

Next Generation Sequencing Technologies and Their Applications

Ku Chee-Seng, Centre for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Loy En Yun, Centre for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Pawitan Yudi, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Chia Kee-Seng, Centre for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Advanced article

Article Contents

- Introduction
- Revolution in the Approaches for Genomics Studies
- Next Generation Sequencing Technologies
- Applications in Structural and Functional Genomics Studies
- Future Perspectives and Summary

Online posting date: 19th April 2010

The advances in next generation sequencing (NGS) technologies have tremendous impacts on the studies of structural and functional genomics. Sequencing-based approaches like ChIP-Seq and RNA-Seq have started taking the place of microarray experiments to study protein–DNA (deoxyribonucleic acid) interactions and transcriptomic profiling, respectively. The arrival of NGS technologies has also enabled several whole human genome resequencing studies to be completed efficiently at an affordable price. The major strengths of NGS technologies are their ultra high-throughput production, characterized by their ability to generate several hundred megabases to tens of gigabases of sequencing data per instrument run, and more importantly, the steep reduction in cost compared to the traditional Sanger sequencing method. Hence, NGS technologies have rapidly become the primary choice for large scale as well as genome-wide sequencing studies. The new sequencing-based approaches to explore structural and functional genomics have produced important information and significantly expanded our knowledge in these areas.

Introduction

The rapid developments in *sequencing* technologies have transformed the approaches in the studies of structural and functional genomics. The studies of structural genomics focus on identifying various genetic variations or mutations, whereas functional genomics studies aim to interrogate and annotate the functional and regulatory elements or sequences in the human genome. The *next generation sequencing (NGS) technologies* have started substituting traditional Sanger sequencing methods in many large scale or genome-wide sequencing studies. These new sequencing technologies have been attracting a considerable amount of interest from researchers since they have been commercially marketed. The major attractions are their ultra high-throughput production, characterized by their ability to simultaneously sequence millions of DNA (deoxyribonucleic acid) fragments and produce gigabases of sequencing data per instrument run, and more importantly, the steep reduction in cost compared to the traditional sequencing method.

Revolution in the Approaches for Genomics Studies

Previously, the molecular genomics studies mainly relied on microarray technologies such as gene expression microarrays and the ChIP-chip method (i.e. *chromatin immunoprecipitation* coupled with microarray) for genome-wide interrogation. However, this was swiftly replaced by sequencing-based methods, namely RNA-Seq (to measure transcripts or ribonucleic acids (RNAs) expression levels) and ChIP-Seq (to study protein–DNA

ELS subject area: Genetics and Disease

How to cite:

Chee-Seng, Ku; En Yun, Loy; Yudi, Pawitan; and Kee-Seng, Chia (April 2010) Next Generation Sequencing Technologies and Their Applications. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.

DOI: 10.1002/9780470015902.a0022508

interactions like identifying transcription factor-binding sites and interrogating histone modifications), respectively (Wang *et al.*, 2009; Park, 2009).

There are a number of limitations in using microarrays compared to sequencing-based methods. For example, conventional microarrays do not allow a truly comprehensive interrogation of the whole genome, because the selection of probes to be synthesized and immobilized on the solid surface of microarrays requires some prior knowledge and reference genome sequences are also needed. The probes are needed to detect and measure the abundance of DNA or RNA targets through hybridization. In other words, microarray-based methods are limited to interrogating those genomic regions that are probed by the microarrays. It is obvious from the conventional gene expression microarray studies where the gene expression levels could not be measured unless there are probes to capture them, and the probes are usually synthesized to capture known annotated protein-coding genes. Therefore unknown transcripts or those transcripts from noncoding sequences in the *transcriptome* could not be assessed. Similarly for ChIP-chip experiments, the DNA fragments that are pulled down by immunoprecipitation would be undetected if no complementary probes are designed to capture them. On the contrary, theoretically sequencing-based approaches are able to capture all the DNA fragments that are isolated by immunoprecipitation (ChIP-Seq), and all the transcripts (coding and noncoding transcripts) that are available in the transcriptome (RNA-Seq) including the low abundance transcripts, if the sequencing depth is sufficient (Wang *et al.*, 2009; Park, 2009).

Likewise in structural genomics studies, microarray-based methods such as comparative genomic hybridization (CGH) and single nucleotide polymorphism (SNP) arrays have poor sensitivity to detect smaller sizes of copy number variations (CNVs) like those of < 10 kb, and these methods are unable to detect copy neutral variations like balanced translocations and inversions. Furthermore, microarray-based methods have limited resolution to define the breakpoints of CNVs and *structural variations*. However, these limitations have been overcome by sequencing-based methods like paired-end mapping (Korbel *et al.*, 2007). These new and innovative sequencing-based approaches to studying structural and functional genomics have produced important information and have significantly expanded our knowledge in each area.

Next Generation Sequencing Technologies

Sanger dideoxynucleotide or chain termination sequencing has been the most widely used sequencing method for the past three decades since it was invented in late 1970s until the first NGS platform was marketed in 2005. Sanger sequencing has been used for various applications such as mutations discovery, genotyping and *serial analysis of gene*

expression (SAGE) for measuring gene expression levels, and more importantly, it was used to complete the Human Genome Project (International Human Genome Sequencing Consortium, 2004). **See also:** [Human Genome Project: Importance in Clinical Genetics](#); [Sequencing the Human Genome: Novel Insights into its Structure and Function](#); [Whole Genome Resequencing and 1000 Genomes Project](#)

Shortly after the first next generation sequencer was introduced by Roche® 454 Life Science, the Genome Sequencer 20 (GS 20) System (it was subsequently replaced by GS FLX System with further improvements, i.e. higher throughput and longer sequence read length, to the preceding system), another two biotechnology companies also marketed their sequencing platforms: Illumina® Genome Analyzer (GA) and Applied Biosystems® (ABI) Supported Oligonucleotide Ligation Detection System (SOLiD). The simultaneous advent of several next generation sequencers created intense competition in the sequencing market; with each technology having its own strengths and limitations. This article focuses on the NGS technologies because they have been widely used for various applications unlike the newer third generation sequencing instrument, the Heliscope Single Molecule Sequencer, which has only recently been introduced. The following sections described the main features of NGS technologies.

