3 Next-Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine

Huixiao Hong, Wenqian Zhang, Zhenqiang Su, Jie Shen, Weigong Ge, Baitang Ning, Hong Fang, Roger Perkins, Leming Shi, and Weida Tong

Abstract

Personalized medicine can improve healthcare by selecting treatments that are more efficacious or induce less adverse responses in stratified cohorts sharing differentiating genetic traits. Personalized medicine has advanced quickly, providing both opportunities and challenges for the pharmaceutical industry and regulatory agencies in the twenty-first century. Pharmacogenomics is the key to the identification of personalized medicine biomarkers useful for efficacy and safety that can ultimately be clinically applied for diagnosis, prognosis, and treatment selection. The requisite technologies and approaches needed for pharmacogenomics have steadily advanced over more than a decade in terms of both capability and cost. In 2005, 454 Life Sciences announced their sequencing-by-synthesis technology, the first next-generation sequencing (NGS) platform, proclaiming the breakthrough in sequencing technology. NGS is revolutionizing pharmacogenomics and personalized medicine, with several NGS platforms commercially available. Illumina, Roche 454, and Applied Biosystems are the current major vendors. This chapter will characterize the technical assessments of NGS, including comparative analyses across platforms, experimental protocols, algorithms for mapping short reads to reference

The views presented in this article do not necessarily reflect those of the US Food and Drug Administration.

H. Hong (✉) • Z. Su • J. Shen • W. Ge • R. Perkins
L. Shi • W. Tong
Division of Bioinformatics and Biostatistics,
National Center for Toxicological Research,
US Food and Drug Administration, 3900 NCTR Road,
Jefferson, AR 72079, USA
e-mail: huixiao.hong@fda.hhs.gov

W. Zhang
Beijing Genomic Institute, Beishan Industrial Zone,
Yantian District, Shenzhen, Guangdong 518083, China

B. Ning
Division of Systems Biology, National Center
for Toxicological Research, US Food and Drug
Administration, 3900 NCTR Road, Jefferson,
AR 72079, USA

H. Fang
Office of Scientific Coordination, National Center
for Toxicological Research, US Food and Drug
Administration, 3900 NCTR Road, Jefferson,
AR 72079, USA

D. Barh et al. (eds.), *Omics for Personalized Medicine*,
DOI 10.1007/978-81-322-1184-6_3, © Springer India 2013

genomes, strategies for quantitatively measuring expression levels, and methods for detecting single nucleotide polymorphisms (SNPs). Different pipelines and software packages for analyzing NGS data will be reviewed. Examples and a prospective outlook on applications of NGS in pharmacogenomics and personalized medicine will be given.

1 Introduction

1.1 Personalized Medicine

The term "personalized medicine" has been used for some 13 years (Langreth and Waldholz 1999) in the context widely understood today of "the right drug for the right patient with the right dose at the right time through the right route." The President's Council of Advisors on Science Technology (PCAST) has a more comprehensive definition for personalized medicine: "'Personalized medicine' refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not" (President's Council of Advisors on Science Technology 2008).

The rapid increase in the number of published articles per year referring to personalized medicine, shown in Fig. 3.1, is a testament to the substantial and accelerating scientific interest (Jorgensen 2009) that is driven by real medical needs and fostered by rapid advances in information-rich and high-throughput technologies. While technology trends and palpable scientific excitement has moved personalized medicine from futuristic to realistic, a lack of demonstrable progress in the clinical appearance of personalized medicine related biomarkers portends substantial challenges remain for both the pharmaceutical industry and regulatory agencies.

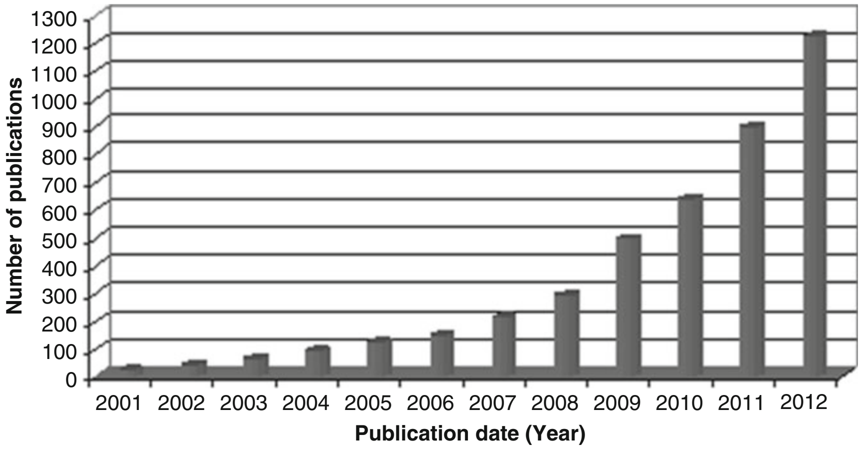

Fig. 3.1 Annual number of publications related to personalized medicine from 2000 to 2011 based on a keyword search in PubMed. Keyword used: personalized medicine. Fields searched: title and abstract (Search was conducted on August 31, 2012. The number of publications in 2012 was projected from 815 for the first 8 months to 1,219 for the whole year)

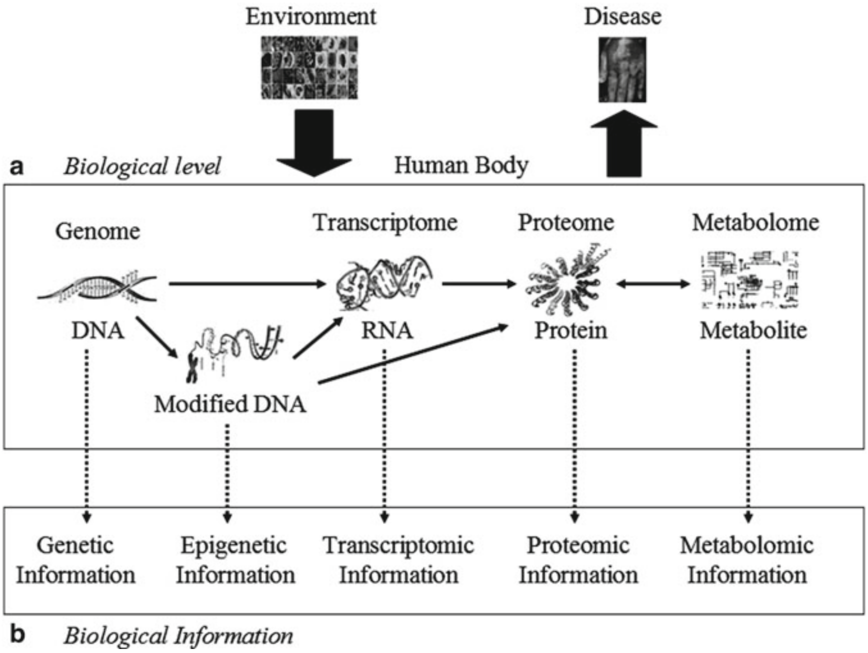

Fig. 3.2 Taxonomy of molecules used for biomarkers. *Box a* depicts the biological levels and associated molecules that can be identified as biomarkers with the appropriate technologies. *Box b* gives the corresponding categories of the molecular biomarkers based on the type of molecule class

1.2 Pharmacogenomics

Adequate individuals' biological and lifestyle (environment) information is an essential prerequisite for realizing personalized medicine applications. The biological information comprises the molecules and their measurements for individuals that genetically distinguish a subpopulation and can therefore be translated into clinical practices for personalized medicine. Figure 3.2 provides a taxonomy of molecular classes pertinent for personalized medicine. Most active pharmacogenomics research fully utilizes these molecular classes based on DNA and RNA measurements yielding genetic, epigenetic, and transcriptomic information.

Pharmacogenomics is a scientific research field that attempts to explain how genomic and genetic variations affect a patient's response to a drug. Motulsky first reported a pharmacogenomics study in 1957 (Motulsky 1957). That study found enzymes for the metabolism of a number of drugs exhibited variation in activity among patients that were correlated with the adverse reactions to those drugs. Since then, pharmacogenomics has enormously advanced in accordance with technologies, leading to the discovery of genetic variants that can be used as markers to monitor the efficacy and safety of drugs on the market (Hong et al. 2010).

Some 3.1 million common SNPs in human populations have been identified by the HapMap project (The International HapMap Consortium 2007). In concert, state-of-art high-throughput SNP genotyping technology enables simultaneously genotyping of hundreds of thousands of SNPs, making genome-wide association studies (GWAS) a promising pharmacogenomics research field for linking genetics with biological response. Some GWAS have found associations between genetic variations of patients and their therapeutic responses to drugs such as thiazide diuretic (Turner et al. 2008), warfarin (Takeuchi et al. 2009), and iloperidone (Lavedan et al. 2008). Other GWAS have found associations between genetic variants and the adverse events caused by drugs such as the elevation of serum alanine aminotransferase by ximelagatran (Kindmark et al. 2008) and drug-induced liver injury due to flucloxacillin (Daly et al. 2009). Both types of studies are important as risk assessment considers both efficacy and safety.

1.3 Biomarkers in Pharmacogenomics

In this book chapter, a biomarker in pharmacogenomics is a molecule or a set of molecules of DNA or RNA that can be measured and used for purposes such as characterizing financial risk in pharmaceutical drug development, quantifying risk or benefit to inform regulatory decisions, and ultimately predicting clinical response (including selecting clinical treatment or tailoring dose to patient) (Hong et al. 2010). Here we classify pharmacogenomics biomarkers into three types: genetic biomarkers, epigenetic biomarkers, and transcriptomic biomarkers, as depicted in Fig. 3.2.

1.3.1 Genetic Biomarkers

Genetic biomarkers are used in personalized medicine practices mainly for selecting those patients who benefit more or have less risk of adverse drug reactions (ADRs) from a particular drug. For example, human leukocyte antigen (HLA) allelic marker, HLA-B*5701, is a genetic biomarker that can be used to predict increased risk of developing hypersensitivity reactions to the antiretroviral drug, abacavir. Thus, it is recommended that the marker's presence is assessed in patients with human immunodeficiency virus (HIV) prior to treatment (Lucas et al. 2007; Mallal et al. 2008; US Food and Drug Administration 2012).

