Current Plant Biology 11–12 (2017) 33–40

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

An update on bioinformatics resources for plant genomics research

Mahesh Kumar Basantani[a,*], Divya Gupta[a], Rajesh Mehrotra[b], Sandhya Mehrotra[b], Swati Vaish[a], Anjali Singh[a]

[a] Institute of Bioscience and Technology, Shri Ramswaroop Memorial University, Lucknow-Deva Road, Barabanki, Uttar Pradesh 225003, India
[b] Department of Biological Sciences, Birla Institute of Technology and Science, Vidyavihar Campus, Pilani, Rajasthan 333031, India

ARTICLE INFO

Keywords:
Genomics
Next-generation sequencing
Plant databases
Sequence assembly
Transcriptomics
SNPs

ABSTRACT

Next-generation sequencing and traditional Sanger sequencing methods are of great significance in unraveling the complexity of plant genomes. These are constantly generating heaps of sequence data to be analyzed, annotated and stored. This has created a revolutionary demand for bioinformatics tools and software that can perform these functions. A large number of potentially useful bioinformatics tools and plant genome databases are created that have greatly simplified the analysis and storage of vast amounts of sequence data. The information garnered using the available bioinformatics methods have greatly helped in understanding the plant genome structure. Despite the availability of a good number of such tools, the information pouring from single gene-sequencing, and various whole-genome sequencing projects is overwhelming; thus, further innovations and improved methods are needed to sift through this sequence data, and assemble genomes. The current review focuses on diverse bioinformatics approaches and methods developed to systematically analyze and store plant sequence data. Finally, it outlines the bottlenecks in plant genome analysis, and some possible solutions that could be utilized to overcome the problems associated with plant genome analysis.

1. Introduction

Ever since the publication of *Arabidopsis thaliana* genome sequence, the first plant sequence, in 2000 [1], there has been a deluge of plant genome sequencing projects spawning a vast amount of sequencing data on a regular basis. Both traditional [2] and next-generation sequencing techniques [3] have made significant contributions in plant genome sequencing. The analysis of data generated from these projects would not have been possible without the development of sophisticated bioinformatics approaches. They have greatly simplified the entire process of plant whole-genome sequencing, right from doing a sequencing run to data analysis, to sequence assembly, annotation, storage, and publication of the genome. The data generated from these projects have helped in understanding the architecture, complexity, and dynamic nature of the plant genome [4]. The information gleaned from plant genome sequences has proven useful for generating high-density genetic maps, genome-wide association studies (GWAS), allele mining, genotype-by-sequencing (GBS), better assessment of plant diversity, etc. All these are contributing towards better plant breeding and plant improvement programs.

2. History of DNA sequencing

2.1. Alanine tRNA was the first nucleic acid to be sequenced

The identification of amino acid sequence of insulin [5,6], elucidation of DNA double-helical structure [7] and sequencing of *Escherichia coli* alanine tRNA [8] were perhaps the three most significant developments that laid the foundation for DNA sequencing. Another major development was the use of oligonucleotide primers in DNA sequencing reactions [9]. The discovery of type II restriction enzymes [10,11] was another major milestone on the path to DNA sequencing.

2.2. Sanger's 'plus and minus' method for DNA sequencing used polyacrylamide gels

Sanger introduced the 'plus and minus' method for DNA sequencing in 1975 [12]. This method of DNA sequencing was a critical step leading to the development of modern day sequencing methods. Maxam and Gilbert introduced the chemical sequencing method [13] that was an improvement over Sanger's 1975 technique. Both these methods had their problems and pitfalls and enjoyed only a limited success. The main disadvantage of Maxam–Gilbert method was the use of radioisotopes and highly poisonous chemical for the chemical cleavage.

* Corresponding author.
E-mail address: mkbasantani@gmail.com (M.K. Basantani).

https://doi.org/10.1016/j.cpb.2017.12.002
Received 20 August 2017; Received in revised form 4 December 2017; Accepted 9 December 2017

2.3. Bacteriophage ΦX174 genome was the first genome to be sequenced completely

The Sanger's dideoxy chain termination method [2] was used to sequence bacteriophage ΦX174 genome: This was the first DNA-based genome to be sequenced completely [14]. The ΦX174 genome was resequenced by the dideoxy method in 1978. Soon after the success with ΦX174 genome, complete sequences of simian virus SV40, human mitochondrial genome, phage lambda genome, Epstein-Barr virus, and human CMV genome became available [15]. The sequencing of both short DNA fragments and whole- genomes flourished in leaps and bounds after the introduction of this method. Sanger's sequencing generates biological bias due to the cloning or PCR. It is difficult to analyze allele frequencies and heterozygous SNPs that are not represented as 1:1 ratios.

2.4. Next-generation DNA sequencing

In the past few years, new sequencing technologies have emerged that can generate sequence read lengths much greater than which is possible by Sanger's method. These techniques are mostly employed for whole-genome sequencing, genome resequencing, exome sequencing, ChIP-sequencing, RNA sequencing, epigenome characterization and other similar projects where extensive nucleic acid sequence coverage is the objective. Both, the possibility of assessing a broad range of biological phenomena and a general progress in technology have spurred an interest in, and so also the growth of, these new sequencing methodologies [16]. Over the past several years these high throughput massively-parallel DNA sequencing techniques (a) have become widely available, and (b) have significantly reduced the cost of DNA sequencing. Some of these next generation sequencing (NGS) technologies are 454 sequencing, Illumina sequencing, SOLiD sequencing, ion torrent semiconductor sequencing, DNA nanoball sequencing, HeliScope Single Molecule Real Time (SMRT) sequencing technology, nanopore sequencing etc. A large number of publications utilizing NGS for applications as diverse as whole-genome sequencing, RNA sequencing, analysis of DNA methylation, etc have appeared over the past several years.

2.4.1. Illumina sequencing technology

Illumina is based on the sequencing by synthesis (SBS) chemistry and clonal amplification. It utlilizes a bridge amplification strategy for DNA sequencing [17]. This method relies on identification of nucleotides while they are being incorporated in the growing nucleic acid chain. Various variations or upgraded versions of the Illumina technology have evolved in the past few years, which are MiniSeq series, MiSeq Series, NextSeq Series and NovaSeq Series [18].

2.4.2. Minion nanopore sequencing

Oxford's Nanopore MinION is a very small hand held sequencing device, which is based on nanopores. A membrane containing nanopores is positioned on a detection grid. A change in the ionic current occurs when a DNA molecule to be sequenced passes through nanopores. These changes occur due to the shifting nucleotide that occupies the nanopore space during the movement. The sensors measure this change in current and an algorithm is used to deduce the sequence of the DNA molecule [19].

Biological nanopores formed by *Mycobacterium smegmatis* porin A (MspA) protein have generated lot of interest as a tool for nanopore sequencing [20]. In addition, MspA nanopores have been successfully employed for the detection of unnatural bases, dNaM and d5SICS, in DNA molecules [21].