Sequencing throughput and cost

Currently, Sanger sequencing machines (e.g. ABI® 3730xl) have been largely supplanted by next generation sequencers in many large genomics institutes worldwide. This was mainly due to the ultra high-throughput production of NGS technologies which is several orders of magnitude higher than Sanger sequencing method. One of the major differences between modern and traditional sequencing is the ability of next generation sequencers to simultaneously sequence one million to several hundred millions of DNA fragments in contrast to the 96-capillary Sanger sequencer. Therefore, NGS is also known as massively parallel sequencing technologies. This feature has enormously increased the amount of the production or the number of nucleotides or bases that it can sequence compared to the Sanger sequencer in one experiment or per instrument run. For example, the latest developments in Illumina® GA and ABI® SOLiD have further increased the throughput production generating more than 10 gigabases of sequencing data per instrument run in a few days, whereas Roche® GS FLX can generate several hundred megabases per run in 10 h. In contrast, Sanger sequencer like ABI® 3730xl which is commonly used in most of the research laboratories can only produce ~100 kb per run in 3 h (see [Table 1](#) for the summary of the features of NGS technologies) (Shendure and Ji, 2008; Tucker *et al.*, 2009).

The sequencing chemistry of NGS technologies together with their ultra high-throughput production has also reduced the sequencing cost significantly, making large-

Table 1 Summary of the features of NGS technologies

Feature	Roche® 454 GS FLX	Illumina® GA	ABI® SOLiD
The year of the first sequencer that commercially marketed	2005	2006	2007
Current generation of the sequencer	Roche® 454 GS FLX Titanium	Illumina® GA II	ABI® SOLiD 3.0
Massively parallel sequencing (number of DNA fragments)	Several hundred thousand to one million	Several hundred millions	Several hundred millions
Sequencing throughput per instrument run	Several hundred megabases per run in 10 h	> 10 Gb per run in a few days	> 10 Gb per run in a few days
Sequencing cost per megabase (US\$)	~ \$80	~ \$6	~ \$6
Differences in cost in relative to Sanger sequencing (\$500 per megabase)	~ 6-fold	~ 80-fold	~ 80-fold
<i>In vitro</i> amplification method	Emulsion PCR	Bridge amplification on solid surface	Emulsion PCR
Sequencing approach	Sequencing by synthesis mediated by polymerase – pyrosequencing	Sequencing by synthesis mediated by polymerase – sequencing by reversible terminator chemistry	Sequencing by ligation of dinucleotide probes mediated by ligase
Sequencing reagent	Four types of dNTPs	Four types of ddNTPs labelled by four different fluorescent colours	16 types of dinucleotide probes labeled by 4 different fluorescent colours
Detection method of the incorporated nucleotides	Emission of chemiluminescent light	Fluorescent colours	Fluorescent colors
Sequence read length	400–500 bases	75–125 bases	50 bases
Read base or base calling error rate	0.5–1.5%	0.2–2%	< 0.1%
Error type	Insertion or deletion of nucleotides in homopolymer sequences	Substitution of nucleotides	Substitution of nucleotides

scale sequencing studies affordable nowadays. Currently, Illumina® GA and ABI® SOLiD have already achieved a sequencing cost of \$6 per megabase as compared to Roche® GS FLX, which is offered at \$80 per megabase. In general, the sequencing cost of NGS technologies was substantially decreased by several folds to nearly 100-fold compared to Sanger sequencing, which costs about \$500 for the same amount of sequencing data (Shendure and Ji, 2008; Tucker *et al.*, 2009). It is noteworthy that the cost of sequencing is changing continuously; therefore the prices cited here may not be the latest in the market. Regardless, this provides some useful information on differences in sequencing cost between Sanger sequencing and NGS. Undoubtedly, both sequencing production and cost would be continuously improved. The developments of third generation sequencing technologies are on the horizon and the instruments are expected to be marketed soon which would certainly decrease the sequencing cost further and eventually achieve the ultimate goal of \$1000 per genome sequencing (Von Bubnoff, 2008).

On top of the considerations of sequencing throughput and cost, the other concern is logistics. As the amount of sequencing data produced by a next generation sequencer is equivalent to tens of Sanger sequencers, a large area or space would be needed to accommodate the instruments. This can only be feasibly attained by large genomics laboratories or institutes. Furthermore, the maintenance of tens of sequencing instruments will also be substantial and this has not taken into account costs of labour or manpower to operate the instruments.

Sequencing chemistry: *in vitro* amplification

The advances in sequencing technologies have enabled several whole human diploid genome resequencing studies to be completed efficiently. Besides the genome of James Watson (Wheeler *et al.*, 2008), several genomes of anonymous individuals have also been sequenced; they are two Koreans (AK1 and SJK) and one individual each of

Han Chinese (YH), African (NA18507) and European (P0) ancestries (Kim *et al.*, 2009; Ahn *et al.*, 2009; Wang *et al.*, 2008; Bentley *et al.*, 2008; Mckernan *et al.*, 2009; Pushkarev *et al.*, 2009). All these genomes were sequenced by NGS technologies except the genome of the European individual P0 which was sequenced by Heliscope Single Molecule Sequencer. In contrast, the diploid genome of Craig Venter was sequenced by the Sanger sequencing method (Levy *et al.*, 2007). The whole genome resequencing studies using next and third generation sequencing technologies were completed at a cost of tens of thousands to several hundred thousands of dollars compared to Venter's genome which cost millions of dollars.

One of the major limitations in whole genome resequencing using the Sanger sequencing method is the *in vivo* amplification of DNA fragments using bacterial cloning. This is unlike targeted sequencing studies, where conventional polymerase chain reaction (PCR) is commonly used to amplify the regions of interest to be sequenced. The bacterial cloning procedures can introduce host cloning-related biases; for example, it could affect the genome representation in the sequencing of organism genomes because some of the DNA fragments failed to be cloned. Moreover, these steps are tedious and labour intensive. However, this method has since been eliminated and is replaced by the *in vitro* amplification of millions of DNA fragments simultaneously by NGS technologies, that is emulsion PCR for Roche® GS FLX and ABI® SOLiD, and bridge amplification on solid surface for Illumina® GA (Mardis, 2008; Strausberg *et al.*, 2008; Ansorge, 2009).

In emulsion PCR, the single-stranded DNA fragments or templates are attached to the surface of beads using adaptors or linkers, and one bead is attached to a single DNA fragment from the DNA library. The DNA library is generated through random fragmentation of the genomic DNA. The surface of the beads contains oligonucleotide probes with sequences that are complementary to the adaptors binding the DNA fragments. After that, the beads will be compartmentalized into separate water-oil emulsion droplets. In the aqueous water-oil emulsion, each of the droplets capturing one bead will serve as a PCR microreactor for amplification steps to take place and produce clonally amplified copies of the DNA fragment.

