

REVIEW

Open Access

# Next generation sequencing in cancer research and clinical application

Derek Shyr<sup>1</sup> and Qi Liu<sup>2,3\*</sup>

## Abstract

The wide application of next-generation sequencing (NGS), mainly through whole genome, exome and transcriptome sequencing, provides a high-resolution and global view of the cancer genome. Coupled with powerful bioinformatics tools, NGS promises to revolutionize cancer research, diagnosis and therapy. In this paper, we review the recent advances in NGS-based cancer genomic research as well as clinical application, summarize the current integrative oncogenomic projects, resources and computational algorithms, and discuss the challenge and future directions in the research and clinical application of cancer genomic sequencing.

**Keywords:** Next generation sequencing, Cancer research, Clinical application

## Introduction

Sanger sequencing has dominated the genomic research for the past two decades and achieved a number of significant accomplishments including the completion of human genome sequence, which made the identification of single gene disorders and the detection of targeted somatic mutation for clinical molecular diagnostics possible [1,2]. Despite Sanger sequencing's accomplishments, researchers are demanding for faster and more economical sequencing, which has led to the emergence of "next-generation" sequencing technologies (NGS). NGS's ability to produce an enormous volume of data at a low price [3,4] has allowed researchers to characterize the molecular landscape of diverse cancer types and has led to dramatic advances in cancer genomic studies.

The application of NGS, mainly through whole-genome (WGS) and whole-exome technologies (WES), has produced an explosion in the context and complexity of cancer genomic alterations, including point mutations, small insertions or deletions, copy number alternations and structural variations. By comparing these alterations to matched normal samples, researchers have been able to distinguish two categories of variants: somatic and germ line. The Whole transcriptome approach (RNA-Seq) can not only quantify

gene expression profiles, but also detect alternative splicing, RNA editing and fusion transcripts. In addition, epigenetic alterations, DNA methylation change and histone modifications can be studied using other sequencing approaches including Bisulfite-Seq and ChIP-seq. The combination of these NGS technologies provides a high-resolution and global view of the cancer genome. Using powerful bioinformatics tools, researchers aim to decipher the huge amount of data to improve our understanding of cancer biology and to develop personalized treatment strategy. Figure 1 shows the workflow of integrating omics data in cancer research and clinical application.

## Cancer research

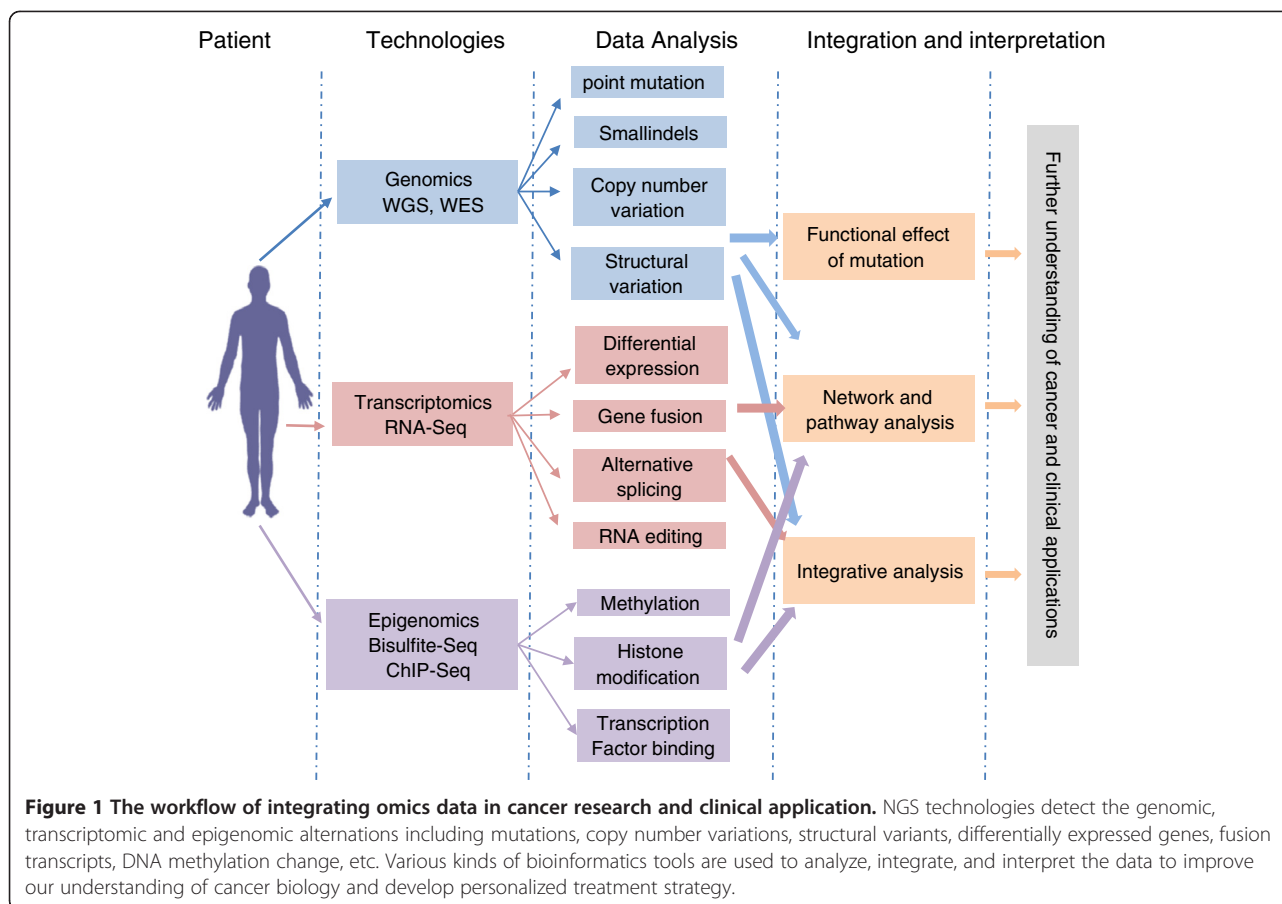
In the last several years, many NGS-based studies have been carried out to provide a comprehensive molecular characterization of cancers, to identify novel genetic alterations contributing to oncogenesis, cancer progression and metastasis, and to study tumor complexity, heterogeneity and evolution. These efforts have yielded significant achievements for breast cancer [5-12], ovarian cancer [13], colorectal cancer [14,15], lung cancer [16], liver cancer [17], kidney cancer [18], head and neck cancer [19], melanoma [20], acute myeloid leukemia (AML) [21,22], etc. Table 1 summarizes the recent advances in cancer genomics research applying NGS technologies.

