Turkish Journal of Agriculture and Forestry

http://journals.tubitak.gov.tr/agriculture/

Review Article

Turk J Agric For
(2015) 39: 361-376
© TÜBİTAK
doi:10.3906/tar-1409-93

# Sequencing of plant genomes – a review

**Mine TÜRKTAŞ[1], Kuaybe YÜCEBİLGİLİ KURTOĞLU[2], Gabriel DORADO[3], Baohong ZHANG[4], Pilar HERNANDEZ[5], Turgay ÜNVER[1,*]**

[1]Department of Biology, Faculty of Science, Çankırı Karatekin University, Çankırı, Turkey
[2]Department of Biology, Faculty of Arts and Sciences, Marmara University, İstanbul, Turkey
[3]Department of Biochemistry and Molecular Biology, Agrifood Campus of International Excellence, University of Córdoba, Córdoba, Spain
[4]Department of Biology, College of Arts and Sciences, East Carolina University, Greenville, NC, USA
[5]Institute for Sustainable Agriculture-National Research Council (IAS-CSIC), Córdoba, Spain

**Received:** 16.09.2014 • **Accepted:** 19.11.2014 • **Published Online:** 12.06.2015 • **Printed:** 30.06.2015

**Abstract:** The scientific revolution that started with the human-genome sequencing project, carried out with first-generation sequencing technology, has initiated other sequencing projects, including those for plant species. Different technologies have been developed together with the second- and third-generation sequencing platforms called "next-generation" sequencing. This review deals with the most relevant second-generation sequencing platforms, advanced analysis tools, and sequenced plant genomes. To date, a number of plant genomes have been sequenced, with many more projected for the near future. Using the new techniques and developed advanced bioinformatics tools, several studies including both plant genomics and transcriptomics were carried out. Likewise, completion of reference genome sequences and high-throughput resequencing projects presented opportunities to better understand the genomic nature of plants and accelerated the process of crop improvement. Modern sequencing and bioinformatics approaches have led to overcome the challenges that arose mainly in plant genomes with large size, high CG content, heterozygosity, transposable elements, repetitive DNA, and homopolymers or polyploidy, as may be the case with the most important crops. There is no doubt that the rest of the species will also benefit from such breakthroughs, which also include direct RNA sequencing without requiring cDNA synthesis. In fact, we are not in a postgenomic era as is sometimes stated, but rather in the beginning of a genomic revolution.

**Key words:** ChIP-Seq, deep sequencing, high-throughput sequencing technologies, RNA-Seq

## 1. Introduction

In the year 2000, researchers announced the first whole-genome sequence of a plant species. Sequencing of *Arabidopsis thaliana* was a cutting-edge achievement in the field of plant genomics. The impact of that study was so great that it boosted the demand for genomic information. However, using the conventional Sanger method (first-generation technology), sequencing a whole genome is time-consuming, laborious, and expensive work. In 2005, sequencing-by-synthesis technology developed by 454 Life Sciences revolutionized sequencing technology and started the second-generation sequencing era. Both required previous amplification in vivo (molecular cloning) or in vitro (e.g., polymerase chain reaction (PCR)). This was followed by the third-generation sequencing platforms, capable of sequencing single molecules without previous amplification. The sequencing generations following Sanger's approach are also known as next-generation sequencing (NGS), although this is rather ambiguous terminology for obvious reasons. The new sequencing strategies greatly reduced the necessary effort, time, and cost, also allowing for unprecedented throughput.

In the beginning, the read length of the 454 system was about 100 bases, which was increased up to 10-fold longer within a decade. In a short time, other new strategies were developed and appeared on the market. Within a few years, many genomes were sequenced, and several strategies have been developed to overcome certain problems like large genome size, high CG content, high heterozygosity, transposable elements, repetitive DNA, and homopolymers or polyploidy. One of the biggest challenges was that sequencing of large genomes required immense experimental work and elaborate analyses. However, scientists succeeding in sequencing large genomes, like that of Norway spruce (*Picea abies*), which is 20 Gbp in size (Nystedt et al., 2013 ). Thus, with the promises offered by the new sequencing technologies, a new trend for the life sciences was shaped. As a consequence, genomics

[*] Correspondence: turgayunver@karatekin.edu.tr

is experiencing its golden age. Indeed, we are not in a postgenomic era as sometimes indicated, but rather in the beginning of a genomic revolution.

In this review, we focus on 3 commercial sequencing systems: Roche/454 Life Sequencing, ABI/SOLiD, and Solexa/Illumina technologies. There are other methodologies that are outside of the scope of the present work, including the Life Technologies Ion Torrent, as well as new third-generation sequencing platforms (mostly in active current development), like the Helicos BioSciences true single-molecule, Pacific Biosciences real-time, Complete Genomics combinatorial, or Oxford Nanopore GridION/MiniION sequencing. We describe the different sequencing approaches by comparing the platforms. Since the new sequencing systems provide large amounts of data, analyses of them may become bottlenecked. Fortunately, computing has also experienced significant development in the recent years, both in terms of hardware and software (Galvez et al., 2010; Diaz et al., 2014). Consequently, several bioinformatics tools have been developed, and here we summarize the methodologies used for assembly and other analyses. In order to provide broader perspectives, we present different application areas of sequencing technologies in relation to some recent sequencing studies. We draw attention to the whole-genome sequencing of plants, breakthrough outcomes, and great impacts on the understanding of several important biological phenomena.

## 2. Current sequencing technologies

Genome sequencing is being revolutionized by developments in high-throughput technologies. Intense competition between new sequencing technologies has given rise to remarkable innovations. The basic concepts of the currently best-known sequencing platforms are described below.

### 2.1. Roche/454 Life Sciences sequencing

454 Life Sciences (a subsidiary of Roche) developed the first commercial second-generation sequencing platform with the motto of "one fragment-one bead-one read" (http://www.454.com). The backbone of this high-throughput pyrosequencing platform is emulsion-based clonal amplification. The first step of the method is preparation of a single-stranded template DNA library, which involves fragmentation of the genome, ligation of 2 specific adaptors to fragments, and their selection. The protocol continues with emulsion PCR (emPCR), a technique in which the DNA fragments are clonally amplified on beads within a water-in-oil emulsion, followed by enrichment. The emPCR takes place in conditions favoring the binding of only one fragment to individual beads and generates millions of clonally amplified sequencing templates on each bead. In the next step, DNA beads are deposited into a PicoTiterPlate device, which enables loading one bead per each well, and the sequencing run starts. The signal is acquired by the sequencing-by-synthesis principle. The bases are flowed sequentially across the device, and when there is complementation with the template, a pyrophosphate signal is generated and recorded by a charge-coupled device camera. Accordingly, the simultaneous sequencing of the entire genome in picoliter-sized plates occurs.

Depending on the complexity of the genome of interest, the 454 sequencing system offers shotgun alone and in combination with paired-end sequencing approaches for whole-genome sequencing. Additionally, targeted resequencing, epigenetic, metagenomic, and transcriptome sequencing studies have been achieved with this system. The first study using this technique was reported in 2005 (Andries et al., 2005). Since then, more than 445/2000 studies applying the Roche 454 Life Sequencing system for various organisms have been published (http://454.com/publications/publications.asp?postback=true). Recently, the platform was upgraded with longer read capacity of up to 1000 b and higher performance (http://454.com/products/gs-flx-system/index.asp).

### 2.2. ABI/SOLiD sequencing

In 2008, a new massively parallel sequencing technology, SOLiD (Sequencing by Oligonucleotide Ligation and Detection), was developed by Life Technologies. The process starts with fragment library or mate-paired library preparation. As with Roche 454 sequencing, amplification of the template is also achieved by emPCR in this system. After clonal amplification of the template on beads and their enrichment are achieved, beads with extended templates are immobilized onto a flow-cell surface followed by sequencing reaction. The sequencing-by-ligation chemistry is applied in the SOLiD system.

Subsequent ligation, detection, and cleavage of a set of 4 fluorescently labeled 8-mer probes to sequencing primers are performed. The first 2 bases are complementary to the template; the next 3 bases are degenerate, consisting of 64 possible combinations, and the last 3 nucleotides are universal for each probe. Following the incorporation of the first 2 bases, the other 3 bases of the probe are cleaved, leaving a free 5'-phosphate group ready for further ligation. Therefore, the bases at positions 1, 2 and 6, 7 and 11, 12 (and so on) are determined. In the next round, the primer complementary to position n – 1 of the adapter sequence is annealed, which is followed by 4 more further rounds until annealing of primer at position n – 4. At this point, there are 4 dinucleotides for each fluorescent dye to encode. Since each base is interrogated twice by 2 different primers, it is possible to determine which base is at which position. Taking advantage of the 2-color base encoding, the system offers a high sequencing accuracy.

