10×) genome sequence data from accessions of interest and compares the sequences against a reference genome to detect various kinds of variation – single nucleotide polymorphisms (SNPs), insertion-deletions (InDels), presence-absence variants (PAVs), copy number variations (CNVs), and other structural variants – to understand the genetic diversity of a crop species. In plants, the 1,001 genome project in Arabidopsis [36] demonstrated the value of resequencing to enhance understanding of a species and soon several large-scale resequencing projects were initiated in crop plants like rice [37], maize [38, 39], soybean [40], and sorghum [41]. These resequencing data were able to provide unprecedented information about the variation existing within each crop species that can be utilized for improvement of these crops. Such resequencing data are now routinely used to find novel alleles for genes of interest [42–45], find the signals of domestication, provide background data to build genomic selection models, and form the basis for generation of tailored populations such as multi-parent advanced generation inter-cross (MAGIC) and nested association mapping (NAM) populations. Many of these applications are discussed in detail later in this volume.

Sequencing several accessions from a crop has demonstrated the presence of extensive structural variations within crop species [37, 38] leading to the recognition of the importance of generating multiple de novo assembled genomes (e.g., soybean, rice) [34, 35, 46]. Although high-throughput NGS technologies have shown advantages in generating variants and draft assemblies at low cost, the incompleteness of these assemblies and their reliance on a single existing reference genome makes it challenging to comprehensively identify structural variations.

In 2015, sequence entries archived in NCBI showed an interesting pattern: the number of entries for WGS surpassed the general sequence entries submitted to GenBank (Fig. 1). The dramatic increase of WGS data has been a result of re-sequencing driven by ever-decreasing sequencing cost (Fig. 2). Biologists have been using the resequencing approach for across a wide range of species and for varied research goals. For all, the ability to sequence across multiple individuals is a powerful approach, albeit with logistical challenges.

2.5 Data Management and Visualization

When the first plant genome became available, efforts in data management and visualization were primarily focused on making the sequence data and the corresponding annotations available to a broader scientific community and enabling

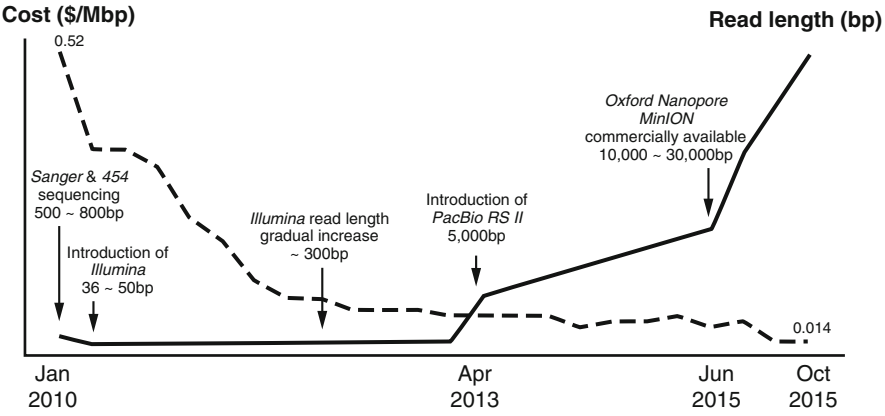


Fig. 2 Submitted sequence entries for GenBank and WGS archived in NCBI. Data source: NCBI (http://www.ncbi.nlm.nih.gov/genbank/statistics/)

the use of genome sequences to address specific research questions. With the large influx of genome sequencing data from NGS technologies, tools for data analysis, storage, and management soon became a critical need of the scientific research community. Initial developments centered on developing data standards and guidelines so that data could be easily shared and accessed.

2.5.1 Variant Data Standards

The human 1,000 Genomes Project led the way and provided invaluable insights into genetic variants in humans, as well as established some of the early standards to manage and analyze large-scale variant data that soon became the standard for later large-scale studies in all organisms [47–49]. New file formats to compress and store sequence alignment data and tools that could manipulate these file formats quickly came into existence and widely spread within the bioinformatics community [50]. The 1,000 Genomes Project created the Variant Call Format – a format that has become the standard for managing and manipulating variant data obtained by comparing re-sequencing data to reference genomes [51]. The initial development of VCFtools and the more recent vcfR and PyVCF tools enabled scientists using three major programming languages used in the bioinformatics community to embrace VCF as the system to manage and analyze variants [51, 52]. These tools in combination with SnpEff, a tool for annotating functional impacts of SNPs, provide a toolkit to utilize variant data in the pursuit of answers to deeper scientific questions [53].

