Next-generation Sequencing: Basic Concepts and Applications in Genome Analysis

Genome: *A genome is an organism's complete set of DNA or hereditary material. It contains all of the information needed to build and maintain that organism.*

Transcriptome: *The transcriptome comprises of all the RNA molecules, including mRNA, rRNA, tRNA and non-coding RNA, produced in one or a population of cells.*

Epigenome: *The epigenome includes modifications associated with chromatin which do not alter the DNA sequence, such as DNA methylation, and histone tail modifications, which can change the structure of chromatin and thereby regulate gene expression.*

Proteome: *The proteome consists of the set of expressed proteins in a given type of cells at a given time under defined conditions.*

SAGE: *Serial Analysis of Gene Expression is a technique used by molecular biologists to produce a snapshot of the messenger RNA population in a sample of interest in the form of small tags that correspond to fragments of those transcripts.*

CAGE: *Cap Analysis Gene Expression is a technique used in molecular biology to produce a snapshot of the 5' end of the messenger RNA population in a biological sample.*

MPSS: *Massively parallel signature sequencing is a sequence based approach that can be used to identify and quantify mRNA transcripts present in a sample by generating small fragment signatures of each mRNA species.*

The last twenty years have seen a phenomenal increase in genomic data output from large facilities and small labs. The major hallmark of the past decade was the completion of the human *genome* project in 2003, which realized the possibility of applying sequencing techniques on the genome-wide scale (Fig. 1) This set the stage for future projects: comparison of genomes of different species, discovering the conservation and variation in sequences between species and individuals, and comparing the genomes of different cells of the same organism, and those of cells during different stages of development, or in normal state and in disease [2,7,9].

Moreover, it became possible to sample the *transcriptome*, *epigenome*, and the *proteome* with the development of techniques like *SAGE, CAGE, MPSS* and *ChIP*. These techniques can provide a snapshot of the expression and regulation of the information content in the cell, and made apparent the need to go beyond the sampling, into an "in depth" analysis of the genome, transcriptome, epigenome, and proteome, which was not previously possible due to technical limitations, and also time and cost related constraints of the old sequencing methods. The birth of the next-generation sequencing (NGS) technology ushered in the promise of high throughput sequencing of multiple genomes and transcriptomes in entirety, while being time efficient and more affordable. It also led to the simultaneous development of algorithms and bioinformatics tools to store, interpret and analyze the terabases of data produced [7,13].

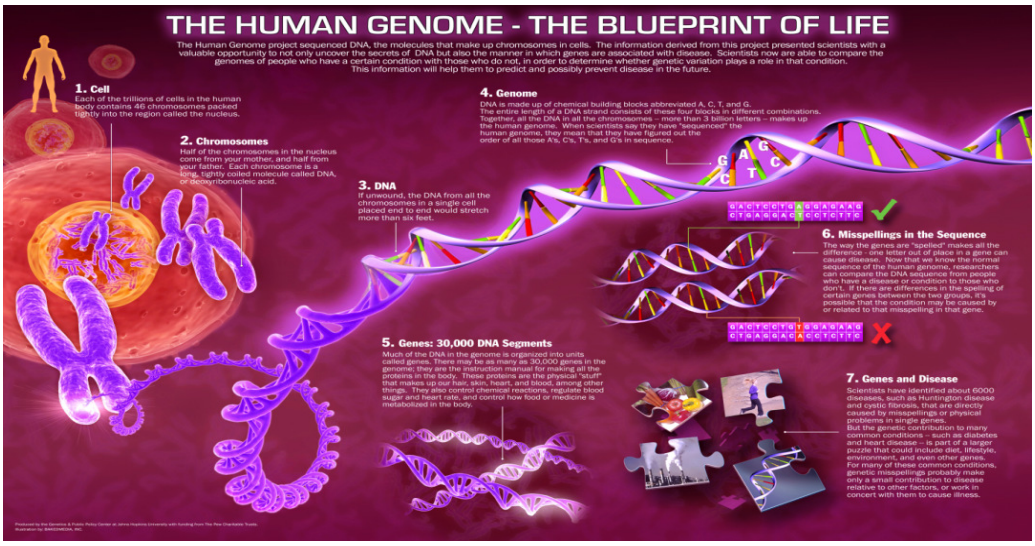

Fig. 1 Human Genome Project. www.dnapolicy.org/resources/human_genome_final.pdf

***ChIP**: Chromatin Immunoprecipitation is an experimental form of immunoprecipitation used to investigate the interaction between proteins and DNA in the cell and determine specific locations of histone modifications in a genome.*

***cDNA:** complementary DNA is DNA synthesized from a mature mRNA template in a reaction catalyzed by the enzyme reverse transcriptase.*

***Template:** A fragment of DNA which contains the region to be amplified or sequenced .*

***PCR:** Polymerase Chain Reaction is a technique to amplify a single or few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence. The method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA.*

***Microarray:** It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, containing specific DNA sequence, known as probes (or reporters) that are used to hybridize a cDNA sample (called target) under high-stringency conditions.*

***PHRED**: Phred quality scores are assigned to each base call in automated sequencer traces and are used to characterize the quality of DNA sequences, and can be used to compare the efficacy of different sequencing methods.*

The complete human genome was first sequenced using the Sanger sequencing (chain termination) method, which has dominated the field for the last two decades. In the chain termination method (Fig. 2), whole genomic DNA or *cDNA*, is fragmented, and each fragment is cloned in a plasmid vector and amplified in *E.coli.* Millions of fragments, produced from a single bacterial colony, are then isolated and purified. Each single strand of DNA is elongated using a labeled primer in the presence of DNA polymerase and 2'-deoxynucleotides (dNTP's) and 2',3'-dideoxynucleotides (ddNTP's). The ddNTP's are labeled with four different fluorescent dyes (one for each base type) and added to the same reaction mixture. The ddNTP's serve as non-reversible synthesis terminators: when a ddNTP is incorporated into the growing chain, DNA synthesis is terminated due to lack of the 3'OH group required for binding the incoming nucleotide. The fragments are then separated by capillary gel electrophoresis and the fluorescent dye determines the DNA sequence (each base type is tagged to a different dye). The use of capillary arrays has increased the efficiency and throughput of the automated Sanger method over the traditional Sanger method. However, one disadvantage is that it involves a bacterial cloning step for amplification of the *template*, which can result in host related biases. Further, the chain termination method is time consuming, labor intensive, and very expensive, and therefore it limits the extent of the study to regions of interest rather than whole genome scale undertakings. The NGS platforms have supplanted the Sanger technology by performing massively parallel "in depth" genome sequencing for a lower cost, opening new frontiers for smaller organizations [10].

