# Sequencing a bacterial genome: an overview

**J. Duan, J.J. Heikkila, B.R. Glick**

Department of Biology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada

Since the first bacterial genome, *Haemophilus influenzae*, was fully sequenced in 1995, over 1000 complete bacterial genome sequences have been determined. DNA sequencing technology has dramatically improved from the first generation, automated Sanger DNA sequencing, which dominated this field for almost two decades, to the current Next-generation sequencing protocols. This newer technology dramatically reduces both the time and cost of DNA sequencing, making it possible for a small laboratory to completely sequence the genome of their favorite bacterium. With the enormous amount of information obtained from whole genome sequencing, scientists can readily address a wide range of biological questions that were hitherto beyond their capabilities. In this chapter, strategies of how to sequence genomic DNA as well as how to assemble and annotate a bacterial genome are reviewed and discussed.

**Keywords** bacterial genomics; next generation sequencing; de novo assembly; bacterial genome annotation

## 1.Genome sequencing

As of May 2010, 1,072 complete published bacterial genomes have been reported in the Genomes Online Database and another 4,289 bacterial genome projects are known to be ongoing (www.genomesonline.org). The underlying reasons for sequencing the genome of various bacteria are either because they are highly virulent to humans, animals or plants, or they can be applied to bioremediation or bioenergy production. In 2009, a new initiative called 'Genomic Encyclopedia of Bacteria and Archaea' (GEBA) was reported by Eisen and colleagues [1]. The project aims to provide a more complete picture of bacterial and archaeal genomic diversity by systematically filling in the gaps in the tree of life [1, 2]. With the assistance of continuously evolving DNA sequencing technologies, the goal of generating reference genomes for every major and minor group of bacteria should be achieved in the near future.

### 1.1 Sanger DNA Sequencing

The Sanger DNA sequencing technique has been an important method in sequencing bacterial genomes. The sequencing chemistry is based on the use of the DNA chain terminator dideoxynucleotide (ddNTP), which is a molecule lacking a hydroxyl group at the 3' carbon of the deoxyribose sugar [3]. During DNA synthesis, an incoming deoxyribonucleotide (dNTP) can form a phosphosdiester bond between its 5' α-phosphate group and the 3' hydroxyl group of the last nucleotide. However, if a dideoxynucleotide is incorporated at the end of the growing strand, DNA chain growth terminates. The four dideoxynucleotides used in Sanger sequencing are labeled with four fluorescent dyes with each dye representing a particular nucleotide [4]. The dye-labeled ddNTPs are added into the reaction mixture containing single-stranded DNA template, primer, DNA polymerase and all four dNTPs. The polymerase chain reaction (PCR) products are separated by capillary electrophoresis according to their masses. Each fluorescent dye emits light following its activation by a laser at the end of the capillary. Therefore, the DNA sequence can be determined by the order of the fluorescent signals [5]. Detailed schematic representations of these steps can be found in [6; Figures 4.20-4.22, 4.26].

Using Sanger DNA sequencing chemistry to sequence the entire genome of an organism is comprised of three major steps: DNA library preparation, template purification and DNA sequencing. For shotgun *de novo* sequencing, DNA is randomly fragmented to generate small (2 Kb) and large (15~20 Kb) fragments which are subsequently cloned into a high-copy-number plasmid. The plasmids are then used to transform *Escherichia coli* cells. After transformation, resultant colonies are transferred into either 96- or 384- well plates. Plasmid purification occurs directly on the plates. Then PCR-based DNA sequencing is performed [7].

After decades of improvement of this methodology, sequences of up to ~1 Kb DNA can be obtained by Sanger sequencing with an accuracy as high as 99.999% [8]. Nevertheless, an individual lab may encounter substantial expense and many months of work to complete the sequencing of a microbial genome. These limitations have encouraged scientists to develop and utilize a variety of new sequencing technologies.

### 1.2 Next-generation DNA sequencing technologies

#### 1.2.1 454/Pyrosequencing

Pyrosequencing was commercialized by Roche/454 in 2005 and was the first next-generation sequencing (NGS) platform on the market [9]. The basis of this technique is the measurement of the release of inorganic pyrophosphate by converting it into visible light during DNA synthesis [10, 11]. The sequencing chemistry consists of a series of

enzymatic reactions. First, pyrophosphate, released from the growing DNA strand, combines with adenosine-5'-phosphosulfate catalyzed by ATP sulfurylase to form ATP. Then, the ATP is used by luciferase to convert luciferin to oxyluciferin to generate light. Before the next nucleotide is added, it is necessary to remove the unused ATP and the unincorporated deoxynucleoside triphosphate; this is done by the enzyme apyrase. In addition, a thio-modified dATP, deoxyadenosine α-thiotriphosphate (dATPαS), is used as a substitute for natural dATP to avoid creating a false positive signal [10, 11].

