Research Paper
Impact Factor: 3.567
Peer Reviewed Journal

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

SIGNIFICANCE OF MASSIVELY PARALLEL SEQUENCING STRATEGIES AND *DE NOVO* ASSEMBLY ALGORITHMS IN WHOLE GENOME SEQUENCING

H.W. Rajitha Udakara Sampath* Dr.R.S. Dassanayake *
**Department of Chemistry, Faculty of Science, University of Colombo, Sri Lanka.*

Abstract
Massively parallel sequencing (MPS) is a novel approach for sequencing genomes and it provides a significantly higher throughput when compared with the conventional sequencing platforms. Therefore, it has become a better solution for achieving genome sequences of particular organisms with a greater accuracy and precision. Genome assembly is the immediate process which is performed subsequently to genome sequencing and if the genome of a particular organism has not been sequenced previously, de novo assembly approach is the sole mode of acquiring the complete genome. When considering a plausible model organism, it is crucial to procure the complete genome sequence in order to provide a better biological insight to that particular organism. This article therefore, begins to discuss the significance of acquiring the complete genome sequence of a plausible model organism and then the major approaches that are available for genomes sequencing showing the suitability of massively parallel sequencing approaches for such exercise. Thereafter, the comparison of next generation sequencing platforms and sequencing assembly algorithms showing the importance of de novo assembly approach in achieving genome sequencing are discussed while highlighting the importance of quality assessment and validation procedures for sequenced genomes. Finally, the challenges and their countermeasures in whole genome sequencing of model organisms are addressed.

Keywords:Genome sequencing, Next Generation Sequencing, Genome assembly, De novo assembly, Sequence validation.

Introduction

Genome sequencing is the process in which the nucleotide order of a particular genome is achieved in terms of four different nucleotides that are Adenine, Guanine, Thymine, and Cytosine. There are numerous advantages of acquiring the complete genome sequence of an organism, especially in research purposes since it represents the entire biological and biochemical insight into the organism of interest [1]. As a result of that most of the biological and biochemical research projects that are based on animals or plants, are demanding for the sequence information in order to ascertain biological processes or biochemical pathways that are being taken place within them.

In the past, genome sequencing and genome assembly processes were highly challenging due to the time consumption and technological issues. However, with the development of the sequencing technology and improvements in the field of computer science, the complexity of sequencing and assembly processes have been drastically reduced [2]. For an instance, human genome project (HGP) which was the first massive genome sequencing project started in the year 1990, acquired more than ten years to complete the genome sequence [3]. On the other hand, with the rapid improvement of next generation sequencing technology, currently it takes less than a single day in order to get the complete sequence of a human genome. However, when it is performed in a *de novo* fashion, still it would take a comparatively longer period of time than that of template-directed sequencing. Fundamentally, *de novo* sequencing is a process in which the whole genome sequence of a particular organism is achieved by merely considering overlapping sequences without utilizing any template genome sequences [4]. Therefore, the organisms which have not been sequenced previously or the organisms which do not possess appropriate “models” are required to be necessarily undergone *de novo* sequencing process with the use of plausible model organisms.

Genome sequencing and significance of acquiring the complete genome sequence

Genomic sequencing is a laboratory process which determines the complete DNA sequence of an organism's genome at a single time. With the advancement of genomics technologies, sequencing genomes have been widely improved and various sequencing platforms have come into the picture.

The most preliminary method for sequencing is “Sanger sequencing” which is also known as dideoxynucleotide chain termination method. In this case, termination of DNA polymerization reaction in the presence of dideoxynucleotides is considered as the driving force of sequencing and gel electrophoresis is utilized for the prediction of the order of nucleotide. Sanger sequencing and automated versions of Sanger sequencing were utilized for a long period of time and then requirement for a rapid and high throughput sequencing methodology was prominently emerged [5]. After that, generation-wise improvement of sequencing technologies began and three important generations of sequencing were introduced (Second generation, Third generation and Fourth generation of sequencing). Out of them, a combination of second and third generation sequencing methods is particularly termed as Next Generation Sequencing (NGS) which is the most abundant, rapid and high throughput sequencing technology. Fourth generation sequencers are not widely used for routine sequencing

purposes and most of them are still at the research level. However fourth generation sequencers such as Oxford Nanopore sequencer, have eliminated most of the drawbacks of Next Generation Sequencing methodologies and it has provided a further rapid and reliable platform for sequencing genomes in research purposes [6].

Genome sequencing is a vital process for a model organism thus the complete biological and biochemical insight to the particular organism can be investigated via complete genome sequence. Novel functionalities that are encoded by genomes, evolutionary linkages, alterations in conserved gene products *etc.* can be readily and more reliably investigated by using the whole genome sequence of a particular organism. Further, the complete genome sequence is the major requirement of functional genomics studies and it demands an error-free genome sequence as a starting point. Regardless of the definition of functional genomics, all downstream works beginning with genome annotation is greatly facilitated by a complete, high-quality DNA sequences. This is true in the case of defining gene coordinates in a genome; identifying paralogous gene families or designing PCR primers and probes for microarray analysis *etc.* Robust annotation of any genome sequence will ultimately require experimental work that will proceed more quickly and economically with a prevailing complete genome sequence as a starting point [7]. Another important point is that the process of comparative genomics is meaningful only if the complete gene (or genome) sequences are taken into consideration, because merely the draft genome or partial genome sequence will not represent the entire set of features of the complete gene (or genome). Therefore, it is necessary to obtain the whole genome sequence of a particular organism in order to comment on the genetics, genomic organization gene functions *etc.* Moreover, the genome sequence of a particular organism is a permanent and valuable scientific resource. That means, if an organism is sufficiently important to study at a particular instance, then a complete genome sequence of at least one strain provides the basis for future investigations. Primarily, the complete genomic sequence represents a permanent snapshot of one moment in evolutionary history and therefore it remains accurate even though the particular organism will continue to evolve [7].

1.2.1. Major approaches of genome sequencing

Basically, there are three major approaches for sequencing genomes that are, clone by clone shotgun sequencing; whole genome shotgun sequencing and hybrid strategies for shotgun sequencing. With the improvement of chemistry, biochemistry and nanotechnology *etc.* sequencing platforms have been gradually evolved and currently genome sequencing process is in a rapidly developing phase. There are numerous sequencing platforms that are being developed day by day and many research projects are being designed or conducted upon the introduction and improvement of novel sequencing platforms.