The technology supports a wide range of applications that includes whole genome and transcriptome sequencing, methylation analyses, chromatin immunoprecipitation sequencing, small RNA sequencing, and metagenomic studies (http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing/Next-Generation-Sequencing/Publications-Literature.html).

### 2.3. Solexa/Illumina sequencing

The third sequencing platform is Illumina, which is capable of sequencing hundreds of millions of fragments. The genome analyzer instrument was commercialized in 2006 by Solexa/Illumina. The sequencing chemistry is based on reversible terminators. Modified dNTP containing a fluorescently labeled terminator that allows only a single-base extension is used in the sequencing reaction. The method consists of 3 stages. As with the other platforms, the Illumina sequencing workflow starts with library preparation, including fragmentation of DNA and adaptor ligation. The library is then flowed across a solid surface, and the fragments (each around 200 bp long) bind to this surface, following "bridge amplification" of the templates to generate clusters.

Two different primers complementary to the adaptors are also attached to the surface, and 1 of the primers has a cleavage site. Thus, the single-stranded DNA (ssDNA) molecules can twist and hybridize to PCR primers, forming bridges. This allows the ssDNA to be extended to form double-stranded DNA (dsDNA). After denaturing and washing-up steps, dense clusters of ssDNA fragments stay on the surface. This solid-phase amplification creates 1000 copies of each fragment in close proximity on the surface. Amplification of templates on a solid surface is the major innovation of this system, which favors signal detection. To prevent extension of DNA molecules onto each other, 3'-ends of the fragments are blocked by terminal transferase, following addition of 4 types of terminator bases. After washing of nonincorporated nucleotides, the fluorescent signals are recorded, terminators are removed, and the next round of one-base extension starts. Since one base is added at a time, the read lengths are equivalent.

This method has been widely used for whole-genome sequencing (Potato Genome Sequencing Consortium, 2011), transcript profiling of both protein-coding genes and small RNAs (Eldem et al., 2012), and gene regulation studies (Yanik et al., 2013). With the latest improvements in 2011, Solexa/Illumina has significantly enhanced the platform, increasing read length and overall throughput (http://www.illumina.com/technology/solexa_technology.ilmn).

### 2.4. Which sequencing method to choose?

We have been witnessing the beginning of a new era in genome research with the arrival of the new high-throughput sequencing technologies. Since a variety of sequencing platforms are available, it raises the question of which method is best. It must be said that there is no definitive answer for this question. The decision depends on numerous factors, involving the research goal, the starting material to be sequenced, and the available budget.

The different sequencing platforms differ in several ways, such as read length and sequencing chemistry (Table 1). Each of them has pros and cons. For that reason, in some studies, different platforms have been used simultaneously (Potato Genome Sequencing, 2011; Brenchley et al., 2012; Tomato Genome Consortium, 2012).

Whole-genome shotgun sequencing is a common sequencing strategy. It has been successfully implemented on a variety of eukaryotic genomes. These include poplar (Tuskan et al., 2006), papaya (Ming et al., 2008), cucumber (Huang et al., 2009 ), apple (Velasco et al., 2010), *Brachypodium* (International Brachypodium Initiative, 2010), soybean (Schmutz et al., 2010), and potato (Potato Genome Sequencing, 2011), among others. On the other hand, several factors may complicate whole-genome sequencing, especially in plant genomes that may have certain characteristics that complicate sequencing studies. These include large genome size (>1 Gbp), high

**Table 1.** Technical properties of the 3 second-generation platforms.

| Properties | Roche/454 | ABI/SOLiD | Solexa/Illumina |
|---|---|---|---|
| Sequencing chemistry | Pyrosequencing | Bridge amplification | Sequencing-by-synthesis |
| Read length (b) | 1000 | 75 | 2 × 101 |
| Number of reads | 1 million | 5 billion | 3 and 6 billion (single and paired-end reads, respectively) |
| Total throughput | 700 Mb | 120 Gb | 540–600 Gb |
| Base-calling error rate (%) | 1–3 | 0.01 | 0.1 |
| Run time | 23 h | 14 days | 8.5 days |
| Price per Mb ($) | 8 | 0.05 | 0.02 |

CG content, polyploidy, high heterozygosity, large number of transposable elements, and repetitive nature of the genome, which arise as big challenges for the whole-genome shotgun approach.

For instance, it has been suggested that short-read shotgun strategies should be avoided when assembling particularly highly repetitive plant genomes (Feuillet et al., 2011; Taudien et al., 2011). As longer reads are preferable for accurate assembling and for interpreting repetitive sequences, the Sanger method (first-generation sequencing platform) would be the best, but the cost, time, labor, and equipment required would be prohibitive. Hence, the Roche/454 technology, offering the longest read-length capacity of the second-generation platforms, appears as the method of choice for those studies without considering the total sequencing cost differences between such platforms. Additionally, having the highest speed, the Roche/454 technology has an excellent advantage for analysis of massive sample sets, at least until the third-generation sequencing platforms are fully developed.

Sequence-variation detection represents one of the major research goals of the sequencing applications. Nevertheless, errors in base-calling may lead to both false positives and false negatives. In this respect, the 2-base color coding of the SOLiD system has the highest accuracy compared to the others, and consequently it emerges as the choice for detection of variations in sequencing (Liu et al., 2012).

On the other hand, the new sequencing technologies have greatly increased the potential of epigenomic research. Though short reads may cause ambiguities for particular applications, such as de novo assembly, they are acceptable for chromatin immunoprecipitation sequencing (ChIP-Seq). Thus, the highest throughput of the Illumina system makes it the preferred platform for such studies of DNA–protein interactions (Park, 2009).

The new sequencing technologies greatly benefit from their deep coverage, which may compensate for their failure rate in general. However, when the repetitive sequence is longer than the read length, deeper coverage is not enough to avoid the generation of gaps during assembly. In such cases, paired-end sequencing, in which both ends of fragments are sequenced, is needed to span those gaps (Schatz et al., 2010). Moreover, paired-end sequencing is also advantageous, especially for de novo sequence assembly (Wang et al., 2010; Wang S et al., 2012). This way, more detailed and accurate information about the sequenced fragment is achieved. Currently, most of the new sequencing devices offer both standard and paired-end sequencing; hence, it should not be a restricting criterion for most platforms.

On the other hand, the bacterial artificial chromosome (BAC) approach known as BAC-by-BAC couples physical mapping with sequencing and may allow sequencing of complex genomes, as in the case of maize (Schnable et al., 2009). Therefore, the BAC-by-BAC approach served to improve whole-genome sequencing assembly (Haiminen et al., 2011). Additionally, the isolation and sequencing of chromosomes and even their arms has been developed as an alternative approach to sequence large and polyploid genomes, such as hexaploid wheat (Dolezel et al., 2007; Paux et al., 2008; Hernandez et al., 2012).

## 3. Genome-sequence analysis tools

While developments in sequencing technology make it possible to obtain large-scale sequence data in a short time, the assembly and analysis of sequences remains a challenging task. Thus, much of the effort in recent years has been dedicated to developing and improving bioinformatics tools.

Different scenarios may cause erroneous base-calling in the sequencing platforms. For instance, most of the errors that come from indels in 454 reads are caused by incorrect homopolymer length calls. On the other hand, the sequencing chemistry of Illumina ensures that only one nucleotide is incorporated in each cycle, avoiding such homopolymer issues. However, this technology may suffer from wrong identification of the incorporated nucleotide. Finally, areas in the genome with a high single-nucleotide polymorphism (SNP) density may get lower coverage with the ABI/SOLiD system. Thus, the sequencing data are managed and analyzed with advanced bioinformatics tools. Currently, a number of bioinformatics software packages are available, which are essentially used for different purposes, including alignment, assembly, annotation, and sequence-variation detection (e.g., identification of SNPs) (Imelfort and Edwards, 2009; Scheibye-Alsing et al., 2009; Lerat, 2010; Paszkiewicz and Studholme, 2010; Bao et al., 2011).

The first step of assembly is to control the quality of the raw sequences. Since most of the machines produce the data in FASTA or FASTQ formats, the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) emerge as useful tools for the preprocessing steps.

After quality check and trimming (such as removing adapter sequences and short reads), the next step of sequencing data analysis is assembly of the sequences. The genome-assembly process can be divided into 2 steps: draft assembly and assembly improvement (finishing). In the majority of the cases, 98% of the genome is covered by draft assembly with an error rate of 1/2000 b, while this ratio is 5-fold lower in finished assemblies (Lapidus, 2009).