The general pyrosequencing workflow includes DNA library preparation, emulsion PCR, DNA sequencing and data analysis [9]. Briefly, bacterial genomic DNA is fractionated by nebulization, in which the DNA is forced through a small hole, and DNA fragments in the range of 300- to 800-bp are selected. After DNA repair and end polishing to generate blunt ends, short 3' and 5' DNA adaptors are added to each fragment. In the next step, each fragment is immobilized onto a 28 μm bead which has sulphurylase and luciferase attached to it. PCR amplification is performed within droplets of an oil-water emulsion. As a result, several thousand copies of the same template sequence are generated on each bead. The beads are then deposited into titanium-coated PicoTiterPlate wells. The diameter of the PicoTiterPlate wells are designed to allow for only one bead per well. During pyrosequencing, individual dNTPs are added sequentially in a predetermined order. The amount of light generated is proportional to the number of dNTPs added. The bioluminescent images are recorded by a charge-coupled device (CCD) camera. Because the linear relationship between light intensity and the number of dNTP incorporated can only hold up to 6 nucleotides, pyrosequencing has high error rates (insertions and deletions) in homopolymer repeats [9, 12]. Schematic representations of this approach can be found in [6; Figures 4.29-4.31, 4.38-4.41].

The aforementioned DNA library preparation method generates sequences that can be assembled into a number of unordered and unoriented "contigs". In order to close the gaps in bacterial genome sequences between the contigs, construction of a paired-end DNA library is usually recommended [13]. First, bacterial genomic DNA is sheared randomly and certain size fragments are selected, i.e. 3 Kb, 8 Kb or 20 Kb. The fragments are methylated to avoid EcoRI cleavage, and hairpin adaptors are ligated onto both ends of the DNA. Subsequent exonuclease digestion removes all of the DNA fragments that are not protected by hairpins. In addition, the hairpin adaptors are biotinylated and contain EcoRI recognition sites that are not methylated. Therefore, after digestion with EcoRI, the DNA can be circularized by self-ligation. The EcoRI digestion step also removes the terminal hairpin structures from the DNA. Second, the circularized DNA is fragmented by nebulization and fragments containing the added adaptors are selected using streptavidin, which has very tight biotin-binding capability [14]. Eventually, a DNA library consisting of true paired end reads is generated, with a 44-mer adaptor sequence in the middle, flanking with ~ 150 bp sequences on average. The two flanking 150 bp sequences are fragments of DNA that were originally located approximately 3 Kb, 8 Kb or 20 Kb in the genome of interest. This library is now ready for emulsion PCR and DNA sequencing. Using paired-end reads, scaffolds can be obtained from the ordered and oriented contigs and this greatly facilitates the complete sequencing of the genome.

Currently, the average read length from pyrosequencing is ~ 400 bp. However, Roche/454 has announced that this technology is expected to extend read lengths up to 1,000 bp. In addition, a recently introduced bench top high-throughput sequencing platform should fit the needs of small to medium sized research laboratories.

### 1.2.2 Illumina/Solexa

llumina/Solexa's Genome Analyzer is currently the most widely used DNA sequencing platform. It is based on four-colour cyclic reversible terminators that are blocked at the 3' end using modified nucleotides such as 3'-O-azidomethyl-dNTPs [15]. The sequencing chemistry behind this technique is based on the Sanger DNA sequencing technique [16]. In this procedure, four fluorescently labeled nucleotides are simultaneously added into the reaction mixture and DNA polymerase incorporates the nucleotide that is complementary to the template base. DNA synthesis terminates after the addition of one nucleotide. The unincorporated nucleotides are washed away and fluorescence is recorded to determine the incorporated nucleotide. Then, the 3' blocking group is cleaved from the incorporated nucleotide to restore a 3'-OH group and the fluorophore is also removed from the base. Therefore, two chemical bonds need to be cleaved before the next cycle begins [16]. Also, since the 3' blocked terminators cannot be readily incorporated into the growing DNA strand by the native form of DNA polymerase, a modified form of DNA polymerase, created by site-directed mutagenesis, is used to perform the reaction [17].