\* Correspondence: qiliu@vanderbilt.edu

<sup>2</sup>Center for Quantitative Sciences, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

<sup>3</sup>Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

Full list of author information is available at the end of the article



### Discovery of new cancer-related genes

Cancer is primarily caused by the accumulation of genetic alterations, which may be inherited in the germ line or acquired somatically during a cell's life cycle. The effects of these alterations in oncogenes, tumor suppressor genes or DNA repair genes, allows cells to escape growth and regulatory control mechanisms, leading to the development of a tumor [23]. The progeny of the cancer cell may also undergo further mutations, resulting in clonal expansion [24]. As clonal expansion continues, clones eventually become invasive to its surrounding tissue and metastasize to distant areas from the primary tumor [25].

The sequencing of cancer genomes has revealed a number of novel cancer-related genes, especially in breast cancer. Recently, six papers reported their findings on large breast cancer dataset: TCGA performed exome sequencing on 510 samples from 507 patients [5], Banerji et al. conducted exome sequencing on 103 samples and whole genome sequencing on 17 samples, Ellis et al. did exome sequencing on 31 samples and whole genome sequencing on 46 samples [7], Stephens et al. applied exome sequencing on 100 samples, Shah et al. performed whole genome/exome and RNA sequencing on 65 and 80 samples of triple-negative breast cancers [11],

and Nik-Zainal et al. performed whole genome sequencing on 21 tumor/normal pairs [12]. Besides confirming recurrent somatic mutations in TP53, GATA3 and PIK3CA, these studies discovered novel cancer-related mutations. Although novel mutations occur at low frequency (less than 10%), mutations of specific genes are enriched in the subtype of breast cancers and could be grouped into cancer-related pathways. For example, mutations of MAP3K1 frequently occur in luminal A subtype [5,7]. Pathways involving p53, chromatin remodeling and ERBB signaling are overrepresented in mutated genes [11]. Furthermore, some mutations indicate therapeutic opportunities such as the mutant GATA3, which might be a positive predictive marker for aromatase inhibitor response [7].

Genomic sequencing has also helped characterize the mutation profile of colorectal cancer. For example, exome sequencing performed on 72 tumor-normal pairs identified 36,303 protein-altering somatic mutations. Further analysis for significantly mutated genes led to 23 candidates that included expected cancer genes such as KRAS, TP53 and PIK3CA and novel genes such as ATM, which regulates the cell cycle checkpoint. RNA sequencing identified recurrent R-spondin fusions,

**Table 1 Recent NGS-based studies in cancer**

Cancer	Experiment Design	Description	ref
Colon cancer	72 WES, 68 RNA-seq, 2 WGS	Identify multiple gene fusions such as RSPO2 and RSPO3 from RNA-seq that may function in tumorigenesis	[15]
Breast cancer	65 WGS/WES, 80 RNA-seq	36% of the mutations found in the study were expressed. Identify the abundance of clonal frequencies in an epithelial tumor subtype	[11]
Hepatocellular carcinoma	1 WGS, 1 WES	Identify TSC1 nonsense substitution in subpopulation of tumor cells, intra-tumor heterogeneity, several chromosomal rearrangements, and patterns in somatic substitutions	[17]
Breast cancer	510 WES	Identify two novel protein-expression-defined subgroups and novel subtype-associated mutations	[5]
Colon and rectal cancer	224 WES, 97 WGS	24 genes were found to be significantly mutated in both cancers. Similar patterns in genomic alterations were found in colon and rectum cancers	[14]
squamous cell lung cancer	178 WES, 19 WGS, 178 RNA-seq, 158 miRNA-seq	Identify significantly altered pathways including NFE2L2 and KEAP1 and potential therapeutic targets	[16]
Ovarian carcinoma	316 WES	Discover that most high-grade serous ovarian cancer contain TP53 mutations and recurrent somatic mutations in 9 genes	[13]
Melanoma	25 WGS	Identify a significantly mutated gene, PREX2 and obtain a comprehensive genomic view of melanoma	[20]
Acute myeloid leukemia	8 WGS	Identify mutations in relapsed genome and compare it to primary tumor. Discover two major clonal evolution patterns	[21]
Breast cancer	24 WGS	Highlights the diversity of somatic rearrangements and analyzes rearrangement patterns related to DNA maintenance	[8]
Breast cancer	31 WES, 46 WGS	Identify eighteen significant mutated genes and correlate clinical features of oestrogen-receptor-positive breast cancer with somatic alterations	[7]
Breast cancer	103 WES, 17 WGS	Identify recurrent mutation in CBFB transcription factor gene and deletion of RUNX1. Also found recurrent MAGI3-AKT3 fusion in triple-negative breast cancer	[6]
Breast cancer	100 WES	Identify somatic copy number changes and mutations in the coding exons. Found new driver mutations in a few cancer genes	[9]
Acute myeloid leukemia	24 WGS	Discover that most mutations in AML genomes are caused by random events in hematopoietic stem/progenitor cells and not by an initiating mutation	[22]
Breast cancer	21 WGS	Depict the life history of breast cancer using algorithms and sequencing technologies to analyze subclonal diversification	[12]
Head and neck squamous cell carcinoma	32 WES	Identify mutation in NOTCH1 that may function as an oncogene	[19]
Renal carcinoma	30 WES	Examine intra-tumor heterogeneity reveal branch evolutionary tumor growth	[18]

which might potentiate Wnt signaling and induce tumorigenesis [15]. Another example includes exome sequencing performed on 224 tumor and normal pairs. This study identified 15 highly mutated genes in the hypermutated cancers and 17 in the non-hypermutated cancers. Among the non-hypermutated cancers, novel frequent mutations in SOX9, ARID1A, ATM and FAM123B were detected besides the known APC, TP53 and KRAS mutations. The analysis of the mutations and functional roles of SOX9, ARID1A, ATM and FAM123B suggested they are highly potential colorectal cancer-related genes. Non-hypermutated colon and rectum cancers were found to have similar patterns in genomic alternation. Whole genome sequencing of 97 tumors

with matched normal samples identified the recurrent NAV2-TCF7L1 fusion [14].

#### Tumor heterogeneity and evolution

What makes cancer a difficult disease to conquer has much to do with the evolution of cancer that results from the selection and genetic instability occurring in each clone, leading to heterogeneity in tumors [26]. This idea was first proposed by Peter Nowell in 1976 as the clonal evolution model of cancer, which attempted to explain the increase in tumor aggressiveness over a period of time. Further work by other researchers in the 1980s supported this theory with studies of metastatic subclones from a mouse sarcoma cell line [26].

The wide application of NGS has revealed substantial insights into tumor heterogeneity and tumor evolution. Variations between tumors are referred to as intertumor heterogeneity, while variations within a single tumor are intratumor heterogeneity. Intertumor heterogeneity is recognized by different morphological phenotype, expression profiles and mutation and copy number variation patterns, categorizing tumors into different subtypes [27-31]. The mRNA-expression subtype was found to be associated with somatic mutation landscapes in the recent TCGA and Eillis et al.'s studies. [5,7]. As a huge amount of somatic mutations generated by NGS, the picture emerges like that individual tumor is unique, each containing distinct mutation patterns. For instance, Stephens et al. found that there were 73 different combination possibilities of mutated cancer genes among the 100 breast cancers [9].