2.4.3. PacBio sequencing

PacBio Sequencing is a method for real-time sequencing. It does not stop between the reads. It is based on Single molecule real time (SMRT) sequencing. This technology utilizes the DNA replication process and monitors DNA synthesis in real-time. SMRT sequencing is based on zero-mode waveguides (ZMWs) and phospholinked dye labeled nucleotides [22]. The template is created by adding hairpin adaptors to both ends of the target double-strand DNA molecule and called the SMRTbell. The SMRTbell is placed in a SMRT cell. DNA polymerase/template complex is immobilized on the bottom of a well and ZMWs are attached to the DNA polymerase. Immobilized DNA polymerase synthesizes DNA strand, which is imaged in real time with the help of phospholinked dye labeled nucleotides. In the improved Sequel platform based on SMRT technology, SMRT cells includes one million ZMWs as compared to the PacBioRSII, which contains 150,000 ZMWs. With about seven times more ZMWs Sequel System has added scalability as compared to the PacBio® RS II System. The Sequel system can be utilized for producing *de novo* assemblies of whole-genome for large genomes.

The reader is referred to [15,16] for excellent reviews on the history of DNA sequencing and NGS technologies.

3. Computational biology for plant genomics

Motivated by the growth of traditional sequencing and development of NGS technologies, both single gene and whole-genome sequencing projects have become commonplace, which has led to a torrent of nucleic acid sequence information available to the scientific community. The question is how best to analyze and utilize this information.

Rice is the first crop species for which whole genome sequence became available [23,24]. Twelve years have elapsed since then and the list of complete plant genome sequences available is growing ever since [25]. The sequence length varies from the smallest published genome of the carnivorous bladderwort (*Utricularia gibba*) at 82 Mb to Norway Spruce (*Picea abies*) at 19,600 Mb, compared to the second largest of maize at 2300 Mb [25].

This deluge of data has prompted scientists to develop sophisticated computational methods capable of extracting biologically meaningful information from a very large amount of data. Several bioinformatics tools are available that are capable of anatomizing sequence data churned out by sequencers on a regular basis.

3.1. Genome assembly

Genome assembly refers to the reconstruction of the whole genome sequence by aligning and merging sequence reads generated from the current genome sequencing technologies. It is needed (a) because current NGS technologies mostly generate short DNA read lengths (25–400 bp depending on the NGS platform), (b) to handle terabytes of sequencing data, (c) to rectify errors generated during sequencing and (d) to resolve repetitive sequences, which is perhaps the biggest challenge faced by genome assembly methods [26]. A large number of genome assembly computer programs have been written that stitch together entire chromosomes from short fragmented reads of DNA. Genome assembly programs use the data from single and paired reads to assemble a genome [27]. Single reads are continuous sequenced fragments that can be joined up through overlapping regions into 'contigs'. Paired reads are the two ends of the same DNA molecule, which come from sequencing one end of DNA and then sequencing it from the other end. Paired-read data can indicate the size of repetitive regions.

Genome assembly programs use two classes of algorithms: overlap–layout–consensus (OLC) and de-bruijn-graph (DBG) [28]. OLC approach first searches for overlap amongst all the reads, then creates graph layout of all the overlaps and reads, and, finally, generates the consensus sequence. A number of programs such as Arachne, Celera Assembler, CAP3, PCAP, Phrap, Phusion and Newbler employ the OLC approach for genome assembly. DBG utilizes short reads to assemble genomes. It works by breaking down sequence reads into shorter *k*-

mers, and use these *k*-mers to assemble the genome [29]. Many short-read genome assemblers like Euler-USR, Velvet, ABySS, AllPath-LG and SOAPdenovo have been developed based on DBG. OLC algorithm is more applicable to low-coverage long reads, whereas DBG is more suitable for high-coverage short reads. However, the assembly results vary among genomes sequenced, and sequencing technologies used.

3.1.1. Plant genomes are hard to assemble

Plant genomes pose some unique challenges for genome assembly [30]. They are difficult to assemble because of (a) large and complex genomes (plant genomes can be 100 times larger than many mammalian genomes sequenced), (b) high ploidy, (c) high heterozygosity, and (d) the presence of a large number of repeat sequences. Generally, the assemblers used for mammalian genome assembly have been employed successfully for plant genomes as well. Some of these genome assemblers are TIGR assembler [31], CAP3 [32], string graph assembly [33], Newbler [34], SSAKE [35], VCAKE [36], SHARCGS [37], ALLPATHS-LG [38], Edena [39], Velvet [40], CABOG [41], ABySS [42], SOAPdenovo2 [43], Genomic Analysis Toolkit [44], etc.

3.1.1.1. CAP3 is the third generation of contig assembly program. CAP3 program was created to meet the challenges associated with the assembly of large genomes like mouse, humans and maize [32]. It is the third generation of contig assembly program (CAP).

CAP3 shows significant improvements over its predecessors. It has the capability to clip 5′ and 3′ low-quality regions of reads. It uses more efficient algorithms to identify and compute overlaps between reads. It employs forward-reverse constraints to correct assembly errors and link contigs. CAP3 assembly algorithm consists of 3 major phases: In the first phase, 5′ and 3′ low-quality regions of reads, and false overlaps are identified and removed; in the second phase, reads are joined to form contigs. Finally, the multiple sequence alignment of reads is constructed and a consensus is computed for each contig. PlantGDB, a database of plant genomes and plant EST (expressed sequence tags) sequences, employs CAP3, besides Vmatch and PaCE, for EST assembly [45]. CAP3 has been successfully used for EST assembly of several plant species including *Arabidopsis thaliana*, rice, wheat, *Ricinus communis*, maize, *Glycine max*, *Medicago truncatula*, *Brassica* sps., etc [46,47]. CAP3 along with MIRA was used to assemble the 454 reads in a study to compare the Compositae crops and their wild relative; greater divergence was found between the self-incompatible progenitors. Postzygotic isolation between pairs of taxa also led to greater divergence [48].

3.1.1.2. Newbler assembles sequence data generated by 454 sequencing platform. Newbler is a software package for *de novo* assembly of DNA sequence data generated by the Roche 454 pyrosequencing platform [49]. It runs via Java GUI (graphical user interface) or the command line, and works with the .SFF data output by the sequencer; however, it can also accept FASTA files of nucleotide sequences. It has been used to assemble pyrosequencing ESTs of *Arabidopsis thaliana* [50], *Eucalyptus grandis* [51], *Castanea dentata* and *C. mollissima* [52].

3.1.1.3. SSAKE is the first published short-read sequence assembly algorithm. Short Sequence Assembly by progressive K-mer search and 3′ read Extension (SSAKE) is a tool for assembling millions of short nucleotide sequences. It is the first published algorithm for genome assembly with short DNA sequences. It runs on Linux and is written in PERL. It progressively searches for the longest possible overlap between any two sequences and assembles them into contigs. VCAKE (Verified Consensus Assembly by K-mer Extension) [36], an extension of SSAKE, has been developed to efficiently handle sequencing errors. Another assembler built on the methods employed for SSAKE and VCAKE is SHARCGS (SHort-read Assembler based on Robust Contig extension for Genome Sequencing). It is also capable of assembling millions of very short reads, and copes well with sequencing errors. All three of these assemblers follow a prefix tree-based approach introduced with SSAKE.