Clone by clone approach involves shotgun sequencing of individually mapped clones and it is one of the most utilized sequencing approaches for genomic sequencing purposes. In human genome project and other major genome projects such as yeast and nematodes *etc.* clone by clone approach has been utilized. The most important feature in this technology is map construction process that is done prior to sequencing phase. In clone by clone approach, initially the genome sequence is divided into numerous clones and subclones and then those individual clones/subclones are sequenced. Thereafter, the sequenced segments reassembled on to the reference genome by using the map constructed prior to sequencing. As a consequence of that, clone by clone shotgun sequencing approach is referred to as “Hierarchical sequencing or Map based shotgun sequencing”. Principally, genome map can be either physical map or a genetic map. In the process of construction of genetic maps, fundamental techniques in genetics such as calculation of recombination frequencies, cross breeding experiments, and analysis of family histories are utilized. Therefore, genetic maps are considered to acquire a poor resolution when compared to the physical maps which are constructed by utilizing more specific and advanced techniques [8]. However, clone by clone approach necessarily demands genetic maps in certain instances, because particular organisms such as humans are not possible to be subjected to physical mapping.

Whole genome shotgun sequencing approach is rather different from clone by clone shotgun sequencing approach. The most prominent feature of whole genome shotgun sequencing approach is the absence of a genome map that guides the genome assembly process at the end of the sequencing phase [9]. In this case, by using chemical or physical methods, the particular genome is fragmented so that it gives a large number of overlapping library fragments and then sequence reads are generated in a genome wide fashion. As a result of that the requirement for the construction of genome map is avoided and depending on the redundancy of coverage, the quality or the accuracy of the sequence assembly is determined [10]. Genome assembly is a process that demands an intense computational power. The reason is that, the enormous amount of sequence data generated through massively parallel sequencing (MPS) cannot be manually handled and therefore, highly advanced computer algorithms should be utilized for the sequence assembly process to be done properly. Even though whole genome shotgun sequencing provides a quick and easy solution for genome assembly, there are several prominent defects with this sequencing approach. Out of them, the misassembly of genomic fragments due to the presence of repeating sequences is considered to be

a major source of errors. Basically, Tandem repeats and Genome-wide repeats are the major types of repeating sequences that should be taken into consideration in the process of genome assembly. Tandem repeats are the stretches of sequences that are present in head to tail fashion within a particular genome [11,12]. Most of the time they are localized repeating sequences and significant portions of those tandem repeats could be omitted due to the errors in sequence assembly. Genome-wide repeating sequences are situated far apart in the genome and failure to identify them, can lead to omission of large parts of a particular genome.

Hybrid approach between clone by clone shotgun sequencing and whole genome shotgun sequencing has become the most accepted methodology for sequencing the complete genome of a particular organism. Most of the novel genome sequencing projects are conducted by using hybrid approach as to why there are several preferable features with this approach [12]. As mentioned earlier, there are few drawbacks with clone by clone approach and whole genome shotgun approach when they are utilized separately. However, most of the drawbacks have been overcome by combining those two sequencing approaches. For an instance, when considering clone by clone approach, the major issue was the necessity of a template or a genomic map for the assembly process to be carried out. However, map construction is a laborious process requiring a considerable time [9]. On the other hand, in whole genome shotgun sequencing approach, the requirement of genome mapping has been eliminated and instead of that computer based *de novo* assembly algorithms were utilized for acquiring the whole genome sequence. The major drawback with whole genome shotgun sequencing was the difficulty of sequence assembly due to the presence of tandem and genome-wide repeating sequences [11]. However, when the two approaches are combined, the requirement for the genomic map (clone by clone approach) and the errors occurred due to the presence of repetitive sequences (whole genome shotgun sequencing approach) are minimized [13]. Further, depending on the type of genome and the redundancy of coverage, the appropriate rations between those two approaches are determined [11]. For an instance, when considering a complex genome with highly repetitive sequences, whole genome shotgun sequencing bias is reduced and clone by clone sequencing bias should be necessarily improved over those repetitive sequences. Thereby, an error free genome assembly can be obtained regardless of the presence of repetitive sequences.

1.2.1.1. Massively parallel sequencing, a better approach for sequencing the whole genome

As mentioned earlier, sequencing technologies have been developed in a generation-wise manner and three major generations have come into the picture from the beginning of the time. Out of them, a combination of second and third generation sequencing methodologies is considered as Next Generation Sequencing (NGS) [14]. Besides, second generation sequencing platforms are particularly referred as massively parallel sequencing (MPS) and they perform millions of sequencing reactions simultaneously within a single sequencing run [19-21]. In general, each NGS platform acquires a complex interplay of enzymology, chemistry, high-resolution optics, hardware, and software engineering. Also, these platforms allow highly efficient and reliable sample preparation steps prior to DNA sequencing, which provides a significant time saving and a minimal requirement for associated equipment. This is an important feature of NGS related methodologies when compared with the highly automated, multistep pipelines necessary for clone-based high throughput sequencing techniques such as capillary Sanger sequencing. Although clone-based library preparation is not done in NGS related platforms, amplification of the genomic fragments is necessarily done prior to sequencing process. In this case, library fragments are formed by annealing platform-specific linkers to the blunt ended cleaved fragments of genomic DNA or DNA source of interest. Thereafter, the entire set of library fragments can be subjected to either bridge amplification process or emulsion PCR amplification so that eventually it gives a large number of fragment libraries with high copy numbers [14]. Therefore, NGS related platforms have become more prevalent due to the lack of the more cumbersome and laborious bacterial cloning steps or bacterial intermediates for the amplification of the library fragments. Fundamentally, 454 pyrosequencing (Roche sequencer), Illumina (Solexa) sequencing, SOLiD sequencing and Ion Torrent semiconductor sequencing are the four major types of second generation sequencing platforms which have been recently introduced and their sequencing capacities have been improved rapidly with the development of technology. Further, Pacific Bio sequencer (Single Molecule Real Time sequencing or SMRT) and Heliscope single molecule sequencer are third generation sequencing platforms which are highly sophisticated and rapidly emerging sequencing platforms that are available for genome sequencing. However, it is important to notice that, even though both second and third generation sequencing platforms are collectively considered as next generation sequencing (NGS) methodologies, the basic principles and the sequencing cascades are significantly different within those two generations thus the accuracy and the precision of sequencing products are significantly influenced [15,22].

In the process of sequencing, the amplification step is not performed in third generation sequencers whereas extensive amplification steps are widely utilized in second generation sequencers. This extensive amplification phase is considered to be the major source of error in second generation sequencing platforms and it further, reduces the accuracy of the sequence reading. Instead of amplification, third generation platforms utilize a set of oligo-dT primers immobilized either on a flow

cell surface or else immobilized DNA polymerase enzyme at the bottom of zero-mode wave-guides (ZMWs) [18,19]. Thereafter, fluorescent labeled dNTPs are added to the sequencing reaction mixture thus incorporation of particular types of bases can be readily observed. Further, the absence of amplification process improves the reliability and efficiency of the sequencing process while reducing the cost per base in third generation sequencers. Also, comparatively higher read length of the third generation sequencing platforms as in the case of Pacific bio-sequencer (SMRT) make the reassembling process more effective while further improving the accuracy of the sequencing process [5,15,21,24].