Investigation of biomarkers in drug-metabolizing genes may lead to a better understanding of drug efficacy. Drug-metabolizing genes encode proteins responsible for metabolizing exogenous toxicants including drugs; genetic polymorphisms are prevalent in these genes. Gene deletions, missense, nonsense, and splice site mutations can alter or abolish enzyme activity, whereas mutations causing amino acid substitutions can lead to markedly modified enzyme action. A good example is the anticoagulant warfarin (a commonly used blood thinner). Warfarin is metabolized by the enzyme CYP2C9 that exhibits wide genetically based activity variation among patients (Sanderson et al. 2005; Takahashi et al. 2006). There is a poor efficacy to safety margin with this drug and, since individual response to warfarin varies so greatly, that patients have to be closely monitored to adjust the dose during treatment to prevent hemorrhage that can result in fatal strokes, etc. In addition, the target of warfarin is vitamin K epoxide reductase and polymorphisms in its gene (VKORC1) account for some of the dosing variation among individuals, with certain haplotype groups requiring larger doses and other haplotype groups requiring lower doses, to reach safe therapeutic levels. Studies have identified variants associated with warfarin adverse reactions that resulted in the FDA updating the label for warfarin, stating that "the patient's *CYP2C9* and *VKORC1* genotyping information, when available, can assist in selection of starting dose" (Kim et al. 2009).

Another example of biomarkers enabling personalized medicine is the genetic testing of mutations in the KRAS gene (Kirsten *ras*) to avoid treatment of patients with metastatic colorectal cancer who cannot benefit from the use of panitumumab (Vectibix®) and cetuximab (Erbitux®). KRAS is a protein encoded by the KRAS gene (McGrath et al. 1983); mutations in the KRAS gene make potent oncogenes and the protein products of oncogenes play a role in many cancers (Kranenburg 2005). Pharmacogenetics studies revealed that the presence of mutations in the KRAS gene was associated with the poor response to panitumumab or cetuximab therapy in patients with colorectal cancer. Accordingly, drug labels were updated by the US Food and Drug Administration (FDA) and suggest that when these two anti-epidermal growth factor receptor (EGFR) antibody drugs are used for the treatment of patients with metastatic colorectal cancer, gene tests for *KRAS* mutations are recommended (Hong et al. 2010).

Table 3.1 lists the genetic biomarkers mentioned in some drug labels approved by the FDA. In addition, to promote drug efficacy and drug

Table 3.1 Genetic biomarkers mentioned in labels of drug products approved by FDA (Hong et al. 2010)

Biomarker	Drug	Section in label
CYP2C19	Plavix®	Clinical pharmacology, precautions, dosage and administration
	Vfend®	Clinical pharmacology
	Effient®	Use in specific populations, clinical pharmacology, Clinical Studies
	Celebrex®	Clinical pharmacology
CYP2C9	Celebrex®	Clinical pharmacology
	Effient®	Use in specific populations, clinical pharmacology, clinical studies
	Coumadin®	Clinical pharmacology, precautions
CYP3A4	Celebrex®	Clinical pharmacology
	Codeine sulfate	Drug interactions, clinical pharmacology
CYP3A5	Effient®	Use in specific populations, clinical pharmacology, clinical studies
CYP2B6	Effient®	Use in specific populations, clinical pharmacology, clinical studies
CYP2D6	Strattera®	Dosage and administration, warnings and precautions, drug interactions, clinical pharmacology
	Prozac®	Clinical pharmacology, precautions
	Codeine sulfate tablets	Warnings and precautions, drug interactions, use in specific populations, clinical pharmacology
VKORC1	Coumadin®	Clinical pharmacology, precautions
UGT1A1	Camptosar®	Clinical pharmacology, warnings, dosage and administration
	Tasigna®	Drug interactions, clinical pharmacology
HLA-B*1502	Tegretol®	Warnings, precautions
HLA-B*5701	Ziagen®	Warnings and precautions
Deletion 5q	Revlimid®	Hematologic toxicity, clinical studies, precautions, adverse reactions

safety, FDA maintains a database (http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm) of genetic variants that affect the treatment outcomes of some drugs. Included are genetic biomarkers and related advices/warnings that are listed on the drug labels that indicate efficacy differences and possible adverse reactions among patient with certain genotypes.

1.3.2 Epigenetic Biomarkers

While DNA sequence of an organism may not be changed, nongenetic factors, environmental, and lifestyle-related influences such as nutrition and exposure to stress can induce epigenetic alteration that causes the organism's genes to behave (or express themselves) differently (Bird 2007). The genetic background of a person can provide information at the individual level on the possibility of developing some diseases and responses (efficacy and safety) to some drugs, while epigenetic alterations might be more directly associated with the phenotype observed. Consequently, epigenetic biomarkers hold great promise for personalized medicine.

Numerous mechanisms are involved in gene regulation, among which, chromatin remodeling (such as DNA methylation and histone methylation), microRNAs (miRNAs, the term will be used hereafter), and other noncoding RNAs are the most studied (Bird 2007; Berger et al. 2009). Chromatin remodeling is an approach of epigenetic alteration that is achieved either by the post translational modification of the amino acids that make up histone proteins or by adding methyl groups to the DNA (most likely at CpG sites) to convert cytosine to 5-methylcytosine. Though epigenetic biomarkers of histone modification (Kondo et al. 2008) and noncoding RNAs (He et al. 2005; Lu et al. 2005) have been reported, the most common ones are the biomarkers identified based on DNA methylation. The epigenetic biomarkers being pursued so far are mainly utilized for the diagnosis and prognosis of cancers, some of which will be discussed in detail below.

Cancer epigenetic biomarkers can be simply categorized as diagnostic and prognostic biomarkers based on their clinical applications.

Table 3.2 Fluids and tissues that have been used in discovery of epigenetic biomarkers

Fluid/tissue	Cancer type	Epigenetic biomarkers (methylated genes)
Biopsy	Prostate	GSTP1, RARβ2, APC, TIG1
Ejaculate	Prostate	GSTP1
Nipple fluid	Breast	RASSF1
Peritoneal	Ovary	BRCA1, RASSF1
	Breast	RASSF1, APC, DAPK
Serum	Colorectum	SEPT9, TMEEF2, NGFR
	Ovary	BRCA1, RASSF1
Sputum	Lung	$P16^{INK4A}$, RASSF1, MGMT
Stool	Colorectum	SFRP2, CDKN2A, hMLH1
Urine	Bladder	RASSF1, APC, $p14^{ARF}$, DAPK, BCL2, TERT
	Prostate	GSTP1, RARβ2, APC, RASSF1

It has been observed that DNA methylation patterns are often altered in early-stage tumors (Weisenberger et al. 2006; Irizarry et al. 2009). For example, methylation detection in conjunction with a cervical Pap test is more sensitive to diagnose lethal cervical cancers than the Pap test alone (Kahn et al. 2008). For the most part, to be potentially useful in clinical practice, diagnostic epigenetic biomarkers identified from easily obtainable and readily available body fluids and tissues are the most valuable. Table 3.2 lists some fluids and tissues that have been used in epigenetic biomarkers development (Duffy et al. 2009). However, one of the challenges for developing diagnostic biomarkers is the requirement of low false positive rates when biomarkers are used for screening in an overall healthy population since a moderately low false positive rate can lead to a substantial number of unnecessary follow-up examinations in healthy individuals. In contrast, epigenetic biomarkers used as prognostic indicators do not present the same concern in terms of specificity. In addition, many epigenetic biomarkers were also developed for classifying disease subtypes or disease stages among already diagnosed individuals (Weisenberger et al. 2006).

1.3.3 Transcriptomic Biomarkers

DNA microarrays were introduced for analyzing gene expression profiles approximately 17 years ago (Schena et al. 1995; Lockhart et al. 1996). They enable measuring expression levels of thousands of genes in a single experiment. This technology has become ubiquitously applied across research and drug development. The basic principle behind DNA microarrays experiments is the base-pairing hybridization between two DNA strands as complementary nucleic acid sequences form hydrogen bonds between complementary nucleotide base pairs. A typical experiment for measuring gene expressions using DNA microarrays consists of the following steps for a two color array: mRNA molecules in the cells/tissues of interest are extracted and collected, the mRNA molecules are labeled by attaching fluorescent nucleotides to the cDNAs that are generated by using a reverse transcriptase enzyme (usually samples labeled with special fluorescent dyes), the labeled cDNAs are added onto a DNA microarray slide and hybridized to their complementary DNAs attached on the microarray slide, and finally, a special scanner is used to measure the fluorescent intensity for each spot/areas on the microarray slide. In this manner, DNA microarrays represent a mature technology for measuring transcriptome expression for specific cell/tissue and differential expressions between disease and healthy populations. Microarrays can thus be applied to discover transcriptomic biomarkers for personalized medicine.

