Genome Medicine

REVIEW

Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine

Benjamin J Raphael[1,2*], Jason R Dobson[1,2,3], Layla Oesper[1] and Fabio Vandin[1,2]

Abstract

High-throughput DNA sequencing is revolutionizing the study of cancer and enabling the measurement of the somatic mutations that drive cancer development. However, the resulting sequencing datasets are large and complex, obscuring the clinically important mutations in a background of errors, noise, and random mutations. Here, we review computational approaches to identify somatic mutations in cancer genome sequences and to distinguish the driver mutations that are responsible for cancer from random, passenger mutations. First, we describe approaches to detect somatic mutations from high-throughput DNA sequencing data, particularly for tumor samples that comprise heterogeneous populations of cells. Next, we review computational approaches that aim to predict driver mutations according to their frequency of occurrence in a cohort of samples, or according to their predicted functional impact on protein sequence or structure. Finally, we review techniques to identify recurrent combinations of somatic mutations, including approaches that examine mutations in known pathways or protein-interaction networks, as well as *de novo* approaches that identify combinations of mutations according to statistical patterns of mutual exclusivity. These techniques, coupled with advances in high-throughput DNA sequencing, are enabling precision medicine approaches to the diagnosis and treatment of cancer.

Challenges of cancer genome sequencing and analysis

Cancer is driven largely by somatic mutations that accumulate in the genome over an individual's lifetime, with additional contributions from epigenetic and transcriptomic alterations. These somatic mutations range in scale from single-nucleotide variants (SNVs), insertions and deletions of a few to a few dozen nucleotides (indels), larger copy-number aberrations (CNAs) and large-genome rearrangements, also called structural variants (SVs). These genomic alterations have been studied for decades using low-throughput approaches such as targeted gene sequencing or cytogenetic techniques, which have led to the identification of a number of highly recurrent somatic mutations [1,2]. Importantly, a subset of these mutations have been successfully targeted therapeutically; for example, imatinib has been used to target cells expressing the *BCR-ABL* fusion gene in chronic myeloid leukemia [3], and gefitinib has been used to inhibit the epidermal growth factor receptor in lung cancer [4]. Unfortunately, highly recurrent mutations with a corresponding drug treatment are unknown for most cancer types, in part due to our lack of comprehensive knowledge of somatic mutations present in different patients from a variety of cancer types.

In the past few years, high-throughput DNA sequencing has revolutionized the identification of somatic mutations in cancer genomes. Whole-genome sequencing reveals somatic mutations of all types, whereas whole-exome sequencing identifies coding mutations at a lower cost, but does not allow the analysis of non-coding regions or the detection of SVs. When applied to many samples of the same cancer type, these technologies enable the identification of novel recurrent somatic mutations, a subset of which present new targets for cancer diagnostics and treatment [5-15]. These advances hold promise for precision medicine, or precision oncology, where a cancer treatment could be tailored to a patient's mutational profile [16]. Fulfilling this promise of precision oncology will require researchers to overcome several challenges in the analysis and interpretation of sequencing data.

In this review, we focus on three key challenges in cancer genome sequencing. First is the issue of identifying somatic mutations from the short sequence reads generated by high-throughput technologies, particularly in the presence of intra-tumor heterogeneity. Second is the problem of distinguishing the relatively small

* Correspondence: braphael@brown.edu
[1]Department of Computer Science, Brown University, 115 Waterman Street, Providence, RI 02912, USA
Full list of author information is available at the end of the article

number of driver mutations that are responsible for the development and progression of cancer from the large number of passenger mutations that are irrelevant for the cancer phenotype. Third is the challenge of determining the biological pathways and processes that are altered by somatic mutation. We survey recent computational approaches that address each of these challenges.

The rapid advances in high-throughput DNA sequencing technologies and their application to cancer genome sequencing has led to a proliferation of approaches to analyze the resulting data. Moreover, there are multiple signals in sequencing data that can be used to address the challenges listed above, and different computational methods use different combinations of these signals. This rapid pace of progress, the diversity of strategies and the lack, for the most part, of rigorous comparisons among different methods explain why a standard pipeline for the analysis of high-throughput cancer genome sequencing data has yet to emerge. Hence, we are able to include only a fraction of possible approaches. Moreover, we restrict attention to methods for DNA sequencing data and do not discuss the analysis of other high-throughput sequencing data, such as RNA sequencing data, that also provide key components for precision medicine [17].

Detection of somatic mutations

Many of the recent advances in our understanding of driver mutations have been the result of the increasing availability and affordability of DNA-sequencing technologies produced by companies such as Illumina, Ion Torrent, 454, Pacific Biosciences, and others. Such technologies enabled the sequencing of the first cancer genome [18] and the subsequent sequencing of thousands of additional cancer genomes, particularly through collaborative projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). Some of these projects employ whole-genome sequencing, whereas others use exome sequencing, a targeted approach that sequences only the coding regions of the genome, enabling deeper coverage sequencing of genes but at the expense of ignoring non-coding regions. At the moment, the dominant approach is to perform whole-exome sequencing using one of several target-enrichment protocols followed by Illumina sequencing. However, the cost-benefit analysis of different technologies and approaches is continually changing, and we refer the reader to recent surveys for additional information [17,19,20].

The advances in DNA sequencing technologies have been dramatic, but these technologies still face some significant limitations in measuring genomes. In particular, all of the technologies that sequence human genomes at reasonable cost produce millions to billions of short sequences, or reads, of approximately 50–150 bp in length. To detect somatic mutations in cancer genomes, these reads are aligned to the human reference genome and differences between the reference genome and the cancer genome are identified (Figure 1a). A matched normal sample from the same individual is typically analyzed simultaneously to distinguish somatic from germline mutations. The process of detecting somatic mutations from aligned reads is not straightforward. Numerous errors and artifacts are introduced during both the sequencing and the alignment processes including: optical PCR duplicates, GC-bias, strand bias (where reads indicating a possible mutation only align to one strand of DNA) and alignment artifacts resulting from low complexity or repetitive regions in the genome. These lead to somatic mutation predictions containing both incorrect variants (false positives) and missing variants (false negatives) [21].