Fundamentally, all the next generation sequencing platforms are conferring different aspects of DNA polymerization reaction and eventually, the types of incorporated bases are identified or characterized via different approaches such as fluorescent labeling, change in chemical properties (e.g. pH value of the surrounding medium) *etc.* Theoretically, there are four major steps in next generation sequencing platforms and they are termed as fragmentation, tagging, amplification and sequencing [20]. Although, the principles are slightly different among those NGS platforms, the order of the basic steps are necessarily maintained. Subsequently, data analysis process is done with the aid of platform-integrated and highly sophisticated software packages. Next generation sequencing platforms such as Ion torrent PGM and Illumina sequencer are frequently integrated with software packages such as Partek® or SeqMan genome analyzer and thereby highly accurate and efficient data analysis can be achieved [14]. Considerations in assortment of the most appropriate NGS platform for sequencing complex genomes.

1.3 Considerations in assortment of the most appropriate NGS platform for sequencing complex genomes

Most of the next generation sequencing platforms can be effectively utilized for *de novo* sequencing of complex eukaryotic genomes. Therefore, when selecting a particular platform for whole genome sequencing there are several important factors that should be necessarily taken into consideration. Particularly, average read length, reads per run, run time, throughput, number of sensors, instrumental price, reagent price and accuracy are the factors that are being predominantly emerged (Table 1.0). Fundamentally, Illumina sequencer is the most widely utilized and the most developed sequencing platform out of all the other second and third generation sequencing platforms. The other NGS platforms such as Ion torrent sequencer are widely utilized for targeted sequencing rather than whole genome sequencing [21]. However, some of the third generation sequencers and most of the fourth generation sequencers are still under research level and they are being improved gradually with the development of the technology.

Feature	***Roche sequencer***	***Illumina sequencer***	***SOLiD sequencer***	***Ion Torrent sequencer***	***PacBio Sequencer***
Company	Roche	Illumina	Life technologies	Life technologies	Pacific Bioscience
Sequencing Chemistry	Pyro-sequencing	Reversible terminator	Ligation	Proton detection	Real time sequencing
Read length	500-700	2×100	85	200-400	3000 (up to 15000)
Run time (depend on the length of sequence)	8-24 hours	2 days (rapid mode)	8 days	2 hours	20 minutes
Gb per run	0.04 – 0.7	120 (rapid mode)	150	100	3
Reagent cost per Mbp	$10	$0.07-$ 0.5	$0.13	<$1	$7
Instrumental price (Cost per run)	$500000 ($7000)	$130000-$690000 ($1000-$6000)	$500000 ($15000)	$80000-$150000 ($1000)	$695000 ($300)
Library preparation	Emulsion PCR	Bridge amplification	Emulsion PCR	Emulsion PCR	Single molecule
Primary errors	Substitution, Indels	Substitution	AT bias	Indels	CG deletions

Error rate	0.1% - 1.0%	0.2% - 0.8%	0.01%	1.7%	12.86%
Paired end reads	Yes	Yes	Yes	Yes	No
Typical DNA requirement	~ 1 µg	50-1000 ng	~ 1 µg	100-1000 ng	~ 1 µg
Application	*De novo* WGS of microbes, Pathogen discovery, Exome sequencing	Human WGS, Exome seq, RNA seq, Microbial discovery, Targetted capture	Human WGS, Exome seq, RNA seq, Methylation	Microbial discovery, Targetted capture, Exome sequencing	Human WGS, Exome seq, RNA seq, Methylation
Advantages	Long read length	Highest throughput, Lowest cost per Mb	Low cost per base, Higher accuracy	Low cost per sample	Longest read length, No amplification error
Disadvantages	High capital cost, High cost per Mb	High capital cost, High computation needs	Slower method, Lower reads	High cost per	High error rates, comparatively small outputs, High cost per Mb

As a whole, most of the NGS platforms including bench-top sequencers (second generation sequencers) are capable of generating useful draft genome sequences with assemblies that mapped to ≥95% of the reference genome. However, no instrument is gifted with the ability to generate the accurate one-contig-per-replicon assemblies that might equate to a finished genome.

Strategies for assembly of complex genomes

Genome assembly is the process in which sequence reads are assembled on to the reference genome thus it gives an error free sequence assembly that resembles the reference genome. As mentioned earlier, next generation sequencers generate a large number of sequencing reads in a single run and therefore it is crucial to assemble the reads precisely within a short period of time. However, this assembly process cannot be performed manually; instead, complex computational algorithms should be essentially utilized thus more efficient and more accurate genome assembly is achieved.

At the beginning of sequencing and genome assembly era, Sanger sequencing was the sole approach for genome sequencing and there were not highly sophisticated computational algorithms for assembly of sequenced genomic data. Instead, scientists divided large genomes into numerous sets of genomic segments and the resulted segments were treated separately so that the complexity of the assembly process is significantly reduced. Then, there were several evolutionary improvements of Sanger sequencing method such as introduction of laser based instrumentation that allows the automated detection of fluorescently labeled DNA molecules, improvement of biochemical components in sequencing reaction such as thermo stable polymerases, fluorescent dye labeled dideoxy terminators and robust fluorescent dyes *etc.*[22]. Another significant improvement in Sanger sequencing method is the introduction of various robotic systems that have been designed in order to automate particular steps in sequencing process. Those automated steps include systems that facilitate subclone library construction, picking and arraying subclones, template purification and loading samples into slab gel or capillaries [23]. As consequences of them, the amount of data generated per single instrumental run was dramatically increased while making the manual assembly of sequenced fragments is a highly challenging process which demands very much of time and labor. Therefore, a prominent requirement for highly advanced computer algorithms were emerged thus it allows the easy manipulation of large amount of data generated within a short period of time. Afterwards, numerous important software packages came into the picture and those packages were capable of analyzing primary sequence data, carrying out sequence assembly and calling for nucleotide bases at each position of the sequence while assigning a corresponding quality score in order to reflect the statistical likelihood that the indicated base call is correct. Software packages such as “Phred”, “Phrap” and “Consed” are well known packages which are designed for base calling, sequence assembly and viewing the sequence assembly [24,25].

1.2.2. *De novo* assembly of complex genomes

By definitions, *de novo* assembly is the process in which the sequencing reads generated via massively parallel sequencing platforms are pieced back together with the utility of highly sophisticated computer algorithms thus it results in organism’s chromosome. Unlike other strategies, the most prominent feature of *de novo* assembly is the absence of a reference genome in the sequence assembly process. Reference genome provides the template for sequence assembly by guiding the sequencing

reads to be assembled in a predetermined manner. Therefore, the approaches which do not depend upon reference genomes, acquire several advantages over the other prevailing strategies. In the case of *de novo* assembly, there is a greater potential to detect a more complete set of genetic variations, especially novel sequences and structural variations even in relatively well studied genomes [26]. However, template directed assemblers are not capable of performing the above functionalities at a comparable level with *de novo* assembly.

