Deep Sequencing in Cancer Research

Kenichi Yoshida, Masashi Sanada and Seishi Ogawa*

Cancer Genomics Project, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

*For reprints and all correspondence: Seishi Ogawa. E-mail: sogawa-tky@umin.ac.jp

Received August 26, 2012; accepted November 4, 2012

Cancer is caused by alterations in the cellular genome including single-nucleotide variations, small insertions and deletions (indels), copy number changes and other structural variations and, as such, their detection in a comprehensive manner is of critical importance for fully understanding cancer pathogenesis, improvement of diagnosis as well as the development of novel therapeutics. In this point of views, the recent development of massively parallel (or ‘next-generation’) sequencing technologies has provided an unprecedented possibility to accomplish this need by enabling single-nucleotide resolution analysis of the entire genome of cancer cells as well as more targeted analysis of coding sequencing or transcriptomes. Through international co-operations, a wide variety of cancer cell types have now been analyzed using these technologies to help unmask their pathogenesis. In this review, we briefly overview the recent advances in cancer research obtained through the massive effort of sequencing cancer genomes.

Key words: massively parallel sequencing – cancer genome – somatic mutation

INTRODUCTION

Cancer is caused by accumulation of genomic abnormalities (1). Therefore, comprehensive detection of various types of genetic changes, such as point mutations, copy number alterations and structural variations, are critically important for the understanding cancer biology, the improvement of diagnosis as well as the development of targeted therapies.

Cancer genomes are highly complex, which are composed of as many as 3 billion bases, including ~24 000–25 000 of protein-coding genes. The initial clue to accommodate this complexity of cancer genomes was obtained by the construction of the first draft sequence of the human genome in 2001 through the international collaborative effort under the human genome project (2). Since then, the sequence has been used as the reference human genome, based on which researches have been analyzing somatic mutations in cancers. However, the real breakthrough to facilitate the detection of cancer-specific gene mutations/alterations has been brought about by the recent development of massively parallel sequencing technologies; the 13 years required to read an initial human genome has been dramatically reduced to ~2 weeks with typical next-generation sequencing. This revolution of sequencing technologies envisages a new era of cancer genetics and biology. In this review, we overview the second-generation sequencing technologies and their application to cancer research.

ANALYSIS OF CANCER GENOMES USING SECOND-GENERATION SEQUENCING TECHNOLOGIES

A typical second-generation (also known as ‘next-generation’) sequencer goes through billions of fragmented DNAs (reads) simultaneously in a highly parallel way (massively parallel sequencing). Inevitably high sequencing errors in reading numerous bases are compensated by reading multiple (>30–40 times) DNA fragments for an accurate determination of somatic mutations and allelic status of polymorphisms in diploid sequences. For example, typically ~90 Gb of sequence or 30-fold coverage are obtained...
to call 99% of SNP alleles (3). The need of multiple reads is also realized by the fact that the mutations under interest may exist only in low allele frequencies in many situations; primary cancer samples are more or less ‘contaminated’ with normal tissues and there also exist multiple tumor subclones with different genetic alterations within a tumor, comprising intra-tumor heterogeneity (4–6). It may be necessary to deeply sequence to capture such low-frequency alleles depending on the apparent frequencies of the relevant alleles.

To further compound matters, pathogenesis of cancer is heterogeneous in terms of the spectrum of gene mutations even within the same histology type. Thus, sequencing multiple tumor samples may be necessary to understand the full spectrum of causative mutations depending on the extent of the heterogeneity, except for rare cases where the tumor is highly homogeneous and therefore only a few samples were sufficient to capture the relevant mutations. In fact, several ongoing projects are analyzing large numbers of cancer specimens; in the Cancer Genome Atlas (TCGA) (7) promoted by NIH/NCI in the USA and the International Cancer Genome Consortium (ICGC) participated in by institutes from multiple nations (8), more than 500 specimens for each of the major cancer types will be subjected to integrated genome, epigenome, transcriptome studies (9–14).

WHOLE-GENOME SEQUENCING

In 2008, the first sequencing of an entire cancer genome was reported, providing a comprehensive registry of gene mutations in a case with acute myeloid leukemia (AML) with a normal karyotype. In whole-genome sequencing, fragmented genomic DNA is directly subjected to massively parallel sequencing after minimum processing with several cycles of PCR amplification. Although simple in manipulation, this approach provides least biased, full-range information about genetic alterations at a single-nucleotide level (Fig. 1). Major caveats include the cost and time required for the analysis, compared with sequencing with lower complexity. It also consumes much more computer resources.