While standard pre-processing handles some sources of error (such as the removal of PCR duplicates), most methods for somatic mutation detection address only a subset of the possible sources of error. For instance, the methods MuTect [22] and Strelka [23] for predicting SNVs both employ stringent filtering after initial SNV detection to remove false positives resulting from strand bias or from poor mapping resulting from repetitive sequence in the reference genome. Such filtering may, however, result in high false negatives. On the other hand, the VarScan 2 method [24] does not specifically address either of these issues, but still outperforms the previously mentioned methods on some datasets [25]. These differences demonstrate that the performance of methods can vary by dataset, and suggest that running multiple methods is advisable at present. Table 1 lists a number of publicly available algorithms for the detection of somatic SNVs, CNAs, and SVs from DNA-sequencing data. New methods and further refinements of existing methods for somatic mutation detection continue to be developed.

Intra-tumor heterogeneity

One particular challenge in identifying and characterizing somatic mutations in tumors is the fact that most tumor samples are a heterogeneous collection of cells, containing both normal cells and different populations of cancerous cells [26]. The clonal theory of cancer [27] posits that all cancerous cells in a tumor descended from a single cell in which the first driver mutation occurred, and that subsequent clonal expansions and selective sweeps lead to a tumor with a dominant (majority) population of cancerous cells containing early driver events. Most cancer-genome sequencing studies generate data from a bulk tumor sample that contains both normal cells and one or more subpopulations of tumor cells. This intra-tumor heterogeneity complicates the identification of all types of somatic mutations and

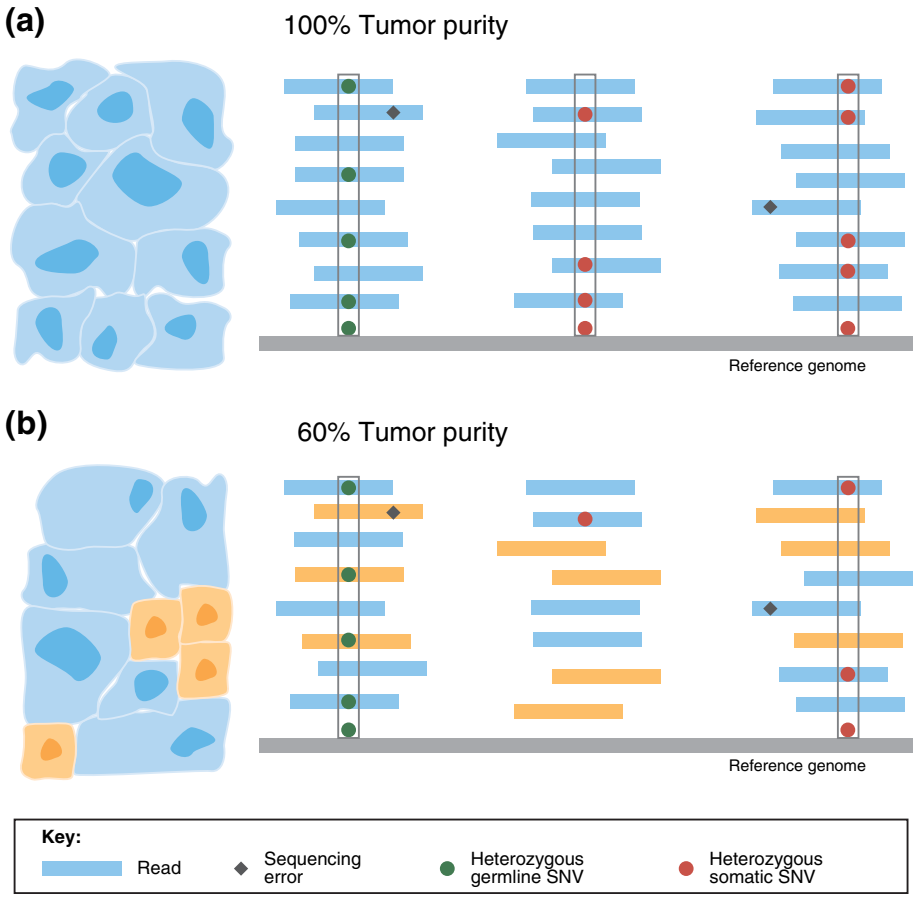

Figure 1 Somatic mutation detection in tumor samples. DNA-sequence reads from a tumor sample are aligned to a reference genome (shown in gray). Single-nucleotide differences between reads and the reference genome indicate germline single-nucleotide variants (SNVs; green circles), somatic SNVs (red circles), or sequencing errors (black diamonds). **(a)** In a pure tumor sample, a location containing mismatches or single nucleotide substitutions in approximately half of the reads covering the location indicates a heterozygous germline SNV or a heterozygous somatic SNV - assuming that there is no copy number aberration at the locus. Algorithms for detecting SNVs distinguish true SNVs from sequencing errors by requiring multiple reads with the same single-letter substitution to be aligned at the position (gray boxes). **(b)** As tumor purity decreases, the fraction of reads containing somatic mutations decreases: cancerous and normal cells, and the reads originating from each, are shown in blue and orange, respectively. The number of reads reporting a somatic mutation decreases with tumor purity, diminishing the signal to distinguish true somatic mutations from sequencing errors. In this example, only one heterozygous somatic SNV and one hetererozygous germline SNV are detected (gray boxes) as the mutation in the middle set of aligned reads is not distinguishable from sequencing errors.

specialized methods [28-31] have been developed to quantify the extent of heterogeneity in a sample. The simplest form of intra-tumor heterogeneity is admixture by normal cells. The tumor purity of a sample is defined as the fraction of cells in the sample that are cancerous. A read from a tumor sample represents a sequence in the cell, or subpopulation of cells, from which the read was derived. Thus, lower tumor purity results in a reduction in the number of sequence reads derived from the cancerous cells, and thus a reduction in the signal that can be used to detect somatic mutations (Figure 1b).

