BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 4. 506–519
Advance Access published on 9 August 2012
doi:10.1093/bib/bbs044

Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives

Qingguo Wang, Junfeng Xia, Peilin Jia, William Pao and Zhongming Zhao

Submitted: 12th April 2012; Received (in revised form): 24th June 2012

Abstract

Gene fusions are important genomic events in human cancer because their fusion gene products can drive the development of cancer and thus are potential prognostic tools or therapeutic targets in anti-cancer treatment. Major advancements have been made in computational approaches for fusion gene discovery over the past 3 years due to improvements and widespread applications of high-throughput next generation sequencing (NGS) technologies. To identify fusions from NGS data, existing methods typically leverage the strengths of both sequencing technologies and computational strategies. In this article, we review the NGS and computational features of existing methods for fusion gene detection and suggest directions for future development.

Keywords: *gene fusion; next generation sequencing; cancer; whole genome sequencing; transcriptome sequencing; computational tools*

INTRODUCTION

The past two decades have witnessed extensive research on human genomic aberrations which are believed to be causal factors of a variety of diseases. Among the most widely studied variations, gene fusions have been of great interest due to their associations with tumorigenesis. A fusion gene, also called a chimeric gene or a hybrid gene, is the juxtaposition of two otherwise separate genes. A canonical example of fusion genes is *BCR-ABL*, the transcript of which is translated into an abnormal tyrosine kinase that drives the development of chronic myelogenous leukemia (CML) [1–3]. A *BCR-ABL* targeting drug, Gleevec/Glivec, has been proven very successful in the treatment of CML [4], prompting the search for other fusion genes to be used as tumor-specific biomarkers or drug targets.

Active research on genomic aberrations and gene fusions has been fueled by the advent and widespread applications of high-throughput next generation sequencing (NGS) technologies, the applications of which have revealed complex landscapes of human cancer [5–9]. NGS accelerates nucleotide acquisition by sequencing tens to hundreds of millions of

Corresponding author. Zhongming Zhao, 2525 West End Avenue, Suite 600, Nashville, TN 37203, USA. Tel: +615-343-9158; Fax: +615-936-8545; E-mail: zhongming.zhao@vanderbilt.edu

Qingguo Wang is a postdoctoral researcher in Dr Zhongming Zhao's group at Vanderbilt University, USA, working mostly on next generation sequencing data and other genomic data.

Junfeng Xia is a postdoctoral researcher in Dr Zhongming Zhao's group at Vanderbilt University, USA, working mostly on next generation sequencing data and other genomic data.

Peilin Jia is a postdoctoral researcher in Dr Zhongming Zhao's group at Vanderbilt University, USA. Her main interest is integrative genomics, next generation sequencing data analysis and method development.

William Pao is a Professor of Medicine, Cancer Biology, & Pathology/Microbiology/Immunology, Director of the Personalized Cancer Medicine Initiative and Director of the Division of Hematology and Oncology at Vanderbilt University. His laboratory focuses on identifying actionable genetic alterations in cancers.

Zhongming Zhao is an Associate Professor at Vanderbilt University and the Chief Bioinformatics Officer at Vanderbilt-Ingram Cancer Center. His research activity focuses on bioinformatics and systems biology in complex disease.

sequence targets simultaneously and thus allows for comprehensive genome-wide analysis at a low cost [10–15]. Additionally, NGS generates digital output, for example, the counts of each transcript in RNA sequencing data rather than the quantitative estimate of signals from microarray gene expression data. The digital nature of NGS gives rise to data of higher resolution that enables mutation analysis to the base-pair level. Another strength in NGS applications is its support for a wide array of applications to human genetic research [16, 17]. Based on the type of input materials, the mainstream applications of NGS can be classified into: whole genome sequencing (WGS), whole exome sequencing (WES) and whole transcriptome sequencing (RNA-Seq), enabling different levels of researches. Besides whole genome and whole transcriptome sequencing, NGS can also be applied to a subset of genes (targeted sequencing). Moreover, the versatility of sequencing platforms, e.g. Illumina Genome Analyzer, Life Technologies SOLiD, Roche 454, as well as the recently debuted Illumina HiSeq and MiSeq systems, provides rich options for biological researches. For an in-depth study of the NGS technologies, interested readers are referred to several recent reviews [10–16, 18].

With the rapid advances of the NGS technologies, the cost of sequencing has dramatically decreased over the past few years and has made the sequencing of human genomes routine at the genome sequencing centers and core facilities in institutes. Advances in sequencing technologies have also stimulated software development. A number of new software tools have quickly emerged to identify structural variants (SVs), as well as gene fusions resulting from these variants. In the year 2011 alone, at least eight new computational tools for the characterization of fusion genes were published in major scientific journals [19–26]. These new methods have led to important discoveries [27–31], e.g. disease-defining fusion *WWTR1-CAMTA1* in 'epithelioid hemangioendothelioma' [28] and a novel case of prostate cancer with unique biological features [31].

To provide guidelines for the rapidly growing number of fusion gene studies through NGS and to foster the development of new algorithms, in this article we present a review of existing computational methods for human gene fusion detection from NGS data, primarily in cancer genomes. The article summarizes the common features of existing methods and discusses important issues that an algorithm needs to consider while dealing with NGS data. The capability, limitations and future directions of this new field are also explored.

ADVANCES IN GENE FUSION DETECTION

A fusion gene can form as a consequence of a SV, which is typically defined as large variation in structure of human genome. The types of SVs that may result in gene fusions include large insertions, deletions, inversions and in particular translocations. For instance, the fusion *BCR-ABL1* is created by a characteristic interchromosomal translocation (termed 'Philadelphia chromosome') that brings together the 5′ part of the *BCR* gene on chromosome 22 and the 3′ part of the *ABL1* gene on chromosome 9 [32]. Genes involved in a fusion can come from the same chromosome too. For example, *TMPRSS2-ERG*, the most common genetic alteration in prostate cancer, is an intrachromosomal fusion of two genes on chromosome 21 [33, 34].

Major advancements have been made in the discovery of fusion genes. As of 21 May 2012, as many as 839 fusion genes have been documented in the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (http://cgap.nci.nih.gov/Chromosomes/Mitelman) [32]. These fusion genes enhance our understanding of the origins and progression factors of cancer. More significantly, a number of fusions have been recognized as important prognostic tools or therapeutic targets in anti-cancer treatments. In a recent study [35], a new urine test was invented to detect the presence of *TMPRSS2-ERG* to indicate the risk of prostate cancer.

The prevalence of known fusion genes in different cancers varies greatly [32]. For the aforementioned two fusions, *BCR-ABL1* is expressed in >90% of patients with CML while *TMPRSS2-ERG* is found in >50% of individuals with prostate cancer [36]. However, the majority of recurrent fusion genes are prevalent at low frequency in patients. For instance, the newly discovered fusion *KIF5B-RET* is estimated in 1–2% of lung adenocarcinomas [37, 38].