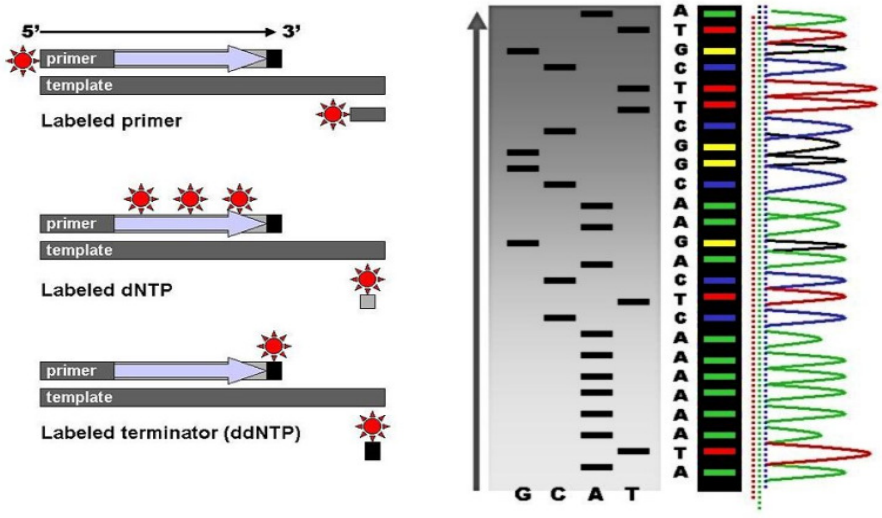

Fig. 2 The chain termination method for DNA sequencing. www.wikipedia.org

***SNP**: A single-nucleotide polymorphism is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G — in the genome or sequence differs between members of a species (or between paired chromosomes in an individual).*

***Structural variations:** These are large non-SNP variations in the genome and can include copy number variations (CNVs), inversions, insertions, deletions and other complex rearrangements.*

***Comparative Genomic Hybridization(CGH) :** It is a molecular-cytogenetic method for the analysis of copy number changes (gains/losses) in the DNA content of a given subject's DNA and often in tumor cells.*

***Pharmacogenomics:** It is the branch of pharmacology which deals with the influence of genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with a drug's efficacy or toxicity.*

***Mutations:** Changes in the DNA sequence: they may be insertions, deletions, single base substitutions, rearrangements or CNVs.*

***Exome:** All the protein coding regions or exons of the genome.*

***Somatic mutations:** Mutations (alterations in the DNA) that are not transmitted to the offspring or changes that arise within individual cells and accumulate throughout a person's lifetime; also called acquired mutations.*

***Synonymous mutations:** Mutations in exons that do not alter the amino acid sequence of a protein.*

***Non-synonymous mutations:** Mutations in exons that alter the amino acid sequence of a protein.*

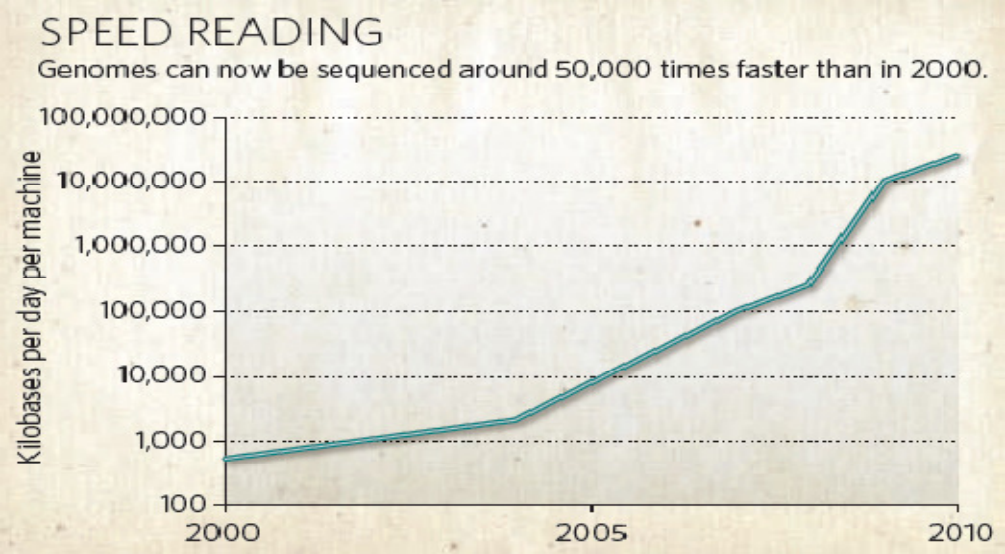

Fig. 3 Next generation sequencers and sequencing speed. NATURE|Vol 464|1 April 2010. Pg. 676

The most widely used NGS platforms are 454/Roche, Illumina/Solexa, ABI/SOLiD, Helicos BioSciences, and the list is still growing. These platforms can perform high throughput sequencing at single base resolution, and can perform several experiments in parallel and multiple folds of the same experiment within a short time period. NGS platforms are thus said to provide deep coverage of the genome, where depth refers to the number of times an individual experiment or run is repeated, and coverage refers to the fraction of the complete genome sequenced. For instance, NGS technology makes it possible to achieve 99.9% genome coverage, at a high speed producing 1-3 Gb per run (Fig. 3). This makes it possible to compare results over multiple runs, which allows for confident sequence assignment. In turn, this gives scientists the opportunity to analyze large data sets and compare results across cells, samples, individuals, species and populations at an unprecedented depth [2,7,9,10,13].

All NGS platforms follow the same basic steps but the methods used for each step differ across different platforms. In general the sequencing procedure can be divided into three major steps: template preparation, sequencing and imaging, and analysis of the sequence data [10].