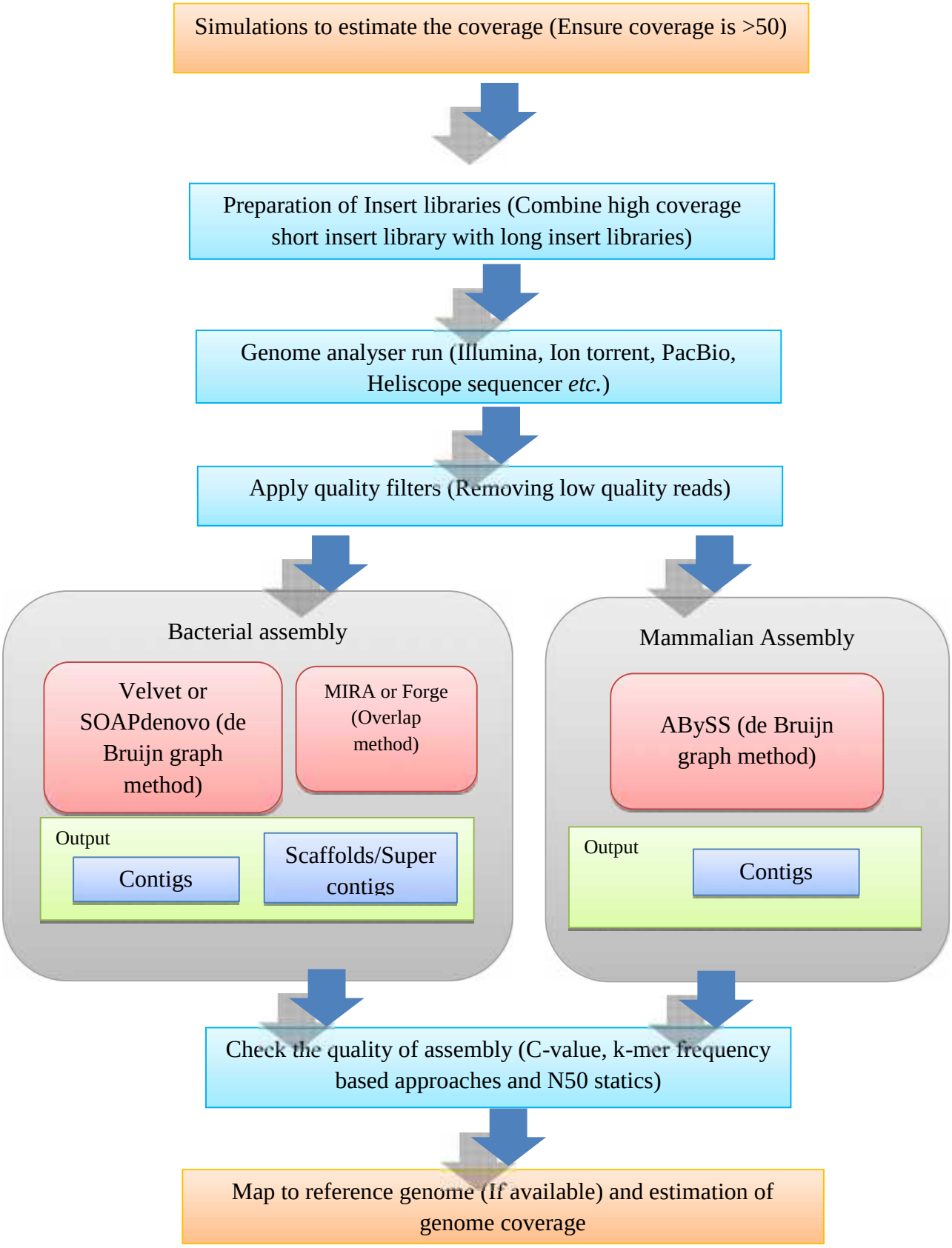
In the beginning of evolution of genome sequencing, clone by clone shotgun sequencing approach was highly utilized and sequencing genomes of complex organisms was an expensive and highly time consuming process. However, with the development of the technology, computer based genome assembly algorithms were rapidly emerged and therefore genome sequencing has become an efficient, cost effective and highly accurate process [20,27]. Another important feature of *de novo* assembly process is the manipulation of a large amount of data generated through massively parallel sequencing platforms. For an instance, when considering human genome project which was conducted utilizing Sanger sequencing, eventually it resulted in a genome sequence with the depth of coverage around seven-fold. It indicates that more than 2.1×10^{10} bp ($7 \times 3 \times 10^{9}$ bp) have been sequenced throughout the genome sequencing process. On the other hand, if a human genome sequencing project which is conducted utilizing massively parallel sequencing approaches, is taken into consideration, the depth of coverage is around hundred-fold. That indicates, 3×10^{11} bp have been sequenced and manipulated during the entire process. Further, the advancement of computer science and the introduction of highly sophisticated computer algorithms for sequence assembly have facilitated the genome sequencing process and therefore whole genome sequencing has become a more convenient process and also it has become the simplest solution for most of the scientific problems such as investigation of evolutionary history and functional studies of different proteins *etc*. [3,42,44].

Fundamentally, genome assembly is a stepwise process that is performed by considering the sequence overlapping, a phenomenon that occurs when the terminal nucleotide stretches of two reads are identical with each other. Thereby, adjacent sequence reads can be recognized and the order of reads over the genome can be predicted. Moreover, depending on the overlap to read length ratio the quality and the accuracy of resulted "contiguous sequences" are determined. When the assembly process reaches the level of "contigs", genome wide consideration of entire set of contigs is done by the assembly algorithm. That means, by considering the sequence overlapping, all the contiguous sequences are subsequently assembled into "Scaffolds" [42,45]. The most prominent issue within this process is the presence of repeating sequences in most of the genomes. As mentioned earlier, repeating sequences can be the major source of errors in sequence assembly, not only it underestimates the size of genome but also it causes erroneous sequence assemblies thus it gives rise to sequencing scaffolds that are not actually found in the original genome. Therefore, highly advanced computer based *de novo* algorithms have been introduced so that the errors in assembly process are minimized. Sequencing scaffolds gives rise to chromosomes at the end of sequence assembly and the whole genome of a particular organism consists of set of individual chromosomes. (The major steps in de novo assembly process have been summarized in "Figure 1.0" and some of the important concepts in *de novo* assembly such as assembly algorithms, quality assessment and validation have been discussed in detail in imminent sessions.)

1.2.2.1. Several major *de novo* assembly algorithms

De novo assembly was introduced in the year 2013 and since then it has been gradually improved with the development of technology. Also, *de novo* assembly has become a better solution for sequencing the complex genomes because of several important features such as that minimizing the sequencing errors, higher depth of coverage, long enough reads and properly spaced paired end reads [4]. Basically, there are three major algorithmic approaches for *de novo* assembly and they are termed as "overlap-layout-consensus", "de Bruijn" and "String graph" algorithm. All those three algorithms are heavily utilized in novel genome projects and out of them, string graph algorithm (SGA) is the most recently introduced algorithm for genomic assembly [30]. Further, all the above mentioned algorithms are based on respective theoretical graph framework and the difference betw een those algorithms are coming from the specificity in their treatment of repeats in assembly process.

