Current Plant Biology xxx (xxxx) xxx–xxx

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*-