OXFORD

Briefings in Bioinformatics, 17(4), 2016, 642–656

doi: 10.1093/bib/bbv068
Advance Access Publication Date: 24 August 2015
Paper

Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes

Feixiong Cheng, Junfei Zhao and Zhongming Zhao

Corresponding author: Zhongming Zhao, Department of Biomedical Informatics, Vanderbilt University School of Medicine, 2525 West End Avenue, Suite 600, Nashville, TN 37203, USA. Tel.: (615) 343-9158; Fax: (615) 936-8545; Email: zhongming.zhao@vanderbilt.edu

Abstract

Cancer is often driven by the accumulation of genetic alterations, including single nucleotide variants, small insertions or deletions, gene fusions, copy-number variations, and large chromosomal rearrangements. Recent advances in next-generation sequencing technologies have helped investigators generate massive amounts of cancer genomic data and catalog somatic mutations in both common and rare cancer types. So far, the somatic mutation landscapes and signatures of >10 major cancer types have been reported; however, pinpointing driver mutations and cancer genes from millions of available cancer somatic mutations remains a monumental challenge. To tackle this important task, many methods and computational tools have been developed during the past several years and, thus, a review of its advances is urgently needed. Here, we first summarize the main features of these methods and tools for whole-exome, whole-genome and whole-transcriptome sequencing data. Then, we discuss major challenges like tumor intra-heterogeneity, tumor sample saturation and functionality of synonymous mutations in cancer, all of which may result in false-positive discoveries. Finally, we highlight new directions in studying regulatory roles of noncoding somatic mutations and quantitatively measuring circulating tumor DNA in cancer. This review may help investigators find an appropriate tool for detecting potential driver or actionable mutations in rapidly emerging precision cancer medicine.

Key words: next-generation sequencing; cancer driver genes; significantly mutated genes; driver mutations; structural genomics; precision cancer medicine; panomics; computational tools

Introduction

Cancer is driven by genetic alterations, including single nucleotide variants (SNVs), small insertions or deletions (indels), gene fusions, copy-number variations (CNVs) and large chromosomal rearrangements (also called structural variants). The revolutionary advances in next-generation sequencing (NGS) technologies, now with high-throughput, much greater speed and much lower cost, have helped investigators generate massive amounts of cancer genomic data, providing somatic mutation landscapes for better understanding cancer biology and improving cancer diagnosis and therapy [1–8]. So far, somatic mutations of >20 cancer types have been systematically explored. The COSMIC (the Catalogue Of Somatic Mutations In Cancer) database deposits >3.1 million of coding mutations. In addition, several national and international cancer genome projects, such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), are ongoing, aiming to complete cancer genome sequencing for >50 types or subtypes

Feixiong Cheng is a postdoctoral researcher in Dr. Zhongming Zhao's group at Vanderbilt University, USA, working mostly on next-generation sequencing data for precision cancer medicine and network-based pharmacogenomics studies.
Junfei Zhao is a postdoctoral researcher in Dr. Zhongming Zhao's group at Vanderbilt University, USA, developing statistical methods and computational approaches for identifying cancer mutations and genes from next-generation sequencing data in cancer genomes.
Zhongming Zhao is Ingram Professor of Cancer Research at Vanderbilt University and the Chief Bioinformatics Officer at Vanderbilt-Ingram Cancer Center. His recent research activity focuses on precision cancer medicine, especially identifying actionable mutations and driver genes from cancer genomes.
Submitted: 31 May 2015; **Received (in revised form):** 2 July 2015

[9–11]. As of May 2015, nearly 13 million somatic mutations have been uncovered by ICGC (https://icgc.org).

The mutation rate in cancer genomes varies dramatically. It varies as greatly as by 1000 folds among different cancer types [12]. Most solid tumor genomes harbor hundreds of sequence-level genetic alterations. The majority of these alterations are expected to be passenger mutations (mutations that have no direct or indirect effect on a selective growth advantage of tumor cells), while few are driver mutations (mutations that have a selective growth advantage in tumor cells) [13]. Although it is easy to define a 'driver mutation' in a physiological role (conferring a selective tumor growth advantage), systematically identifying driver mutations and the significantly mutated genes (SMGs) that mediate tumor physiological roles from large-scale human cancer genomic data remains a monumental challenge [14, 15]. Here, we used SMGs to generally refer to the terms in literature such as driver genes, candidate cancer genes and mutated genes in cancer. We believed this term is more appropriate in the analysis of large-scale somatic mutations by statistical methods and computational tools.

In this review, we focus on the description of computational approaches and tools in identifying driver mutations and SMGs in cancer using NGS data. To our best understanding, we categorized these approaches into five types based on their major features: (1) mutation frequency based, (2) functional impact based, (3) structural genomics based, (4) network or pathway based and (5) data integration based, as shown in Figure 1. It is important to note that many methods and tools have more than one feature above, but we believed this categorization could best reflect the major features among them. In this review article, we first summarized the major biological resources that are commonly used for the development of these tools. Then, we described the main features of the tools in these five types. Next, we discussed some major challenges on identification of driver mutations or SMGs from large number of somatic mutations in cancer NGS data. These challenges include tumor heterogeneity and purity, tumor sample saturation, somatic mutation calling and potential functional roles of synonymous mutations and noncoding mutations in cancer. Inappropriate dealing with such factors may yield false-positive discoveries from computational approaches. Finally, we highlight several new directions, such as the study of noncoding regulatory mutations through integrated pan-cancer analyses of somatic mutations using functional genomics and whole-genome sequencing (WGS) data.

Data resources for method and tool development and evaluation

NGS data resources

COSMIC is the largest somatic mutation repository database. It offers an important resource for exploring the impact of somatic mutations in cancer [16]. As of March 31, 2015 (version v70), COSMIC contained 3 158 657 coding point mutations. TCGA was jointly funded by the National Cancer Institute (NCI) and National Human Genome Research Institute, National Institutes of Health, USA in 2006 [9]. Since then, the success of

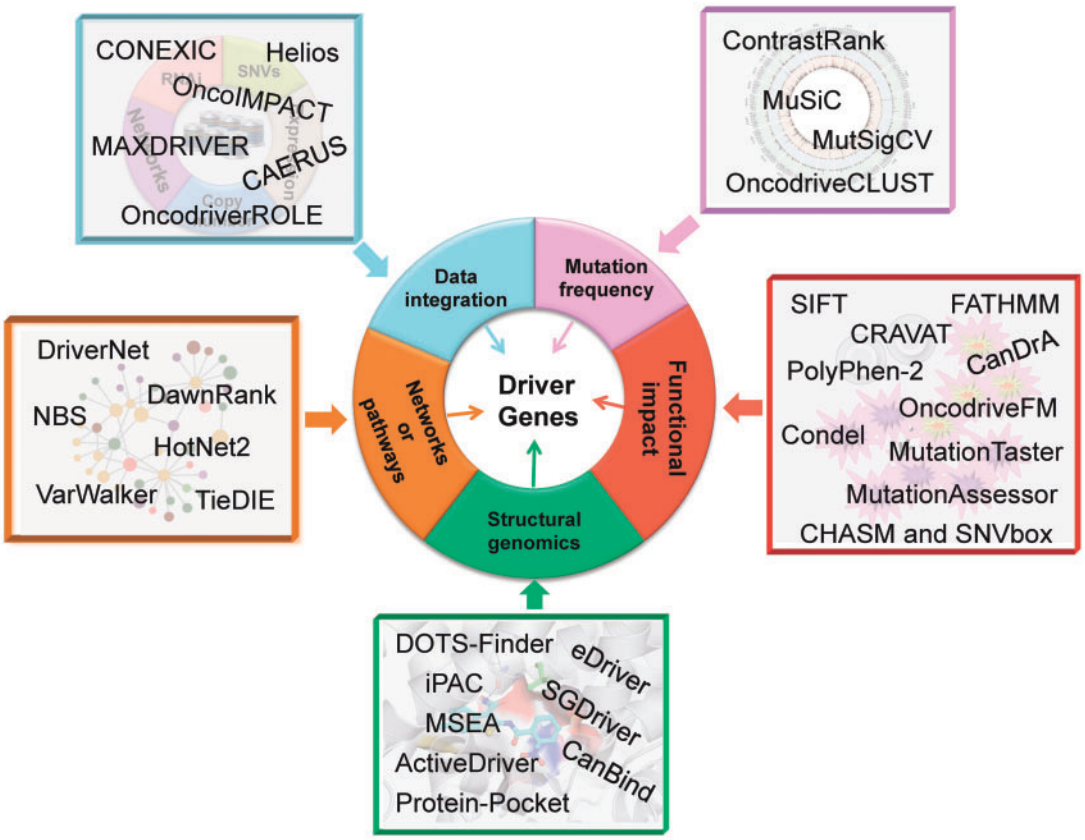

Figure 1. An overview of computational approaches and tools for identifying driver mutations and significantly mutated genes (driver genes) from somatic mutations in cancer genomes. We assigned each method or tool to one of the five categories based on its main feature; however, some tools use the features in more than one category.

the TCGA project has led to characterization of >25 cancer types, providing an important opportunity in evaluating the biological relevance of cancer genomics discovery (http://cancergenome.nih.gov). ICGC aims to systematically catalog the genomic, epigenomic and transcriptomic profiles of >25 000 cancer genomes across 50 different cancer types or subtypes [10, 11]. As of February 2015 (data release 18), ICGC collected >12.9 millions of somatic mutations from 12 807 cancer genomes (including >1000 based on WGS) to provide insights into the landscape of somatic mutations and define the unique genetic signature of an individual tumor (https://dcc.icgc.org). The cBioPortal is a web resource for exploring, visualizing and analyzing multidimensional cancer genomics data [17]. As of April 2015, cBioPortal contained genomic data of 20 958 samples from 89 cancer studies.