The feasibility of applying NGS to detect fusion genes in cancer genomes was first evaluated by Campbell *et al.* [6], who produced sequencing reads from lung cancer cell lines and identified two fusion transcripts in addition to other genomic rearrangements. Then, Ley *et al.* [39] applied WGS to cancer

Table 1: Recent next generation sequencing (NGS) studies that discovered novel gene fusions in human cancers

Cancer	Sequencing technology	# novel fusions detected[a]	Example fusion event(s)	Ref.
Acute myeloid leukemia (AML)	WGS	3	*bcr3 PML-RARA, LOXL1-PML*, and *RARA-LOXL1*	[44]
Breast	RNA-Seq	13	*SEC16A-NOTCH1*	[45]
Breast	RNA-Seq	2	*WWC1-ADRBK2, ADNP-C20orf132*	[30]
Breast	RNA-Seq	24	*VAPB-IKZF3*	[46]
Breast[b]	RNA-Seq	5	*ARHGAP19-DRG1*	[47]
Breast[a]	RNA-Seq	7	*PDCD1LG2-C18orf10*	[41]
Chronic myeloid leukemia (CML)[b]	RNA-Seq	1	*NUP214-XKR3*	[47]
Colorectal	RNA-Seq and targeted sequencing	1	*C2orf44-ALK*	[37]
Colorectal	WGS	1	*VTI1A-TCF7L2*	[48]
Epithelioid hemangioendothelioma	RNA-Seq	1	*WWTR1-CAMTA1*	[28]
Gastric cancer	RNA-Seq	1	*AGTRAP-BRAF*	[49]
Hodgkin lymphoma[b]	RNA-Seq	3	*CIITA-BX648577*	[29]
Hepatocellular carcinoma	WGS	4	*BCORL1-ELF4, CTNND1-STX5, VCL-ADK, CABP2-LOC645332*	[50]
Leukemia	Targeted sequencing	11	*RUNX1–KCNMA1*	[51]
Lung	RNA-Seq	1	*ALK-PTPN3*	[52]
Lung	RNA-Seq and targeted sequencing	1	*KIF5B-RET*	[37]
Lung	RNA-Seq	1	*KIF5B-RET*	[38]
Lung[b]	RNA-Seq	1	*R3HDM2-NFE2*	[53]
Lung[b]	WGS	1	*PVT1–CHD7*	[8]
Melanoma	RNA-Seq	11	*RB1-ITM2B*	[54]
Ovary	RNA-Seq and targeted sequencing	1	*ESRRA-C11orf20*	[55]
Prostate	WGS and RNA-Seq	15	*C15orf21-MYC*	[31]
Prostate	RNA-Seq	8	*MSMB-NCOA4*	[56]
Prostate	RNA-Seq	7	*ALG5-PIGU*	[27]
Prostate	RNA-Seq	3	*SLC45A3-BRAF, ESRP1-RAF1*	[49]
Prostate	RNA-Seq	13	*SLC45A3–ELK4*	[40]
Prostate[b]	RNA-Seq	8	*TIA1-DIRC2, ZDHHC7-ABCB9*	[47]
T lymphoblastic lymphoma (T-ALL) and associated myeloproliferative neoplasm	Targeted sequencing	1	*C6orf204-PDGFRB*	[42]
Therapy-related acute myeloid leukemia	WGS	2	*DGKG-BST1, BST1-DGKG*	[57]

NGS, next generation sequencing; WGS, whole genome sequencing; RNA-Seq, whole transcriptome sequencing. [a]Only validated fusions were counted. [b]Only cell line samples were sequenced and analyzed

patients and established WGS as an unbiased method to study human genomic mutations. Later, breakthroughs made by Maher *et al.* [40] and Zhao *et al.* [41] demonstrated the utility of RNA-Seq in the detection of fusion genes. To explore the power of RNA-Seq in fusion gene discovery, Maher *et al.* [40] not only pinpointed known fusions such as *BCR-ABL1* and *TMPRSS2-ERG* from RNA-Seq data of tumors and cancer cell lines, but also found novel fusions that were subsequently validated experimentally. These pioneer works stimulated NGS applications to cancer research and, accordingly, led to substantial expansion of the realm of fusion gene discovery. Table 1 summarizes recent NGS studies that have resulted in the discovery of novel fusion oncogenes. Most of these studies used an RNA-Seq platform, but detection of gene fusions at the genomic DNA level using the WGS [8] or targeted sequencing approaches [42, 43] is also effective.

Advances in fusion gene detection from NGS data are largely attributed to the development of new computational tools. To meet the challenges of NGS data analysis, which deals with millions of short reads that makes alignment to reference genome difficult and error prone, a large number of computational methods have been developed over the past several years. A summary of these tools is available in multiple review articles

Table 2: Computational tools for gene fusion detection using NGS data

Method	URL	Feature	Ref.
Fusion detection specific			
BreakFusion	http://bioinformatics.mdanderson.org/main/BreakFusion	Identifying gene fusions from paired-end RNA-Seq data	[65]
ChimeraScan	http://code.google.com/p/chimerascan/	Detecting fusion transcripts from RNA-Seq data	[24]
Comrad	http://code.google.com/p/comrad/	Using both RNA-Seq and WGS data to detect genomic rearrangements and aberrant transcripts	[23]
FusionAnalyser	http://www.ilte-cml.org/FusionAnalyser/	Detecting gene fusions from paired-end RNA-Seq data	[64]
deFuse	http://sourceforge.net/apps/mediawiki/defuse/	Identifying gene fusions from RNA-Seq data	[22]
FusionMap	http://www.omicsoft.com/fusionmap/	Using either WGS or RNA-Seq data to detect fusion genes	[20]
FusionHunter	http://bioen-compbio.bioen.illinois.edu/FusionHunter/	Detecting fusion transcripts from RNA-Seq data	[21]
FusionSeq	http://archive.gersteinlab.org/proj/rnaseq/fusionseq/	Identifying fusion transcript from RNA-Seq data	[63]
ShortFuse	https://bitbucket.org/mckinsel/shortfuse	Identifying fusion transcripts from RNA-Seq data	[26]
SnowShoes-FTD	http://mayoresearch.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm	Detecting fusion transcripts from RNA-Seq data	[25]
SOAPfusion[a]	http://soap.genomics.org.cn/SOAPfusion.html	Part of the software SOAP, for genome-wide detection of gene fusions from RNA-Seq data	[66]
TopHat-Fusion	http://tophat-fusion.sourceforge.net/	An enhanced version of TopHat, for detection of fusion transcripts from RNA-Seq data	[19]
Structural variant detection			
BreakDancer	http://genome.wustl.edu/software/	Detecting structural variations from paired-end WGS data	[67]
CREST	http://www.stjuderesearch.org/site/lab/zhang	Identifying structural variations from paired-end WGS data	[68]
2003GASV	http://code.google.com/p/gasv/	Software for identifying structural variations	[69]
HYDRA	http://code.google.com/p/hydra-sv/	Detecting SVs in both unique and duplicated genomic regions	[70]
PEMer	http://sv.gersteinlab.org/pemer/download.html	Using paired-end NGS data to detect structural variation	[71]
R453PlusIToolbox	http://www.bioconductor.org/packages/2.10/bioc/html/R453PlusIToolbox.html	An R/Bioconductor package for the analysis of Roche 454 sequencing data	[72]
SVDetect	http://svdetect.sourceforge.net/Site/Home.html	Detecting structural variations from paired-end/mate pair data	[73]
VariationHunter	http://compbio.cs.sfu.ca/strvar.htm	Identifying structural variations from paired-end WGS data	[74, 75]
Others[b]			
R-SAP	http://www.mcdonaldlab.biology.gatech.edu/r-sap.htm	A parallel method to estimate RNA expression level and to detect gene fusions from RNA-Seq data	[76]
Trans-ABySS[c]	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss	*De novo* assembly of RNA-Seq reads	[77]
Trinity	http://trinityrnaseq.sourceforge.net/	*De novo* assembly of RNA-Seq without using a reference	[78]