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877



In the case of overlap-layout-consensus (OLC) assembly, overlaps between all reads are first identified and then contigs are formed by iteratively merging overlapping reads until a read heuristically determined to be at the boundary of a repeating sequence. Basically, repeats shorter than the minimally expected read overlap are often resolved considering that genome resolution increases with read length [31]. Therefore, imprecise read overlaps are allowed in order to account for sequencing errors and in particular circumstances, this procedure may fragment the assembly even when the genomic repeats are nearly identical. Also, OLC algorithms have performed a major role in human genome project due to the simplicity and accuracy of those algorithms. Parallel contig assembly program (PCAP), Forge, Arachne, MIRA and Celera are known to be widely utilized OLC algoritms for routine genome assembly purposes [32].

De novo assembly algorithms based on de Bruijn graphs begin with the replacement of each read with the set of all-overlapping sequences of shorter, fixed length. The length is often denoted by *k*, and the sequences by *k*-mers. Also, the contigs are formed by merging *k*-mers appearing adjacently in reads halting at *k*-mers from repeat boundaries [47,48]. Highly accurate reads are essentially required in this process and in certain instances; it initially discards some of the ability for reads to resolve repeats longer than *k* bases. One of the most important features of this algorithm is that it does not require the storage of pairwise overlaps or a graph structure representing the repeat structure of the genome. Therefore, de Bruijn assembly has been favored for whole genome sequencing projects in the ALLPATHS, SOAPdenovo, Velvet and ABySS mammalian next-generation sequencing assembly methods [32]. Moreover, when utilizing the algorithms based on de Bruijn graphs, the amount of information lost during this process is minimized due to the usage of *k*-mers which are very short in length. This has become the major reason for selecting these algorithms, particularly for novel genome projects.

The string graph and the related A-Bruijn graph assembly formulations are similar in concept to a “de Bruijn graph”. However, there is a prominent advantage with string graph as the sequences are not decomposed into *k*-mers and instead of that, the full-length of a sequence reads are taken. Those full-length reads are produced based on the operations of read overlap and the removal of transitively inferred overlaps. Also, there is an open-source implementation of string graph assembly called FALCON which has been produced by Pacific Biosciences [14,49]. Not only these three algorithms, several other sophisticated computer algorithms have been developed with the introduction of novel programming languages, improvements in the field of computer science and rapid advancement of sequencing technology. Therefore, sequence assembly has become easier and more reliable process thus whole genome sequencing of complex organisms has become an abundant approach to achieve a complete biological insight to a particular organism. Table 2.0 is included with several important statistical data related to certain *de novo* assembly programs that are widely utilized for novel genome sequencing projects.

Table 2.0 – Comparison of assembly statics for different *de novo* assembly programs considering an eukaryotic genome sequence with the size of approximately 100 Mbp [29,49,54,55].

Program	***Computing features***			***Algorithm***	***Genome assembly parameters***		
	Language	*Memory requirement (Gb)*	*Time used (Hours)*		*Contig N50 (Kbp)*	*Scaffold N50 (Kbp)*	*Reference bases Covered (Mbp)*
MIRA	C, C++	16.3	20	Overlap-Layout-Consensus	12.5	27.9	---
ABySS	C++	14.1	5	de Bruijn	18.4	23.8	95.9
SOAPdenovo	C, C++	38.8	13	de Bruijn	16.0	31.1	95.1
Velvet	Perl and Java	23.0	2	de Bruijn	13.6	31.3	94.8
SGA	C++	4.5	41	String Graph	16.8	26.3	96.2

1.3. Quality assessment and validation procedures for genome assembly

Genome assembly of complex genomes is a very complicated process that should be done by using highly advanced computer algorithms and intense computational power thus the error-free genome sequence of a particular organism is achieved. Once the assembly of a genome is accomplished, quality assessment and validation should be necessarily carried out by utilizing particular indices or else the acquired genome assembly is compared via different validation methods. Genome assembly and validation procedures have become highly challenging because they constitute merely a hypothesis of the true underlining genome sequence and the occurrence of draft genome assembly prior to the original genome sequence of a particular organism during genome sequencing projects.

Quality assessment and validation of an assembled draft genome sequence can be executed by utilizing a variety of metrics reflecting different aspects of the assembly process. Basically, there are two major approaches for these tasks to be accomplished. The first approach demands for the additional information from external data and the other approach is merely based on the information derived from the assembly process itself. However, in particular circumstances, the first approach fails to execute because external information about assemblies is not available in situations such as conservation genomics projects *etc.* [35]. In addition to the above mentioned approaches, there are several basic matrices that are prominently utilized in the process of validation of newly assembled draft genome sequences. Out of them, proportion of genome contained with the assembly, C-value, k-mer frequency based approaches and N50 statics are considered to be the major indices that have been recommended for the newly assembled genome sequences to be tested with [36–38].

The proportion of genome contained with the assembly is one of the most important considerations in the process of validating newly sequenced genomes. In accordance with this index, even though there is an error free genome assembly, it particularly focuses on the relative fraction of the genome that has been assembled. Therefore, for a newly sequenced genome to be considered as a draft genome assembly, there should be at least 80% coverage of the particular genome together with a significant precision [39]. C-value is another important index that is utilized for genome validation and it closely resembles the previously explained index which is the proportion of genome contained with the assembly. Basically, C-value is defined as the amount of DNA contained within a haploid nucleus or in other words it is one half the amount in a diploid somatic cell of a eukaryotic organism. Therefore, it is a kind of estimation about the proportion of the genome that has been sequenced and assembled. Moreover, *K*-mer frequency based methods are further effective in validating and assessing the genome sequences that are contained with a significant amount of repetitive sequences such as that tandem repeats, simple sequence repeats, long terminal repeats (LTRs), segmental duplications, and transposable elements (TEs) [40]. Another important index is N50 value which is a statistical parameter related to the size of contigs (or scaffolds). By definition, N50 value is the length which contains at least half of the total of the lengths of the contigs. In accordance with the scientific evidence, the average length of contigs or scaffolds is not the important factor in the genome assembly. The reason is that one set of contigs can be very long whereas the rest of sequencing contigs might be considerably short in length keeping the average length of the set of contigs at a higher value [35,36]. Therefore, instead of considering the average sizes of the contigs, a parameter such as N50 value is more reliably taken into consideration when assembling and validating newly sequenced genomes. Not only N50 value, there are several other indices such as N90 or NG50 *etc.* and they all are considered as the statistical parameters that are mainly focused on the size of the contigs.