Protein's three-dimensional (3D) structure information is often crucial for identifying driver mutations, especially in kinase domains. Thus, such annotation databases will be useful to decipher the biological consequences between protein 3D structures and driver mutations. Cancer3D is a user-friendly database to analyze somatic missense mutations in the context of protein 3D structures [18]. Mosca *et al*. developed dSysMap, a resource for the systematic mapping of disease-related missense mutations on the structurally annotated binary human interactome [19].

Studies of genome sequences have revealed that protein-coding genes account for <2% of the human genome [20]. Recently, several international functional genomics projects have released massive functional genomics data for studying the regulatory roles of noncoding somatic mutations in cancer. These projects include the Encyclopedia of DNA Elements (ENCODE) [21], NIH Roadmap Epigenomics [22] and the functional annotation of the mammalian genome 5 (FANTOM5) [23]. For example, the NIH Roadmap Epigenomics Consortium generated 111 reference human epigenomes using various assays, such as chromatin immunoprecipitation, DNA digestion by DNase I (DNase), RNA expression and DNA methylation [22]. These data sets, along with the previous 16 epigenomes generated by ENCODE project, provide us with valuable opportunities for regulatory annotations of noncoding mutations in cancer [21, 22]. In addition, the Genotype-Tissue Expression project (GTEx) generated large-scale gene expression (e.g. RNA-Seq) and regulation data across multiple types of human tissues. This enables investigators to study the tissue-specific gene regulatory mechanisms that are altered by somatic mutations in cancer [24, 25].

In summary, the aforementioned data resources (Table 1) provide multidomains of data for systematically exploring the genomic, epigenomic and transcriptomic characteristics of tumor samples. These data not only allow for, but also call for, the development of methods and tools that can efficiently detect cancer-related mutations and genes.

Network and pathway data resources

Recently, network-based analyses have been increasingly applied to decipher the biological consequence of somatic mutations in cancer [44]. Much effort has been made to develop comprehensive pathway-related or protein–protein interaction (PPI)-based databases (Table 1). Several curated cell signaling pathway databases, such as WikiPathways [26], KEGG [27], Reactome [28], Pathway Commons [29], and the Pathway Interaction Database (PID) that is carefully curated by US NCI and Nature Publishing Group, [30], have been widely used to explore the functional roles of disease-causing variants [45] or somatic mutations in cancer [46]. In addition, PPI databases deposit experimental and literature-derived PPIs, kinase-substrate-specific phosphorylation and 3D structural PPIs, providing complementary molecular interaction network resources for deciphering functional consequences of somatic mutations in cancer at a molecular network level. Major PPI databases include BioGRID [31], HPRD [32], MINT [33], IntAct [34], STRING [35], PINA [36], PhosphoSitePlus [37], Phospho.ELM [38], PTMcode [39], Interactome3D [41], Instruct [43] and 3did [42]. The details of pathway or PPI databases are provided in Table 1.

Method and computational tools

Mutation frequency-based approaches

Computational approaches commonly define SMGs in cancer by identifying the genes that harbor significantly more mutations than that based on background mutation model in a given cancer type [12]. Table 2 and Figure 1 summarize several mutation frequency-based computational approaches or tools for identifying SMGs. For example, the Mutational Significant in Cancer (MuSiC) is an integrated mutational analysis pipeline that incorporates standardized sequence-based data with clinical data to infer the relationships among mutations, the affected genes and pathways for prioritizing driver mutations and SMGs [47]. Dees *et al*. applied MuSiC to the TCGA ovarian cancer data set and found 12 SMGs in ovarian cancer [47]. ContrastRank prioritizes putative SMGs in cancer by comparing the putative defective rate of each gene in tumor *versus* normal samples and the data from the 1000 Genomes Project [49]. ContrastRank has been found with reasonable accuracy in its evaluation for prioritizing putative SMGs in colon, lung and prostate adenocarcinoma samples.

However, classical mutation frequency-based approaches often have some limitations owing to tumor heterogeneity and other factors [12]. It is expected that the assumption of a constant background mutation model with low mutation frequency will lead to spuriously false-positive discoveries. To solve this problem, some complementary approaches were proposed. OncodriveCLUST is designed to identify SMGs based on the observation that gain-of-function (GoF) mutations in cancer genes predominantly occur at specific protein residues or active domains [48]. OncodriveCLUST primarily uses silent mutations in the coding regions as the background. However, recent studies have showed that silent mutations may play important functional roles in cancer [84]. In addition, the silent mutation-based background model cannot effectively assess the constraints in some genomic regions owing to the low-recurrence of synonymous mutations. Lawrence *et al*. developed MutSigCV, a popular tool for prioritizing SMGs, using gene expression and replication timing information to build a patient-specific background mutation model [12]. They applied MutSigCV to whole-exome sequencing (WES) data in 3083 tumor-normal pairs across 22 cancer types and found 450 SMGs with a false discovery rate of $q < 0.1$.

Empirically observed local mutation frequency obtained from massive amounts of WES data may also influence the accuracy of the mutation frequency-based approaches like MutSigCV and ContrastRank. These limitations may be partially solved by using large-scale WES or WGS data sets from several human genome projects. For instance, the mutation data from the Icelandic genome project [85], the 100 000 Genomes Project

Table 1. Data resources for development and evaluation of computational tools for prioritizing driver mutations and SMGs in cancer

Name	Brief description	Web site	Ref.
Somatic mutation data			
COSMIC	Comprehensive resources of somatic mutations.	http://cancer.sanger.ac.uk/cosmic	[16]
TCGA		http://cancergenome.nih.gov	[9]
ICGC		https://icgc.org	[11]
cBioPortal		http://www.cbioportal.org	[17]
Cancer3D	Resources for functional roles of somatic mutations through protein 3D structures.	http://www.cancer3d.org	[18]
dSysMap		http://dsysmap.irbbarcelona.org	[19]
ENCODE	Comprehensive resources of functional genomics data.	https://www.encodeproject.org	[21]
NIH Epigenome Roadmap		http://www.roadmapepigenomics.org	[22]
FANTOM5		http://fantom.gsc.riken.jp/5/	[23]
GTEx	A atlas of the tissue-specific gene expression and regulation.	http://www.gtexportal.org/	[25]
Pathway annotations			
WikiPathways	Manually curated biological networks and pathways.	http://www.wikipathways.org/	[26]
KEGG		http://www.genome.jp/kegg/	[27]
Reactome		http://www.reactome.org	[28]
Pathway Common		http://www.pathwaycommons.org/	[29]
PID		http://pid.nci.nih.gov	[30]
PPIs			
BioGRID	Repository for PPIs.	http://thebiogrid.org	[31]
HPRD	Manually curated PPIs.	http://www.hprd.org	[32]
MINT		http://mint.bio.uniroma2.it/mint/	[33]
IntAct		http://www.ebi.ac.uk/intact/	[34]
STRING		http://string-db.org	[35]
PINA		http://cbg.garvan.unsw.edu.au/pina/	[36]
PhosphoSitePlus	Manually curated kinase–substrate interactions with specific phosphorylation sites.	http://www.phosphosite.org/	[37]
Phospho.ELM		http://phospho.elm.eu.org	[38]
PTMcode		http://ptmcode.embl.de	[39]
KinomeNetworkX		bioinfo.mc.vanderbilt.edu/kinomenetworkX/	[40]
Interactome3D	Manually curated protein–protein 3D interactions.	http://interactome3d.irbbarcelona.org	[41]
3did		http://3did.irbbarcelona.org	[42]
Instruct		http://instruct.yulab.org	[43]

from the Genomics England [86], the Human Longevity sequencing initiative [86] and the Exome Sequencing Project [87] may provide additional NGS-based mutation information for building more reliable background mutation models, which may not be well represented in the 1000 Genomes Project data set [88].