Template Preparation: For genome sequencing, DNA methylation and epigenome analysis, the template is mostly fragmented DNA; in transcriptome studies the reverse transcribed cDNA is used for library preparation. The NGS platforms circumvent the bacterial cloning step for amplification of template by using various *PCR* based *in vitro* template amplification methods. DNA is randomly sheared and attached to a solid support using adaptors, resulting in a large number of spatially separated templates which can be amplified simultaneously. Templates may be

clonally amplified by either emulsion PCR (Roche/454) or solid-phase amplification (Illumina/Solexa). Emulsion PCR (EmPCR) is a cell free amplification system in which oil droplets act as “nanoreactors”; a single DNA fragment attached to a bead is isolated in each droplet, and clonally amplified using PCR. After amplification, the emulsion is broken, and the beads, each carrying millions of copies of a unique DNA fragment, are attached to a solid support for sequencing (Fig. 4a). Solid-phase amplification is performed by attaching forward and reverse primers to a glass slide to which the template DNA is hybridized, and then subjected to PCR to produce clonally amplified clusters (Fig. 4b). Some platforms use single molecules (Helicos BioSciences) for analysis and require less than 1µg of starting material, which in turn reduces cost and preparation related biases (Fig. 4c). Single molecule templates are attached to a solid support and sequenced directly without amplification [10].

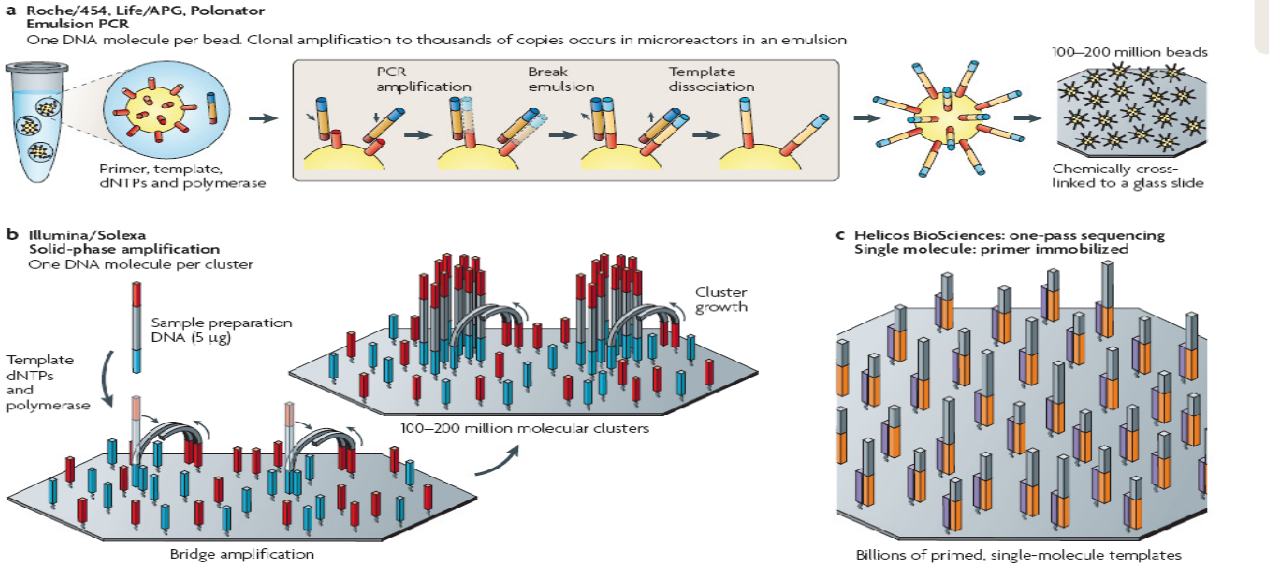

Fig. 4. Template immobilization**.** In emulsion PCR (emPCR) (**a**), a reaction mixture consisting of an oil–aqueous emulsion is created to encapsulate bead–DNA complexes into single aqueous droplets. PCR amplification is performed within these droplets to create beads containing several thousand copies of the same template sequence. EmPCR beads can be chemically attached to a glass slide. Solid-phase amplification (**b**) is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. Immobilizing single-molecule templates to a solid support by a primer (**c**). Michael L. Metzker. Sequencing technologies – the next generation. Nature reviews Genetics 11 January 2010.

Sequencing and Imaging: The immobilized template is then subjected to sequencing using one of several sequencing and imaging methods. In general, the sequence is interpreted from the chemiluminescent (pyrosequencing) or fluorescent signals released during the extension of the oligonucleotide chain upon addition of bases that are complimentary to the sequence of the template (Fig. 5a & 5c). In case of fluorescence detection, each dNTP is labeled with a different fluorescent dye and each addition step is imaged, and the sequence is then determined by analyzing the position and color data (Fig. 5b.). For pyrosequencing (Roche/454), the individual beads are either attached to a glass slide (Fig. 4a) or placed in PicoTiterPlate wells (Fig. 5c-d) to which reporter enzymes are added. Single dNTP’s are then added one at a time, and the incorporation of a base,

which is indicated by the release of inorganic phosphate, is recorded in a graph called a pyrogram. The DNA sequence is then determined from the pyrogram [10].