NGS, next generation sequencing; WGS, whole genome sequencing; RNA-Seq, whole transcriptome sequencing; SV, structural variation. [a]SOAPfusion is an unpublished web-downloadable method. [b]Methods in this category are designed for the general purpose of genetic alteration detection. Gene fusion identification is only a small part of their pipelines. [c]Fusion gene detection using Trans-ABySS is described in its user manual (http://www.bcgsc.ca/downloads/trans-abyss/data/trans-abyss-manual.v1.2.0.doc.pdf), not mentioned in its published paper [77].

[16, 58–62], none of which, however, covers new methods for fusion gene discovery. Along with the progress of computational technologies, software designed specifically to detect gene fusions from NGS data has emerged [63] and this new field is expanding rapidly due to the pressing need to decode the complexity of cancer and other genomes. As shown in Table 2, in the year 2011 alone, at least eight new methods for the characterization of fusion genes were published in major scientific journals [19–26]. Unlike earlier works that used long or the combination of long and short single-end reads for fusion event identification [40, 41], these new methods leverage the strengths of high-throughput short paired-end reads to achieve better accuracy as well as efficiency.

Table 2 also lists eight SV-detecting tools, e.g. BreakDancer [67] and CREST [68], the purpose of which is mainly to provide accurate and comprehensive predictions of SVs in genomes. We include them here because fusion genes are potential products of SVs. The characterization of SVs that result in fusions is an important step in fusion gene calling. For example, BreakFusion [65] utilizes BreakDancer [67] to locate splicing breakpoints, while more commonly fusion-detecting methods develop algorithms targeted specifically at fusion-causing SVs. In comparison with the methods for gene fusion detection that emerged only lately, SV identification is a well-studied field with some methods, such as BreakDancer [67], having been reviewed by a large number of articles [11, 59, 79, 80]. Hence, in the following text, we will not touch on these SV-detecting tools. Instead, we will focus our discussion on the new methods in Table 2 that aim specifically for fusion gene detection.

ISSUES AFFECTING GENE FUSION DETECTION

The ability of an approach to identify fusions from NGS data relies on the types of sequencing data it aims to work on as well as its computational strategies to process the data. In this section, we review these NGS and computational issues. An in-depth discussion of the computational features of existing methods is the topic of the next section.

WGS, RNA-Seq and targeted sequencing

WGS and RNA-Seq are two major NGS technologies for fusion gene detection (Table 1). As the most powerful sequencing technology today, WGS provides the most comprehensive and unbiased characterization of genomic alterations in genomes, especially cancer genomes. Using WGS technology, a variety of fusion genes have been discovered [8, 50, 57], some of which, for example, *VTI1A-TCF7L2*, are believed important for the growth of certain cancer cells [48]. One drawback of WGS, however, is that it requires a great amount of sequencing and intensive computational analysis. The whole process of WGS, from sample preparation to fusion identification and verification, may take months to complete [44]. While we have seen the cost of NGS decreased dramatically during the past few years, it is still expensive compared to RNA-Seq [81]. Finally, the significance of a fusion gene discovered using WGS relies on its effects on expression and on whether it produces fusion transcripts.

Compared to WGS, RNA-Seq only sequences the regions of the genome that are transcribed and spliced into mature mRNA, which is ~2% of the entire genome [63]. Another advantage that makes RNA-Seq ideal for the discovery of expressed fusion genes is that it allows for detection of multiple alternative splice variants resulting from a fusion event. These distinct features of RNA-Seq, together with its low cost and quick turnaround time, make RNA-Seq very popular in fusion gene studies. Over the past 3 years, 21 out of 29 studies found novel onco-fusions through RNA-Seq, in contrast with only 5 out of 29 through WGS (Table 1). However, one main limitation of RNA-Seq is that it cannot detect fusion events involving non-transcribed regions [19]. Other factors that complicate RNA-Seq data analysis are tissue-specificity and the broad dynamic range of expression in the human transcriptome [82]. The dynamic transcription of genes in different tissues and cellular stages makes detection of gene fusions more complicated, especially when transcript expression is low. Moreover, it is a challenge to differentiate fusions of interest from artifacts due to the prevalence of gene readthrough events.

Table 3 summarizes features of fusion gene detection tools that have implemented the methods listed in Table 2. Among these tools, the software FusionMap [20] works with either WGS or RNA-Seq data, while the majority of tools focus on RNA-Seq only, primarily due to the aforementioned strengths of RNA-Seq over WGS and investigators' interest in known gene regions. Different from other software that analyzes one type of sequencing data at a time, Comrad [23] requires both WGS and RNA-Seq data produced for the same sample as input, aiming to leverage the advantages of both sequencing technologies. Comrad [23] does not nominate a fusion event unless both WGS and RNA-Seq reads support the evident. This integrative approach reduces false positive detection of fusions that are plaguing existing RNA-Seq methods and has been shown to be powerful in a recent study of prostate cancer [31]. However, the combination of WGS and RNA-Seq increases cost and computational time. It also limits the applications of Comrad, as investigators may not have samples for both platforms in one study. Furthermore, if the false negative rate is high in any platform (WGS or RNA-Seq), this