Some transcriptomic biomarkers are validated for the application in clinical diagnosis and risk assessments. For example, in 2007, FDA approved a DNA microarray-based diagnostic kit (MammaPrint®) that measures the transcription level of 70 genes in breast cancer patients (Van't Veer et al. 2002; van de Vijver et al. 2002; Glas et al. 2006; Buyse et al. 2006). The profiles of the 70-gene signature are usable by physicians as a prognosis test but only for breast cancer patients who are less than 61 years old with Stage I or Stage II disease, tumor size ≤5.0 cm, and who are lymph node negative. Another example of transcriptomic biomarkers is Onco*type* DX®, a 21-gene assay (Paik et al. 2004; Goldstein et al. 2008) approved by the FDA in 2005 and currently widely used for lymph node negative breast cancer patients whose tumors tested positive for hormone receptors (Harris et al. 2007). Based on

Table 3.3 Representative technologies used in pharmacogenomics

Type	Technology	Detection	Principal	Application
LCA	PCR-RFLP	Gel-based	Restriction site	Genotyping
	AS-PCR	Gel-based and homogeneous fluorescence	AS amplification	Genotyping
	TaqMan-PCR	Homogeneous fluorescence	FRET quenched hydrolysis probes	Genotyping and gene expression
HCA	Microsphere array	Flow cytometry fluorescence	Bead-immobilized oligonucleotide	Genotyping and gene expression
	DNA microarray	Fluorescence	Oligonucleotide hybridization	Genotyping and gene expression
	NGS	Fluorescence	Pyrosequencing and sequencing-by-synthesis or ligation	DNA and RNA sequencing

Abbreviations: *LCA* low content analysis, *HCA* high content analysis, *PCR* polymerase chain reaction, *RFLP* restriction fragment length polymorphism, *AS* allele-specific, *NGS* next-generation sequencing

results from multi-institutional clinical trials, the 21 genes (*MK167*, *STK15 (AURKA)*, *BURC5*, *CCNB1*, *MYBL2*, *MMP11*, *CTSL2*, *GRB7*, *HER2 (ERBB2)*, *GSTM1*, *CD68*, *BAG1*, *ESR1*, *PGR*, *BCL2*, *SCUBE2*, *ACTB*, *GAPDH*, *RPLPO*, *GUS*, and *TFRC*) are validated to be able to predict recurrences, deaths, and responses to chemotherapy for patients with estrogen-positive breast cancer (Paik et al. 2004).

Given the amount of resources (including current RNA sequencing efforts) being applied to find transcriptomic biomarkers, it is expected that the number of qualified biomarkers will grow.

1.4 Technologies for Pharmacogenomics

Rapid advances in technologies for measuring genomic variations that affect patients' responses to drugs have been responsible for accelerating pharmacogenomics research that is potentially extensible to personalized medicine. The technologies that have been used in pharmacogenomics can be conveniently divided into two types based on the number of prospective genomic or genetic markers that can be simultaneously analyzed: low content analysis (LCA) and high content analysis (HCA) technologies. Table 3.3 summarizes a few of the representative technologies; there was no intent to provide a complete list.

LCA technologies are used when a few SNPs within a single gene (or a few genes) need to be genotyped or the gene expression levels for a few genes need to be measured. Typical LCA technologies include methods such as PCR-restriction fragment length polymorphism (PCR-RFLP), allele-specific PCR (AS-PCR), TaqMan-PCR, hybridization probe-melting analysis, and oligonucleotide ligation-PCR reactions (Newton et al. 1989; Syvanen et al. 1990; Holland et al. 1991; Barany 1991).

PCR is a biochemical technique that has been used for some 30 years to generate millions of copies from a single DNA molecule or from a few DNA molecules. A pair of primers, complementary to the 3′ ends on each of the two strands of the DNA molecule, is used in the amplification. A typical PCR experiment consists of 15–40 cycles, with each cycle having several temperature changes. The amplification in PCR is achieved in three major steps: denaturation, annealing, and elongation. First, denaturation is conducted by heating the reaction to a preset temperature in a short time to melt the DNA template by breaking the hydrogen bonds. In the annealing step, the temperature of the reaction is lowered to a preset temperature (depending the primers used) quickly (usually about 15–40 s) to anneal the primers. In the elongation (or extension) step, a new DNA molecule complementary to the DNA template sequence from the primers is synthesized using an enzyme (Tag DNA polymerase) by adding dNTPs (deoxyribonucleotide triphosphate) to the template in the 5′–3′ direction. PCR is a relatively mature and sensitive technology for measuring expression levels of genes. It has been broadly applied in biomedical research and diagnosis of diseases.

PCR-restriction fragment length polymorphism (PCR-RFLP) technology was an old fashion technology that was used for genotyping common SNPs within candidate genes, especially when a variant naturally occurs at a restriction enzyme cutting site (Goldstein and Blaisdell 1996). This technology is labor intensive and difficult to automate. More importantly, errors happen frequently due to incomplete restriction digestion. It has been replaced, generally, by other more efficient and accurate genotyping methods.

In allele-specific PCR (AS-PCR), two homologous primers that vary only in their 3′ ends (one is complementary to the normal allele and another is complementary to the variant allele) are differentially amplified (Newton et al. 1989). Gel or homogeneous fluorescent dyes are then used to detect the PCR reaction products (Higuchi et al. 1992). In the PCR amplification, undesired side reactions such as primer dimerization can generate noise that confounds the signal from the fluorescent dye bound to the expected reaction products. Thus, optimizing PCR amplification selectivity is vital to data quality. However, the homogeneous real-time fluorescent dye-based detection is easier to automate than the gel-based electrophoretic analysis.

Many genotyping methods use direct probe, as opposed to the complement, sequence that is homogeneously amplified for detecting variants. The most popular one is the TaqMan-PCR that uses 5′-nuclease hydrolysis probes (i.e., the TaqMan probes) (Holland et al. 1991). These are designed in a way that a dye in the native oligonucleotide probe quenches the reporter fluorophore. In amplification, a TaqMan probe hybridizes to its complementary oligonucleotide and is cleaved by the 5′-nuclease action of thermostable DNA polymerases in the primer extension and, thus, the quencher and fluorescent reporters are separated and released for the next cycle. The fluorescent signal of the reporter dye from the TaqMan probes can be quantified in real time. The measured fluorescent intensity steadily strengthens with each PCR cycle and reaches a threshold that is preset for determining whether homozygous or heterozygous alleles are contained in the genotyping sample. The number of SNPs that can be simultaneously genotyped in a single TaqMan-PCR is determined by fluorescent signals that can be measured within a single PCR reaction. The TaqMan-PCR instruments on the market are capable of simultaneously genotyping four to six SNPs.

LCA technologies have their limitations in time, cost, and material efficiency. HCA technologies are the solution when a large number of genomic or genetic markers need to be interrogated. HCA technologies are usually based on hybridization or single-base extension technologies, which makes high-throughput SNP genotyping and gene expression profiling possible. The early stage of HCA technologies were array-based assays, such as oligonucleotide (Cronin et al. 1996) and bead-based microarrays (Armstrong et al. 2000) for simultaneously genotyping a large number of genetic variants.

The microsphere array is a bead-based genotyping technology that became available early in the last decade. The technology is based on attaching oligonucleotide anti-tag sequences to polystyrene microspheres with different dyes that can be read using a laser instrument. Tag complementary oligonucleotides covalently combined with allele-specific primer extensions can be detected. This technology allows tens of SNPs to be simultaneously genotyped in a single tube.

GWAS became a promising research approach for pharmacogenomics 5 years ago when the genotypes of more than 3.1 million common SNPs in human populations were determined by the HapMap project (The International HapMap Consortium 2007), and high-throughput SNP genotyping technology was advanced enough to enable simultaneous genotyping of hundreds of thousands of SNPs using high-density oligonucleotide microarrays. However, the common genetic variants that were identified by GWAS typically contributed only to a small portion of the total variation in the phenotype (Frazer et al. 2009). Rare genetic variants having high penetrance were also found to contribute to the phenotypes of interest, e.g., blood pressure (Ji et al. 2008). Consequently, the best way for obtaining biological information on DNA is to have the single-base resolution for the DNA sequence of an individual, which can be used to interrogate all genetic variants, both common and rare. Next-generation sequencing is currently the best choice for most applications and will be discussed below.

2 Next-Generation Sequencing

2.1 Background

The Sanger method was invented in 1977 (Sanger et al. 1977). The method involves DNA synthesis in the presence of chain-terminating inhibitors followed by electrophoresis. It is noted for excellent accuracy and reasonable read length but very low throughput and high expense, rendering it unsuitable for deciphering the human genome, crucial information for realizing personalized medicine. In 2005, 454 Life Sciences announced their breakthrough sequencing-by-synthesis technology, the first next-generation sequencing (NGS) platform (Margulies et al. 2005). Since then there has been remarkable advances in DNA sequencing technologies known as NGS or as massively parallel sequencing (Shendure and Ji 2008; Reis-Filho 2009; Ansorge 2009; Voelkerding et al. 2009; Metzker 2010). Currently, the Illumina HiSeq-2000 and HiScan, the Roche 454 GS-FLX, and the Applied Biosystems SOLiD Analyzer 5500xl are the most used and commercially available platforms, albeit the Illumina platforms dominate the market. In addition to the platforms that already exist on the market, new NGS platforms are under development and mainly adopt single DNA molecule sequencing technology (e.g., nanotechnology and electron microscopy) which can read through DNA templates in real time without amplification, and thus could be more accurate with potentially longer reads (e.g., Pacific BioSciences RS system produces reads of >1,000 bp; nanoAnalyzer from BioNanomatrix, now BioNano Genomics, generates reads of around 400,000 bp (Das et al. 2010)).

2.2 NGS Workflow

To date, NGS has been successfully applied to different aspects of applications, such as RNA sequencing for profiling gene expression (also called RNA-seq or whole transcriptome sequencing) (Mortazavi et al. 2008; Wang et al. 2008a, Nagalakshmi et al. 2008; Cloonan et al. 2008), chromatin immunoprecipitation followed by sequencing (ChIP-seq) for identifying DNA-binding sites of proteins (Johnson et al. 2007; Park 2009; Schmidt et al. 2010), sequencing to identify methylated DNA (methyl-seq) (Brunner et al. 2009; Hormozdiari et al. 2009) for DNA methylation analysis, whole human genome sequencing (Wheeler et al. 2008; Wang et al. 2008b) for determination of genetic variants, and targeted sequencing of specific candidate genes or the entire human exome in large numbers of individuals (Hodges et al. 2007). The common workflow for studies using NGS technologies is depicted in Fig. 3.3, with main aspects of the process described below.