Intratumor heterogeneity can be recognized as non-identical cellular clones or subclones within a single tumor, indicating different histology, gene expression, and metastatic and proliferative potential. The ability to generate high-resolution data makes NGS a particularly useful tool for studying intratumor heterogeneity. A recent NGS-based study on renal cell carcinoma from four patients has successfully illuminated intratumor heterogeneity [18]. For patient 1, the pre-treatment samples of the primary tumor and chest-wall metastasis went through exon-capture multi-region sequencing on DNA. Of the 128 validated mutations found in 9 regions of the primary tumor, 40 were ubiquitous, 59 were shared by some regions, and 29 were unique to specific regions, showing that genetic heterogeneity exists within a tumor and an "ongoing regional clonal evolution" [18]. Most importantly, the study showed that a single biopsy of a tumor only reveals a small part of a tumor's mutational landscape; from a single biopsy, about 55% of all mutations were detected in this tumor and 34% were shared by most regions of the tumor.

The ongoing and parallel evolution of cancer cells may establish and maintain intratumor heterogeneity. For example, phylogenetic relationships of the tumor regions in patient 1 and 2 by the renal cell carcinoma study revealed a branching rather than linear evolution of the tumor [18]. Studies have also shown branching structures of evolution in breast cancer [26]. According to the "Trunk-Branch Model of Tumor Growth" [26], there are somatic events that promote tumor growth, which represents the trunk of the tree in the early stage of tumor development. These somatic aberrations would most likely be ubiquitous at this stage. Over time, other somatic events, known as drivers, cause tumor heterogeneity to occur, which causes branching to take place in tumors as well as in metastatic sites. Later, these branches will evolve and become more isolated, resulting in a 'Bottleneck Effect' that can result in chromosomal instability, allowing further expansion of tumor heterogeneity [26]. This leads to the tumor's ability

to adapt and survive in changing environments, which affects the success of drug treatment [18]. Therefore, it is important to examine tumor clonal structure and identify common mutations located in the trunk of the phylogenetic tree, which may help understand target therapy resistance and discover more robust therapeutic approaches.

### **Clinical application**

Besides allowing researchers to understand mutations in cancer, NGS has already been applied to the clinic in many areas including prenatal diagnostics, pathogen detection, genetic mutations, and more [32]. Although genetic mutations have been identified with Sanger sequencing, PCR, and microarrays in clinical application, these three have limitations that don't apply to NGS. For example, although microarrays can detect single nucleotide variants (SNVs), they have trouble identifying larger DNA aberrations, e.g., large indels and structural rearrangements, which are common in cancer. In contrast, whole exome and whole-genome sequencing can provide the clinician a comprehensive view of the DNA aberrations, genetic recombination, and other mutations [28,32]. Therefore, NGS platforms serve as a good diagnostic and prognostic tool and help clinicians identify specific characteristics in each patient, paving the road towards personalized medicine.

NGS has already been applied in the clinic for cancer diagnosis and prognosis. For example, whole genome sequencing identified a novel insertional fusion that created a classic bcr3 PML-RARA fusion gene for a patient with acute myeloid leukemia and the findings altered the treatment plan for the patient [33]. By sequencing the tumor genome of a patient, clinicians are able to design patient-specific probes that uses DNA in the patient's blood serum to monitor the progress of a patient's treatment and detect for any signs of relapse [27-31]. The discovery of more biomarkers and the development of target-therapies will be essential in helping a clinician choose the best personalized treatment for his or her patients.

There has also been a dramatic increase in the number of clinical trials using NGS technologies since 2010 (Table 2). Ranging from WGS and WES to RNA-seq and targeted sequencing, clinical trials are using NGS to find genetic alterations that are the drivers of certain diseases in patients and apply that knowledge into the practice of clinical medicine. The information gained from these studies may help with drug development and explain the resistance of certain treatments.

### **Methods and resources**

#### ***Pipeline and tools for NGS data analysis***

To analyze and interpret the increasing amount of sequencing data, a number of statistical methods and

**Table 2 Active cancer studies using NGS as the primary outcome measure**

Study Title/Sponsor	NCT#/# Enrolled/ Start Date	Condition	Description	Sequencing Technologies
Tumor Specific Plasma DNA in Breast Cancer/ <i>Dartmouth-Hitchcock Medical Center</i>	NCT01617915/6/ October 2012	Breast Cancer	Analyze chromosomal rearrangements and genomic alterations	Whole genome sequencing
Whole Exon Sequencing of Down Syndrome Acute Myeloid Leukemia/ <i>Children's Oncology Group</i>	NCT01507441/10/ February 2012	Leukemia	Examine DNA samples of patients with Leukemia and Down Syndrome and identify DNA alterations	Whole exome Sequencing
Studying Genes in Samples From Younger Patients with Adrenocortical Tumor/ <i>Children's Oncology Group</i>	NCT01528956/10/ February 2012	Adrenocortical Carcinoma	Study genes from patients with adrenocortical tumor	Whole genome Sequencing
Feasibility Clinical Study of Targeted and Genome-Wide Sequencing/ <i>University Health Network, Toronto</i>	NCT01345513/ 150/March 2011	Solid Tumors	Identify gene mutations in cancer patients	Whole genome sequencing
An Ancillary Pilot Trial Using Whole Genome Sequencing in Patients with Advance Refractor Cancer/ <i>Scottsdale Healthcare</i>	NCT01443390/10/ September 2011	Advanced Cancer	Investigate patients with cancer that are using Phase I drugs and its effect on the patient	Whole genome Sequencing
Cancer Genome Analysis/ <i>Seoul National University Hospital</i>	NCT01458604/ 100/August 2011	Malignant Tumor	Identify and analyze genetic alterations in tumors for therapeutic agents	Targeted Sequencing, whole exome sequencing and RNA-seq
RNA Biomarkers in Tissue Samples From Infants with Acute Meyloid Leukemia/ <i>Children's Oncology Group</i>	NCT01229124/20/ October 2010	Leukemia	Analyze tissue samples and identify biomarkers from RNA	RNA-seq
Molecular Analysis of Solid Tumors/ <i>St. Jude Children's Research Hospital</i>	NCT01050296/ 360/January 2010	Pediatric Solid Tumors	Analyze gene expression profiles of tumor and examine genetic alterations	Whole genome Sequencing
Deep Sequencing of the Breast Cancer Transcriptome/ <i>University of Arkansas</i>	NCT01141530/30/ Sept 2009	Breast Cancer	Examine transcriptional regulation and triple negative breast cancer	RNA-seq

