Journal of Cancer Therapy, **2012, 3, 797-809**
http://dx.doi.org/10.4236/jct.2012.325101 Published Online October 2012 (http://www.SciRP.org/journal/jct)

Novel Methods in the Study of the Breast Cancer Genome: Towards a Better Understanding of the Disease of Breast Cancer

Jian Li[1,2*], Xue Lin[1*], Nils Brünner[3], Huanming Yang[2#], Lars Bolund[1,2#]

[1]Department of Biomedicine, University of Aarhus, Aarhus, Denmark; [2]BGI, Shenzhen, China; [3]Department of Veterinary Pathobiology, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark.
Email: [#]yanghm@genomics.org.cn, [#]bolund@hum-gen.au.dk

Received August 15th, 2012; revised September 17th, 2012; accepted September 29th, 2012

ABSTRACT

Rapidly developing sequencing technologies and bioinformatic approaches have provided us with an unprecedented instrument allowing for an unbiased and exhaustive characterization of the cancer genome in genetic, epigenetic and transcriptomic dimensions. This review introduces recent exciting findings and new methodologies in genomic breast cancer research. With this development, cancer genome research will illuminate new delicate interactions between molecular networks and thereby unravel the underlying biological mechanisms for cancer initiation and progression. It also holds promise for providing a molecular clock for the estimation of the temporal processes of tumorigenesis. These methods in combination with single cell sequencing will make it possible to construct a family tree elucidating the evolutionary lineage relationships between cell populations at single-cell resolution. The anticipated rapid progress in genomic breast cancer research should lead to an enhanced understanding of breast cancer biology and guide us towards novel ways to ultimately prevent and cure breast cancer.

Keywords: Breast Cancer Genome; Massively Parallel Sequencing; Pathway-Oriented Analysis; Mitochondrial Genome; Temporal Order of Aberrations; Single Cell Sequencing; Microbiome

1. Introduction

Breast cancer is the second most commonly diagnosed cancer and seriously threatens women health [1]. As a complex disease, both genetics and environmental causes are implicated in the tumorigenesis of breast cancer. The catalogue of inherited or somatic mutations accumulated in a cancer genome encompasses substitutions of nucleotides, insertions and deletions, translocations and other chromosomal rearrangements as well as copy number changes [2]. Many efforts have been spent in the last decade to identify the spectrum of genes associated with breast cancer [3]. Genes, such as *BRCA*1 and *BRCA*2, with high penetrance mutations are involved in approximately 70% of breast cancers in high-risk families. However, they only account for a minority of all breast cancer cases [4]. In general, <10% of breast cancer cases are thought to be hereditary in a Mendelian fashion and usually a somatic "second hit" in the homologous normal allele is required for disease development.

Thus, to identify low penetrance susceptibility gene variants (inherited or somatically acquired) has become an area of interest in breast cancer research. Genome-wide association studies (GWAS) are commonly used for the search for correlations between disease incidence and genetics. GWAS routinely encompasses tens of thousands of patient samples and scans the full length of the genomes [5]. GWAS have identified 25 genetic loci associated with breast cancer risk [5]. Still, to date, GWAS can only account for 9% - 10% of breast cancers [5]. Even when considering all types of genetic studies, some 70% of breast cancer cases remain unexplained [5,6]. It has become obvious that genetic factors only account for part of the phenotypic variance [7]. Breast cancer development represents a multiple-step process and the risk increases with age. Environmental degenerative factors no doubt play an important role in breast cancer tumorigenesis. Epigenetic changes, including somatically acquired (and sometimes germ line transmitted) chemical modifications of DNA (without DNA sequence changes) as well as DNA binding small RNAs and proteins (e.g. histones), bridge the gap between genetics and the environment significantly improving our understanding of the

[*]These two authors (Jian Li and Xue Lin) equally contributed to this work.
[#]Corresponding authors.

disease of breast cancer [8,9].

The emergence of massively parallel sequencing technology provides researchers with an unprecedented powerful tool for breast cancer research. Currently, there are five commonly applied massively parallel sequencing technologies: 454 Life Sciences (Roche) applies a pyrosequencing approach [10], Illumina/Solexa uses the principle of sequencing by synthesis (SBS) with reversible dye terminators [11], Applied Biosystems SOLiD [12] and Complete Genomics [13] perform sequencing by ligation strategies, and Ion Torrent [14] utilizes an ion-sensitive SBS principle for sequencing. Although these sequencing platforms are technically quite diverse, they share many common features: Similar process of library preparation, amplification of libraries prior to sequencing, and similar process of sequencing by an automated series of enzyme-driven biochemical and fluorescent imaging based data acquisition steps [15]. This allows ultra-deep investigations of breast cancer genomes and their epigenetic modifications in a fast and cost-effective way, without the requirement of abundant amounts of material [2,16]. Here we briefly describe current genomic approaches applied in breast cancer research.

Although array-based approaches remain broadly applied for RNA analyses at present, transcriptome sequencing is becoming increasingly important, as sequencing has a greater dynamic range and provides the possibility to discover new transcripts, sequence variants and splicing events [17,18]. RNA sequencing allows deep mapping of short RNA fragments (17 - 22 nucleotides), thus exponentially increasing our knowledge of the biology, diversity and abundance of small RNA populations [19,20].

Despite the fact that a number of whole breast cancer genomes have already been sequenced [21-23], the analyses of particularly informative sectors of the cancer genome, e.g. sequencing the DNA sequence based on capturing the exomes and DNA sequences coding for known micro-RNAs, are likely to be carried out commonly [24]. Exome sequencing applies affinity-enrichment techniques to enrich exome sequences from the genome before sequencing, thereby allowing a deep characterization of the target sequence for a decreased cost [25]. Massively parallel sequencing also can efficiently sequence small genome fragments that have been randomly collected from the tumor genome to reveal copy number changes (low coverage sequencing) [16]. The relative number of sequenced short DNA fragments in equal-sized bins distributed along the genome, can be regarded as an estimate of the relative copy number at different genomic locations [16].

Massively parallel sequencing has also dramatically increased our ability to survey genome-wide epigenetic markers. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) uses antibodies to pull down target DNA to globallysurvey the DNA binding pattern of a protein of interest [26]. This method is also applied for measurement of histone modifications. DNA methylation as an important epigenetic mechanism has been extensively studied. To date, there are three main approaches that are compatible with massively parallel sequencing for genome-wide mapping of DNA methylation information: 1) endonuclease digestion-based methods such as modified methylation specific digital karyotyping (MMSDK) [27]; 2) affinity enrichment-based methods such as methylated DNA binding domain sequencing (MBD-Seq) [28] and methylated DNA immunoprecipitation sequencing (MeDIP-Seq) [29]; and 3) bisulfite conversion-based methods such as MethylC-Seq (methylome) [30,31] and reduced representation bisulfite sequencing (RRBS) [32]. Exhaustive comparisons of these DNA methylation assays have been recently carried out by several groups, and these studies are invaluable when selecting methods for DNA methylation analysis [33-36]. With the cost of sequencing the whole human genome dropping towards 1000 US dollars (http://www.genome.gov/12513210) [37] in the near future, a revolutionary era of personalized medical care for breast cancer patients will soon become a reality. For example, the elucidation of a number of intrinsic breast cancer subtypes [38] has added significantly to our understanding of breast cancer heterogeneity and also provides tools that can be used to select the right treatment for the right patient at the right time. The important advances in cancer genome analysis brought about by the application of massively parallel sequencing have already been discussed in detail in many other reviews [2,16,39,40]. In the present review, we will introduce and highlight some new research directions, which we expect will lead to an increased understanding of the breast cancer disease.