However, for bridge amplification on solid surface for Illumina® GA, the single-stranded DNA fragments are first attached to a solid surface known as a flowcell using adaptors with complementary probes on the flowcell. Then, the other unattached end of the DNA fragments will create a 'bridge-like structure' by bending over and also hybridize to the probes on the flowcell, which form the template for amplification to generate clonally amplified copies of the DNA fragments on the surface of the flowcell. However, this third generation sequencing is characterized by single DNA molecule sequencing without the need for amplification steps. The first third generation sequencing instrument – Heliscope Single Molecule Sequencer – is now commercially marketed by Helicos Biosciences.

Sequencing chemistry: massively parallel sequencing

The sequencing approaches for NGS technologies can be broadly divided into sequencing-by-synthesis mediated by polymerase enzymes (pyrosequencing for Roche® GS FLX and sequencing by reversible terminator chemistry for Illumina® GA) and sequencing-by-ligation mediated by ligase enzymes (ABI® SOLiD) (Mardis, 2008; Strausberg *et al.*, 2008; Ansorge, 2009).

In pyrosequencing, the adding of dNTPs (deoxynucleotide triphosphate) and reagents for cyclic sequencing is controlled, where each of the four types of dNTPs will flow through the picotiter plate consecutively or sequentially. This means that only one type of dNTP is present per cycle of sequencing or synthesis, followed by another type of dNTP in the next cycle and the cycles repeat. This is totally different from the reversible terminator chemistry sequencing for Illumina® GA where all the four types of ddNTPs labelled by different fluorescent colours are present in each cycle of sequencing. A picotiter plate contains more than one million wells where the beads (attached to clonally amplified copies of DNA fragments) are situated, and one well holds one bead. As such, it allows parallel sequencing of an enormous number of DNA fragments.

The polymerase-based synthesis or incorporation of the complementary dNTPs to the DNA templates will cause the release of inorganic pyrophosphate triggering a series of downstream reactions which eventually produce chemiluminescent light which is captured by a detection system (CCD camera). The detection system records the intensity of light emitted from each well that corresponds to a single DNA fragment. In summary, generally each cycle of sequencing consists of dNTPs incorporation, pyrosequencing reactions and emission of chemiluminescent light and measurement of the light intensity. The sequencing reagents of the previous cycle are washed away before next cycle of sequencing takes place.

The intensity of chemiluminescence is proportional to the amount of inorganic pyrophosphate released and thus the number of dNTPs incorporated to the DNA template. Owing to this factor, pyrosequencing is more susceptible to insertion deletion (indel) errors in homopolymer sequences (i.e. DNA sequences of consecutive identical nucleotides like GGGGG or AAAAA) because of less accurate estimation of the length or the number of nucleotides in homopolymer sequences. This is especially problematic for homopolymers with more than six bases. In pyrosequencing, several dNTPs can be incorporated when there are consecutive identical nucleotides in the sequences; this is in contrast to the sequencing by reversible terminator chemistry where only one ddNTP (dideoxynucleotide triphosphate) is incorporated to the DNA templates per cycle of sequencing. To further illustrate this, for example, for homopolymer GGGGG, five dCTPs (deoxycytidine triphosphate) will be incorporated for pyrosequencing at one time, whereas only one ddCTP (dideoxycytidine

triphosphate) for reversible terminator chemistry sequencing and another ddCTP will be incorporated in the next four cycles of sequencing.

Like Roche® GS FLX, Illumina® GA also employs the sequencing-by-synthesis approach, although it is totally different from pyrosequencing. In reversible terminator chemistry sequencing, all the four types of ddNTPs and sequencing reagents are added onto the flowcell, and these ddNTPs are labelled by four different fluorescent colours corresponding to the four different nucleotides. One flowcell has several hundred million clusters and each cluster contains clonally amplified copies from a single DNA fragment. Similar to the Roche® GS FLX picotiter plate, the format of the flowcell also allows simultaneous sequencing of an enormous number of DNA fragments. However, it is noteworthy that the difference in the number of DNA fragments that gets sequenced in parallel between the two platforms is about several hundred-folds.

The ddNTPs are reversible terminators, allowing for the synthesis of DNA templates in the next cycle of sequencing for the incorporation of other ddNTPs. In this cyclic sequencing approach, one complementary ddNTP will be incorporated to the DNA template at one time, followed by washing steps to remove the excess sequencing reagents. This is then followed by the imaging of the fluorescence signals across the whole flowcell. After imaging, the fluorescent labels will be removed and the 3' blocking group of the ddNTPs is also chemically removed. These steps are then repeated. Since only one ddNTP is incorporated at one time, and the base calling is not proportional to light intensity but is dependent on the fluorescent colours, the reversible terminator chemistry does not have problems in the sequencing of homopolymer sequences. However, it is more prone to substitution errors because all the four types of ddNTPs are present in each cycle of sequencing, unlike in pyrosequencing, where only one specific type of dNTP is present.

It is worthwhile to note that in pyrosequencing, dNTPs are used, whereas in reversible terminator chemistry sequencing, ddNTPs are used and they are reversible terminators. However, Sanger sequencing requires a mixture of both dNTPs and ddNTPs, and the ddNTPs are non-reversible terminators. Although these sequencing approaches are generally based on sequencing-by-synthesis, it is obvious that the sequencing chemistries and approaches are very different. In pyrosequencing, the identity of nucleotides that are incorporated into DNA templates is determined by emission of chemiluminescent light; however, the nucleotides are determined by different fluorescent colours for reversible terminator chemistry and Sanger sequencing.

The sequencing approach of ABI® SOLiD is based on sequencing-by-ligation. Like Roche® GS FLX, ABI® SOLiD also employs emulsion PCR for amplification. The beads containing DNA fragments are then deposited on a glass slide. The sequencing of DNA templates is mediated by ligase. In brief, the sequencing is based on sequential ligation of dinucleotide probes which are labelled by four different fluorescent colours. There are 16 possible combinations of two nucleotides, and these dinucleotide probes

will compete for incorporation into the DNA templates. As such, ligation of one probe will query two nucleotides in the DNA templates. The sequencing of DNA templates is completed by seven ligation cycles for each of the five rounds of primer reset, and at the end produces a sequence read length of 35 bases. Using this unique sequencing approach, every single position or base in the DNA template is interrogated twice, and allowing for distinction between true genetic variations and errors.