Tumor purity is an important parameter in the detection of somatic mutations. To obtain reasonable sensitivity and specificity, methods to predict somatic aberrations must utilize, either implicitly or explicitly, an estimate of tumor purity. The VarScan 2 program [24] for calling somatic SNVs and indels allows a user to provide an estimate of tumor purity in order to calibrate the expected number of reads containing a somatic mutation at a single locus. Conversely, methods such as MuTect [22] and Strelka [23] explicitly model tumor and normal allele frequencies using observed data to calibrate sensitivity. As a result, MuTect and Strelka may provide improved sensitivity for detecting mutations that occur in lower frequencies, especially when tumor purity is unknown *a priori*. The performance of these and other somatic mutation-calling algorithms depends on accurate estimates of tumor purity.

Standard methods for estimating tumor purity involve visual inspection by a pathologist or automated analysis

Table 1 Methods for detecting somatic mutations

Objective	Data	Method	Description
Somatic mutation detection	SNV	MuTect [22]	Designed to detect low-frequency mutations in both whole-genome and exome data.
		Strelka [23]	Can be applied to both whole-genome and whole-exome data. Uses stringent post-call filtration.
		VarScan 2 [24]	Demonstrates high sensitivity for detecting SNVs in relatively pure tumor samples from both whole-genome and exome data.
		JointSNVMix [128]	A probabilistic model that describes the observed allelic counts in both tumor and normal samples.
	CNA or SV	BIC-Seq [129]	Detects CNAs from whole-genome data.
		APOLLOH [130]	Predicts loss of heterozygosity regions from whole-genome sequencing data.
		CoNIFER [131]	Detects CNAs from exome data.
		BreakDancer [132]	Cluster paired-end alignments to detect SVs. One version to detect large aberrations and another to detect smaller indels.
		VariationHunter-CommonLaw [133], HYDRA [70]	Cluster paired-reads, including reads with multiple possible alignments. Support simultaneous analysis of multiple samples.
		GASV/GASVPro [134,135], PeSV-Fisher [136]	Combine paired-read and read-depth analysis to detect SVs.
		Meerkat [130]	Combines paired-end split-read and multiple alignment information to detect structural aberrations.
		Delly [137], Break-Pointer [138]	Combines paired-end and split-read signals to detect structural aberrations.
Tumor purity estimation	SNV	ABSOLUTE [28]	Originally designed for SNP array data, but may be adapted for whole-genome sequencing data. Handles subclonal populations as outliers.
		ASCAT [29]	Designed for SNP array data, but may be adapted for whole-genome sequencing data. Only considers a single tumor population.
	CNA	THetA [30]	Able to consider multiple subclonal tumor populations, but only if they differ by large CNAs. Designed for whole-genome sequencing data.
		SomatiCA [31]	Only uses aberrations that are identified as clonal to estimate tumor purity.

CNA, copy number aberration; SNV, single-nucleotide variant; SV, structural variant.
A representative list of software available for the detection of somatic mutations from high-throughput sequencing data of cancer genomes. Some methods detect more than one type of mutation but are listed only once for clarity.

of cellular images [32]. Recently, several alternative approaches have been developed to estimate tumor purity directly from sequencing data by identifying shifts in the expected number of reads that align to a locus (Table 1). This is not an easy task as most cancer genomes are aneuploid and thus do not contain two copies of each chromosomal locus. The tumor ploidy, defined as the total DNA content in a tumor cell, also results in shifts in the sequencing coverage. Thus, estimation of tumor purity and tumor ploidy are closely intertwined. ABSOLUTE [28] and ASCAT [29] are two algorithms that are used to infer both tumor purity and tumor ploidy from single-nucleotide polymorphism (SNP) array data. Although both methods may be modified to work with DNA-sequencing data [33], they model a tumor sample as consisting of only two populations: normal cells and tumor cells. As they do not directly model the possible existence of multiple distinct tumor subpopulations, the tumor purity estimates that result can be inaccurate, and reflect either an average over all tumor subpopulations or a bias for the dominant tumor subpopulation [30]. Furthermore, accurate identification of tumor subpopulations may provide important information on tumors that do not respond well to treatments [34-36].

Recently, the Tumor Heterogeneity Analysis (THetA) algorithm [30] was developed to infer the composition of a tumor sample (including tumor purity) containing any number of tumor subpopulations directly from DNA-sequencing data. Although THetA overcomes some of the limitations of earlier methods, it is unable to distinguish distinct tumor subpopulations that do not contain CNAs, necessitating the development of additional approaches to identify tumor subpopulations that are distinguished only by SNVs and/or small indels. The identification of somatic mutations and the estimation of intra-tumor heterogeneity are closely related, and so methods that jointly perform these tasks while allowing for multiple tumor subpopulations are desirable for obtaining highly sensitive and specific estimates of all somatic aberrations in tumors.