The sequencing workflow of the Genome Analyzer includes three steps: DNA library preparation, generation of clonal clusters, and sequencing [16]. Genomic DNA is first fragmented by nebulization or sonication. DNA end-repair is performed in order to generate blunt ended DNA. Following phosphorylation of the 5' and 3' ends, an adenosine overhang is added to each end. This facilitates the ligation between the sequencing adapters and the DNA fragments. Next, a flow cell is used to capture template molecules to generate clonal clusters, which are identical copies of each single DNA template within the diameter of one micron. The flow cell is a silica slide with eight channels and each channel can hold up to 12 samples. Different from emulsion PCR on small beads, in solid-phase amplification denatured DNA templates are covalently attached to a lawn of oligonucleotides immobilized on the flow cell surface. Templates bound to the primers are 3'-extended using a high-fidelity DNA polymerase. After denaturation, the original

 

templates are washed off and the amplified copies are left on the flow cell surface. Because there are adapter oligonucleotides on the free ends of the bound templates, this adapter may hybridize to adjacent lawn primer, which is immobilized on the flow cell surface, to form a bridge. DNA polymerase copies the template from the primer to form a double-stranded DNA bridge, which is subsequently denatured and two single-stranded DNAs may hybridize to adjacent lawn primers to form new bridges. This process is repeated to create millions of dense clonal clusters, each containing about 2,000 molecules. Following denaturation of the double-stranded DNA bridges, the reverse strand is removed by cleavage at the reverse strand-specific lawn primers. The 3'-OH ends are blocked to avoid nonspecific priming and sequencing primers are hybridized to the adapter attached to the unbound ends of the DNA templates. Now the flow cell, which contains clusters of ~1,000 copies of single-stranded DNA molecules, is ready for transfer to the Genome Analyzer for sequencing [16]. Detailed diagrams of the procedure can be found in [18; Figure 2.2].

Illumina/Solexa offers three strategies to prepare a DNA library, including single-read, paired-end and mate pair. The single-read method is described in the sequencing workflow section of this review. Mate pair library preparation is essentially the same as paired end library preparation in pyrosequencing, except that in mate pair protocols the templates to be sequenced are separated by 2-5 Kb inserts instead of 3 Kb, 8 Kb or 20 Kb. The protocol for paired-end library preparation for the Illumina Genome Analyzer is completely different from mate pair DNA library preparation. It generates twice the amount of sequencing data compared with single-read and it also requires twice the run time. In addition, another instrument, a Paired-End Module, needs to be attached to the Genome Analyzer. End users may choose the length of the sequencing insert ranging from 200 to 500 bp. During paired-end library sequencing, the forward strand of the DNA template is sequenced in the same way as in single-read sequencing. After denaturation, the newly synthesized partial strand is removed and the 3' ends are unblocked. The free ends can bind to lawn primers to reform bridges and double-stranded DNA clusters are regenerated. This time, the forward strands are cleaved leaving only the newly synthesized reverse strands attached to the flow cell. Subsequent sequencing is performed on the reverse strands to produce paired end data.

Currently, Illumina/Solexa provides three sequencing systems, Genome Analyzer$_{IIe}$, Genome Analyzer$_{IIx}$ and HiSeq 2000. Sequencing read length can be chosen from 35, 50, 75 and 100 base pair, single or paired-end. With this system, the most common error during sequencing is substitutions, especially after a 'G' base [19]. Amplification bias during template preparation could also cause underrepresentation of AT-rich and GC-rich regions [12, 19, 20, 21].

### 1.2.3 Single-Molecule Sequencing

Single-molecule sequencing has the advantage of not requiring amplification of the templates by PCR before sequencing, since clonal amplification of templates may introduce errors. HeliScope developed by Helicos BioSciences was the first commercialized single-molecule sequencer [22]. This technology significantly increased the speed of DNA sequencing, while decreasing the cost.

The HeliScope uses Virtual Terminators, which are 3'-unblocked cyclic reversible terminators [23]. The inhibiting group is a nucleoside analogue that is directly attached to the fluorophore. Using a 3'-unblocked terminator is highly efficient because removal of the fluorophore and terminating group is combined into one step. Furthermore, it is no longer necessary to screen large libraries of mutant DNA polymerase since 3'-unblocked terminators can be incorporated into the growing strand DNA effectively by wild-type DNA polymerase.