Functional impact-based approach

In response to the large volume of mutations being generated from massively parallel sequencing projects, it is urgently needed to find highly efficient ways to prioritize driver mutations that can be further selected for experimental validation and clinical applications. Computational approaches provide us with a fast and inexpensive way to define functional annotation and evaluate the functional impact of mutations. These methods and tools could theoretically be used to help investigators select putative driver mutations that would merit further experimental validation or have potential for clinical applications [89]. If the approaches work well, it will save huge amounts of work for laboratory and physician scientists, thus dramatically promoting translational medicine.

Table 2 and Figure 1 describe major computational approaches or tools for characterizing functional impact of mutations. Most of these tools were developed in the past several years, reflecting the strong need of such tools in the field. The Sorting Intolerant from Tolerant (SIFT) is an algorithm that predicts the potential impact of amino acid substitutions on protein functions based on the degree of conservation of amino acids in sequence alignments derived from the closely related sequences [50, 51]. So far, SIFT has become one of the standard tools for characterizing functional impacts of missense mutations. Polymorphism Phenotyping v2 (PolyPhen-2) is a software tool that predicts the functional impact of protein sequence variants by an integration of eight sequence-based and three structure-based features [52]. It is commonly used in conjunction with SIFT to improve the accuracy. MutationAssessor is a web server that uses a novel functional impact score for characterizing amino-acid residue mutations. It applies combinatorial entropy formalism to define evolutionary conservation patterns that are derived from aligned families and subfamilies of sequence homologs within and between species [54]. Of note, the application of these three methods is limited to nonsynonymous SNVs only.

There are many approaches by integrating multiple-domain information to train machine learning-based models for predicting the functional impact of SNVs. OncodriveFM is a specific approach to identify lowly recurrent candidate SMGs using functional impact features derived from SIFT, PolyPhen-2 and MutationAssessor [57]. MutationTaster is a web-based application for rapid evaluation of disease-causing functional effects of DNA-sequence alterations using information from evolutionary conservation, splice-site changes and loss of protein features or changes [53]. However, MutationTaster cannot evaluate

Table 2. Summary of computational approaches and tools for identifying driver mutations and SMGs in cancer genomes

Name	Brief description	Inventor institute	Year	Ref.
Mutation frequency based				
MuSiC	A pipeline for determining the mutational significance in cancer.	Washington University	2012	[47]
MutSigCV	An integrative approach that corrects for variants using patient-specific mutation frequency and spectrum, and gene-specific background mutation model derived from gene expression and replication timing information.	The Broad Institute	2013	[12]
OncodriveCLUST	Identifying genes with a significant bias toward mutation clustering in specific protein regions using silent mutations as a background mutation model.	Universitat Pompeu Fabra, Spain	2013	[48]
ContrastRank	A method based on estimating the putative defective rate of each gene in tumor against normal and samples from the 1000 Genomes Project data.	University of Alabama at Birmingham	2014	[49]
Functional impact based				
SIFT	A popular tool for predicting the biological effect of missense variations by using protein sequence homology.	J. Craig Venter Institute	2009/2012	[50, 51]
PolyPhen-2	A popular tool using eight sequence-based and three structure-based predictive features to build naïve Bayes classifiers for predicting the functional impacts of protein sequence variants.	Harvard Medical School	2010	[52]
MutationTaster	A web-based tool comprising evolutionary conservation and splice-site change information for predicting the functional impacts of DNA sequencing alterations. It limits on alterations spanning an intron-exon border or indels at most 12 base pairs.	Charite-Universitatsmedizin Berlin, Germany	2010	[53]
MutationAssessor	Predicting functional impact scores based on evolutionary conservation patterns.	Memorial Sloan-Kettering Cancer Center	2011	[54]
Condel	A consensus deleteriousness score for assessing the functional impact of missense mutations.	Universitat Pompeu Fabra, Spain	2011	[55]
CHASM and SNVbox	Python and C++ programs for prioritizing cancer-related mutations using their tumorigenic impact.	Johns Hopkins University	2011	[56]
OncodriveFM	An approach based on functional impact bias using three well-known methods.	Universitat Pompeu Fabra, Spain	2012	[57]
PROVEAN	A tool for predicting the functional effects of SNV and in-frame insertions and deletions.	J. Craig Venter Institute	2012	[58]
CanDrA	A machine learning-based tool based on a set of 95 structural and evolutionary features.	The University of Texas MD Anderson Cancer Center	2013	[59]
FATHMM	A Hidden Markov model-based tool for functional analysis of driver mutations.	University of Bristol, UK	2013	[60]
CRAVAT	A web-based toolkit for prioritizing missense mutations related to tumorigenesis.	Johns Hopkins University	2013	[61]
Structural genomics based				
iPAC	An algorithm using protein 3D structure information for predicting SMGs.	Yale University	2013	[62]
ActiveDriver	An approach for predicting SMGs harboring driver mutations significantly altering protein phosphorylation sites.	University of Toronto, Canada	2013	[63]
CanBind	A computational pipeline for predicting SMGs using protein–ligand binding site information.	Princeton University	2014	[64]
MSEA	MSEA for predicting SMGs based on mutation hotspot patterns on protein domains or any genomic regions.	Vanderbilt University	2014	[65]

(continued)

Table 2. Continued

Name	Brief description	Inventor institute	Year	Ref.
eDriver	A method for predicting SMGs based on the mutation bias between protein domain or intrinsically disordered regions and other regions.	Sanford-Burnham Medical Research Institute	2014	[66]
Protein-Pocket	A method for prioritizing SMGs harboring enriched mutations in its protein pocket regions.	Vanderbilt University	2014	[67]
SGDriver	A method for prioritizing SMGs and druggable mutations in protein–ligand binding sites using a Bayes inference statistical framework.	Vanderbilt University	2015	[68]
Network or pathway based				
PARADIGM	A novel method for detecting consistent pathways in cancers by incorporating patient-specific genetic data into carefully curated NCI pathways.	University of California, Santa Cruz	2010	[69]
PARADIGM-SHIFT	A method for prioritizing downstream pathways altered by a mutation in cancer using a belief-propagation algorithm.	University of California, Santa Cruz	2012	[70]
Personalized Pathway Enrichment Map	A personalized pathway enrichment method for identifying putative cancer genes and pathways from each individual genome.	Vanderbilt University	2012	[71]
DriverNet	A computational framework for identifying driver mutations by estimating their effect on mRNA expression networks.	British Columbia Cancer Agency, Canada	2012	[72]
TieDIE	A network diffusion approach for identifying cancer mutated subnetworks.	University of California, Santa Cruz	2013	[73]
NBS	A somatic mutation network-based approach for stratifying tumor mutations.	University of California, San Diego	2013	[74]
DawnRank	A tool for prioritizing SMGs in a single patient based on the PageRank algorithm.	University of Illinois at Urbana-Champaign	2014	[75]
VarWalker	A novel personalized mutation network analysis approach for prioritizing SMGs.	Vanderbilt University	2014	[76]
HotNet2	A new algorithm uses an insulated heat diffusing process to overcome the limitations of existing single-gene, pathway and network approaches for detecting mutated subnetworks in cancer.	Brown University	2015	[46]
Data integration based				
CONEXIC	A computational framework that integrates copy number variants and gene expression changes for prioritizing SMGs.	Columbia University	2010	[77]
CAERUS	An integrative approach for predicting SMGs using protein structural information, protein networks, gene expression and mutation data.	University of British Columbia, Canada	2011	[78]
MAXDRIVER	An integrated approach for predicting SMGs using the data from copy number variant regions of cancer genomes.	Chinese Academy of Sciences, China	2013	[79]
Helios	An algorithm predicts SMGs by integrating genomic and functional RNAi screening data from primary tumors.	Columbia University	2014	[80]
DOTS-Finder	A functional and frequentist-based tool for predicting SMGs in cancer.	Istituto Italiano di Technologia, Italy	2014	[81]
OncodriverROLE	A machine learning-based approach classifies SMGs into LoF and GoF.	Universitat Pompeu Fabra, Spain	2014	[82]
OncoIMPACT	An integrative framework for prioritizing SMGs based on their phenotypic impacts.	Genome Institute of Singapore, Singapore	2015	[83]

alterations spanning an intron-exon border or insertions/deletions (indels) >12 base pairs. CHASM is a method that predicts the functional significance of somatic missense mutations using a Random Forest classifier trained with 49 predictive features [56, 90]. FATHMM is a Hidden Markov model-based software that distinguishes cancer-associated amino acid substitutions from passenger mutations by integrating the alignment of homologous sequences and conserved protein domain

information [60]. CRAVAT is a web-based toolkit for prioritizing driver mutations and SMGs using CHASM and SNVbox [61]. CanDrA is a supporting vector machine (SVM)-based tool for prioritizing SMGs by incorporating 95 structural and evolutionary features generated by >10 different functional prediction algorithms [59].