2. Breast Cancer Genome Research

2.1. Pathway-Oriented Analysis Based on Integration of Multiple "Omic" Dimensions

One important insight obtained from the large-scale mutational analyses carried out in pioneering large-scale sequencing studies of breast and colon cancer was the importance of taking a pathway-oriented strategy [41-43]. A pathway-oriented model of tumorigenesis is also supported by the observation that although different genes may be mutated in the same type of tumors, these genes often belong to a more limited number of pathways and biological processes [41]. For example, breast and colo-

rectal cancers both have frequent mutations in PIK3CA pathway genes, but these mutations are not always in the same genes [41]. The cancer genome can be dysregulated through multiple mechanisms including mutations in coding and non-coding sequences, alterations in DNA copy number and organization, and aberrations in modifications of DNA and DNA related proteins [16]. The abnormalities may simultaneously occur in a key gene in an independent or synergistic manner, leading to dysfunction of this gene, thereby fueling tumorigenesis. Alternatively, these abnormalities can target different genes that are connected within a pathway and, thereby, through dysfunction of the pathway, ultimately facilitate cancer development. A classic example of this is the tumor suppressor gene, *TP*53, which can be inactivated in three ways: through homozygous deletions in the 17p13.1 region; through hypermethylation of *TP*53 promoter to epigenetically silence the expression; or through mutations that cripple the function of *TP*53. More interestingly, these multiple mechanisms can collaborate to cause dysfunction of this gene; for instance, one allele may be inactivated by mutation whilst the other allele may subsequently be silenced by DNA methylation of its promoter region. Alternatively, one mutated allele in combination with a subsequent copy loss of the second allele or epigenetic silencing of the other allele can eventually completely inactivate the gene function. Allele-specific gene expression regulated by epigenetic mechanisms was previously regarded as mainly constrained to genomic imprinting. A recent study of the DNA methylome of human peripheral blood mononuclear cells demonstrated that the regulation mechanisms of allelic-specific gene expression by allelic-specific DNA methylation may exists a more comprehensive biological phenomenon [31], which underscores the relevance of an integrative analysis involving multiple dimensions of biological information. This can be even more important in cancer genome research, because genomic abnormalities observed in cancers can be associated with abroad range of biological characteristics. Thus, breast cancer is undoubtedly a complex disease, both in its biological mechanisms and in its final biological endpoints.

A deeper understanding of breast cancer therefore requires broad investigations of the breast cancer genome in different dimensions followed by integrative analysis of the findings using a pathway-oriented strategy. Additionally, pathway analyses can facilitate the selection of genes for further functional analyses [44]. Recent breast cancer genome studies have focused their efforts on integrative analysis for large-scale sequencing data. As examples of this, many studies have involved combinational analysis of sequencing data from the genome, exome and/or transcriptome to evaluate the impact of mutations or genome rearrangements on gene expression [45,46]. Integrating DNA copy number, RNA transcriptome, and CpG island methylation profiles, Sun *et al.* systematically examined the genomic features underlying the estrogen receptor positive (ER+) and estrogen receptor negative (ER-) breast cancer phenotypes [47]. These studies demonstrate the transition of the strategy of breast cancer research from focusing on a handful gene sets to multi-dimensional investigations including the whole genome using pathway-oriented models.

2.2. Mitochondrial Genome Analysis

The human mitochondrial genome is a 16.6 kb double-stranded circular DNA molecule presenting a copy number that varies widely according to the cell type [48]. Because of a lack of histone protection, limited repair capacity and proximity to superoxide radicals, mitochondrial DNA (mtDNA) has a higher susceptibility to damage, compared with the nuclear genome [49]. Additionally, the absence of introns in the mitochondrial genome leads to more frequent coding sequence mutations, which affect mitochondrial function. Numerous somatic mutations of mtDNA have been observed in breast cancer [50,51]. The dysfunction of the mitochondria has long been suspected to contribute to the development and progression of cancer [52]. At present, a primary goal is to assess the functional role of the various mitochondrial mutations in the initiation and progression of breast cancer, with a specific focus on identifying mutations associated with acquired adaption for rapid proliferation under hypoxic conditions, as well as mutations related to drug metabolism. Thus, mtDNA mutations may have potential value as cancer biomarkers, for example to predict the metabolism of different chemotherapeutic drugs, *i.e*. to predict sensitivity/resistance to treatment. Moreover, the majority of mtDNA mutations have been observed to be homoplasmic in early preneoplastic and cancerous lesions, *i.e*. the mutated mtDNA predominates and is readily detectable in tumor biopsy material with amounts reported to be 19 - 22 times more abundant than, for example, mutated *TP*53 DNA [53]. The success in identifying mtDNA mutations in material obtained from fine-needle aspirates underscores the potential possibility of using this methodology in clinical practice [51,54]. Distinguishing the spectrum of mutations related to cancer from age related mutations is necessary, since mitochondrial mutations have also been reported to occur as a function of the aging process [55].

The heterogeneity of the mitochondrial genome must be considered during analysis. To address this issue, ultra high-depth sequencing and additional association analyses are required. Still, research into the mitochondrial

genome is relatively neglected, since previous studies using massively parallel sequencing have mainly focused on the characterization of the cancer nuclear genome. Thanks to the abundant copy numbers of mitochondria, obtaining mitochondrial sequences is a common bonus derived from whole cancer nuclear genome sequencing. Taking advantage of this mitochondrial genome information will hopefully provide us with a better understanding of the associations between the breast cancer genome and the diverse range of breast cancer phenotypes. According to our experience, even using low-coverage genomic sequencing (typically one gigabase per sample), we can obtain 100% coverage and more than 300× depth of mitochondrial genome sequence. Alternatively, implementation of custom-designed enrichment assays that specifically capture mtDNA from total isolated DNA can be used to achieve in depth target sequencing.

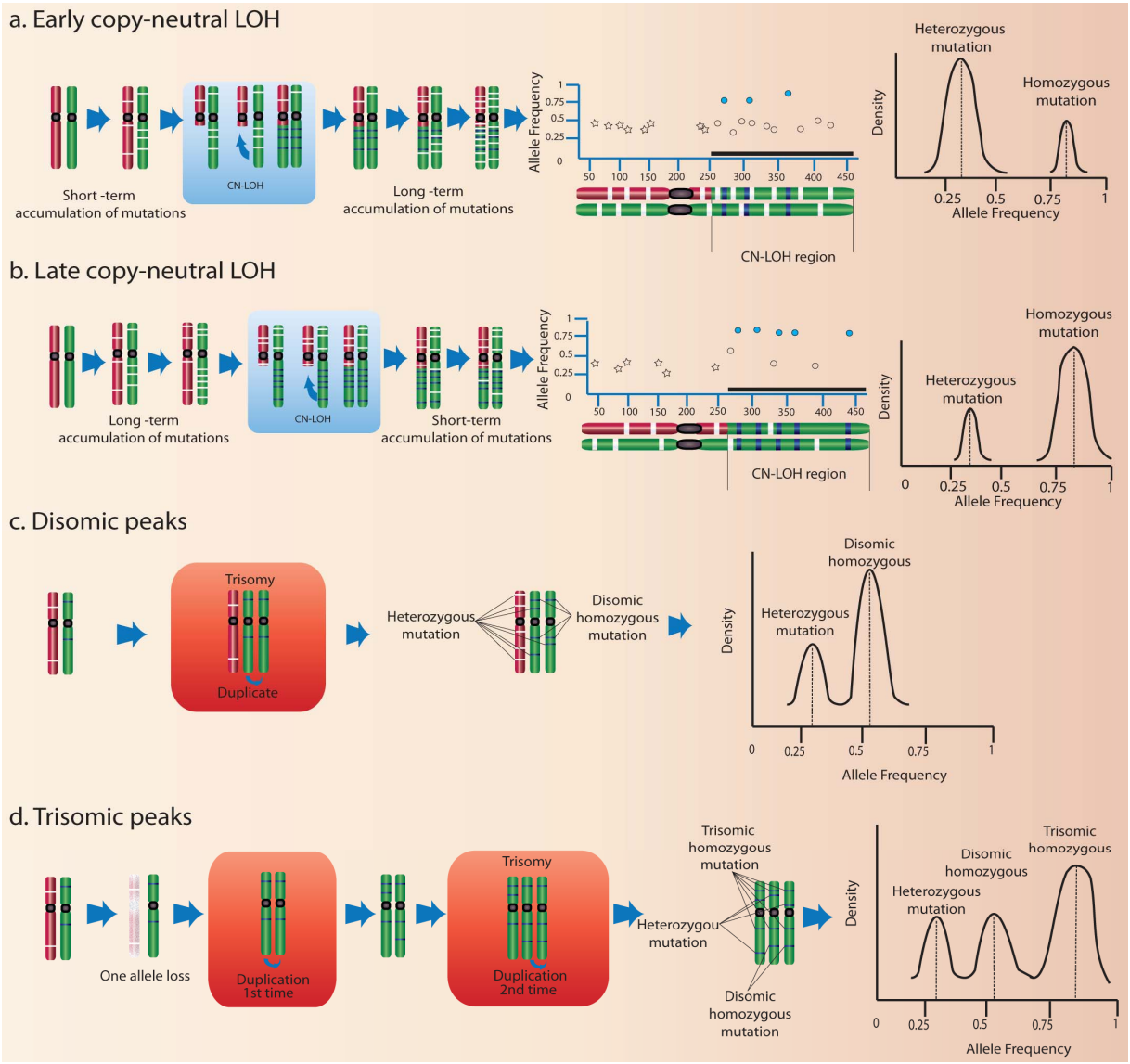
2.3. The Temporal Order of Genome Changes in the Evolution of the Breast Cancer Genome