2.5.2 Variant Data Management Systems

While the VCF file is fairly simple, it can contain essentially complete information about individual variants. However, it is not a user-friendly format for querying as VCF files can easily contain millions of variants, and be 10s or 100s of Gigabytes in size. Solutions employing relational database or indexing schemes are needed to extract information from VCF quickly and efficiently with complex query structures.

One solution to the variant storage and query problem is to utilize relational database systems such as MySQL. One example of this is the maize HapMap project [38]: all variants generated in this project are imported into "Ensembl Variations," in which each variant, SNP or InDel, is stored as an entry in the relational database and contains several attributes linking it to other relevant information (Fig. 3). A user can explore the frequency and genome context, as well as linkage information of variants using the data schema of Ensembl.

The Ensembl MySQL solution is intended for the most widely used genomes that have Gold Standard quality assemblies, extensive annotations, and functional studies. It may not work with the majority of re-sequencing projects, as these projects are often focused on less well studied species whose data do not meet the high quality standards of Ensembl. For such projects, VCF is still the best choice for data retention and downstream analyses, but there are some alternate solutions that do not rely on relational databases. For example, genome browsers such as JBrowse render compressed and indexed VCF to visualize information [54]. The "focused" nature of a genome browser takes advantage of the indices to show only variants in selected genomic intervals.

In many cases, the primary piece of information needed is the impact of the variant, i.e., the functional annotation of the variant: is it in a coding region or non-coding region, is it a synonymous or non-synonymous change, etc. For this purpose, there are a number of solutions [53, 55, 56]. For example, SnpEff is a suite of tools for genetic variant annotation and effect prediction. A primary advantage of SnpEff is the 38,000 genomes supported out-of-the-box, so users can leverage prior annotation efforts of the community. SnpEff also supports VCF files generated by major variant calling pipelines such as SAMtools and GATK and appends the annotation results to the VCF files. The VCF-in and VCF-out workflow for SnpEff enables users to apply existing tools for manipulating VCF files and allows SnpEff to be tightly integrated into analysis pipelines without too much additional effort. Another SNP annotation tool with comparable gene annotation databases is Ensembl's Variant Effect Predictor (VEP). Unlike SnpEff, VEP does not generate VCF files but a unique plain text, closely tied to the unique relational database of Ensembl. By taking advantage of the rich infrastructure of Ensembl's web front end, VEP provides a more user-friendly point-and-click web interface for variant annotation.

Fig. 3 Ensembl variations: explore one variant at a time

2.5.3 Visualization of Variant Data

Many layers of information are stored in the linear string of four nucleotides that make up a genome – from single nucleotides, codons, exons and genes to regulatory units, chromatin structure, and chromosome conformation. Visualization of re-sequencing results at many different levels is a crucial component of such projects. Generally, there are two approaches to visualizing data (primarily reads and/or VCF files) from re-sequencing projects: one is the dedicated application on a desktop or laptop computer; the other is by utilizing an Application Programming Interface (API) for existing web-based genome browsers to work with short-read mapping and variant calling results. Given the amount of data from re-sequencing projects, the key to achieving performance is to create the ability to access only the reads or variants needed for the specific slice of genome that is being viewed.

The champion of read-centric visualization tools is Integrative Genomics Viewer (IGV) [57]. In addition to providing a large number of ready-to-use genomes and annotations, IGV has the best support for visualizing almost every detail of read-mapping information, including the very important but largely overlooked CIGAR string [50]. It also provides a read coloring system that helps users spot split reads and read pairs with abnormal insert sizes between the mates – crucial for the exploration of structural variations. IGV has also gone beyond a standalone desktop application and supports the access of data files from distributed sources via the HTTP protocol. Tablet is another desktop solution for read visualization that stands out from the crowd with its great usability and interface. Tablet works extremely well in terms of zooming in and out, as well as views at different levels in one screen (Fig. 4).

With the need to visualize the large amount of re-sequencing data, the traditional feature-based genome browsers are playing a catch-up game. Genome browsers, such as UCSC browser and GBrowse, have been the data hub and integrator for feature-based data, i.e., genomic data based on genomic intervals for many years [58, 59]. The feature-based data are rich, detailed, but small in size, so the traditional genome browsers have been optimized to primarily handle large numbers of tracks of small sizes. NGS data from re-sequencing projects present the opposite challenge – read alignments files are very big, but information for each read is minimal. Because UCSC and GBrowse both use relational databases in the backend, they had to create database adaptors to handle read alignments, which turned out to be inefficient and awkward, especially when the alignment files are large in size. Subsequently, many new genome browsers have been developed with optimized functionalities for visualizing short reads. The best examples among these are JBrowse, a generic genome browser from GMOD, and Savant Genome Browser, a short-read browser optimized for human genome and medical and diagnostic purposes [54, 60]. Both genome browsers abandoned the old relational database architecture and embraced read alignment formats directly, so they read the alignments and render the reads on-the-fly. Coupled with various indexing schemes, they