Usually, before assembly, repetitive elements are identified and filtered out from the dataset. Repetitive elements are one of the challenging issues for assembly procedures. In fact, the majority of the gaps in an assembly are caused by repeated sequences (Cahill et al., 2010). Sequencing with longer reads emerges as a good way out. Paired-end sequencing is also commonly used for this purpose. Depending on availability, repetitive elements are computationally detected by homology searches to known repeat sequences. REPuter (Kurtz et al., 2001), Tandem Repeat Finder (Benson, 1999), and RepeatMasker (http://www.repeatmasker.org) are among the most common programs for detecting such repetitive elements. When there is a lack of a reference genome, repetitive elements are identified de novo. The basic workflow pipeline is composed of masking the known repeats, de novo repeat finding on the masked genome, and classification of the newly identified repeats. Detailed de novo repeat discovery tools are mentioned elsewhere (Bergman and Quesneville, 2007). RECON (Bao and Eddy, 2002), RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html), RepeatScout (Price et al., 2005), and REPET (Flutre et al., 2011) are examples of the best-known software packages for this purpose.

Presently, a number of assembly approaches are applied for short-read assemblies. The first assemblers are based on a simple strategy known as the greedy algorithm, which is an implementation of finding the shortest common supersequence (Narzisi and Mishra, 2011). The algorithm proceeds as follows: 1) pairwise comparison of all sequences is done to identify overlapping sequences and merge the best overlapped sequences; and 2) these steps are repeated until no more sequences are left to be merged. The greedy algorithm has been used mainly for assembling small genomes. On the other hand, since the algorithm needs local information at each step, the presence of complex repeats may lead to misassemblies. The most accepted packages based on this method are TIGR (http://www.jcvi.org/cms/uploads/media/TIGR-assembler.pdf) (Sutton et al., 1995), PHRAP (http://www.phrap.org/phredphrapconsed.html), CAP3 (Huang and Madan, 1999), PCAP (Huang and Yang, 2005), Phusion (Mullikin and Ning, 2003), SSAKE (Warren et al., 2007), and VCAKE (Jeck et al., 2007).

With the advent of sequencing technologies, new assemblers have been developed, particularly for more complex genomes. The overlap-layout-consensus (OLC) approach analyzes the overlap graph of the sequencing reads and searches for a consensus genome. When applied to short reads, the main drawback of this approach is that it shows low performance, as too many overlaps have to be calculated. Examples of genome-assembly software packages applying the OLC approach are ARACHNE (Batzoglou et al., 2002) and Atlas (Havlak et al., 2004).

Since the computer memory required by the OLC approach is quite high, alternative methods were developed. The most recent assemblers generally use De Bruijn graphs. The method compresses redundant sequences and does not need all reads to perform the alignment. The principle is based on k-mer graphs. Thus, the reads are partitioned into certain k-mers. Each edge of linking nodes is a unique subsequence of k-mer length, and the nodes of the graph are assigned as common subsequences of k –1 length. Since the analysis is strictly dependent on the k-mer size, the main critical point of this approach is setting the optimal parameters. Compared to the OLC method, shared k-mers are generally easier to find. Hence, the method is much faster and needs much less computational power to perform the assemblies. Since the publication of EULER (Pevzner et al., 2001), the first assembler using De Bruijn graphs, many other packages such as Velvet (Zerbino and Birney, 2008) and ABySS (Simpson et al., 2009) have been released.

On the other hand, the string graph method (Myers, 2005) is a variant of the OLC approach. In this approach, overlaps between sequences are found, and the constructed graph is transformed into a string graph. The sequences are not fragmented into k-mers. Therefore, it is a memory-efficient strategy. EDENA (Hernandez et al., 2008) was the first assembling software implementing the string graph approach. Read Joiner (Gonnella and Kurtz, 2012) and SGA (Simpson and Durbin, 2012) are the other string graph-based assemblers.

Many tools and algorithms relevant to bioinformatics analyses of sequencing data have been published. Two classes of assemblies are carried out: map-based and de novo. Map-based assemblies refer to the reconstruction of sequences by alignments to previously resolved reference sequences. Although the BLAST (Altschul et al., 1990) and Blat (Kent, 2002) analysis tools can be used for alignments, more multifaceted software programs have been developed. For this purpose, Maq (http://maq.sourceforge.net/maq-man.shtml), Bowtie (Langmead et al., 2009), SOAPaligner (http://soap.genomics.org.cn/soapaligner.html), and BWA (http://bio-bwa.sourceforge.net/bwa.shtml#13) (Li and Durbin, 2009) are among the most preferred programs.

The de novo assemblies define the reconstruction of sequences without a reference sequence. SOAPdenovo (http://soap.genomics.org.cn/soapdenovo.html) and Velvet (http://www.ebi.ac.uk/~zerbino/velvet) are common de novo assembling programs for short reads. Additionally, the GS De Novo Assembler and GS Reference Mapper programs were developed by 454 Life Sciences to assemble shotgun reads into contigs and to map them against a reference sequence, respectively. On the other hand, Illumina developed a genome alignment program called ELAND for map-based assembly purposes.

In the last step of assemblies, the assembling results are statistically evaluated. Thus, the length distribution of contigs, the average and largest contig sizes, and N50 and N80 sizes are considered as the major indicators of a sequence assessment (Zhang et al., 2011a).

## 4. Sequencing applications

NGS technologies have contributed a series of genetic improvements in plant breeding and biotechnology. In contrast to first-generation sequencing, second- and third-generation technologies produce an enormous volume of sequence data at a much lower cost, making the system versatile for plenty of applications (Metzker, 2009; Llaca, 2012). Today, second-generation sequencing is extensively used in the discovery of genetic markers, gene expression profiling through mRNA sequencing, and comparative and evolutionary studies to answer a diverse set of biological questions (Wang et al., 2009; Jia et al., 2013; Nystedt et al., 2013; Dohm et al., 2014; Sierro et al., 2014). Even more promising for the immediate future is third-generation sequencing, being mostly in active development nowadays.

### 4.1. Whole-genome sequencing

The broadest application of the new sequencing approaches to plant species may be whole-genome sequencing (WGS) to reveal the full sequence and genetic structure of genomes. In WGS projects such as those for strawberry (Shulaev et al., 2011) and wheat (Brenchley et al., 2012), whole-genomic DNA content was first randomly cut into fragments of different sizes. BAC-end sequencing was then carried out and the obtained reads were assembled using powerful bioinformatics tools. The WGS approach can be accomplished not only for resequencing, but also for de novo projects.

Although it takes more time, the de novo sequencing of whole DNA or mRNA is useful for producing draft genomes when the plant genome of interest is unknown. For instance, draft genomes of several crop species such as einkorn (Ling et al., 2013), as well as wheat and *A. tauschii* (Jia et al., 2013), were produced using the WGS approach. Apart from this, resequencing is mostly used in transcriptome profiling and SNP discovery for marker development (Llaca, 2012). Thus, a high-quality reference genome of potato was revealed utilizing the WGS approach and SNP identification was performed to compare a homozygous doubled-monoploid line with its heterozygous diploid line (Xu et al., 2011). More recently, several accessions of watermelon were resequenced and compared with each other. Thus, a total of 6,784,860 SNPs were identified, representing the genetic diversity of the crop species (Guo et al., 2013).

### 4.2. Transcriptome sequencing

So-called RNA sequencing (RNA-Seq) is rapidly becoming the method of choice for gene expression analysis, replacing other profiling approaches such as microarrays. It must be noted that RNA-Seq is not a type of direct RNA sequencing, but rather is done after cDNA generation via reverse transcriptase. True and direct RNA sequencing can be accomplished with third-generation sequencing platforms, which are beyond the scope of this review. The rationale behind RNA-Seq is that the coverage depth of a particular sequence is proportional to its expression level (Jain, 2012). In transcriptome sequencing, total mRNA isolated from a diverse set of cells or tissues subjected to different conditions is first converted to cDNA fragments as indicated above, and then randomly sheared, followed by end-sequencing (Wang et al., 2009; Marguerat and Bähler, 2010). Adapting the new sequencing platforms to transcriptome sequencing brought about several advantages, such as producing cost-effective transcriptome reads in a relatively short time (Góngora-Castillo et al., 2012). Differently from genome sequencing, it is possible to obtain a repertoire of transcripts present in a specific sample under a predefined stress or condition using RNA-Seq (Hirsch and Buell, 2013). In other words, RNA-Seq data represent all expressed sequences of the plant in a spatiotemporal manner.