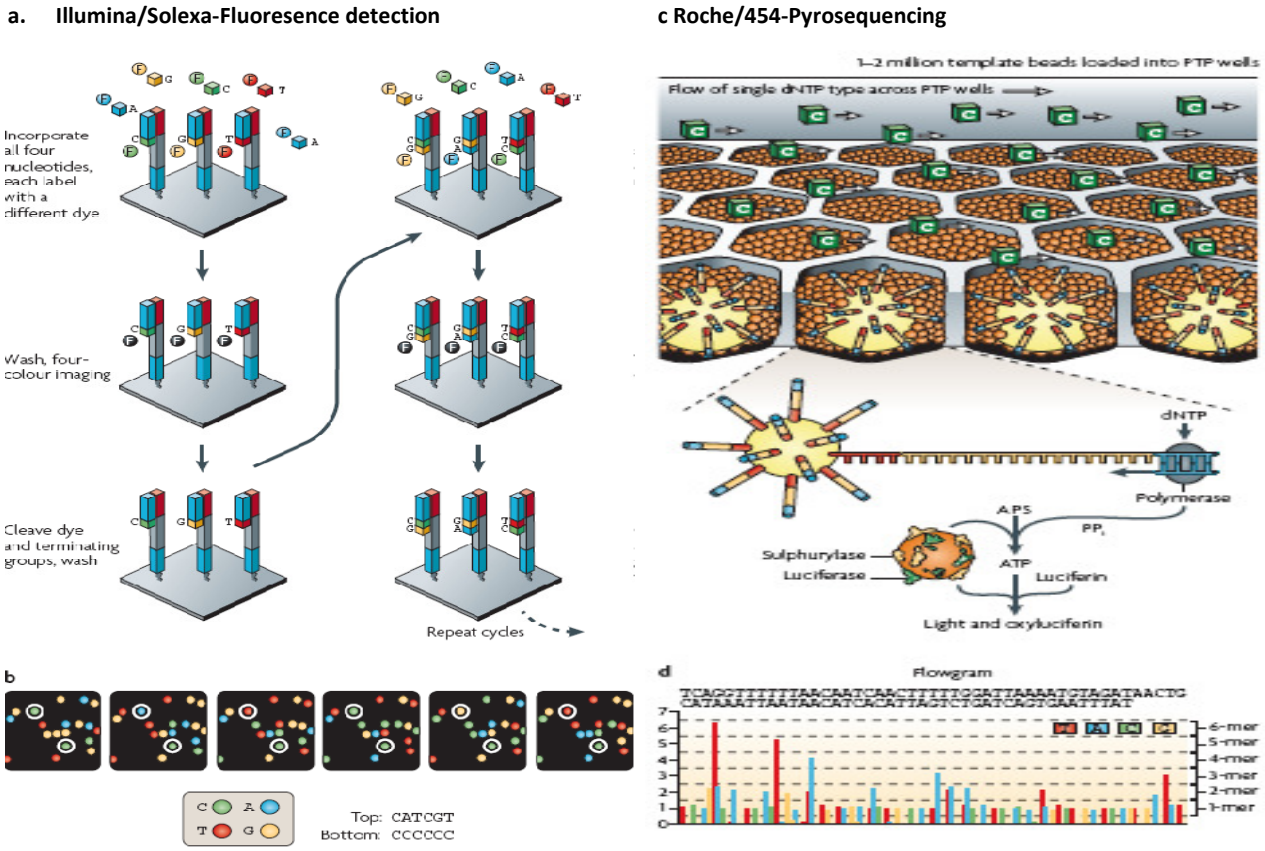

Fig. 5 Sequencing and Imaging: a) Fluorescence detection using differently labeled dNTP's. b) Image acquisition: color and position data from sequencing. c) Pyrosequencing: Flow of single dNTP over the PicoTiterPlate which contains DNA and enzymes. Reporter enzymes release light upon incorporation of each nucleotide. d) Pyrogram: Release of light recoded as series of peaks from which sequence is determined. Michael L. Metzker. Sequencing technologies – the next generation. Nature reviews Genetics 11 January 2010.

Sequence Data Analysis: The data generated from pyrosequencing or fluorescence detection is then interpreted, aligned and analyzed with the help of sophisticated algorithms [10].

The same set of techniques can be applied to different experiments and across different species, therefore several studies can be performed on a single NGS platform. NGS technology can thus detect novel sequences without any prior knowledge of the sequence. This confers it a major advantage over alternative methods like *microarrays*, which use hybridization of template sequences to specifically designed probes, thus scope to sequences that are similar to known genes. On the other hand, the read length for most NGS platforms is very short; a maximum of ~100 bp when compared to the Sanger sequencer whose maximum read length is 900 bp. This is a major impediment when assembling sequences without a reference genome and aligning repetitive sequences from complex genomes. The former is slightly relieved by the fact that each run can be

performed multiple times, and to overcome the latter some NGS platforms can perform “paired-end reads” in which the short template is read from both ends making it possible to position multiple reads in context of the adjacent unique sequence. Another limiting factor, which is due to its recent development, is the lack of reliable standardized tools to compare results across platforms, such as the *PHRED* quality score used for Sanger sequencing. However, many new platforms are in the pipeline, which allow increased read length and perform paired-end reads. Combined with the simultaneous development of standardized tools for data analysis, these developments have the potential to take next-generation sequencing to the limit [2,7,9,10,13].

Next-generation Sequencing: Applications in Genome Analysis

Next-generation sequencing can generate huge amounts of sequence data per run, and has so far been applied in various fields of genomics, such as *de novo* genome sequencing, personal genome sequencing (resequencing), *SNP* discovery, ancient DNA sequencing, mutation mapping, DNA binding site discovery, DNA methylation patterns, and other emerging applications [2,7,9,13].

