Mini-review

Whole-exome sequencing reveals recurrent somatic mutation networks in cancer

Xiaoping Liu a,1, Jiguang Wang b,d,1, Luonan Chen a,c,*

a Key Laboratory of Systems Biology, SIBS-Novoc Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
b Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China
c Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan
d Department of Biomedical Informatics, Columbia University College of Physicians and Surgeons, New York, NY 10034, USA

A R T I C L E   I N F O

Keywords:
Whole-exome sequencing
Somatic mutation network
Cancer

A B S T R A C T

The second-generation sequencing technologies have been extensively used to reveal the mechanism of tumorigenesis and find critical genes in cancer progression that can be potential targets of clinic treatment. Exome is a part of genome formed by exons which are the protein-coding portions of genes. The whole-exome sequencing information can reflect the mutations of the protein-coding region in the genome and depict the causal relationship between the mutations and phenotypes. Now, many network-based methods have been developed to identify cancer driver modules or pathways, which not only provide new insights into molecular mechanism of disease progression at network level but also can avoid low coverage or lowly recurrent on disease samples in contrast to individual driver genes. In this review, we focus on the recent advances on network-based methods for identifying cancer driver modules or pathways, including methods of whole-exome sequencing, somatic mutation detection, driver mutation identification, and mutation network reconstruction.

© 2012 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Cancers are known medically as a malignant neoplasm, caused by the accumulation of genomic alterations or dysfunction of multiple biological processes. During the past decades, the high-throughput technologies, e.g. microarray and mass spectrometry, have provided a challenge to depict the mechanism of human cancer, and many methods have been developed to analyze and mine critical genes from microarray [1–6]. However, there are some defects for the microarray technology. For example, it is mainly used to identify the expression variation of known genes, rather than to search new genes; it can be used to find the difference of known SNP (Single Nucleotide Polymorphism), but it is not designed to detect new somatic mutations.

Clearly, genome sequencing technology can remedy the defects of microarray. Although the first-generation sequencing technology alleviated those problems of microarray, the high cost limited its application to the large-scale sequencing of multi-samples. Nevertheless, the second-generation sequencing technology allows the widespread use of massively parallel sequencing with lower cost and higher throughput. This technological advance is of significant importance for improving our understanding of the malignant neoplasm because cancer is basically considered as a disease of genome [7]. The data by the second-generation sequencing can not only be utilized to find the new genes and somatic mutations, but also reflect the expression level of every gene that cannot be measured by the first-generation sequencing technology. In particular, the genomic somatic alterations that present in cancer cell but not in germ line can lead to the development of cancer [7,8], so more and more discoveries about new cancer related genes [9–12] and regulation relationship [13–16] have been revealed based on the analysis of recurrent somatic mutation by the second-generation sequencing data.

The exome is an important functionally relevant part of genome formed by exons that are the coding portions of genes. On the other hand, the exome is the genetic blueprints of the synthesis of proteins and other functional products, which implies that the exome is more likely to contribute to phenotype of an organism. Since the exome only occupies approximately 1% regions of whole human genome or about 30 megabases (Mb) in length [17], the whole-exome sequencing can achieve higher sequence depth with less raw sequence and lower cost than whole-genome sequencing by using second-generation sequencing platforms. For example, 30-fold average coverage is achieved for the whole genome with
90 gigabases (Gb) of sequence, while only 3 Gb is required to obtain 75-fold average coverage of the exome [7,18]. Actually, the number of bases that can be sequenced for a given cost has increased one million-fold since 1990, and it is more than double increasing every year and twice as fast as Moore’s law for semiconductors [19].

As a result, many cancer mechanisms have been revealed by the whole-exome sequencing technology based on detecting the recurrent somatic mutation in cancer samples [4,11,13,14]. In the conventional methods, the “driver” mutation genes are expected to be identified by the frequency of recurrent somatic mutations (Fig. 1A), but it is difficult to distinguish the “driver” mutations from the “passenger” mutations which generally have no effects on the cancer development. To overcome this problem, recently, many network reconstruction based methods (Fig. 1B) have been developed to identify not “driver” gene but “driver” pathway, which is more meaningful to delineate the mechanism of cancer development and also has clear advantage to explore whole genome information from system viewpoint [20–23].