The workflow of single-molecule sequencing may be summarized as following [24]. A DNA sample is sheared into short strands of about 100 to 200 nucleotides in length before a poly-A universal priming sequence is added to the 3' end of each DNA strand which is then labeled with a fluorescent adenosine nucleotide. The labeled strands serve as templates for the single molecule sequencing chemistry. The DNA strands are hybridized to the Helico's flow cell which contains billions of oligo-T universal capture sites that are immobilized on the flow cell surface. Because the HeliScope sequencer detects single molecules, the templates can be packed at very high density, i.e. billions of templates per run. After the DNA sequences have been hybridized to the flow cell surface, they are loaded into the Heliscope instrument. A laser illuminates the surface of the flow cell, highlighting the location of each fluorescently labeled template. A CCD camera then produces a map of the template on the flow cell surface. After the template has been imaged, the template label is cleaved and washed away. The sequencing reaction begins by introducing a DNA polymerase and a fluorescently (Cy5) labeled nucleotide with the oligo-T capture sites serving as sequencing primers. DNA polymerase catalyzes the addition of Cy5-labeled nucleotides to the primers in a template directed manner. A washing step then removes the polymerase and any unincorporated nucleotides. The billions of single molecule templates that have incorporated a particular nucleotide are then visualized by illuminating and imaging the entire flow cell surface. After imaging, the fluorescent labels are cleaved and removed. The process continues with each of the remaining bases and repeats until the desired read length has been achieved. Unlike amplification-based sequencing technologies, the single-molecule sequencing process is asynchronous. Every strand is unique and is sequenced independently. Illustration of this sequencing approach is shown in [12; Figure 2].

Paired-end reads can be obtained from individual single molecules as well (www.helicosbio.com). Unlike the traditional paired-end library preparation, the procedure does not involve cloning, circularization and digestion of the sheared genomic DNA sample. Briefly, after fragmentation of genomic DNA, an adaptor sequence is ligated to the 5' ends of the fragments. Then poly-A tails which will hybridize to the poly-T immobilized on the flow cell surface are

created on the 3' ends of the fragments. Following the completion of sequencing the 3' end of the template, the template is copied to the end by DNA polymerase and all four natural nucleotides. The template DNA is removed by denaturation, leaving only the reverse strand bound to the flow cell surface. A universal primer can then hybridize to the adaptor sequence which is at the free end of the DNA fragment and sequencing can be performed from the free end of DNA template, i.e. the 3' end of the reverse strand.

Currently, the average single read length that the HeliScope procedure generates is 35 bp. The error rate for substitution, insertion and deletion are 0.2%, 1.5%, and 3.0%, respectively.

#### 1.2.4 Sequencing by ligation

In contrast to DNA sequencing by synthesis, sequencing by ligation uses DNA ligase to determine the identity of a nucleotide in a DNA sequence. DNA ligase can join two DNA strands that have a double-strand break. It can also link the ends on only one of the two strands, providing that the incoming single strand nucleotides are perfectly complementary to the reverse strand [25].

The technology of sequencing by ligation was commercialized by Applied Biosystems in 2007; its platform is named support oligonucleotide ligation detection (SOLiD) [26]. DNA templates are prepared in a manner similar to pyrosequencing technology. The DNA is sheared by nebulization or sonication, and ligated to oligonucleotide adapters. One adapter is then hybridized to another adapter, called the P1 adapter, which is immobilized onto one-micron diameter paramagnetic beads. The DNA library is diluted before the hybridization between the two adapters to ensure that only one DNA template attaches to each bead. In the next step, DNA templates are clonally amplified by emulsion PCR, followed by bead purification. In contrast to pyrosequencing where a PicoTiterPlate is used to catch the beads, in the present technique a flow cell glass slide is used. Before bead deposition, it is necessary to modify the 3' end of the DNA template on the beads to allow its covalent attachment to the slide. Because there is another adapter attached to the free end of the DNA template, the modification can be made by attaching a polystyrene bead which has complementary adapter sequences on its surface. Currently, three types of slides are available, which allow for the analysis of 1, 4, or 8 samples on a single slide. In order to further extend the throughput, barcoding is introduced and up to 16 libraries can be loaded on one region of an 8 region glass slide. The SOLiD system can hold two independent flow cell slides at once. Therefore, up to 256 samples can be analyzed in a single run.

Three types of DNA libraries can be used for sequencing: a fragment library, a mate-paired library (insert size from 600 bp ~ 10 Kb) and a paired-end library. The strategies that are used to construct fragment and mate-paired libraries are similar to those are used in the Illumina/Solexa technology, and up to 75 base pair read lengths can be generated from the SOLiD system. Paired-end library sequencing involves sequencing of both the forward and reverse direction of DNA templates using DNA ligase and provides read lengths of up to 35 base pairs at the 5' and 3' ends.