3.1.1.4. ALLPATHS assembles genomes from microreads. ALLPATHS has been developed to efficiently assemble genomes from short-reads generated from platforms like Illumina-Solexa and ABI-SOLiD. The software can perform *de novo* assembly from the reads of size 25–50 bases. It follows the de Bruijn graph representation for genome assembly. Despite the efficacy and widespread applicability of ALLPATHS to short read data, it can be applied for large sequence reads as well [38]. It has been used successfully for *de novo* assembly of *Arabidopsis thaliana*. [53]. ALLPATHS-LG was used for the assembly of Illumina and 454 reads generated from *M. truncatula* genome sequencing. *M. truncatula* genome (version Mt4.0) has 50,894 genes which is about 82% similar for the annotated genes of the version Mt3.5 [54].

3.1.1.5. ABySS is a parallel assembler for short reads that can efficiently identify polymorphic sequences. Massively parallel sequencing platforms like Illumina, Inc. Genome Analyzer, Applied Biosystems SOLiD System, 454 Life Sciences (Roche) GS FLX, etc produce high-quality short reads from 25 to 500 bp in length. Although this read length is much shorter than the traditional capillary-based sequencing technology, but the total number of bases sequenced in a given run is orders of magnitude higher. The problem of *de novo* short read assembly has been handled successfully by widespread application of de Bruijn graphs [55,56] in short read assemblers like Velvet [40], ALLPATHS [38], and EULER-SR [57], etc.

ABySS (Assembly By Short Sequencing) has been created to assemble large data sets produced by sequencing of very large genomes. It has been used successfully to identify insertions and deletions and novel sequences in the human genome. It has been used to quickly and accurately assemble 3.5 billion short sequence reads generated from whole-genome sequencing of a Yoruban male on the Illumina platform [42]. It is particularly useful for the assembly of those genomes for which reference sequence is not available. It utilizes the distributed representation of a de Bruijn graph: This allows parallel computation of the assembly algorithm across a cluster of computers. It performs the assembly in two steps. First, contigs are extended without the paired-end information; in the second-step paired-end information is utilized to merge contigs. It is implemented in C++.

ABySS was used for the assembly of the whole-genome shotgun sequencing data of the mitochondrial and plastid genomes of white spruce (*Picea glauca*). The mitochondrial (5.9-Mb) and plastid (123-kb) genome reads were more abundant than the nuclear genome reads in the sequencing data. Coding genes and RNAs (rRNA and tRNA) were annotated from the sequence data. This approach has been found to be helpful for the assembly of organellar genomes of different plant species [58]. Mitochondrial genomic diversity exists between the upland cotton *Gossypium hirsutum* and Sea Island cotton *Gossypium barbadense* L. Mitochondrial genome sequencing in these cotton species was accomplished with Solexa using paired-end, 90 bp read. The sequencing reads were assembled using ABySS. It was found that *G. barbadense* mitochondrial genome has a total of 40 protein-coding genes, 6 rRNA genes, and 29 tRNA genes [59]. The study highlights the applications of ABySS for organellar genome sequence assembly in plant species.

3.1.1.6. Velvet is a set of algorithms for genomic sequence assembly. Velvet is set of algorithms designed to perform short read assembly by manipulating de Bruijn graph to eliminate errors and resolve repeats [40]. These two tasks are done separately.

Velvet implements “Tour Bus” and “Breadcrumb” algorithms for removing bubbles (created by sequencing errors or SNPs), and resolution of repeats, respectively. Velvet along with ABySS and CLCBio was used to assemble 95% complete genome of *Ostreococcus tauri*, a marine alga and the smallest free-living photosynthetic eukaryote. This assembly filled 930 gaps, which were left in the original genome assembly [60].

3.1.1.7. CABOG is a modification of celera assembler. The sequencing data generated from two different platforms could be of significant use for a competent *de novo* genome assembly. The two read types could complement each other in a "hybrid" assembly approach. However, the reads generated from two diverse platforms have inherent characteristics specific to each platform. The assembly software designed to handle data from varied platforms should be designed such that they can accommodate different characteristics of each platform. CABOG (Celera Assembler with the Best Overlap Graph) is one such assembler that has been designed to handle data from two different sequencing platforms [41]. It is capable of performing *de novo* genome assembly from pyrosequencing and Sanger sequencing reads. It is found to be more efficient than Newbler for the resolution of repeat sequences. Both Velvet and CABOG assemblers were employed for *de novo* assembly of the wild strawberry (*Fragraria vesca*) genome using the sequencing data generated from 454, Illumina and ABI SOLiD platforms [61].

3.1.1.8. Canu assembles long-read single-molecule sequencing data. Though besotted with high error rates single-molecule sequencing has revolutionized genome sequencing, generating read lengths up to 10 kbp. PacBio single-molecule real-time (SMRT) sequencing [19] and Oxford Nanopore Sequencing [22] have been successfully employed to resolve the problem of high error rates associated with long-read single-molecule sequencing. The assembly of noisy long-read sequence data requires specialized computational strategy. Canu has ben developed to address the challenges associated with assembling long-read single-molecule sequencing datasets [62]. It is a successor of Celera Assembler. It is the most efficient single-molecule read assembler for large genomes, and requires a low runtime. It can generate genome assemblies from both PacBio and Nanopore sequencing platforms. It is particularly efficient at assembling large repeat sequences, a hallmark of plant genomes. Canu performs the sequence analysis in three stages—correction, trimming, and assembly.

Canu has been used successfully for the *de novo* assembly of *Solanum pennellii* genome sequence using the data generated from Nanopore Sequencing [63]. It has also been used for the *de novo* assembly of an indica rice genome Shuhui498 (R498) using the data generated from PacBio SMRT sequencing platform [64].

3.1.1.9. Miniasm also assembles long-read sequencing data. Miniasm also tackles the challenges and issues associated with long-reads [65]. Miniasm, like Canu, can generate assemblies from both PacBio and Nanopore sequencing reads. It is based on OLC approach for assembly, performing only the O and L steps. It is faster than many existing assemblers, and assembles genomes without an error correction stage.

3.1.1.10. FALCON assembles noninbred genomes. One of the major problems of whole genome sequencing is the inability to capture heterozygosity, thereby resulting in the loss of variation between homologous chromosomes from the sequencing data. FALCON assembler has been developed to resolve this problem [66]. The utility of FALCON has been assessed on *Arabidopsis thaliana* and *Vitis vinifera* cv. Cabernet Sauvignon.