2.2.1 Library Generation

The first step in NGS is to make a library from a DNA or a cDNA sample. Different NGS platforms use divergent protocols and reagent kits for the library preparation. In practice, the instruction from a specific NGS vendor should be strictly followed to pursue the quality of the library. Here we briefly review the principle and procedure of the library preparation for the Illumina HiSeq-2000 platform, the most popular platform on the market, though readers are recommended to follow Illumina's instruction for detailed procedures for library preparation.

The Illumina HiSeq-2000, an upgrade of the Illumina GA system that was the first short-read sequencing platform, currently dominates the NGS platform market (Metzker 2010). Figure 3.4 depicts the principle and procedure of library

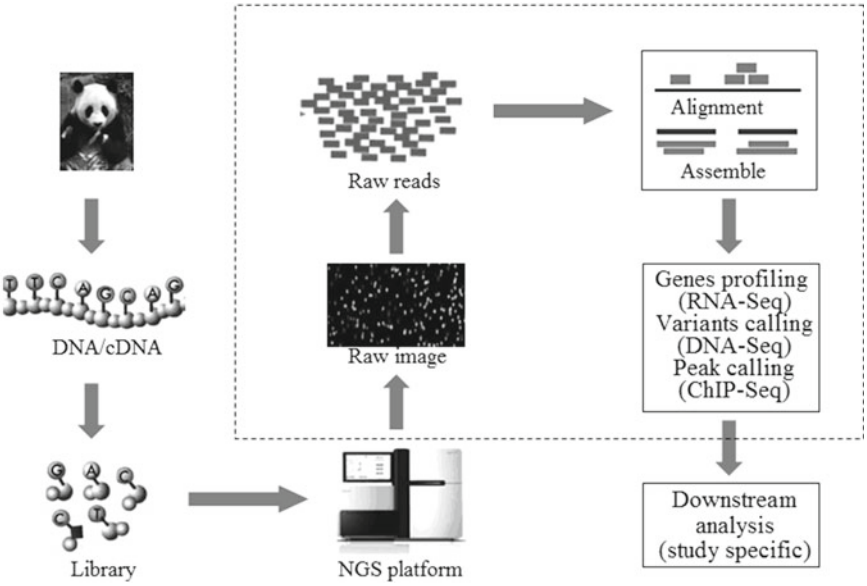
Raw reads
Alignment
Assemble
DNA/cDNA
Raw image
Genes profiling (RNA-Seq)
Variants calling (DNA-Seq)
Peak calling (ChIP-Seq)
Library
NGS platform
Downstream analysis (study specific)

Fig. 3.3 Overview of a common NGS workflow

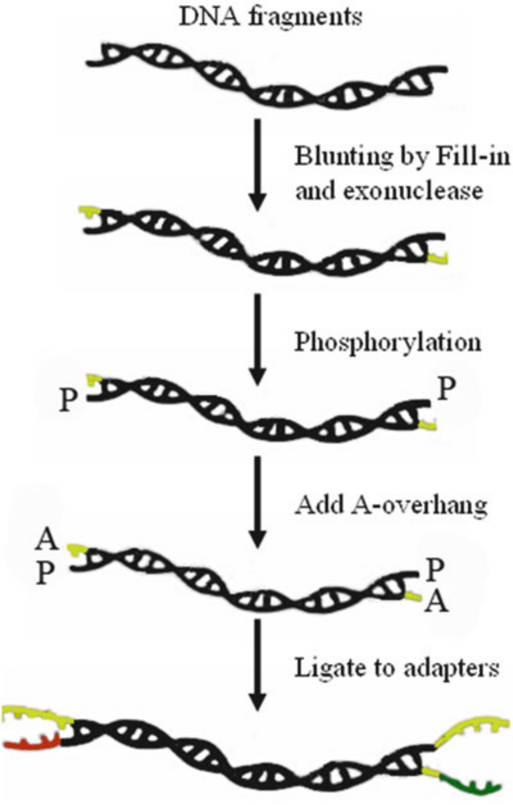

Fig. 3.4 Library preparation of Illumina NGS platform

preparation for the Illumina platform. DNA samples are first sheared into fragments that are then end-repaired to generate 5′-phosphorylated blunt ends. The Klenow fragment of DNA polymerase is then used to attach a single “A” base to the 3′ end of each DNA fragments, which enables the DNA fragments for ligation to oligonucleotide adapters. After ligation to adapters at both ends, the DNA fragments are denatured, and single-stranded DNA fragments are attached to reaction chambers that are optically transparent solid surfaces called a flow cell. The attached DNA fragments are extended and amplified by bridge PCR amplification.

2.2.2 Sequencing

Different chemical reactions for sequencing are used in current NGS platforms. The main features of three most popular NGS platforms are compared in Table 3.4. It is important to note that these metrics are constantly changing when newer models of the same platforms are released. Both the Illumina HiSeq-2000 system and the SOLid 5500xl system use the short-read sequencing technologies, while the Roche 454 FLX system provides relatively longer reads that could be advantageous for de novo sequencing of new genomes.

Table 3.4 Comparison of NGS platforms

	Platform		
	Illumina	Roche 454	SOLid
Sequencing reaction	Sequencing- by-synthesis	Pyrosequencing	Ligation-based sequencing
Amplification	Bridge PCR	Emulsion PCR	Emulsion PCR
Read length	~100 bp	~700 bp	~75 bp
Paired ends/separation	Yes/200 bp	Yes/3,000 bp	Yes/3,000 bp
Reads per run (in millions)	3,000	1	1,500
Run time	11 days	23 h	8 days
Comments	Most widely used	Longer reads, fast run, higher cost	Good data quality

The Illumina HiSeq-2000 system currently is the most widely used short-read sequencing platform. It uses the sequencing-by-synthesis (SBS) method. By using the SBS technology, the bases of a DNA fragment are determined using a proprietary reversible terminator-based method when they are synthesized to “grow” the DNA strands. Since many synthetic reactions are carried out simultaneously in a reaction chamber (flow cell), the sequences of millions of DNA fragments are determined in parallel. To be specific, when synthesizing a base, a fluorescence-labeled terminator-bound dNTP is added to the strand and imaged. Then, the terminator is cleaved for the synthesis of the next base. Since there are four reversible terminator-bound dNTPs (dATP, dCTP, dGTP, and dTTP) in each sequencing cycle, natural competition minimizes incorporation bias. During each sequencing cycle, bases are called directly from the measured fluorescence intensities of each base. The final data obtained are the base-by-base information of the fragments.

The Roche 454 Genome Sequencer FLX is based on pyrosequencing technology. DNA fragments are sheared into shorter segments that are then ligated to specific oligonucleotide adapters for amplification by emulsion PCR on the surfaces of agarose beads. The current read length of DNA sequencing produced by the Roche 454 platform is about 700 bp and is the longest read among the three short-read NGS platforms. Therefore, the Roche 454 could be more suitable for applications requiring longer reads, such as RNA isoform identification in RNA-seq and de novo assembly of microbes in metagenomics (Mocali and Benedetti 2010).

The Applied Biosystems SOLiD 5500xl sequencer technology is based on the principle of sequencing-by-ligation. In this method, sheared DNA fragments are amplified by an emulsion PCR approach with small magnetic beads. DNA fragments on the surface of each magnetic bead are then “sequenced” by detection of the oligonucleotide ligation. One notable drawback of this platform is its tedious and time-consuming procedures for DNA library preparation prior to sequencing, currently amounting to some 5 days.

2.2.3 Data Analysis

In the workflow of an NGS project (Fig. 3.3), once the samples are sequenced using a sequencing platform, the task becomes one for bioinformaticians. Therefore, NGS data analysis (the rectangle in dash line, Fig. 3.3) is a crucial step in an NGS project.

Mapping or Assembling Short Reads to a Reference Genome

The first step in NGS data analysis is to align or to assemble the huge amount of short reads to a reference genome. Although NGS is a powerful sequencing tool, the short length of the reads generated from NGS technology limits its biological applications (Li and Homer 2010). Therefore, accurate alignment or assembly of the millions of short reads is an important determinate of experiment success. A variety of algorithms and software packages have been specifically developed for the task (Chistoserdova 2010), some of the most popular of which are listed in Table 3.5.

Table 3.5 Short-read sequence alignment tools

Name	Description
Bowtie	Uses a Burrows–Wheeler transform to create a permanent, reusable index of the genome; faster run for short sequence alignment to reference genome
BWA	Slower than bowtie but allows indels in alignment
MAQ	Performs only un-gapped alignments and allows up to three mismatches
SeqMap	Up to five mixed substitutions and insertions/deletions. Various tuning options and input/output formats
SOAP	Allows up to three gaps and mismatches. SOAP2 uses bidirectional BWT to build the index of references and increases the running speed
TopHat	Splice junction mapper for RNA-Seq reads

Bowtie (http://bowtie.cbcb.umd.edu/) is a program for aligning short read sequences to a genome that is very fast without burdensome amounts of computer memory (Langmead et al. 2009). The reference genome is first indexed using a scheme based on the Burrows–Wheeler index; the short reads are then mapped to the indexed genome, making its memory footprint small. Backtracking and double indexing are the two major algorithmic strategies Bowtie uses to rapidly align short reads to a genome. Bowtie allows mismatches and favors high-quality alignments through backtracking, while the double indexing strategy makes Bowtie avoid excessive backtracking. One of the drawbacks of this program is that gaps are not allowed in alignment. However, its new version, Bowtie 2, overcomes this and can do gapped alignment.

BWA (http://maq.sourceforge.net/) is a fast "lightweight" software package for aligning short sequencing reads against a large reference sequence such as the human genome (Li and Durbin 2009). This alignment algorithm uses a backward search strategy with the Burrows–Wheeler transform (BWT) algorithm to align sequences to the reference genome. It allows mismatches and gaps in mapping. BWA can align both single-end and paired-end reads to a reference. It uses a mapping quality index for assessing the goodness of a read aligned with a specific region in the reference genome. Therefore, it provides choices for a read mapping to the best location in the reference or to multiple regions if requested.