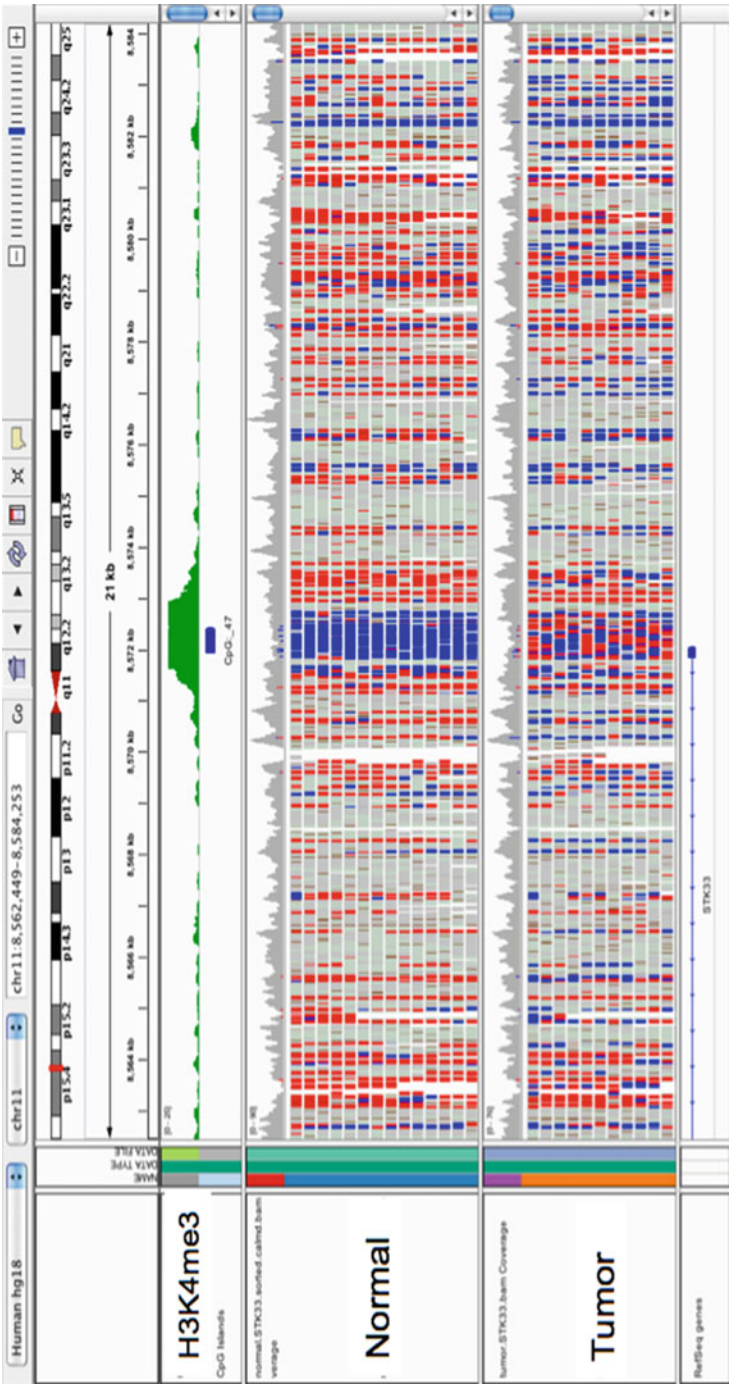


Fig. 4 IGV and Tablet visualization of reads and read alignments. Data sources: Broad Institute (software.broadinstitue.org), Tablet (ics.hutton.ac.uk/tablet/)