bioinformatics tools have been developed. For WGS and WES, the analysis generally includes read alignment, variant detection (point mutation, small indels, copy number variation and structural rearrangement) and variant functional prediction (Table 3). Reads are mapped back to the human reference genomes using MAQ [34], BWA [35,36], Bowtie2 [37], BFAST [38], SOAP2 [39], Novoalign/NovoalignCS, SSAHA2 [40], SHRiMP [41], etc. These methods differ in their computational efficiency, sensitivity and ability to accurately map noisy reads, to deal with long or short reads and pair-end reads. Having aligned the reads to the genome, mutation calling identifies the sites in which at least one of the bases differs from a reference sequence by GATK [42], SAMtools [43], SOAPsnp [44], SNVMix [45], Varscan [46], etc. Differing in the underlying statistical models, the performances of these methods are comparable and vary on sequencing depths [47-49]. Detecting somatic mutation involves mutation calling in paired tumor-normal DNA, coupled with comparison to the reference. A naïve somatic mutation caller applies standard calling tools on the normal and tumor samples separately and then selects mutations detected in tumor but not in normal. Alternatively, a complicated caller jointly analyzes tumor-normal pair data such as Varscan2 [50], Somaticsniiper [51] and JointSNVMix [52]. SIFT [53], PolyPhen [54], CHASM [55] and ANNOVAR [56] have been developed to

understand the impact of the mutations on gene function and to distinguish between driver and passenger mutations. For WGS, various kinds of structural variations can be discovered using BreakDancer [57], VariationHunter [58], PEMer [59] and SVDetect [60]. RNA-seq data analysis generally includes reads alignment, gene expression quantification, differentially expressed genes/isoforms or alternative splicing detection and novel transcripts discovery (Table 4). There are two major approaches to map RNA-seq reads. One is to align reads to the reference transcriptome using standard DNA-seq reads aligner. The alternative is to map reads to the reference genome allowing for the identification of novel splice junctions using a RNA-seq specific aligner, such as TopHat [61], MapSplice [62], SpliceMap [63], GSNAP [64], and STAR [65]. Having aligned reads, expression values are quantified by aggregating reads into counts and differential expression analysis is performed based on counts (DEseq [66], edgeR [67]) or FPKM/RPKM values (CuffLinks [68,69]). Estimating isoform-level expression is very difficult since many genes have multiple isoforms and most reads are shared by different isoforms. To deal with read assignment uncertainty, Alexa-seq [70] counts only the reads that map uniquely to a single isoform, while Cufflinks [68,69] and MISO [71] construct a likelihood model that best explains all the reads obtained in the experiment. In addition, fusion transcripts can be detected

**Table 3 Computational tools for cancer genomics**

Category	Program	URL	Ref
Alignment	MAQ	<a href="http://maq.sourceforge.net/">http://maq.sourceforge.net/</a>	[34]
	BWA	<a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a>	[35,36]
	Bowtie2	<a href="http://bowtie-bio.sourceforge.net/bowtie2/">http://bowtie-bio.sourceforge.net/bowtie2/</a>	[37]
	BFAST	<a href="http://bfast.sourceforge.net">http://bfast.sourceforge.net</a>	[38]
	SOAP2	<a href="http://soap.genomics.org.cn/soapaligner.html">http://soap.genomics.org.cn/soapaligner.html</a>	[39]
	Novoalign/NovoalignCS	<a href="http://www.novocraft.com/">http://www.novocraft.com/</a>	
	SSAHA2	<a href="http://www.sanger.ac.uk/resources/software/ssaha2/">http://www.sanger.ac.uk/resources/software/ssaha2/</a>	[40]
	SHRIMP	<a href="http://compbio.cs.toronto.edu/shrimp/">http://compbio.cs.toronto.edu/shrimp/</a>	[41]
Mutation calling	GATK	<a href="http://www.broadinstitute.org/gatk/">http://www.broadinstitute.org/gatk/</a>	[42]
	Samtools	<a href="http://samtools.sourceforge.net/">http://samtools.sourceforge.net/</a>	[43]
	SOAPSnp	<a href="http://soap.genomics.org.cn/soapsnp.html">http://soap.genomics.org.cn/soapsnp.html</a>	[44]
	SNVMix	<a href="http://compbio.bccrc.ca/software/snvmix/">http://compbio.bccrc.ca/software/snvmix/</a>	[45]
	VarScan	<a href="http://varscan.sourceforge.net/">http://varscan.sourceforge.net/</a>	[46,50]
	Somaticsniper	<a href="http://gmt.genome.wustl.edu/somatic-sniper/">http://gmt.genome.wustl.edu/somatic-sniper/</a>	[51]
	JointSNVMix	<a href="http://compbio.bccrc.ca/software/jointsnvmix/">http://compbio.bccrc.ca/software/jointsnvmix/</a>	[52]
	SV detection	BreakDancer	<a href="http://breakdancer.sourceforge.net/">http://breakdancer.sourceforge.net/</a>
VariationHunter		<a href="http://variationhunter.sourceforge.net/">http://variationhunter.sourceforge.net/</a>	[58]
PEMer		<a href="http://sv.gersteinlab.org/pemer/">http://sv.gersteinlab.org/pemer/</a>	[59]
SVDetect		<a href="http://svdetect.sourceforge.net/">http://svdetect.sourceforge.net/</a>	[60]
Function effect of mutation	SIFT	<a href="http://sift.jcvi.org/">http://sift.jcvi.org/</a>	[53]
	CHASM	<a href="http://wiki.chasmssoftware.org">http://wiki.chasmssoftware.org</a>	[55]
	PolyPhen-2	<a href="http://genetics.bwh.harvard.edu/pph2/">http://genetics.bwh.harvard.edu/pph2/</a>	[54]
	ANNOVAR	<a href="http://www.openbioinformatics.org/annovar/">http://www.openbioinformatics.org/annovar/</a>	[56]

Source: [www.clinicaltrials.gov](http://www.clinicaltrials.gov).

using SOAPfusion, TopHat-Fusion [72], BreakFusion [73], FusionHunter [74], deFuse [75], FusionAnalyser [76], etc. To obtain a more complete view of cancer genome, an integrative approach to study diverse mutations, transcriptomes and epigenomes simultaneously on the pathways or networks is much more informative and promising. A growing number of pathway-oriented tools is now becoming available, including PARADIGM [77], NetBox [78], MEMo [79], CONEXIC [80], etc.