MAQ (http://maq.sourceforge.net/) is a freely available aligner for rapidly mapping short reads from NGS to a reference genome. It generates genotype calls according to the consensus sequence of a diploid genome by using quality scores (Li et al. 2008b). MAQ searches for the un-gapped match with lowest mismatch score. This program calculates a quality score for each alignment to measure the probability that the true alignment differs from the one the program found. It then searches for the alignment with the lowest score. It uses un-gapped mapping. Users can use MAQ for aligning reads, calling consensus sequences such as SNPs and indel variants, simulating diploid genomes, and processing alignment results in various ways.

SeqMap (http://www.stanford.edu/group/wonglab/jiangh/seqmap/) is an efficient sequence mapping program for mapping millions of short reads to a reference genome (Jiang and Wong 2008). It finds all possible locations in a whole reference genome for each of the short read sequences from NGS. SeqMap allows up to five substitutions and insertions/deletions. The FASTA file format is used for inputting data and alignment results and output is in various formats such as the SAM. SeqMap supports parallel computing and, thus, can be run on a cluster of computers. This program is very efficient, usually taking just a few hours on a desktop PC for a typical alignment job of NGS data. One distinct feature of this aligner is that it indexes and hashes the input short reads before mapping them to the reference genome.

SOAP (http://soap.genomics.org.cn/) is efficient for aligning short read sequences from NGS to a reference genome (Li et al. 2008a). It allows both gapped and un-gapped alignments and can align both single-read and paired-end reads. The seed-and-hash lookup table algorithm is used to accelerate its alignment process. Briefly, a two-bits-per-base encoding strategy is used to transform the short read sequences and the reference sequence into a numeric format. The base difference between a short read sequence and the reference sequence is then calculated using the

Table 3.6 Programs for de novo assembly

Name	Developer	Website
ABySS	Simpson JT et al.	www.bcgsc.ca/platform/bioinfo/software/abyss
ALLPATHS	Butler J et al.	ftp.broadinstitute.org/pub/crd/ALLPATHS/
Cufflinks	Trapnell C et al.	cufflinks.cbcb.umd.edu/
Edena	Hernandez D et al.	www.genomic.ch/edena.php
MIRA	Chevreux B et al.	sourceforge.net/apps/mediawiki/mira-assembler
SOAPdenovo	Li R et al.	soap.genomics.org.cn/soapdenovo.html

lookup table. SOAP is a command-driven program and easily used for batch processing with a user-specific script. It also provides a multithread option and thus supports parallel computing. It accepts FASTA format for the reference genome and both FASTA and FASTQ formats for the input short reads.

TopHat (http://tophat.cbcb.umd.edu/) is designed to align short reads from RNA-Seq to a reference genome (Trapnell et al. 2009). TopHat is designed mainly for finding exon–exon splicing junctions, although it can be used for other tasks. In the TopHat pipeline, the input short read sequences are first aligned to the reference genome using Bowtie. A short read can contiguously align to more than one exon in the genome since many exons are shorter than the length of the reads obtained from current NGS. Therefore, TopHat splits an input short read into shorter segments, and then independently aligns these segments to different exons. The aligned exons are connected together to generate the complete alignment. While mainly developed for aligning short reads from an Illumina platform such as HiSeq-2000 (~100 bp in length), it can align reads up to 1,024 bp. It can handle both single-end and paired-end reads. Users need to be aware of the parameters used in the program. The default values are optimized for alignment of short reads to the mammalian genomes. When aligning to other genomes, the parameters should be reset to suitable values.

De Novo Assembly of Short Reads

De novo assembly refers to the process of assembling short reads to an unknown genome or to a much larger reference sequence by directly “sewing” the sequences of short reads that are produced using an NGS platform. De novo assembly of NGS data is different from reference-based assembly, and it has more challenges than mapping assemblies, in terms of complexity and time consuming. Current de novo assembly algorithms have difficulty assembling large genomes, and are generally more suitable for small genomes such as found in bacteria (Chaisson et al. 2009). Table 3.6 lists some widely used de novo assembly tools. In this section, we briefly describe them.

The ABySS program (Simpson et al. 2009) is written in C++ and performs de novo sequence assembly from short paired-end reads. It has two versions, the single-processor version is designed for assembly small genomes with a size up to 100 Mbp. The parallel version is designed for assembly of large genomes with a size greater than 100 Mbp, and it is implemented using MPI (message passing interface).

ALLPATHS (Butler et al. 2008) is a de novo assembler that assembles short sequences such as the reads generated from NGS platforms into high-quality genome assemblies. One distinct feature of ALLPATHS is that its assemblies can be nonlinear. Actually, the assembling results are presented in graphs that can retain ambiguities in the input reads arising from polymorphism, sequencing errors, unresolved repeats, etc. It works better on the reads produced by Illumina platform such as HiSeq-2000 than on other platforms. ALLPATHS is not the best choice for assembling long reads such as those from the Roche 454 FLX. The program requires high sequence coverage of the genome in order to compensate for the short length of the reads. The required depth of raw reads coverage, before any error correction or filtering, depends on both the length and quality of the paired-end reads. As a rule of thumb, the required coverage for a human genome is one flow cell of data produced by the Illumina HiSeq-2000.

Cufflinks (Trapnell et al. 2010) is a program designed for multiple functions: assembling transcripts, estimating abundances of assembled transcripts, and quantifying differential expression of the transcripts based on RNA-Seq data. It assembles the short reads into a parsimonious set of transcripts and then estimates the relative abundances of these transcripts, taking into account the biases in library preparation protocols. Cufflinks can ignore fragments that map to the genome more than a specified number of times, but by default it uses all fragments in the alignment file. This program fills gaps smaller than 50 bp and joins the "transfrags" in coverage when assembling transcripts. Before testing for differential expression or regulation of genes and transcripts, Cuffdiff, another program in the package, checks goodness of the variance model for the gene or transcript. Because a positional bias correction was reducing accuracy on certain data sets in some genes, it was modified with an option to model sequence-specific bias. Its new library size normalization model is based on the geometric mean. It uses Eigen package for matrix operations that makes good use of vector registers in modern processors, speeding up the numerical routines used during abundance estimation.

Edena (Exact DE Novo Assembler) (Hernandez et al. 2008) is mainly designed for assembling short reads generated from the Illumina NGS platforms, including HiSeq-2000. It is based on the traditional overlap-layout-consensus paradigm. The program is still in the development stage, thus some features are incomplete and new functions continue to be added. Currently, both Linux (64 and 32) and Windows versions are available.

MIRA (Mimicking Intelligent Read Assembly) (Chevreux et al. 2004) was developed for the Linux operating system. This program was designed for hybrid de novo assembling short reads generated from different NGS platforms such as Roche 454 and Illumina. Therefore, it assembles reads instead of a mix of (eventually shredded) consensus sequence and reads. One restriction is the length of reads that must be less than 15 kbp.

SOAPdenovo (Li et al. 2010a) performs de novo assembly of large genomes from short reads generated from NGS platforms. This program uses the De Bruijn graph in which each node is a k-mer (the suggested k-mer is 25-mer, but it supports up to 127-mer to utilize long reads). The longer the k-mer that is used, the lower the quantity of nodes is obtained, and the more the memory is consumed. Dijkstra's algorithm is used to detect bubbles that are then merged into a single path if the sequences of the parallel paths are very similar. It produces data by its Data Preparation Module that is necessary to run the "map" and "scaff" steps using the contigs generated by SOAPdenovo or other assemblers. At beginning, the program was specially designed to assemble short reads from Illumina NGS platform such as HiSeq-2000, but now it works for short reads from other NGS platforms.

Visualizing and Annotating the Mapped or Assembled Results

Once mapping or assembly is completed, the immediate and imperative task is visualizing mapping and assembly results that are in the form of huge text or binary files. The complex and rich information requires graphical display commensurate with the task of interrogating and interpreting the data, and much effort has been expended by the scientific community toward this end. Table 3.7 lists some popular visualization and annotation tools for NGS data analysis. Below we will briefly introduce EagleView (Huang and Marth 2008) and IGV (Robinson et al. 2011).

EagleView (http://bioinformatics.bc.edu/marthlab/EagleView) is a genome assembler viewer and can be used for data integration. It can display many types of information such as base qualities and genome feature annotations. Users can easily use it to visually examine the quality of a genome assembly and to validate polymorphism candidate sites (e.g., SNPs) identified by polymorphism discovery programs. EagleView can be used to interpret assembly results and to produce hypotheses. It is written in C++ and is available for Windows, Linux, and Mac operating systems. EagleView accepts different types of data files such as the standard ACE genome assembly file and optional READS, EGL, and MAP files. It uses the optional files to visualize a genome

Table 3.7 Programs for visualizing and annotating mapping and assuming results

Name	Description and reference
EagleView	EagleView is developed at Boston College (Huang and Marth 2008). It is an information-rich genome assembler viewer that can display a dozen different types of information including base quality
IGV	IGV (Integrative Genomics Viewer) developed at MIT (Robinson et al. 2011). It is a high-performance visualization tool for interactive exploration of large, integrated genomic data sets. It supports a wide variety of data types, including next-generation sequence data, and genomic annotations
SeqTools	SeqTools was developed at Wellcome Trust Sanger Institute and can be obtained from www.broadinstitute.org/software/igv/home (Sonnhammer and Hollich 2005). It contains three tools: Blixem, Dotter, and Belvu which can be used independently and called from other tools as part of a software pipeline. In a typical application, Blixem is called to analyze a set of alignments in more detail, and Dotter is called within Blixem to give a graphical representation of a particular alignment
MapView	MapView developed at Sichuan University in China (Bao et al. 2009). It is a short-reads alignment viewer with genetic detection capability for NGS data analysis. It supports a compact alignment view for both single-end and paired-end short reads, multiple navigation and zoom modes, and multithread processing. It can handle large-scale data with high computational efficiency. Moreover, it offers automated genetic variation detection
SAM	SAM (Sequence Assembly Manager) is a WGA (whole genome assembly) management and visualization tool developed at Canada's Michael Smith Genome Sciences Centre (Warren et al. 2005). It provides a generic platform for manipulating, analyzing, and viewing WGA data, regardless of input type (www.bcgsc.ca/platform/bioinfo/software/sam)

assembly using the additional information contained in the standard files. Users can define the features to be visualized using EagleView by adding the features in the data files easily as the input files are simple text format (EagleView documentation contains the detailed format for each type file). EagleView automatically detects and opens the associated optional files from the same directory when it opens an ACE file. The key features of EagleView include (1) fast and efficient memory usage; (2) platform-specific signals pinpoint views of base quality; (3) pinpoints views of read identifier and strand distinct marks for discrepancy sites; (4) provides genome annotation; (5) navigation by read ID, contig ID, genomic features, or user defined locations, and both unpadded and padded positions; (6) data utility tools; and (7) customizable font and color.