Molecular characterization of human cancers usually gives a catalogue of genomic and epigenetic abnormalities reflecting years of somatic changes until the sampling time point [56]. Efforts to elucidate the temporal order of aberrations are performed by examination of a series of samples such as paired matched primary tumors and metastases or sporadic samples ordered according to different clinicopathological stages. These studies have revealed mutations associated with tumor progression and metastasis [57,58]. The differences between the primary tumor and its metastasis actually present a molecular profile of the late stages of tumorigenesis, and the molecular characterization of progression between sporadic samples intrinsically contains bias from different genetic backgrounds.

Copy neutral loss of heterozygosity (CN-LOH), as a frequently observed event in tumors, offers a unique opportunity for illustrating the longitudinal evolution of somatic events, beginning early in tumorigenesis in a single cancer [56,59]. CN-LOH, also referred to as uniparental disomy (UPD), is a loss of one copy (allele) of a heterozygous chromosomal region followed by a duplication of the other allele, yielding a homozygous chromosomal region without a copy number change (**Figures 1(a)** and **1(b)**). The process of CN-LOH can reveal important information contained in the evolutionary history of somatic aberrations: If a mutation precedes a regional UPD duplication, its copy number is doubled, *i.e.* homozygous, and mutations following such a duplication event appear in haploid copy number, *i.e.* heterozygous [56,59]. Based on this principle, simple mutations preceding a chromosomal duplication event show discretely higher copy numbers compared to those occurring after duplications and the ratio of heterozygous to homozygous mutations in CN-LOH regions directly reflects the temporal order of the duplication in tumorigenesis (**Figures 1(a)** and **1(b)**) [56]. In practical analysis, mutants can be discretely classified as homozygous mutations (high allele frequency) and heterozygous mutations (low allele frequency). The difference in allele frequency, *i.e.* shifts between homozygous and heterozygous mutations, can reveal the temporal order of genetic events that occurred in different regions in a single cancer genome. Individual somatic homozygous mutations accompanied by abundant heterozygous mutations in a CN-LOH region, implies that the homozygous mutations are early events in tumorigenesis, since a long period after a duplication event would allow this region to accumulate numerous new heterozygous mutations. On the other hand, a majority of homozygous mutations with a concurrent minority of heterozygous mutations implies that a new duplication event has occurred in the recent past, in which the previous heterozygous mutations have been lost and only one allele's information is retained and doubled. Thus new heterozygous mutations are quite limited due to the short period of accumulation after the duplication event.

This principle is also valid for trisomic regions. A trisomic region can be obtained through two different patterns: It can be the result of a simple duplication in which one allelic chromosomal region is doubled. In this case, the trisomic region harbors both heterozygous and disomic homozygous mutations (**Figure 1(c)**). Alternatively, a CN-LOH event could be followed by a secondary duplication to generate a trisomic region. In this scenario, the trisomic region harbors three types of mutations; heterozygous, disomic homozygous and trisomic homozygous mutations (**Figure 1(d)**).

Taken together, combining the information of the allelic frequency of mutations and the corresponding chromosomal copy number allows the measurement of the relative order of progressive events determining a cancer's individuality [56,59]. Durinck *et al.* recently applied this principle to delineate the temporal order in cancer evolution of skin and ovarian cancers [56]. In that study, based on investigations of the allele frequency of the mutations and the corresponding copy number profile, the mutation of *TP*53 was revealed as an initial event prior to the substantial numbers somatic mutations in tumor development of both types of cancers. Notably, the method introduced by Durinck *et al.* can sharply delineate the wide spread genomic instability for any type of cancer, setting the stage for determining the genetic events in the progression of breast cancer too. Delineation of the temporal order in cancer evolution will offer important information for future characterization of the succession of molecular changes and identification of



(a) and (b) show the principle of determining the temporal order of point mutations and copy-neutral loss of heterozygosity (CN-LOH) events [56]. In (a), homologous chromosomes (one chromosome is in red and its homolog is in green) accumulate different mutations in their alleles. These mutations are heterozygous (yellow line). A CN-LOH event (highlighted by blue rectangle) occurs at an early stage. Thus, the number of heterozygous mutations (yellow lines) is limited due to a relatively short time allowed for mutations to accumulate. During the CN-LOH event, the loss of one allelic chromosomal region is compensated by duplication of its homolog. The previously heterozygous mutations on the homolog become homozygous (dark blue lines) by this duplication event and, thus can be classified as early. The heterozygous mutations not located in CN-LOH regions remain intact. The new mutations arising after the CN-LOH event are heterozygous (yellow lines). Since this CN-LOH occurred early, a long period allowed for accumulation of more new heterozygous mutations prior to the sampling time point (the left pane in (a)). By contrast, a majority of homozygous mutations with a concurrent minority of heterozygous mutations implies that a duplication event has occurred in the recent past, in which the previous heterozygous mutations have become homozygous and new heterozygous mutations are limited due to the short period of accumulation after the duplication event (the left pane in (b)). By using sequencing or microarray technologies, heterozygous mutations (newly accumulated, indicated by open circles) and homozygous mutations (generated by the duplication, indicated by blue solid circles) in CN-LOH region (indicated by a horizontal thick solid black line) can be identified by their allele frequency. Mutations located in non-CN-LOH regions are shown by open stars (the middle panes in (a) and (b). A statistical model is applied to determine the temporal order of CN-LOH by calculating the densities for the allele frequency of heterozygous and homozygous mutations. The ratio of heterozygous to homozygous mutations in CN-LOH regions directly reflects the temporal order of the duplications in tumorigenesis (the right panes in (a) and (b). This principle can also be applied to determine the temporal order for trisomic regions. A trisomic region can be acquired by two distinct types of events: It can be the result of a simple duplication (highlighted by a red rectangle) in which one allelic chromosomal region is doubled. In this case, the trisomic region harbors both heterozygous and disomic homozygous mutations (c).Alternatively, a CN-LOH event (highlighted by a red rectangle) is followed by a secondary duplication (highlighted by a second red rectangle) to generate a trisomic region. In this scenario, the trisomic region harbors three types of mutations; heterozygous: disomic homozygous and trisomic homozygous mutations (d).

Figure 1. Conceptual framework defining the temporal order of genetic events in cancer genome evolution based on the relationship between point mutations and chromosome aberrations.

driver mutations in early breast cancer tumorigenesis, thereby supporting the development of novel cancer detection assays and the establishment of new innovative targeted treatment modalities [56].