#### Comprehensive cancer projects and resources

The vast amount of oncogenomics data are generated from large scale collaborative cancer projects (Table 5). The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) are the two largest representatives of such coordinated efforts. Beginning as a three-year pilot in 2006, TCGA aims to comprehensively map the important genomic changes that occur in the major types and subtypes of cancer. TCGA will examine over 11,000 samples for 20 cancer types (<http://cancergenome.nih.gov/>). ICGC launched in 2008 and its goal is 'to obtain a comprehensive description of genomic, transcriptomic

and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe' (<http://icgc.org/icgc>). The Cancer Genome Project (CGP) has many efforts at the Sanger Institute and aims to identify sequence variants/mutations critical in the development of human cancers (<http://www.sanger.ac.uk/genetics/CGP/>). The NCI's Cancer Genome Anatomy Project (CGAP) seeks to determine the gene expression profiles of normal, precancer and cancer cells, leading eventually to improved detection, diagnosis and treatment for the patient (<http://cgap.nci.nih.gov/>). Recently, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) has launched to systematically identify proteins that derive from alterations in cancer genomes using proteomic technologies (<http://proteomics.cancer.gov/>). The combination of genomic and proteomic initiatives is anticipated to produce a more comprehensive inventory of the detectable proteins in a tumor and advance our understanding of cancer biology.

The data and the results from these projects are freely available to the research community (Table 5). A number of databases and frameworks have been developed to make

**Table 4 Computational tools for cancer transcriptomics**

Category	Program	URL	ref
Spliced alignment	TopHat	<a href="http://tophat.cbcb.umd.edu/">http://tophat.cbcb.umd.edu/</a>	[61,69]
	MapSplice	<a href="http://www.netlab.uky.edu/p/bioinfo/MapSplice">http://www.netlab.uky.edu/p/bioinfo/MapSplice</a>	[62]
	SpliceMap	<a href="http://www.stanford.edu/group/wonglab/SpliceMap/">http://www.stanford.edu/group/wonglab/SpliceMap/</a>	[63]
	GSNAP	<a href="http://research-pub.gene.com/gmap/">http://research-pub.gene.com/gmap/</a>	[64]
	STAR	<a href="http://gingeraslab.cshl.edu/STAR/">http://gingeraslab.cshl.edu/STAR/</a>	[65]
Differential expression	CuffDiff	<a href="http://cufflinks.cbcb.umd.edu/">http://cufflinks.cbcb.umd.edu/</a>	[68,69]
	EdgeR	<a href="http://www.bioconductor.org/packages/2.11/bioc/html/edgeR.html">http://www.bioconductor.org/packages/2.11/bioc/html/edgeR.html</a>	[67]
	DESeq	<a href="http://www-huber.embl.de/users/anders/DESeq/">http://www-huber.embl.de/users/anders/DESeq/</a>	[66]
	Myrna	<a href="http://bowtie-bio.sourceforge.net/myrna/index.shtml">http://bowtie-bio.sourceforge.net/myrna/index.shtml</a>	[81]
Alternative splicing	CuffDiff	<a href="http://cufflinks.cbcb.umd.edu/">http://cufflinks.cbcb.umd.edu/</a>	[68,69]
	MISO	<a href="http://genes.mit.edu/burgelab/miso/">http://genes.mit.edu/burgelab/miso/</a>	[71]
	DEXseq	<a href="http://watson.nci.nih.gov/bioc_mirror/packages/2.9/bioc/html/DEXSeq.html">http://watson.nci.nih.gov/bioc_mirror/packages/2.9/bioc/html/DEXSeq.html</a>	[82]
	Alexa-seq	<a href="http://www.alexaplatform.org/alexa_seq/">http://www.alexaplatform.org/alexa_seq/</a>	[70]
Gene fusion	SOAPfusion	<a href="http://soap.genomics.org.cn/SOAPfusion.html">http://soap.genomics.org.cn/SOAPfusion.html</a>	
	TopHat-Fusion	<a href="http://tophat.cbcb.umd.edu/fusion_index.html">http://tophat.cbcb.umd.edu/fusion_index.html</a>	[72]
	BreakFusion	<a href="http://bioinformatics.mdanderson.org/main/BreakFusion">http://bioinformatics.mdanderson.org/main/BreakFusion</a>	[73]
	FusionHunter	<a href="http://bioen-compbio.bioen.illinois.edu/FusionHunter/">http://bioen-compbio.bioen.illinois.edu/FusionHunter/</a>	[74]
	deFuse	<a href="http://sourceforge.net/apps/mediawiki/defuse/">http://sourceforge.net/apps/mediawiki/defuse/</a>	[75]
	FusionAnalyser	<a href="http://www.ilte-cml.org/FusionAnalyser/">http://www.ilte-cml.org/FusionAnalyser/</a>	[76]

the data and the results easily and directly accessible. For example, the results from CGP are collated and stored in COSMIC [83]. The cBio Cancer Genomics Portal, containing dataset from TCGA and published papers, is specifically designed to interactively explore multidimensional cancer genomics data, including mutation, copy number variations, expression changes (microarray and RNA-seq), DNA methylation values, and protein and phosphoprotein levels [84]. Intogen is also a framework that facilitates the analysis and integration of multidimensional data for the identification of genes and biological modules critical in cancer development [85]. The Broad GDAC Firehose, designed to coordinate the various tools utilized by TCGA, provides level 3 and level 4 analyses and enables researchers to easily incorporate TCGA data into their projects. Table 5 also includes resources useful for cancer research but not built on NGS data, e.g., Progenetix [86].

### Challenges and perspective

Although NGS has already helped researchers discover a plethora of information in the field of cancer, challenges in translating the large amounts of oncogenomics data into information that can be easily interpretable and accessible for cancer care still lie ahead. From a computational point of view, many technical and statistical issues remain unsolved. For example, repetitive DNA represents a major obstacle for the accuracy of read alignment and assembly, as well as structure variation detection [87].

Furthermore, it is difficult to distinguish rare mutations in tumor from sequencing and alignment artifacts, especially when a tumor has low purity. Despite new methods to comprehensively catalogue genomic variants, the prediction of their functional effect and the identification of disease-causal variants are still in an early phase [88]. Current algorithms for quantifying isoform expression are not computationally trivial and are incredibly difficult to explain. Although the concept of integrative analysis is not new, predictive networks or pathway models that combine various omics data are still underway. Most importantly, since sequencing technologies and methodologies are both evolving rapidly, it is a difficult challenge to store, analyze and present the data in a method that is transparent and reproducible [89]. On the other hand, tumor complexity and heterogeneity make the analysis and the interpretation of sequencing data even harder. Heterogeneity is dynamic and evolves over time. This challenges the simple notion of binning mutations as tumorigenesis ‘driver’ and neutral ‘passenger’, since some passengers are also drivers just waiting for the right context [90].