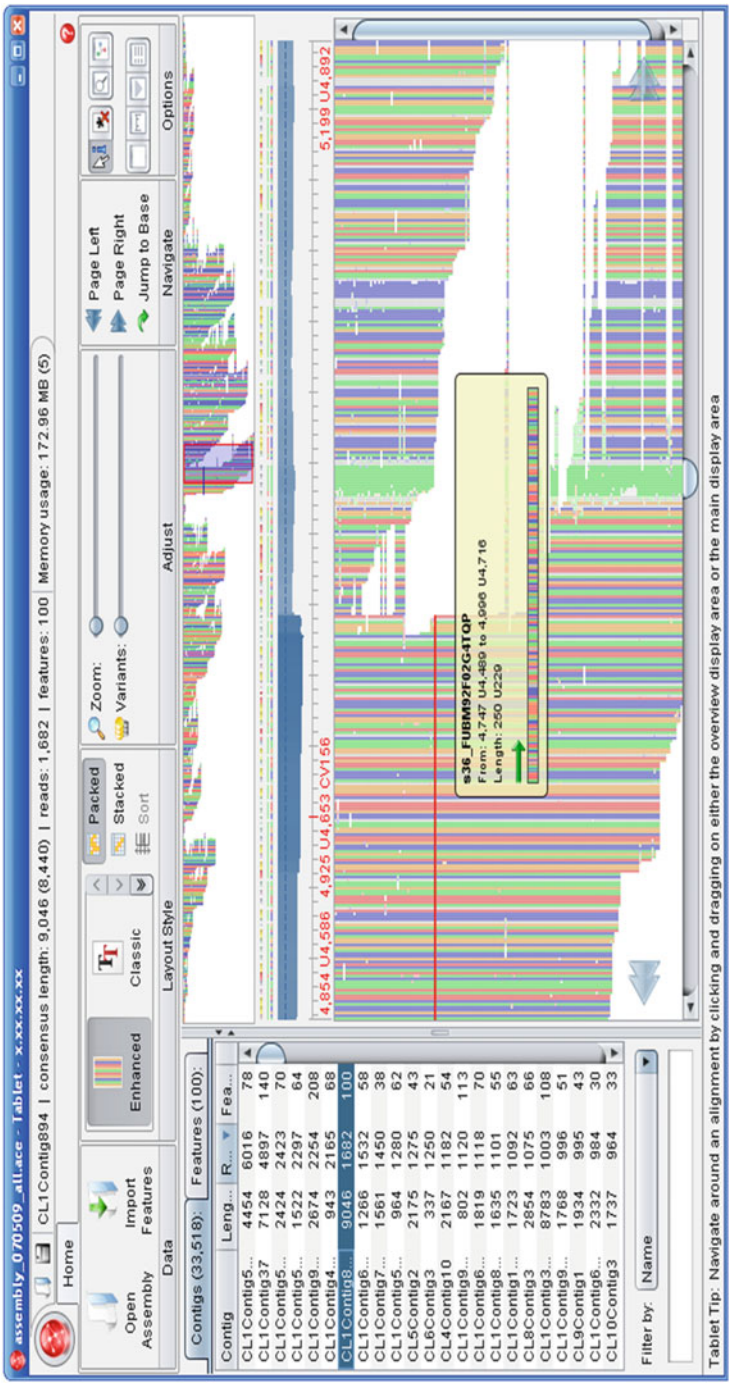


Fig. 4 (continued)

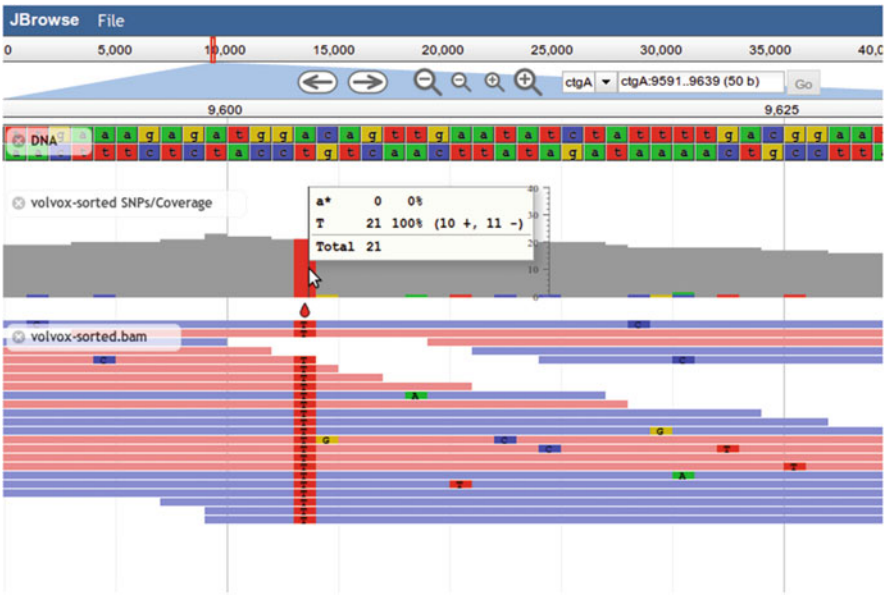


Fig. 5 Rendered reads and SNP in a JBrowse view, adapted from Ref. [54]

provide intuitive navigating functions to explore read mapping and variant calling results (Fig. 5).