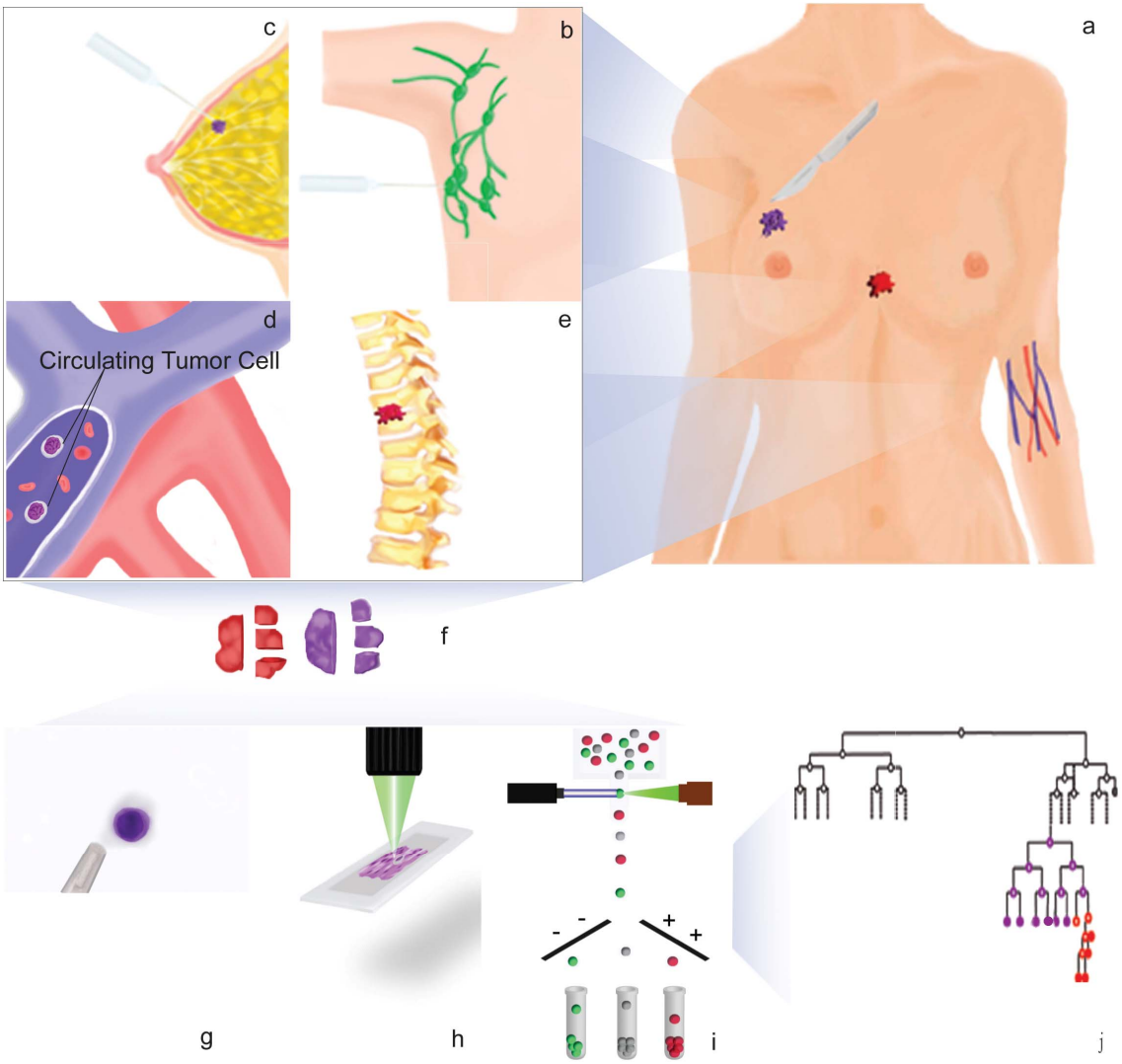
2.4. Single-Cell Sequencing

Breast cancer is a complex disease in part because the progression of breast cancer is a dynamic evolutionary process in the temporal dimension, and in part because breast cancer neoplasms contain highly heterogeneous cell populations in the spatial dimension. Tumor heterogeneity is an unavoidable fact in cancer research, because it is related to many of the important features of tumorigenesis including tumor progression, metastasis and therapeutic resistance [60,61]. Breast cancer is a typically heterogeneous cancer type, composed of diverse malignant epithelial subpopulations mixed with non-malignant tissues, such as infiltrating stromal cells, and cells from the immune system, such as infiltrating lymphocytes [62]. In some scenarios, normal cell populations may contribute to more than 50% of the total extracted DNA or RNA [63]. To address tumor heterogeneity, one solution is to select samples enriched for tumor content (at least 80%) and perform in depth sequencing to obtain sufficient sequenced data for characterization of dominant cancerous populations. This strategy is not optimal for studies aimed at reconstructing the evolutionary history and revealing the hierarchical structures in cell populations, since the subtle, important information from special rare subpopulations of cells may be masked, or even lost, in the data obtained from mixed bulk populations. Recently developed sequencing approaches for single cells at transcriptomic [64], genomic (DNA copy number profile) [65] and exomic [66,67] levels provide a new strategy for improved characterization of tumorigenesis. These approaches also offer promising tools for the early detection of compromised genes involved in cancer initiation, deciphering intratumour heterogeneity, monitoring the most malignant cells and capturing circulating tumor cells, thus guiding clinical therapy [63] (**Figure 2**).

Isolation of individual cellsis a prerequisite for single-cell genomic and transcriptomic analyses. Several attempts have been made to stratify cell subpopulations using regional macrodissection, fluorescence-activated cell sorting (FACS), laser capture microdissection (LCM) and other forms of micromanipulation (**Figure 2**). Macrodissection can retain the anatomical information, thereby providing a possibility to clarify the relationship between cells in special proximity where they share the same microenvironment. Additionally, this method is easily performed without the requirement of special equipment. However, one drawback of this is that it only can provide a gross stratification rather than single-cell resolution. FACS can collect cells according to the fluorescent characteristics of each cell, but selected cell populations based on limited number of labeled features may remain heterogeneous according to other cellular or molecular properties. In addition, anatomical information would be lost in the procedure of making the suspension of cells from dissociated tissue. LCM enables users to individually collect target cells, thereby providing an ideal and well characterized biological material for subsequent analysis. But LCM is labor-intensive and time-consuming. Micromanipulation can also capture single cells from cultured cell, dissociated tissue or biopsy material according to a given feature, but with the same shortcomings as LCM.

The amount of material isolated from individual cells using the above approaches is usually very small. Thus, an amplification step of the DNA or mRNA (through amplification of cDNA) extracted from captured single cells is necessary. Whole genome DNA amplification approaches, such as PCR-based amplification [68,69] and isothermal multiple displacement amplification [70], provide tools for relatively unbiased increasing of DNA material from single cells. The method of isothermal multiple displacement in particular has been demonstrated to ensure a highly efficient and good quality representative amplification of the template genome [70]. RNA is prone to degradation, thus stabilization of RNA is necessary for single-cell transcriptomic analysis. To maximize the sensitivity of subsequent sequencing analyses, elimination of genomic DNA contamination is also recommended. The methods for increasing the amounts of RNA include linear in vitro transcription (IVT)-based and exponential PCR-based methods [71]. With improvement in methodologies, ~10 pg of total RNA and ~0.1 pg of mRNA, in a typical mammalian cell, can be converted to up to 3-kb fragments of cDNA, followed by uniform amplification that can increase the yield around ten million-fold to match the requirement for downstream analyses with a high reproducibility [64,71,72]. The techniques and methods applied in single-cell transcriptome analyses have recently been highlighted and discussed by Tang *et al.* [71].

In early studies, single-cell genomic and transcriptomic analyses mainly utilized microarray-based technologies such as array-CGH and gene expression microarrays. Massively parallel sequencing not only ensures deeper measurement of DNA copy number and transcriptomic profiles, but also directly provides sequence information. Recently, some attempts utilizing the application of massively parallel sequencing platforms for single-cell analysis, have been reported [64-67]. Navin and his colleagues applied single-nucleus sequencing (SNS) to



Multiple approaches can be used to obtain single cells for different analyses [64-67]. Samples can be obtained as follows: Surgical removal of primary breast cancer (purple) (a); Fine needle biopsy of axillary lymph node (b); Fine needle biopsy of primary breast tumor (c); Capturing of circulating tumor cells from blood (d); Surgical dissection of thoracic vertebral metastasis (red) (e); Macrodissection collects primary breast cancer (purple) and distant metastasis (red) (f); Isolation of cancer single cells can be performed by micromanipulation; g), laser capture microdissection (h); and fluorescence-activated cell sorting (FACS) (i); Polygenetic analysis at the single-cell level is applied to uncover the evolutionary relationship between cancer cell populations (j).