Decoding the cancer genomics is such a rapidly moving field, in this Letter we aim to briefly report recent advances and also important findings about the application of whole-exome sequencing technology in cancer research, focusing on recurrent somatic mutation networks. Following the general pipeline of mutation network detection, Section 2 firstly introduces important progresses in whole-exome sequencing techniques, and then Sections 3 and 4 describe and compare some popular methods to detect somatic mutations and driver mutations. Section 5 focuses on how to properly construct biologically meaningful mutation networks (or pathways) based on large batch of patient sequencing data and prior knowledge on biological systems.

2. Whole-exome sequencing methods

The development of massively parallel DNA enrichment technologies (genomic capture) has made it possible to capture the specific areas of human genome and sequence these targeted regions based on nucleic acid baits, including interesting genes and linkage regions, whole-exome, or whole-transcriptome. Most large-scale approaches for sequencing targeted regions utilize a variation of hybrid selection method, which uses the nucleic acid baits to capture the genes or regions of interest in the total DNA or RNA pool [24]. Recently, several approaches of them have been further extended to capture human exome.

2.1. Solid-phase capture method

Solid-phase capture method generally utilizes the bait probes to “fish” the sequences of interest by the principle of complementary base pairing, and the bait is usually affixed to a solid support (Fig. 2A), such as microarrays [25–28] or filters [29]. In these methods, the genomic DNA is firstly digested to short fragments, and then these fragments are used to hybridize with bait probes and then the desired fragments will be captured by these bait probes. The non-targeted fragments are subsequently washed away, and the desired fragments are enriched in solid support and then eluted for sequencing (Fig. 2A). The efficiency of these methods has been improved by multiple enrichment cycles [30], recently. Solid-based capture method is a time-saving and cost-effective method [25,26], and can effectively capture large scale of exome, e.g. a whole-exome of human (approximately 30 Mb) [18,31]. However, the shortcoming of this method includes the requirement of expensive hardware and relatively more DNA information than other methods [18,31]. Agilent, Roche/Nimblegen and Febit also offer the commercial kits implementing the solid-phase hybridization methods [24].

2.2. Liquid-phase capture method

Liquid-phase capture method is similar to solid-phase capture method. In this method, the bait probes are no longer fixed to a solid support, but combined with biotin (Fig. 2B). In this method, the bait probes are firstly marked by biotin, and then the biotinylated probes are mixed with digested DNA fragments. The bait probes will hybridize to the desired fragments in solution, and then the magnetic streptavidin beads are added to the solution and bind to the biotinylated probes. Subsequently,
the non-desired fragments are washed away and the desired DNA fragments which are bound to the bait probes are kept with bead–probe complexes. The desired DNA fragments are eluted from bait probes for sequencing [32] (Fig. 2B). Liquid-phase capture method is an effective method for target capture with low agent cost, much greater flexibility and less DNA requirement than solid-based method [18,32]. The shortcoming of this method is that the results only with small scale target sizes (≤3.5 Mb) have high uniformity and specificity [18,33]. The commercial kits are also available from Agilent and Roche/Nimblegen [24].

2.3. Polymerase capture methods

Another serial of methods to capture desired sequence are polymerase capture methods. Although polymerases take part in most methods to amplify captured fragments, in these methods, the polymerases are employed by a more integral way. Here, we only describe the basic mechanisms of two types of polymerase capture methods. The Molecular Inversion Probes (MIPs) technology has been developed by extending the padlock probe technology [34]. In this method, the genomic DNA is not digested into fragments and the bait probes can bind to the flanks of desired regions in genomic DNA (Fig. 2C). The two arms of the probes are designed to complement the flanks of desired regions in genome. After the hybridization between probe’s arms and desired regions, the polymerase and ligase are used to fill the gap between the two arms and form closed circular DNA. Subsequently, the genomic DNA is digested by exonucleases and only the closed circular products with the desired sequence can be left [35,36] (Fig. 2C). The advantages of MIP technology include low DNA requirement, DNA quality less important and more dynamic range [37,38]. Due to two complement arms for every probe, the limitations of this technology include the low uniformity for capture targets and the high cost with covering large target sets [33]. Primer extension capture (PEC) is sim-