Several RNA-Seq projects have been undertaken for crop plants. These studies enable gene discovery, SNP detection (Novaes et al., 2008; Angeloni et al., 2011), and transcript annotation and quantification (Der et al., 2011), as well as comparative gene expression analyses (Strickler et al., 2012). In one of those studies, differential expressions between homologs in 3 different genomes of wheat were observed by investigating their transcriptomes (Leaungthitikanchana et al., 2013). Similarly, comparative gene-expression analyses have been performed in the garden pea (*Pisum sativum*) (Franssen et al., 2011b) and bracken fern (*Pteridium aquilinum*) (Der et al., 2011) employing the 454 sequencing platform. The transcriptomes of tomato and its wild relatives were also dissected for differential gene expression and SNP detection using Illumina sequencing (Koenig et al., 2013). Additionally, large-scale transcriptome profiling studies such as the 1000-plant genome-sequencing project can give insights about the adaptation of plants to differing environmental conditions (Franssen et al., 2011a), among other scientific insights.

### 4.3. Small-RNA deep sequencing

Small RNA (sRNA) belong to a class of noncoding RNA (ncRNA), being ~21 nucleotide-long nonprotein-coding molecules that have important roles in living cells, including plant development and metabolism. The majority of sRNA can be grouped as microRNA (miRNA), which have posttranscriptional regulatory functions, and small interfering RNA (siRNA), mainly responsible for gene-silencing mechanisms (Vaucheret, 2006; Kurtoglu, 2013). Sequencing of small RNA libraries prepared from different tissue types under different conditions became a widely used method for sRNA identification and functional studies. Prior to sequencing of small RNA molecules, they are first isolated and size-selected utilizing a polyacrylamide gel electrophoresis system, followed by reverse transcription and an optional PCR step. The implementation of the new sequencing technologies resulted in considerable increase in the number of studies based on deep-sequencing of sRNA libraries constructed from plant tissues grown under normal or stressed conditions (Cantu et al., 2010; Kenan-Eichler et al., 2011; Eldem et al., 2012; Gupta et al., 2012; Tang et al., 2012; Yao and Sun, 2012; Li et al., 2013; Yanik et al., 2013).

### 4.4. Probing DNA-protein interaction (ChIP-Seq)

Chromatin immunoprecipitation followed by direct sequencing is a widely used method to determine genome-wide profiles of DNA–protein interactions (Wold and Myers, 2008; Park, 2009; Varshney et al., 2009). With the advent of the new sequencing technologies, ChIP sequencing has surpassed the microarray-based ChIP-Chip method, which was previously used in such studies, offering a tremendous data throughput increase with low cost. Performing strong bioinformatic analyses on these data helps to reveal gene-regulation and epigenetic-modification mechanisms.

Thus, protocols have been developed for ChIP-Seq in plant species to study interactions between transcription factors (TFs) and DNA in vivo (Kaufmann et al., 2010). For instance, following this procedure, the chromatin complexes of soybean seedlings were isolated and DNA was treated with antibodies developed against YABBY or NAC TF. DNA was recovered by dissociating precipitated DNA–antibody complexes. ChIP-Seq was performed using the Illumina HiSeq 2000 platform. Thus, identification of genome-wide NAC and YABBY TF binding sites has contributed to a better understanding of the transcriptional gene regulation networks in soybean cotyledons about to develop into photosynthetic tissue (Shamimuzzaman and Vodkin, 2013). In another line of research, MADS-domain TF complexes in *Arabidopsis* flower development were also characterized using the same protocol (Smaczniak et al., 2012).

### 4.5. Exome sequencing

Exome sequencing is a technique in which only the protein-coding stretches of genes are being sequenced. Thus, the method first requires the selection of all the protein-encoding DNA regions (exons), which are then sequenced using one of the new platforms. It has the advantage of producing sequencing data in a quicker and cheaper way than WGS, since the exome comprises only a small (and sometimes even very small) portion of the genome.

Exome sequencing is usually used to identify mutations in protein-coding genes (Schneeberger, 2014). In a recent study, exome capture and sequencing coupled with custom-developed bioinformatics tools was used to identify mutations in mutant populations of rice (*Oryza sativa*) and wheat (*Triticum aestivum*). This provided a method for large-scale mutation discovery, allowing generation of useful polymorphism database resources in a quick and rather inexpensive way (Henry et al., 2014). Nucleotide polymorphism and copy-number variant detection utilizing this method was conducted in another study on the switchgrass *Panicum virgatum* (Evans et al., 2014). In that study, a total of 1,395,501 SNPs and 8173 putative copy-number variants were detected. Hence, the applicability of exome capture for genomic variation studies in polyploid species with large, repetitive, and heterozygous genomes was shown. In a similar study carried out in hexaploid wheat (*T. aestivum*), a total of 10,251 SNP markers were developed, employing targeted resequencing of the wheat exome to produce large amounts of genomic data for 8 varieties. These exome-based SNP markers provide a prominent source of information, especially for wheat breeders (Allen et al., 2013).

## 5. Sequenced plant genomes

Along with the breakthrough in sequencing technology, there has been a great accumulation of genome-sequence data of plant species (Figure 1). The application of the new sequencing technologies to plant genomes gave rise to rapid improvements in crop science. Genomic-sequence availability and easy access to such data enabled researchers to discover and develop genetic markers, improve knowledge of breeding, and reveal evolutionary relationships between the sequenced species via comparative genomic analysis in general and synteny approaches in particular. Currently, bread wheat (*Triticum aestivum* 'Chinese Spring', 2n = 6x = 42), which is a major staple food with annual production of approximately 700 × $10^6$ t (http://www.fao.org), is being sequenced by the International Wheat Genome Sequencing Consortium (IWGSC), adopting a chromosome-by-chromosome approach. Due to the huge size and complex nature of the wheat genome (17 Gbp, AABBDD), researchers have sorted chromosomes and performed synteny with model grass genomes (Choulet et al., 2014).

Much effort has been carried out in elucidating genomic backgrounds in order to improve grain yield and quality against some of the limiting factors, such as biotic and abiotic stresses. Thus, 454 pyrosequencing was used to survey individual chromosomes (Vitulo et al., 2011; Hernandez et al., 2012; Poursarebani et al., 2014; Sergeeva et al., 2014). Recently, a bread wheat (*T. aestivum*) genome draft was obtained by Illumina sequencing of the flow-sorted chromosomes (International Wheat Genome Sequencing Consortium, 2014) and was simultaneously

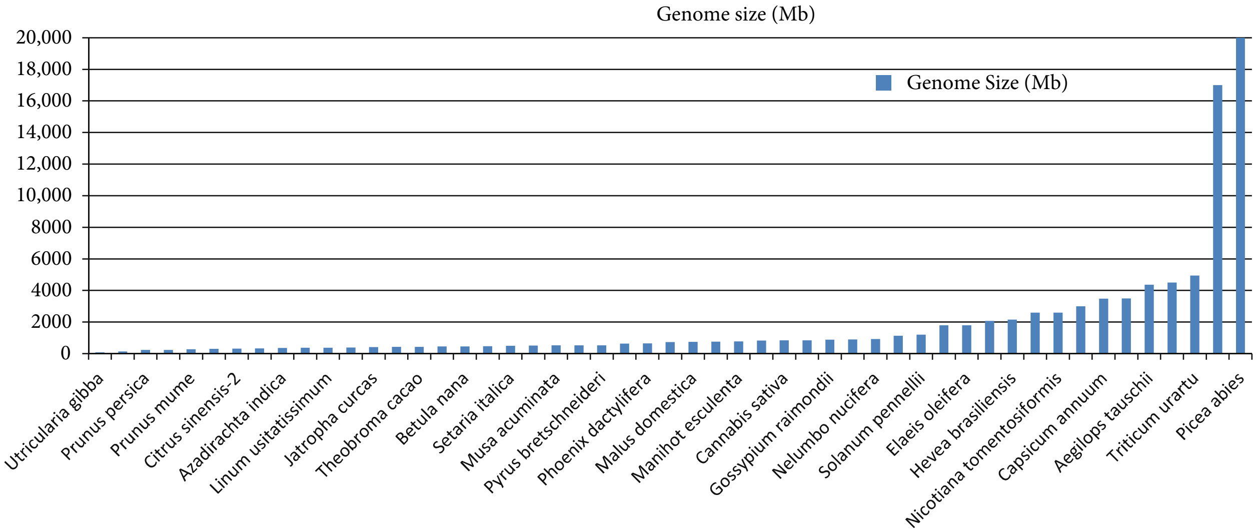

**Figure 1.** Plant genome sizes.

published with the first wheat-chromosome (3B) reference sequence (Choulet et al., 2014). Comparative gene analyses of wheat subgenomes and extant diploid and tetraploid wheat relatives showed that both a high sequence similarity and a structural conservation are retained, with limited gene loss after polyploidization. The study showed evidence of dynamic gene gain, loss, and duplication across the genomes. Such alterations would have a critical role in wheat adaptation in a diverse set of climatic conditions (Langridge, 2012).