Several computer algorithms have been introduced for the assessment and validation of newly sequenced genomes. With the advancement of sequencing technology (such as the introduction of next generation sequencing) and computer based sequence analysis tools, the number of genomes that are being sequenced and the amount of data generated in genome sequencing projects have been rapidly increased. Further, numerous genome sequencing projects are launched annually and “draft” or “permanent draft” genome assemblies have become the major output of those genome projects. Therefore, the validation and quality assessment processes have become one of the most prominent requirements of a genome project. As a result of that universal genome assembly tools such as REAPR has been introduced and the precision of validation process has been significantly improved with the use of those novel tools [39].

1.4. Challenges in *de novo* assembly of complex genomes

As mentioned earlier, *de novo* assembly of complex genomes is highly challenging due to the absence of a reference genome for sequence assembly process to be carried out. In this case, the entire genome is obtained via a route directed by the presence of overlapping fragments whereas repeating sequences such as genome wide repeats or tandem repeats have become a prominent drawback in the process of sequence assembly. Not only the above mentioned factors, but also several other

factors like heterozygosity, contaminated samples and sequencing errors (e.g. homopolymer runs, substitutions *etc.*) can be the possible sources of errors in genome sequencing and assembly [41,42].
The size and the complexity of the genome are rudimentary considerations in *de novo* assembly. For an instance, when comparing microbial and higher eukaryotic genomes, it is very convenient to deal with microbial genomes in *de novo* assembly due to comparatively less complexity and trivial size of microbial genomes. Thus the requirement of computational power, time and the cost of genome sequencing and assembly are significantly reduced. Therefore, it is evident that size and the complexity of the genome are very crucial factors in sequencing and assembly procedures. Moreover, genome size will affect the sequencing depth of coverage and in certain instances sequencing platforms are determined depending on the size of the genome. For an instance, if a particular genome is 100 Mb in size and if it requires 100× coverage for *de novo* assembly, the sequencing machine should be capable of generating around 100 × 100 Mb (1 Gb) of data for the required coverage to be obtained [20,31]. If it is a larger genome like human genome, it is very difficult to achieve a large depth of coverage when sequencing and assembling the whole genome. However, on the other hand when the size of the genome becomes smaller, it is convenient and technologically less demanding for sequencing, assembly and storage of sequencing data. With the development of technology, sequencing capacity has been dramatically improved and currently scientists are along the direction towards improving genomic assembly algorithms and storage of assembled genomes.

Another important consideration is the length of genomic segment that is sequenced at a single sequencing run or the "read length". Fundamentally, read length is classified under two major categories that are termed as "short reads" and "long reads". Short reads are within the size range of < 200-400 bp and several types of second generation sequencing platforms such as Illumina, SOLiD, IonTorrent are considered to be the short read platforms. On the other hand, long reads are greater than 400 bp in size and the sequencing platforms such as Sanger, 454 pyrosequencer, novel versions of IonTorrent are known as platforms which give long read lengths [21,36,42]. However, several types of third and fourth generation sequencers such as Pacific biosequencer (PacBio sequencer) and Oxford nanopore sequencer are capable of producing reads with unusual lengths, which are predominately around 10 kb. When selecting the most appropriate read length for genome sequencing purposes, there are several important facts to be considered. Short reads acquire favorable features over long reads in certain aspects and long or synthetic long reads possess several other important features which make them more suitable for genome sequencing purposes [44].

Fundamentally, short read lengths are utilized when it requires high-quality deep coverage and higher accuracy within small to large genomes and also it facilitates recovery of various genomic parts with a higher efficiency. According to the novel scientific investigations, it has been discovered that short reads tend to concentrate more on the two arms than the mid-part of the chromosome. Further, the two arms of a chromosome consist of a higher density of single nucleotide polymorphisms (SNPs) and repetitive sequences such as telomeres and therefore, shorter reads are more capable of resolution of those SNPs and repetitive sequences than that of long reads which are mostly concentrated onto the mid-part of the chromosome [45,46]. However, there are certain issues with short read sequencing platforms. Out of them, the most prominent concern is the complexity in genomic assembly due to the redundancy of the genomic sequences and the presence of tandem or genome-wide repeats. Furthermore, short read length limits its capability to resolve complex regions with heterozygous sequences and as a result of that, important biological sequences such as genes or promoter regions are often highly fragmented using short-read sequencing while making other computations like sequencing entire RNA transcripts or entire 16S rRNA gene sequences in metagenomics projects difficult or impossible. Therefore, even though the process is simple at the phase of sequencing, when it comes to the assembly phase it becomes highly complicated with short reads [44].

Long reads or synthetically long reads acquire number of favorable features over short reads and particularly novel genome sequencing projects are carried out utilizing long reads that facilitate the sequence assembly process at a significant level. Moreover, gap finishing of genomes, identification of complex structural variations and certain applications such as genotyping, expression profiling, systematic identification of DNA binding sites *etc.* are considered to be the specific utilities of long read sequencing [47,48]. When considering third generation sequencing which acquires reads of unusual length, even though the genome assembly is greatly facilitated, the error rate of sequencing process is approximately 15% -18%. It indicates that there is an erroneous nucleotide within every nine nucleotide that has been sequenced and it has become the major drawback of this approach. However, in the case of synthetic long read technology, the short reads have been preliminary assembled thus it gives rise to a considerably long reads (approximately 1000 bp) within a short period of time and error rate has been drastically reduced [49]. Therefore, certain platforms which possess synthetically long reads (e.g. "Illumina HiSeq") have become emerging solutions for sequencing novel genomes. Further, *de novo* assembly process has been rapidly boosted via the introduction of third generation sequencing with reads of unusual length and synthetic long reads of second generation sequencers so that the requirement for the novel genome sequences can be efficiently fulfilled.

However, the finished genome or the draft genome sequence that has been generated is the final output of a combination of pieces of information generated via different sequencing platforms possessing short and long sequencing read.

1.5. Conclusion

Model organisms play a huge role in biological and biochemical research projects either by acting as an effective template or else providing the information about basal metabolism. Therefore, acquiring a model organism together with the complete genome sequence has become an added advantage for research projects to be conducted successfully. As a result of that, scientists are highly keen on discovering model organisms for particular disease conditions and sequencing their whole genome in order to unravel the biological and biochemical aspects to facilitate most of downstream processes such as drug designing.

With the development of technology, complete genome has become a readily approachable property of an organism and genome sequencing can be done with a greater accuracy and precision. Furthermore, emergence of paired end sequencing and sequencing platforms acquiring higher depths of coverage has revolutionized the field of genomics while improving the quality and the validity of the resulted genome sequences. Moreover, sequence assembly process has been dramatically improved with the advancement of computer science and currently novel assembly algorithms with a greater capacity are being emerged. However, it is essential to perform quality assessment and validation procedures followed by each and every genome project thus it confirms the overall quality and validity of the sequenced genome.