Table 3: Features of computational tools for fusion gene detection

Method	Input data				Reference[f]		Fusion junction detection[g]		Assembly[h]
	Type[d]		Format[e]						
	WGS	RNA-Seq	Single-end	Paired-end	Transcriptome	Genome	Split-read	Spanning-read	
Fusion detection specific									
BreakFusion[a]		•		•	•	•			•
ChimeraScan		•		•	•	•	•	•	
Comrad[b]	•	•		•	•	•	•	•	
FusionAnalyser[a]		•		•	•	•	•	•	
deFuse		•		•	•	•	•	•	
FusionMap	•	•	•	•	•	•	•		
FusionHunter		•		•	•	•	•	•	
FusionSeq		•		•	•	•	•	•	
ShortFuse		•		•	•	•	•	•	
SnowShoes-FTD		•		•	•	•	•	•	
SOAPfusion		•		•	•	•	•	•	
Tophat-Fusion		•	•	•	•	•	•		
Structural variant detection									
BreakDancer[c]	•			•		•	•	•	
CREST	•			•		•	•		
GASV	•			•		•		•	
HYDRA	•			•		•		•	
PEMer	•			•		•		•	
R453PlusIToolbox	•		•	•		•	•		
SVDetect	•			•		•		•	
VariationHunter	•			•		•		•	
Others									
R-SAP		•	•	•	•	•	•	•	
Trans-ABySS		•		•			•	•	•
Trinity		•		•			•	•	•

[a]The input of FusionAnalyser and BreakFusion are alignment files (typically BAM format). [b]Comrad requires both WGS and RNA-Seq data of the same sample for fusion gene characterization. [c]Although BreakDancer is applied in BreakFusion to find splicing breakpoints from RNA-Seq data, it is designed for detection of genomic structural variation. [d]Column 'Type' is explained in detail in 'WGS, RNA-Seq, and Targeted Sequencing' section. [e]Column 'Format' corresponds to 'Single-End Versus Paired-End' section. [f]Column 'Reference' is explained in 'Reference Sequences' section. [g]Column 'Fusion Junction Detection' is explained in 'Single-End Versus Paired-End' and 'Procedure For Detection of Gene Fusions from NGS Data' sections. [h]Column 'Assembly' corresponds to 'Mapping-First Versus Assembly-First' section.

strategy would miss the chance to identify true fusion genes.

Finally, besides whole genome and whole transcriptome, targeted sequencing is also effective for the detection of gene fusions [42, 43]. For example, we took advantage of an observation that many breakpoints occur upstream of a conserved GXGXXG kinase motif and targeted the region upstream of the exon containing the motif for DNA capture. Our approach successfully identified several tyrosine kinase fusions in cancer cell lines [43].

Reference sequences

To detect SVs that may result in gene fusions, each read (pair) needs to be aligned to a reference genome sequence in order to determine its genomic location. Currently, the reference genome used in existing methods is NCBI build 37/36 (http://www.ncbi.nlm.nih.gov/projects/mapview/) or UCSC hg19/hg18 (http://hgdownload.cse.ucsc.edu/downloads.html#human). One limitation with the use of reference genome is that fusion genes involving novel sequences that are not represented in the reference will be missed.

For RNA-Seq data, in addition to alignment to the reference genome, reads are also mapped to a transcriptome library so that the genes involved in each fusion can be identified. The reference transcriptome used widely in the community includes the UCSC annotations of known genes [83] and the Ensembl human gene model RefSeq [84]. One potential problem with those methods relying on a known transcriptome library is that they consider only the candidates involving annotated exons. The fusion genes with novel exons cannot be detected.

To alleviate the reliance on known transcriptomes, FusionHunter [21] proposes to map sequencing reads to the human reference genome and then to identify putative exons by clustering reads mapped to the same genomic regions. However, in its current implementation, FusionHunter still needs a transcriptome library to make its results more reliable.

Single-end versus paired-end

The earlier RNA-Seq studies used single-end data to detect fusion genes [40, 41]. We call a read that harbors a fusion junction a 'split read'. By analyzing the alignments of 'split reads', which do not map to the reference sequences directly, gene fusion events can be characterized [19, 20].

If both ends of a set of long DNA/cDNA fragments are sequenced, the resulting paired short reads are called paired-end (or mate-pair). We call a pair of reads that harbor a fusion junction within its insert sequence a 'spanning pair' (or 'spanning reads'). When two ends of a 'spanning pair' are aligned to two different genes, a discordant mapping is produced. The discordant mappings of paired-end reads are therefore characteristic of fusion genes. In the landmark work by Maher *et al.* [47], discordantly mapped reads were used to detect fusions from paired-end RNA-Seq data. FusionSeq [63], the first publically accessible method for fusion gene detection, also exploits discordantly mapped reads for fusion gene identification. As the most influential method in this field, FusionSeq has contributed to the discovery of multiple novel fusion genes, e.g. *ALG5-PIGU* in prostate cancer [27] and *WWTR1-CAMTA1* in 'epithelioid hemangioendothelioma' [28]. It is also a starting point of several fusion-detecting tools including FusionAnalyzer [64] in Table 2. Besides FusionSeq, FusionAnalyzer and all other software in Table 3 works with paired-end reads and two of them, FusionMap [20] and TopHat-Fusion [19], can work on single-end data.

Using both 'spanning' and 'split reads' to identify fusion candidates, Maher *et al.* [47] and Ha *et al.* [30] reported improved sensitivity if paired-end data is used. However, if only 'split reads' are utilized to identify fusion junctions, it was shown that the ability to characterize fusion genes using single-end reads is as good as with paired reads [19, 20].

Mapping-first versus assembly-first

Based on the computational strategies for fusion gene detection, the methods in Table 3 can be grouped into two categories, mapping-first [22] and assembly-first [78]. The mapping-first approach first aligns reads to reference DNA/RNA sequences and then finds fusion breakpoints from the resulting alignment patterns [22]. All the software customized for fusion gene characterization (i.e. 'Fusion detection specific') in Table 3 falls into this category. Compared to the assembly-first approach, the mapping-first approach is faster and has dominated the field of NGS-based gene fusion studies.

The second computational strategy, the assembly-first approach, first assembles reads that overlap. The long sequences assembled, i.e. contigs, are then mapped to reference sequences for structure alteration identification. For an assembly algorithm, e.g. Trans-ABySS [77] or Trinity [78], if it assembles short reads directly without mapping them to the references, then it is called *de novo* assembly. The exclusive advantage of *de novo* assembly is that it does not need a reference genome/transcriptome for fusion detection (by comparing directly the assembled sequences of the sample with the assembled ones of control). Its disadvantage is that the assembly of short sequences is too time-consuming and too error prone to enable this approach for practical biomedical applications. None of the studies in Table 1 employed short-read assembly to identify fusion genes.

The assembly approach can be combined with reference alignment. For instance, the aforementioned method BreakFusion [65] performs targeted/localized *de novo* assembly after short read mapping so as to balance sensitivity, specificity and computational efficiency of fusion gene calling.