Sequence read length and error

The NGS technologies have a number of advantages over Sanger sequencing, but they are not without limitations. The new sequencing technologies are characterized by shorter sequence read lengths compared to Sanger sequencing, that is 125 bases or less for Illumina® GA and ABI® SOLiD, as well as for Heliscope Single Molecule Sequencer. As a result, NGS technologies are not suitable for *de novo* sequencing of large and complex genomes like the human genome as the assembly of billions of short sequence reads into large contigs would be difficult and challenging. Relatively longer sequence read lengths are needed to obtain larger contigs with fewer gaps in between in the assembled consensus sequence. However, the latest improvements in sequencing chemistry and system have enabled Roche® GS FLX to achieve sequence read lengths of 400–500 bases on average, but it is still half of that that can be achieved by Sanger sequencing, which is approximately 800 bases to 1 kb in length (Mardis, 2008; Strausberg *et al.*, 2008; Ansorge, 2009).

The feature of short sequence read lengths also makes NGS technologies like Illumina® GA and ABI® SOLiD inadequate for metagenomics studies in investigating bacterial diversity. It is crucial to have longer sequence read lengths to achieve sufficient discriminatory power of the sequences derived from different bacterial species in a sample, determining the presence of diverse species by mapping the sequences against different reference bacterial genomes. As a result, Roche® GS FLX has become the primary choice for this kind of studies. Nevertheless, the feature of short sequence read length is just nice for 'sequence census methods or applications' like ChIP-Seq and RNA-Seq. These sequence census methods do not require full sequence or long sequence read lengths, but rather, lengths sufficient to align or map the sequences uniquely to the reference genome sequence (Wold and Myers, 2008).

Although Illumina® GA and ABI® SOLiD are less suitable for metagenomics studies at the time, they appear to be more ideal for studies like ChIP-Seq and RNA-Seq compared to Roche® GS FLX. This is because of their ability to generate several hundred millions of short sequence reads compared to several hundred thousand to one million longer sequence reads for Roche® GS FLX. In the applications like ChIP-Seq and RNA-Seq, the number of sequence reads is more crucial than the length of sequence reads for 'counting' purposes, as far as the length is sufficient to align uniquely to the reference genome sequence.

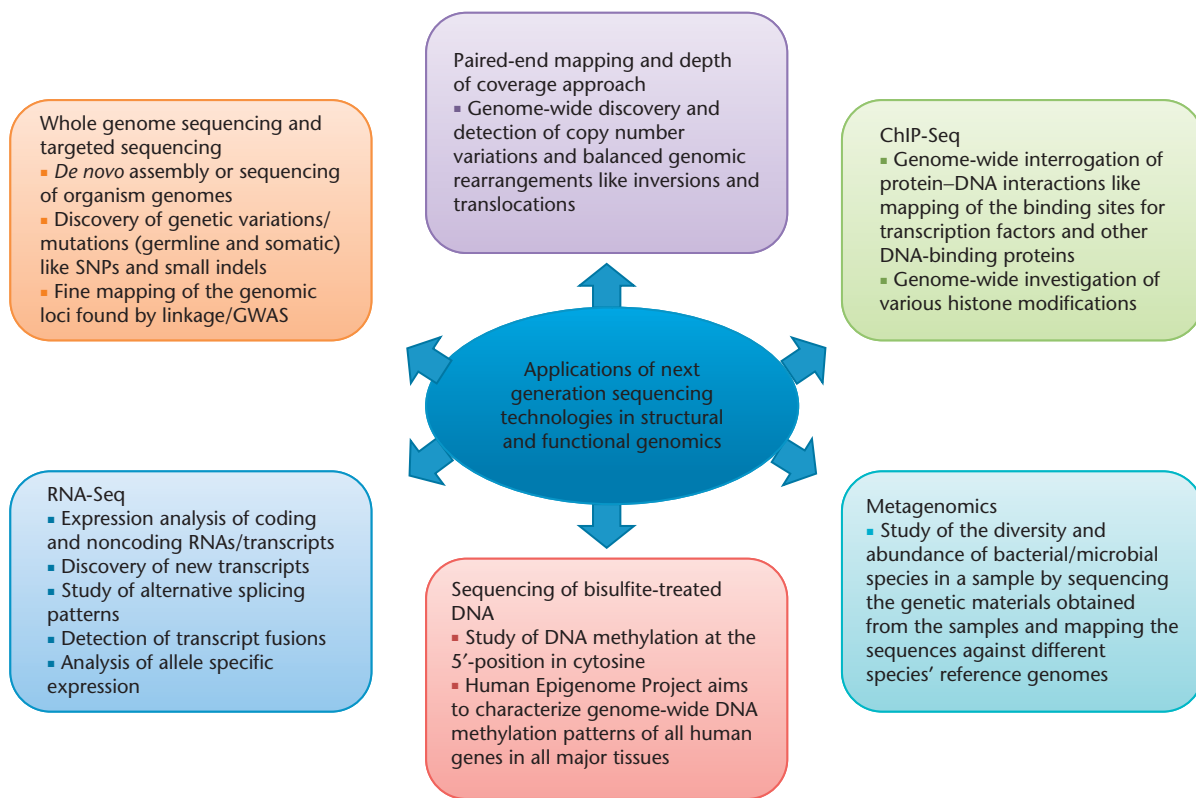


Figure 1 Application of next generation sequencing technologies in structural and functional genomics.

In addition to the limitation in sequence read length, the NGS technologies were also reported to have higher read base or base calling error rates, although it has been improving. ABI® SOLiD has achieved the highest accuracy with <0.1% error rate among the NGS technologies, whereas the read base error rates for Illumina® GA and Roche® GS FLX are within 0.2–2% and 0.5–1.5%, respectively (Li and Wang, 2009). The differences seem to be small and insignificant in terms of the percentage, but when the error rates are transmitted to whole genome sequencing of six billion bases for a human diploid genome, it will generate hundreds of thousands to millions of errors in base calling and this will cause a detrimental effect in identifying genetic variations like SNPs. Fortunately, results from whole genome resequencing studies suggest that the SNP calling error rate decreases significantly with greater sequencing depth (Wang *et al.*, 2008). Therefore, it seems that the remedy is to increase the sequencing depth, but one has to bear in mind that this will also add to the sequencing cost.

Applications in Structural and Functional Genomics Studies

Since the arrival of first NGS technology in 2005, these new sequencing platforms have contributed much to the

progress in the research of structural and functional genomics. The NGS technologies have been used in various research areas besides the standard sequencing applications such as whole genome sequencing; they have also been increasingly applied in detecting structural variations (paired-end mapping), studies of protein–DNA interactions and histone modifications (ChIP-Seq), and transcriptomic profiling of messenger RNAs (mRNAs) and noncoding RNAs (RNA-Seq). These are the most common applications built on the NGS data and will be the focus of our discussion (Figure 1). This article also focuses on these applications in human genomics studies, although NGS technologies have also been widely used for genomics studies of plants and other model organisms. The new and innovative applications of NGS technologies have contributed remarkably to the advancement in human genomics studies.