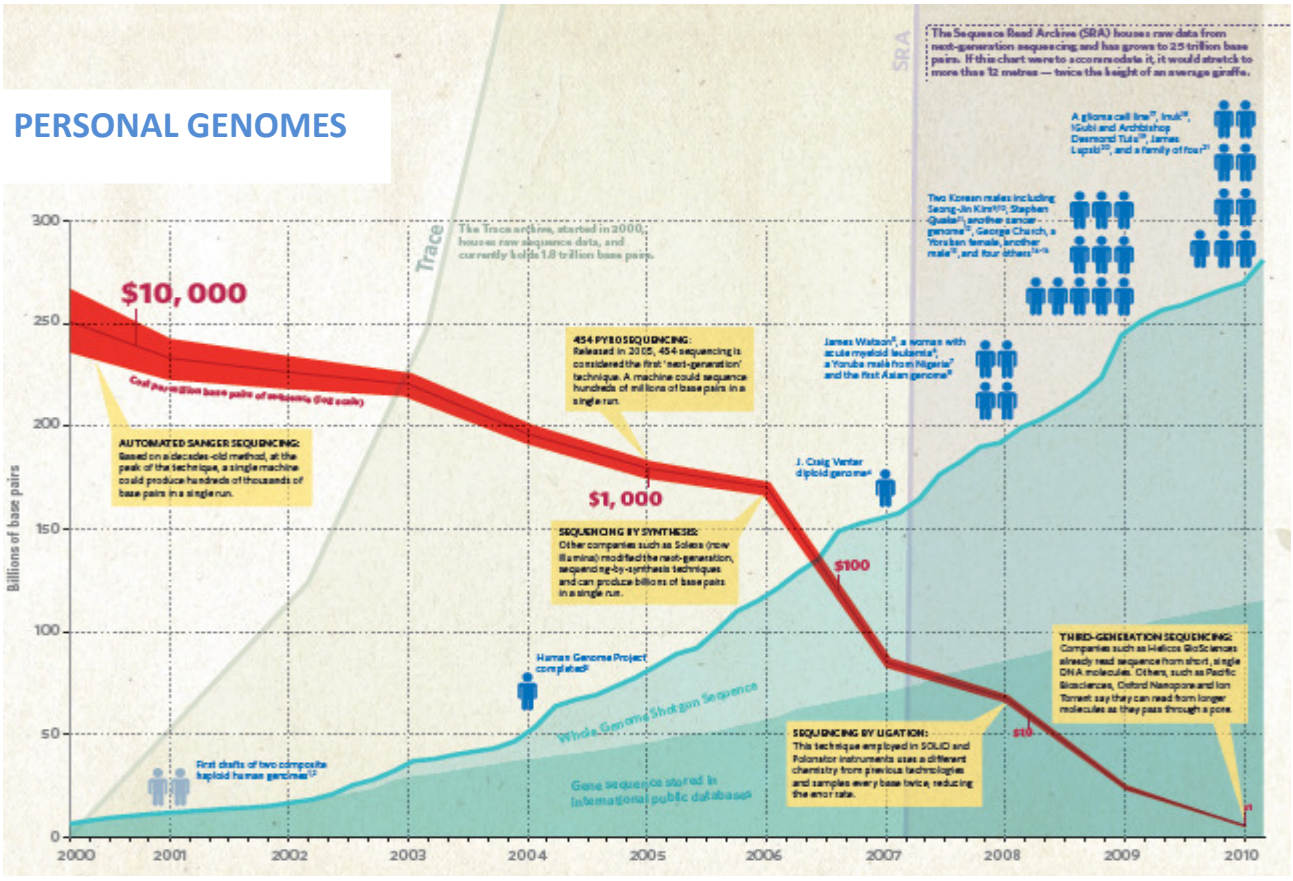

Fig. 6 Personal genomes sequenced in the last decade. The Sequence Explosion NATURE|Vol 464|1 April 2010 pg.671

Personal genome sequencing

The recent improvements in NGS platforms over the pioneering technology have facilitated comprehensive interrogation of personal genomes. Since the release of the first complete human genome sequence, called the reference genome [5], most of the human genome sequencing is done relative to this reference genome and is known as "resequencing". The main aim of most resequencing projects is to detect single nucleotide polymorphism (*SNP*) and *structural variations* in genomes. SNPs are relatively easy to detect and SNPs can be used to conduct genome wide association studies (GWAS) which help determine the link between genotypes and diseases, and they can also be utilized for *pharmacogenomics.* SNP tiling arrays have been used for similar projects but the scope of discovery is limited to sequences that have high degree of similarity to the probe sequence. As a result, SNPs in regions that contain new sequences and in regions with high sequence variability with reference to the probe can be missed. Structural variations consist of large non-SNP variations in the genome and can include insertions, deletions, inversions and copy number variation (CNV). CNVs are a result of deletion, insertion or duplication of genes in the genome, and they are not easy to detect. Large CNVs can be detected with the help of *comparative genomic hybridization* (*CGH)* but many CNVs are small and cannot be resolved by CGH. Until the recent advance in NGS technology which has enabled detection of large range of CNVs, their wide spread existence in the genome was not appreciated [2,5,7,8,11,17].

In the last few years several labs have sequenced personal genomes (Fig. 6). Pushkarev et al. [15] sequenced the genome of a European male on the Heliscope single molecule sequencer (Helicos BioSciences). The single molecule sequencing platform has several advantages over the other NGS platforms: First, it is simple and requires less starting material. Second, it does not require any cloning or amplification, thus reducing cloning vector related biases and probe cross hybridization. Third, a single operator using a single instrument can sequence the entire genome of an individual over a very short period at a great depth, and for a low cost ($48,000; the least expensive of all resequencing projects so far) making it affordable to small labs and medical facilities [10,15].

The Pushkarev et al. [15] study found approximately 2.8 million SNPs. Previously, most genome resequencing projects focused on sites of known genetic variation due to cost and technological limitations. However, in this study with the help of the single molecule sequencer they were able to detect a large number of SNPs in the genome. Profiling and analysis of SNPs across genomes can detect not only new alleles that may be associated with disease, but it can also determine normal genetic variation among individuals, and thus link genotypes with phenotypes [15].

Surprisingly, the Pushkarev study also found 752 CNVs in the genome. This was possible due to deep sequencing by NGS platforms which can provide quantitative information (read density—the number of reads that map to a particular region of the genome) required to determine the copy number of genes in the genome. The widespread existence of CNVs in the genome has come to light only in recent years and their role is still not well understood. CNVs have been linked to genomic

disorders like Charcot-Marie-Tooth disease and the susceptibility or predisposition of individuals to cancer, HIV/AIDS, and they have also been implicated in development of complex traits like schizophrenia and autism. It is speculated that CNVs may alter phenotypes through gene dosage related effects, which may confer beneficial traits to the individual and help in evolution of new phenotypes, or produce abnormal phenotypes due to over or under expressed gene products, which may result in disease. Detection and profiling of CNVs across several individuals with the help of NGS and parallel GWAS that link the CNVs with phenotypes will help elucidate their role in evolution and disease [8,15,17].

The single molecule sequencer can detect a large number of SNPs and also a wide range of CNVs when compared to other conventional methods (microarrays, CGH). However, it cannot detect all the SNPs and it is not very sensitive to CNVs. An increase in read length and paired-end reads will facilitate sensitive identification of all the SNPs, CNVs and other structural variations that exist in the genome. As of now there are no single molecule sequencers that perform paired-end reads and their development along with an increase in read length will be instrumental in addressing these shortcomings. Furthermore, the development of new algorithms, which are less prone to errors when compared to the old software will help realize the full potential of future single molecule sequencers [10,15].