Advances in DNA-sequencing technologies have also enabled the direct quantification of intra-tumor

heterogeneity. One approach is to perform targeted, ultra-deep-coverage sequencing of SNVs, followed by clustering of the read counts for each SNV into distinct subpopulations [37,38]. Ding *et al.* [37] identified two distinct clonal evolution patterns for acute myeloid leukemia (AML) patients: a relapse sample evolved either from the founding clone in the primary tumor or from a minor subclone that survived initial treatment. Shah *et al.* [38] demonstrated extreme variability in the total number of tumor subpopulations (ranging from 1–2 to more than 15 subpopulations) in tumors from a large cohort of breast cancer patients. Another approach to measure intra-tumor heterogeneity is to sequence samples from multiple regions within the same tumor. Gerlinger *et al.* [39] sequenced multiple regions from several kidney tumors and found that a majority (63-69%) of the somatic mutations identified were present in only a subset of the sequenced regions of the tumor. Navin and colleagues [40,41] found similar heterogeneity in the CNAs present within different regions of breast tumors. These results demonstrate that a single sample from a tumor might not fully represent the complete landscape of somatic mutations (including driver mutations) present in the tumor.

Finally, Nik-Zainal *et al.* [42] demonstrated how careful computational analysis can reveal information about the composition of a tumor sample, including the identification of clonal mutations that are present in nearly all cells of the tumor (and thus presumably are early events in tumorigenesis) and subclonal mutations that are present in a fraction of tumor cells. Using high-coverage (188X) whole-genome DNA sequencing of a breast tumor, they inferred the proportion of tumor cells containing somatic SNVs and CNAs and grouped these proportions into several clusters, demonstrating different mutational events during the evolutionary progression from the founder cell of the tumor to the present tumor cell population. Eventually, single-cell sequencing technologies [41,43-47] promise to provide a comprehensive view of intra-tumor heterogeneity, but these approaches remain limited by artifacts introduced during whole-genome amplification [47]. In the interim, there is an immediate need for better methods to detect somatic mutations that occur in heterogeneous tumor samples.

Computational prioritization of driver mutations

Following the sequencing of a cancer genome, the next step is to identify driver mutations that are responsible for the cancer phenotype. Ultimately, the determination that a mutation is functional requires experimental validation, using *in vitro* or *in vivo* models to demonstrate that a mutation leads to at least one of the characteristics of the cancer phenotype, such as DNA repair deficiency, uncontrolled proliferation and growth, or immune evasion. As a result of advances in DNA-sequencing technology, the measurement of somatic

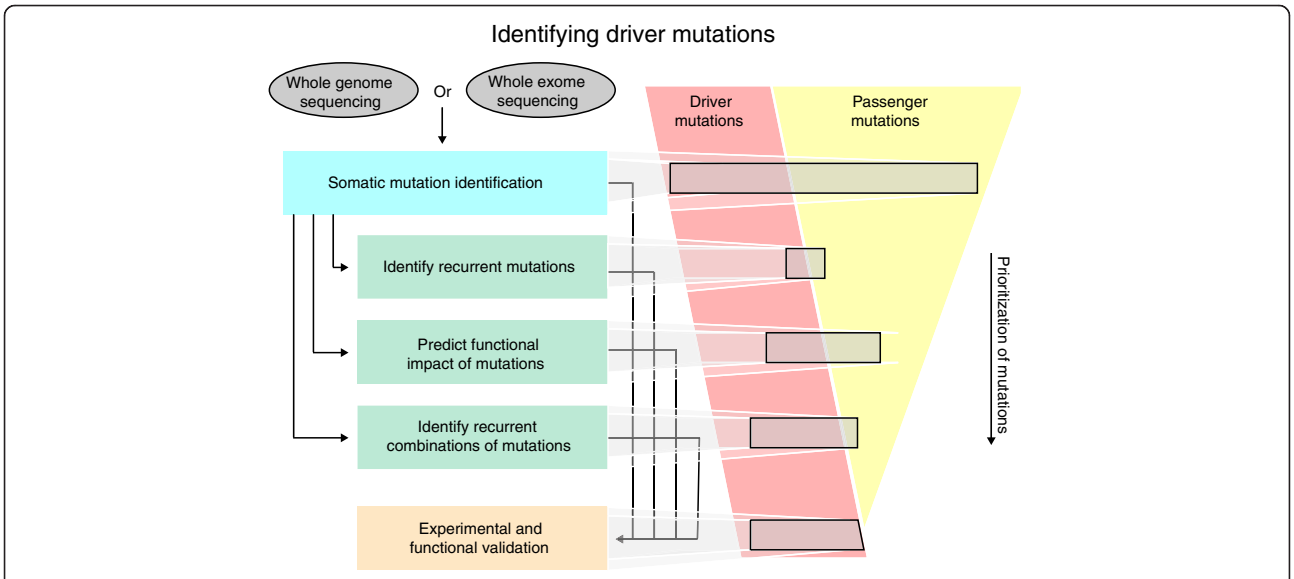

Figure 2 Overview of strategies for cancer-genome sequencing. A cancer-genome sequencing project begins with whole-genome or whole-exome sequencing. Various methods are used to detect somatic mutations in the resulting sequence (see Table 1), yielding a long list of somatic mutations. Several strategies can then be employed to prioritize these mutations for experimental or functional validation. These strategies include: testing for recurrent mutations, predicting functional impact, and assessing combinations of mutations (see Table 2). None of these approaches are perfect, and each returns a subset of driver mutations as well as passenger mutations. The mutations returned by these approaches can then be validated using a variety of experimental techniques.