The advantage of generic genome browsers over a specialized short reads viewer is that it is very easy to incorporate feature-based genomic data and much more into a holistic genomic view. Instead of adding short read functionalities as an afterthought, this new generation of genome browsers puts short reads in the center and builds genomic resources around them. Moreover, a lot of the new genome browsers have started to utilize cloud platforms to store and manage sequencing data, majority of them re-sequencing data (https://cloud.google.com/genomics/, https://aws.amazon.com/). This should not be surprising since the "big" nature of re-sequencing data fits nicely to the concept of "Big Data" advocated by cloud technologies.

3 Trends, Advanced Technologies, and Strategies

3.1 Sequencing Technologies

The second-generation sequencing technologies such as Illumina are very useful for resequencing studies to understand the variability of a crop species. However, due to their short reads, it is challenging to generate finished quality reference genomes. The recent emergence of long-read sequencing technologies such as PacBio (http://www.pacb.com/) and Oxford Nanopore (https://www.nanoporetech.com/), and technologies that focus on providing long-range genomic linking information such

as Dovetail Genomics (http://dovetailgenomics.com/), 10× Genomics (http://www.10xgenomics.com/) and BioNano Genomics (http://www.bionano-genomics.com/) are making it feasible to generate good quality reference genomes faster and cheaper.

PacBio single-molecule real-time (SMRT) sequencing captures the sequence information during the replication process of a DNA molecule that is tracked in a zero-mode waveguide (ZMV) on a SMRT cell. The DNA molecule is circularized by adding the adapters on both ends and diffused into a ZMV with DNA polymerase immobilized at the bottom. Four fluorescent bases are flowed through the SMRT cell and a distinct light pulse is produced for each base that is recorded as a movie. The movie can then be analyzed to extract DNA sequence. PacBio can produce reads in the average range of 20 kb and is being routinely used to finish microbial genomes [61]. Until recently, it has been very expensive to use PacBio data alone for a large crop genome, and thus many hybrid strategies have been deployed that combine PacBio sequences with other short read data to improve genome assemblies [62], and new algorithmic strategies are being developed to better utilize these long reads in assembly processes both in hybrid strategies and alone [63–66]. The new SEQUEL system from PacBio can deliver up to 50 Gb sequences for a few thousand dollars at an average read length of ~20 kb and consensus accuracy >99.999%, making it an attractive option for crop reference genomes. For example, the genome of adzuki bean (*Vigna angularis*) was assembled using SMRT sequencing technology and the PacBio assembly produced 100 times longer contigs with 100 times fewer gaps compared to the NGS-based assemblies [67]. Efforts are currently underway to improve the B73 reference genome of maize and build high-quality reference genomes for 23 species of rice using PacBio SMRT Sequencing and create new resources for crop improvement (http://www.pacb.com/wp-content/uploads/agi-rod-wing-corelab.pdf).

Oxford Nanopore (ONT) sequencing is the latest long-read sequencing technology that offers a lot of promise for generating de novo assemblies of complex plant genomes. This technology passes a long DNA molecule through a charged protein nanopore and measures the changes in current as the molecule passes through the nanopore. The changes in current or "squiggleplot" are then input into a basecaller to produce DNA sequence information. ONT is very promising technology with reads as long as 150 kb having been reported by early users, although average read lengths are much lower. The technology is deployed in two forms – a small mobile sequencer, the minION, which is approximately the size of a stapler that has flowcells with 512 nanopores, and a much larger format called the promethION, which can house 48 flowcells, each with 3,000 nanopores. MinION has been commercially available since May 2015 and has been applied to the rapid identification of viral pathogens [42, 68], 16S sequencing [69], and haplotype sequencing [70]. At the time of writing, nearly 50 publications have used or developed tools for the ONT platform. As the accuracy and throughput continue to improve, de novo sequencing of large complex crop genomes will become practical soon.

The parallel development of several long-range sequencing technologies from Dovetail Genomics and 10× Genomics, or long-range mapping technologies from BioNano Genomics, can provide the contiguity information in a genome. The long-range information when complemented with sequences from long-read single

molecule technologies can deliver high quality assemblies with fewer gaps and megabase-long contigs for complex plant genomes without the need to construct traditional physical and genetic maps. In view of the significant structural variation in crop species, these new technologies can redefine our understanding of genomes and help pinpoint the underlying genetics of complex traits.