IGV (http://www.broadinstitute.org/software/igv/home) provides concurrent visualization of multiple data types of samples, as well as a correlation of integrated data sets with clinical and phenotypic data. Sample annotations and associations with data tracks can be defined using a tab-delimited text file format in which sample identifier (used to link different types of data for the same sample), phenotype, outcome, cluster membership, and other clinical data can be the content. IGV uses the annotations that are visualized as a heatmap to group, sort, filter, and overlay diverse data types to generate a comprehensive picture of the integrated data set. Diverse genomic data types such as aligned sequence reads, identified mutations and copy numbers, RNA interference screens, gene expression, methylation, and genomic annotations can be easily integrated in IGV. Users can navigate a data set in a way similar to Google Maps: the genome can be zoomed and paned at any level of detail from whole genome to a single-base pair. One useful feature of IGV is that users can visualize their own genomic data sets alongside publicly available data because data sets can be input from a user's computer or loaded remotely from online resources such as cloud-based repositories. This makes IGV particularly effective for collaborative projects as it allows collaborators to share data remotely over the Internet. The current IGV version utilizes Java 6 or later that has to be installed on a user's computer. Several genomes are hosted in IGV genome server, including human (UCSC hg16–19 from UCSC Genome Bioinformatics, genome.ucsc.edu; Assembly b37 from 1000 Genomes, www.1000genomes.org), mouse (UCSC mm7–9 from UCSC Genome Bioinformatics), and rat (UCSC Baylor 3.4/rn4 based on version 3.4 produced by the Atlas group at Baylor Human Genome). Users can load a hosted genome into IGV by selecting the genome from the drop-down list in the tool bar. IGV provides the igvtools utility that contains several tools for preprocessing data. (1) tool *tile* can be used to convert a sorted data input file to a binary tiled data (.tdf) file and to preprocess large data sets for improved IGV performance (supported input file formats are .wig, .cn, .snp, .igv, .res, and .gct). (2) Tool *count* can be used to calculate average alignment or feature density for over a specified window size across the genome and to create a track that can be displayed in IGV, for example, as a bar chart (supported input file formats are .sam, .bam, .aligned, .psl, .pslx, and bed). (3) Tool *index* can be used to convert an ASCII alignment file or feature file to an index file that is required for loading alignment files into IGV. Index files can significantly improve performance for large feature files (supported input file formats are .sam, .aligned, .vcf, .psl, and .bed). (4) Tool *sort* that can be used to sort the input file by start position and to prepare data files for tools that required sorted input files (supported input file formats are .cn, .igv, .sam, .aligned, .psl, .bed, and .vcf).

3 NGS Applications

NGS technology is used to determine sequences of short DNA fragments in a very high-throughput manner. Theoretically, there is no limitation for application of NGS to scientific research so long as short oligonucleotide fragments are of high quality. NGS has already been widely applied to many scientific research areas, such as determination of sequences of genomes of many species (Harris et al. 2008; Chaisson and Pevzner 2008; Li et al. 2010b); identification of genetic variants by re-sequencing whole genome or targeted genes (Yeager et al. 2008; Yi et al. 2010); characterization of transcriptomes for cells, tissues, and organisms by sequencing their mRNA content (RNA-Seq) (Denoeud et al. 2008; Pan et al. 2008; Hiller et al. 2009; Guttman et al. 2010; Trapnell et al. 2010); identification of DNA-binding proteins and epigenetic markers by ChIP-Seq (Robertson et al. 2007; Kharchenko et al. 2008; Kaufmann et al. 2010); and pinpointing of the genomic content in a complex sample by metagenomics (Turnbaugh et al. 2010; Fierer et al. 2010). Many comprehensive review articles on NGS technologies and their applications have been published (Wang et al. 2009; Kato 2009; Voelkerding et al. 2009). This section does not intend to cover a comprehensive of applications of NGS technologies, and rather focuses on NGS technologies with the most potential applications in pharmacogenomics.

3.1 RNA Sequencing

Gene expression profiles of patients can be used to interrogate their responses to drugs for both efficacy and safety considerations. Identification of a gene or a panel of genes that are associated with a patient's response to a specific drug is an important topic in pharmacogenomics. NGS is thought to be more accurate and precise in measuring expression levels of genes or transcripts, compared to other technologies such as PCR and DNA microarray. Therefore, it is expected that NGS will accelerate pharmacogenomics research and translation of the findings in pharmacogenomics into personalized medicine.

Directly sequencing RNAs/cDNAs offers an alternative approach for high-throughput transcriptome analysis (Marioni et al. 2008; Wilhelm et al. 2010). RNA sequencing (simplified as RNA-Seq) is revolutionary in its abilities to precisely measure expression of genes at the whole genome level (Li et al. 2010c). Its higher resolution output makes NGS a promising tool for discovery of novel transcripts, exploration of allele-specific expression, and identification of alternative splice variants and posttranscriptional mutations and isoforms compared to the conventional Sanger sequencing and microarray-based approaches (Sultan et al. 2008; Chepelev et al. 2009; Perkins et al. 2009; Tang et al. 2009; Hittinger et al. 2010). RNA-Seq has been used to characterize RNA populations and provided more complicated pictures of RNA regulation and expression through alternative splicing, alternative polyadenylation, and RNA editing (Nagalakshmi et al. 2008; Guttman et al. 2009; Li et al. 2009a). NGS has expanded our understanding of the extent and complexity of gene expression and the mechanisms of RNA expression regulation in both eukaryotic and prokaryotic genomes (Jacquier 2009; Sorek and Cossart 2010; Licatalosi and Darnell 2010).

The most active field of applying RNA-seq in pharmacogenomics is the discovery of cancer-related genomic biomarkers. Since the Human Genome Project finished sequencing the human genome and published the first human genome map draft (a landmark in genomics) (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), the understanding and discovery of cancer genomic biomarkers became feasible. The recent advances in RNA-seq technology have made it possible to more precisely profile genes at the whole genome level and discover new RNA molecules such as new gene transcripts, small RNAs, alternative splicing products, and gene fusions products (Lipson et al. 2012). Sinicrop et al. developed and optimized the RNA-Seq library chemistry as well as bioinformatics methods for whole transcriptome profiling (Sinicropi et al. 2012). This work led to not only the re-discovery of RNA biomarkers for disease recurrence risk that were previously identified by RT-PCR analysis of a cohort of 136 patients but also the identification of a new group of recurrence risk biomarkers that were not previously discovered using DNA microarrays in a separate cohort of patients (Sinicropi et al. 2012).

3.2 DNA Sequencing

Identifying genetic variants (such as SNPs and mutations) is possible after successfully aligning (or assembling) NGS short reads to a genome. However, distinguishing the causal variants for a phenotype from the large number of apparently novel genetic variants present by chance in any human genome can be difficult. Identification of rare mutations introduced into the population that differ from low-frequency alleles descendent from ancient ancestors is still more difficult when analyzing DNA sequencing data. One of the key factors for the success of application of DNA sequencing in pharmacogenomics is the reliability of software packages for detecting genetic variants such as SNPs from the NGS data.

Many software packages for detecting SNPs from DNA sequencing data have been developed. For example, SOAPsnp (Li et al. 2009b) is a member of the Short Oligonucleotide Analysis Package (http://soap.genomics.org.cn/soapsnp.html). SOAPsnp is a re-sequencing utility that can be used to assemble a consensus sequence for a genome after the raw reads are aligned to the reference genome. It identifies SNPs and determines their genotypes by comparing the aligned reads with the reference genome using the consensus sequence. SOAPsnp is a command line driven program written in C/C++ that generally runs under the 64-bit Linux system. However, this program has been tested on other operation systems. It only needs ~0.5 GB or even less memory to run but its output might be very large because it consumes a lot of hard disk space when memory is small. In its text output mode, the output file may be 60 times the genome size (e.g., 180G free space is required to run a human genome). The program was evaluated using the Asian genome re-sequencing project data, based on the Illumina HapMap 1 M BeadChip Duo genotyping sites. Over 99 % of the known SNPs were identified and 99.9 % of the genotypes for the covered SNPs are consistent, indicating its reliability for SNP calling.

DNA sequencing has been actively applied in pharmacogenomics and many interesting findings have been obtained. Some representative studies and their major finding are revisited below. Pleasance et al. sequenced the genomes of a malignant melanoma and a lymphoblastoid cell line obtained from the same patient to provide somatic mutation information from cancer cells (Pleasance et al. 2010). Sequence data analysis detected 33,345 somatic base substitutions, 680 small deletions, 303 small insertions, 51 somatic rearrangements, and all well-validated somatic copy number alterations and regions of loss of heterozygosity. Lee et al. sequenced a genome from a primary lung tumor and a normal one from adjacent normal tissue of the same patient (Lee et al. 2010). Comparative analysis identified more than 50,000 high-confidence single nucleotide variants (SNV) as somatic variants. Shah et al. sequenced the genome and transcriptome of an estrogen-receptor-positive metastatic lobular breast cancer (Shah et al. 2009). DNA sequence data analysis led to the identification of 32 somatic non-synonymous coding mutations. Tran et al. sequenced the blood, tumor biopsy, and archived tumor samples of 50 patients collected from four different cancer centers by using targeted exon sequencing (Tran et al. 2013). Combining with multiplex somatic mutation genotyping using Sequenom MassARRAY and Sanger sequencing, they assessed the feasibility of incorporating real-time analysis of somatic mutations within exons of 19 genes into patient management. Moreover, the identified actionable mutations were similar between archival and biopsy samples, implying that the cancer mutations do not change much across clinical stages making them good predictors of drug response. Their results demonstrated that the use of next-generation sequencing for real-time genomic profiling in advanced cancer patients is feasible.

