

Preface

DNA sequencing is a fast-moving science with technologies and platforms being updated at breathtaking speed. The hallmark of next generation sequencing (NGS) has been a massive increase in throughput and a decrease in price compared with previous technologies. The first next-generation DNA sequencing machine was introduced to the market by 454 Life Sciences (Basel, Switzerland) in 2005. The technology is based on a large-scale parallel pyrosequencing system, which relies on fixing nebulized and adapter-ligated DNA fragments to small DNA-capture beads in a water-in-oil emulsion. The Illumina's (CA, USA) Genome Analyzer was released in 2007 and marked a true revolution for genome sequencing in which short reads became significant to genomic applications. The technology is based on reversible dye terminators. DNA molecules are first attached to primers on a slide and amplified so that local clonal colonies are formed. Life Technologies' (CA, USA) SOLiD™ technology employs sequencing by ligation. In this technology, a pool of all possible oligonucleotides of a fixed length is labeled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal that is informative of the nucleotide at that position.

So-called 'third-generation' technologies directly sequence individual DNA molecules rather than relying on any amplification prior to sequencing. The recently released PacBio system can produce 35–45 Mb of data per cell with an average read length of 1,500 bp. The Ion Torrent Personal Genome Machine (PGM) is another third-generation platform that uses standard sequencing chemistry, but with a novel, semiconductor-based detection system. This technology already claims read lengths of approximately 200 bp with high accuracy, and the latest PGM 318 chip can produce 1.0 Gb of data in a 2-h run. When the implications of NGS technology became apparent, several assemblers were designed to deal with the new problems, i.e., assembly of short NGS reads in order to reconstruct the main longer sequences. Assembly process can be done either having a reference genome available (mapping) or without having a reference genome available (de Novo assembly). De Novo assembly algorithms, discussed in more detail in this book, can be classified into three main categories: greedy algorithms, Overlap-Layout-Consensus (OLC) methods, and De Bruijn graph approaches. The Euler assembler was the first to employ de Bruijn graphs for

whole genome shotgun (WGS) assembly, and proved capable of assembling bacterial genomes. Velvet and ALLPATHS improved assembly in terms of speed, contig and scaffold length, and avoidance of misassembly. ABySS followed the innovations with de Bruijn methods, but also introduced a distributed representation of the graph, allowing message passing interface parallelization. The CABOG and variant MSR-CA pipelines are updates of the Celera overlap-based assembler designed for a combination of read types, which showed some success with short-read data for genomes in the 100 Mb range. The String Graph Assembler (SGA) is the first to make assembly of mammalian-sized genomes practical using the string graph approach. This observation on the current tradeoff between accuracy and continuity suggests avenues for future improvements in assembly. There is room for other improvements at the scaffolding stage, where, as has happened at the assembly stage, we witness a move from naïve and greedy algorithms to more subtle graph-based techniques.

In this book, we briefly introduce the history of first, second, and third generation sequencing technologies and also describe drawbacks of the old techniques which now are not suitable due to their cost and the need for automation which could not be achieved in those methods. In [Sect. 2](#) major NGS methods—namely Roche/454 FLX, Illumina/Solexa Genome Analyzer, and Applied Biosystems SOLiD System, etc.—are described in detail. Also, after bringing the latest and most predominant technologies in NGS, nanopore DNA sequencing and Pacific single molecule real time (SMRT) DNA sequencing, which does not need an amplification step, are described. Latest subsections of this section are devoted to information about sequencing costs, file formats of the output, a comparison of methods, and their drawbacks, and finally application of NGS technologies. The second two sections, i.e. [Sects. 3](#) and [4](#), provide an overview of the algorithmic view of the assembly problem. Our main focus in these two sections will be on de Novo assembly algorithms of NGS reads. In [Sect. 3](#), we generally define the assembly problem and mention the challenges involved in the assembly process, including errors propagated from sequencing process beside computational challenges. Appropriate use of paired-end read data, which helps to overcome the challenges regarding short length of reads, and also preprocessing that helps to eliminate some other issues regarding inaccurate data, is the next topic discussed in this section. Using all these techniques to reduce problems, there will still be errors in assembly, and relevant assembly algorithms are needed to be validated in a standard way: These are the final topics which will be discussed in [Sect. 3](#). Finally, in [Sect. 4](#), an exact view of the assembly algorithm is given as to how the problem can be mapped to a graph and how different kind of graphs are treated in finding the solution, which is the final assembled genome. Concerning each of the assembly approaches, several example algorithms are then described in detail and, finally, a comparison of these methods is provided in [Sect. 4](#).

Next Generation Sequencing and Sequence Assembly
Methodologies and Algorithms

Masoudi-Nejad, A.; Narimani, Z.; Hosseinkhan, N.

2013, X, 86 p. 46 illus., 43 illus. in color., Softcover

ISBN: 978-1-4614-7725-9