The first step of sequencing by ligation is to ligate a probe to a sequencing primer [26]. Each probe consists of eight base pairs (octamer), of which the first two at the 3' end are the ones providing the measuring information. The remaining six nucleotides are degenerate nucleotides with one of four fluorescent labels linked to the 5' end. Since dinucleotides can generate sixteen different combinations, and only four colors are used for measurement, a two-base encoding strategy is employed. For instance, blue represents the combination of AA, CC, GG or TT; green represents the combination of CA, AC, TG or GT; yellow corresponds to the combination of GA, AG, TC or CT; and red corresponds to the combination of TA, AT, GC or CG. As long as the first nucleotide is known, the second nucleotide can be determined based on the color observed. In the first sequencing step, the probes representing all 16 possible two-base combinations are added into the reaction mixture. Annealing only occurs when the probe is complementary to the sequences immediately adjacent to the sequencing primer. Then ligation is performed and the unbound probes are washed away. In the next step, unextended reactions are capped by dephosphorylation, making them unavailable to participate the future reactions. The last three bases and the fluorescent moiety from the probes are then cleaved with $AgNO_3$. Now the probe is reduced to 5 nucleotides with a free phosphate group. Ligation is repeated up to 15 cycles to obtain a sequence of 75 base pairs. Fifteen cycles of ligation is referred to as a "round". Primer reset is carried out where the extended sequences melt off the template and a new primer which is one base inset closer to the bead than the starting primer is hybridized to the adapter. Then the same set of probes is used to measure different pairs of dinucleotides. Primers are reset for five rounds in total and each new primer has a successive offset, i.e. n-1, n-2 and so on. Using this approach, each nucleotide on the template is sequenced twice by different dye-labeled probes, thereby reducing sequencing errors dramatically. Eventually, 75 color space sequence information is collected which will be taken forward to obtain a 75 nucleotides sequence. In order to convert the color space sequence to a base pair sequence, the first nucleotide has to be known. This information can be easily obtained from the first cycle of the second round of sequencing because the first base pair of probe is the complement of the last nucleotide of the sequencing primer. Diagrammatic representation of this procedure is found in [12; Figure 3].

In order to obtain paired-end or mate-pair sequences, sequencing of the reverse direction of the template is performed [26]. First, 3'-hydroxylated primer is annealed to the adapter region of the templates and probes that are 5' phosphorylated are ligated to the primer. To prevent dephasing, primer that is unextended is capped by a ddNTP which is introduced by polymerase. After cleavage by $AgNO_3$, only 5 nucleotides from the probe will remain. The 3' phosphate is removed and the cycles are repeated until the desired read length is obtained.

 

Applied Biosystems provide three next generation sequencing instruments and they are the SOLiD 4 System, SOLiD 4hq System and SOLiD PI System. Among the three systems, SOLiD PI system is a bench top instrument and designed for smaller laboratories.

Similar to Genome Analyzer of Illumina/Solexa, the most common error type created by SOLiD is substitution [21]. In addition, SOLiD data also reveals an underrepresentation of AT-rich and GC-rich regions [21].

#### 1.2.5 Single Molecule Real Time Sequencing

Single molecule real time (SMRT) sequencing is a sequencing by synthesis technology that has been developed quite recently [27]. The technology involves monitoring the incorporation of fluorescent dye-labeled nucleotides continuously during DNA synthesis [27].

The SMRT DNA sequencing system designed by Pacific Biosciences was released to a small group of research institutes in February, 2010. The official commercial launch of the machine is expected to be in the second half of 2010. The most attractive feature of this sequencing system is that it can generate more than a thousand base pairs of sequence information in fast cycle times, since DNA polymerase synthesizes DNA continuously without termination [12, 27, 28].

Sample preparation is unique for SMRT sequencing. The DNA sample is first sheared to the desired size; the ends of the fragments are then repaired so that a hairpin structure can be ligated to each end. Following purification, the fragments with hairpin adaptors are selected and ready for sequencing. In contrast to other sequencing by synthesis approaches in which templates are attached to a solid surface, in SMRT sequencing, single DNA polymerase molecules are immobilized in a nanoscale well [29, 30]. Parallel sequencing is achieved using a chip containing thousands of wells to capture individual DNA polymerase molecules [27]. A modified $\phi$29 DNA polymerase is chosen for this sequencing platform because it can incorporate phospholinked dNTPs efficiently into the growing strand. In addition, $\phi$29 polymerase is able to synthesize DNA in a strand-displacement manner so that the template can be sequenced multiple times to ensure accuracy. With this approach, read accuracy was improved from <80% for a 499 bp template to >99% by circular consensus sequencing for 15 times or more [27]. The nucleotides used in SMRT are phospholinked hexaphosphate nucleotides and have a distinct character in that a fluorophore is linked to the terminal phosphate rather than to the base. The formation of phosphodiester bond leads to the release of the dye-labeled pentaphosphate and the signal is recorded immediately before it diffuses away [27]. A schematic representation of this sequencing method is shown in [12; Figure 4].