From a clinical point of view, a major challenge is to assess genomic variants as potential therapeutic targets. Although many diverse variants are demonstrated to converge on similar deregulated pathways, there is still a lack of pathway-targeted therapies. With the discovery of intra-tumor heterogeneity, questions have been raised about how well a glimpse of a tumor’s genomic

**Table 5 Comprehensive cancer projects and resources**

Name	Description	URL
Comprehensive cancer projects		
The Cancer Genome Atlas	A joint effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies	<a href="http://cancergenome.nih.gov/">http://cancergenome.nih.gov/</a>
International Cancer Genome Consortium	International consortium with the goal of obtaining comprehensive description of genomic, transcriptomic, and epigenomic changes in 50 different cancer types and/or subtypes of clinical and societal importance across the globe	<a href="http://icgc.org/icgc">http://icgc.org/icgc</a>
Cancer Genome Anatomy Project	Interdisciplinary program to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient	<a href="http://cgap.nci.nih.gov/">http://cgap.nci.nih.gov/</a>
Cancer Genome Project	To identify somatically acquired sequence variants/mutations and hence identify genes critical in the development of human cancers	<a href="http://www.sanger.ac.uk/genetics/CGP/">http://www.sanger.ac.uk/genetics/CGP/</a>
The Clinical Proteomic Tumor Analysis Consortium	A comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of proteomic technologies	<a href="http://proteomics.cancer.gov/">http://proteomics.cancer.gov/</a>
Resources		
COSMIC	Catalogue of Somatic Mutations in Cancer	<a href="http://www.sanger.ac.uk/genetics/CGP/cosmic/">http://www.sanger.ac.uk/genetics/CGP/cosmic/</a>
Progenetix	Copy number abnormalities in human cancer from CGH experiments	<a href="http://www.progenetix.org/cgi-bin/pgHome.cgi">http://www.progenetix.org/cgi-bin/pgHome.cgi</a>
MethyCancer	An information resource and analysis platform for study interplay of DNA methylation, gene expression and cancer	<a href="http://methycancer.psych.ac.cn/">http://methycancer.psych.ac.cn/</a>
IntOGen	Integrates multidimensional OncoGenomics Data for the identification of genes and groups of genes involved in cancer development	<a href="http://www.intogen.org/">www.intogen.org/</a>
Oncomine	A cancer microarray database and integrated data-mining platform	<a href="http://www.oncomine.org/">www.oncomine.org/</a>
cBio	Provides visualization, analysis and download of large-scale cancer genomics data sets	<a href="http://www.cbioportal.org/">www.cbioportal.org/</a>
Firehose	Provides L3 data and L4 analyses packaged in a form amenable to immediate algorithmic analysis	<a href="https://confluence.broadinstitute.org/display/GDAC/Home">https://confluence.broadinstitute.org/display/GDAC/Home</a>
UCSC Cancer Genomics Browser	A suite of web-based tools to visualize, integrate and analyze cancer genomics and its associated clinical data	<a href="https://genome-cancer.soe.ucsc.edu/">https://genome-cancer.soe.ucsc.edu/</a>
Cancer Genome Workbench	Hosts mutation, copy number, expression, and methylation data from a number of projects, including TCGA, TARGET, COSMIC, GSK, NCI60. It has tools for visualizing sample-level genomic and transcription alterations in various cancers.	<a href="https://cgwb.nci.nih.gov/">https://cgwb.nci.nih.gov/</a>

landscape can steer the treatment. Currently, many clinicians decide a treatment based on the genetic markers from a few biopsies. Whether these markers are over- or under-represented in the tumor is unknown, causing the selection of treatment to be difficult [29]. In addition to heterogeneity, the tumor's ability to evolve allows it to have more opportunities to adapt and survive to various treatments. Some researchers hope that with current target therapies, intratumor heterogeneity will decrease to a certain point [29] so that clinicians can then target the non-responsive clones before a tumor re-growth and more mutations can occur; however, choosing an appropriate target therapy will be a challenge. A few researchers have already shown certain treatments, such as the cytotoxic therapies, that have increased genome instability and diversity, resulting in a faster tumor evolution rate and, thus, heterogeneity. The fact is that this area of

cancer is understudied [26]; however, one of the key challenges researchers must solve is identifying branched subclones are resistant to which target therapies. More knowledge of network medicine and the interaction between the trunk and branch mutations may lead to appropriate target therapies and personalized therapeutic strategies that can prevent drug resistance and effectively eradicate cancer [26,91].

To accelerate the rate of translating genomic data into clinical practice, a sustained collaboration among multiple centers and effective communication among bioinformaticians, statistical geneticists, molecular biologists and physician are required. Bioinformaticians and statistical geneticists are responsible for providing reproducible and accurate analysis, identifying 'drivers' in the unstable and evolving cancer genome and building powerful and flexible integrative model to consider interactions among



genomic, transcriptomic, metabolomics, proteomics and epigenomic alterations in the context of tumor microenvironment. Biologists interpret and confirm the functional relevance of variants to cancer. Physicians assess relationships of variants to cancer prognosis and response to therapy. Appropriate infrastructure within each research institution that integrates the clinic for patient samples, wet lab for sequencing, and Bioinformatics for data analysis should allow the sequenced data to be processed efficiently, producing results that can create effective personalized therapies applicable to the clinic. In addition, easily accessible and understandable databases that connect genomic findings with clinical outcome are also required. With these efforts and developments, NGS will greatly potentiate genome-based cancer diagnosis and personalized treatment strategies.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

QL led the project. DS drafted the manuscript and QL revised the manuscript. All authors read and approved the final manuscript.

#### Acknowledgements

This work was supported by National Cancer Institute grants U01 CA163056, P30 CA068485, P50 CA098131, and P50 CA090949 and QL's work was partially supported by the State Key Program of National Natural Science of China (no. 31230058) and the National Natural Science Foundation of China (no. 31070746).

#### Author details

<sup>1</sup>Washington University, St. Louis, MO 63130, USA. <sup>2</sup>Center for Quantitative Sciences, Vanderbilt University School of Medicine, Nashville, TN 37232, USA. <sup>3</sup>Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA.