Latest advances in NGS have also revolutionized the field of paleogenomics, i.e., ancient DNA sequencing. Recently, Rasmussen et al. [16] sequenced an ancient human genome, belonging to an extinct Palaeo-Eskimo, from approximately 4,000 years old permafrost-preserved hair, using the Illumina Genome Analyzer. They were the first to sequence an ancient human nuclear genome at 20 fold depth, which is necessary to differentiate between DNA damage, sequencing errors and genetic variation in the genome. The depth of sequencing achieved by NGS platforms is a major advantage when working with ancient DNA samples, which are subject to postmortem DNA damage. Post mortem DNA damage is characterized by base substitutions which result from cytosine to uracil deamination. In order to isolate damaged DNA from the sample, they sequenced genomic libraries resulting from two different DNA polymerase enzymes: one which cannot replicate through uracil, and the other which can replicate through uracil. This method allowed them to effectively detect and isolate damaged DNA. Another challenge in ancient DNA studies is the contamination of the sample with modern human DNA. Their comparison of the ancient DNA sequence with that of modern human DNA sequence data indicated that the contamination due to handling was less than 0.8% and they were able to recover 79% of the diploid genome for further sequencing [16].

Analysis of DNA sequence data helped them deduce interesting traits of the individual. For example, the presence of A1 antigen allele indicated that the individual had A+ blood type, which has a high incidence in ethnic groups from east cost Siberia to mid China. They detected 353,151 SNPs in the genome and by comparing functional SNPs (SNPs associated with specific known phenotypes) in the ancient genome with those of present day populations, they associated certain phenotypes with the individual. Interestingly, the presence of a combination of four SNPs at the HERC2-OCA2 locus,

which are associated with thick, dark color hair and darker skin found in present day Native American and Asian populations enabled them to predict that the individual had thick dark hair (which was also evident from the examination of the sample) and perhaps darker skin. Further, by looking at a set of 12 SNPs which control metabolism and body mass index they suggested that the individual lived in a cold climate. There are very few human remains from this extinct population that settled in Greenland and genotyping of ancient human remains is probably the only way of deducing physical characteristics of this extinct population which has left few marks behind [16].

SNP genotyping analysis can also be used to determine ancestry and the demographic history of a population by comparing against that of existing populations. Surprisingly, they found that the SNP profile of the individual was similar to the SNP profiles of present day populations living on both sides of the Bering Sea. Therefore, they suggest that the individual did not descend from the Native Americans or Inuits, but rather that his ancestors were Arctic north-east Asians who had migrated across the Bering Strait separately from the Native Americans. This finding is in favor of the theory that the Saqqaq population migrated to the New-World separately and at a time much later than the Native Americans or Inuits. Therefore, by analyzing DNA sequence data alone they were able to predict phenotypic traits, genetic origin and relationship to present day populations [16].

Analysis of data from recent whole genome sequencing projects indicates that a large amount of genetic variation exists in genomes and that it is mainly dominated by SNPs and CNVs. It is already known that SNPs may be associated with disease or with normal phenotypic differences between individuals. A large number of SNPs can now be detected with the help of NGS. Besides linking variation with disease, SNP profiling via NGS can be used to determine phenotypic characteristics associated with specific genotypes, to determine ancestry and demographic history, and this data can also be utilized for population genetics and pharmacogenomics. Similarly, with the advent of quantitative NGS platforms that can determine read density, it is now possible to detect a wide range of CNVs in the genome. However, the role of CNVs is still not clear. CNVs are present in the genomes of both normal and diseased individuals, and it is suspected that some CNVs may have a shielding effect against certain diseases, while others may be responsible for susceptibility to certain diseases. It is also believed that CNVs may be responsible for morphological or other phenotypic differences between normal individuals and may thus play a significant role in evolution. Therefore, it is not possible to determine the consequence of the variation by looking at the sequence alone, and GWAS along with other molecular and genetic methods will be necessary to confirm the contribution of SNPs and CNVs to disease and evolution. Furthermore, a single individual is not representative of an entire population and a larger pool of samples will have to be analyzed to obtain concrete information about the characteristics of the population. Nevertheless, in the context of ancient genome sequencing, NGS has opened a window into the past and realized the possibility of recovering useful information from trace archeological human remains. The clinical or archeological significance, respectively, of these findings maybe debatable, but they prove that

genomic research is greatly impacted by the advent of new technology and with future advances in NGS platforms, many more avenues of genomic research will emerge [8,15-17].

Cataloguing of Somatic Mutations in Cancer

The DNA sequence carries the signature of all the mutational events that take place during the development of a disease. Cataloguing mutations from individual cancers and subsequent examination of *mutations* (the type of mutations--insertion, deletion, transition, transversion, CNV, and the nature of mutations--*synonymous, non-synonymous*, homozygous, heterozygous) can help retrieve this information. Deep sequencing of genomes from several cancers and the analysis of mutations across several tissues and individuals can help determine the causative events and identify new targets for medical intervention [6,14].

Acute myeloid leukaemia (AML) is a cancer which is characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells. Certain forms of AML have normal cytogenetics with 46XX/XY karyotype, and no somatic copy number changes can be detected with the help of microarrays. Ley et al. [6] sequenced the *exome* (all the exons or coding regions of the DNA) of tumor cells and normal skin cells of a patient with AML and catalogued all the tumor related *somatic mutations*. Cataloguing of mutations from cancer genomes can be a powerful tool for discovering driver mutations that are responsible for oncogenesis. However, until the recent development of high throughput and cost effective NGS platforms, whole genome sequence analysis of clinical samples was not feasible [6].

Ley et al. [6] used the Illumina Genome Analyzer to sequence both tumor and normal skin diploid exomes. Most acquired mutations in cancer are known to be heterozygous, and diploid genome sequencing, which includes both alleles at a given position, is required to reveal the presence of such mutations. Diploid genome sequencing is now feasible due to the development of high speed NGS platforms. Another major advantage of the Illumina platform is that it requires only 1μg of DNA, which allows direct sequencing of the primary tumor genome without the need to maintain cell lines. This facilitates the recognition of primary tumor induced somatic (acquired) mutations and precludes variations that occur due to evolution or immortalization of cell lines [6].