A number of studies have been undertaken to compare the efficiency and applicability of different genome assemblers. One such study compared eight short reads assemblers, against two types of simulated Solexa short reads datasets derived from four different genomes [67]. The assemblers were tested for computational time and memory cost, assembly accuracy, completeness and size distribution of assembled contigs. The study proposed that for small genomes (microorganisms) SSAKE, QSRA, and Edena assemblers perform better with very short reads (36 bp) when the RAM size is less than 16GB. However, with a RAM size of more than 16 GB Taipan assembler gives good results with both very short (36 bp) and short (75 bp) reads. For large genomes (eukaryotes), with a RAM size of more than 16GB, SOAPdenovo assembler gives better results with very short (36 bp) reads, and ALL-PATHS-LG gives better results with short (75 bp) reads. An exhaustive comparative assessment of assemblers to address the quality and accuracy issues associated with whole-genome sequence assembly problem was performed by [68]. The study compared the sequence assemblers both for low-coverage long reads and high-coverage short reads. Feature-Response curve (FRC), an improved comprehensive metric, was introduced in the study. This metric captures the trade-offs between contigs' quality against their sizes. It does not require reference genome sequence for validation, and, therefore, makes it very useful for *de novo* sequencing projects. Different genome assemblers possess different properties in relation to accuracy and completeness of genome assembly. A comparison of different genome assemblers on pyrosequencing data was carried out by [69]. It was found that CABOG produced more complete and continuous assemblies with fewer and larger contigs and scaffolds, with some regions reconstructed incorrectly. Newbler, on the other hand, generated assemblies with very few errors. However, the assemblies were more incomplete and discontinuous.

3.2. Transcriptome assembly

NGS technologies have proved instrumental in understanding various facets of plant genome architecture and complexity. These technologies are now becoming important for plant gene expression analysis, but with a limited use [70]. These technologies are a method of choice for gene expression analysis because of their high sensitivity and throughput. High-throughput mRNA sequencing (RNA-Seq) has proved instrumental in transcript discovery and abundance estimation [71] and is quickly superseding the microarray-based and EST sequencing approaches for studying gene expression. Gene expression profiling using RNA-seq assumes that the depth of coverage of a sequence is proportional to the expression of corresponding gene of interest [70].

Transcript assembly is much more complex than genome assembly. Several factors such as varying gene expression levels, splice variants, exon-intron boundaries and multiple reading frames are responsible for this. *De novo* assembly (assembly of transcripts where no genome sequence exists) or reference-based (alignment with a reference genome) approaches are utilized for transcriptome assembly. Several programs are available for transcriptome assembly implementing different methods and algorithms.

3.2.1. QPALMA is a reference-based transcriptome assembler

QPALMA is implemented in C++ and Python [72]. It aligns short reads to genomic sequences. It utilizes the Smith–Waterman algorithm for alignment of transcripts to the reference genome. It is capable of computing accurate spliced alignments over exon boundaries. It has been used for *Arabidopsis thaliana* transcriptome assembly using the data generated from Illumina Genome analyzer.

3.2.2. TopHat aligns RNA sequencing reads to a reference genome

TopHat is a free open-source software package that identifies splice sites *ab initio* by large-scale mapping of high-throughput RNA-seq reads [73]. It allows identification of new genes and splice variants, and comparison of gene and transcript expression under different conditions. It aligns RNA-Seq reads to large genomes using the high-throughput short read aligner Bowtie. TopHat is implemented in C^{++} and runs on both Linux and Mac OS X.

3.2.3. Trinity is a de novo transcriptome assembler

Large-scale cDNA sequencing provides ample opportunities to understand genome complexities at the transcriptional level. However, precise transcript reconstruction or assembly depends upon the availability of a reference genome. Trinity assembler overcomes this issue [74]. It is a highly efficient *de novo* transcriptome assembler, which is capable of transcript assembly without a reference genome. It is capable

of assembling large transcriptomes rich in alternatively spliced isoforms, and/or duplicated genes.

Trinity is developed at the Broad Institute and the Hebrew University of Jerusalem. It combines three independent software modules, Inchworm, Chrysalis, and Butterfly, to process large volumes of RNA-seq reads, and efficiently recovers more full-length transcripts across different levels of spatial and temporal expression patterns.

3.2.4. Oases: an improved de novo transcriptome assembler

Oases software package is another *de novo* transcriptome assembler that accounts for alternative splicing and varied expression levels [75]. It receives as input a preliminary assembly produced by the Velvet assembler. As compared to its predecessors Oases performs better at handling incomplete or uneven coverage and alternative splicing events. *Plantago ovata* seed husk is used in the pharmaceutical, food and cosmetic industry. Illumina Genome Analyzer platform was used to decipher the mucilage biosynthesis pathway transcriptome of *P. ovata* ovary. Data was assembled using Oases followed by velvet. There were 46,955 non-redundant transcripts ($\geq$100 bp), which were involved in the different metabolic pathways for mucilage biosynthesis. Along with these transcripts there were several non-coding RNAs, sequence-repeat motifs, and transcription factors present in the dataset [76].

Comparison of transcriptome assemblers has been performed to assess the quality and entirety of assembled transcripts. Such studies may be helpful in choosing the most appropriate assembler to achieve superior quality transcriptome assembly. A comparative analysis of four *de novo* transcriptome assemblers, TransAbyss, Trinity, SOAPdenovo-Trans, and Oases was performed on *Nicotiana benthamiana* transcript sequencing carried out on Illumina HiSeq2000 instrument [77]. *N. benthamiana*, and allo-tetraploid plant, poses specific challenges for *de novo* transcriptome assembly because of the presence of large number of homeologous and duplicated gene copies. The study demonstrated that some transcripts were more completely assembled with TransAbyss as compared to Trinity, and vice-versa. It was found that overall TransAbyss generated the highest number of transcripts assembled to more than 80% target sequence length, followed by Oases, Trinity and SOAPdenovo-Trans.

3.3. Molecular markers

Molecular markers have become an integral component of plant genome analysis and plant improvement programs. Marker assisted selection (MAS) has been employed in many crop improvement programs [78]. Molecular markers are used for phylogeny and evolution studies, analysis of exotic germplasm diversity, cultivar genotyping, etc. Molecular markers have taken the centre stage in plant improvement programs since NGS and traditional DNA sequencing have become routine.

3.3.1. Pyrobayes is a modification of PolyBayes

PolyBayes is a computer program for the automated analysis of single-nucleotide polymorphism (SNP). It is built on Bayesian inference engine [79]; it calculates the probability that differences at a given location of a multiple alignment represent actual sequence variations and not sequencing errors. It is developed in a Unix environment. The output from the PolyBayes program consists of a list of candidate SNPs, each with an SNP probability score.

PyroBayes is a modification of PolyBayes [80]. It is designed to read pyrosequencing data generated from 454 sequencing technology platforms. It is more accurate at SNP calling than PolyBayes.

3.3.2. GS reference mapper aligns reads to the reference genome

GS Reference Mapper is a part of the GS Data Analysis Software package provided with the GS Junior and GS FLX System sequencing platforms from 454 Life Sciences. It accurately aligns reads to any reference genome and identifies differences compared to the reference. The program is extremely useful for identifying genome variations including SNPs, and insertions and deletions. It has been used for SNP detection in *Eucalyptus grandis* [51].