Finally, whole genome sequence has become an emerging requirement of an organism which acquires a higher plausibility of being a model, thus it bestows the complete biological and biochemical insight to the particular organism. Further, when selecting the sequencing platforms there are several facts that should be necessarily considered such that sequencing depth of coverage, read length of the platform, assembly algorithms, quality assessment and validation procedures *etc.* However, regardless of the throughput of the sequencing platform, the accuracy and precision of resulted genome sequence has a significant influence on most of the downstream processes such as investigation of SNPs, designing primers for polymerase chain reactions (PCRs) etc.

References

[1] K. Gull, The biology of kinetoplastid parasites: Insights and challenges from genomics and post-genomics, Int. J. Parasitol. 31 (2001) 443–452. doi:10.1016/S0020-7519(01)00154-0.
[2] N.J. Gnerre, C. Constantinidou, J.Z.M. Chan, M. Halachev, M. Sergeant, C.W. Penn, E.R. Robinson, M.J. Pallen, High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity, Nat. Rev. Microbiol. 10 (2012) 599–606. doi:10.1038/nrmicro2850.
[3] DNA Sequencing Technologies Key to the Human Genome Project, (n.d.).
[4] M. Baker, De novo genome assembly: what every biologist should know, Nat. Methods. 9 (2012) 333–337. doi:10.1038/nmeth.1935.
[5] L.T.C. França, E. Carrilho, T.B.L. Kist, A review of DNA sequencing techniques., Q. Rev. Biophys. 35 (2002) 169–200. doi:10.1017/S0033583502003797.
[6] E.R. Mardis, Next-Generation DNA Sequencing Methods, (2008). doi:10.1146/annurev.genom.9.081307.164359.
[7] C.M. Fraser, J.A. Eisen, K.E. Nelson, I.T. Paulsen, S.L. Salzberg, The Value of Complete Microbial Genome Sequencing (You Get What You Pay For), 184 (2002) 6403–6405. doi:10.1128/JB.184.23.6403.
[8] P.H. Dear, Genome Mapping, Life Sci. (2001) 1–7. doi:10.1038/npg.els.0001467.
[9] G. Myers, Whole genome DNA sequencing, Comput. Sci. Eng. 1 (2002) 33–43. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=764214.
[10] S. Batzoglou, {ARACHNE}: A Whole-Genome Shotgun Assembler, Genome Res. 12 (2002) 177–189. doi:10.1101/gr.208902.7.
[11] E.D. Green, Strategies for the of Complex Genomes, Genetics. 2 (2001) 573–83. doi:10.1038/35084503.
[12] M.L. Metzker, Sequencing technologies — the next generation, Nat. Rev. Genet. 11 (2009) 31–46. doi:10.1038/nrg2626.
[13] K. Scheibye-Alsing, K. Scheibye-Alsing, S. Hoffmann, S. Hoffmann, a Frankel, a Frankel, P. Jensen, P. Jensen, P.F. Stadler, P.F. Stadler, Y. Mang, Y. Mang, N. Tommerup, N. Tommerup, M.J. Gilchrist, M.J. Gilchrist, a.-B. a-B. Nygård, a.-B. a-B. Nygård, S. Cirera, S. Cirera, C.B. Jørgensen, C.B. Jørgensen, M. Fredholm, M. Fredholm, J. Gorodkin, J. Gorodkin, Sequence assembly, Comput. Biol. Chem. 33 (2009) 121–136. doi:http://dx.doi.org/10.1016/j.compbiolchem.2008.11.003.
[14] E.R. Mardis, Next-Generation Sequencing Platforms, Annu. Rev. Anal. Chem. 6 (2013) 287–303. doi:10.1146/annurev-anchem-062012-092628.

[15] D. a Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G.T. Roth, X. Gomes, K. Tartaro, F. Niazi, C.L. Turcotte, G.P. Irzyk, J.R. Lupski, C. Chinault, X. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D.M. Muzny, M. Margulies, G.M. Weinstock, R. a Gibbs, J.M. Rothberg, The complete genome of an individual by massively parallel DNA sequencing., Nature. 452 (2008) 872–6. doi:10.1038/nature06884.

[16] S. Gnerre, I. Maccallum, D. Przybylski, F.J. Ribeiro, J.N. Burton, B.J. Walker, T. Sharpe, G. Hall, T.P. Shea, S. Sykes, A.M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E.S. Lander, D.B. Jaffe, High-quality draft assemblies of mammalian genomes from massively parallel sequence data., Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 1513–8. doi:10.1073/pnas.1017351108.

[17] J. Shendure, H. Ji, Next-generation DNA sequencing, 26 (2008) 1135–1145. doi:10.1038/nbt1486.

[18] E.E. Schadt, S. Turner, A. Kasarskis, A window into third-generation sequencing, Hum. Mol. Genet. 19 (2010) 227–240. doi:10.1093/hmg/ddq416.

[19] H. Lee, J. Gurtowski, S. Yoo, M. Nattestad, S. Marcus, S. Goodwin, W.R. McCombie, M. Schatz, Third-generation sequencing and the future of genomics, bioRxiv. (2016) 48603. doi:10.1101/048603.

[20] S. Srinivasan, J. Batra, Next Generation : Sequencing & Applications Four Generations of Sequencing - Is it Ready for the Clinic Yet ?, Next Gener. Seq. Appl. 1 (2014) 1–7. doi:10.4172/jngsa.1000107.

[21] M. Quail, M.E. 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 Genomics. 13 (2012) 1. doi:10.1186/1471-2164-13-341.

[22] J.M. Prober, G.L. Trainor, R.J. Dam, F.W. Hobbs, C.W. Robertson, R.J. Zagursky, A.J. Cocuzza, M.A. Jensen, K. Baumeister, A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides., Science. 238 (1987) 336–41. http://www.ncbi.nlm.nih.gov/pubmed/2443975 (accessed July 4, 2016).

[23] J. Shendure, R.D. Mitra, C. Varma, G.M. Church, Advanced sequencing technologies: methods and goals., Nat. Rev. Genet. 5 (2004) 335–344. doi:10.1038/nrg1325.

[24] D. Gordon, C. Abajian, P. Green, Consed: A graphical tool for sequence finishing, Genome Res. 8 (1998) 195–202. doi:10.1101/gr.8.3.195.

[25] M. Machado, W.C. Magalhães, A. Sene, B. Araújo, A.C. Faria-Campos, S.J. Chanock, L. Scott, G. Oliveira, E. Tarazona-Santos, M.R. Rodrigues, Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies., Investig. Genet. 2 (2011) 3. doi:10.1186/2041-2223-2-3.

[26] Y. Li, Y. Hu, L. Bolund, J. Wang, State of the art de novo assembly of human genomes from massively parallel sequencing data., Hum. Genomics. 4 (2010) 271–277. doi:10.1186/1479-7364-4-4-271.

[27] H.P.J. Buermans, J.T. Den Dunnen, Next generation sequencing technology: Advances and applications ☆, (2014). doi:10.1016/j.bbadis.2014.06.015.