Before the bread wheat genome draft, the draft genome sequences of 2 progenitors of the hexaploid wheat had been simultaneously published: *Triticum urartu* and *Aegilops tauschii* (Jia et al., 2013; Ling et al., 2013). *Triticum urartu* (AA, 2n = 2x = 14), the progenitor of the A genome of wheat (Chantret et al., 2005; Dvorak and Akhunov, 2005), was sequenced on the Illumina platform using the whole-genome shotgun strategy, resulting in 448.49 Gbp of high-quality sequence data corresponding to ~91× coverage of an estimated 4.94 Gbp genome size. Additionally, a total of 34,879 protein-coding gene models were predicted using transcriptome-sequence data obtained from the same study (Ling et al., 2013). Additionally, *Aegilops tauschii* (DD, 2n = 2x = 14) was sequenced using the same Illumina whole-genome shotgun strategy. Jia et al. generated 398 Gbp of high-quality reads (90× coverage), representing 97% of the genome size of 4.36 Gbp. A 117-Mb transcriptome assembly was generated from RNA-Seq data obtained from different tissues and used to predict 34,498 high-confidence protein-coding loci (Jia et al., 2013). The data revealed in these articles identified genes that are of agronomical importance, such as resistance to abiotic stresses and nutritious quality. Hence, these developments help to understand the environmental adaptation of wheat, together with its genomic nature. Additionally, the strategy developed for genome sequencing and assembly of wheat could also be adapted to other large and complex plant genomes, as well.

On the other hand, cotton, as one of the most economically important crops for the textile industry, was another genome sequenced with the new technologies. Wang et al. published a draft genome of *Gossypium raimondii* (2n = 2x = 26), a putative D-genome donor, employing an Illumina paired-end sequencing strategy. A total of 78.7 Gbp Illumina reads were produced, with 103.6× genome coverage. The draft sequence was 775.2 Mbp, accounting for 88.1% of the estimated genome size. Combining ab initio predictions, homology searches, and EST alignment methods, a total of 40,976 protein-coding genes were identified and 92.2% of them were supported by transcriptome-sequencing data. Comparative analysis with *T. cacao*, *A. thaliana*, and *Zea mays* showed that *G. raimondii* contains a high proportion of transposable elements and a lower gene density than the other species, although they all have a similar number of gene families. Another finding of this study revealed the evolutionary relationships between *G. raimondii* and *T. cacao*, which probably diverged 33.7 million years ago. The authors also claimed that both of these draft sequences will serve as a reference for the assembly of the tetraploid *G. hirsutum* genome and a useful source for genetic improvement of cotton quality and yield (Wang K et al., 2012).

Sugar beet (*Beta vulgaris*) is another important crop, which substantially contributes to world-wide sugar production. In 2013, the reference genome sequence of this species was released, representing 85% of its 576-Mbp genome size. A combination of 454, Illumina, and Sanger sequencing platforms were utilized in that study. In total, 27,421 protein-coding genes were identified and evidenced by RNA-Seq data. Based on intraspecific genomic analysis of 5 different sugar beet species, 7 million genomic variants were identified, together with large constant regions. The availability of the sugar beet genome enables the discovery of agronomically important traits that may increase the quality and productivity of the plant. The genome sequences would also contribute to comparative studies with Caryophyllales species and other flowering plants (Dohm et al., 2014).

Conifers, as the largest division of gymnosperms, have had widespread distribution in forests for almost 200 million years (Nystedt et al., 2013). Besides the economic value of conifers as a source of timber, they are of great ecological importance, since a high proportion of plant photosynthesis is met by these woody plants. However, genomic studies of conifers require much effort, due to their huge genome size and repetitive nature. In a recent study, de novo sequencing of the coniferous tree Norway spruce (*Picea abies*) was performed using the Illumina technology, following a whole-genome shotgun approach. A hierarchical genome-assembly strategy was developed to combine haploid and diploid genomic and RNA-Seq data. The genome size of *P. abies* was estimated as 19.6 Gbp. On the contrary, only 28,354 high-confidence protein-coding sequences were predicted from EST and transcriptome data, which is similar to the almost 40-times smaller sugar beet genome. In this case, the large genome size was interpreted as a result of the accumulation of transposable elements (TEs) and, especially, long terminal repeats, due to the possibility of lacking an efficient elimination mechanism. Furthermore, a model for conifer genome evolution has been proposed, which suggests that the TE removal is less active than in most other plant species (Bennetzen et al., 2005), with TE insertions into genes resulting in large introns and pseudogenes (Nystedt et al., 2013). Additional genome sequencing of conifer species would enable comparative analyses and provide further resources to understand the evolution of important traits for seed plants.

Additionally, *Eucalyptus* is one of the most widespread tree genera, with more than 20 × $10^6$ ha of land planted throughout the world. This noteworthy diversity and adaptability of eucalyptus can be exploited as a sustainable energy source, mostly providing cellulose for the paper industry. Myburg et al. (2014) sequenced and assembled a reference sequence for *Eucalyptus grandis*. They used Sanger WGS, paired BAC-end sequencing, and a high-density genetic linkage map (Myburg et al., 2014). The *E. grandis* genome size was estimated to be 640 Mbp, and 36,376 protein-coding loci were predicted. For further gene-expression analyses, RNA-Seq reads were obtained from diverse sets of *E. grandis* tissues by Illumina sequencing. This was the first reference genome published for the eudicot order of Myrtales, providing a resource to gain insights about the genetic nature of large woody perennials.

Tobacco (*Nicotiana tabacum,* 2n = 4x = 48) is a widely cultivated nonfood crop used as a model organism in molecular plant studies (Zhang et al., 2011b). In a recent study, 3 inbred varieties were sequenced using an Illumina WGS approach. Estimated genome sizes were reported as 4.41 Gbp for *N. tabacum* TN90, 4.60 Gbp for *N. tabacum* K326, and 4.57 Gbp for *N. tabacum* BX (with 49×, 38×, and 29× coverage, respectively). Based on NGS transcriptome data, protein-coding sequences ranging from 81,000 to 94,000 were identified in the 3 varieties. The *N* gene and *va* allele responsible for hypersensitive response to the tobacco-mosaic virus and potyvirus were also investigated in these lines. The authors foresaw that the draft genomes would significantly contribute to functional genomic studies of the *N. tabacum* model organism (Sierro et al., 2014).

Watermelon (*Citrullus lanatus*) is one of the most consumed fresh fruits, with annual production of 90 × $10^6$ t. A high-quality draft genome sequence was published recently. De novo sequencing was generated utilizing the Illumina platform, resulting in reads of 46.18 Gbp, corresponding to 108.6× coverage of the estimated 425-Mbp genome size of this species. Subsequently, a total of 23,440 protein-coding genes were identified using ab initio predictions, cDNA/EST, and homology-mapping methods. Furthermore, 20 watermelon accessions were resequenced following the paired-end Illumina strategy. Among them, 6,784,860 candidate SNPs and 965,006 small indels were identified, representing a germplasm biodiversity that can contribute to the breeding of the species. Additionally, the comparative analyses of the transcriptome data should contribute to the understanding of the genetic diversity and molecular mechanisms underlying some biological processes in watermelon populations. Thus, the evolutionary scenario proposed in this study should shed light on the genetic backgrounds of modern cultivars (Guo et al., 2013).

In addition to the draft and reference genomes mentioned above, more than 50 plant species have been sequenced so far, as listed in Table 2 and Figure 2.