Other issues

Additional issues on gene fusion detection include sequencing coverage, size of the insert sequences, read length, rate of sequencing error, as well as the length and location of the fusion gene to be discovered. Overall, in the case of paired-end sequencing, the increase in sequencing coverage, read length or the decrease in sequencing error improves the probability of detecting fusion genes [85]. For an in-depth exploration into these features, interested readers are referred to reference [85].

PROCEDURE FOR DETECTION OF GENE FUSIONS FROM NGS DATA

In this section, we take a closer look at the methods designed specifically for gene fusion identification.

Table 4: Mapping tools and parameters used in methods for fusion gene detection[a]

Method	Mapping tool	Intra-chromosomal distance cutoff *D* (kb)[b]	*Split-read* mapping		# supporting reads[c]	Features of scoring function
			# segments	Length (bp) of end segments/ seeds		
BreakFusion	BLAT					A function of BLAT alignment scores
ChimeraScan	Bowtie		Various	25	3	
Comrad	Bowtie and BLAT			25	1 WGS + 5 RNA-Seq	
FusionAnalyser	BWA		2	Various		A scoring algorithm that ranks fusion candidates based on their coverage and annotation status
deFuse	Bowtie		≥2	Various	5 spanning + 3 split	A classifier trained using an Adaboost algorithm on 11 features
FusionMap	Bowtie and GSNAP	5	2	Various	1	A score based on read density and mapping quality
FusionHunter	Bowtie	600	2	Various	2 spanning + 1 split	
FusionSeq	Eland/Bowtie and BLAT				various	A score of the normalized and expected number of supportive reads
ShortFuse	Bowtie	10		22	1/20 × coverage	A score based on mapping quality and insert size distribution
SnowShoes-FTD	BWA	100		32	10 spanning + 2 split	
SOAPfusion					6	
TopHat-Fusion	Bowtie	100	3	25	Various	A weighted sum of read coverage, number of supporting reads, etc.

[a]Only the software in Table 2 that is designed specifically for fusion gene detection is compared in this table. [b]*D* is defined as the minimally allowed distance (kb) between genes that form intra-chromosomal fusion. [c]Default/minimum number of supporting reads, which can be *split reads*, *spanning reads*, supportive WGS reads and/or RNA-Seq reads.

These methods overall follow a three-step procedure to detect gene fusions: (i) mapping and filtering, (ii) fusion junction detection and (iii) fusion gene assembly and selection. The major computational features of these methods are provided in Table 4.

Mapping and filtering

All the methods in Table 4 take mapping as their initial analysis step. As the most important step in the mapping-first approach, mapping is a well-investigated problem in computational biology [60, 61, 86]. However, different from NGS applications in other fields, e.g. SNP and indel callings, the methods in Table 4 exploit primarily unmapped reads and/or discordantly mapped reads, the analysis of which is much more challenging than those of the mapped ones.

Table 4 lists the mapping tools utilized in existing fusion detection software. Among them, Bowtie [87] is applied in most methods to short read alignments. The popularity of Bowtie in the field is ascribed to its speed and ability to find all possible mapping loci for each read pair. In the broader areas of RNA-Seq applications, however, ELAND, which is shown more accurate than Bowtie [88], is used most widely for mapping [89]. ELAND is rarely used for fusion detection methods primarily because ELAND is a commercial product and not free to the research community.

After mapping, the alignment of each read (pair) is evaluated and the reads unrelated to fusions are removed from further consideration. The methods that are based primarily on 'split reads', like Tophat-Fusion [19] and FusionMap [20], filter out all mapped reads. However, for methods that exploit

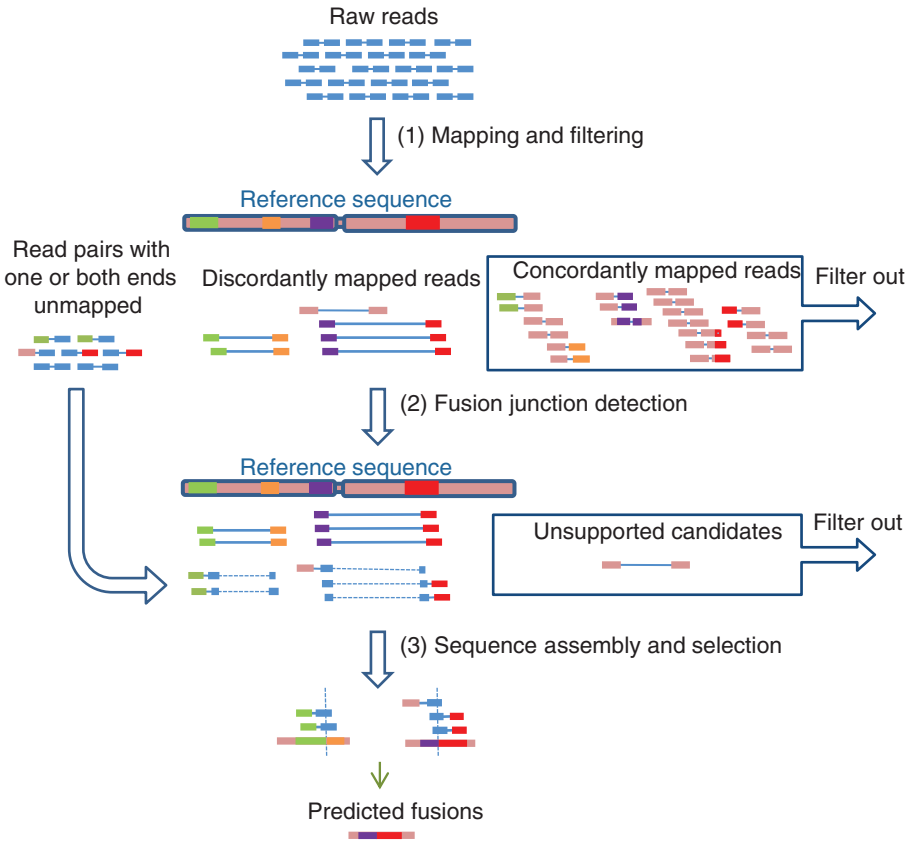

Figure 1: A procedure to detect gene fusions from paired-end NGS data through 'split read' mapping. (1) Reads are mapped to reference sequences. Those mapped concordantly are discarded. To differentiate mapped reads in the figure from unmapped ones, which remain blue throughout the pipeline, the color of mapped reads is changed from blue to the color of reference sequence region they are aligned to. (2) Reads mapped discordantly are used to infer approximate fusion boundaries. For each pair of reads with at least one end unaligned to reference sequence, the unmapped end is cut into multiple segments (two in this case) to be aligned independently to the approximate fusion boundaries. In the figure, two segments from the same read are connected by a dashed line. Fusion junctions are identified through 'split-reads', whose two segments are mapped to two different chromosomes/genes. (3) The sequences of the fusion candidates are assembled and the ones with the highest likelihood to be real fusions are selected as final outputs. In the figure, we use vertical dotted lines to represent fusion breakpoints.