Received: 6 February 2013 Accepted: 9 February 2013

Published: 13 February 2013

#### References

1. Taylor BS, Ladanyi M: Clinical cancer genomics: how soon is now? *J Pathol* 2011, **223**:318–326.
2. Sosman JA, Kim KB, Schuchter L, Gonzalez R, Pavlick AC, Weber JS, McArthur GA, Hutson TE, Moschos SJ, Flaherty KT, Hersey P, Kefford R, Lawrence D, Puzanov I, Lewis KD, Amaravadi RK, Chmielowski B, Lawrence HJ, Shyr Y, Ye F, Li J, Nolop KB, Lee RJ, Joe AK, Ribas A: **Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib.** *N Engl J Med* 2012, **366**:707–714.
3. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11**:31–46.
4. Wold B, Myers RM: **Sequence census methods for functional genomics.** *Nat Methods* 2008, **5**:19–21.
5. Cancer Genome Atlas Research Network: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490**:61–70.
6. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, Peng S, Ardlie KG, Auclair D, Bautista-Pina V, Duke F, Francis J, Jung J, Maffuz-Aziz A, Onofrio RC, Parkin M, Pho NH, Quintanar-Jurado V, Ramos AH, Rebollar-Vega R, Rodriguez-Cuevas S, Romero-Cordoba SL, Schumacher SE, Stransky N, Thompson KM, Uribe-Figueroa L, Baselga J, Beroukhi R, Polyak K, Sgroi DC, Richardson AL, Jimenez-Sanchez G, Lander ES, Gabriel SB, Garraway LA, Golub TR, Melendez-Zajgla J, Toker A, Getz G, Hidalgo-Miranda A, Meyerson M: **Sequence analysis of mutations and translocations across breast cancer subtypes.** *Nature* 2012, **486**:405–409.
7. Ellis MJ, et al: **Whole-genome analysis informs breast cancer response to aromatase inhibition.** *Nature* 2012, **486**:353–360.
8. Stephens PJ, et al: **Complex landscapes of somatic rearrangement in human breast cancer genomes.** *Nature* 2009, **462**:1005–1010.
9. Stephens PJ, et al: **The landscape of cancer genes and mutational processes in breast cancer.** *Nature* 2012, **486**:400–404.
10. Nik-Zainal S, et al: **The life history of 21 breast cancers.** *Cell* 2012, **149**:994–1007.
11. Shah SP, et al: **The clonal and mutational evolution spectrum of primary triple-negative breast cancers.** *Nature* 2012, **486**:395–399.
12. Nik-Zainal S, et al: **Mutational processes molding the genomes of 21 breast cancers.** *Cell* 2012, **149**:979–993.
13. Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474**:609–615.
14. Cancer Genome Atlas Research Network: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487**:330–337.
15. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, Chaudhuri S, Guan Y, Janakiraman V, Jaiswal BS, Guillery J, Ha C, Dijkgraaf GJ, Stinson J, Gnad F, Huntley MA, Degenhardt JD, Haverty PM, Bourgon R, Wang W, Koeppen H, Gentleman R, Starr TK, Zhang Z, Largaespada DA, Wu TD, de Sauvage FJ: **Recurrent R-spondin fusions in colon cancer.** *Nature* 2012, **488**:660–664.
16. Hammerman PS, Hayes DN, Wilkerson MD, Schultz N, Bose R, Chu A, Collisson EA, Cope L, Creighton CJ, Getz G, Herman JG, Johnson BE, Kucherlapati R, Ladanyi M, Maher CA, Robertson G, Sander C, Shen R, Sinha R, Sivachenko A, Thomas RK, Travis WD, Tsao MS, Weinstein JN, Wigle DA, Baylín SB, Govindan R, Meyerson M: **Comprehensive genomic characterization of squamous cell lung cancers.** *Nature* 2012, **489**:519–525.
17. Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, Tsutsumi S, Sonoda K, Totsuka H, Shirahara T, Sakamoto H, Wang L, Ojima H, Shimada K, Kosuge T, Okusaka T, Kato K, Kusuda J, Yoshida T, Aburatani H, Shibata T: **High-resolution characterization of a hepatocellular carcinoma genome.** *Nat Genet* 2011, **43**:464–469.
18. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, Varela I, Phillimore B, Begum S, McDonald NQ, Butler A, Jones D, Raine K, Latimer C, Santos CR, Nohadani M, Eklund AC, Spencer-Dene B, Clark G, Pickering L, Stamp G, Gore M, Szallasi Z, Downward J, Futreal PA, Swanton C: **Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.** *N Engl J Med* 2012, **366**:883–892.
19. Agrawal N, Frederick MJ, Pickering CR, Bettgowda C, Chang K, Li RJ, Fakhry C, Xie TX, Zhang J, Wang J, Zhang N, El-Naggar AK, Jasser SA, Weinstein JN, Trevino L, Drummond JA, Muzny DM, Wu Y, Wood LD, Hruban RH, Westra WH, Koch WM, Califano JA, Gibbs RA, Sidransky D, Vogelstein B, Velculescu VE, Papadopoulos N, Wheeler DA, Kinzler KW, Myers JN: **Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1.** *Science* 2011, **333**:1154–1157.
20. Berger MF, et al: **Melanoma genome sequencing reveals frequent PREX2 mutations.** *Nature* 2012, **485**:502–506.
21. Ding L, et al: **Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing.** *Nature* 2012, **481**:506–510.
22. Welch JS, et al: **The origin and evolution of mutations in acute myeloid leukemia.** *Cell* 2012, **150**:264–278.
23. Wong KM, Hudson TJ, McPherson JD: **Unraveling the genetics of cancer: genome sequencing and beyond.** *Annu Rev Genomics Hum Genet* 2011, **12**:407–430.
24. Cahill DP, Kinzler KW, Vogelstein B, Lengauer C: **Genetic instability and darwinian selection in tumours.** *Trends Cell Biol* 1999, **9**:M57–M60.
25. Brosnan JA, Iacobuzio-Donahue CA: **A new branch on the tree: next-generation sequencing in the study of cancer evolution.** *Semin Cell Dev Biol* 2012, **23**:237–242.
26. Swanton C: **Intratumor heterogeneity: evolution through space and time.** *Cancer Res* 2012, **72**:4875–4882.
27. Russnes HG, Navin N, Hicks J, Borresen-Dale AL: **Insight into the heterogeneity of breast cancer through next-generation sequencing.** *J Clin Invest* 2011, **121**:3810–3818.
28. Samuel N, Hudson TJ: *Translating Genomics to the Clinic.* Clinical chemistry: Implications of Cancer Heterogeneity; 2012.
29. Almendro V, Fuster G: **Heterogeneity of breast cancer: etiology and clinical relevance.** *Clinical & translational oncology: official publication of the*

- Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico* 2011, **13**:767–773.
30. Yancovitz M, Litterman A, Yoon J, Ng E, Shapiro RL, Berman RS, Pavlick AC, Darvishian F, Christos P, Mazumdar M, Osman I, Polsky D: **Intra- and inter-tumor heterogeneity of BRAF(V600E) mutations in primary and metastatic melanoma.** *PLoS One* 2012, **7**:e29336.
  31. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowitz F, Grauphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature* 2012, **486**:346–352.
  32. Desai AN, Jere A: **Next-generation sequencing: ready for the clinics?** *Clin Genet* 2012, **81**:503–510.
  33. Welch JS, Westervelt P, Ding L, Larson DE, Klco JM, Kulkarni S, Wallis J, Chen K, Payton JH, Fulton RS, Veizer J, Schmidt H, Vickery TL, Heath S, Watson MA, Tomasson MH, Link DC, Graubert TA, DiPersio JF, Mardis ER, Ley TJ, Wilson RK: **Use of whole-genome sequencing to diagnose a cryptic fusion oncogene.** *JAMA* 2011, **305**:1577–1584.
  34. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**:1851–1858.
  35. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–1760.
  36. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26**:589–595.
  37. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**:357–359.
  38. Homer N, Merriman B, Nelson SF: **BFAST: an alignment tool for large scale genome resequencing.** *PLoS One* 2009, **4**:e7767.
  39. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinformatics* 2009, **25**:1966–1967.
  40. Ning Z, Cox AJ, Mullikin JC: **SSAHA: a fast search method for large DNA databases.** *Genome Res* 2001, **11**:1725–1729.
  41. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M: **SHRIMP: accurate mapping of short color-space reads.** *PLoS Comput Biol* 2009, **5**:e1000386.
  42. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Phillipakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43**:491–498.
  43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–2079.
  44. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K: **SNP detection for massively parallel whole-genome resequencing.** *Genome Res* 2009, **19**:1124–1132.
  45. Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, Huntsman D, Murphy KP, Aparicio S, Shah SP: **SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors.** *Bioinformatics* 2010, **26**:730–736.
  46. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L: **VarScan: variant detection in massively parallel sequencing of individual and pooled samples.** *Bioinformatics* 2009, **25**:2283–2285.
  47. Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, Snyder M: **Detecting and annotating genetic variations using the HigeSeq pipeline.** *Nat Biotechnol* 2012, **30**:226–229.
  48. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y: **Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data.** *BMC Genomics* 2012, **13**:58.
  49. Wang W, Wei Z, Lam TW, Wang J: **Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions.** *Sci Rep* 2011, **1**:55.
  50. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.** *Genome Res* 2012, **22**:568–576.
  51. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L: **SomaticSniper: identification of somatic point mutations in whole genome sequencing data.** *Bioinformatics* 2012, **28**:311–317.
  52. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliani R, Bashashati A, Hirst M, Turashvili G, Oloumi A, Marra MA, Aparicio S, Shah SP: **JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data.** *Bioinformatics* 2012, **28**:907–913.
  53. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4**:1073–1081.
  54. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**:248–249.
  55. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R: **CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer.** *Bioinformatics* 2011, **27**:2147–2148.
  56. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38**:e164.
  57. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.** *Nat Methods* 2009, **6**:677–681.
  58. Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC: **Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery.** *Bioinformatics* 2010, **26**:i350–i357.
  59. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB: **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.
  60. Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-ne P, Nicolas A, Delattre O, Barillot E: **SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data.** *Bioinformatics* 2010, **26**:1895–1896.
  61. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105–1111.
  62. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J: **MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.** *Nucleic Acids Res* 2010, **38**:e178.
  63. Au KF, Jiang H, Lin L, Xing Y, Wong WH: **Detection of splice junctions from paired-end RNA-seq data by SpliceMap.** *Nucleic Acids Res* 2010, **38**:4570–4578.
  64. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads.** *Bioinformatics* 2010, **26**:873–881.
  65. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29**:15–21.
  66. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**:R106.
  67. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**:139–140.
  68. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq.** *Nat Biotechnol* 2012, **31**:46–53.
  69. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7**:562–578.
  70. Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJ, Tai IT, Marra MA: **Alternative expression analysis by RNA sequencing.** *Nat Methods* 2010, **7**:843–847.
  71. Katz Y, Wang ET, Airoidi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation.** *Nat Methods* 2010, **7**:1009–1015.

72. Kim D, Salzberg SL: **TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.** *Genome Biol* 2011, **12**:R72.
73. Chen K, Wallis JW, Kandoth C, Kalicki-Weizer JM, Mungall KL, Mungall AJ, Jones SJ, Marra MA, Ley TJ, Mardis ER, Wilson RK, Weinstein JN, Ding L: **BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data.** *Bioinformatics* 2012, **28**:1923–1924.
74. Li Y, Chien J, Smith DJ, Ma J: **FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq.** *Bioinformatics* 2011, **27**:1708–1710.
75. McPherson A, Hormozdiari F, Zayed A, Giuliani R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Marra MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP: **deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data.** *PLoS Comput Biol* 2011, **7**:e1001138.
76. Piazza R, Pirola A, Spinelli R, Valletta S, Redaelli S, Magistrini V, Gambacorti-Passerini C: **FusionAnalyser: a new graphical, event-driven tool for fusion rearrangements discovery.** *Nucleic Acids Res* 2012, **40**:e123.
77. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM: **Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.** *Bioinformatics* 2010, **26**:i237–i245.
78. Cerami E, Demir E, Schultz N, Taylor BS, Sander C: **Automated network analysis identifies core pathways in glioblastoma.** *PLoS One* 2010, **5**:e8918.
79. Ciriello G, Cerami E, Sander C, Schultz N: **Mutual exclusivity analysis identifies oncogenic network modules.** *Genome Res* 2012, **22**:398–406.
80. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D: **An integrated approach to uncover drivers of cancer.** *Cell* 2010, **143**:1005–1017.
81. Langmead B, Hansen KD, Leek JT: **Cloud-scale RNA-sequencing differential expression analysis with Myrna.** *Genome Biol* 2010, **11**:R83.
82. Anders S, Reyes A, Huber W: **Detecting differential usage of exons from RNA-seq data.** *Genome Res* 2012, **22**:2008–2017.
83. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA: **COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.** *Nucleic Acids Res* 2011, **39**:D945–D950.
84. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.** *Cancer Discov* 2012, **2**:401–404.
85. Gundem G, Perez-Llamas C, Jene-Sanz A, Kedziarska A, Islam A, Deu-Pons J, Furney SJ, Lopez-Bigas N: **IntOGen: integration and data mining of multidimensional oncogenomic data.** *Nat Methods* 2010, **7**:92–93.
86. Baudis M, Cleary ML: **Progenetix.net: an online repository for molecular cytogenetic aberration data.** *Bioinformatics* 2001, **17**:1228–1229.
87. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nat Rev Genet* 2012, **13**:36–46.
88. Cooper GM, Shendure J: **Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.** *Nat Rev Genet* 2011, **12**:628–640.
89. Nekrutenko A, Taylor J: **Next-generation sequencing data interpretation: enhancing reproducibility and accessibility.** *Nat Rev Genet* 2012, **13**:667–672.
90. Eisenstein M: **Reading cancer's blueprint.** *Nat Biotechnol* 2012, **30**:581–584.
91. Katsios C, Papaloukas C, Tzaphlidou M, Roukos DH: **Next-generation sequencing-based testing for cancer mutational landscape diversity: clinical implications?** *Expert Rev Mol Diagn* 2012, **12**:667–670.

doi:10.1186/1480-9222-15-4

Cite this article as: Shyr and Liu: Next generation sequencing in cancer research and clinical application. *Biological Procedures Online* 2013 15:4.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