3.3.3. AutoSNP and AutoSNPdb

AutoSNP is a program to detect SNPs and insertion/deletion polymorphisms (indels) in EST data. The program is particularly good at handling sequencing errors, particularly those associated with the process of reverse transcription. The program is written in PERL. The program was used successfully to identify a total of 14832 candidate polymorphisms in maize EST sequence data [81]. AutoSNPdb combines SNP discovery software and sequence annotation in a relational database for the efficient identification of SNP and indel polymorphisms related to specific genes or traits [82]. The program has been employed for SNP identification in barley, rice and *Brassica* EST sequences.

The reader is referred to [83] for an excellent review on SNP bioinformatics resources.

3.3.4. VarScan detects variants in a gene

VarScan is an open-source software tool that detects variants of a gene in NGS data. This is platform-independent mutation caller software that can be utilized for the detection of somatic mutations, somatic copy number alterations and germline variants, SNPs, insertion/deletions with their chromosomal locations and read counts. It analyzes variants in individual samples as well as pooled samples. It detects low (1%) frequency variants in the pooled samples. It discards the reads that are present at multiple locations and shows low identity upon alignment in the reference sequence. The best alignment for each read is screened for sequence changes. Multiple reads are used for the variant detection and then the data from these are then joined into unique SNPs and insertion/deletions. To predict variants VarScan identifies overall coverage, total number of supporting reads and strands observed for each allele and average base quality [84]. In a study for the detection of the somatic mutations in tumor samples to resolve subclones. VarScan2 proved better as compared to the other software available [85].

3.3.5. Genomic analysis toolkit (GATK)

It is a collection of tools that includes SNP/insertion deletion caller, quality score recalibrator and local realigner [44]. GATK suite was primarily developed for the processing of exome and whole genome data. This was primarily developed for identifying SNP/INDEL in human genome sequencing projects, but now it can be used for the analysis for genome data of several different organisms.

GATK is based on MapReduce, and is provided as a Java framework. It can read sequence data from any sequencing platform, and has been tested on Illumina, ABI SOLiD System, etc. GATK is organized into traversals, which provide the division and preparation of data, and walkers, which are analysis modules.

The URLs of various sequence assemblers, tools and resources mentioned in this review are given in Table 1.

3.4. Plant genome databases

With the increasing availability and augmented computational capacities to analyze sequencing data, the demand for dedicated databases to store this data is ever growing. This has led to the formation of dedicated plant genome repositories. These repositories are not mere warehouses that store the sequence data, but have huge computational capabilities that can anatomize this data and extract relevant information. Both general and species-specific plant databases are available that are valuable for plant sequencing data storage and analysis.

A list of some representative plant databases is given in Table 2.

4. Discussion

NGS technologies and the sequence data analysis have transformed

Table 1
URLs of some important tools and resources for plant genome analysis.

	Bioinformatics tool	URL	Reference
	Assemblers		
1	ABySS	http://www.bcgsc.ca/platform/bioinfo/software/abyss	Simpson et al. [42]
2	Velvet	http://www.ebi.ac.uk/~zerbino/velvet	Zerbino and Birney [40]
3	CABOG	http://wgs-assembler.sf.net/	Miller et al. [41]
4	Canu	https://github.com/marbl/canu	Koren et al. [62]
5	Miniasm	https://github.com/lh3/miniasm	Li [65]
6	FALCON	https://github.com/PacificBiosciences/FALCON/	Chin et al. [66]
7	GRIDSS	https://github.com/PapenfussLab/gridss	Cameron et al. [108]
8	QPALMA	http://www.fml.mpg.de/raetsch/projects/qpalma	De Bona et al. [72]
9	TopHat	http://ccb.jhu.edu/software/tophat/index.shtml	Trapnell et al. [73]
10	Trinity	http://trinityrnaseq.sourceforge.net/	Grabherr [74]
11	Oases	http://www.ebi.ac.uk/~zerbino/oases/	Schulz et al. [75]
12	MaSuRCA	http://www.genome.umd.edu/masurca.html	Zimin et al. [102]
	Cloud computing		
13	Contrail	http://sourceforge.net/projects/contrail-bio/	Kumari et al. [110]
14	DDBJ Read Annotation Pipeline	https://p.ddbj.nig.ac.jp/pipeline/Login.do	Nagasaki et al. [111]

the field of plant genomics in a radical way in the past few years. Single-gene sequences, and whole-genome sequences of both model and non-model plant species have burst upon the plant genomics data landscape, revolutionizing the way plant genomes are analyzed and understood. At least 183 plant reference sequences have been published till date [97]. Bioinformatics has provided the necessary framework to systematically analyze and store this data. The capability and efficiency to handle this vast amount of data have increased several folds. Despite substantial advancements made and improved software and algorithms developed for sequence assembly, data storage and visualization, detection of sequence variations, etc, there is no dearth of challenges posed by plant genomes. Although we have come a long way since the first sequence assembly algorithms were developed [98], yet repetitive sequences and genome size are still the major issues that need to be addressed. Assemblathon1, Genome Assembly Gold-Standard Evaluation (GAGE) and Assemblathon2 were undertaken to evaluate the status of currently available assembly methods [99–101]. These efforts concluded that the quality of data generated has a bearing on the quality of assembled genome, and there are variations in the degree of contiguity, and correctness of an assembly. However, the genomes analyzed in these studies were simple and small as compared to large and complex plant genomes. It is therefore important to perform such large-scale studies on plant genomes, hallmarks of which are large size, higher ploidy and heterozygosity. MaSuRCA (Maryland Super-Read Celera Assembler) is a recently developed genome assembler that performs better than its predecessors Allpaths-LG and SOAPdenovo2 [102]. It was successfully employed to assemble the massive, 22-Gb, loblolly pine genome [103–105]. Another significant milestone is the development of JR-Assembler that reduces the computer memory and time required to assemble large genomes [106]. In order to improve the accuracy of genome assembly, and handle large and complex genomes, assembly reconciliation tools have also been created. They are specialized algorithms that produce a higher quality assembly by merging two or more draft assemblies [107]. Identification of genomic rearrangements and structural variants (SVs) in high-throughput sequencing data still remains a major challenge. GRIDSS (Genome Rearrangement IDentification Software Suite) assembler has been recently developed to resolve this problem [108]. GRIDSS is based on de Bruijn graph-based assembly approach. Even with the significant improvements in genome assembly approaches, only a few plant genome assemblies are on the chromosome-level; most of the unassembled or incompletely assembled genome sequences consist of fragmented contigs and scaffolds, with no chromosomal locations assigned to them [97]. Thus from the bioinformatics perspective, the major questions that remain unanswered as yet are: a) which will be the best assembly algorithm, b) what are the parameters to be considered when running an assembly software, and c) how to tell if the assembly generated is good? Another limiting factor is the computational power required to handle plant genomes. The cost of DNA sequencing has come down significantly in the recent years [109]; however, the computational might required to perform heavy-duty calculations has not matched up to the number of bases that can be sequenced per unit cost. Cloud computing comes to the rescue here: multiple-computer server farms with terabytes of shared storage can be employed to analyze sequence runs generated on NGS platforms. This could a) be particularly useful for species with large genomes, b) significantly reduce the time required to complete the assembly and c) provide cost effective genome assembly pipeline for smaller academic research groups or laboratories that lack the infrastructure to build their own computer clusters. Recently, Contrail, an assembler based on cloud computing was developed and tested on zebrafish [110]. In an initiative taken by DDBJ, a cloud computing based annotation service, DDBJ Read Annotation Pipeline (DDBJ Pipeline), has been established [111]. It performs high-throughput annotation of NGS reads.