'spanning reads', such as SnowShoes-FTD [25], which is a powerful analytic pipeline for identification of fusion transcripts in breast cancer and for assessment of tumor subtype-specific distribution in primary tumors [90], all discordantly mapped pairs are preserved. In addition to discordantly mapped reads, these methods also keep unmapped reads (potentially 'split reads') in order to assist in the selection of fusion candidates [24–26, 63]. Figure 1 illustrates a procedure for identification of gene fusions through paired-end reads.

To further discard reads that are less likely to harbor fusions, all the methods in Table 4 developed filtering techniques. For instance, the software FusionSeq [63] eliminates spurious fusion candidates through more than 10 filters, e.g. sequence similarity filter, repetitive regions filter, abnormal insert size filter, ribosomal filters, etc. In essence, the two steps discussed below are filtering techniques too, considering the removal of fusion candidates as a consequence of these two steps. One commonly used filter that may be worth mentioning relates to intra-chromosomal fusions. Two neighboring genes on the same chromosome may produce a read-through transcription. To differentiate true intra-chromosomal rearrangements from read-through events, a common practice is to define a threshold D. A fusion candidate is discarded if the distance between its fusion partner genes is smaller than D. An appropriate value of D is important for fusion gene calling in that more read-through events are preserved if a smaller D value is used,

whereas true intra-chromosomal fusions may be discarded with a greater D value. The default values of D in existing methods are provided in Table 4.

Fusion junction detection

Step 2 of the procedure in Figure 1 illustrates the detection of fusion junctions through 'split read' mapping [19–22, 64]. The unmapped reads are cut into multiple pieces. The first and last segments of each 'split read' are then mapped against the reference sequences independently. Partitioning of the reads increases the chance of the reads to be aligned to the references. If the two end segments of a 'split read' are mapped to two different chromosomes or genes, then the read is potentially from a fusion gene. Once this alignment pattern is detected, the precise location of the fusion junction can then be found by adjusting the boundaries of the original fragments and performing realignment.

One factor that influences 'split read' mapping is the length of the partitioned segments. Shorter segments improve sensitivity for nominating fusions but increase false positive rate. To balance sensitivity and false positive rates, these methods either split a read evenly into two segments if the read is not long [21, 64], or simply use a fixed length (e.g. 25 bp) of end segments, as shown in Table 4. For example, TopHat-Fusion [19] splits an 80-bp read into three segments with lengths 25, 30 and 25 bp, respectively. The two end segments, which are 25 bp in length, are remapped to the reference sequences.

A different strategy to detect fusion junctions is to infer fusion breakpoints from 'spanning reads' and then select those predictions that are likely to be real using 'split reads' [24–26, 63]. Discordant alignments are first grouped into clusters, each consisting of a maximal set of reads that share the same pair of breakpoints. Then, the boundary region of each candidate fusion junction is identified from its cluster. Next, fusion junction loci are inferred and putative fusion transcripts are predicted. Finally, unmapped reads are aligned to the predicted fusion transcripts. The predictions, to which the highest number of unmapped reads is aligned, are nominated as candidate fusion genes.

Fusion gene assembly and selection

After identifying fusion junctions, the sequence of each candidate fusion gene can be derived by stitching directly two partner genes together. Then, the reads unmapped previously are aligned to the candidate fusion genes. Reads mapped in this step provide additional confidence in the candidate. Hence, they are called supporting reads. Besides 'split reads', 'spanning reads' can also serve as supporting evidence, as they encompass fusion junctions in their insert sequences. Existing methods all require the presence of supporting reads as a prerequisite to nominate a fusion, although the required number of supports varies greatly, as illustrated in Table 4. Note that the requirement for supporting reads should be adjusted based on the scale of the sequencing data. Theoretically, the limit on the number of supporting reads should be increased for larger data sets, such as those generated by the Illumina HiSeq system.

The requirement for more supporting reads removes more inauthentic candidates, however, by risking discarding true fusion genes of low transcription level or coverage. To help distinguish true fusions from candidates of uneven expression/coverage, it is common practice in existing methods to develop scoring functions to rank fusion candidates [19, 20, 22, 26, 63]. The candidates with the maximum likelihood to be real fusions are selected as final outputs. As shown in Table 4, these scoring functions are mostly based on features including mapping quality, number of supporting reads and read depths. The scores are either derived analytically and empirically (e.g. FusionSeq [63]), or learned from known data using machine learning techniques (e.g. the AdaBoost classifier in deFuse [22], which, with improved accuracy over FusionSeq, was applied to detection of fusion genes in various studies [29, 91]).

PERSPECTIVES

The active development of algorithms to identify driver fusion genes in human cancer has resulted in a variety of software, the sensitivity and specificity of which are subject to today's NGS platforms, sequencing protocols, mapping tools, fine tuning of filtering parameters, as well as other issues as discussed in this work. Even with a wise combination of the most up-to-date technologies, the capability of existing methods to detect fusions in cancer genomes still needs to be improved. As one example, one of the latest software, SnowShoes-FTD [25], identified five novel fusions in the breast cancer cell line MCF7 but failed to characterize several fusion genes that were previously discovered by Maher *et al.* [47].

To increase the accuracy of fusion gene characterization, one direction is to improve mapping quality by adopting mapping tools that are more accurate than what are commonly used currently. Due to the critical role of short-read mapping in fusion gene detection, it is expected that more accurate mapping will improve the performance of fusion gene calling. Another direction is to develop new *de novo* assembly algorithms that, unlike today's mainstream methods, do not rely on reference sequences and, hence, are potentially more sensitive in fusion detection.

Improvements in NGS sequencing throughput, data accuracy and read lengths will continue at an unprecedented pace. For example, the HiSeq 2000 sequencing system (http://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf) unveiled recently by Illumina is capable of generating 25 Gb of data per day, a 5-fold increase in data generation rate from its previous version (GAIIx). Although the identification of fusions of low expression level from transcriptome sequencing still remains a challenge despite higher coverage data, deeper coverage overall will give rise to higher sensitivity in fusion gene detection. Therefore, we expect more interesting fusion genes to be identified from differentially expressed systems. Further, increased read length will reduce the ambiguity of short-read mapping and, thus, lower the false positive rate of fusion candidate callings.