Some well-known mutations in BRCA1 and BRCA2 (BRCA1/2) genes can markedly elevate risk of developing breast and ovarian cancers. However, mutations in unscreened regions of BRCA1/2 and other predisposition genes need identification. Ozcelik et al. recently applied long-range PCR and NGS for BRCA1/2 mutation analysis (Ozcelik et al. 2012). By sequencing the genomic DNA from 12 cancer patients, all 19 distinct (51 total) BRCA1 and 35 distinct (63 total) BRCA2 sequence alterations detected by the Sanger sequencing were confirmed, with no false-negatives. Moreover, variants from introns and untranslated regions were identified, demonstrating that NGS can provide accurate and more comprehensive genetic information.

These demonstrable successes portend that advances in DNA sequencing will further enable molecular diagnostics, accelerate pharmacogenomics research, and expand translation of genetic biomarkers into personalized medicine.

3.3 miRNA Sequencing

miRNAs have been found to be involved in many diseases (Jiang et al. 2009), including diabetes (Hennessy and O'Driscoll 2008), cardiomyopathies (van Rooij et al. 2006), psychiatric disorders such as schizophrenia (Barbato et al. 2008; Beveridge et al. 2009), and cancer (Medina and Slack 2008). High-throughput sequencing has been applied for profiling known and novel small RNAs such as miRNAs that regulate gene expression and play a major role in most biological processes. They could be used as biomarkers and as therapeutic agent targets (Fasanaro et al. 2010; Dreyer 2010; Nana-Sinkam and Croce 2011) for drug development. Consequently, identifying miRNAs and measuring their expression using NGS has been a recent pharmacogenomics research field.

Theoretically, the expression of a therapeutic target protein involved in a disease can be blocked by inhibitors that repress the protein itself or by particular miRNAs that repress the expression of the protein. Thus, administration of miRNAs mimetics can simulate the endogenous miRNAs population repressing a detrimental gene and its protein products. Therefore, miRNA inhibitors and some miRNA mimics can be used as therapeutic candidates for various diseases, including cardiovascular disease, neurological disorders, and viral infection (Nana-Sinkam and Croce 2011). Drugs that target miRNAs are in the discovery and development pipelines of many pharmaceutical companies. Table 3.8 lists some miRNAs that are used as the targets of a number of new drug products in preclinical studies and in clinical trials.

Several miRNA biomarkers for diagnosis and prognosis of disease have recently been identified using NGS technologies. One example is the differential expression of the oncogenic miRNAs of the *miR17-92* cluster and the *miR-181* family among different types of neuroblastoma patients, which was measured using the SOLid NGS platform. The expression levels of those miRNAs were higher in the five unfavorable neuroblastoma patients. In contrast, the expression levels of the tumor suppressive miRNAs of *miR-542-5p* and

Table 3.8 Examples of miRNAs in drug development

Company	Generic name	Target	Indication	Status
Santaris Pharma	Miravirsen	miR-122	Hepatitis C virus	Phase IIA
Regulus Therapeutics	Unspecified	miR-21	Fibrosis	Clinical
Regulus Therapeutics	Unspecified	miR-21	Cancer	Clinical
Regulus Therapeutics	Unspecified	mi-R122	Hepatitis C virus	Clinical
Regulus Therapeutics	Unspecified	mi-155	Inflammation	Preclinical
Regulus Therapeutics	Unspecified	miR-33a	Metabolic diseases	Preclinical
miRagen Therapeutics	MGN-9103	miR-208/499	Chronic Heart Failure	Preclinical
miRagen Therapeutics	MGN-1374	miR-15/195	Post-MI Remodeling	Preclinical
miRagen Therapeutics	MGN-4893	miR-451	Polycythemia vera	Preclinical
miRagen Therapeutics	MGN-4420	miR-29	Cardiac fibrosis	Lead optimization
Mirna Therapeutics	Unspecified	Let-7	Lung cancer	Preclinical
Mirna Therapeutics	Unspecified	miR-34	Prostate cancer	Preclinical
Rosetta Genomics	TCDD	miR-191	Hepatocellular carcinoma	Preclinical
Rosetta Genomics	Unspecified	miR-34a	Liver cancer	Preclinical

miR-628 measured using NGS were much higher in the five favorable neuroblastoma patients compared to the five unfavorable neuroblastoma patients whose expressions of these two miRNAs were virtually absent (Schulte et al. 2010).

4 Future Perspectives

GWAS are based on the "common disease–common variants hypothesis" and had a great wave of research activities 6 years ago. The huge prospect for GWAS to advance pharmacogenomics and personalized medicine has so far largely proved disappointing. Fortunately, next-generation sequencing is fundamentally changing the way in which genomic information of individuals at the base level is being obtained for a holistic understanding of the human genome. Sequencing target genes of diseases or whole genomes of patients is expected to reveal additional instances of lower-frequency, higher-penetrance alleles and is renewing prospects for more effective personalized therapy. It is also expected that more and more genomic biomarkers will be identified and used in the development of drug products in the future. The challenges in the future are not only the discovery of genomic biomarkers, but also how to incorporate them into drug development, regulatory decision making, and personalized medicine. To enhance FDA's scientific base for regulatory decision making on pharmacogenomics findings from NGS for personalized medicine, FDA initiated SEQC (sequence quality control) project, also known as MAQC-III (http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/default.htm#MAQC-IIIalsoknownasSEQC) to address various issues on the reliability of NGS technologies for pharmacogenomics.

The fields of pharmacogenomics and personalized medicine are undergoing a revolution today because of the advances in NGS technologies. Both common genetic variants that have been explored by GWAS and rare genetic variants that have not been explored can be identified using NGS. The authors believe this approach will drive the next wave of pharmacogenomics and personalized medicine advances. We expect NGS to make significant contributions to our understanding of pharmacogenomics and will redefine the field of personalized medicine. As the pharmacogenomics knowledge base expands, translation of pharmacogenomics findings into personalized medicine will be within reach.

References

Ansorge WJ (2009) Next-generation DNA sequencing techniques. Nat Biotechnol 25:195–203

Armstrong B, Stewart M, Mazumder A (2000) Suspension arrays for high throughput, multiplexed single nucleotide polymorphism genotyping. Cytometry 40:102–108

Bao H, Guo H, Wang J, Zhou R, Lu X, Shi S (2009) MapView: visualization of short reads alignment on a desktop computer. Bioinformatics 25:1554–1555

Barany F (1991) Genetic disease detection and DNA amplification using cloned thermostable ligase. Proc Natl Acad Sci USA 88:189–193

Barbato C, Giorge C, Catalanotto C, Cogoni C (2008) Thinking about RNA? MicroRNAs in the brain. Mamm Genome 19:541–551

Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A (2009) An operational definition of epigenetics. Genes Dev 23:781–783

Beveridge NJ, Gardiner E, Carroll AP et al (2009) Schizophrenia is associated with an increase in cortical microRNA biogenesis. Mol Psychiatry 15:1176–1189

Bird A (2007) Perceptions of epigenetics. Nature 447: 396–398

Brunner AL, Johnson DS, Kim SW et al (2009) Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res 19:1044–1056

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18:810–820

Buyse M, Loi S, Van't Veer L et al (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 98:1183–1192

Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18:324–330

Chaisson MJ, Brinza D, Pevzner PA (2009) De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res 19:336–346

Chepelev I, Wei G, Tang Q, Zhao K (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res 37:e106

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG et al (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14:1147–1159

Chistoserdova L (2010) Recent progress and new challenges in metagenomics for biotechnology. Biotechnol Lett 32:1351–1359

Cloonan N, Forrest AR, Kolle G et al (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5:613–619

Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG (1996) Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 7:244–255

Daly AK, Donaldson PT, Bhatnagar P et al (2009) HLA-B*5701 genotype is a major determinant of drug induced liver injury due to flucloxacillin. Nat Genet 41:816–819

Das SK, Austin MD, Akana MC, Deshpande P, Cao H, Xiao M (2010) Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. Nucleic Acids Res 38:e177

Denoeud F, Aury JM, Da Silva C et al (2008) Annotating genomes with massive-scale RNA sequencing. Genome Biol 9:R175

Dreyer JL (2010) New insights into the roles of microRNAs in drug addiction and neuroplasticity. Genome Med 2:92

Duffy MJ, Napieralski R, Martens JW et al (2009) Methylated genes as new cancer biomarkers. Eur J Cancer 45:335–346

Fasanaro P, Greco S, Ivan M, Capogrossi MC, Martelli F (2010) MicroRNA: emerging therapeutic targets in acute ischemic diseases. Pharmacol Ther 125:92–104

Fierer N, Lauber C, Zhou N et al (2010) Forensic identification using skin bacterial communities. Proc Natl Acad Sci USA 107:6477–6481

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251

Glas AM, Floore A, Delahaye LJ et al (2006) Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics 7:278

Goldstein JA, Blaisdell J (1996) Genetic tests which identify the principal defects in CYP2C19 responsible for the polymorphism in mephenytoin metabolism. Methods Enzymol 272:210–218

Goldstein LJ, Gray R, Badve S et al (2008) Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features. J Clin Oncol 26:4063–4071

Guttman M, Amit I, Garber M et al (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–227

Guttman M, Garber M, Levin JZ et al (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28:503–510