The strides and innovations made towards better sequence analysis, whether it is the development of faster and more accurate assemblers, or the sequencing and assembly of the loblolly pine genome, the largest ever genome sequenced, or development of cloud computing, or creation of databases, all have generated a wealth of meaningful

Table 2
List of some representative plant databases.

	Databases		
1	PlantGDB	http://www.plantgdb.org/	Dong et al. [45]
2	Wheatgenome.info	http://www.wheatgenome.info/	Lai et al. [86]
3	Phytozome	http://www.phytozome.net	Goodstein et al. [87]; Zhang et al. [88]
4	EuroPineDB	http://www.scbi.uma.es/pindb/	Fernández-Pozo et al. [89]
5	TreeGenes	http://dendrome.ucdavis.edu/treegenes/	Wegrzyn et al. [90]
6	Gramene	http://www.gramene.org/	Ware et al. [91]; Tello-Ruiz et al. [92]
7	Sol Genomics	http://github.com/solgenomics/; http://solgenomics.net/	Bombarely et al. [93]; Fernandez-Pozo et al. [94]
8	Legume Information System	http://legumeinfo.org	Gonzales et al. [96]; Dash et al. [95]

information. The knowledge gained from these efforts could be utilized to further understand plant genome design and intricacy, ultimately leading to a) the determination of functions of genes involved in environmental stress resistance, b) the generation of plants with better fruit quality, c) the superior use of genetic diversity that could help produce better plants and crops for the future.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interests statement

The authors have no conflict of interest.