Figure 2. Schematic indicating the framework of single cell sequencing analysis for breast cancer.

investigate tumor population structure and evolution in two human breast cancer cases through the investigation of copy number profiles [65]. SNS was demonstrated to be a reproducible method the sequencing result from a single-cell showed a high correlation ($R^2 > 0.85$) with that from a million cells [65]. Tang *et al* first reported the transcriptome analysis of single cell mouse blastomeres in combination with massively parallel sequencing technology [71]. In their study, numerous known transcripts and splicing isoform expression patterns were identified at single cell resolution. Notably, thousands of previously unknown exon exon junctions were found in the transcriptome, indicating the potential value of this application in transcriptomic analysis for cancer single cells [64]. Recently, single-cell exome sequencing method was introduced [66,67]. Hou and his colleagues carried out whole-exome single-cell sequencing of a JAK2-negative myeloproliferative neoplasm [67] and Xu and his colleagues carried out single-cell exome sequencing of a clear cell renal cell carcinoma (ccRCC) [66]. These two studies opened the way for detailed analyses of a variety of tumor types and other complex diseases, thereby sup-

porting the development of more effective therapies, which are targeted to the relevant cells [66,67].

Phylogenetic analysis is a commonly used bioinformatic tool in research of the evolutionary relationship between cancer cell subpopulations [57,58,65]. Cancer progression can be regarded as a micro-evolutionary process: A cancer begins with an initiating aberration in a normal cell that confers a selective growth advantage. Subsequently, successive clonal expansions occur fueled by the acquisition of additional aberrations, corresponding to progression stages. At the same time, there is a massive loss of clones with lower fitness. In the late phases of tumorigenesis, founder cells within the cancer give rise to seeding clones that can colonize distant organs and hence initiate a disease stage characterized by metastatic lesions [73]. In phylogenetic analysis, if single cells have similar DNA sequences they likely originate from a common ancestor and locate in a lineage branch with short evolutionary distances in a phylogenetic tree. The lineage branch will be split, when a 'speciation' (founder cell) event occurs, in which a single ancestral lineage gives rise to two or more daughter lineages (extended clones). Consequently, through phylogenetic analysis of data generated by sequencing of multiple samples ordered by the progressive stages of cancer, such as the normal epithelium, carcinoma in situ, infiltrating carcinoma, lymph node metastasis, and distant metastasis, it would be possible to construct the evolutionary relationships between single cells, identify the founders responsible for initiating next stage and determine their molecular features as well as estimate time intervals between the successive stages.

Single-cell sequencing and related bioinformatic analyses open a new avenue for breast cancer research. These methods may have great importance for future breast cancer genome studies-especially with a continuous reduction in sequencing costs and the emergence of more powerful sequencing technologies. Limited to current conditions, there are some drawbacks in the present methodologies, such as the relatively low coverage in single-cell sequencing and sequence information not being fully exploited [65]. Compared with genomic information, transcriptomes from single cells present with more variability due to the influences from epigenetic events, the circadian clock, the cell cycle, microenvironmental niches as well as "transcriptional noise" [71]. The evidence of stochastic characteristics in gene expression among single cells underscores the importance and necessity of applying multiple single-cell transcriptomic analyses, and also highlights the challenge in understanding and interpreting the gene expression results from individual cells [71]. Epigenetic abnormalities may also contribute to breast cancer progression, but DNA methylation analysis for single cells has not yet been developed, mainly due to lack of proper amplification methods. At present, no DNA amplification method is able to properly retain the DNA methylation information in newly amplified DNA copies. If a technical breakthrough can occur in single cell epigenetic analysis, the evolutionary models currently being constructed on the basis of single-cell genomic data will be improved by addition of epigenetic information. Simultaneous analysis of DNA, DNA modification and mRNA from the same individual cells will be an ideal strategy for the comprehensive and precise interpretation of the functional alterations occurring in single cancer cells.

Accomplishing the above goal will depend on advances in sequencing technology. Nanopore DNA sequencing is one of a number of promising single molecule sequencing approaches that can directly sequence DNA or RNA molecules using tiny amounts of material without the requirement of an amplification and labeling step [74,75]. DNA methylation information would be available in the direct readout by precisely distinguishing unmethylated cytosines from methylated cytosines in the DNA sequence [76]. Therefore, we believe single-cell sequencing in combination with novel sequencing technologies will bring a revolutionary change in breast cancer research.

2.5. The Microbiome

Beside the aforementioned progress, microbiome and metagenomic studies will be other promising fields in cancer research. Microbes inhabiting the human body, including eukaryotes, archea, bacteria and viruses, are collectively known as microbiome. Bacteria alone are estimated to outnumber human cells by an order of magnitude and the gene set of a microbiome is approximately 150 times larger than the human gene complement [77,78]. Increasing evidences implicate the microbiome as crucially important for metabolism, immune defense, and the development of diverse disorders including cancers. In recent years, microbiome research has been boosted through such large-scale sequence-based human microbiome projects as Metagenomics of the Human Intestinal Tract (MetaHIT) and the Human Microbiome Project (HMP). A variety of microbial communities have been characterized by massively parallel sequencing, sequence analysis and functional studies [77-79]. Following the establishment of microbiome catalogs and references as well as the development of laboratory and bioinformatic approaches-especially, investigations of the correlation with host phenotype-the microbiome will become an important aspect in cancer research. In the context of breast cancer research, the next effort will be to establish cause and effect relationship between the microbiome and breast cancer susceptibility.

3. Challenges and Progress

Rapid development of improved methods for studying the breast cancer genome poses many future challenges. Some challenges will arise from analysis of numerous short reads the amounts of which are several magnitudes higher than those traditionally obtained by Sanger sequencing. Thus, the first challenge is to meet growing computational requirements such as sufficient storage, data transfer and assembly. Secondly, there is an urgent need for fast, accurate and user-friendly bioinformatic approaches for data mining to realize the full potential of these improved sequencing technologies. Numerous recently published bioinformatic tools offer a wide variety of options for broad "omics" analysis, but also result in questions on which method provides the best results. Thus, exhaustive comparisons between algorithms, incorporating miscellaneous analytic methods into an integrative pipeline, evaluating the statistical power, sensitivity and specificity of software developed for the same analyses, will be required for standardization of bioinformatics inanalyses of the breast cancer genome.

In addition to bioinformatic methods, as the cornerstone for cancer genome research, more representative human reference genomes are greatly required. With growing number of published reference genomes and an increasing knowledge of the variations in the normal human genome, the previous single consensus representation of the genome is not sufficient, especially in regions with complex allelic diversity. This challenge is being addressed by an effort to create assemblies that better represent the diversity (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/). At the same time, functional annotation projects will provide the necessary information for elucidating dysfunctions of protein-coding genes (GENCODE project (http://www.sanger.ac.uk/gencode/)) [80] and defining functional elements (http://encodeproject.org/ENCODE/) [81]. Another important resource for cancer genome research is well-annotated databases. Advances in understanding the cancer genome depend on the access to comprehensive catalogues of variations in the human genome in normal populations. These normal variations are well collected, curated and updated by many different databases, according to different variation patterns, for example, SNPs in dbSNP [82] and the HapMap database [83], copy number variations in DGV [84], dbVar (http://www.ncbi.nlm. nih.gov/dbvar/) and comprehensive human genome variations in The 1000 Genomes Project [85].

A more difficult challenge is the defining of normal epigenetic references, because epigenetic information is reversible and presents in highly tissue-specific and developmentally associated patterns. Recently, the Epigenomics Mapping Consortium has been working to produce a public resource of epigenomic maps (DNA methylation, histone modifications and related chromatin features) for stem cells and primary *ex vivo* human fetal and adult tissues representative of normal human biology, thereby offering the normal counterpart for cancer research [86]. These databases, either presenting the repertoire of oncogenic variations [3] or collections of normal variations (see above), which are well-curated and periodly updated, have provided profound value for cancer genome research by providing comprehensive references and aiding in identifying novel aberrations for individual studies. The relationship between large-scale sequencing projects for the construction of reference databases and the many milestone events of cancer genome sequencing has been well described in a recent review [87].

Large-scale sequencing of cancer genomes, including breast cancer, is rapidly providing an astronomical amount of data, which will offer many new candidates that will be assumed to play pivotal roles for a given cancer phenotype. Careful functional studies of mutated genes are required for ultimate proof of the relationship between cancer gene status and clinical behavior [41]. How to validate these candidate genes will become a crucial challenge for researchers using routine assays such as cell lines or animal models. High-throughput RNA interference screens in combination with the adaptation of existing model systems, will be a promising tools for refining the potential candidates provided by large-scale sequencing by further functional studies [16].