Another prominent development is the development of third generation sequencing (TGS) technologies, which promise to provide dramatically longer read lengths, shorter times for data generation and lower costs than NGS [17, 92]. Representative platforms of TGS of today include Helicos, Pacific Biosciences and Ion Torrent (acquired by Life Technologies Corporation in 2010). Long reads will dramatically simplify aberration analysis algorithms. If the TGS reaches a throughput and error rate comparable to that of the NGS technologies of today, it is expected that the techniques for fusion gene detection will greatly accelerate the studies of human cancer and other cellular systems (e.g. somatic mutations in tissues).

Key points

- Research on gene fusions in human cancer has been greatly accelerated due to next generation sequencing technologies and the development of detection algorithms and software tools.
- The sensitivity and specificity of existing methods for gene fusion discovery are subject to issues such as today's NGS platforms, sequencing protocols, mapping tools and fine tuning of filtering parameters.
- New sequencing technologies, especially the emerging third generation sequencing (TGS) technologies, will improve fusion gene calling.

ACKNOWLEDGEMENTS

We would like to thank Dr Jingchun Sun and Dr Min Zhao in the Bioinformatics and Systems Medicine Laboratory for valuable suggestions and discussion on this work.

FUNDING

This work was supported by the Stand Up To Cancer-American Association for Cancer Research Innovative Research Grant [SU2C-AACR-IRG0109] and National Institutes of Health Grants [P30CA68485, R01LM011177]. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Bartram CR, Deklein A, Hagemeijer A, *et al*. Translocation of c-abl oncogene correlates with the presence of a philadelphia-chromosome in chronic myelocytic-leukemia. *Nature* 1983;**306**:277–80.
2. Deklein A, Vankessel AG, Grosveld G, *et al*. A cellular oncogene is translocated to the philadelphia-chromosome in chronic myelocytic-leukemia. *Nature* 1982;**300**:765–67.
3. Lugo TG, Pendergast AM, Muller AJ, *et al*. Tyrosine kinase-activity and transformation potency of bcr-abl oncogene products. *Science* 1990;**247**:1079–82.
4. Capdeville R, Buchdunger E, Zimmermann J, *et al*. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nat Rev Drug Discov* 2002;**1**:493–502.
5. Berger MF, Lawrence MS, Demichelis F, *et al*. The genomic complexity of primary human prostate cancer. *Nature* 2011; **470**:214–20.
6. Campbell PJ, Stephens PJ, Pleasance ED, *et al*. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 2008;**40**:722–29.
7. Stephens PJ, McBride DJ, Lin ML, *et al*. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 2009;**462**:1005–10.
8. Pleasance ED, Stephens PJ, O'Meara S, *et al*. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 2010;**463**:184–90.
9. Pleasance ED, Cheetham RK, Stephens PJ, *et al*. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010;**463**:191–96.
10. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;**11**:31–46.

11. Koboldt DC, Ding L, Mardis ER, *et al*. Challenges of sequencing human genomes. *Brief Bioinform* 2010;**11**: 484–98.
12. Robison K. Application of second-generation sequencing to cancer genomics. *Brief Bioinform* 2010;**11**:524–34.
13. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genom Hum G* 2008;**9**:387–402.
14. Shendure J, Ji HL. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45.
15. Ansorge WJ. Next-generation DNA sequencing techniques. *New Biotechnol* 2009;**25**:195–203.
16. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 2010;**11**:685–96.
17. Glenn TC. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 2011;**11**:759–69.
18. Fuller CW, Middendorf LR, Benner SA, *et al*. The challenges of sequencing by synthesis. *Nat Biotechnol* 2009;**27**: 1013–23.
19. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* 2011;**12**:R72.
20. Ge HY, Liu KJ, Juan T, *et al*. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* 2011;**27**:1922–28.
21. Li Y, Chien J, Smith DI, *et al*. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics* 2011;**27**:1708–10.
22. McPherson A, Hormozdiari F, Zayed A, *et al*. deFuse: An algorithm for gene fusion discovery in tumor RNA-Seq data. *Plos Comput Biol* 2011;**7**:e1001138.
23. McPherson A, Wu C, Hajirasouliha I, *et al*. Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. *Bioinformatics* 2011;**27**:1481–8.
24. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* 2011;**27**:2903–4.
25. Asmann YW, Hossain A, Necela BM, *et al*. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. *Nucleic Acids Res* 2011;**39**:e100.
26. Kinsella M, Harismendy O, Nakano M, *et al*. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics* 2011;**27**:1068–75.
27. Pflueger D, Terry S, Sboner A, *et al*. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Res* 2011;**21**:56–67.
28. Tanas MR, Sboner A, Oliveira AM, *et al*. Identification of a disease-defining gene fusion in Epithelioid Hemangioendothelioma. *Sci Transl Med* 2011;**3**:98ra82.
29. Steidl C, Shah SP, Woolcock BW, *et al*. MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature* 2011;**471**:377–81.
30. Ha K, Lalonde E, Li L, *et al*. Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. *BMC Med Genomics* 2011;**4**:75.
31. Wu C, Wyatt AW, Lapuk AV, *et al*. Integrated genome and transcriptome sequencing identifies a novel form of hybrid and aggressive prostate cancer. *J Pathol* 2012;**227**: 53–61.
32. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* 2007;**7**:233–45.
33. Tomlins SA, Laxman B, Dhanasekaran SM, *et al*. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* 2007;**448**: 595–99.
34. Tomlins SA, Rhodes DR, Perner S, *et al*. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005;**310**:644–48.
35. Tomlins SA, Aubin SMJ, Siddiqui J, *et al*. Urine TMPRSS2:ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA. *Sci Transl Med* 2011;**3**: 94ra72.
36. Nam RK, Sugar L, Yang W, *et al*. Expression of the TMPRSS2: ERG fusion gene predicts cancer recurrence after surgery for localised prostate cancer. *Brit J Cancer* 2007;**97**:1690–95.
37. Lipson D, Capelletti M, Yelensky R, *et al*. Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. *Nat Med* 2012;**18**:382–84.
38. Kohno T, Ichikawa H, Totoki Y, *et al*. KIF5B-RET fusions in lung adenocarcinoma. *Nat Med* 2012;**18**:375–77.
39. Ley TJ, Mardis ER, Ding L, *et al*. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008;**456**:66–72.
40. Maher CA, Kumar-Sinha C, Cao XH, *et al*. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009; **458**:97–101.
41. Zhao Q, Caballero OL, Levy S, *et al*. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc Natl Acad Sci USA* 2009;**106**:1886–91.
42. Chmielecki J, Peifer M, Viale A, *et al*. Systematic screen for tyrosine kinase rearrangements identifies a novel C6orf204-PDGFRB fusion in a patient with recurrent T-ALL and an associated myeloproliferative neoplasm. *Genes Chromosomes Cancer* 2012;**51**:54–65.
43. Chmielecki J, Peifer M, Jia P, *et al*. Targeted next-generation sequencing of DNA regions proximal to a conserved GXGXXG signaling motif enables systematic discovery of tyrosine kinase fusions in cancer. *Nucleic Acids Res* 2010;**38**:6985–96.
44. Welch JS, Westervelt P, Ding L, *et al*. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *Jama* 2011;**305**:1577–84.
45. Robinson DR, Kalyana-Sundaram S, Wu YM, *et al*. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat Med* 2011;**17**:1646–51.
46. Edgren H, Murumagi A, Kangaspeska S, *et al*. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol* 2011;**12**:R6.
47. Maher CA, Palanisamy N, Brenner JC, *et al*. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci USA* 2009;**106**:12353–58.
48. Bass AJ, Lawrence MS, Brace LE, *et al*. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* 2011;**43**:964–68.
49. Palanisamy N, Ateeq B, Kalyana-Sundaram S, *et al*. Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nat Med* 2010;**16**: 793–98.