Harris L, Fritsche H, Mennel R et al (2007) American Society of Clinical Oncology 2007 update of recommendations for the use of tumour markers in breast cancer. J Clin Oncol 25:5287–5312

Harris TD, Buzby PR, Babcock H et al (2008) Single-molecule DNA sequencing of a viral genome. Science 320:106–109

He L, Thomson JM, Hemann MT et al (2005) A microRNA polycistron as a potential human oncogene. Nature 435:828–833

Hennessy E, O'Driscoll L (2008) Molecular medicine of microRNAs: structure, function, and implications for diabetes. Expert Rev Mol Med 10:e24

Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop. Genome Res 18:802–809

Higuchi R, Dollinger G, Walsh PS, Griffith R (1992) Simultaneous amplification and detection of specific DNA sequences. Biotechnology (N Y) 10:413–417

Hiller D, Jiang H, Xu W, Wong WH (2009) Identifiability of isoform deconvolution from junction arrays and RNA-Seq. Bioinformatics 25:3056–3059

Hittinger CT, Johnston M, Tossberg JT, Rokas A (2010) Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc Natl Acad Sci USA 107:1476–1481

Hodges E, Xuan Z, Balija V et al (2007) Genome-wide in situ exon capture for selective re-sequencing. Nat Genet 39:1522–1527

Holland PM, Abramson RD, Watson R, Gelfand DH (1991) Detection of specific polymerase chain reaction product by utilizing the 5′–3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc Natl Acad Sci USA 88:7276–7280

Hong H, Goodsaid F, Shi L, Tong W (2010) Molecular biomarkers: a US FDA effort. Biomarkers Med 4:215–225

Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res 19:1270–1278

Huang W, Marth G (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res 18:1538–1543

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Irizarry RA, Ladd-Acosta C, Wen B et al (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 41:178–186

Jacquier A (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet 10:833–844

Ji W, Foo JN, O’Roak BJ et al (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40:592–599

Jiang H, Wong WH (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24:2395–2396

Jiang Q, Wang Y, Hao Y et al (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 37:D98–D104

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. Science 316:1497–1502

Jorgensen JT (2009) New era of personalized medicine: a 10-year anniversary. Oncologist 14:557–558

Kahn SL, Ronnett BM, Gravitt PE, Gustafson KS (2008) Quantitative methylation-specific PCR for the detection of aberrant DNA methylation in liquid-based Pap tests. Cancer 114:57–64

Kato K (2009) Impact of the next generation DNA sequences. Int J Clin Exp Med 2:193–202

Kaufmann K, Muino JM, Osteras M et al (2010) Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nat Protoc 5:457–472

Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26:1351–1359

Kim MJ, Huang SM, Meyer UA, Rahman A, Lesko LJ (2009) A regulatory science perspective on warfarin therapy: a pharmacogenetic opportunity. J Clin Pharmacol 49:138–146

Kindmark A, Jawaid A, Harbron CG et al (2008) Genome-wide pharmacogenetic investigation of a hepatic adverse event without clinical signs of immunopathology suggests an underlying immune pathogenesis. Pharmacogenomics J 8:186–195

Kondo Y, Shen L, Cheng AS et al (2008) Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. Nat Genet 40:741–750

Kranenburg O (2005) The KRAS oncogene: past, present, and future. Biochim Biophys Acta 1756:81–82

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Langreth R, Waldholz M (1999) New era of personalized medicine: targeting drugs for each unique genetic profile. Oncologist 4:426–427

Lavedan C, Licamele L, Volpi S et al (2008) Association of the *NPAS3* gene and five other loci with response to the antipsychotic iloperidone identified in a whole genome association study. Mol Psychiatry 14:804–819

Lee W, Jiang Z, Liu J et al (2010) The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 465:473–477

Li R, Li Y, Kristiansen K, Wang J (2008a) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714

Li H, Ruan J, Durbin R (2008b) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Li JB, Levanon EY, Yoon JK et al (2009a) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science 324:1210–1213

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J (2009b) SNP detection for massively parallel whole-genome re-sequencing. Genome Res 19:1124–1132

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–483

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z et al (2010a) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272

Li R, Fan W, Tian G et al (2010b) The sequence and de novo assembly of the giant panda genome. Nature 463:311–317

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN (2010c) RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26:493–500

Licatalosi DD, Darnell RB (2010) RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 11:75–87

Lipson D, Capelletti M, Yelensky R et al (2012) Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. Nat Med 18:382–384

Lockhart DJ, Dong H, Byrne MC et al (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680

Lu J, Getz G, Miska EA et al (2005) MicroRNA expression profiles classify human cancers. Nature 435:834–838

Lucas A, Nolan D, Mallal S (2007) HLA-B*5701 screening for susceptibility to abacavir hypersensitivity. J Antimicrob Chemother 59:591–593

Mallal S, Phillips E, Carosi G et al (2008) HLA-B*5701 screening for susceptibility to abacavir hypersensitivity. N Engl J Med 358:568–569

Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517

McGrath JP, Capon DJ, Smith DH et al (1983) Structure and organization of the human Ki-ras proto-oncogene and a related processed pseudogene. Nature 304:501–506

Medina PP, Slack FJ (2008) MicroRNAs and cancer: an overview. Cell Cycle 7:2485–2492

Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11:31–46

Mocali S, Benedetti A (2010) Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. Res Microbiol 161:497–505

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628

Motulsky AG (1957) Drug reactions enzymes and biochemical genetics. J Am Med Assoc 165:835–837

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349

Nana-Sinkam SP, Croce CM (2011) MicroRNAs as therapeutic targets in cancer. Transl Res 157:216–225

Newton CR, Graham A, Heptinstall LE et al (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res 17:2503–2516

Ozcelik H, Shi X, Mc C et al (2012) Long-range PCR and next-generation sequencing of Brca1 and Brca2 in breast cancer. J Mol Diagn 14:467–475

Paik S, Shak S, Tang G et al (2004) A multi-gene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351:2817–2826

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415

Park PJ (2009) ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680

Perkins TT, Kingsley RA, Fookes MC et al (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. PLoS Genet 5:e1000569

Pleasance ED, Cheetham RK, Stephens PJ et al (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463:191–196

President's Council of Advisors on Science Technology (2008) Priorities for personalised medicine. http://www.whitehouse.gov/files/documents/ostp/PCAST/pcast_report_v2.pdf. Accessed 28 Aug 2012

Reis-Filho JS (2009) Next-generation sequencing. Breast Cancer Res 11(Suppl 3):S12

Robertson G, Hirst M, Bainbridge M et al (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26

Sanderson S, Emery J, Higgins J (2005) CYP2C9 gene variants, drug dose, and bleeding risk in warfarin-treated patients: a HuGEnet™ systemic review and meta-analysis. Genet Med 7:97–104

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Schmidt D, Wilson MD, Ballester B et al (2010) Five-vertebrate ChIP–seq reveals the evolutionary dynamics of transcription factor binding. Science 328:1036–1040

Schulte JH, Marschall T, Martin M et al (2010) Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. Nucleic Acids Res 38:5919–5928

Shah SP, Morin RD, Khattra J et al (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature 461:809–813

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123

Sinicropi D, Qu K, Collin F et al (2012) Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin- embedded tumor tissue. PLoS One 7:e40092

Sonnhammer EL, Hollich V (2005) Scoredist: a simple and robust protein sequence distance estimator. BMC Bioinformatics 6:108

Sorek R, Cossart P (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat Rev Genet 11:9–16

Sultan M, Schulz MH, Richard H et al (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 32:956–960

Syvanen AC, Aalto-Setala K, Harju L, Kontula K, Soderlund H (1990) A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. Genomics 8:684–692

Takahashi H, Wilkinson GR, Nutescu EA et al (2006) Different contributions of polymorphisms in VKORC1 and CYP2C9 to intra- and inter-population differences in maintenance doses of warfarin in Japanese, Caucasians and African Americans. Pharmacogenet Genomics 16:101–110

Takeuchi F, McGinnis R, Bourgeois S et al (2009) A genome-wide association study confirms *VKORC1*, *CYP2C9*, and *CYP4F2* as principal genetic determinants of warfarin dose. PLoS Genet 5, E1000433

Tang F, Barbacioru C, Wang Y et al (2009) MRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–862

Tran B, Brown AM, Bedard PL et al (2013) Feasibility of real time next generation sequencing of cancer genes linked to drug response: results from a clinical trial. Int J Cancer 132:1547–1555. doi:10.1002/ ijc.27817

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111

Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515

Turnbaugh PJ, Quince C, Faith JJ et al (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proc Natl Acad Sci USA 107:7503–7508

Turner ST, Bailey KR, Fridley BL et al (2008) Genomic association analysis suggests chromosome 12 locus influencing antihypertensive response to thiazide diuretic. Hypertension 52:359–365

US Food and Drug Administration (2012) Table of valid genomic biomarkers in the context of approved drug labels. http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm. Accessed 28 Aug 2012

Van't Veer LJ, Dai H, van de Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536

van de Vijver MJ, He YD, Van't Veer LJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009

van Rooij E, Sutherland LB, Liu N et al (2006) A signature pattern of stress-responsive microRNAs that can evoke cardiac hypertrophy and heart failure. Proc Natl Acad Sci USA 103:18255–18260

Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291: 1304–1351

Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55:641–658

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008a) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476

Wang J, Wang W, Li R et al (2008b) The diploid genome sequence of an Asian individual. Nature 456:60–65

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Warren RL, Butterfield YS, Morin RD, Siddiqui AS, Marra MA, Jones SJM (2005) Management and visualization of whole genome shotgun assemblies using SAM. Biotechniques 38:715–720

Weisenberger DJ, Siegmund KD, Campan M et al (2006) CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet 38:787–793

Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Wilhelm BT, Marguerat S, Goodhead I, Bahler J (2010) Defining transcribed regions using RNA-seq. Nat Protoc 5:255–266

Yeager M, Xiao N, Hayes RB et al (2008) Comprehensive re-sequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet 124:161–170

Yi X, Liang Y, Huerta-Sanchez E et al (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75–78