In conclusion, NGS has becoming a powerful tool for decoding the entire genome of a plant species as well

**Table 2.** Draft plant genomes using the second-generation sequencing.

| Common name | Species | Genome size (Mb) | Year | Reference |
|---|---|---|---|---|
| Cucumber | *Cucumis sativus L.* | 367 | 2009 | (Huang et al., 2009 ) |
| *Brachypodium* | *Brachypodium distachyon* | 355 | 2010 | (International Brachypodium Initiative 2010) |
| Apple | *Malus domestica* | 742 | 2010 | (Velasco et al., 2010) |
| Jatropha | *Jatropha curcas* | 410 | 2010 | (Sato et al., 2010) |
| Salt cress | *Thellungiella parvula* | 140 | 2011 | (Dassanayake et al., 2011) |
| Peach | *Prunus persica* | 230 | 2011 | (Ahmad et al., 2011) |
| Strawberry | *Fragaria vesca* | 240 | 2011 | (Shulaev et al., 2011) |
| Cacao | *Theobroma cacao* | 430 | 2011 | (Argout et al., 2011) |
| Barrel medic | *Medicago truncatula* | 475 | 2011 | (Young et al., 2011) |
| Canola | *Brassica rapa* | 516 | 2011 | (Wang et al., 2011) |
| Palm | *Phoenix dactylifera* | 650 | 2011 | (Al-Dous et al., 2011) |
| Pigeonpea | *Cajanus cajan* | 833 | 2012 | (Varshney et al., 2012) |
| *Cannabis* | *Cannabis sativa* | 843 | 2011 | (van Bakel et al., 2011) |
| Potato | *Solanum tuberosum* | 844 | 2011 | (Xu et al., 2011) |
| Flax | *Linum usitatissimum* | 373 | 2012 | (Wang Z et al., 2012) |
| Dwarf birch | *Betula nana* | 462 | 2013 | (Wang et al., 2013) |
| Chinese plum | *Prunus mume* | 280 | 2012 | (Zhang Q et al., 2012) |
| Millet | *Setaria italica* | 490 | 2012 | (Zhang G et al., 2012) |
| Banana | *Musa acuminata* | 523 | 2012 | (D'Hont et al., 2012) |
| Cotton | *Gossypium raimondii* | 880 | 2012 | (Wang K et al., 2012) |
| Tomato | *Solanum lycopersicum* | 900 | 2012 | (Tomato Genome Consortium, 2012) |
| Bread wheat | *Triticum aestivum* | 17,000 | 2014 | (International Wheat Genome Sequencing Consortium, 2014) |
| *Nicotiana benthamiana* | *Nicotiana benthamiana* | 3000 | 2012 | (Bombarely et al., 2012) |
| Melon | *Cucumis melo* | 450 | 2012 | (Garcia-Mas et al., 2012) |
| Cassava | *Manihot esculenta* | 770 | 2012 | (Prochnik et al., 2012) |
| Sunflower | *Helianthus annuus* | 3500 | 2012 | (Staton et al., 2012) |
| Neem | *Azadirachta indica* | 364 | 2012 | (Krishnan et al., 2012) |
| Sugar beet | *Beta vulgaris* | 758 | 2014 | (Dohm et al., 2014) |
| Orange | *Citrus sinensis*-1 | 380 | 2013 | (Xu et al., 2013) |
| Watermelon | *Citrullus lanatus* | 425 | 2013 | (Guo et al., 2013) |
| Pear | *Pyrus bretschneideri* | 528 | 2013 | (Wu et al., 2013) |
| Chickpea | *Cicer arietinum* | 738 | 2013 | (Varshney et al., 2013) |
| Bamboo | *Phyllostachys heterocycla* | 2075 | 2013 | (Peng et al., 2013) |
| Rubber tree | *Hevea brasiliensis* | 2150 | 2013 | (Rahman et al., 2013) |
| Tausch's goatgrass | *Aegilops tauschii* | 4360 | 2013 | (Jia et al., 2013) |
| Einkorn wheat | *Triticum urartu* | 4940 | 2013 | (Ling et al., 2013) |
| Norway spruce | *Picea abies* | 20,000 | 2013 | (Nystedt et al., 2013 ) |
| Mulberry tree | *Morus notabilis* | 330 | 2013 | (He et al., 2013) |

**Table 2.** (Continued).

| Common name | Species | Genome size (Mb) | Year | Reference |
|---|---|---|---|---|
| Oil palm (African) | *Elaeis guineensis* | 1800 | 2013 | (Singh et al., 2013) |
| Oil palm (South American) | *Elaeis oleifera* | 1800 | 2013 | (Singh et al., 2013) |
| Wild rice | *Oryza brachyantha* | 300 | 2013 | (Chen et al., 2013) |
| Woodland tobacco | *Nicotiana sylvestris* | 2600 | 2013 | (Sierro et al., 2013) |
|  | *Nicotiana tomentosiformis* | 2600 | 2013 | (Sierro et al., 2013) |
| Hot pepper | *Capsicum annuum* | 3480 | 2014 | (Kim et al., 2014) |
| Tobacco | *Nicotiana tabacum* | 4500 | 2014 | (Sierro et al., 2014) |
| Pineapple | *Ananas comosus* | 526 | 2014 | (Zhang et al., 2014) |
| Eucalyptus | *Eucalyptus grandis* | 640 | 2014 | (Myburg et al., 2014) |
| Wild tomato | *Solanum pennellii* | 1207 | 2014 | (Bolger et al., 2014) |
| Lotus | *Nelumbo nucifera* | 929 | 2013 | (Ming et al., 2013) |
| Bladderwort plant | *Utricularia gibba* | 82 | 2013 | (Ibarra-Laclette et al., 2013) |
| Oilseed | *Brassica napus* | 1130 | 2014 | (Chalhoub et al., 2014) |
| Sweet orange | *Citrus sinensis*-2 | 319 | 2014 | (Wu et al., 2014) |

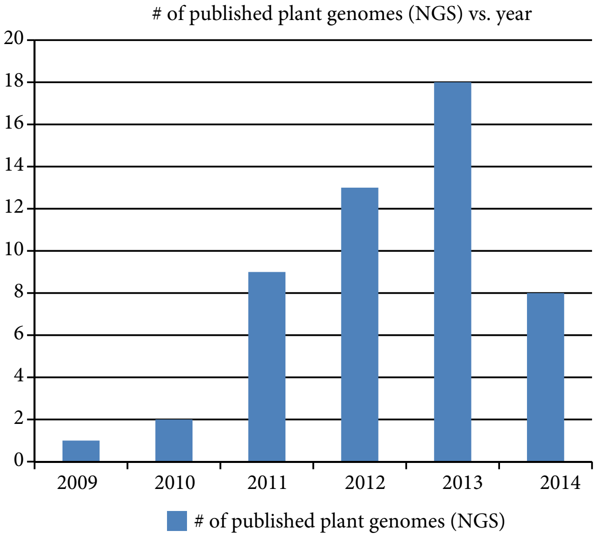

**Figure 2.** Chronology of published plant genomes.

as investigating gene expression profiles and SNPs. As techniques develop, more sequencing strategies will be formed, and selecting and comparing the different NGS platforms will be a challenge. In the past years, more than 50 plant species have been sequenced, providing new resources for plant improvement. However, more bioinformatics tools need to be developed for better use of the data generated from NGS. Sequencing the genome is not the purpose; the final goal should be using this genome to improve crop yield and quality and better understand the evolutionary history.

## 6. Future perspectives

Many new de novo and resequenced plant genomes are expected in the near future for plants in general and crop species in particular, using second- and mostly third-generation sequencing platforms. Further work is needed to complete the biggest and most complex genome drafts while achieving high-quality reference sequences for most plant genomes. This genome knowledge will be coupled with deep gene-expression analyses (RNA-Seq and true RNA sequencing), uncovering alternative splicing, copy-number variations, etc. ChIP-Seq and microRNA-Seq availability for an increasing number of crops will further expand the emerging field of epigenomics. These are all necessary tools for food production and security in a climate-change scenario.

## Acknowledgments

MT and TÜ were funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) with grant numbers 111O036, 112O502, and, 113O016. PH and GD were funded by the Ministerio de Economía y Competitividad (MINECO grants AGL2010-17316 and BIO2011-15237-E) and the Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (MINECO and INIA RF2012-00002-C02-02); Consejería de Agricultura y Pesca (041/C/2007, 75/C/2009 and 56/C/2010), Consejería de Economía, Innovación y Ciencia (P11-AGR-7322 and P12-AGR-0482), and Grupo PAI (AGR-248) of Junta de Andalucía; and Universidad de Córdoba (Ayuda a Grupos), Spain.

## References

Ahmad R, Parfitt D, Fass J, Ogundiwin E, Dhingra A, Gradziel T, Lin D, Joshi N, Martinez-Garcia P, Crisosto C (2011). Whole genome sequencing of peach (Prunus persica L.) for SNP identification and selection. BMC Genomics 12: 569.

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J et al. (2011). De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nat Biotech 29: 521–527.

Allen AM, Barker GLA, Wilkinson P, Burridge A, Winfield M, Coghill J, Uauy C, Griffiths S, Jack P, Berry S et al. (2013). Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). Plant Biotechnol J 11: 279–295.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J Mol Biol 215: 403–410.

Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, Van Gestel J, Timmerman P, Zhu M, Lee E et al. (2005). A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. Science 307: 223–227.