In summary, most of the above tools (Table 2) primarily rely on estimating the deleterious effects of SNVs via evaluating amino acid conservation at the corresponding positions. There are some limitations in the clinical settings for those approaches and tools. For example, most tools are machine learning based, such as PolyPhen-2, CHASM and CanDrA. Building a gold-standard positive data set (experimentally validated functional mutations) is always a difficult task for machine learning-based tool development. Furthermore, selection of high-quality negative data set (nonfunctional mutations) is another big challenge owing to few negative results published in current literature. Martelotto *et al.* systematically evaluated 15 mutation impact prediction algorithms using experimentally validated cancer missense mutations [89]. They found that no algorithm was able to accurately predict SNVs that should be taken forward for further experimental or clinical testing, while combination of different tools could modestly improve accuracy and significantly reduce false-negative predictions. Gonzalez-Perez and Lopez-Bigas presented a consensus deleteriousness (Condel) score using a weighted average of the normalized scores derived from five complementary tools [55]. Condel score outperforms a single approach in predicting functional impact of missense mutations. Put together, taking advantage of complementarity of different approaches or tools is a useful strategy to improve the predictions of functional impacts of somatic mutations in cancer.

Structural genomics-based approach

Owing to the rapid advancement of structural genomic technologies, such as nuclear magnetic resonance and X-ray, large-scale, high-quality protein 3D structure data have been generated and carefully curated, and such data are made available in the databases like Protein Data Bank (PDB) [91]. Because the mutations at the structurally important sites are more likely linked to disease or drug targets, we have witnessed many computational methods and tools that are recently developed by using such features. The functional features implemented in such methods and tools include specific protein regions (e.g. protein domain, intrinsically disordered regions), posttranslational modification (PTM) sites (e.g. phosphorylation sites), protein pockets and protein–ligand binding sites.

MSEA, mutation set enrichment analysis, was implemented by two novel modules (MSEA-domain and MSEA-clust) to predict putative SMGs. MSEA-domain is based on mutation hotspot patterns on protein domains, while MSEA-clust is to screen mutation hotspot regions by scanning any genomic regions [65]. Yang *et al.* systematically investigated the mutation frequency distribution in protein domains using thousands of tumor genomes across 21 cancer types [92]. They observed the expected patterns that protein domain mutations (e.g. both known and new cancer hotspot mutations in kinase domains) are recurrently mutated in both oncogenes and tumor suppressor genes.

In addition to protein domain information, protein PTM sites (e.g. phosphorylation sites) play essential roles in regulating cellular signaling pathways [40]. Considering the large number of PTM sites being identified by proteomic approaches, there is good rationale to develop new approaches or tools for pinpointing putative SMG products that harbor mutations leading to significant PTM site changes. Among these methods, Cheng *et al.* constructed a global kinase–substrate interaction network containing 7346 pairs connecting 379 kinases and 1961 substrates harboring 36 576 phosphorylation sites. Based on the global network analysis, they found a high anticancer drug resistance risk that might be caused by the distinct network centrality of kinases owing to feedback or crosstalk mechanisms within cellular networks [40]. ActiveDriver [63] is an approach to search for SMGs based on the hypothesis that cancer driver mutations are more likely to alter protein's phosphorylation sites [93]. ActiveDriver was demonstrated to successfully identify dozens of SMGs (e.g. *ASF1*, *FLBN* and *GRM1*) based on the cancer genomics data in 800 cancer genomes across eight cancer types. Furthermore, the same group applied ActiveDriver to analyze known phosphorylation sites mutated by SNVs in ~3200 cancer genomes across 12 cancer types from the TCGA pan-cancer data set, and they identified 150 SMGs by a gene-centric analysis [94]. So far, ActiveDriver only analyzes missense point mutations; other types of mutations like truncated mutations have not yet been implemented. In addition, ActiveDriver uses all literature-reported phosphorylation sites from different cell or tissue types as a mixture training set. However, while some phosphorylation sites are cancer specific, others may not. Massive high-quality cancer-specific phosphoproteomic data, such as that generated by Clinical Proteomic Tumor Analysis Consortium, may provide new opportunities in this research direction [95].

Understanding the biological consequences of somatic mutations at the protein structural and functional levels is a promising research field. Several studies have demonstrated the close relationship between protein structures and their function altered by cancer-related missense mutations [96, 97]. Vuong *et al.* presented a protein pocket-based computational pipeline to study the functional consequences of somatic mutations in cancer [67]. Protein pocket regions are where small molecules and drugs bind with the protein; thus, mutations at such sites are likely to alter protein function, leading to disease such as cancer. By mapping 1.2 million somatic mutations across 36 cancer types from the COSMIC database and TCGA onto the computationally predicted protein pocket regions for >5000 protein 3D structures, they found that gene products (proteins) harboring missense mutations located in their protein pocket regions were more likely to be cancer proteins. Furthermore, they identified four putative cancer genes (*RWDD1*, *NCF1*, *PLEK* and *VAV3*), whose expression levels were associated with overall poor survival rates in lung, melanoma or colorectal cancer patients. Furthermore, based on the close relationship between somatic mutations and protein 3D structures, Zhao *et al.* developed a protein structural genomics-based approach, SGDriver, to prioritize SMG products and druggable mutations [68]. SGDriver incorporates the somatic missense mutations into the protein–ligand binding sites using a Bayes inference statistical framework. They applied SGDriver to analyze missense mutations in 4997 cancer genomes across 16 cancer types from TCGA. SGDriver identified ~300 proteins (adjusted P-value < 0.05) harboring mutations that were significantly enriched at protein–ligand binding sites through both pan-cancer and individual cancer analyses. One utility of SGDriver is to identify promising druggable mutations that can be further studied in the emerging field of precision cancer medicine.

CanBind is a computational approach to prioritize SMGs that harbor enriched mutations by altering their nucleic acid, small molecules and ion or peptide binding sites [64]. iPAC, namely Identification of Protein Amino acid Clustering, is an algorithm

that characterizes nonrandom somatic mutations in protein by using its 3D structure information [62]. Evaluation of iPAC using the data from the PDB and COSMIC databases indicated that it could identify both well-known cancer driver genes (e.g. *KRAS* and *PIK3CA*) and new cancer driver genes (e.g. *EIF2AK2*). Another tool, eDriver, was developed to prioritize SMGs based on the comparison of the internal distribution of somatic missense mutations between the protein's domains or intrinsically disordered regions and other domains of the same protein [66].

Collectively, development of new computational approaches or tools that can efficiently prioritize SMGs in cancer according to their functional effects on protein 3D structures will provide us with unprecedented opportunities for clinical applications of the cancer genomic data. While the demand on such tools is strong, so far, most approaches have focused on only SNVs, rather than all types of mutations, including indels and gene fusions [98]. Previous observations suggested that truncated mutations (e.g. nonsense mutation, out-of-frame indels and splicing) also play critical roles in cancer molecular networks [99]. The other challenge of structural genomics-based approaches is the limited number of proteins having high-resolution 3D structures available (~15% human proteins having known 3D structures) when compared with the whole human genome [91].

Network- or pathway-based approach

Cells consist of various molecular structures that form complex, plastic and dynamic networks [100, 101]. Under the molecular network framework, a genetic aberration may cause network architectural change by affecting or removing a node or its connection within the network or by changing the biochemical properties of a node (e.g. protein) [99, 102, 103]. The abundance of cancer genomics data from NGS studies provides biologists with huge opportunities to gain a network- or systems-level understanding of tumor initiation and progression. One of the major findings from TCGA project is that cancer is a complex disease, with many changes altered at the network and pathway levels, not simply a point mutation. Therefore, there has been strong interest in prioritizing driver mutations and SMGs using the network- and pathway-based approaches.

PARADIGM is a novel method for detecting consistent pathways in cancer by incorporating patient-specific genetic data (CNVs and gene expression) into carefully curated NCI pathways [69]. PARADIGM outperforms a previous method in identifying cancer-related pathways based on evaluation of both breast cancer and glioblastoma multiforme data sets. The authors further expanded PARADIGM to PARADIGM-SHIFT, which infers downstream pathways altered by a given mutation in cancer by incorporating somatic mutations, CNVs and gene expression into an integrated pathway using a belief-propagation algorithm [70]. Importantly, PARADIGM-SHIFT could identify potential functional effects such as neutral, loss-of-function (LoF) and GoF for a given mutation in individual cancer patient. In addition, Jia and Zhao developed a personalized pathway enrichment method for identifying putative cancer genes and pathways from each individual genome using NGS-based mutation data [71].