The many applications and analyses using massively parallel sequencing platforms have not yet been fully optimized, standardized and systematically evaluated for samples routinely processed in cancer pathology in clinical practice. This poses a gap between bench and bedside. To address this important matter, comprehensive coordinated international collaboration is necessary for the standardization of laboratory endeavors and bioinformatic analyses [24].

4. Conclusion

The completion of the draft of the human genome signaled the ushering in of the genomic era [88]. Thereafter, revolutionary breakthroughs in sequencing technology, a spectacular blossoming of bioinformatics and an accelerating accumulation of sequencing data, bring unprecedented opportunities as well as challenges to cancer research. Recently, the International Cancer Genome Consortium (ICGC) was launched to coordinate the large-scale sequencing of the genomes, epigenomes, and transcriptomes for 50 different cancer types and/or subtypes [24]. The goal of the project is to define catalogues of cancer genomic abnormalities and translate the findings of these genomic analyses into clinical utility [24]. This

project not only has a profound influence on present cancer research, but more importantly, it heralds the start of the era of personalized medicine [24]. Consequently, we can anticipate that sequencing and genomic analysis will play an important role in clinical practice. In the not too distant future, sequencing may become a population screening approach for the early detection of breast cancer, and sequencing of the breast cancer genome of individual patients may be routinely applied to confer guidelines for personalized breast cancer patient management.

5. Acknowledgements

We thank for the support from the project "Molecular Tools for Optimal Personalized Treatment of Breast Cancer" under the auspices of Sino-Danish Breast Cancer Research Centre, financed by the Danish National Research Foundation (Grundforskningsfonden), and the National Natural Science Foundation of China (30890032, 31161130357). We are also grateful to the Chinese 863 Program (2012AA02A201, 2012AA02A502), Guangdong Innovative Research Team Program (2009010016) and A Race Against Breast Cancer.

REFERENCES

[1] J. Ferlay, H. R. Shin, F. Bray, *et al.*, "Estimates of Worldwide Burden of Cancer in 2008: Globocan 2008," *International Journal of Cancer*, Vol. 127, No. 12, 2010, pp. 2893-2917.

[2] M. R. Stratton, P. J. Campbell and P. A. Futreal, "The Cancer Genome," *Nature*, Vol. 458, No. 7239, 2009, pp. 719-724. doi:10.1038/nature07943

[3] S. A. Forbes, N. Bindal, S. Bamford, *et al.*, "Cosmic: Mining Complete Cancer Genomes in the Catalogue of Somatic Mutations in Cancer," *Nucleic Acids Research*, Vol. 39, No. Database Issue, 2011, pp. D945-D950.

[4] M. Song, K. M. Lee and D. Kang, "Breast Cancer Prevention Based on Gene-Environment Interaction," *Molecular Carcinogenesis*, Vol. 50, No. 4, 2011, pp. 280-290. doi:10.1002/mc.20639

[5] A. Petherick, "Environment and Genetics: Making Sense of the Noise," *Nature*, Vol. 485, No. 7400, 2012, pp. S64-S65. doi:10.1038/485S64a

[6] N. Mavaddat, A. C. Antoniou, D. F. Easton, *et al.*, "Genetic Susceptibility to Breast Cancer," *Molecular Oncology*, Vol. 4, No. 3, 2010, pp. 174-191. doi:10.1016/j.molonc.2010.04.011

[7] L. M. Butcher and S. Beck, "Future Impact of Integrated High-Throughput Methylome Analyses on Human Health and Disease," *Journal of Genetics and Genomics*, Vol. 35, No. 7, 2008, pp. 391-401. doi:10.1016/S1673-8527(08)60057-0

[8] J. Jovanovic, J. A. Ronneberg, J. Tost, *et al.*, "The Epigenetics of Breast Cancer," *Molecular Oncology*, Vol. 4, No. 3, 2010, pp. 242-254. doi:10.1016/j.molonc.2010.04.002

[9] Y. Huang, S. Nayak, R. Jankowitz, *et al.*, "Epigenetics in Breast Cancer: What's New?" *Breast Cancer Research*, Vol. 13, No. 6, 2011, p. 225. doi:10.1186/bcr2925

[10] M. Margulies, M. Egholm, W. E. Altman, *et al.*, "Genome Sequencing in Microfabricated High-Density Picolitre Reactors," *Nature*, Vol. 437, No. 7057, 2005, pp. 376-380.

[11] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, *et al.*, "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry," *Nature*, Vol. 456, No. 7218, 2008, pp. 53-59. doi:10.1038/nature07517

[12] J. Shendure, G. J. Porreca, N. B. Reppas, *et al.*, "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome," *Science*, Vol. 309, No. 5741, 2005, pp. 1728-1732. doi:10.1126/science.1117389

[13] R. Drmanac, A. B. Sparks, M. J. Callow, *et al.*, "Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays," *Science*, Vol. 327, No. 5961, 2010, pp. 78-81. doi:10.1126/science.1181498

[14] J. M. Rothberg, W. Hinz, T. M. Rearick, *et al.*, "An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing," *Nature*, Vol. 475, No. 7356, 2011, pp. 348-352. doi:10.1038/nature10242

[15] H. Stranneheim and J. Lundeberg, "Stepping Stones in DNA Sequencing," *Biotechnology Journal*, 2012.

[16] L. Chin and J. W. Gray, "Translating Insights from the Cancer Genome into Clinical Practice," *Nature*, Vol. 452, No. 7187, 2008, pp. 553-563. doi:10.1038/nature06914

[17] A. Mortazavi, B. A. Williams, K. McCue, *et al.*, "Mapping and Quantifying Mammalian Transcriptomes by Rna-Seq," *Nature Methods*, Vol. 5, No. 7, 2008, pp. 621-628. doi:10.1038/nmeth.1226

[18] F. Ozsolak and P. M. Milos, "Rna Sequencing: Advances, Challenges and Opportunities," *Nature Reviews Genetics*, Vol. 12, No. 2, 2011, pp. 87-98. doi:10.1038/nrg2934

[19] C. Lu, S. S. Tej, S. Luo, *et al.*, "Elucidation of the Small Rna Component of the Transcriptome," *Science*, Vol. 309, No. 5740, 2005, pp. 1567-1569. doi:10.1126/science.1114112

[20] K. P. McCormick, M. R. Willmann and B. C. Meyers, "Experimental Design, Preprocessing, Normalization and Differential Expression Analysis of Small Rna Sequencing Experiments," *Silence*, Vol. 2, No. 1, 2011, p. 2. doi:10.1186/1758-907X-2-2

[21] L. Ding, M. J. Ellis, S. Li, *et al.*, "Genome Remodelling in a Basal-Like Breast Cancer Metastasis and Xenograft," *Nature*, Vol. 464, No. 7291, 2010, pp. 999-1005. doi:10.1038/nature08989

[22] S. P. Shah, R. D. Morin, J. Khattra, *et al.*, "Mutational Evolution in a Lobular Breast Tumour Profiled at Single Nucleotide Resolution," *Nature*, Vol. 461, No. 7265, 2009, pp. 809-813. doi:10.1038/nature08489

[23] P. J. Stephens, D. J. McBride, M. L. Lin, *et al.*, "Complex Landscapes of Somatic Rearrangement in Human Breast Cancer Genomes," *Nature*, Vol. 462, No. 7276, 2009, pp. 1005-1010. doi:10.1038/nature08645

[24] T. J. Hudson, W. Anderson, A. Artez, *et al.*, "International Network of Cancer Genome Projects," *Nature*, Vol. 464, No. 7291, 2010, pp. 993-998. doi:10.1038/nature08987

[25] G. J. Porreca, K. Zhang, J. B. Li, *et al.*, "Multiplex Amplification of Large Sets of Human Exons," *Nature Methods*, Vol. 4, No. 11, 2007, pp. 931-936. doi:10.1038/nmeth1110

[26] G. Robertson, M. Hirst, M. Bainbridge, *et al.*, "Genome-Wide Profiles of Stat1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing," *Nature Methods*, Vol. 4, No. 8, 2007, pp. 651-657. doi:10.1038/nmeth1068

[27] J. Li, F. Gao, N. Li, *et al.*, "An Improved Method for Genome Wide DNA Methylation Profiling Correlated to Transcription and Genomic Instability in Two Breast Cancer Cell Lines," *BMC Genomics*, Vol. 10, 2009, pp. 223.