Table 2 Methods for prediction of driver mutations and genes

Objective	Data	Method	Description
Recurrent somatic mutation identification	SNV	MutSigCV [48]	Uses coverage information and genomic features (e.g. DNA replication time) to estimate the background mutation rate of a gene.
		MuSiC [49]	Uses a per-gene background mutation rate; allows for user-defined regions of interest.
		Youn *et al.* [51]	Includes predicted impact on protein function in determining recurrent mutations.
		Sjöblom *et al.* [52]	Defines a cancer mutation prevalence score for each gene.
		DrGaP [139]	Uses Bayesian approach to estimate background mutation rate; helpful for cancer types with low mutation rate.
	CNA	GISTIC2 [61], JISTIC [63]	Uses 'peel-off' techniques to find smaller recurrent aberrations inside larger aberrations.
		CMDS [62]	Identifies recurrent CNAs from unsegmented data.
		ADMIRE [65]	Multi-scale smoothing of copy number profiles.
Functional impact prediction	General	SIFT [72]	Uses conservation of amino acids to predict functional impact of a non-synonymous amino-acid change.
		Polyphen-2 [74]	Infers functional impact of non-synonymous amino-acid changes through alignments of related peptide sequences and a machine-learning-based probabilistic classifier.
		MutationAssessor [75]	Uses protein homologs to calculate a score based on the divergence in conservation caused by an amino-acid change.
		PROVEAN [73]	Benchmarks favorably against MutationAssessor, Polyphen-2 and SIFT.
	Cancer-specific	CHASM [77]	Uses a machine-learning approach to classify mutations as drivers or passengers based on sequence conservation, protein domains, and protein structure.
		Oncodrive-FM [79]	Combines scores from SIFT, Polyphen-2, and MutationAccessor into a single ranking.
	Positional or structural clustering	NMC [83]	Finds clusters of non-synonymous mutations across patients. Typically used with missense mutations to detect so-called 'activating' mutations.
		iPAC [84]	Extends the NMC approach to search for clusters of mutations in three-dimensional space using crystal structures of proteins.
Pathway analysis and combinations of mutations	Known pathways	GSEA [92]	A general technique for testing ranked lists of genes for enrichment in known gene sets. Can be used on rankings derived from significance of observed mutations.
		PathScan [95]	Finds pathways with excess of mutations in a gene set (pathway), by combining *P*-values of enrichment across samples.
		Patient-oriented gene sets [94]	Tests known pathways using a binary indicator for a pathway in each patient.
	Interaction networks	NetBox [140]	Finds network modules in a user-provided list of genes. Significance depends only on the topology of the genes in the network, and not on mutation scores.
		HotNet [102]	Finds subnetworks with significantly more aberrations than would be expected by chance, using both network topology and user-defined gene or protein scores.
		MEMo [104]	Finds subnetworks whose interacting pairs of genes have mutually exclusive aberrations [105]; recommends including only recurrent SNVs and CNAs in the analysis.
	De novo	Dendrix [102]	Identifies groups of genes with mutually exclusive aberrations.
		Multi-Dendrix [112]	Simultaneously finds multiple groups of genes with mutually exclusive aberrations.
		RME [110]	Finds groups of genes with mutually exclusive aberrations by building from gene pairs; best results obtained when restricting to genes with high mutation frequencies (e.g. > 10%).

CNA, copy number aberration; SNV, single-nucleotide variant.
A representative list of software available to predict driver mutations or genes by detecting their recurrence across multiple samples, functional impact, or interactions with other mutations in pathways or combinations. Some methods fall into multiple categories but are listed only once for clarity.

mutations is now significantly cheaper and faster than the functional characterization of a mutation. Moreover, as cancer-genome sequencing moves from the research laboratory into the clinic, there is a strong need to automate the categorization of mutations to prioritize rapid, accurate diagnoses and treatments for patients. Unfortunately, distinguishing driver from passenger mutations solely from the resulting DNA-sequence change is extremely complicated, as the effect of most DNA-sequence changes is poorly understood, even in the simplest case of single nucleotide substitutions in coding regions of well-studied proteins.

In the following sections, we describe three approaches for computational prioritization of driver mutations: identifying recurrent mutations; predicting the functional impact of individual mutations; and assessing combinations of mutations using pathways, interaction networks, or statistical correlations. These approaches provide alternative strategies to filter the long list of measured somatic mutations, and to identify a smaller subset enriched for driver mutations to undergo further experimental and functional validation (Figure 2).

Statistical tests for recurrent mutations

One approach to prioritize mutations for further experimental characterization is to identify recurrent mutations. Each cancer sample has undergone an independent evolutionary process in which acquired driver mutations that provide selective advantage result in clonal expansion of these lineages [27]. As these mutational processes converge to a common oncogenic phenotype, the mutations that drive cancer progression should appear more frequently than expected by chance across patient samples. Recurrence may be revealed at different levels of resolution, such as an individual nucleotide, a codon, a protein domain, a whole gene, or even a pathway. In this section, we describe the techniques and difficulties in identifying recurrently mutated driver genes.

Statistical tests for genes with recurrent single-nucleotide mutations

Several methods have been designed to find recurrent mutations in a cohort of cancer patients, including MutSigCV [48], MuSiC [49], and others [50-53] (Table 2). The fundamental calculation in all these approaches is to determine whether the observed number of mutations in the gene is significantly greater than the number expected according to a background mutation rate (BMR). The BMR is the probability of observing a passenger mutation in a specific location of the genome. From the BMR and the number of sequenced nucleotides within a gene, a binomial model can be used to derive the probability of the observed number of mutations in a gene across a cohort of patients (Box 1).

Box 1. The binomial model: a statistical test for detecting recurrent mutations.

Using the background mutation rate (BMR) and the number n of sequenced nucleotides within a gene (g), the probability (Pg) that a passenger mutation is observed in g is given by $Pg = 1 - (1- BMR)$. Since somatic mutations arise independently in each sample, the occurrences of passenger mutations in g are modeled by flipping a biased coin with probability pg of heads (mutation). Thus, if somatic mutations have been measured in m samples, the number of patients in which gene g is mutated is described by a binomial random variable $B(m, Pg)$ with parameters m and Pg. From $B(m, Pg)$, it is possible to compute the probability that the observed number or more samples contain passenger mutations; this is the P-value of the statistical test. A multiple-hypothesis testing correction is applied when examining multiple genes.