It has been reported that using the SMRT system to sequence an *E. coli* genome, it is possible to achieve 99.3% genome coverage with average read lengths of 964 bp and at high accuracy, i.e. >99.999% [12]. Although the technology is still in its infancy and requires additional improvements to accommodate large genome sequencing, the current SMRT platform satisfies the needs for sequencing small viral and bacterial genomes.

Due to the intrinsic limitations and biases of each of the currently available sequencing technologies, it has been suggested that using a combined sequencing strategies is more practical, considering both the quality of sequencing results and cost [31, 32, 33]. In addition, for a genome having a large number of repetitive regions, the use of paired-ends or mate-pairs is necessary because the addition of a large amount of relatively short reads won't help to reduce the gaps between sequenced regions.

## 2. Bacterial genome assembly

Currently, sequencing reads generated by different sequencing technologies range from 35-1000 bp. In order to obtain a complete bacterial genome, the fragments need to be aligned and joined together using a computer program called an assembler. Assemblers can join sequences together based on overlapping regions between the sequences, assuming that the two sequence reads have originated from the same place in the genome. After assembly, a collection of contiguous pieces (contigs) instead of an entire chromosome are usually obtained. This is often due to non-random shearing of DNA, intrinsic cloning bias and repeated regions in the genome. Increasing the sequencing coverage of the genome will help to reduce the number of contigs. For example, Lander and Waterman [34] showed that sequencing a 1 Mbp genome using Sanger chemistry resulted in a small number of contigs (~5) if 8 to 10 times genome coverage was attained.

In addition to the assembler, another computer program called a scaffolder is used if paired-end or mate-pair reads are available. Scaffolder can link distant sequences together based on the distance between the two ends of the original template, i.e. 3 kb or 8 kb. Therefore the scaffolder is able to define the size of gaps between contigs and orient the contigs into a draft genome [35].

To assemble next-generation sequencing reads, *de novo* assembly becomes more of a challenge because only short overlaps can be considered. Therefore, higher coverage of the genome is required for assembly of shorter reads, resulting in large volumes of data [36]. However, one study has shown that it is possible to assemble a large portion of the *E. coli* K12 MG1655 genome using read lengths of 20-50 nucleotides, given that the reads were error-free [37]. Since then, several bacterial genomes have been *de novo* assembled using short read sequence data, including *Helicobacter acinonychis* [38, 39], *Staphylococcus aureus* [39], *Bacillus subtilis* [40, 41], *Pseudomonas aeruginosa*

[42], *Pseudomonas syringae* [43, 44], and *Erwinia pyrifoliae* [45]. Furthermore, combining the read data from more than one type of sequencing platform can improve *de novo* assembly of a bacterial genome significantly [33, 43, 45].

## 3. Bacterial genome annotation

In general, genome annotation can be performed at two main levels, static and dynamic [46]. At the static level, the main features of the genome can be obtained, such as protein coding genes, functional RNA products, GC content, codon usage, genomic islands, motifs, chemical and structural properties of proteins and their subcellular localisation. In contrast, the dynamic view can exhibit gene context, gene order, regulatory networks, protein interaction networks, metabolic networks as well as phyloprofile and gene fusion/fission information obtained by comparative genomics [46]. Nowadays, bacterial genomes are often annotated using automated pipelines, which can be either web-based or run locally [46, 47]. Web-based annotation pipelines are more convenient for small research groups or smaller sequencing facilities that lack computing resources and expertise that is necessary to maintain or implement the software [47]. As an example, BASys (**B**acterial **A**nnotation **Sys**tem) is a automated bacterial annotation web server (http://wishart.biology.ualberta.ca/basys), which uses more than 30 programs to determine nearly 60 annotation subfields for each gene [48]. With this system, results can be generated in about 24 h for a 5 megabase genome and can be browsed and evaluated using a navigable graphical map. However, BASys is not able to analyze partially assembled genomes [48]. In 2008, the RAST (**R**apid **A**nnotation using **S**ubsystem **T**echnology) Server (http://rast.nmpdr.org/), a fully automated annotation service for complete or draft archaeal and bacterial genome, was built [49]. Annotations provided by the service include protein-encoding, rRNA and tRNA genes, gene function and metabolic network. On the completion of annotation, the annotated genome can be downloaded in a variety of formats or browsed in SEED-Viewer for up to 120 days [49]. The SEED is an annotation/analysis tool provided by the Fellowship for Interpretation of Genomes (FIG) [50]. More recently, a metagenomics RAST server was constructed (mg-RAST) [51]. Mg-RAST (http://metagenomics.nmpdr.org/) is used to generate pylogenetic and functional summaries from metagenome data. The user can also perform comparative metagenomic analyse using the tools incorporated into the annotation pipeline [51]. WeGAS (http://ns.smallsoft.co.kr:8051) is another web-based microbial genome annotation system [52]. Like the RAST Server, it is capable of handling both ongoing and completed microbial genome projects. The user can start genome annotation with contigs and the process can be monitored during each analysis. The annotation pipeline includes seven major modules, which are gene prediction, homology search, promoter search, pathway mapping, motif search, COG (Clusters of Orthologous Groups of proteins) assignment and GO (Gene Ontology) assignment. A genome browser is also available to view the detailed results [52]. Recently, another prokaryotic genome annotation web server called Integrative Services for Genomics Analysis (ISGA) became available to the community (http://isga.cgb.indiana.edu/) [53]. Unlike many other web-based annotation servers, ISGA can be installed on a desktop computer in order to better serve the needs of scientists. This is particularly beneficial because the increasing demand of web users may eventually cause deteriorating performance of public web servers [53].