Whole genome sequencing

The completion of several whole human genome resequencing studies has yielded important scientific findings and new insights into *human genetic variations* (Wheeler *et al.*, 2008; Kim *et al.*, 2009; Ahn *et al.*, 2009; Wang *et al.*, 2008; Bentley *et al.*, 2008; Mckernan *et al.*, 2009; Pushkarev *et al.*, 2009). It is equally important that they also served as proof-of-concept studies demonstrating the feasibility of using NGS and third generation sequencing technologies

to decode the DNA sequence of human genome efficiently and at an affordable price per genome. Moreover, these studies have also addressed important questions and issues surrounding the experimental design and data analysis, such as the preparation of DNA libraries for sequencing, assessment of the sequencing depth that is needed to provide adequate coverage of the reference genome sequence and to minimize SNP calling error rate, and the quality control criteria for the detection of genetic variations like SNPs, indels and structural variations. For example, Wang *et al.* (2008) found that at a sequencing depth of greater than 10-fold, the assembled consensus sequence covered ~83% of the NCBI human reference genome using single-end reads and ~95% coverage using paired-end reads, and greater sequencing depth has minimally increased in the coverage. However, the SNP calling error rate decreases significantly with greater sequencing depth.

The findings from several whole genome resequencing studies have also deepened our understanding of human genetic variations. These studies revealed an abundance of various genetic variations in the human genome, namely SNPs, indels and structural variations. Although the finding of several million SNPs in each individual genome is not new, more interesting is the fact that the studies have identified several hundreds of thousands of new SNPs that have not been catalogued in dbSNP. For example, about one million new SNPs were identified in the African genome (NA18507) and approximately half a million SNPs for the other genomes of Caucasian and Asian ancestry (Bentley *et al.*, 2008; Wheeler *et al.*, 2008; Wang *et al.*, 2008; Kim *et al.*, 2009; Ahn *et al.*, 2009).

Apart from SNPs, whole genome resequencing studies also identified several hundred thousand of short indels with sizes ranging from several bases to tens of bases. The Han Chinese (YH) genome contained approximately 135 000 indels within 1–3 bp, and approximately 400 000 indels defined from 1 to 16 bp were found in the African NA18507 genome. However, Ahn and colleagues identified the indels within a size range from –29 to +14 bp and found nearly 343 000 entries for the Korean genome SJK (Bentley *et al.*, 2008; Wang *et al.*, 2008; Ahn *et al.*, 2009). The effort to catalogue short indels in the human genome was far less devoted than that for SNPs, where more than 50% of the identified indels have not been catalogued, whereas only less than 30% of the identified SNPs are new. Similarly some new discoveries have also been made for structural variations, where several thousands of them were identified. The large-scale sequencing studies like whole genome resequencing and *1000 Genomes Project* would not have been feasible without the advances in NGS technologies. **See also:** [Copy Number Variation in the Human Genome](#); [Genetic Variation: Human](#); [Single Nucleotide Polymorphism \(SNP\)](#)

In addition to the aforementioned whole genome resequencing of nondisease genomes, the cancer genome of acute myeloid leukaemia has also been sequenced to study the *de novo* somatic mutations (Ley *et al.*, 2008). Apart from germline genetic variations, the importance of

somatic mutations in carcinogenesis is also well established. Therefore, focusing merely on germline genetic variations will not be sufficient to fully decipher the genetic basis of cancers. It is noteworthy that the genome-wide association studies (GWAS) only interrogated the germline genetic variations of cancer and that the whole genome SNP genotyping arrays used in GWAS are not designed to study somatic mutations. Direct sequencing is required for detecting somatic mutations; hence, sequencing approach provides an additional advantage in dissecting the cancer genome compared to genotyping.

Paired-end mapping of structural variations

The ubiquity of CNVs in the human genome was first reported several years ago (Sebat *et al.*, 2004; Iafrate *et al.*, 2004), and many more have since been found. Previous studies have used poor sensitivity methods to detect CNVs leading to high false negative rates (Scherer *et al.*, 2007). Most of the CNV data were generated by microarray-based methods such as CGH and SNP arrays where the signal intensity information is used to detect deletions and duplications. Because of the reliance on relative or differences in signal intensities to detect copy number variable regions, these methods are unsuitable for detecting other structural variations like inversions and translocations (also known as balanced chromosomal rearrangements). Furthermore, due to the limitations in density or resolution of CGH and SNP arrays, the methods are lacking in sensitivity to detect smaller sizes of CNVs (< 50 kb). The discovery of smaller sizes of CNVs is crucial as they are predicted to be more abundant than the larger CNVs (Estivill and Armengol, 2007). The latest developments in SNP genotyping arrays, namely increased probe density and uniformity of distribution in the genome, and also included copy number probes to cover regions lacking of SNPs, have improved the sensitivity compared to earlier arrays. Nonetheless, the SNP arrays still suffer from poor sensitivity to detect CNVs smaller than 5–10 kb even using the highest density SNP arrays such as Illumina® Human 1M Beadchip and Affymetrix® 6.0 SNP Arrays (McCarroll *et al.*, 2008; Cooper *et al.*, 2008). Therefore, higher resolution and sensitivity methods are needed to detect CNVs and also balanced structural variations.

The proof-of-concept study using NGS technologies to detect structural variations was published in 2007, and the sequencing-based method was known as paired-end mapping (Korbel *et al.*, 2007). In this method, a library of DNA fragments of fixed insert sizes is prepared, both ends of the DNA fragments are sequenced, and the sequence information is used to map against the human reference genome. The underlying principle of the paired-end mapping approach to detect structural variations is simple; it is based on the discrepancies in length or orientation of the DNA fragments to be sequenced. In other words, when both ends of the DNA fragment that map against the reference sequence show discordances in terms of size or

length, this is an indication for deletion and insertion, whereas discordance in orientation suggests the presence of inversion. Since the insert size of the library is known, both ends of DNA fragments that map to the reference is shorter than expected; this indicates the presence of insertion; conversely, longer than the insert size suggests the presence of deletion. Korbelt and colleagues prepared libraries of 3 kb insert size for two female individuals, and using the aforementioned mapping approach and Roche® 454 sequencing, they found 1297 structural variations, including 853 deletions, 322 insertions and 122 inversions. After this study, several whole genome resequencing studies have also used the paired-end mapping strategy and identified thousands of structural variations (Wang *et al.*, 2008; Ahn *et al.*, 2009).