Angeloni F, Wagemaker C, Jetten M, Op den Camp H, Janssen-Megens E, Francoijs KJ, Stunnenberg H, Ouborg N (2011). De novo transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques. Mol Ecol Resour 11: 662–674.

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN et al. (2011). The genome of *Theobroma cacao*. Nat Genet 43: 101–108.

Bao S, Jiang R, Kwan W, Wang B, Ma X, Song YQ (2011). Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet 56: 406–414.

Bao Z, Eddy SR (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12: 1269–1276.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES (2002). ARACHNE: A whole-genome shotgun assembler. Genome Res 12: 177–189.

Bennetzen JL, Ma J, Devos KM (2005). Mechanisms of recent genome size variation in flowering plants. Ann Bot 95: 127–132.

Benson G (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573–580.

Bergman CM, Quesneville H (2007). Discovering and detecting transposable elements in genome sequences. Brief Bioinform 8: 382–392.

Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, Lichtenstein G et al. (2014). The genome of the stress-tolerant wild tomato species *Solanum pennellii*. Nat Genet 46: 1034–1038.

Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller LA, Martin GB (2012). A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. Mol Plant Microbe In 25: 1523–1530.

Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D et al. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491: 705–710.

Cahill MJ, Koser CU, Ross NE, Archer JA (2010). Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies. PLoS One 5: e11518.

Cantu D, Vanzetti LS, Sumner A, Dubcovsky M, Matvienko M, Distelfeld A, Michelmore RW, Dubcovsky J (2010). Small RNAs, DNA methylation and transposable elements in wheat. BMC Genomics 11: 408.

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

Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P et al. (2005). Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). Plant Cell 17: 1033–1045.

Chen J, Huang Q, Gao D, Wang J, Lang Y, Liu T, Li B, Bai Z, Goicoechea JL, Liang C (2013). Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. Nat Commun 4: 1595.

Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, Yun DJ, Bressan RA, Zhu JK, Bohnert HJ et al. (2011). The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet 43: 913–918.

Der JP, Barker MS, Wickett NJ, Wolf PG (2011). De novo characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. BMC Genomics 12: 99.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488: 213–217.

Diaz D, Esteban FJ, Hernandez P, Caballero JA, Guevara A, Dorado G, Galvez S (2014). MC64-ClustalWP2: A highly-parallel hybrid strategy to align multiple sequences in many-core architectures. PLoS One 9: e94044.

Dohm JC, Minoche AE, Holtgrawe D, Capella-Gutierrez S, Zakrzewski F, Tafer H, Rupp O, Sorensen TR, Stracke R, Reinhardt R et al. (2014). The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature 505: 546–549.

Dolezel J, Kubalakova M, Paux E, Bartos J, Feuillet C (2007). Chromosome-based genomics in the cereals. Chromosome Res 15: 51–66.

Dvorak J, Akhunov ED (2005). Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *Aegilops*-*Triticum* alliance. Genetics 171: 323–332.

Eldem V, Çelikkol Akçay U, Ozhuner E, Bakır Y, Uranbey S, Unver T (2012). Genome-wide identification of miRNAs responsive to drought in peach (*Prunus persica*) by high-throughput deep sequencing. PLoS One 7: e50298.

Evans J, Kim J, Childs KL, Vaillancourt B, Crisovan E, Nandety A, Gerhardt DJ, Richmond TA, Jeddeloh JA, Kaeppler SM et al. (2014). Nucleotide polymorphism and copy number variant detection using exome capture and next-generation sequencing in the polyploid grass *Panicum virgatum*. Plant J 79: 993–1008.

Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K (2011). Crop genome sequencing: lessons and rationales. Trends Plant Sci 16: 77–88.

Flutre T, Duprat E, Feuillet C, Quesneville H (2011). Considering transposable element diversification in de novo annotation approaches. PLoS One 6: e16526.

Franssen SU, Gu J, Bergmann N, Winters G, Klostermeier UC, Rosenstiel P, Bornberg-Bauer E, Reusch TBH (2011a). Transcriptomic resilience to global warming in the seagrass *Zostera marina*, a marine foundation species. P Natl Acad Sci USA 108: 19276–19281.

Franssen SU, Shrestha RP, Bräutigam A, Bornberg-Bauer E, Weber AP (2011b). Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. BMC Genomics 12: 227.

Galvez S, Diaz D, Hernandez P, Esteban FJ, Caballero JA, Dorado G (2010). Next-generation bioinformatics: using many-core processor architecture to develop a web service for sequence alignment. Bioinformatics 26: 683–686.

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González VM, Hénaff E, Câmara F, Cozzuto L, Lowy E et al. (2012). The genome of melon (*Cucumis melo* L.). P Natl Acad Sci USA 109: 11872–11877.

Góngora-Castillo E, Fedewa G, Yeo Y, Chappell J, DellaPenna D, Buell CR (2012). Genomic approaches for interrogating the biochemistry of medicinal plant species. Method Enzymol 517: 139–159.

Gonnella G, Kurtz S (2012). Readjoiner: a fast and memory efficient string graph-based sequence assembler. BMC Bioinformatics 13: 82.

Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z et al. (2013). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet 45: 51–58.

Gupta OP, Permar V, Koundal V, Singh UD, Praveen S (2012). MicroRNA regulated defense responses in *Triticum aestivum* L. during *Puccinia graminis* f.sp. *tritici* infection. Mol Biol Rep 39: 817–824.

Haiminen N, Feltus FA, Parida L (2011). Assessing pooled BAC and whole genome shotgun strategies for assembly of complex genomes. BMC Genomics 12: 194.

Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Weinstock GM, Gibbs RA (2004). The Atlas genome assembly system. Genome Res 14: 721–732.

He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, Lee TH, Wang X, Cai Q, Li D et al. (2013). Draft genome sequence of the mulberry tree *Morus notabilis*. Nat Commun 4: 2445.

Henry IM, Nagalakshmi U, Lieberman MC, Ngo KJ, Krasileva KV, Vasquez-Gross H, Akhunova A, Akhunov E, Dubcovsky J, Tai TH et al. (2014). Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. Plant Cell 26: 1382–1397.

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 18: 802–809.

Hernandez P, Martis M, Dorado G, Pfeifer M, Galvez S, Schaaf S, Jouve N, Simkova H, Valarik M, Dolezel J et al. (2012). Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. Plant J 69: 377–386.

Hirsch CN, Buell CR (2013). Tapping the promise of genomics in species with complex, nonmodel genomes. Annu Rev Plant Biol 64: 89–110.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P et al. (2009). The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41: 1275–1281.

Huang X, Madan A (1999). CAP3: A DNA sequence assembly program. Genome Res 9: 868–877.

Huang X, Yang SP (2005). Generating a genome assembly with PCAP. Curr Protoc Bioinformatics 11: 11.13.

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juárez MJA, Simpson J et al. (2013). Architecture and evolution of a minute plant genome. Nature 498: 94–98.

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

International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463: 763–768.

International Wheat Genome Sequencing Consortium (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345: 1251788.

Jain M (2012). Next-generation sequencing technologies for gene expression profiling in plants. Briefings in Functional Genomics 11: 63–70.

Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD (2007). Extending assembly of short DNA sequences to handle error. Bioinformatics 23: 2942–2944.

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X et al. (2013). *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496: 91–95.

Kaufmann K, Muino JM, Osteras M, Farinelli L, Krajewski P, Angenent GC (2010). Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nat Protocols 5: 457–472.

Kenan-Eichler M, Leshkowitz D, Tal L, Noor E, Melamed-Bessudo C, Feldman M, Levy AA (2011). Wheat hybridization and polyploidization results in deregulation of small RNAs. Genetics 188: 263–272.

Kent WJ (2002). BLAT--The BLAST-like alignment tool. Genome Res 12: 656–664.

Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, Seo E, Choi J, Cheong K, Kim KT et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat Genet 46: 270–278.

Koenig D, Jiménez-Gómez JM, Kimura S, Fulop D, Chitwood DH, Headland LR, Kumar R, Covington MF, Devisetty UK, Tat AV et al. (2013). Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. P Natl Acad Sci USA 110: E2655–E2662.

Krishnan NM, Pattnaik S, Jain P, Gaur P, Choudhary R, Vaidyanathan S, Deepak S, Hariharan AK, Krishna PB, Nair J et al. (2012). A draft of the genome and four transcriptomes of a medicinal and pesticidal angiosperm *Azadirachta indica*. BMC Genomics 13: 464.

Kurtoglu KY KM, Lucas SJ, Budak H (2013). Unique and conserved microRNAs in wheat chromosome 5D revealed by next-generation sequencing. PLoS ONE 8: e69801.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29: 4633–4642.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

Langridge P (2012). Genomics: decoding our daily bread. Nature 491: 678–680.