3.2 Assembly Strategies/Technologies

Developers of assembly algorithms are focusing on developing methods that will enable the seamless integration of long-read and long-range data into the assembly process. The long-read technologies in their initial growth cycles have higher error rates, and algorithms must take these into account. Various software tools have been developed to handle multiple scenarios involving longer reads. PBJelly2 is effective on low coverage ($<15\times$) PacBio data and has the ability to use the long-read data to link scaffolds and close gaps for existing short-read assemblies [65]. Tools such as ECTools [66], SPAdes [63], and PBcR [64] can handle 20–30$\times$ coverage PacBio data either in combination with Illumina reads or on their own. If more than 50$\times$ coverage is available, the PacBio sequences can be assembled de novo without short-read sequences using packages such as HGAP [71] and Canu [64]. In some of the most recent algorithms, with $>30\times$ coverage of PacBio sequences, an overlap-layout-consensus approach can be used to assemble corrected sequences. A final polishing step is used to correct the errors in the consensus with raw PacBio sequences, which can improve the consensus accuracy to 99.999%. Similar assembly and error correction strategies can be applied to ONT data. As each platform continues to improve accuracy towards a 1% error rate, error correction will not be necessary.

3.3 Genome Project Strategies

The development of new third-generation sequencing technologies is leading to a trend of combining these data with second generation technologies in genome sequencing projects. Due to a relatively lower throughput and higher cost (per Gb) of the long-read technologies, current genome sequencing strategies typically combine lower coverage long-read/long-range data to with higher coverage of short read data to improve the qualities of genome assemblies especially for large, complex crop genomes. Ultimately, the selection of a strategy is driven by what can provide the highest quality for a given cost combined with the perception of what is an acceptable cost for a genome project. As the technologies continue to develop further, error rates and cost are expected to drop, which will change both what is possible and the perception of what is reasonable.

3.4 Resequencing Strategies

Short-read technologies have been heavily used by projects to generate understanding of the variation within a crop species. However, since they rely on a reference genome, these projects have had limitations in identifying large structural variations among accessions within a species. Crop species like maize and soybean have been shown to have large variation in their genome content between lines – almost to the extent of 30% [72]. Resequencing that relies on a single reference genome is not able to adequately capture the full extent of these variations. As long-read technologies continue to improve and drop in price, we expect projects to generate de novo reference assemblies for multiple lines within a species – perhaps for all lines within a species if the price drops far enough. These types of data will enable us to better characterize and catalog the variation within a species.

3.5 Data Management, Visualization, and Storage

The availability of multiple reference genomes for crops, the widespread re-sequencing efforts, and the resulting variant data are constantly pushing the limits of VCF files, as well as the tools and infrastructure for handling them. For example, one VCF file containing millions of SNP/InDels from hundreds of thousands of samples could not be effectively managed by any of the tools previously described. Further evolution of variant storage is needed. One recent advance is BGT, a flexible genotype query tool that works with large scale multi-sample VCF files [73]. The key to these tools is to generate indices of variant genotypes that can be harnessed for rapid retrieval in a flexible manner – whether it be for a subset of genomic locations, or a subset of samples or any combination of both. Moreover, BGT supports encoded phenotypes associated with the samples, which creates opportunities to slice and group the samples based on phenotypes, a convenient way to conduct local and small-scale association studies.

Meshing the concepts contained within VCFtools, PyVCF, vcfR, and BGT, there are unlimited possibilities to create new tools to extract the information in VCF and utilize the information in plant genetics and crop breeding. To date, information from VCF files has been queried, summarized, and manipulated for GWAS, LD analyses, small-scale association studies, as well as variant data dissemination through various data visualization frameworks. We expect this trend to continue with ever more sophisticated data manipulation tools and approaches.

3.6 Beyond Individual Variants: Alleles, Haplotypes, LD Blocks, and Pan-Genomes

Interpretation of the biological meaning of information contained within sequence data is not the exclusive domain of bioinformaticians. It takes collaborations between biologists of all kinds, which is particularly true as the complexity of the

data and data structure increases. Information stored in genomes comes at several levels, and the bulk of what we have discussed in this chapter is focused on the discovery and description of individual variants. But this just scratches the surface of the full impact of genomic diversity. There are relationships between variants that can be identified, stored, and visualized.

As an example, consider a gene as a unit within which combinations of multiple variants create the variant of that gene that may have a phenotypic impact. Each combination of variants that form a single version of that gene can be considered as a group to define a single allele of that gene. This requires a more complex form of annotation and visual representation than is found in current genome browsers. Genetic linkage beyond gene boundaries forms yet another higher level of information, indicating longer segments that are segregating and propagating in the real world, called haplotypes. Stretches of such haplotype segments form Linkage Disequilibrium (LD) blocks, in which most native traits harbor. Above the LD blocks and haplotypes, there are chromatin structure and chromosome conformations. Characterization of these even higher levels lies beyond the scope of simple re-sequencing and requires other technologies such as methylation profiling by bisulfate sequencing and Hi-C profiles [74, 75].