## Summary

In this chapter, an overview of genome sequencing technologies, genome assembly strategies as well as bacterial genome annotation approaches have been discussed. This is not intended to provide an exhaustive review of each topic, but rather focus on the strategies that may be useful for a small microbial laboratory that is interested in sequencing bacterial genomes. Currently, sequencing cost and sequence quality are still the main concerns of bacterial genome sequencing. However, with constantly developing sequencing technology, assembly algorithms, and annotation software, it is possible that in the near future, sequencing, assembling and annotating bacterial genomes will become a routine procedure in every microbial laboratory.

**Acknowledgements** This research was supported by Natural Sciences and Engineering Research Council Grants to B.R.G. and J.J.H.

## References

[1] Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng JF, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk HP, Eisen JA. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*. 2009;462:1056-1060.

[2] Williams D, Gogarten JP, Lapierre P. Filling the gaps in the genomic landscape. *Genome Biology*. 2010;11:103.

[3] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*. 1977;74:5463-5467.

[4] Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE. Fluorescence detection in automated DNA sequence analysis. *Nature*. 1986;321:674-679.

 

[5] Swerdlow H, Wu SL, Harke H, Dovichi NJ. Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *Journal of Chromatography*. 1990;516:61-67.

[6] Glick BR, Pasternak JJ, Patten CL. *Molecular Biotechnology: Principles and Applications of Recombinant DNA*. 4th ed. Washington, DC: American Society for Microbiology Press; 2010:118-120, 124, 127, 129, 138-141.

[7] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC. Whole-Genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995;269:496-498+507-512.

[8] Shendure J, Ji H. Next-generation DNA sequencing. *Nature Biotechnology*. 2008;26:1135-1145.

[9] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picoliter reactors. *Nature*. 2005;437:376-380.

[10] Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*. 1996;242:84-89.

[11] Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. *Science*. 1998;281:363-365.

[12] Metzker ML. Sequencing technologies-the next generation. *Nature Reviews Genetics*. 2010;11:31-46.

[13] Jarvie T, Harkins T. 3K long-tag paired end sequencing with the Genome Sequencer FLX system. *Nature Methods*. 2008;5:i-ii

[14] Chaiet L, Wolf FJ. The properties of streptavidin, a biotin-binding protein produced by *Streptomycetes*. *Archives of Biochemistry and Biophysics*. 1964;106:1-5

[15] Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, Marma MS, Meng Q, Cao H, Li X, Shi S, Yu L, Kalachlkov S, Russo JJ, Turro NJ, Ju J. Four-color DNA sequencing with 3'-*O*-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences*. 2008;105:9145-9150.

[16] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Cheetham RK, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ. Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, BuermannDH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi MCE, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Furey WS, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ng BL, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Pinkard DC, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Rodriguez AC, Roe PM, Rogers J, Bacigalupo MCR, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Sohna JES, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, vandeVondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008:456:53-59.

[17] Chen F, Gaucher EA, Leal NA, Hutter D, Havemann SA, Govindarajan S, Ortlund EA, Benner SA. Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. *Proceedings of the National Academy of Sciences*. 2010;107:1948-1953.

[18] Lakdawalla A, VanSteenhouse H. Illumina Genome Analyzer II System. In: Janitz M, eds. *Next-Generation Genome Sequencing: Towards Personalized Medicine*. Weinheim, Germany: Wiley-VCH; 2008:18.