Because of its uniform coverage across all the genomic regions except for GC-rich sequences, whole-genome sequencing captured mutations that occur in both coding and non-coding regions, including promoters, enhancer, introns and non-coding RNAs as well as intergenic sequences. Not only detecting single-nucleotide substitutions, whole-genome sequencing can also identify a wide variety of structural abnormalities, such as chromosomal translocations, deletions and insertions, as well as copy number alterations, in a comprehensive manner. Especially, currently no other methods are available for detecting chromosomal translocations systematically, which would be underscored by the fact that a number of fusion genes caused by chromosomal translocations play a critical role in the cancer pathogenesis and are also exploited for contriving novel diagnostics and therapeutics (15,16). Moreover, whole-genome sequencing can reveal a more complex aspect of cancer genomes, in which tens to hundreds of genomic rearrangements involved localized genomic regions (‘chromothripsis’), which were thought to be acquired in an apparently one-off cellular crisis and reported to occur in 2–3% of all cancers and ~25% of bone cancers (Fig. 2) (17).

Finally, whole-genome sequencing is capable of detecting non-human sequencing derived from microbial genome by

---

**Figure 1.** Various types of genetic alterations that can be detected by whole-genome sequencing. Sequenced fragments (reads) are depicted as bars and paired-end reads connected with a dotted line are mapped to different genomic positions in the case of translocation. Copy number alterations can be detected by change in the number of sequence reads (depth) compared with a normal control.
analyzing sequences that are not mapped to human reference genome. Sung et al. performed whole-genome sequencing of 76 HBV-positive hepatocellular carcinomas and detected HBV integration in 76.4% of tumor samples. They also revealed that HBV integrations in \textit{TERT}, \textit{MLL4} and \textit{CCNE1} genes accounted for 40% of cases with HBV integrations \cite{18}. As the cost of sequencing continues to reduce, whole-genome sequencing is becoming a standard method of cancer genome sequencing feasible for many researchers.

**TARGETED/WHOLE-EXOME SEQUENCING**

Considering the current cost and throughput of sequencing, alternative to whole-genome sequencing would be focusing on particular sequences under interest, for example, entire coding sequences that account for <1.3% of the total genomes, which are now widely used for cancer genome sequencing (targeted sequencing). In whole-exome sequences, entire protein-coding sequences encompassing from 50 to 62 Mb are captured using liquid–solid phase hybridization to chemically synthesized nucleotides or ‘baits’ (Fig. 3), which were subjected to massively parallel sequencing. By limiting the targets to approximately one-hundredth of the entire genome, much higher sequence coverage ($\approx$ 100 times) can be achieved with considerably less sequence data compared with whole-genome sequencing, which far outperforms capillary sequencing-based approach employed in exome sequencing of glioblastoma \cite{19}.

Potential drawbacks in whole-exome sequencing are additional costs and procedures required for target capture and compromised sensitivity due to less uniform coverage and inability to detect fusion genes and other structural variations. Nevertheless, whole-exome sequencing provides a reasonable alternative to whole-genome sequencing, in terms of a high throughput of analyzing multiple samples at higher coverage and lower costs and the higher sensitivity of detecting mutations with lower allele frequencies, which may exist only in minor subclones or be associated with lower tumor burden, although it is challenging to detect somatic mutations in cases where germline control sample is contaminated with tumor tissues. In fact, whole-exome sequencing has been successfully applied to identify a number of novel targets of recurrent mutations \cite{10,20,24,25,26}.

**TRANSCRIPTOME SEQUENCING**

One of the major shortcomings in whole-exome sequencing is its inability to detect gene fusions. The deficit could be complemented in part by transcriptome sequencing (RNA sequencing, RNA-seq) at an acceptable cost (Fig. 4). In fact, transcriptome sequencing has been shown to provide a powerful method to detect biologically relevant fusion genes in that the detected fusion genes are actually expressed in tumor cells. The exact sequences of the expressed fusion transcripts can be directly determined. Steidl et al. performed RNA-seq of Hodgkin lymphoma cell lines and identified gene fusion involving MHC class II transactivator \textit{CIITA} \cite{25}, which were also found in primary mediastinal B-cell lymphoma (38%) and Hodgkin lymphoma samples (15%). Several algorithms are available for detecting fusion genes.