Furthermore, the paired-end sequencing method has also been used to interrogate somatic genomic rearrangements in cancer (Campbell *et al.*, 2008). In the study, Illumina® GA was used to perform the sequencing of both ends of DNA fragments derived from the genomes of two individuals with lung cancer, and they identified 306 germline structural variants and 103 somatic rearrangements to the single nucleotide level of resolution. The cancer genome is well characterized by genomic instability, with the presence of numerous structural variations and complex genomic rearrangements, and these genetic aberrations are not well captured by microarray hybridization methods. However, this study has now shown the feasibility and advantages of paired-end sequencing method to decipher the cancer genome. This mapping approach is undoubtedly a promising strategy to harvest new cancer genes. The paired-end sequencing approach takes the advantages of the short sequence reads produced by NGS technologies to map against human reference genome, it is an application of 'census sequence methods'. Nonetheless, one major limitation of paired-end mapping is the inability to detect insertions larger than the insert size of the library.

Recently a new and innovative method of using NGS data to detect CNVs has been developed. The approach is based on the depth of coverage of the sequence reads, and some promising results have been obtained showing that it is effective to search for copy number variable regions. The principle underlying the depth of coverage approach is not complicated. This approach assumes that the sequencing is uniform, and that the number of sequence reads mapping to a region follows a Poisson distribution. As such, the number of reads should be proportional to the number of times that a particular region appears in the genome. Therefore, it is expected that a duplicated region will have more number of reads mapping to it, and the converse is true for deletions (Yoon *et al.*, 2009; Medvedev *et al.*, 2009).

Studies comparing the results between the depth of coverage approach and the paired-end mapping approach found that only a minority of the CNVs had overlapped between the two methods. Furthermore, the identified CNVs that are specific to the former method are more greatly enriched in segmental duplications than the paired-

end mapping-specific CNVs. This suggests that both methods in identifying CNVs are complementary to each other and that the combination of the methods will certainly further improve the sensitivity of detection throughout the genome. In fact, both methods have their own advantages and limitations (Yoon *et al.*, 2009; Medvedev *et al.*, 2009).

ChIP-seq for studying protein–DNA interactions and histone modifications

Previously, the studies of protein–DNA interactions like identifying transcription factor binding sites, have relied on some low-throughput methods, and focused on some specific genomic regions. However, with the advent of microarray technologies, for the first time, a comprehensive interrogation of the whole genome has become feasible. In the era of microarrays, the genome-wide studies of protein–DNA interactions and histone modifications were performed using a method known as ChIP-chip.

The ChIP or chromatin immunoprecipitation experiment consists of several steps. First, the protein (e.g. a transcription factor of interest) and its binding DNA sequences or genomic regions are chemically cross-linked by treating the cells with formaldehyde. Then the genomic DNA is extracted and fragmented before adding the specific antibody interacting with the protein of interest. The function of the antibody is to selectively isolate the antibody–protein–DNA complexes by immunoprecipitation. After the immunoprecipitation, the cross-linking between protein and DNA is reversed to obtain the DNA sequences. The identity of isolated DNA sequences can be determined by methods such as Southern blot, quantitative PCR (qPCR), microarray (ChIP-chip) or sequencing (ChIP-Seq). Chromatin immunoprecipitation requires a highly specific antibody for the DNA-binding protein of interest.

Before microarrays were available, most of the studies of protein–DNA interactions were designed to answer simple questions like whether a genomic region (e.g. the promoter region of a gene of interest) is bound to a transcription factor thus regulating the transcription levels, that is locus-specific experiment. These studies require some prior knowledge to design the experiments and the immunoprecipitated DNA sequences are analysed by Southern blot or qPCR to determine whether the genomic region was indeed immunoprecipitated. However, the arrival of microarray technologies has enabled a different question to be asked. Since the scope of ChIP-chip experiments is not restricted to specific regions, the question that is posed is where the transcription factor binds to in the human genome, that is to identify all the regions where the transcription factor might have regulatory roles. In ChIP-chip experiments, the isolated DNA fragments are labelled fluorescently and hybridized to the probes on microarrays. Undeniably, the developments of microarrays have enabled interrogation on a genome-wide scale, but the

detection of the isolated DNA sequences is still dependent on the availability of the probes to capture them. Although the developments of high-density tiling arrays, where oligonucleotide probes are placed in high density throughout the whole genome, have improved the sensitivity of the ChIP-chip, the cost for such tiling arrays is expensive especially for large genomes like the human genome.

In contrast, for ChIP-Seq, the isolated DNA sequences are not hybridized on microarrays (hence avoiding the inherent problems in probe hybridization experiments); instead they are directly sequenced to detect their presence and abundance. This allows detection of all the DNA fragments or sequences that are isolated in the sample without biases of probe selection. Actually, both the methods, microarray- and sequencing-based experiments, rely on the reference genome sequence, the former method requiring it for synthesizing the probes, and the later method requiring the reference genome for alignments of DNA sequences that it sequenced (Park, 2009; Farnham, 2009).

The earliest two ChIP-Seq studies were first published in 2007 to identify the genome-wide binding sites for DNA-binding proteins, NRSF (neuron restrictive silencer factor) and STAT1 (signal transducer and activator of transcription 1) (Johnson *et al.*, 2007; Robertson *et al.*, 2007). These papers served as proof-of-concept studies for the new approach in studying protein–DNA interactions. Both studies used Illumina® GA to sequence the immunoprecipitated DNA sequences. The identification of the previously known binding sites in both the studies serves as the validation of the approach, and the detection of novel-binding sites shows the higher sensitivity of ChIP-Seq compared to ChIP-chip. The studies have shown some promising results; for example, a total of 1946 locations were identified in the human genome for NRSF, and more importantly, the sequencing data provide a sharp resolution of the binding sites. This approach will certainly facilitate the annotation of the binding sites in the genome for other DNA-binding proteins as well.

The first paper investigating histone methylations using ChIP-Seq also appeared in 2007 (Barski *et al.*, 2007). The study performed genome-wide mapping of 20 different types of histone modifications in the human genome and also used Illumina® GA to perform the sequencing. The high-resolution maps of histone modifications generated by sequencing methods are important in expanding our knowledge on how this mechanism regulates the expression of genes in the human genome. The development of ChIP-Seq is a major stride in functional genomics as the studies of genome-wide protein–DNA interactions like transcription factor binding sites and studies of epigenetics like histone modifications are essential in our understanding of the transcriptional regulatory network. Nonetheless, ChIP-Seq is not without its own challenges and limitations (Park, 2009; Farnham, 2009).