In order to detect the tumor induced mutations they sequenced both normal and tumor genomes from the same patient, and after subtracting the mutations common to both genomes they were able to identify mutations exclusive to the tumor. After filtering out *synonymous mutations*, they found ten *non-synonymous* somatic mutations. Two were well established AML mutations and the other eight were new mutations which were never previously reported in AML. Four of the new mutations were in genes which were previously not implicated in cancer [6].

This study demonstrates how comprehensive interrogation of the whole exome of tumor and normal cells of the same individual, which is now possible due to the massively parallel sequencing performed on NGS platforms, is crucial to uncovering previously unknown acquired mutations. For

instance, unbiased sequencing made it possible to detect four new genes which may be involved in AML pathogenesis. Profiling of several cancer exomes can help discover new genes involved in cancer pathogenesis and comparison of datasets can help narrow the list of genes with driver mutations and discover new targets for cancer therapy [6].

Nevertheless, the protein coding genes constitute only 1-2% of the entire genome and several mutations in non-coding regions have been implicated in cancer pathogenesis. Therefore, it is imperative to sequence the entire genome to nail down all mutations that can provide selective growth advantage to the cancer cells by indirect regulation of coding genes or other regulatory mechanisms. Pleasance et al. [14] took this step by cataloguing most of the somatic mutations from malignant melanoma (COLO-829/cancer cell line) and lymphoblastoid (COLO-829BL/normal cell line) cell lines derived from a single individual. Malignant melanoma is a cancer of the melanocytes, (which are responsible for producing the dark pigment, melanin, in normal skin) and there are no clinical tests to predict melanomas and the best measure is early detection and surgical removal of thin tumors, which if left undiagnosed can become metastatic and fatal over time. Sequencing complete genomes from normal and cancer cells of a single patient, and cataloguing somatic mutations in coding and non-coding regions of the genome, can help detect variations that play an important role in cancer pathogenesis [6,14].

Pleasance et al. [14] used the Illumina Genome Analyzer with the paired end sequencing feature to detect structural variations in the genome. They sequenced both the cancer and normal genomes and mapped the variants with respect to the reference genome. When a paired end read did not align with the reference genome it was classified as a rearrangement. With the help of this method they found 37 rearrangements in the genome, most of which were CNVs. Surprisingly they detected 8 to 12 fold amplification of a region on chromosome 3 which included a set of four genes, and 4-6 times amplification of a 0.5 Mb region on chromosome 15 which included two genes. Amplification of these genes had previously not been associated with cancer [14].

Further, they also detected several single base substitutions like C>T transitions which are indicative of DNA damage due to exposure to UV light (known risk factor for melanoma) and some G>T transversions which hint at DNA damage due to reactive oxygen species. Interestingly, they scrutinized the whole genome somatic mutation data and retrieved important information like the order of certain mutational and duplication events that may have occurred during the development of the cancer. For instance, the condition in which one of the alleles of a particular gene is missing and the other allele is inactivated due to mutation is referred to as loss of heterozygosity (LOH). When a LOH region of the chromosome re-duplicates the mutation is also re-duplicated and is present as a homozygous mutation, whereas a mutation that occurred after the re-duplication will be heterozygous. When examining a region of chromosome 1q they found few homozygous substitutions and many heterozygous substitutions from which they deduced that the chromosome had an early re-duplication. In addition they report that most of the C>T transitions were homozygous and that most of the G>T transversions were heterozygous. Following the above logic

they suggest that the UV related damage which resulted in C>T transitions (homozygous), occurred at an early stage prior to the re-duplication and somehow stopped after re-duplication. Whereas the G>T transversions (heterozygous) which occurred due to exposure to reactive oxygen species may have occurred at a later stage after re-duplication [14].

This study illustrates how whole genome sequence data can be utilized to detect structural variations (CNVs) which may harbor driver mutations, and how other mutations (transitions or transversions) can provide information about the source of mutations, and also how examination of the nature of mutations (heterozygous or homozygous) can help deduce the sequence of events that take place during the development of a complex disease like cancer. Moreover, it draws attention to the presence of large scale amplification in coding regions of the cancer genome which was previously not known, and could possibly be a common feature of many other cancers. Unbiased whole genome sequencing via NGS thus has the potential to detect most of the somatic mutations that initiate or are involved in tumorigenesis [14].

Cataloguing somatic mutations from normal and cancer diploid genomes of a single patient has become feasible due to the low cost and high throughput of NGS platforms. Diploid genome sequencing is required for detection of heterozygous mutations which account for most of the acquired mutations in cancer, and comparison of normal and cancer genome variants from a single patient is critical for distinguishing the acquired mutations from inherited ones. The acquired mutations (tumor induced mutations) which mainly occur in the exons (coding regions) harbor the driver mutations and exon sequencing can help shortlist the genes with driver mutations and also identify new genes that may be involved in tumorigenesis. However, the non-coding regions contain other mutations whose analysis can help predict the source of mutations and also the sequence of events that lead to cancer. Therefore cataloguing all the somatic mutations in the diploid genome is essential for uncovering all the embedded information and this insight can be instrumental in our fight against cancer. Future cataloguing projects facilitated by NGS will provide enough data for detection of causative genes which can be further validated and targeted for cancer therapy. Eventually, with the driver mutations determined, it will be possible to look at an individual's genomic profile and predict their risk for cancer. In cancer, where the stage of discovery determines the prognosis, this may ultimately help save many lives and shift the emphasis of healthcare towards prevention rather than treatment [6,14].

Next-generation Sequencing: Emerging Applications

Biomedical Applications

Next generation sequencing has been used for diverse biomedical applications like sequencing of mitochondrial DNA to detect mitochondrial disease and sequencing of circulating nucleic acids (CNA) to detect prion diseases in cattle. Mitochondrial diseases are confounding due to mitochondrial DNA heteroplasmy which results in inter-tissue differences, and the heterogeneity

increases with the accumulation of new mutations over the life span of an individual. There is no definite age of onset and a wide range of organs and systems may be involved making diagnosis difficult. Screening of a large number of nuclear and mitochondrial genomes from diseased individuals has the potential to detect disease specific markers that can be used for clinical diagnosis of mitochondrial diseases. In the case of prion diseases, deep sequencing of CNA from serum can be used as early detection test in cattle. It is alleged that CNA result from apoptosis of cells in diseased tissues and they have been shown to carry prion disease specific motifs. These motifs can be detected in blood test samples from live cattle four months prior to the appearance of clinical symptoms. This non-invasive method allows large scale screening of live animals and early detection can prevent the spread of disease [3,19].