TieDIE uses a network diffusion approach to predict gene expression changes altered by genomic alteration in cancer [73]. Specifically, TieDIE identified a cancer-specific subnetwork by incorporating genomic and transcriptomic data into networks from PPIs, computationally predicted transcription factor-to-target connections and manually curated interactions from literature. Comparing with other approaches, TieDIE can identify pathways that are related to the downstream transcriptional changes altered by somatic alterations in cancer.

DriverNet is a computational framework to identify candidate driver mutations by modeling their effect on mRNA expression networks [72]. A useful feature of DriverNet is to identify rare driver mutations mediating oncogenic and metabolic networks. VarWalker is the first personalized network tool by using somatic mutations from individual genome to prioritize putative SMGs. It incorporates large-scale cancer genomic data into PPI network using the random walk with restart algorithm [76]. The unique features of VarWalker include its use of the somatic mutations from individual cancer genomes and adjustment of the gene length biases by resampling mutations from each individual genome. Network-based stratification (NBS) is a novel network-based approach that stratifies cancer subtypes based on the profiles of the somatic mutations presented in individual tumor [74]. A recent tool, DawnRank, is a computational approach to prioritize SMGs on an individual patient using a PageRank algorithm [75]. HotNet detects significantly mutated pathways in cancer based on the context of a genome-scale gene interaction network using a network diffusion approach [104]. However, most of aforementioned network-based approaches, such as VarWalker, NBS, DawnRank and HotNet, are proposed based on network propagation processes (e.g. random walk with restart algorithm or PageRank). Hub genes (the nodes with high degree in the network) are often yielded with the highly predicted scores. Thus, development of novel algorithms by investigating the significance of the predicted SMGs regardless of network topology biases would be more appropriate. The same group of HotNet further developed HotNet2, a new algorithm for detecting mutated subnetworks in cancer by using an insulated heat diffusing process to overcome the limitations of existing single-gene, pathway and network approaches [46]. They identified 16 significantly mutated subnetworks comprising well-known cancer signaling pathways during pan-cancer analysis. In addition, HotNet2 can identify subnetworks containing genes that are rarely mutated in both individual cancer type and pan-cancer data sets.

Although network or pathway-based approaches have been successfully used for studying the biological consequences of somatic mutations in cancer, these approaches have limitations too. First, current PPI networks detected by high-throughput technologies may only cover 20–30% of all potential pairwise PPIs in the human cells [105, 106], suggesting that the current human interactome map might be up to 80% incomplete [107]. Second, the network is often error prone because it is built based on large-scale experimental data, computational prediction data or both. Such data are always mixed, rather being cell type specific, tissue specific or condition specific. Third, structural variants, noncoding variants, gene expression and methylation data are often not considered in the majority of the aforementioned approaches. Thus, developing an integrative framework by incorporating somatic mutations, structural variations, gene expression and methylation into the improved knowledge of the human interactome would provide a more comprehensive catalog of significantly mutated networks or pathways in cancer.

Data integration-based approach

Cancer 'panomics' data, including somatic mutations, transcriptome, methylation and proteomics profiles of a patient's tumor and matched normal tissue generated from

high-throughput technologies, enable investigators to have systematic investigation of SMGs and driver mutations for precision cancer medicine [108, 109]. The details of several data integration-based approaches are provided in Table 2. Diver Oncogene and Tumor Suppressor (DOTS)-Finder identifies SMGs in cancer by an integration of three aspects of a mutated gene: mutation pattern (the genome position of the observed mutations), the functional effect of mutations on gene product and mutation frequency [81]. An important feature of DOTS-Finder is that it can predict SMGs as specific as tumor suppressor genes or oncogenes.

In addition to SNVs, structural variants, such as deletions, duplications and CNVs, often alter DNA sequences. For example, as much as 15% of the human genomes falls into CNV regions [110]. Wong et al. proposed a novel computational pipeline, SVMerge, for detection of structural variants and breakpoints by integrating several existing structural variant calling algorithms and local assembly information [111]. Detailed descriptions of structural variant detection can be found in several recent review articles [15, 112]. Development of an integrated approach to prioritize driver mutations or SMGs by using structural variant data such as CNVs is a promising direction. CONEXIC identifies driver mutations related to cancer progression by integrating CNVs (amplifications and deletions) and gene expression data from matched tumor-normal samples [77]. They have successfully identified known SMGs and multiple tumor dependences (e.g. *TBC1D16* and *RAB27A*) in melanoma via CONEXIC. The same group further developed Helios, an algorithm that identifies SMGs within large recurrently amplified regions of DNA by incorporating cancer genomics data into data from functional RNA interference (RNAi) screening studies [80]. They pinpointed a set of candidate SMGs in breast cancer and further experimentally validated that RSF-1-mediated tumorigenesis and metastasis *in vivo*. Helios assesses candidate drivers by a transfer learning technique that does not require any prior list of driver genes. Thus, it does not suffer from the prior knowledge biases.

MAXDRIVER identifies putative SMGs using several optimization strategies to construct a heterogeneous network through an integration of a fused gene functional similarity network and an existing gene–cancer association network [79]. However, incompleteness and data noise of currently known gene–cancer associations may yield false-positive discoveries in MAXDRIVER. OncodriverROLE is a machine learning-based approach to classify SMGs into LoF and activated (Act) genes [82]. Construction of the gold-standard positive and negative LoF and Act gene sets is a big challenge for machine learning-based classification model that is implemented in OncodriverROLE. OncoIMPACT is a data integration framework for predicting patient-specific SMGs based on their phenotypic impacts [83]. OncoIMPACT can predict patient-specific drivers.

The existing computational tools are often developed based on different biological hypotheses. Combining two or more methods by their complementary biological hypotheses may improve the prediction accuracy of each individual tool. With this rationale, Tamborero *et al.* systematically identified 291 high-confidence SMGs using cancer genomics data in 3205 tumors across 12 different cancer types from TCGA using a combination of four complementary methods, including MuSiC, OncodriveFM, OncodriveCLUST and ActiveDriver [113]. They demonstrated that the combinations of different approaches using their complementary hypothesis could outperform each individual method.

Challenges on current approaches

Tumor heterogeneity and sample saturation

So far, the widely used computational methods are designed to identify SMGs that have more mutations than the expected based on background mutation model. However, tumor intra-heterogeneity often leads to false-positive discoveries [113–115]. A subtype of colorectal cancer (namely, stem/serrated/mesenchymal transcriptional subtype) was reported to be driven by stromal cells rather than tumor cells [116]. Batile *et al.* drew a similar conclusion that a poor-outcome colorectal cancer subtype was driven by the genes expressed in tumor-associated stromal cells [117]. Yoshihara *et al.* described ESTIMATE, a computational method to infer the fraction of stromal and immune cells in tumor samples using gene expression signatures [118]. Thus, evaluation of tumor purity and intra-heterogeneity is a critical part when we distinguish SMGs and driver mutations from passenger mutations in cancer. In addition, tumor sample saturation also limits the creation of a comprehensive catalog of SMGs based on the currently limited sequencing data. For instance, a recent mathematical model was proposed to estimate the minimum number of samples for detecting SMGs [119]. The mathematical analysis revealed that building a comprehensive catalog of SMGs would need to sequence an average of nearly 2000 tumors for each of the at least 50 cancer types.

Although some cancer driver genes are mutated at high frequencies (>20%), most cancer mutations occur at intermediate frequencies (2–20%) or lower [119]. An analysis based on 183 lung adenocarcinoma samples suggested that 15% of patients lacked even a single mutation affecting well-known cancer genes (e.g. *EGFR*, *KRAS* and *ALK*) [120]. Vogelstein *et al.* estimated that a typical tumor contained two to eight driver mutations; the remaining mutations are passengers that do not contribute to the tumor growth advantage [13]. Tomasetti *et al.* suggested that only three sequential mutations are required to develop colon and lung cancers based on genome-wide sequencing data [121]. Identifying the exact number of driver mutations in a typical tumor would be helpful for our understanding of tumor initiation and progression.

High quality of sequencing data is important when predicting SMGs and driver mutations using computational tools. However, existing methods typically miss low allele frequency mutations that occur in only a small subset of the sequenced cells owing to tumor heterogeneity [122]. In addition, the coverage and false discovery rate of sequencing data from WES are two other critical issues for pinpointing driver mutations and SMGs. A recent study showed that WGS is more powerful than WES for detecting potential disease-causing mutations within WES regions, particularly those due to SNVs [123]. This observation only focused on disease-causing mutations in six unrelated patients. The accuracy of the detection of somatic mutations of tumor-normal matched samples using WGS and WES may be different [86]. Although WGS is currently more expensive than WES, its cost is expected to decrease dramatically, and coverage in WGS is expected to increase as well. Another issue is the false-negative discovery—those cancer mutations that could not easily be detected by NGS technologies. The advances in high-throughput NGS technologies, both second-generation and third-generation, will help solve this problem. Therefore, reliability of computational tools will improve with higher quality of data derived from cancer genomes in the near future.