Fig. 2. Diagrammatic sketch for the target sequence capture. (A) Solid-phase capture method. The bait probes are colored by the light blue shadow and the desired sequences are indicated by the red bar. (B) Liquid-phase capture method. The bait probes are colored by the light blue shadow; yellow stars represent biotin; the black points are used to represent the streptavidin beads and the desired sequences are the red bar. (C) MIP method. The red bar is the desired sequence; the yellow oval represents the polymerase; the blue shadows are the arms of probe and the red shadow is complementary sequence of the desired sequence. (D) PEC method. The bait probes are colored by the light blue shadow; yellow stars represent biotin; yellow ovals are the polymerase; the black points are used to represent the streptavidin beads and the desired sequences are the red bar. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
ilar to liquid-phase capture method and this method uses biotinylated probes to capture the DNA of interest (Fig. 2D). In this method, the digested DNA fragments are mixed with bait probes and the desired fragments hybridize to the bait probes in solution. The biotinylated probes act as primers to begin elongation with DNA polymerases and create tighter interactions with desired fragments. After the biotin binds to the streptavidin beads, the non-targeted fragments will be washed away. The desired DNA fragments are eluted and then sequenced [39] (Fig. 2D). PEC method is quick, sensitive and specific. However, the sensitivity of capture decreases, when the numbers of bait probes in a multiplex capture reaction increase [40]. Hence, PEC method is not suitable to capture large regions, but is fitted to capture small target regions or SNP position from many individuals in a rapid fashion [40].

3. Somatic mutation detection

A somatic mutation is an alteration in DNA which occurs in a somatic cell after conception. Somatic mutations cannot be inherited because they occur in any of the cells of the body except the germ cells. Cancer related somatic mutations can be detected by measuring the alterations between tumor and normal samples. At present, germline mutation analyses are more mature and most of tools which are used to detect somatic mutations were developed for germline mutations, e.g. SAMtools [41], GATK [42] and VarScan [43]. However, somatic mutation detection is more complex than germline mutation because cancer samples vary in their purity and ploidy [7]. In general, a clinical cancer sample usually contains a mixture of tumor and normal cells, and therefore, it is a mixture of cancer and normal genomes (and exomes). Furthermore, the cancers themselves may be highly heterogeneous and contain different clones that have different genomes [44]. Recently, some methods have been developed to identify the ratio between tumor and normal cells in a tumor sample [45].

Primary methods to detect the somatic mutations are based on directly comparing the difference between tumor sample's sequence and reference sequence, counting the number of variant reads and calculating the ratio of variant reads in every position of a tumor sample, e.g. Maq [46], Bowtie [47] and VarScan [43]. In these methods, if the ratio of variant reads in a position is greater than a threshold, the position would be considered as a somatic mutation in this tumor sample. These methods can simply and efficiently detect the somatic mutations from tumor samples, but they usually lose some information by discarding the reads of mapping to multiple locations or low identity [43]. Some improved methods based on statistical model can solve the problem of information lost, e.g. SNVMix [48], JointSNVMix [49], SomaticSniper [50] and VarScan 2 [51].

Cancer genomes vary substantially in their sequences and structures compared to normal genomes and among themselves, and every cancer genome is abnormal in its own way. Therefore, the ploidy of allelic fractions in tumor genome may be not an integer. For example, in germline analysis, most mutations have an allelic fraction of either 1/2 for heterozygous events or 1 for homozygous events. By contrast, the fraction of a somatic mutation may differ significantly from 1/2, leading the somatic mutations from germline variants [52,53]. SNVMix is the first report of a method specific for tumor sample's somatic mutation detection [7.48] by implementing a probabilistic Binomial mixture model and the corresponding tool package has been developed [48]. But SNVMix only uses the tumor sample's sequence to compare with public reference sequence, and does not consider about the germline variant of normal samples. In fact, somatic mutations are relatively rare compared with germline variants [52,53], so it is necessary to distinguish the somatic mutations from germline variants. JointSNVMix is an improved tool for SNVMix with a novel model to detect the somatic mutations between tumor and normal sequencing data [49], and it can distinguish the somatic mutations from germline variants.

SomaticSniper is another method to identify somatic mutations from germline variants based on a Bayesian comparison of the genotype likelihoods between the tumor and normal by sequencing data [50] and can efficiently detect the potential somatic point mutations from sequencing samples. VarScan 2 is a new method which was designed to discover the somatic mutations in tumor by comparing with tumor and normal sequences based on exome sequencing data [51], and it comprehensively considers about the difference between exome and genome sequencing, e.g. difference depth and coverage.