[28] D. Serre, B. H. Lee and A. H. Ting, "Mbd-Isolated Genome Sequencing Provides a High-Throughput and Comprehensive Survey of DNA Methylation in the Human Genome," *Nucleic Acids Research*, Vol. 38, No. 2, 2010, pp. 391-399. doi:10.1093/nar/gkp992

[29] F. V. Jacinto, E. Ballestar and M. Esteller, "Methyl-DNA Immunoprecipitation (Medip): Hunting down the DNA Methylome," *Biotechniques*, Vol. 44, No. 1, 2008, pp. 35, 37, 39 passim. doi:10.2144/000112708

[30] R. Lister, M. Pelizzola, R. H. Dowen, *et al.*, "Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences," *Nature*, Vol. 462, No. 7271, 2009, pp. 315-322. doi:10.1038/nature08514

[31] Y. Li, J. Zhu, G. Tian, *et al.*, "The DNA Methylome of Human Peripheral Blood Mononuclear Cells," *PLoS Biology*, Vol. 8, No. 11, 2010, p. e1000533. doi:10.1371/journal.pbio.1000533

[32] A. Meissner, T. S. Mikkelsen, H. Gu, *et al.*, "Genome-Scale DNA Methylation Maps of Pluripotent and Differentiated Cells," *Nature*, Vol. 454, No. 7205, 2008, pp. 766-770.

[33] N. Li, M. Ye, Y. Li, *et al.*, "Whole Genome DNA Methylation Analysis Based on High Throughput Sequencing Technology," *Methods*, Vol. 52, No. 3, 2010, pp. 203-212. doi:10.1016/j.ymeth.2010.04.009

[34] P. W. Laird, "Principles and Challenges of Genome-Wide DNA Methylation Analysis," *Nature Reviews Genetics*, Vol. 11, No. 3, 2010, pp. 191-203. doi:10.1038/nrg2732

[35] C. Bock, E. M. Tomazou, A. B. Brinkman, *et al.*, "Quantitative Comparison of Genome-Wide DNA Methylation Mapping Technologies," *Nature Biotechnology*, Vol. 28, No. 10, 2010, pp. 1106-1114. doi:10.1038/nbt.1681

[36] R. A. Harris, T. Wang, C. Coarfa, *et al.*, "Comparison of Sequencing-Based Methods to Profile DNA Methylation and Identification of Monoallelic Epigenetic Modifications," *Nature Biotechnology*, Vol. 28, No. 10, 2010, pp. 1097-1105. doi:10.1038/nbt.1682

[37] S. T. Bennett, C. Barnes, A. Cox, *et al.*, "Toward the 1000 Dollars Human Genome," *Pharmacogenomics*, Vol. 6, No. 4, 2005, pp. 373-382. doi:10.1517/14622416.6.4.373

[38] T. Sorlie, C. M. Perou, R. Tibshirani, *et al.*, "Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications," *Proceedings of the National Academy of Sciences of the United States*, Vol. 98, No. 19, 2001, pp. 10869-10874. doi:10.1073/pnas.191367098

[39] M. Meyerson, S. Gabriel and G. Getz, "Advances in Understanding Cancer Genomes through Second-Generation Sequencing," *Nature Reviews Genetics*, Vol. 11, No. 10, 2010, pp. 685-696. doi:10.1038/nrg2841

[40] P. A. Cowin, M. Anglesio, D. Etemadmoghadam, *et al.*, "Profiling the Cancer Genome," *Annual Review of Genomics and Human Genetics*, Vol. 11, No., 2010, pp. 133-159.

[41] T. Sjoblom, "Systematic Analyses of the Cancer Genome: Lessons Learned from Sequencing Most of the Annotated Human Protein-Coding Genes," *Current Opinion in Oncology*, Vol. 20, No. 1, 2008, pp. 66-71. doi:10.1097/CCO.0b013e3282f31108

[42] T. Sjoblom, S. Jones, L. D. Wood, *et al.*, "The Consensus Coding Sequences of Human Breast and Colorectal Cancers," *Science*, Vol. 314, No. 5797, 2006, pp. 268-274. doi:10.1126/science.1133427

[43] L. D. Wood, D. W. Parsons, S. Jones, *et al.*, "The Genomic Landscapes of Human Breast and Colorectal Cancers," *Science*, Vol. 318, No. 5853, 2007, pp. 1108-1113. doi:10.1126/science.1145720

[44] J. Lin, C. M. Gan, X. Zhang, *et al.*, "A Multidimensional Analysis of Genes Mutated in Breast and Colorectal Cancers," *Genome Research*, Vol. 17, No. 9, 2007, pp. 1304-1318. doi:10.1101/gr.6431107

[45] Q. Zhao, E. F. Kirkness, O. L. Caballero, *et al.*, "Systematic Detection of Putative Tumor Suppressor Genes through the Combined Use of Exome and Transcriptome Sequencing," *Genome Biology*, Vol. 11, No. 11, 2010, p. R114. doi:10.1186/gb-2010-11-11-r114

[46] K. Inaki, A. M. Hillmer, L. Ukil, *et al.*, "Transcriptional Consequences of Genomic Structural Aberrations in Breast Cancer," *Genome Research*, Vol. 21, No. 5, 2011, pp. 676-687. doi:10.1101/gr.113225.110

[47] Z. Sun, Y. W. Asmann, K. R. Kalari, *et al.*, "Integrated Analysis of Gene Expression, Cpg Island Methylation, and Gene Copy Number in Breast Cancer Cells by Deep Sequencing," *PLoS One*, Vol. 6, No. 2, 2011, p. e17490. doi:10.1371/journal.pone.0017490

[48] S. Anderson, A. T. Bankier, B. G. Barrell, *et al.*, "Sequence and Organization of the Human Mitochondrial Genome," *Nature*, Vol. 290, No. 5806, 1981, pp. 457-465. doi:10.1038/290457a0

[49] D. L. Croteau and V. A. Bohr, "Repair of Oxidative Damage to Nuclear and Mitochondrial DNA in Mammalian Cells," *The Journal of Biological Chemistry*, Vol. 272, No. 41, 1997, pp. 25409-25412.

doi:10.1074/jbc.272.41.25409

[50] D. J. Tan, R. K. Bai and L. J. Wong, "Comprehensive Scanning of Somatic Mitochondrial DNA Mutations in Breast Cancer," *Cancer Research*, Vol. 62, No. 4, 2002, pp. 972-976.

[51] P. Parrella, Y. Xiao, M. Fliss, *et al.*, "Detection of Mitochondrial DNA Mutations in Primary Breast Cancer and Fine-Needle Aspirates," *Cancer Research*, Vol. 61, No. 20, 2001, pp. 7623-7626.

[52] A. Chatterjee, E. Mambo and D. Sidransky, "Mitochondrial DNA Mutations in Human Cancer," *Oncogene*, Vol. 25, No. 34, 2006, pp. 4663-4674. doi:10.1038/sj.onc.1209604

[53] M. S. Fliss, H. Usadel, O. L. Caballero, *et al.*, "Facile Detection of Mitochondrial DNA Mutations in Tumors and Bodily Fluids," *Science*, Vol. 287, No. 5460, 2000, pp. 2017-2019. doi:10.1126/science.287.5460.2017

[54] C. Isaacs, L. R. Cavalli, Y. Cohen, *et al.*, "Detection of Loh and Mitochondrial DNA Alterations in Ductal Lavage and Nipple Aspirate Fluids from Hngh-Risk Patients," *Breast Cancer Research and Treatment*, Vol. 84, No. 2, 2004, pp. 99-105. doi:10.1023/B:BREA.0000018406.03679.2e

[55] N. G. Larsson, "Somatic Mitochondrial DNA Mutations in Mammalian Aging," *Annual Review of Biochemistry*, Vol. 79, 2010, pp. 683-706.