In summary, the number of discovered SMGs has been steadily increasing with the accumulation of high-quality sequencing data [119]. Importantly and timely, the Precision

Medicine Initiative launched by US President Obama earlier this year with a $215 million initial investment will catalog mutations in 1 million patients including cancer patients; this will dramatically accelerate the catalog of cancer driver mutations and SMGs by using the flood of sequencing data and computational approaches in the near future [124].

The accuracy of somatic mutation calling

Somatic mutation calling in cancer genomes is an important prerequisite step for the identification of driver mutations or SMGs [125]. There are many tools available for detection of SNVs, indels, CNVs, gene fusions and large structural variants [15]. Genome Analysis Toolkit is a broad and widely used toolkit developed by the Broad Institute for NGS data processing and variant discovery in the 1000 Genomes Project and TCGA [126]. Calling somatic mutations is a harder problem than calling germline variants because of high variability such as tumor heterogeneity and tumor subclonality [127]. Cibuskis *et al.* proposed a Bayesian classifier-based approach, namely MuTect, for detecting somatic mutations having low allele fractions [122]. They showed that MuTect has higher sensitivity with similar specificity in comparison with several previous approaches. In a recent study, O'Brien *et al.* suggested the inconsistency and features for SNV detection in WES versus transcriptome sequencing data [128]. They found a low overlap of ~14% SNVs called in WES and RNA-Seq data from 27 tumor-normal pairs in lung cancer as a case study. In addition to the evaluation of SNV calling in MuTect, Wang *et al.* systematically evaluated the performance of six tools (EBCall, JointSNVMix, MuTect, SomaticSniper, Strelka and VarScan 2) for somatic point mutation detection based on real WGS and WES data as well as the simulation data [129]. The found that MuTect detected most low allelic-fraction somatic point mutations, while VarScan 2 identified more somatic point mutations than other tools, suggesting a potential room for improvement of somatic mutation detection. In addition, the ICGC-TCGA DREAM Somatic Mutation Calling Challenge benchmark evaluated 248 submissions from 21 research teams [130]. They found that different algorithms exhibit characteristic error profiles and false-positive discoveries: recall ranged from 0.559 to 0.994, F-score from 0.046 to 0.975 and precision from 0.101 to 0.997. The authors suggested that robust ensemble learners might eventually improve the accuracy of somatic mutation detection [130]. We believe that combining different somatic mutation calling approaches or tools would enhance the somatic mutation calling and, thus, benefit the identification of driver mutations and SMGs in cancer genomes.

Functional synonymous mutations in cancer

Most recently, efforts have been made in identifying functional mutations from synonymous variants, those mutations in coding regions but do not change amino acids or protein sequences. Supek *et al.* performed a large-scale cancer genomics analysis using 3851 cancer exomes from TCGA and COSMIC [84]. They showed that synonymous mutations also contributed to human cancer. For instance, they found that synonymous mutations recurrently mutated in oncogenes. One possible mechanism of recurrent mutations in oncogenes contributing to cancer is that synonymous mutations recurrently target exonic splicing motifs and cause abnormal oncogene splicing. For instance, they showed that recurrent synonymous mutations in *TP53* are adjacent to splice sites and inactivate splice sites. Thus, consideration of functional roles of synonymous mutations as part of further development of computational approaches may both increase the use of the existing data and find additional cancer SMGs with different molecular mechanisms. This direction is relatively new; however, the initial findings are promising. And efforts on using other 'nonfunctional' mutations, like those in noncoding regions, in computational analysis, are ongoing as well.

Pan-cancer analysis

By the end of 2015, the TCGA research network will have achieved the ambitious goal of analyzing the genomic, epigenomic and gene expression profiles of >10 000 specimens from >25 different tumor types [131]. As of late 2014, TCGA scientists had nearly completed sequencing the protein-coding regions for most tumor types and had completed WGS of >1000 tumor samples. During the past several years, many investigators have successfully applied the TCGA data for their own research projects or clinical applications, resulting in publishing over 2700 peer-review articles, many of which appeared in high profiling journals (http://cancergenome.nih.gov/publications). Now, TCGA expanded to its pan-cancer analysis. In August 2014, TCGA pan-cancer project of a multiplatform analysis of 12 cancer types was published in Cell [132]. This pan-cancer analysis showed that some tumors were more likely to be genetically and molecularly similar owing to the type of their arising cells rather than from the tissue site of origin [132]. To expand the TCGA's pan-cancer analysis, two new projects (PanCanAtlas and Pan-Cancer Analysis of Whole Genomes) are currently underway. Put together, a systematic pan-cancer analysis based on panomics data generated by NGS and other platforms will allow investigators to have more clinical relevance discoveries across different cancer types [131, 133].

Perspectives

Several future directions are attracting more and more attention. First, most cancer genomic studies have focused on protein-coding regions of the genome. However, functional genomics projects, such as the ENCODE [21], NIH Roadmap Epigenomics [22] and FANTOM5 [23], are elucidating the functional elements of the human genome, including promoter and enhancer regions. These projects will provide researchers with huge opportunities for deciphering the regulatory landscape of somatic mutations across the whole genome covering both coding and noncoding regions. For example, two independent groups performed genome-wide analysis of noncoding regulatory mutations in cancer, and found that *TERT* promoter recurrent mutations play crucial roles in multiple cancer types [134, 135]. Second, high-throughput functional screening technologies, such as RNAi and CRISPR-Cas9, would provide us with new innovative strategies for identifying SMGs and driver mutations with high accuracy. For instance, Schramek *et al.* found Myosin IIa as a tumor suppressor of squamous cell carcinoma using *in vivo* RNAi screens [136]. Konermann *et al.* proposed a structure-guided engineering of a CRISPR-Cas9 complex to identify candidate genes mediating the resistance to a BRAF inhibitor in both cell lines and patient-derived samples [137]. Taken together, an integration of genomics data with other data from functional screens (e.g. RNAi and CRISPR-Cas9) would create a new promising research field for identifying new SMGs and driver mutations in cancer.

Third, circulating tumor DNA (ctDNA) is a promising biomarker for noninvasive monitor of the tumor burden for diagnosis, prognosis and treatment selection [138–140]. Currently, ctDNA detecting methods do not have high sensitivity, limiting their broad clinical applicability. In addition, ctDNA detection approach needs to overcome the limitation in isolating rare circulating tumor cells and sequencing low volume of circulating cell-free DNA materials [140]. Recently, Newman *et al.* proposed a cancer personalized profiling via deep sequencing approach (CAPP-Seq) for detecting ctDNA by combining optimized library preparation methods and a multiphase bioinformatics approach [138]. They demonstrated that CAPP-Seq could detect ctDNA in 100% of non-small cell lung cancer patients with stages II–IV. Thus, an approach effectively combining both technologies (e.g. NGS) and bioinformatics methods is promising for enhancing the ctDNA detection, and this will provide new ways for prognosis and precision treatment of cancer [140, 141]. Forth, development of a new integrated approach using single-cell genomics data would reduce the false-positive discoveries caused by tumor purity and tumor intra-heterogeneity [142–145]. Finally, the genomics landscape of individual tumors enables systematic investigation of antitumor immunotherapeutic responses driven by somatic mutations [146–148]. For example, Rooney *et al.* quantified the cytolytic activities of the local immune infiltrate across 18 cancer types using large-scale genomic data sets from solid tumor biopsies [149]. Furthermore, the large volume of genomic alterations has made it possible to examine the immune response to patient-specific neoantigens, advancing development of personalized cancer immunotherapy [150, 151].

Key Points

- Massive amounts of somatic mutation data have been generated from large-scale cancer genome sequencing projects; how to identify driver mutations and significantly mutated genes (SMGs) remains a great challenge.
- This review describes recent development and advances of methods and computational tools for identifying driver mutations and SMGs from whole-exome, whole-genome and whole-transcriptome sequencing data.
- Studying noncoding regulatory mutations through the integrated analysis of functional genomics and whole-genome sequencing data will be a promising research direction for precision cancer medicine.

Acknowledgements

We thank Barbara O'Brien for English polishing an earlier draft of the manuscript. We thank the members in Bioinformatics and Systems Medicine Laboratory for valuable discussion on this topic. We apologize that we cannot include and cite all related studies owing to the limited space of manuscript space.