The main differences between methods for identifying recurrently mutated genes are in how they estimate the BMR and how many different mutational contexts they analyze. Regarding the former, the BMR is not constant across the genome, but depends on the genomic context of a nucleotide [52] and the type of mutation [7]. Moreover, the BMR of a gene is correlated with both its rate of transcription [54] and replication timing [55,56]. The BMR is also not constant across patients, and cancer cohorts often present hypermutated samples [6]. Finally, certain genomic regions may display localized somatic hypermutation, termed kataegis [57]. Different combinations of these effects can cause the BMR to vary by as much as an order of magnitude across different genes.

The estimated BMR greatly affects the identification of recurrent mutations, as an estimate that is higher than the true value fails to identify recurrent mutations (false negatives), whereas an estimate that is lower than the true value would leads to false positives. Of course, if a driver gene is mutated in a very high percentage of samples (more than 20%, for example), even an inaccurate estimate of the BMR is sufficient to correctly identify such a gene as recurrently mutated. Thus, well-known cancer genes (such as *TP53*) are readily identified as recurrently mutated genes by all computational methods. The priority now is to identify rare driver mutations that are important for precision oncology. The tools that are currently available often report different rare mutations as drivers, and more work is needed in order to improve the sensitivity in the detection of rare driver mutations and to compare and combine the results from different tools [58]. In general, reporting rarely mutated genes as recurrently mutated with high confidence requires either better estimates of the BMR and/or much larger patient cohorts.

Statistical tests for genes with recurrent copy number and structural aberrations

The identification of genes with recurrent copy number or structural aberrations presents different challenges. Somatic copy number aberrations (SCNAs) show large variation in their position and length across different samples. For example, an oncogene may be amplified in one sample because of a whole-chromosome gain, whereas in another sample, the amplification may be focal and include only the oncogene. Thus, determining whether CNAs in two individuals are the 'same' is not a straightforward task. Moreover, recent evidence suggests that SCNAs are not distributed uniformly over the genome but are biased by chromosome organization and DNA replication timing [59,60]. Because of these difficulties, no accurate model to identify CNAs has been developed. Thus, methods for predicting recurrent CNAs generally take a non-parametric approach. Early approaches looked for minimal common regions, regions of shared aberrations across individuals. The statistical significance of such overlaps was then assessed by fixing the lengths of the aberrations but independently permuting their position across individuals. More recent approaches, such as GISTIC2 [61], CMDS [62], JISTIC [63], DiNAMIC [64], and ADMIRE [65] (see Table 2), use more sophisticated models to separate and assess the statistical significance of overlapping CNAs of different lengths.

Recurrent structural aberrations such as translocations, inversions, and other genome rearrangements are typically straightforward to detect when: (1) the breakpoints of these aberrations are closely located in different individuals; and (2) these breakpoints are outside of repetitive or low-complexity regions that present difficulties for read alignment. Examples of rearrangements that are readily detectable include highly recurrent fusion genes such as *BCR-ABL* in leukemias and *TMPRSS2-ERG* in prostate cancers. In some cases, it is possible to detect recurrent fusion genes directly from microarray data that does not involve sequencing the breakpoints [66]. At the other extreme, mechanisms such as chromothripsis [67] or chromoplexy [68], which lead to simultaneous rearrangement of multiple genomic loci, result in complicated sets of overlapping breakpoints. Such complex rearrangements demand specialized techniques for analysis [69,70] and are difficult to assess for recurrence across individuals.

Prediction of functional impact

Another approach for distinguishing driver mutations from passenger mutations is to predict the functional impact of a mutation using additional biological information about the sequence and/or structure of the protein encoded by the mutated gene. The advantage of such approaches is that they can be applied to mutations that are present in only a single individual. These methods are applied to non-silent SNV (nsSNVs) that result in changes in the amino-acid sequence of the corresponding protein. These changes include missense mutations that substitute one amino-acid residue, nonsense mutations that introduce a stop codon, frame-shift mutations that alter the reading-frame of the transcript, in-frame insertions or deletions that may alter the function of the protein, and splice-site mutations that alter splice donor or acceptor sites. Nonsense and frame-shift mutations are typically assumed to be inactivating mutations, and therefore highly likely to have a functional impact. Thus, these mutations are not further annotated with respect to functional impact. Splice-site mutations require specialized techniques for interpretation that address the complexities of alternative splicing [71]. In this section, we briefly highlight methods for predicting the functional impact of missense mutations (Table 2).

Several methods have been developed to predict the effect of germline SNPs. Popular methods include SIFT [72], PROVEAN [73], and Polyphen-2 [74]. More recently, MutationAssessor [75] and the algorithm of Fischer *et al.* [76] have been designed to combine evolutionary conservation and protein-domain information in order to infer the functional impact of somatic mutations and therefore distinguish driver from passenger mutations. Other recent methods focus specifically on somatic mutations. These include CHASM [77], which uses machine-learning algorithms trained on known driver mutations and the algorithm presented by Li *et al.* [78], which uses a combination of clustering of nsSNVs and conservation of residues at nsSNVs. Similarly, Oncodrive-FM [79] combines scores from SIFT, Polyphen-2 and MutationAssesor and looks for bias in these scores across a collection of patients (typically having the same cancer type).

Another approach to predict functional impact is to examine whether missense mutations cluster in the protein sequence. The motivation for examining positional clustering comes from examples of activating mutations in oncogenes that show strong positional preferences (such as the V600E mutation in *BRAF* [80] and mutations in residues 12, 13, and 61 of *KRAS* [81]) or inactivating mutations such as those observed in the DNA-binding domain of *TP53* [82]. Approaches such as NMC [83] and iPAC [84] identify clustering of missense mutations in protein sequence (two-dimensional space) and protein structure (three-dimensional space), respectively. NMC can be run on any sequence, but iPAC requires that the crystal structure of the protein has been solved. Although the percentage of solved protein structures is rapidly increasing, this requirement limits three-

dimensional analysis to well-studied proteins and thus reduces the ability of iPAC to discover novel cancer-related genes.