[56] S. Durinck, C. Ho, N. J. Wang, *et al.*, "Temporal Dissection of Tumorigenesis in Primary Cancers," *Cancer Discovery*, Vol. 1, No. 2, 2011, pp. 137-143. doi:10.1158/2159-8290.CD-11-0028

[57] P. J. Campbell, S. Yachida, L. J. Mudie, *et al.*, "The Patterns and Dynamics of Genomic Instability in Metastatic Pancreatic Cancer," *Nature*, Vol. 467, No. 7319, 2010, pp. 1109-1113. doi:10.1038/nature09460

[58] S. Yachida, S. Jones, I. Bozic, *et al.*, "Distant Metastasis Occurs Late during the Genetic Evolution of Pancreatic Cancer," *Nature*, Vol. 467, No. 7319, 2010, pp. 1114-1117. doi:10.1038/nature09515

[59] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, *et al.*, "A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome," *Nature*, Vol. 463, No. 7278, 2010, pp. 191-196. doi:10.1038/nature08658

[60] M. G. Daidone, R. Silvestrini, B. Valentinis, *et al.*, "Proliferative Activity of Primary Breast Cancer and of Synchronous Lymph Node Metastases Evaluated by [3h]-Thymidine Labelling Index," *Cell and Tissue Kinetics*, Vol. 23, No. 5, 1990, pp. 401-408.

[61] D. L. Dexter and J. T. Leith, "Tumor Heterogeneity and Drug Resistance," *Journal of Clinical Oncology*, Vol. 4, No. 2, 1986, pp. 244-257.

[62] M. Aubele and M. Werner, "Heterogeneity in Breast Cancer and the Problem of Relevance of Findings," *Analytical Cellular Pathology*, Vol. 19, No. 2, 1999, pp. 53-58.

[63] N. Navin and J. Hicks, "Future Medical Applications of Single-Cell Sequencing in Cancer," *Genome Medicine*, Vol. 3, No. 5, 2011, pp. 31. doi:10.1186/gm247

[64] F. Tang, C. Barbacioru, Y. Wang, *et al.*, "Mrna-Seq Whole-Transcriptome Analysis of a Single Cell," *Nature Methods*, Vol. 6, No. 5, 2009, pp. 377-382. doi:10.1038/nmeth.1315

[65] N. Navin, J. Kendall, J. Troge, *et al.*, "Tumour Evolution Inferred by Single-Cell Sequencing," *Nature*, Vol. 472, No. 7341, 2011, pp. 90-94. doi:10.1038/nature09807

[66] X. Xu, Y. Hou, X. Yin, *et al.*, "Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor," *Cell*, Vol. 148, No. 5, 2012, pp. 886-895. doi:10.1016/j.cell.2012.02.025

[67] Y. Hou, L. Song, P. Zhu, *et al.*, "Single-Cell Exome Sequencing and Monoclonal Evolution of a Jak2-Negative Myeloproliferative Neoplasm," *Cell*, Vol. 148, No. 5, 2012, pp. 873-885. doi:10.1016/j.cell.2012.02.028

[68] V. G. Cheung and S. F. Nelson, "Whole Genome Amplification Using a Degenerate Oligonucleotide Primer Allows Hundreds of Genotypes to Be Performed on Less than One Nanogram of Genomic DNA," *Proceedings of the National Academy of Sciences of the United States*, Vol. 93, No. 25, 1996, pp. 14676-14679. doi:10.1073/pnas.93.25.14676

[69] H. Telenius, N. P. Carter, C. E. Bebb, *et al.*, "Degenerate Oligonucleotide-Primed Pcr: General Amplification of Target DNA by a Single Degenerate Primer," *Genomics*, Vol. 13, No. 3, 1992, pp. 718-725. doi:10.1016/0888-7543(92)90147-K

[70] J. M. Lage, J. H. Leamon, T. Pejovic, *et al.*, "Whole Genome Analysis of Genetic Alterations in Small DNA Samples Using Hyperbranched Strand Displacement Amplification and Array-Cgh," *Genome Research*, Vol. 13, No. 2, 2003, pp. 294-307. doi:10.1101/gr.377203

[71] F. Tang, K. Lao and M. A. Surani, "Development and Applications of Single-Cell Transcriptome Analysis," *Nature Methods*, Vol. 8, No. 4 Suppl, 2011, pp. S6-11.

[72] K. Kurimoto, Y. Yabuta, Y. Ohinata, *et al.*, "An Improved Single-Cell Cdna Amplification Method for Efficient High-Density Oligonucleotide Microarray Analysis," *Nucleic Acids Research*, Vol. 34, No. 5, 2006, pp. e42. doi:10.1093/nar/gkl050

[73] D. Hanahan and R. A. Weinberg, "The Hallmarks of Cancer," *Cell*, Vol. 100, No. 1, 2000, pp. 57-70. doi:10.1016/S0092-8674(00)81683-9

[74] D. Branton, D. W. Deamer, A. Marziali, *et al.*, "The Potential and Challenges of Nanopore Sequencing," *Nature Biotechnology*, Vol. 26, No. 10, 2008, pp. 1146-1153. doi:10.1038/nbt.1495

[75] E. E. Schadt, S. Turner and A. Kasarskis, "A Window into Third-Generation Sequencing," *Human Molecular Genetics*, Vol. 19, No. R2, 2010, pp. R227-240. doi:10.1093/hmg/ddq416

[76] J. Clarke, H. C. Wu, L. Jayasinghe, *et al.*, "Continuous Base Identification for Single-Molecule Nanopore DNA Sequencing," *Nature Nanotechnology*, Vol. 4, No. 4, 2009, pp. 265-270. doi:10.1038/nnano.2009.12

[77] J. Qin, R. Li, J. Raes, *et al.*, "A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequenc-

ing," *Nature*, Vol. 464, No. 7285, 2010, pp. 59-65. doi:10.1038/nature08821

[78] C. Human Microbiome Project, "A Framework for Human Microbiome Research," *Nature*, Vol. 486, No. 7402, 2012, pp. 215-221. doi:10.1038/nature11209

[79] C. Human Microbiome Project, "Structure, Function and Diversity of the Healthy Human Microbiome," *Nature*, Vol. 486, No. 7402, 2012, pp. 207-214. doi:10.1038/nature11234

[80] J. Harrow, F. Denoeud, A. Frankish, *et al.*, "Gencode: Producing a Reference Annotation for Encode," *Genome Biology*, Vol. 7, Suppl. 1, 2006, pp. S41-S49.

[81] R. M. Myers, J. Stamatoyannopoulos, M. Snyder, *et al.*, "A User's Guide to the Encyclopedia of DNA Elements (Encode)," *PLoS Biology*, Vol. 9, No. 4, 2011, p. e1001046. doi:10.1371/journal.pbio.1001046

[82] S. T. Sherry, M. Ward and K. Sirotkin, "Dbsnp-Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation," *Genome Research*, Vol. 9, No. 8, 1999, pp. 677-679.

[83] D. M. Altshuler, R. A. Gibbs, L. Peltonen, *et al.*, "Integrating Common and Rare Genetic Variation in Diverse Human Populations," *Nature*, Vol. 467, No. 7311, 2010, pp. 52-58. doi:10.1038/nature09298

[84] A. J. Iafrate, L. Feuk, M. N. Rivera, *et al.*, "Detection of Large-Scale Variation in the Human Genome," *Nature Genetics*, Vol. 36, No. 9, 2004, pp. 949-951. doi:10.1038/ng1416

[85] 1000 Genomes Project Consortium, "A Map of Human Genome Variation from Population-Scale Sequencing," *Nature*, Vol. 467, No. 7319, 2010, pp. 1061-1073. doi:10.1038/nature09534

[86] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, *et al.*, "The Nih Roadmap Epigenomics Mapping Consortium," *Nature Biotechnology*, Vol. 28, No. 10, 2010, pp. 1045-1048. doi:10.1038/nbt1010-1045

[87] K. M. Wong, T. J. Hudson and J. D. McPherson, "Unraveling the Genetics of Cancer: Genome Sequencing and Beyond," *Annual Review of Genomics and Human Genetics*, 2011.

[88] E. S. Lander, L. M. Linton, B. Birren, *et al.*, "Initial Sequencing and Analysis of the Human Genome," *Nature*, Vol. 409, No. 6822, 2001, pp. 860-921. doi:10.1038/35057062