[19] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*. 2008;36:e105.

[20] Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis E R. Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods*. 2008;5:183-188.

[21] Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*. 2009;10:R32.

[22] Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences*. 2003;100:3960-3964.

[23] Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D, Lowman GM, Marappan S, McInerney P, Platt A, Roy A, Siddiqi SM, Steinmann K, Thompson JF. Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods*. 2009;6:593-595.

[24] Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, DiMeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z. Single-molecule DNA sequencing of a viral genome. *Science*. 2008;320:106-109.

[25] Tomkinson AE, Vijayakumar S, Pascal JM, Ellenberger T. DNA ligases: structure, reaction mechanism and function. *Chemical Reviews*. 2006;106:687-699.

[26] Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research*. 2008;18:1051-1063.

[27] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Viecelli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323:133-138.

[28] Metzker ML. Sequencing in real time. *Nature Biotechnology*. 2009;27:150-151.

[29] Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*. 2003;299:682-686.

[30] Foquet M, Samiee KT, Kong X, Chauduri BP, Lundquist PM, Turner SW, Freudenthal J, Roitman DB. Improved fabrication of zero-mode waveguides for single-molecule detection. *Journal of Applied Physics*. 2008;103:034301.

[31] Goldberg SMD, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R, Halpern A, Khouri H, Kravitz SA, Lauro FM, Li K, Rogers YH, Strausberg R, Sutton G, Tallon L, Thoman T, Venter E, Frazier M, Venter JC. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences*. 2006;103:11240-11245.

[32] McCutcheon JP, Moran NA. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proceedings of the National Academy of Sciences*. 2007;104:19392-19397.

[33] Aury JM, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, Poulain J, Anthouard V, Scarpelli C, Artiguenave F, Wincker P. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics*. 2008;9:603.

[34] Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*. 1988;2:231-239.

[35] Pop M, Kosack DS, Salzberg SL. Hierarchical scaffolding with Bambus. *Genome Research*. 2004;14:149-159.

[36] Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics*. 2010;95:315-327.

[37] Whiteford N, Haslam N, Weber G, Prügel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C. An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*. 2005;33:e171.

[38] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Research*. 2007;17:1697-1706.

[39] Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research*. 2008;18:802-809.

[40] Srivatsan J, Han Y, Peng J, Tehranchi AK, Gibbs R, Wang JD, Chen R. High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genetics*. 2008;4:e1000139.

[41] Nishito Y, Osana Y, Hachiya T, Popendorf K, Toyoda A, Fujiyama A, Itaya M, Sakakibara Y. Whole genome assembly of a natto production strain *Bacillus subtilis* natto from very short read data. *BMC Genomics*. 2010;11:243.

[42] Salzberg SL, Sommer DD, Puiu D, Lee VT. Gene-boosted assembly of a novel bacterial genome from very short reads. *PLoS Computational Biology*. 2008;4:e1000186.

[43] Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD, Dangl JL. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Research*. 2009;19:294-305.

[44] Studholme DJ, Ibanez SG, MacLean D, Dangl JL, Chang JH, Rathjen JP. A draft genome sequence and functional screen reveals the repertoire of type III secreted proteins of *Pseudomonas syringae* pathovar *tabaci* 11528. *BMC Genomics*. 2009;10:395.

[45] Smits THM, Jaenicke S, Rezzonico F, Kamber T, Goesmann A, Frey JE, Duffy B. Complete genome sequence of the fire blight pathogen *Erwinia pyrifoliae* DSM 12163$^T$ and comparative genomic insights into plant pathogenicity. *BMC Genomics*. 2010;11:2.

[46] Médigue C, Moszer I. Annotation, comparison and databases for hundreds of bacterial genomes. *Research in Micribiology*. 2007;158:724-736.

[47] Stothard P, Wishart DS. Automated bacterial genome analysis and annotation. *Current Opinion in Microbiology*. 2006;9:505-510.

[48] Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Research*. 2005;33:W455-459.

[49] Aziz, RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. The RAST server: rapid annotations using subsystems technology. *BMC Genomics*. 2008;9:75.

[50] Overbeek, R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*. 2005;33:5691-5702.

 

[51] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. 2008;9:386.
[52] Lee D, Seo H, Park C, Park K. WeGAS: a web-based microbial genome annotation system. *Bioscience, Biotechnology, and Biochemistry*. 2009;73:213-216.
[53] Hemmerich C, Buechlein A, Podicheti R, Revanna KV, Dong Q. An Ergatis-based prokaryotic genome annotation web server. *Bioinformatics*. 2010;26:1122-1124.