Although these approaches are useful in prioritizing mutations, they assume that *a priori* information, such as evolutionary conservation, known protein domains, non-random clustering of mutations, protein structure, or some combination thereof, will help to distinguish passenger from driver mutations. These data may not, however, provide enough information to allow prediction of a mutation's oncogenic impact; for example, the specific epitopes of phospho-kinases and signal transduction proteins can be quite complex [85]. Thus, these approaches may miss important oncogenic mutations; for example, MutationAssessor [75] assigns a low score to a well-known activating mutation (H1047R) in *PIK3CA* [86].

Combinations of mutations: pathways, interaction networks, and *de novo* approaches

Genes and their protein products rarely act in isolation. Rather, they interact with other genes or proteins in various signaling, regulatory, and metabolic pathways, as well as in protein complexes. Cancer research over the past few decades has characterized a number of these key pathways and has provided information about how these pathways are perturbed by somatic mutation [1,87]. At the same time, the complexity of this interacting network of genes or proteins presents a major confounding factor for identifying driver mutations in genes using statistical patterns of recurrence. For instance, if cancer progression requires the deregulation of a particular pathway (such as those involved in apoptosis) there are a large number of known and unknown genes whose mutation would perturb this pathway. While some of the genes in these pathways may be frequently altered, other genes may be mutated rarely in a collection of patients with a given cancer type. This idea explains the long tail phenomenon that is apparent from recent cancer genome studies: only a few genes are mutated frequently and many more are mutated at frequencies that are too low to be statistically significant [2]. Consequently, in order to identify rare driver mutations that are crucial for precision oncology, it is advantageous to identify groups or combinations of genes that are recurrently mutated.

In the following sections, we consider three approaches that have been used to identify such combinations: first, the identification of recurrent mutations in pre-defined gene sets using databases of known pathways, protein complexes, or other functional groupings; second, the identification of recurrent mutations in genome-scale interaction networks; and third, the identification of recurrent combinations of mutations *de novo* without any prior knowledge of gene sets. These three approaches sequentially reduce the amount of prior knowledge that must be available on the gene sets under consideration, thus enabling the discovery of novel combinations of mutated genes. This potential benefit comes, however, at the expense of an increase in the number of hypotheses that are considered, resulting in computational and statistical issues that must be addressed appropriately (Figure 3).

Known pathways

A direct approach to assess whether groups of genes are recurrently mutated in a cohort of sequenced cancer genomes or exomes is to examine the frequency of mutation in gene sets determined by prior biological knowledge of functionally related genes (Table 2). The most straightforward approach is to determine whether a list of mutated genes shares significant overlap with known gene sets. Any of the many tools used for the analogous analysis of gene expression, such as DAVID [88,89], FaTiGO [90] or GoStat [91], may be used. To use these tools, an appropriate list of mutated genes must first be defined; often this is accomplished by relaxing the threshold for statistical significance in one of the tests for recurrently mutated genes. An alternative approach is to rank the list of mutated genes, and then apply a method such as Gene Set Enrichment Analysis (GSEA) [92] that assesses whether a pre-defined set of genes has more high-ranking genes than would be expected by chance. Lin *et al.* [93] used this approach, ranking genes by their Cancer Mutation Prevalence (CaMP) scores [52]; the resulting method was called CaMP-GSEA. Since then, similar approaches have been taken, in which different scores are applied in combination with the GSEA algorithm to determine enrichment of mutations in certain pathways or cellular functions.

Recently, more sophisticated methods that consider the variability in the mutation rate in individual patients have been developed [94,95]. The method of Boca *et al.* [94] focuses on patient-oriented gene sets, defining a per-patient score for a gene set and then combining these scores across all of the patients. PathScan [95] evaluates the enrichment for mutations in a gene set separately for each patient (also accounting for the length of each gene in the set), and then combines the results of these tests across all of the patients.

These tests of known gene sets overcome some of the difficulties in tests of individual genes, but they have four major limitations. First, many annotated gene sets are large, containing dozens of genes. Enrichment and rank statistics may not deem mutations in a smaller subset of these genes to be significant. Second, pathways do not act in isolation; pathways themselves are interconnected in larger signaling and regulatory networks. This crosstalk between pathways is itself important; as stated by Frank McCormick, the