Cancer is a heterogeneous disease which has cellular differences within a single neoplasm [54]. The cancer cells may be highly heterogeneous and contain different genotypes which have different genomes [44]. Generally, the somatic mutation detection software may give different results, depending on different proportion of various genomes, so effective heterogeneity detection can improve the accuracy of somatic mutation detection methods. Actually, a heterogeneity detection method called “ABSOLUTE” has been proposed to carry out such detection on the proportion of subclonal heterogeneity [45], which demonstrated the effectiveness by considering this factor.

4. “Driver” mutation identification

To depict the cancer mechanism, whole-genome or whole-exome measures of somatic mutations in large numbers of cancer genomes have been done by the second-generation sequencing technologies. A key challenge for these researches is to distinguish “driver” mutations that lead to cancer development from “passenger” mutations which are functionally neutral and do not contribute to tumorigenesis [55]. A common method for identifying driver mutations is to find recurrent mutations (or recurrently mutated genes) with significant frequency in a large cohort of cancer samples. This method has found many important mutations for cancer, e.g. KRAS, BRAF and ERBB2, but it cannot identify all of the driver mutations in individual cancers. Because many driver genes are mutated at low frequency in these tumors, it is difficult or impossible to distinguish them from “passenger” mutations only based on the frequency. Other methods, such as machine learning methods that are trained to classify and predict the “driver” somatic mutations based on prior known cancer-causing mutations, do not require the information of the recurrent mutations in cancer samples [56,57]. It has been proved by recent sequencing studies that the complexity and diversity of “driver” mutations at the gene level can be substantially reduced at the pathway level, because cancer genes tend to come into a limited set of essential biological pathways [58,59]. Hence, new methods based on the integrative pathway or network analysis are developed to identify both “driver” mutations (genes) and potential “driver” pathway by exploring network-related information.

5. Mutation network reconstruction

Initially, highly and recurrently cancer genes such as P53, Myc, PTEN and IDH1 were discovered and considered as key regulators of cancer or “driver” genes, but many biological evidences demonstrate that a complex disease is generally not resulted from the mutations of single molecules. Actually, a biological function is generally carried out not by individual genes but by their pathways or networks. Instead of focusing on highly recurrent genes, identifying highly potential “driver” pathways or subnetworks that can reduce the complexity and diversity of identifying “driver” mutations [58,59], becomes more important and effective with the rapid accumulation of exome-sequencing data. Many methods have been developed to identify the causal genes or critical pathways...
Most of the network reconstruction methods based on prior knowledge mainly focus on the identification of the significant modules with mutation-enriched genes. They usually use the existing molecular network or functional annotation information, and try to extract the significant pathways or subnetworks which are enriched to the “driver” mutations or mutation genes. Boca et al. utilized the known functional gene sets or pathways to construct a score principle, and scored each gene in every gene set by combining the mutation data [60]. The score of every gene is defined by the total number of mutations across all tumor samples compared to the “passenger” mutation rate, and gene sets are identified as the critical gene sets whose scores are significantly higher than those predicted by chance [55]. Cerami et al. developed another method to identify the core pathway in cancer by second-generation data [20]. They used the protein–protein interactions network and the signal transduction pathways to merge them into a new human interaction network, and then mapped all mutation genes to the new human interaction network. The shortest paths connecting all mutation genes were extracted from the interaction network and genes which are not mutated or do not appear at the shortest paths were eliminated from network to form a subnetwork. Subsequently, the network modules were identified from the subnetwork as candidate “driver” (Fig. 3A) [20]. Torkamani and Schork reconstructed the network by the gene co-expression and then found that the distinct co-expression modules contain a large number of mutation genes than expected by chance [21]. It is likely that the co-expression modules bear an excess of somatic mutations to alter the signaling pathways which are known to be important for tumorigenesis [21].