[28] M. V Olson, The human genome project, Proc. Natl. Acad. Sci. U. S. A. 90 (1993) 4338–4344. doi:10.1073/pnas.90.10.4338.

[29] M.D. Cao, S.H. Nguyen, D. Ganesamoorthy, A.G. Elliott, M. Cooper, L.J.M. Coin, Scaffolding and Completing Genome Assemblies in Real-time with Nanopore Sequencing, bioRxiv. (2016). doi:10.1101/054783.

[30] J.R. Miller, S. Koren, G. Sutton, Assembly algorithm for next-generation sequencing data., Genomics. 95 (2010) 315–327. doi:10.1016/j.ygeno.2010.03.001.Assembly.

[31] 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. Genomics. 11 (2012) 25–37. doi:10.1093/bfgp/elr035.

[32] M.J.P. Chaisson, R.K. Wilson, E.E. Eichler, Genetic variation and the de novo assembly of human genomes, Nat. Rev. Genet. 16 (2015) 627–640. doi:10.1038/nrg3933.

[33] I. Sovic, K. Skala, M. Sikic, Approaches to DNA de novo assembly, Inf. Commun. …. (2013) 351–359. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6596281.

[34] Y.-T. Huang, C.-F. Liao, Integration of string and de Bruijn graphs for genome assembly., Bioinformatics. 32 (2016) 1301–7. doi:10.1093/bioinformatics/btw011.

[35] R. Ekblom, J.B.W. Wolf, A field guide to whole-genome sequencing, assembly and annotation, Evol. Appl. 7 (2014) 1026–1042. doi:10.1111/eva.12178.

[36] S. Meader, L.W. Hillier, D. Locke, C.P. Ponting, G. Lunter, Genome assembly quality: Assessment and improvement using the neutral indel model, Genome Res. 20 (2010) 675–684. doi:10.1101/gr.096966.109.

[37] H.H. Lin, Y.C. Liao, Evaluation and validation of assembling corrected pacbio long reads for microbial genome completion via hybrid approaches, PLoS One. 10 (2015) 1–13. doi:10.1371/journal.pone.0144305.

[38] W. Xiao, L. Wu, G. Yavas, V. Simonyan, B. Ning, H. Hong, Challenges, Solutions, and Quality Metrics of Personal Genome Assembly in Advancing Precision Medicine, Pharmaceutics. 8 (2016) 15. doi:10.3390 /pharma

ceutics8020015.

[39] M. Hunt, T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, T.D. Otto, REAPR: a universal tool for genome assembly evaluation., Genome Biol. 14 (2013) R47. doi:10.1186/gb-2013-14-5-r47.

[40] S. Kurtz, A. Narechania, J.C. Stein, D. Ware, A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes., BMC Genomics. 9 (2008) 517. doi:10.1186/1471-2164-9-517.

[41] M.C. Schatz, J. Witkowski, W.R. McCombie, Current challenges in de novo plant genome sequencing and assembly, Genome Biol. 13 (2012) 243. doi:10.1186/gb4015.

[42] A. Alexeyenko, B. Nystedt, F. Vezzi, E. Sherwood, R. Ye, B. Knudsen, M. Simonsen, B. Turner, P. de Jong, C.-C. Wu, J. Lundeberg, Efficient de novo assembly of large and complex genomes by massively parallel sequencing of Fosmid pools., BMC Genomics. 15 (2014) 439. doi:10.1186/1471-2164-15-439.

[43] D. Sims, I. Sudbery, N.E. Ilott, A. Heger, C.P. Ponting, Sequencing depth and coverage : key considerations in genomic analyses, Nat. Publ. Gr. 15 (2014) 121–132. doi:10.1038/nrg3642.

[44] P.F. Dimond, GEN | Insight & Intelligence^{TM}: The Long and the Short of DNA Sequencing, (n.d.). http://www.genengnews.com /insight-and-intelligenceand153/the-long-and-the-short-of-dna-sequencing/77899725/.

[45] J.C. Dohm, C. Lottaz, T. Borodina, H. Himmelbauer, Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, Nucleic Acids Res. 36 (2008). doi:10.1093/nar/gkn425.

[46] J. a Reinhardt, D. a Baltrus, M.T. Nishimura, W.R. Jeck, C.D. Jones, J.L. Dangl, De novo assembly using low coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae, Genome Res. (2008) 294–305. doi:10.1101/gr.083311.108.

[47] H. Lee, J. Gurtowski, S. Yoo, Error correction and assembly complexity of single molecule sequencing reads, bioRxiv. (2014) 1–17. doi:10.1101/006395.

[48] R.J. Roberts, M.O. Carneiro, M.C. Schatz, The advantages of SMRT sequencing., Genome Biol. 14 (2013) 405. doi:10.1186/gb-2013-14-6-405.

[49] R. Li, C.-L. Hsieh, A. Young, Z. Zhang, X. Ren, Z. Zhao, Illumina Synthetic Long Read Sequencing Allows Recovery of Missing Sequences even in the “Finished” C. elegans Genome., Sci. Rep. 5 (2015) 10814. doi:10.1038/srep10814.

[50] Pitcher, D. G.; Saunders, N. A.; Owen, R. J. Rapid Extraction of Bacterial Genomic DNA with Guanidium Thiocyanate. Lett. Appl. Microbiol. 1989, 8 (4), 151–156 DOI: 10.1111/j.1472-765X.1989.tb00262.x

[51] Wang, T. Y.; Wang, L.; Zhang, J. H.; Dong, W. H. A Simplified Universal Genomic DNA Extraction Protocol Suitable for PCR. Genet. Mol. Res. 2011, 10 (1), 519–525 DOI: 10.4238/vol10-1gmr1055.

[52] Hamid Kheyrodin. DNA Purification and Isolation of Genomic DNA from Bacterial Species by Plasmid Purification System. African J. Agric. Res. 2012, 7 (3), 433–442 DOI: 10.5897/AJAR11.1370.

[53] Ilie, L.; Haider, B.; Molnar, M.; Solis-Oba, R. SAGE: String-Overlap Assembly of GEnomes. BMC Bioinformatics 2014, 15 (1), 302 DOI: 10.1186/1471-2105-15-302.

[54] Butler, J.; MacCallum, I.; Kleber, M.; Shlyakhter, I. a; Belmonte, M. K.; Lander, E. S.; Nusbaum, C.; Jaffe, D. B. ALLPATHS: De Novo Assembly of Whole-Genome Shotgun Microreads. Genome Res. 2008, 18 (5), 810–820 DOI: 10.1101/gr.7337908.

[55] Ye, C.; Ma, Z. S.; Cannon, C. H.; Pop, M.; Yu, D. W. SparseAssembler: De Novo Assembly with the Sparse de Bruijn Graph. Arxiv Prepr. arXiv11062603 2011.