Strategies to explore and exploit the alleles, haplotypes and LD blocks present within crop plants are active areas of development, both in academia and commercial breeding contexts. The common goal is understanding the genetic structure of populations at a more sophisticated level than individual variants, which can then enable a mix-and-match approach to the traits required in breeding and efficient and accurate trait characterizations at the molecular level.

One other recent trend is a gradual shift away from the concept of a single reference used as a basis for all future studies of variation within that species. This is happening for a number of reasons: the reference accession is often the most "well-behaved" accession for research, rather than the best representative of the species; our perceptions of what is possible have changed along with sequencing costs; data from early re-sequencing studies showed that it is unlikely that any single reference could represent a species, even with a robust way of describing variants. The concept of the composite genome of a species, rather than an accession, is known as the pan-genome [12, 13, 76]. Depending on the evolutionary history and divergence among the individuals, pan-genomes can be very simple in closely related individuals, or very complex in groups with tremendous genetic diversity. In the latter case, a pan-genome captures the true nature of genetic variations in a way that a single reference could not, because the variation is far beyond single nucleotide changes. Maize is an example of such a species with extremely rich diversity among varieties and accessions. Variants discovered by comparing re-sequencing data to the B73 reference missed significant fractions of the true variation [38].

Unfortunately, although a simple concept, representing pan-genomes in a file format is not a simple task. The definition and implementation of pan-genomes is still in its infancy. One promising idea is to present the genomes in a graph with genomes represented by a "path" in a hypothetical space and variations represented by "bubbles" that bulge on the sides of the paths [77] (https://www.technologyreview.

com/s/537916/rebooting-the-human-genome/). The graph theory supporting the pan-genomes, by its nature, is capable of capturing and presenting information at multiple levels from single nucleotides to large linkage blocks. This will lend itself naturally to representing haplotype LD blocks useful for exploitation of variation in breeding programs. As more and more data are added, the pan-genome concept implemented with effective visualizations and query tools are going to be essential in order to gain biological insights from these incredibly valuable datasets.

4 Conclusion and Outlook

As described in this chapter, technology advances over the last decade have been tremendous and have provided great benefits across the spectrum of biological research. The low cost and ease with which sequence data can be generated has led to larger and larger experiments being imagined by more and more individual scientists. No longer do genome-wide studies require international consortia. To coin an overused phrase – we are seeing the democratization of genomics. While the challenges for individual experiments may be shrinking, the challenges for community-wide management of these enormous stockpiles of sequence data are expanding. To truly enable the next wave of genomics-enabled research, the next advances will need to be not in sequence technology but in the management, access, and analysis of exabytes of data (1 exabyte = 1 million terabytes). Just as success in automated sequencing required biologists to recruit the skills of engineers, physicists, and chemists, success in this next phase will require us to embrace those skilled in computer and data science, and computational and network infrastructure. Only one thing is certain, there will be more data tomorrow than yesterday – which is fortunate, because in science, tomorrow always unveils more questions for us to answer.

References

1. Sanger F et al (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463–5467
2. Mardis ER (2013) Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto, Calif) 6:287–303
3. Goodwin S et al (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333–351
4. Margulies M et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380
5. Li Z et al (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics 11(1):25–37
6. Jaffe DB et al (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res 13(1):91–96
7. Myers EW et al (2000) A whole-genome assembly of *Drosophila*. Science 287 (5461):2196–2204
8. Huang XQ et al (2003) PCAP: a whole-genome assembly program. Genome Res 13 (9):2164–2170