Funding

This work was partially supported by National Institutes of Health grants (R01LM011177, P50CA095103, P50CA098131 and P30CA068485), The Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation (to Z.Z.) and Ingram Professorship Funds (to Z.Z.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Eisenstein M. Startups use short-read data to expand long-read sequencing market. *Nat Biotechnol* 2015;**33**:433–35.
2. Wong KM, Hudson TJ, McPherson JD. Unraveling the genetics of cancer: genome sequencing and beyond. *Annu Rev Genomics Hum Genet* 2011;**12**:407–30.
3. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 2010;**11**:685–96.
4. Eifert C, Powers RS. From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets. *Nat Rev Cancer* 2012;**12**:572–8.
5. Lovly CM, McDonald NT, Chen H, *et al.* Rationale for co-targeting IGF-1R and ALK in ALK fusion-positive lung cancer. *Nat Med* 2014;**20**:1027–34.
6. Xia J, Jia P, Hutchinson KE, *et al.* A meta-analysis of somatic mutations from next generation sequencing of 241 melanomas: a road map for the study of genes with potential clinical relevance. *Mol Cancer Ther* 2014;**13**:1918–28.
7. Jia P, Jin H, Meador CB, *et al.* Next-generation sequencing of paired tyrosine kinase inhibitor-sensitive and -resistant EGFR mutant lung cancer cell lines identifies spectrum of DNA changes associated with drug resistance. *Genome Res* 2013;**23**:1434–45.
8. Dahlman KB, Xia J, Hutchinson K, *et al.* BRAF(L597) mutations in melanoma are associated with sensitivity to MEK inhibitors. *Cancer Discov* 2012;**2**:791–7.
9. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat Med* 2011;**17**:297–303.
10. Gonzalez-Perez A, Mustonen V, Reva B, *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods* 2013;**10**:723–9.
11. International Cancer Genome C, Hudson TJ, Anderson W, *et al.* International network of cancer genome projects. *Nature* 2010;**464**:993–8.
12. Lawrence MS, Stojanov P, Polak P, *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;**499**:214–8.
13. Vogelstein B, Papadopoulos N, Velculescu VE, *et al.* Cancer genome landscapes. *Science* 2013;**339**:1546–58.
14. Raphael BJ, Dobson JR, Oesper L, *et al.* Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med* 2014;**6**:5.
15. Ding L, Wendl MC, McMichael JF, *et al.* Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet* 2014;**15**:556–70.
16. Forbes SA, Beare D, Gunasekaran P, *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015;**43**:D805–11.
17. Gao J, Aksoy BA, Dogrusoz U, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1.
18. Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res* 2015;**43**:D968–73.
19. Mosca R, Tenorio-Laranga J, Olivella R, *et al.* dSysMap: exploring the edgetic role of disease mutations. *Nat Methods* 2015;**12**:167–8.

20. Venter JC, Adams MD, Myers EW, *et al*. The sequence of the human genome. *Science* 2001;**291**:1304–51.
21. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
22. Roadmap Epigenomics C, Kundaje A, Meuleman W, *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30.
23. Andersson R, Gebhard C, Miguel-Escalada I, *et al*. An atlas of active enhancers across human cell types and tissues. *Nature* 2014;**507**:455–61.
24. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015;**348**:648–60.
25. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;**45**:580–5.
26. Kelder T, van Iersel MP, Hanspers K, *et al*. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res* 2012;**40**:D1301–7.
27. Kanehisa M, Goto S, Sato Y, *et al*. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;**42**:D199–205.
28. Croft D, Mundo AF, Haw R, *et al*. The reactome pathway knowledgebase. *Nucleic Acids Res* 2014;**42**:D472–7.
29. Cerami EG, Gross BE, Demir E, *et al*. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 2011;**39**:D685–90.
30. Schaefer CF, Anthony K, Krupa S, *et al*. PID: the Pathway Interaction Database. *Nucleic Acids Res* 2009;**37**:D674–9.
31. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, *et al*. The BioGRID interaction database: 2013 update. *Nucleic Acids Res* 2013;**41**:D816–23.
32. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human Protein Reference Database–2009 update. *Nucleic Acids Res* 2009;**37**:D767–72.
33. Licata L, Briganti L, Peluso D, *et al*. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012;**40**:D857–61.
34. Kerrien S, Aranda B, Breuza L, *et al*. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 2012;**40**:D841–6.
35. Franceschini A, Szklarczyk D, Frankild S, *et al*. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013;**41**:D808–15.
36. Cowley MJ, Pinese M, Kassahn KS, *et al*. PINA v2.0: mining interactome modules. *Nucleic Acids Res* 2012;**40**:D862–5.
37. Hornbeck PV, Kornhauser JM, Tkachev S, *et al*. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 2012;**40**:D261–70.
38. Diella F, Cameron S, Gemund C, *et al*. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* 2004;**5**:79.
39. Minguez P, Letunic I, Parca L, *et al*. PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic Acids Res* 2015;**43**:D494–502.
40. Cheng F, Jia P, Wang Q, *et al*. Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. *Oncotarget* 2014;**5**:3697–710.
41. Mosca R, Ceol A, Aloy P. Interactome3D: adding structural details to protein networks. *Nat Methods* 2013;**10**:47–53.
42. Mosca R, Ceol A, Stein A, *et al*. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* 2014;**42**:D374–9.
43. Meyer MJ, Das J, Wang X, *et al*. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 2013;**29**:1577–9.
44. Cheng F, Jia P, Wang Q, *et al*. Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol Biol Evol* 2014;**31**:2156–69.
45. Jia P, Zhao Z. Network.assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Hum Genet* 2014;**133**:125–38.
46. Leiserson MD, Vandin F, Wu HT, *et al*. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 2015;**47**:106–14.
47. Dees ND, Zhang Q, Kandoth C, *et al*. MuSiC: identifying mutational significance in cancer genomes. *Genome Res* 2012;**22**:1589–98.
48. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* 2013;**29**:2238–44.
49. Tian R, Basu MK, Capriotti E. ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples. *Bioinformatics* 2014;**30**:i572–8.
50. Sim NL, Kumar P, Hu J, *et al*. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* 2012;**40**:W452–7.
51. 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**:1073–81.
52. Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9.
53. Schwarz JM, Rodelsperger C, Schuelke M, *et al*. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;**7**:575–6.
54. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;**39**:e118.
55. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 2011;**88**:440–9.
56. Wong WC, Kim D, Carter H, *et al*. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* 2011;**27**:2147–8.
57. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* 2012;**40**:e169.
58. Choi Y, Sims GE, Murphy S, *et al*. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 2012;**7**:e46688.
59. Mao Y, Chen H, Liang H, *et al*. CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLoS One* 2013;**8**:e77945.
60. Shihab HA, Gough J, Cooper DN, *et al*. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* 2013;**29**:1504–10.
61. Douville C, Carter H, Kim R, *et al*. CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics* 2013;**29**:647–8.
62. Ryslik GA, Cheng Y, Cheung KH, *et al*. Utilizing protein structure to identify non-random somatic mutations. *BMC Bioinformatics* 2013;**14**:190.

63. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol* 2013;**9**:637.
64. Ghersi D, Singh M. Interaction-based discovery of functionally important genes in cancers. *Nucleic Acids Res* 2014;**42**:e18.
65. Jia P, Wang Q, Chen Q, *et al*. MSEA: detection and quantification of mutation hotspots through mutation set enrichment analysis. *Genome Biol* 2014;**15**:489.
66. Porta-Pardo E, Godzik A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* 2014; **30**:3109–14.
67. Vuong H, Cheng F, Lin CC, *et al*. Functional consequences of somatic mutations in cancer using protein pocket-based prioritization approach. *Genome Med* 2014;**6**:81.
68. Zhao J, Cheng F, Wang Y, *et al*. Systematic prioritization of druggable mutations in ~5,000 genomes across 16 cancer types using a structural genomics-based approach. *Mol Cell Proteomics* 2016;**15**:642–56.
69. Vaske CJ, Benz SC, Sanborn JZ, *et al*. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 2010;**26**:i237–45.
70. Ng S, Collisson EA, Sokolov A, *et al*. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* 2012;**28**:i640–6.
71. Jia P, Zhao Z. Personalized pathway enrichment map of putative cancer genes from next generation sequencing data. *PLoS One* 2012;**7**:e37595.
72. Bashashati A, Haffari G, Ding J, *et al*. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* 2012;**13**:R124.
73. Paull EO, Carlin DE, Niepel M, *et al*. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 2013;**29**:2757–64.
74. Hofree M, Shen JP, Carter H, *et al*. Network-based stratification of tumor mutations. *Nat Methods* 2013;**10**:1108–15.
75. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. *Genome Med* 2014;**6**:56.
76. Jia P, Zhao Z. VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLoS Comput Biol* 2014;**10**:e1003460.
77. Akavia UD, Litvin O, Kim J, *et al*. An integrated approach to uncover drivers of cancer. *Cell* 2010;**143**:1005–17.
78. Zhang KX, Ouellette BF. CAERUS: predicting CAncER oUtcomeS using relationship between protein structural information, protein networks, gene expression data, and mutation data. *PLoS Comput Biol* 2011;**7**:e1001114.
79. Chen Y, Hao J, Jiang W, *et al*. Identifying potential cancer driver genes by genomic data integration. *Sci Rep* 2013;**3**:3538.
80. Sanchez-Garcia F, Villagrasa P, Matsui J, *et al*. Integration of genomic data enables selective discovery of breast cancer drivers. *Cell* 2014;**159**:1461–75.
81. Melloni GE, Ogier AG, de Pretis S, *et al*. DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes. *Genome Med* 2014;**6**:44.
82. Schroeder MP, Rubio-Perez C, Tamborero D, *et al*. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. *Bioinformatics* 2014;**30**:i549–55.
83. Bertrand D, Chng KR, Sherbaf FG, *et al*. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res* 2015;**43**:e44
84. Supek F, Minana B, Valcarcel J, *et al*. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 2014;**156**:1324–35.
85. Gudbjartsson DF, Helgason H, Gudjonsson SA, *et al*. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 2015;**47**:435–44
86. Jones S, Anagnostou V, Lytle K, *et al*. Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci Transl Med* 2015;**7**:283ra53.
87. Fu W, O'Connor TD, Jun G, *et al*. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 2013;**493**:216–20.
88. Genomes Project C, Abecasis GR, Altshuler D, *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–73.
89. Martelotto LG, Ng CK, De Filippo MR, *et al*. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol* 2014;**15**:484.
90. Carter H, Chen S, Isik L, *et al*. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* 2009;**69**:6660–7.
91. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
92. Yang F, Petsalaki E, Rolland T, *et al*. Protein domain-level landscape of cancer-type-specific somatic mutations. *PLoS Comput Biol* 2015;**11**:e1004147.
93. Wang Y, Cheng H, Pan Z, *et al*. Reconfiguring phosphorylation signaling by genetic polymorphisms affects cancer susceptibility. *J Mol Cell Biol* 2015;**7**:187–202.
94. Reimand J, Wagih O, Bader GD. The mutational landscape of phosphorylation signaling in cancer. *Sci Rep* 2013;**3**:2651.
95. Ellis MJ, Gillette M, Carr SA, *et al*. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* 2013;**3**:1108–12.
96. Shi Z, Moult J. Structural and functional impact of cancer-related missense somatic mutations. *J Mol Biol* 2011;**413**:495–512.
97. Stehr H, Jang SH, Duarte JM, *et al*. The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol Cancer* 2011;**10**:54.
98. Wang Q, Xia J, Jia P, *et al*. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Brief Bioinform* 2013;**14**:506–19.
99. Zhong Q, Simonis N, Li QR, *et al*. Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* 2009;**5**:321.
100. Pe'er D, Hacohen N. Principles and strategies for developing network models in cancer. *Cell* 2011;**144**:864–73.
101. Rolland T, Tasan M, Charloteaux B, *et al*. A proteome-scale map of the human interactome network. *Cell* 2014; **159**:1212–26.
102. Huang S, Ernberg I, Kauffman S. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. *Semin Cell Dev Biol* 2009;**20**:869–76.
103. Sahni N, Yi S, Taipale M, *et al*. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 2015;**161**:647–60.
104. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 2011;**18**:507–22.
105. Stumpf MP, Thorne T, de Silva E, *et al*. Estimating the size of the human interactome. *Proc Natl Acad Sci USA* 2008;**105**:6959–64.

106. Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? *Genome Biol* 2006;**7**:120.
107. Menche J, Sharma A, Kitsak M, *et al*. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 2015;**347**:1257601.
108. Lopez Villar E, Wang X, Madero L, *et al*. Application of oncoproteomics to aberrant signalling networks in changing the treatment paradigm in acute lymphoblastic leukaemia. *J Cell Mol Med* 2015;**19**:46–52.
109. Brunak S, De La Vega FM, Rätsch G, *et al*. Cancer panomics: Computational methods and infrastructure for integrative analysis of cancer high-throughput "omics" data-session introduction. *Pac Sym Biocomput* 2014;**19**:1–2.
110. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010;**61**:437–55.
111. Wong K, Keane TM, Stalker J, *et al*. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol* 2010;**11**:R128.
112. Alkodsi A, Louhimo R, Hautaniemi S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Brief Bioinform* 2015;**16**:242–54.
113. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, *et al*. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep* 2013;**3**:2650.
114. Alizadeh AA, Aranda V, Bardelli A, *et al*. Toward understanding and exploiting tumor heterogeneity. *Nat Med* 2015;**21**:846–53.
115. Yadav VK, De S. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Brief Bioinform* 2015;**16**:232–41.
116. Isella C, Terrasi A, Bellomo SE, *et al*. Stromal contribution to the colorectal cancer transcriptome. *Nat Genet* 2015;**47**:312–9.
117. Calon A, Lonardo E, Berenguer-Llergo A, *et al*. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet* 2015;**47**:320–9.
118. Yoshihara K, Shahmoradgoli M, Martinez E, *et al*. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013;**4**:2612.
119. Lawrence MS, Stojanov P, Mermel CH, *et al*. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 2014;**505**:495–501.
120. Imielinski M, Berger AH, Hammerman PS, *et al*. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 2012;**150**:1107–20.
121. Tomasetti C, Marchionni L, Nowak MA, *et al*. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci USA* 2015;**112**:118–23.
122. Cibulskis K, Lawrence MS, Carter SL, *et al*. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;**31**:213–9.
123. Belkadi A, Bolze A, Itan Y, *et al*. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci USA* 2015;**112**:5473–78.
124. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;**372**:793–5.
125. Kim SY, Speed TP. Comparing somatic mutation-callers: beyond Venn diagrams. *BMC Bioinformatics* 2013;**14**:189.
126. McKenna A, Hanna M, Banks E, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.
127. O'Rawe J, Jiang T, Sun G, *et al*. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 2013;**5**:28.
128. O'Brien TD, Jia P, Xia J, *et al*. Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer. *Methods* 2015;**83**:118–27.
129. Wang Q, Jia P, Li F, *et al*. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med* 2013;**5**:91.
130. Ewing AD, Houlahan KE, Hu Y, *et al*. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* 2015;**12**:623–30
131. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, *et al*. The cancer genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
132. Hoadley KA, Yau C, Wolf DM, *et al*. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 2014;**158**:929–44.
133. Ledford H. End of cancer-genome project prompts rethink. *Nature* 2015;**517**:128–9.
134. Fredriksson NJ, Ny L, Nilsson JA, *et al*. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* 2014;**46**:1258–63.
135. Weinhold N, Jacobsen A, Schultz N, *et al*. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* 2014;**46**:1160–5.
136. Schramek D, Sendoel A, Segal JP, *et al*. Direct in vivo RNAi screen unveils myosin IIa as a tumor suppressor of squamous cell carcinomas. *Science* 2014;**343**:309–13.
137. Konermann S, Brigham MD, Trevino AE, *et al*. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* 2015;**517**:583–8.
138. Newman AM, Bratman SV, To J, *et al*. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* 2014;**20**:548–54.
139. Dawson SJ, Tsui DW, Murtaza M, *et al*. Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* 2013;**368**:1199–209.
140. Lohr JG, Adalsteinsson VA, Cibulskis K, *et al*. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat Biotechnol* 2014;**32**:479–84.
141. Leary RJ, Sausen M, Kinde I, *et al*. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med* 2012;**4**:162ra54.
142. Navin N, Kendall J, Troge J, *et al*. Tumour evolution inferred by single-cell sequencing. *Nature* 2011;**472**:90–4.
143. Levine JH, Lin Y, Elowitz MB. Functional roles of pulsing in genetic circuits. *Science* 2013;**342**:1193–200.
144. Karr JR, Sanghvi JC, Macklin DN, *et al*. A whole-cell computational model predicts phenotype from genotype. *Cell* 2012;**150**:389–401.
145. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* 2013;**14**:618–30.
146. Gubin MM, Zhang X, Schuster H, *et al*. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* 2014;**515**:577–81.
147. Heemskerk B, Kvistborg P, Schumacher TN. The cancer antigenome. *EMBO J* 2013;**32**:194–203.

148. Gerlinger M, Rowan AJ, Horswell S, *et al*. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012;**366**:883–92.
149. Rooney MS, Shukla SA, Wu CJ, *et al*. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 2015;**160**:48–61.
150. Snyder A, Makarov V, Merghoub T, *et al*. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med* 2014;**371**:2189–99.
151. Rizvi NA, Hellmann MD, Snyder A, *et al*. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 2015;**348**:124–8.