5.2. Mutual exclusivity principle

Mutual exclusivity principle is the gene mutation or alterations that affect the same pathway not tending to co-occur in the same patient [22]. In the exome or genome sequencing, it is found that the non-synonymous somatic mutations cannot simultaneously appear at two genes in the same important pathway of cancer development. One of explanations for mutual exclusivity principle is that once a mutation alters a pathway, the selective advantage incurred by a second mutation in the same pathway is minimal. And large scale sequencing studies about cancer proved an additional support for the mutual exclusivity principle [61,62]. The mutual exclusivity is an important principle for depicting the mechanism of tumorigenesis and many methods utilized this principle to develop new schemes to identify the potential cancer causal genes or cancer related pathway or subnetwork.

5.2.1. Exacting subnetwork based on mutual exclusivity principle

The methods which are based on mutual exclusivity principle are different from the other methods which are not based on mutual exclusivity principle to identify critical modules in cancer. The mutual exclusivity based methods usually require the genes in the module to follow the mutual exclusivity principle. In other words, only one mutation gene is allowed to exist in the module in one sample. Now, many novel methods use the mutual exclusivity principle to identify the functional modules for cancer development. For example, Vandin et al. developed software called Den-
drix to search the causal module or pathway in cancer. Dendrix identifies the module based on two basic rules: The first is high coverage that means that most samples have at least one mutation gene in the module or gene set. The second is high exclusivity that means that nearly all samples have no more than one mutation in the module or gene set [63]. Other method is similar to Dendrix by Ciriello et al., and it lists three principles: The first, member genes are recurrently altered across a set of tumor samples. The second, member genes are known to participate in the same biological process. The third, alteration events within the modules are mutually exclusive. The significant recurrent mutation genes were determined in every sample and mapped to the human reference network to form a subnetwork. The modules are extracted by searching the genes which are involved in the same functional process in the subnetwork (Fig. 3B) [22].

5.2.2. Constructing network based on mutual exclusivity principle

The methods which are based on the mutual exclusivity principle usually choose the critical pathway or subnetwork by using the existing interactions between genes in publically available interaction network. But the mutual exclusivity principle can also be used to construct the interaction network. Miller et al. developed a method to identify functional modules by recurrent and mutually exclusive mutational patterns in tumors [23]. In this method, the gene network was firstly constructed by the mutual exclusivity principle, and the exclusivity score between each pair of genes was calculated by the number of samples where exactly one of the pair is mutated divided by the number of samples where at least one of the pair is mutated. Then a gene exclusivity network can be constructed based on the exclusivity score which can act as the weight of edge between every pair of genes. In the gene exclusivity network, the mutual exclusivity modules were extracted and used to the following statistical analysis (Fig. 3C) [23].

6. Discussion and conclusion

Some approaches described in this review can also be applied to whole genome sequencing, but most of the prior knowledge based approaches cannot be directly used for such a purpose due to the lack of prior knowledge hided in “Junk DNA”. Furthermore, in contrast to whole genome sequencing, the main advantage of whole-exome sequencing in cancer research field is that exome sequencing can test more samples and provide higher sequencing coverage and depth by same cost. Generally, many methods are in common used for both exome sequencing and genome sequencing.

The second-generation sequencing technologies have obviously promoted the process of research to elucidate cancer mechanism. In other words, they can capture and sequence selectively the portion of genome we wanted, e.g. whole-exome sequencing technology is able to capture only the protein-coding parts of whole-genome, thereby significantly cutting the cost for cancer search. On the other hand, “Driver” mutation identification aims to find the causal mutation of cancer development, but the frequency of recurrent mutation on cancer samples is very low. Hence, network-based analysis becomes a promising tool to detect the critical modules or pathways [51-53] which are critical for cancer initiation and progression. Although many existing network-based methods can be used to the analysis of exome sequencing data, as described in this review, the methods based on mutual exclusivity principle provide a different new aspect to network reconstruction of recurrent somatic mutations. As a future topic, it is an important direction to integrate expression data (gene expression, protein expression, methylation, metabolic data, etc.) [64,65], in particular, time-course information [66] for revealing dynamical mechanism [67] of disease development and progression, e.g. detecting network biomarkers [64,65], dynamical network biomarkers [66] of complex diseases, and the leading networks [67] of critical transitions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants Nos. 91029301, 61134013 and 61072149; and by the Chief Scientist Program of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences (CASs) with Grant No. 2009CSP002; by Shanghai Pujiang Program. This research was also partially supported by National Center for Mathematics and Interdisciplinary Sciences of CAS; and by the FIRST program from JSPS initiated by CSTP.

References