References

[1] The Arabidopsis Genome Initiative Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, Nature 408 (2000) 796–815.
[2] F. Sanger, S. Nicklen, A.R. Coulson, DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. U. S. A. 74 (1977) 5463–5467.
[3] M. Thudi, Y. Li, S.A. Jackson, G.D. May, R.K. Varshney, Current state-of-art of sequencing technologies for plant genomics research, Brief. Funct. Genom. 11 (2012) 3–11.
[4] R. Wóycicki, J. Witkowicz, P. Gawroński, J. Dąbrowska, A. Lomsadze, et al., The genome sequence of the North-European cucumber (*Cucumis sativus* L.) unravels evolutionary adaptation mechanisms in plants, PLoS One 6 (2011) e22728.
[5] F. Sanger, The terminal peptides of insulin, Biochemical J. 45 (1949) 563–574.
[6] F. Sanger, H. Tuppy, The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates, Biochem. J. 49 (1951) 481–490.
[7] J.D. Watson, F.H.C. Crick, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid, Nature 171 (1953) 737–738.
[8] R.W. Holley, J. Apgar, G.A. Everett, J.T. Madison, M. Marquisee, S.H. Merrill, J.R. Penswick, A. Zamir, Structure of a ribonucleic acid, Science 147 (1965) 1462–1465.
[9] R. Padmanabhan, R. Wu, Nucleotide sequence analysis of DNA. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis, Biochem. Biophys. Res. Commun. 48 (1972) 1295–1302.
[10] T.J. Kelly Jr, H.O. Smith, A restriction enzyme from Hemophilus influenzae. II, J. Mol. Biol. 51 (1970) 393–409.
[11] H.O. Smith, K.W. Wilcox, A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties, J. Mol. Biol. 51 (1970) 379–391.
[12] F. Sanger, A.R. Coulson, A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, J. Mol. Biol. 94 (1975) 441–448.
[13] A.M. Maxam, W. Gilbert, A new method for sequencing DNA, Proc. Natl. Acad. Sci. U. S. A. 74 (1977) 560–564.
[14] F. Sanger, G.M. Air, B.G. Barrell, N.L. Brown, A.R. Coulson, J.C. Fiddes, C.A. Hutchison, P.M. Slocombe, M. Smith, Nucleotide sequence of bacteriophage φX174 DNA, Nature 265 (1977) 687–695.
[15] C.A. Hutchison III, DNA sequencing: bench to bedside and beyond, Nucl. Acids Res. 35 (2007) 6227–6237.
[16] J. Shendure, H. Ji, Next-generation DNA sequencing, Nat. Biotechnol. 26 (2008) 1135–1145.
[17] J. Shin, G. Ming, H. Song, Decoding neural transcriptomes and epigenomes via high-throughput sequencing, Nat. Neurosci. 17 (2014) 1463–1475.
[18] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, Y. Gu, A tale of three next generation sequencing platforms: comparison of Ion Torrent Pacific Biosciences and Illumina MiSeq sequencers, BMC Genom. 13 (2012) 341.
[19] H. Lu, F. Giordano, Z. Ning, Oxford nanopore MinION sequencing and genome assembly, genomics, Proteom. Bioinform. 14 (2016) 265–279.
[20] A.H. Laszlo, I.M. Derrington, J.H. Gundlach, MspA nanopore as a single-molecule tool: from sequencing to SPRNT, Methods 105 (2016) 75–89.
[21] J.M. Craig, A.H. Laszlo, I.M. Derrington, B.C. Ross, H. Brinkerhoff, I.C. Nova, et al., Direct detection of unnatural DNA nucleotides dNaM and d5SICS using the MspA nanopore, PLoS One 10 (2015) e0143253.
[22] A. Rhoads, K.F. Au, PacBio sequencing and its applications, Genom. Proteom. Bioinform. 13 (2015) 278–289.
[23] J. Yu, S. Hu, J. Wang, et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*), Science 296 (2002) 79–92.
[24] S.A. Goff, D. Ricke, T.H. Lan, et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*), Science 296 (2002) 92–100.
[25] T.P. Michael, S. Jackson, The first 50 plant genomes, Plant Genome 6 (2013) 1–7.
[26] M.C. Schatz, A.L. Delcher, S.L. Salzberg, Assembly of large genomes using second-generation sequencing, Genome Res. 20 (2010) 1165–1173.
[27] M. Baker, De novo genome assembly: what every biologist should know, Nat. Methods 9 (2012) 333–337.
[28] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, W. Fan, Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph, Brief. Funct. Genom. 11 (2012) 25–37.
[29] R.M. Idury, M.S. Waterman, A new algorithm for DNA sequence assembly, J. Computational. Biol. 2 (1995) 291–306.
[30] M.C. Schatz, J. Witkowski, R.W. Mccombie, Current challenges in *de novo* plant genome sequencing and assembly, Genome Biol. 13 (2012) 243.
[31] G.G. Sutton, O. White, M.D. Adams, A.R. Kerlavage, T.I.G.R. Assembler, A new tool for assembling large shotgun sequencing projects, Genome Sci. Technol. 1 (1995) 9–19.
[32] X. Huang, A. Madan, CAP3: a DNA sequence assembly program, Genome Res. 9 (1999) 868–877.
[33] E.W. Myers, The fragment assembly string graph, Bioinformatics 21 (2005) ii79–ii85.
[34] M. Margulies, M. Egholm, W.E. Altman, et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.
[35] R.L. Warren, G.G. Sutton, S.J.M. Jones, R.A. Holt, Assembling millions of short DNA sequences using SSAKE, Bioinformatics 23 (2007) 500–501.
[36] W.R. Jeck, J.A. Reinhardt, D.A. Baltrus, M.T. Hickenbotham, V. Magrini, E.R. Mardis, J.L. Dangl, C.D. Jones, Extending assembly of short DNA sequences to handle error, Bioinformatics 23 (2007) 2942–2944.
[37] J.C. Dohm, C. Lottaz, T. Borodina, H. Himmelbauer, SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing, Genome Res. 17 (2007) 1697–1706.
[38] J. Butler, I. Maccallum, M. Kleber, I.A. Shlyakhter, M.K. Belmonte, E.S. Lander, C. Nusbaum, D.B. Jaffe, ALLPATHS: De novo assembly of whole genome shotgun microreads, Genome Res. 18 (2008) 810–820.
[39] D. Hernandez, P. Francois, L. Farinelli, M. Osteras, J. Schrenzel, De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer, Genome Res. 18 (2008) 802–809.
[40] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829.
[41] J.R. Miller, A.L. Delcher, S. Koren, E. Venter, B.P. Walenz, A. Brownley, J. Johnson, L. Li, C. Mobarry, C. Sutton, Aggressive assembly of pyrosequencing reads with mates, Bioinformatics 24 (2008) 2818–2824.
[42] J.T. Simpson, K. Wong, S.D. Jackman, J.E. Schein, S.J. Jones, I. Birol, ABySS: a parallel assembler for short read sequence data, Genome Res. 19 (2009) 1117–1123.
[43] R. Luo, B. Liu, Y. Xie, et al., SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler, GigaScience 1 (2012) 18.
[44] A. Mckenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. Depristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.
[45] Q. Dong, S.D. Schlueter, V. Brendel, PlantGDB, plant genome database and analysis tools, Nucl. Acids Res. 32 (2004) D354–D359.
[46] J. Duan, C. Xia, G. Zhao, J. Jia, X. Kong, Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data, BMC Genom. 13 (2012) 392.
[47] Y. Yang, S. Smith, Optimizing de novo assembly of short-read RNA-seq data for phylogenomics, BMC Genom. 14 (2013) 328.
[48] K.A. Hodgins, Z. Lai, L.O. Oliveira, D.W. Still, M. Scascitelli, M.S. Barker, N.C. Kane, H. Dempewolf, A. Kozik, R.V. Kesseli, J.M. Burke, R.W. Michelmore, L.H. Rieseberg, Genomics of Compositae crops: reference transcriptome assemblies and evidence of hybridization with wild relatives, Mol. Ecol. Resour. 14 (2014) 166–177.
[49] S. Kumar, M.L. Blaxter, Comparing de novo assemblers for 454 transcriptome data, BMC Genom. 11 (2010) 571.
[50] A.P.M. Weber, K.L. Weber, K. Carr, C. Wilkerson, J.B. Ohlrogge, Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing, Plant Physiol. 144 (2007) 32–42.
[51] E. Novaes, D.R. Drost, W.G. Farmerie, G.J. Pappas Jr., D. Grattapaglia, R.R. Sederoff, M. Kirst, High throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome, BMC Genom. 9 (2008) 312.
[52] A. Barakat, D.S. Diloreto, Y. Zhang, C. Smith, K. Baier, W.A. Powell, N. Wheeler, R. Sederoff, J.E. Carlson, Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection, BMC Plant Biol. 9 (2009) 51.
[53] K. Schneeberger, S. Ossowski, F. Ott, J.D. Klein, X. Wang, C. Lanz, L.M. Smith, J. Cao, J. Fitz, N. Warthmann, S.R. Henz, D.H. Huson, D. Weigel, Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 10249–10254.
[54] H. Tang, V. Krishnakumar, S. Bidwell, B. Rosen, A. Chan, S. Zhou, L. Gentzbittel, K.L. Childs, M. Yandell, H. Gundlach, K.F. Mayer, D.C. Schwartz, C.D. Town, An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*, BMC Genom. 15 (2014) 312.
[55] P.A. Pevzner, H. Tang, Fragment assembly with double-barreled data, Bioinformatics 17 (2001) S225–S233.
[56] P.A. Pevzner, H. Tang, M.S. Waterman, An Eulerian path approach to DNA fragment assembly, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 9748–9753.
[57] M.J. Chaisson, P.A. Pevzner, Short read fragment assembly of bacterial genomes, Genome Res. 18 (2008) 324–330.
[58] S.D. Jackman, R.L. Warren, E.A. Gibb, et al., Organellar genomes of white spruce (*Picea glauca*): assembly and annotation, Genome Biol. Evol. 8 (2015) 29–41.
[59] M. Tang, Z. Chen, C.E. Grover, Y. Wang, et al., Rapid evolutionary divergence of Gossypium barbadense and G. hirsutum mitochondrial genomes, BMC Genom. 16