Leaungthitikanchana S, Fujibe T, Tanaka M, Wang S, Sotta N, Takano J, Fujiwara T (2013). Differential expression of three *BOR1* genes corresponding to different genomes in response to boron conditions in hexaploid wheat (*Triticum aestivum* L.). Plant Cell Physiol 54: 1056–1063.

Lerat E (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 104: 520–533.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li YF, Zheng Y, Jagadeeswaran G, Sunkar R (2013). Characterization of small RNAs and their target genes in wheat seedlings using sequencing-based approaches. Plant Sci 203–204: 17–24.

Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y et al. (2013). Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496: 87–90.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364.

Llaca V (2012). Sequencing technologies and their use in plant biotechnology and breeding. In: Munshi A, editor. DNA Sequencing–Methods And Applications. Rijeka, Croatia: InTech, pp. 35–60.

Marguerat S, Bähler J (2010). RNA-seq: from technology to biology. Cell Mol Life Sci 67: 569–579.

Metzker ML (2009). Sequencing technologies—the next generation. Nat Rev Genet 11: 31–46.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452: 991–996.

Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li LT, Zhang Q, Kim MJ, Schatz MC, Campbell M et al. (2013). Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). Genome Biol 14: R41.

Mullikin JC, Ning Z (2003). The phusion assembler. Genome Res 13: 81–90.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D et al. (2014). The genome of *Eucalyptus grandis*. Nature 510: 356–362.

Myers EW (2005). The fragment assembly string graph. Bioinformatics 21 (Suppl. 2): ii79–85.

Narzisi G, Mishra B (2011). Comparing de novo genome assembly: the long and short of it. PLoS One 6: e19175.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M (2008). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9: 312.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A et al. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature 497: 579–584.

Park PJ (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669–680.

Paszkiewicz K, Studholme DJ (2010). De novo assembly of short sequence reads. Brief Bioinform 11: 457–472.

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W et al. (2008). A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322: 101–104.

Peng Z, Lu Y, Li L, Zhao Q, Feng Q, Gao Z, Lu H, Hu T, Yao N, Liu K et al. (2013). The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). Nat Genet 45: 456–461.

Pevzner PA, Tang H, Waterman MS (2001). An Eulerian path approach to DNA fragment assembly. P Natl Acad Sci USA 98: 9748–9753.

Potato Genome Sequencing Consortium (2011). Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195.

Price AL, Jones NC, Pevzner PA (2005). De novo identification of repeat families in large genomes. Bioinformatics 21 (Suppl. 1): i351–358.

Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T et al. (2012). The cassava genome: current progress, future directions. Trop Plant Biol 5: 88–94.

Rahman AYA, Usharraj A, Misra B, Thottathil G, Jayasekaran K, Feng Y, Hou S, Ong SY, Ng FL, Lee LS et al. (2013). Draft genome sequence of the rubber tree *Hevea brasiliensis*. BMC Genomics 14: 75.

Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N et al. (2010). Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. DNA Res 18: 65–76.

Schatz MC, Delcher AL, Salzberg SL (2010). Assembly of large genomes using second-generation sequencing. Genome Res 20: 1165–1173.

Scheibye-Alsing K, Hoffmann S, Frankel A, Jensen P, Stadler PF, Mang Y, Tommerup N, Gilchrist MJ, Nygard AB, Cirera S et al. (2009). Sequence assembly. Comput Biol Chem 33: 121–136.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115.

Schneeberger K (2014). Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat Rev Genet 15: 662–676.

Shamimuzzaman M, Vodkin L (2013). Genome-wide identification of binding sites for NAC and YABBY transcription factors and co-regulated genes during soybean seedling development by ChIP-Seq and RNA-Seq. BMC Genomics 14: 477.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). Nature Genet 43: 109–116.

Sierro N, Battey JN, Ouadi S, Bovet L, Goepfert S, Bakaher N, Peitsch MC, Ivanov NV (2013). Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. Genome Biol 14: R60.

Sierro N, Battey JND, Ouadi S, Bakaher N, Bovet L, Willig A, Goepfert S, Peitsch MC, Ivanov NV (2014). The tobacco genome sequence and its comparison with those of tomato and potato. Nature Commun 5: 3833.

Simpson JT, Durbin R (2012). Efficient de novo assembly of large genomes using compressed data structures. Genome Res 22: 549–556.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009). ABySS: A parallel assembler for short read sequence data. Genome Res 19: 1117–1123.

Singh R, Ong-Abdullah M, Low ETL, Manaf MAA, Rosli R, Nookiah R, Ooi LCL, Ooi SE, Chan KL, Halim MA et al. (2013). Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature 500: 335–339.

Smaczniak C, Immink RGH, Muiño JM, Blanvillain R, Busscher M, Busscher-Lange J, Dinh QD, Liu S, Westphal AH, Boeren S et al. (2012). Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. P Natl Acad Sci USA 109: 1560–1565.

Staton SE, Bakken BH, Blackman BK, Chapman MA, Kane NC, Tang S, Ungerer MC, Knapp SJ, Rieseberg LH, Burke JM (2012). The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J 72: 142–153.

Strickler SR, Bombarely A, Mueller LA (2012). Designing a transcriptome next-generation sequencing project for a nonmodel plant species. Am J Bot 99: 257–266.

Tang Z, Zhang L, Xu C, Yuan S, Zhang F, Zheng Y, Zhao C (2012). Uncovering small RNA-mediated responses to cold stress in a wheat thermosensitive genic male-sterile line by deep sequencing. Plant Physiol 159: 721–738.

Taudien S, Steuernagel B, Ariyadasa R, Schulte D, Schmutzer T, Groth M, Felder M, Petzold A, Scholz U, Mayer KF et al. (2011). Sequencing of BAC pools by different next generation sequencing platforms and strategies. BMC Res Notes 4: 411.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635–641.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313: 1596–1604.

van Bakel H, Stout J, Cote A, Tallon C, Sharpe A, Hughes T, Page J (2011). The draft genome and transcriptome of *Cannabis sativa*. Genome Biol 12: R102.

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan G, Whaley AM et al. (2012). Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30: 83–89.

Varshney RK, Nayak SN, May GD, Jackson SA (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27: 522–530.

Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B et al. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31: 240–246.

Vaucheret H (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. Gene Dev 20: 759–771.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D et al. (2010). The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet 42: 833–839.

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S et al. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44: 1098–1103.

Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Featherstone AW, Pellicer J, Buggs RJA (2013). Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. Mol Ecol 22: 3098–3111.

Wang S, Wang X, He Q, Liu X, Xu W, Li L, Gao J, Wang F (2012). Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. Plant Cell Rep 31: 1437–1447.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43: 1035–1039.

Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y (2010). De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). BMC Genomics 11: 726.

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R et al. (2012). The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. Plant J 72: 461–473.

Warren RL, Sutton GG, Jones SJ, Holt RA (2007). Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23: 500–501.

Wold B, Myers RM (2008). Sequence census methods for functional genomics. Nat Meth 5: 19–21.

Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol 2: 656–662.

Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, Khan MA, Tao S, Korban SS, Wang H et al. (2013). The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res 23: 396–408.

Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao WB, Hao BH, Lyon MP et al. (2013). The draft genome of sweet orange (*Citrus sinensis*). Nature Genet 45: 59–66.

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J et al. (2011). Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195.

Yanik H, Turktas M, Dundar E, Hernandez P, Dorado G, Unver T (2013). Genome-wide identification of alternate bearing-associated microRNAs (miRNAs) in olive (*Olea europaea* L.). BMC Plant Biol 13: 10.

Yao Y, Sun Q (2012). Exploration of small non coding RNAs in wheat (*Triticum aestivum* L.). Plant Mol Biol 80: 67–73.

Young ND, Debelle F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H et al. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480: 520–524.

Zerbino DR, Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829.

Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W et al. (2012). Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. Nature Biotechnol 30: 549–554.

Zhang J, Chiodini R, Badr A, Zhang G (2011a). The impact of next-generation sequencing on genomics. J Genet Genomics 38: 95–109.

Zhang J, Liu J, Ming R (2014). Genomic analyses of the CAM plant pineapple. J Exp Bot 65: 3395–3404.

Zhang J, Zhang Y, Du Y, Chen S, Tang H (2011b). Dynamic metabonomic responses of tobacco (*Nicotiana tabacum*) plants to salt stress. J Proteome Res 10: 1904–1914.

Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, Tao Y, Wang J, Yuan Z, Fan G et al. (2012). The genome of *Prunus mume*. Nat Commun 3: 1318.