9. Ewing B et al (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8(3):175–185
10. Simpson JT et al (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123
11. Gnerre S et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108(4):1513–1518
12. Li R et al (2010) Building the sequence map of the human pan-genome. Nat Biotechnol 28 (1):57–63
13. Li RQ et al (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20(2):265–272
14. Fleischmann RD et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269(5223):496–512
15. Fraser CM et al (1995) The minimal gene complement of Mycoplasma genitalium. Science 270 (5235):397–403
16. Sutton GG et al (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1(1):9–19
17. Hamilton JP, Buell CR (2012) Advances in plant genome sequencing. Plant J 70(1):177–190
18. Matsumoto T et al (2005) The map-based sequence of the rice genome. Nature 436 (7052):793–800
19. Schnable PS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956):1112–1115
20. Schmutz J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463 (7278):178–183
21. Goff SA et al (2002) A draft sequence of the rice genome (Oryza sativa L. ssp japonica). Science 296(5565):92–100
22. Ming R et al (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452(7190):991–U997
23. Paterson AH et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457(7229):551–556
24. Vogel JP et al (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463(7282):763–768
25. Yu J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp indica). Science 296 (5565):79–92
26. Michael TP, Jackson S (2013) The first 50 plant genomes. Plant Genome 6(2)
27. Chalhoub B et al (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345(6199):950–953
28. Prochnik S et al (2012) The cassava genome: current progress, future directions. Trop Plant Biol 5(1):88–94
29. Sato S et al (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485(7400):635–641
30. Wang M et al (2014) The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat Genet 46(9):982–988
31. International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345 (6194):1251788
32. Wang XW et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43(10):1035–U1157
33. Wang K et al (2012) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44 (10):1098–1103
34. Li FG et al (2014) Genome sequence of the cultivated cotton Gossypium arboreum. Nat Genet 46(6):567–572
35. Li YH et al (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32(10):1045–1052

36. Cao J et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43(10):956–963
37. Xu X et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30(1):105–111
38. Chia JM et al (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 44(7):803–807
39. Jiao Y et al (2012) Genome-wide genetic changes during modern breeding of maize. Nat Genet 44(7):812–815
40. Patil G et al (2016) Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. Sci Rep 6:19199
41. Mace ES et al (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun 4:2320
42. Bradley P et al (2015) Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun 6:10063
43. Brozynska M et al (2016) Genomics of crop wild relatives: expanding the gene pool for crop improvement. Plant Biotechnol J 14(4):1070–1085
44. Leung H et al (2015) Allele mining and enhanced genetic recombination for rice breeding. Rice (N Y) 8(1):34
45. Yang J et al (2015) Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. Plant J 84(3):587–596
46. Schatz MC et al (2014) Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. Genome Biol 15(11):506
47. Genomes Project Consortium et al (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073
48. Genomes Project Consortium et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65
49. Genomes Project Consortium et al (2015) A global reference for human genetic variation. Nature 526(7571):68–74
50. Li H et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25 (16):2078–2079
51. Danecek P et al (2011) The variant call format and VCFtools. Bioinformatics 27 (15):2156–2158
52. Knaus BJ, Grunwald NJ (2016) VCFR: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour 17(1):44–53
53. Cingolani P et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6(2):80–92
54. Skinner ME et al (2009) JBrowse: a next-generation genome browser. Genome Res 19 (9):1630–1638
55. McLaren W et al (2016) The ensembl variant effect predictor. Genome Biol 17(1):122
56. Wang K et al (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164
57. Robinson JT et al (2011) Integrative genomics viewer. Nat Biotechnol 29(1):24–26
58. Donlin MJ (2009) Using the generic genome browser (GBrowse). Curr Protoc Bioinformatics Chapter 9: Unit 9.9
59. Kent WJ et al (2002) The human genome browser at UCSC. Genome Res 12(6):996–1006
60. Fiume M et al (2010) Savant: genome browser for high-throughput sequencing data. Bioinformatics 26(16):1938–1944
61. Koren S, Phillippy AM (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 23:110–120
62. Ming R et al (2015) The pineapple genome and the evolution of CAM photosynthesis. Nat Genet 47(12):1435–1442

63. Bankevich A et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19(5):455–477
64. Berlin K et al (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing (vol 33, pg 623, 2015). Nat Biotechnol 33(10):1109–1109
65. English AC et al (2012) Mind the gap: upgrading genomes with Pacific biosciences RS long-read sequencing technology. PLoS One 7(11):e47768
66. Koren S et al (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30(7):692–700
67. Sakai H et al (2015) The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. Sci Rep 5:16780
68. Quick J et al (2016) Real-time, portable genome sequencing for Ebola surveillance. Nature 530 (7589):228–232
69. Benitez-Paez A et al (2016) Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION (TM) portable nanopore sequencer. Gigascience 5:4
70. Ammar R et al (2015) Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. F1000Res 4:17
71. Chin CS et al (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10(6):563
72. Gore MA et al (2009) A first-generation haplotype map of maize. Science 326 (5956):1115–1117
73. Li H (2016) BGT: efficient and flexible genotype query across many samples. Bioinformatics 32(4):590–592
74. Belton JM et al (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 58(3):268–276
75. van Berkum NL et al (2010) Hi-C: a method to study the three-dimensional architecture of genomes. J Vis Exp 39
76. Hirsch CN et al (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26 (1):121–135
77. Lu F et al (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun 6:6914