(2015) 770.
[60] R. Blanc-mathieu, B. Verhelst, E. Derelle, et al., An improved genome of the model marine alga *Ostreococcus tauri* unfolds by assessing Illumina de novo assemblies, BMC Genom. 15 (2014) 1103.
[61] V. Shulaev, D.J. Sargent, R.N. Crowhurst, et al., The genome of woodland strawberry (*Fragaria vesca*), Nat. Genet. 43 (2011) 109–116.
[62] S. Koren, B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman, A.M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, Genome Res. 27 (2017) 722–736.
[63] M.H.W. Schmidt, A. Vogel, A.K. Denton, et al., De novo assembly of a new *Solanum pennellii* accession using nanopore sequencing, Plant Cell 29 (2017) 2336–2348.
[64] H. Du, Y. Yu, Y. Ma, Sequencing and *de novo* assembly of a near complete indica rice genome, Nat. Commun. 8 (2017) 1–12.
[65] H. Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, Bioinformatics 32 (2016) 2103–2110.
[66] C.S. Chin, P. Peluso, F.J. Sedlazeck, et al., Phased diploid genome assembly with single-molecule real-time sequencing, Nat. Methods 13 (2016) 1050–1054.
[67] W. Zhang, J. Chen, Y. Yang, et al., A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies, PLoS One 6 (2011) e17915.
[68] G. Narzisi, B. Mishra, Comparing *de novo* genome assembly: the long and short of it, PLoS One 6 (2011) e19175.
[69] F. Finotello, E. Lavezzo, P. Fontana, et al., Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data, Brief. Bioinform. 13 (2012) 269–280.
[70] M. Jain, Next-generation sequencing technologies for gene expression profiling in plants, Brief. Funct. Genom. 11 (2011) 63–70.
[71] C. Trapnell, B.A. Williams, G. Pertea, et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol. 28 (2010) 511–515.
[72] F. De Bona, S. Ossowski, K. Schneeberger, G. Rätsch, Optimal spliced alignments of short sequence reads, Bioinformatics 24 (2008) i174–i180.
[73] C. Trapnell, L. Pachter, S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics 25 (2009) 1105–1111.
[74] M.G. Grabherr, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652.
[75] M.H. Schulz, D.R. Zerbino, M. Vingron, E. Birney, Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels, Bioinformatics 28 (2012) 1086–1092.
[76] S. Kotwal, S. Kaul, P. Sharma, et al., De novo transcriptome analysis of medicinally important *Plantago ovata* using RNA-Seq, PLoS One 11 (2016) e0150273.
[77] K. Nakasugi, R. Crowhurst, J. Bally, P. Waterhouse, Combining transcriptome assemblies from multiple *de novo* assemblers in the allo-tetraploid plant *Nicotiana benthamiana*, PLoS One 9 (2014) e91776.
[78] P.M. Jonah, L.L. Bello, O. Lucky, A. Midau, S.M. Moruppa, Review: the importance of molecular markers in plant breeding programmes, Global J. Sci. Front. Res. 11 (2011) 5–12.
[79] G.T. Marth, I. Korf, M.D. Yandell, et al., A general approach to single nucleotide polymorphism discovery, Nat. Genet. 23 (1999) 452–456.
[80] A.R. Quinlan, D.A. Stewart, M.P. Stromberg, G.T. Marth, Pyrobayes: an improved base caller for SNP discovery in pyrosequences, Nat. Methods 5 (2008) 179–181.
[81] G. Barker, J. Batley, H. O'Sullivan, et al., Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP, Bioinformatics 19 (2003) 421–422.
[82] C. Duran, N. Appleby, T. Clark, et al., AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants, Nucl. Acids Res. 37 (2009) D951–953.
[83] A.D. Johnson, SNP bioinformatics: a comprehensive review of resources, Circulation: Cardiovasc. Genet. 2 (2009) 530–536.
[84] D.C. Koboldt, K. Chen, T. Wylie, et al., VarScan: variant detection in massively parallel sequencing of individual and pooled samples, Bioinformatics 25 (2009) 2283–2285.
[85] L.F. Stead, K.M. Sutton, G.R. Taylor, et al., Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: applications in tumor subclone resolution, Hum. Mutat. 34 (2013) 1432–1438.
[86] K. Lai, P.J. Berkman, M.T. Lorenc, et al., WheatGenome.info: an integrated database and portal for wheat genome information, Plant Cell Physiol. 53 (2012) e2.
[87] D.M. Goodstein, S. Shu, R. Howson, et al., Phytozome: a comparative platform for green plant genomics, Nucl. Acids Res. 40 (2012) D1178–D1186.
[88] S.D. Zhang, L.Z. Ling, T.S. Yi, Evolution and divergence of SBP-box genes in land plants, BMC Genom. 16 (2015) 787.
[89] N. Fernández-Pozo, J. Canales, D.N. Guerrero-Fernández, et al., EuroPineDB: a high-coverage web database for maritime pine transcriptome, BMC Genom. 12 (2011) 366.
[90] J.L. Wegrzyn, J.M. Lee, B.R. Tearse, D.B. Neale, TreeGenes: a forest tree genome database, Inte. J. Plant Genom. 2008 (2008) 1–7.
[91] D.H. Ware, P. Jaiswal, J. Ni, et al., Gramene, a tool for grass genomics, Plant Physiol. 130 (2002) 1606–1613.
[92] M.K. Tello-Ruiz, J. Stein, S. Wei, et al., Gramene 2016: comparative plant genomics and pathway resources, Nucl. Acids Res. 44 (2016) D1133–40.
[93] A. Bombarely, N. Menda, I.Y. Tecle, et al., The sol genomics network (solgenomics.net): growing tomatoes using perl, Nucl. Acids Res. 39 (2011) D1149–1155.
[94] N. Fernandez-Pozo, N. Menda, J.D. Edwards, et al., The sol genomics network (SGN)-from genotype to phenotype to breeding, Nucl. Acids Res. 43 (2015) D1036–D1041.
[95] S. Dash, J.D. Campbell, E.K. Cannon, et al., Legume information system (LegumeInfo. org): A key component of a set of federated data resources for the legume family, Nucl. Acids Res. 44 (2016) D1181–D1188.
[96] M.D. Gonzales, E. Archuleta, A. Farmer, et al., The Legume Information System (LIS): an integrated information resource for comparative legume biology, Nucl. Acids Res. 33 (2005) D660–D665.
[97] W. Jiao, K. Schneeberger, The impact of third generation genomic technologies on plant genome assembly, Curr. Opin. Plant Biol. 36 (2017) 64–70.
[98] M. Imelfort, D. Edwards, De novo sequencing of plant genomes using second-generation technologies, Brief. Bioinform. 10 (2009) 609–618.
[99] D. Earl, K. Bradnam, J.S. John, et al., Assemblathon 1: a competitive assessment of de novo short read assembly methods, Genome Res. 21 (2011) 2224–2241.
[100] S.L. Salzberg, A.M. Phillippy, A. Zimin, et al., GAGE: a critical evaluation of genome assemblies and assembly algorithms, Genome Res. 22 (2012) 557–567.
[101] K.R. Bradnam, J.N. Fass, A. Alexandrov, et al., Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species, GigaScience 2 (2013) 10.
[102] A.V. Zimin, G. Marçais, D. Puiu, et al., The MaSuRCA genome assembler, Bioinformatics 29 (2013) 2669–2677.
[103] D.B. Neale, J.L. Wegrzyn, K.A. Stevens, et al., Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies, Genome Biol. 15 (2014) R59.
[104] J.L. Wegrzyn, J.D. Liechty, K.A. Stevens, et al., Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation, Genetics 196 (2014) 891–909.
[105] A. Zimin, K.A. Stevens, M.W. Crepeau, et al., Sequencing and assembly of the 22-Gb loblolly pine genome, Genetics 196 (2014) 875–890.
[106] T. Chu, C. Lu, T. Liu, et al., Assembler for de novo assembly of large genomes, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) E3417–3424.
[107] H. Alhakami, H. Mirebrahim, S. Lonardi, A comparative evaluation of genome assembly reconciliation tools, Genome Biol. 18 (2017) 93.
[108] D.L. Cameron, J. Schröder, J.S. Penington, et al., GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly, Genome Res. 27 (2017) 2050–2060.
[109] L.D. Stein, The case for cloud computing in genome informatics, Genome Biol. 11 (2010) 207.
[110] P. Kumari, R. Mazumder, V. Simonyan, K. Krampis, Advantages of distributed and parallel algorithms that leverage Cloud Computing platforms for large-scale genome assembly, F1000Research 4 (2015) 20.
[111] H. Nagasaki, T. Mochizuki, Y. Kodama, et al., DDBJ Read Annotation Pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data, DNA Res. 20 (2013) 383–390.