Transcriptomic profiling

Studies of gene expression are important because they are the immediate molecular traits that are directly affected by

genetic variations in DNA sequence and epigenetics regulations. The term gene expression usually refers to expression levels of protein-coding genes, or mRNAs. Previous studies were mainly focused on mRNAs expression, because this class of RNAs is important as they serve as the templates to synthesize proteins through the process of translation, and proteins are the functional molecules involved in diverse cellular functions and biological processes. However, this perception has been changed after the completion of the pilot phase of the ENCODE (Encyclopedia of DNA Elements) Project. The project revealed a pervasive transcription pattern in the 1% of the human genome that was interrogated (ENCODE Project Consortium, 2007; Carninci and Hayashizaki, 2007). It had been previously thought that only the protein-coding regions or sequences (i.e. genes) will undergo transcription followed by translation. However, the ENCODE Project showed that transcription also occurs in nonprotein coding regions as well.

Following the findings, the importance and existence of noncoding RNAs is getting appreciated and research has been devoted to identify and characterize them in the transcriptome. In contrast to mRNAs, the noncoding RNAs only undergo transcription, but are not translated into protein. As such, the transcriptome profiling encompassed both the coding RNAs (mRNAs) and noncoding RNAs. One of the well-known noncoding RNAs is *microRNAs*.

Traditionally, gene expression levels were measured by the Northern blot method and reverse transcription quantitative PCR (RT-qPCR) before the introduction of microarray technologies. Nevertheless, both of them are low-throughput methods where expression profiling of all the known annotated genes in the human genome is not feasible.

The arrival of microarray technologies has enabled for the first time the interrogation of several thousand genes simultaneously in a single experiment, and whole genome expression studies of all the known genes have also become feasible. Although microarrays have been the method of choice for whole genome gene expression profiling for more than a decade, there are a number of inherent limitations or problems in microarray studies. The conventional gene expression microarrays are mainly focused on the expression levels of known annotated protein-coding genes. Like the ChIP-chip experiment, the developments of tiling arrays where the probes were designed to cover the genome systematically in high resolution regardless of the gene annotation have been used in discovering unknown or novel transcripts, although the cost for tiling arrays is expensive. Besides gene expression microarrays, further developments have also enabled microarrays to be used for studies of alternative splicing and microRNAs expression. Currently, a variety of microarrays is commercially available for transcriptomic applications by companies like Affymetrix® and Illumina®.

The microarray method is based on the hybridization of fluorescent labelled targets and probes, and the expression

levels are inferred indirectly from fluorescent intensity. Therefore, the method suffers from certain levels of cross hybridization and noise, generating artefacts which will complicate the interpretation of results. Sequencing-based approach like SAGE was developed before the arrival of NGS technologies and offered a number of advantages over microarrays, such as the ability to detect novel transcripts and the direct measurement of the abundance of transcripts instead of relying on hybridization intensities. In SAGE, the abundance of mRNAs is estimated or measured by counting of sequence tags derived from the 3' end of mRNAs. Nonetheless, this method also has several major limitations such as the costly Sanger sequencing and laborious cloning procedure.

The arrival of NGS technologies has brought about another breakthrough in the approaches to explore the transcriptome, and this sophisticated method is known as RNA-Seq. RNA-Seq is based on NGS technologies that offered several advantages over the previous methods like microarray or SAGE for transcriptomic studies. First, unlike the microarray hybridization method, the detection capability of RNA-Seq is not limited by the probes synthesized on the microarray to capture the corresponding transcripts in the transcriptome, but it is instead influenced by the depth of sequencing. Secondly, since RNA-Seq is not based on hybridization to detect and measure transcript expression, it avoids the background noise resulting from cross-hybridization.

Furthermore, RNA-Seq provides the highest resolution to a single-base resolution which precisely maps the transcription boundaries, and it can also identify sequence variations like SNPs in the transcribed regions. In addition, RNA-Seq can be used to study fusion transcripts and alternative splicing. Although special microarrays like exon microarrays where the probes are designed to span exon junctions can be used to study alternative splicing as well, they are subject to inherent limitations of microarray methods.

RNA-Seq directly sequences and maps the transcripts to the reference genome to measure transcript expression by counting the number of sequence reads. Therefore, RNA-Seq has the largest dynamic range of expression levels, from low abundance to highly expressed transcripts (if the sequencing depth is sufficient for low-abundance transcripts). The number of sequence reads that map to a genomic region corresponds to the level of expression from that region. The performance of RNA-Seq has also been evaluated by benchmarking against the gold standard method, that is RT-qPCR for measuring the expression levels, and has been shown to be highly accurate. Besides this parameter, high reproducibility of the results obtained from RNA-Seq has also been shown. Finally, RNA-Seq allows the studying of the expression of mRNAs and noncoding RNAs, and is also able to detect and identify new transcripts (coding and noncoding) that have not been annotated. However, no method is perfect; RNA-Seq is also not without its problems and challenges (Wang *et al.*, 2009; Morozova *et al.*, 2009).

In summary, besides gene expression profiling, the applications of sequencing-based approaches in transcriptomic studies have been expanded to genome annotation, discovery of new transcripts, investigation of the alternative splicing patterns, detection of gene fusions in cancer, allele-specific expression analysis, as well as the discovery and measurement of noncoding RNA expression (Denoeud *et al.*, 2008; Pan *et al.*, 2008; Maher *et al.*, 2009; Heap *et al.*, 2009; Bar *et al.*, 2008; Morin *et al.*, 2008). The number of publications using sequencing for transcriptomic applications has been growing rapidly. The high-throughput production and significantly cheaper cost are the main factors for NGS technologies to be quickly adopted in transcriptomic studies. The ability to study coding and noncoding RNA expression, alternative splicing, protein–DNA interactions and histone modifications effectively on a genome-wide scale holds a great promise to significantly advance our knowledge in this complex field of transcriptional regulations.

Future Perspectives and Summary

Large-scale sequencing studies have become more feasible and affordable nowadays. In the recent few years since the NGS technologies were introduced, we have seen their tremendous impacts on transforming the approaches in the studies of structural and functional genomics. Moreover, sequencing-based approaches have already yielded numerous novel and important findings in research areas like genome-wide mapping of histone modifications and protein–DNA interactions, discovery of genetic variations, and transcriptomics studies even though the approaches are still new and maturing. These new sequencing technologies have enabled researchers to answer old questions in unprecedented detail and have raised new questions. It has also allowed researchers to design various experiments which were unthinkable just a few years ago with Sanger sequencing.