50. Totoki Y, Tatsuno K, Yamamoto S, *et al*. High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet* 2011;**43**:464–69.
51. Grossmann V, Kohlmann A, Klein HU, *et al*. Targeted next-generation sequencing detects point mutations, insertions, deletions and balanced chromosomal rearrangements as well as identifies novel leukemia-specific fusion genes in a single procedure. *Leukemia* 2011;**25**:671–80.
52. Jung Y, Kim P, Keum J, *et al*. Discovery of ALK-PTPN3 gene fusion from human non-small cell lung carcinoma cell line using next generation RNA sequencing. *Genes Chromosomes Cancer* 2012;**51**:590–97.
53. Wang XS, Prensner JR, Chen GA, *et al*. An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nat Biotechnol* 2009;**27**:1005–11.
54. Berger MF, Levin JZ, Vijayendran K, *et al*. Integrative analysis of the melanoma transcriptome. *Genome Res* 2010;**20**: 413–27.
55. Salzman J, Marinelli RJ, Wang PL, *et al*. ESRRA-C11orf20 Is a recurrent gene fusion in serous ovarian carcinoma. *Plos Biol* 2011;**9**:e1001156.
56. Nacu S, Yuan WL, Kan ZY, *et al*. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *Bmc Med Genomics* 2011;**4**:11.
57. Link DC, Schuettpelz LG, Shen D, *et al*. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. *Jama* 2011;**305**:1568–76.
58. Bateman A, Quackenbush J. Bioinformatics for next generation sequencing. *Bioinformatics* 2009;**25**:429–29.
59. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 2009;**6**:S13–20.
60. Bao SY, Jiang R, Kwan WK, *et al*. Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet* 2011;**56**:406–14.
61. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* 2010;**11**: 473–83.
62. Xia J, Wang Q, Jia P, *et al*. NGS catalog: a database of next generation sequencing studies in humans. *Hum Mutat* 2012; **33**:e2341–55.
63. Sboner A, Habegger L, Pflueger D, *et al*. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol* 2010;**11**: R104.
64. Piazza R, Pirola A, Spinelli R, *et al*. FusionAnalyser: a new graphical, event-driven tool for fusion rearrangements discovery. *Nucleic Acids Res* 2012. May 8 (doi:10.1093/nar/gks394; epub ahead of print).
65. Chen K, Wallis JW, Kandoth C, *et al*. BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics* 2012;**28**:1923–24.
66. Short Oligonucleotide Analysis Package (SOAP). http://soap.genomics.org.cn/SOAPfusion.html (9 April 2012, date last accessed).
67. Chen K, Wallis JW, McLellan MD, *et al*. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 2009;**6**:677–81.
68. Wang J, Mullighan CG, Easton J, *et al*. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* 2011;**8**:652–54.
69. Sindi S, Helman E, Bashir A, *et al*. A geometric approach for classification and comparison of structural variants. *Bioinformatics* 2009;**25**:i222–30.
70. Quinlan AR, Clark RA, Sokolova S, *et al*. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* 2010;**20**:623–35.
71. Korbel JO, Abyzov A, Mu XJ, *et al*. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 2009;**10**:R23.
72. Klein HU, Bartenhagen C, Kohlmann A, *et al*. R453Plus1 Toolbox: an R/Bioconductor package for analyzing Roche 454 Sequencing data. *Bioinformatics* 2011;**27**:1162–3.
73. Zeitouni B, Boeva V, Janoueix-Lerosey I, *et al*. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* 2010;**26**:1895–6.
74. Hormozdiari F, Alkan C, Eichler EE, *et al*. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* 2009;**19**: 1270–8.
75. Hormozdiari F, Hajirasouliha I, Dao P, *et al*. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 2010;**26**: i350–7.
76. Mittal VK, McDonald JF. R-SAP: a multi-threading computational pipeline for the characterization of high-throughput RNA-sequencing data. *Nucleic Acids Res* 2012; **40**:e67.
77. Robertson G, Schein J, Chiu R, *et al*. De novo assembly and analysis of RNA-seq data. *Nat Methods* 2010;**7**:909–12.
78. Grabherr MG, Haas BJ, Yassour M, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644–52.
79. Xi R, Kim TM, Park PJ. Detecting structural variations in the human genome using next generation sequencing. *Brief Funct Genomics* 2010;**9**:405–15.
80. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;**12**: 363–76.
81. Sboner A, Mu XJ, Greenbaum D, *et al*. The real cost of sequencing: higher than you think!. *Genome Biol* 2011; **12**:125.
82. Taylor BS, Ladanyi M. Clinical cancer genomics: how soon is now? *J Pathol* 2011;**223**:318–26.
83. Hsu F, Kent WJ, Clawson H, *et al*. The UCSC known genes. *Bioinformatics* 2006;**22**:1036–46.
84. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007;**35**:D61–5.
85. Bashir A, Volik S, Collins C, *et al*. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *Plos Comput Biol* 2008;**4**:e1000051.
86. Horner DS, Pavesi G, Castrignano T, *et al*. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform* 2010;**11**: 181–97.

87. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.

88. Smith E. Data Analysis with CASAVA v1.8 and the MiSeq Reporter. http://www.illumina.com/Documents/seminars/presentations/2011_09_smith.pdf (9 April 2012, date last accessed).

89. Grant GR, Farkas MH, Pizarro AD, *et al*. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* 2011;**27**:2518–28.

90. Asmann YW, Necela BM, Kalari KR, *et al*. Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer Res* 2012;**72**:1921–8.

91. Lee CH, Ou WB, Marino-Enriquez A, *et al*. 14-3-3 fusion oncogenes in high-grade endometrial stromal sarcoma. *Proc Natl Acad Sci USA* 2012;**109**:929–34.

92. Schadt EE, Turner S, Kasarskis A. A window into third generation sequencing. *Hum Mol Genet* 2010;**19**:R227–40.