Figure 2. Schematic overview of chromothripsis and a circos plot of a chromothripsis-positive cancer. (A) Schematic overview of chromothripsis. (B) Chromothripsis detected in a case of clear cell renal carcinoma by whole-genome sequencing. Chromosomes (outside of the circular plot) and chromosomal rearrangements are shown as arcs connecting the two relevant genomic regions in the middle. Chromothripsis rearrangements are indicated by green arrows.
based on transcriptome sequencing, many of which are notorious for high false-positive rates even for recurrently detected candidates, which may necessitate intensive validation tasks using RT–PCRs.

RNA-seq can be used to detect somatic mutations and some important discovery of recurrent mutations in cancer has been made by RNA-seq, such as FOXL2 mutations in ovarian granulosa cell tumors (26). Although biologically relevant mutations tend to be enriched expressed sequences, coverage of sequencing could be highly biased depending on expression levels of different targets, leading to loss of power for detecting some mutations. This is in contrast to whole-genome/exome sequencing, with which relatively unbiased detection is expected. Difficulty in obtaining appropriate normal control tissues is another significant problem with transcriptome-based detection of mutations. RNA-seq also allows analysis of gene expression profiles in higher sensitivity. Novel transcripts or alternative splicing forms could be detected also, although computational analysis could be a challenging task. However, the higher representation of 3′ sequences could prevents unbiased estimation of relative expression levels, especially estimation of large transcripts, in which significant underrepresentation of the 5′ sequences occurs.

DEEP SEQUENCING OF SELECTED TARGETS

Intratumoral heterogeneity seems to be a common feature of cancers, providing a major source of tumor recurrence or resistance to anti-cancer therapies. It could be best evaluated by deep sequencing in combination with whole-genome/exome sequencing, where accurate allele frequencies of a set of mutations detected by genome/exome sequencing are
determined by reading at higher depth (typically >1000 or 10,000). Target DNA was amplified by individual PCRs or captured by high-throughput hybridization to the target baits. As low as 1% of minor alleles could be sensitively detected by target deep sequencing depending on experimental conditions and sequencing platforms, although background errors inherent to sequencing chemistries would prevent further detection of even lower allele frequencies. Saha et al. performed deep sequencing (median 20,000 times) of somatic mutations in triple-negative breast cancers detected by whole-exome sequencing, in which existence of subclones with different mutant allele frequencies were clearly demonstrated in most of cases (5). Walter et al. also showed the presence of subpopulation in secondary AML cases derived from myelodysplastic syndromes by whole-genome sequencing followed by target deep resequencing (average 640 times) of 304–872 candidates of somatic mutations using solid-phase target capture (6).

Another application of target deep sequencing is mutation screening involving a large number of genes and tumor samples. Actually, it becomes more and more unrealistic in detecting mutations of a number of large genes among hundreds of specimens using Sanger sequencing and plausible application of high-throughput sequencing. Deep sequencing also enabled a sensitive detection of minor mutant alleles with low frequencies together with an accurate determination of allele frequencies at the same time (10,20,27).

### SUMMARY AND FUTURE DIRECTIONS

Considering the ever-decreasing cost for sequencing, major mutational targets will be identified for common cancer types within a few years. However, many issues should be remained to be unanswered. For many newly identified mutations, their biological/clinical significance as well as their molecular mechanics in cancer development is unknown. There is an urgent need to achieving the clinical application of these technologies for diagnosis and treatment stratification, eventually leading to cure for cancer, although special attention should be paid before clinical use. For example, validation of the candidate mutations using alternative methods such as the conventional Sanger sequencing is indispensable at the current moment considering a relatively high error rate of second-generation sequencing. While understanding of the intratumoral heterogeneity is one of the critical issues in carcinogenesis, especially with regard to recurrence or resistance to anti-cancer therapies, higher resolution/highthroughput sequencing technologies, ideally of single-cell genomes, will be required to approach this issue. Promising news is the recent announcement of the coming newer generation sequencing platforms, including Ion torrent’s and Nanopore technologies (28,29). Roles of epigenetic abnormalities together with their interactions with gene mutations need to be clarified in further details, for which another technological breakthrough should be warranted.

### Funding

This work was supported by Grant-in-Aids from the Ministry of Health, Labor and Welfare of Japan and from the Ministry of Education, Culture, Sports, Science and Technology, and the Japan Society for the Promotion of Science (JSPS) through the ‘Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program),’ initiated by the Council for Science and Technology Policy (CSTP) and also by ‘project for development of innovative research on cancer therapies (p-direct).’

### Conflict of interest statement

None declared.

### References