Further improvements in various aspects of current NGS technologies such as throughput, read length and accuracy and reduction in cost are anticipated. The NGS technologies have shown their potential of being dominant in future genomics studies. It is evident from several international projects using NGS technologies like the ENCODE Project, 1000 Genomes Project and cancers sequencing project by the International Cancer Genome Consortium. Each of the projects has its own specific aim: the ENCODE project aims to annotate all the functional elements, whereas 1000 Genomes Project aims to construct the most comprehensive map of genetic variations in the human genome. The cancers sequencing project intends to study somatic genetic aberrations like point mutations and chromosomal rearrangements in the cancer genome. It is clear that the approaches based on NGS fit in all research areas, and in fact NGS have become an indispensable tool for genomics studies.

It is only a matter of time before achieving the goal of \$1000 per whole genome sequencing. This should not be

too far from now given the progresses in the development of third generation sequencing technologies. The arrival of single molecule DNA sequencing technologies like nanopore sequencing will certainly bring about another breakthrough (Gupta, 2008). In fact, a recent study has shown that whole human genome sequencing can be done at a cost of US\$4400 using a new sequencing platform (Drmanac *et al.*, 2009). Although the \$1000 genome will technically make sequencing of thousands of human genomes a reality, the substantial cost that will be incurred for data storage, powerful computational packages and analytical softwares has to be borne in mind. However, beyond affordability, what are left behind are the bioinformatics challenges in processing and analysing the huge amount of sequencing data (Flicek and Birney, 2009; Pepke *et al.*, 2009; Pop and Salzberg, 2008).

References

- Ahn SM, Kim TH, Lee S *et al.* (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Research* **19**: 1622–1629.
- Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology* **25**: 195–203.
- Bar M, Wyman SK, Fritz BR *et al.* (2008) MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. *Stem Cells* **26**: 2496–2505.
- Barski A, Cuddapah S, Cui K *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837.
- Bentley DR, Balasubramanian S, Swerdlow HP *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.
- Campbell PJ, Stephens PJ, Pleasance ED *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics* **40**: 722–729.
- Carninci P and Hayashizaki Y (2007) Noncoding RNA transcription beyond annotated genes. *Current Opinion in Genetics and Development* **17**: 139–144.
- Cooper GM, Zerr T, Kidd JM *et al.* (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genetics* **40**: 1199–1203.
- Denoeud F, Aury JM, Da Silva C *et al.* (2008) Annotating genomes with massive-scale RNA sequencing. *Genome Biology* **9**: R175.
- Drmanac R, Sparks AB, Callow MJ *et al.* (2009) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**: 78–81.
- ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.
- Estivill X and Armengol L (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genetics* **3**: 1787–1799.
- Farnham PJ (2009) Insights from genomic profiling of transcription factors. *Nature Reviews. Genetics* **10**: 605–616.
- Flicek P and Birney E (2009) Sense from sequence reads: methods for alignment and assembly. *Nature Methods* **6**: S6–S12.
- Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology* **26**: 602–611.
- Heap GA, Yang JH, Downes K *et al.* (2009) Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. *Human Molecular Genetics* **19**: 122–134.
- Iafraite AJ, Feuk L, Rivera MN *et al.* (2004) Detection of large-scale variation in the human genome. *Nature Genetics* **36**: 949–951.
- International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.
- Johnson DS, Mortazavi A, Myers RM and Wold B (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**: 1497–1502.
- Kim JI, Ju YS, Park H *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**: 1011–1015.
- Korbel JO, Urban AE, Affourtit JP *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.
- Levy S, Sutton G, Ng PC *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biology* **5**: e254.
- Ley TJ, Mardis ER, Ding L *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**: 66–72.
- Li Y and Wang J (2009) Faster human genome sequencing. *Nature Biotechnology* **27**: 820–821.
- Maher CA, Kumar-Sinha C, Cao X *et al.* (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**: 97–101.
- Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**: 387–402.
- McCarroll SA, Kuruvilla FG, Korn JM *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**: 1166–1174.
- McKernan KJ, Peckham HE, Costa GL *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research* **19**: 1527–1241.
- Medvedev P, Stanciu M and Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**: S13–S20.
- Morin RD, O'Connor MD, Griffith M *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research* **18**: 610–621.
- Morozova O, Hirst M and Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. *Annual Review of Genomics and Human Genetics* **10**: 135–151.
- Pan Q, Shai O, Lee LJ *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40**: 1413–1415.
- Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews. Genetics* **10**: 669–680.
- Pepke S, Wold B and Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. *Nature Methods* **6**: S22–S32.
- Pop M and Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24**: 142–149.

- Pushkarev D, Neff NF and Quake SR (2009) Single-molecule sequencing of an individual human genome. *Nature Biotechnology* **27**: 847–852.
- Robertson G, Hirst M, Bainbridge M *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**: 651–657.
- Scherer SW, Lee C, Birney E *et al.* (2007) Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39**: S7–S15.
- Sebat J, Lakshmi B, Troge J *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.
- Shendure J and Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–1145.
- Strausberg RL, Levy S and Rogers YH (2008) Emerging DNA sequencing technologies for human genomic medicine. *Drug Discovery Today* **13**: 569–577.
- Tucker T, Marra M and Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. *American Journal of Human Genetics* **85**: 142–154.
- Von Bubnoff A (2008) Next-generation sequencing: the race is on. *Cell* **132**: 721–723.
- Wang J, Wang W, Li R *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.
- Wang Z, Gerstein M and Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics* **10**: 57–63.
- Wheeler DA, Srinivasan M, Egholm M *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872–876.
- Wold B and Myers RM (2008) Sequence census methods for functional genomics. *Nature Methods* **5**: 19–21.
- Yoon S, Xuan Z, Makarov V *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**: 1586–1592.

Further Reading

- ABI® The SOLiD System: http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiDSystemSequencing/index.htm
- Illumina® Sequencing Technology: http://www.illumina.com/technology/sequencing_technology.ilmn
- Kahvejian A, Quackenbush J and Thompson JF (2008) What would you do if you could sequence everything? *Nature Biotechnology* **26**: 1125–1133.
- MacLean D, Jones JD and Studholme DJ (2009) Application of ‘next-generation’ sequencing technologies to microbial genetics. *Nature Reviews. Microbiology* **7**: 287–296.
- Morozova O and Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**: 255–264.
- Roche® 454 Sequencing: <http://www.454.com/>