Metagenomics

Metagenomics is the study of genetic material directly recovered from environmental samples. NGS is suitable for parallel sequencing of a large number of genomes and its scope is not limited to model organisms. Therefore it is suitable for metagenomic studies which range from sampling of microbes from different niches like the human gut (to study symbiosis and division of labor in the microbial community) and deep sea vents (to study adaptation to extreme environments). Metagenomic analyses can be used to study the effect of industrial pollutants on ecosystems, to monitor the development of disease resistance in infectious strains, and to help develop beneficial microbial industrial strains. The next generation sequencing platforms allow direct sequencing of microbes which are not amenable to culture and thus can detect novel microorganisms which cannot be detected by other means. In agriculture, metagenomic sequencing via NGS platforms has been used to detect viruses and bacteria that infect plants. The early detection of viruses and timely development of pathogen specific treatment can help prevent huge economic losses and also identify novel pathogens [4,9,12].

Transcriptomics

Deep sequencing of the transcriptome relies on a sequence census method referred to as RNA-Seq. The abundance of a transcript is determined by counting the number of reads that map to a particular region of the genome. RNA-Seq is quantitative and has a wide dynamic range. When compared to microarrays, which are subject to hybridization related biases and have a limited dynamic range, RNA-Seq permits sequencing and annotation of rare and transient transcripts. Profiling of RNA facilitates the detection of previously unknown transcriptional and post-transcriptional gene regulatory mechanisms and reveals the complex nature of the transcriptome [2,7].

Epigenomics

Epigenetics is the study of heritable changes associated with chromatin that alter gene expression without changing the DNA sequence. The two main types of epigenetic modifications are DNA

methylation and histone tail modifications. MethylC-Seq and ChIP-Seq are deep sequencing techniques that are used to profile epigenetic modifications. In Methyl-Seq, bisulfite treated DNA is sequenced to reveal the position of methylated cytosines in the genome. In case of ChIP-Seq, regions of DNA that are attached to histones with a particular histone tail modification (methylation, acetylation) are extracted using ChIP and then deep sequenced. Regions of DNA associated with epigenetic modifications have altered transcriptional state and they can play an important role in cancer and development [2,7,9,13].

NGS platforms allow sequencing of genome, transcriptome and epigenome on a single machine, and comparing data generated from different experiments performed on the same sample, can help detect correlation between the different regulatory elements and also facilitate the discovery of new mechanisms of gene regulation [7].

Next Generation Applications of NGS

In the future the completion of the 1000 genomes project (http://www.1000genomes.org), The Cancer Genome Atlas (http://cancergenome.nih.gov/), and many more personal genome and cancer genome sequencing ventures, will provide vast resources for comparison of genetic variations across populations. Furthermore, advances in single molecule sequencing technology which can effectively sequence primary DNA from a single cell will enable somatic evolutionary genomics. The sampling of several genomes within an individual and within tumor tissues can reveal intra-tissue genetic heterogeneity which seems more common and widespread than previously thought. This insight will change the way we look at the initiation and progression of disease. Large scale GWAS that link genetic variations with disease will enable personalized medicine and pharmacogenetics [2,9,10].

The future of NGS looks very bright with the exponential drop in price of genome sequencing (Fig. 6) it will soon be possible to sequence entire genomes, transcriptomes, epigenomes, metagenomes for $1000 or less. The improved and more robust NGS platforms of the future will facilitate advanced applications like forensic detection, astobiological surveys and other applications which are limited only by the imagination [2,7,9,13].

References

1. Bell D.W. Our changing view of the genomic landscape of cancer. Journal of pathology 220, 231-243 (2010).
2. Gilad, Y. et al. Characterizing natural variation using next generation sequencing technologies. Trends in Genetics. 25, 463-471 (2009).
3. Gordon et al. Disease-specific motifs can be identified in circulating nucleic acids from live elk and cattle infected with transmissible spongiform encephalopathies. Nucleic Acids Research, 37, 550–556 (2009).

4. Kreuze, J.F. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. Virology. 388, 1-7 (2009).
5. Lander et al. Initial sequencing and analysis of the human genome. Nature. 409,860-921 (2001).
6. Ley et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 456, 66-72 (2008).
7. Lister, R. et al. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Current Opinion in Plant Biology. 12, 107-118 (2009).
8. Lupski, J.R. et al. Copy Number Variation in Human Health, Disease and Evolution. Annual Rev. Genomics and Human Genetics. 10,451–81 (2009).
9. Mardis, E.R. The impact of next-generation sequencing technology on genetics. Trends in Genetics. 24, 133-141 (2008).
10. Metzeker, M.L. Sequencing technologies – the next generation. Nature Reviews Genetics. 11,31-46 (2010).
11. Michael, S. et al. Personal genome sequencing current approaches and challenges. Genes and Development. 24, 423-431 (2010).
12. Mitchelson, K.R. et al. Overview: Developments in DNA sequencing. In: Mitchelson, K.R. editor, Perspectives in bioanalysis, Vol 2. New high throughput technologies for DNA sequencing and genomics. Oxford: Elsevier. Pg 29-30. 2007.
13. Morozova, O., Marra, M.A. Applications of next-generation sequencing technologies in functional genomics. Genomics. 92, 255-264 (2008).
14. Pleasance et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 463, 191-196 (2010).
15. Pushkarev, D., Neff, N., Quake, S. Single-molecule sequencing of an individual human genome. Nature Biotechnology. 27: 847–852 (2009).
16. Rasmussen et al. Ancient human genome sequence of an extinct Paleo-Eskimo. Nature. 463, 757-762 (2010).
17. Stankiewicz, P. and Lupski, J.R. Structural Variation in the Human Genome and its Role in Disease. Annual Review of Medicine. 61, 437-455 (2010).
18. Frank, S.A. Somatic evolutionary genomics: Mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. PNAS. 107, 1725-1730 (2010).
19. Vasta, V. et al. Next generation sequence analysis for mitochondrial disorders. Genome Medicine. Volume 1, Issue 10, Article 100 (2009).
20. Venter, J.C. Multiple personal genomes await. Nature. 464, 676-677 (2010).