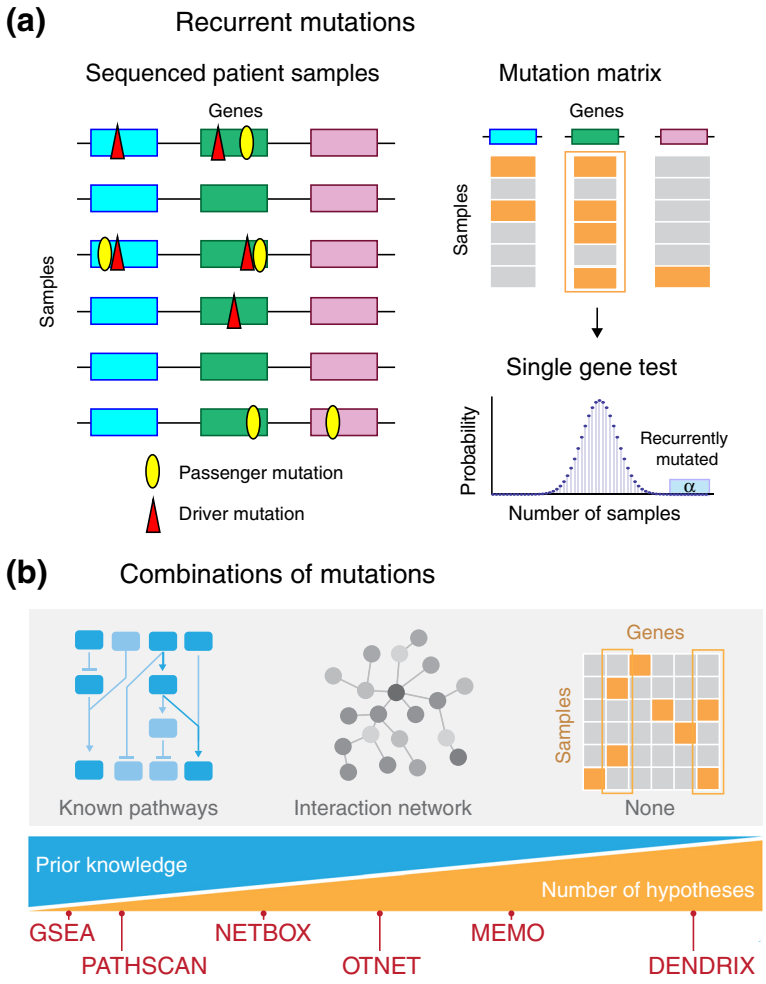

Figure 3 Overview of approaches to predict driver mutations. (a) Recurrent mutations that are found in more samples than would be expected by chance are good candidates for driver mutations. To identify such recurrent mutations, a statistical test is performed (see Table 2), which usually collapses all of the non-synonymous mutations in a gene into a binary mutation matrix that indicates the mutation status of a gene in each sample. **(b)** Assessing combinations of mutations overcomes some limitations of single-gene tests of recurrence. Three approaches to identify combinations of driver mutations are: (1) to identify recurrent mutations in predefined groups (such as pathways and protein complexes from databases); (2) to identify recurrent mutations in large protein-protein interaction networks; (3) *de novo* identification of combinations, without relying on *a priori* definition of gene sets. These approaches sequentially decrease the amount of prior information in the gene sets that are tested, thus allowing the discovery of novel combinations of driver mutations. However, the decrease in prior knowledge comes at the expense of a steep increase in the number of hypotheses considered, posing computational and statistical challenges. Different methods to identify combinations of driver mutations lie on different positions of the spectrum that represents the trade-off between prior knowledge and number of hypotheses tested.

genes involved in the development of cancer 'affect multiple pathways that intersect and overlap' [96]. Third, gene-set methods ignore the topology of interactions, instead considering all genes within a pathway equally. Finally, restricting attention to known pathways, or gene sets, does not allow the discovery of novel combinations of mutated genes and reduces the power to detect driver mutations in less-characterized and less-studied pathways.

Interaction networks

An alternative to examining mutations in previously defined gene sets is to examine mutations on large-scale protein-protein interaction networks. Examples of such networks are HPRD [97], BioGrid [98], Reactome [99], STRING [100], and iRefIndex [101]. These networks include some combinations of experimentally characterized interactions, interactions derived by high-throughput approaches (such as yeast two-hybrid screens or mass spectrometry), and/or interactions derived by automated curation of interactions reported in the literature. Some networks integrate interaction information from multiple sources. Although these networks currently provide only a partial picture of the interactions among proteins, network approaches can potentially

overcome some of the limitations of pathway analysis noted in the previous section.

Driver mutations perturb signaling, regulatory or metabolic pathways that promote the development and progression of cancer. Therefore, a desirable goal is to identify all significantly mutated subnetworks (which comprise connected sets of proteins) in a biological interaction network, but this is a complicated task. A naive approach to the problem (not based on prior knowledge) is to test all possible subnetworks for recurrent mutations using the gene-set approaches described earlier. However, the number of such subnetworks is enormous (for example, there are than 10^{14} subnetworks with at least eight proteins in a moderately sized interaction network); this presents major computational and statistical testing issues. Further complicating this type of approach is the fact that within most biological networks there are a few proteins that have an extreme number of interacting partners compared to the average protein in the network. These high-degree nodes cause many proteins to be connected via a small number of 'hops' in the graph, which implies that straightforward tests of network connectivity may lead to erroneous conclusions.

The HotNet [102] algorithm addresses many of these problems by using a heat-diffusion model to encode both the frequency of mutations in genes and the local topology of the interaction network. Furthermore, to overcome statistical issues HotNet also employs a novel statistical test (Figure 4). HotNet is able to identify subnetworks containing genes that are mutated in a relatively small number of samples (too few to be identified as recurrently mutated genes), but whose interactions indicate that these mutations are clustered on a small set of interacting proteins. HotNet's statistical test avoids the explicit testing of the huge number of subnetworks present in the interaction network, as well as the corresponding naive multiple hypothesis correction that would greatly reduce the power of detecting significantly mutated subnetworks. Two examples of significantly mutated subnetworks that have been identified by HotNet are the Notch pathway in TCGAs ovarian serous adenocarcinoma study [7] and several members of the SWI/SNF chromatin remodeling complex in the TCGA study of renal cell carcinoma [9]. In addition to the ovarian and kidney studies, HotNet has been used in a prostate cancer study [103] and in the TCGA study of AML [10].

The MEMo [104] algorithm takes a different approach in which subnetworks (called modules) of proteins that share multiple interacting partners in an interaction network are partitioned such that the genes encoding proteins in the module demonstrate a significant pattern of mutual exclusivity in their mutations. MEMo is generally run using a short list of genes (< 100) that are recurrently mutated (with SNVs or CNAs), and whose expression level is concordant with identified CNAs [105]. When used in this way, MEMo is unlikely to identify any novel genes that are not already reported as significantly mutated. Nonetheless, MEMo has been used in several TCGA studies [5,8,9,11,12] to identify exclusive mutations in the TP53 signaling pathway in breast cancer [12] among others.

Network analysis is less restrictive than testing known pathways or gene sets, but these analyses remain limited

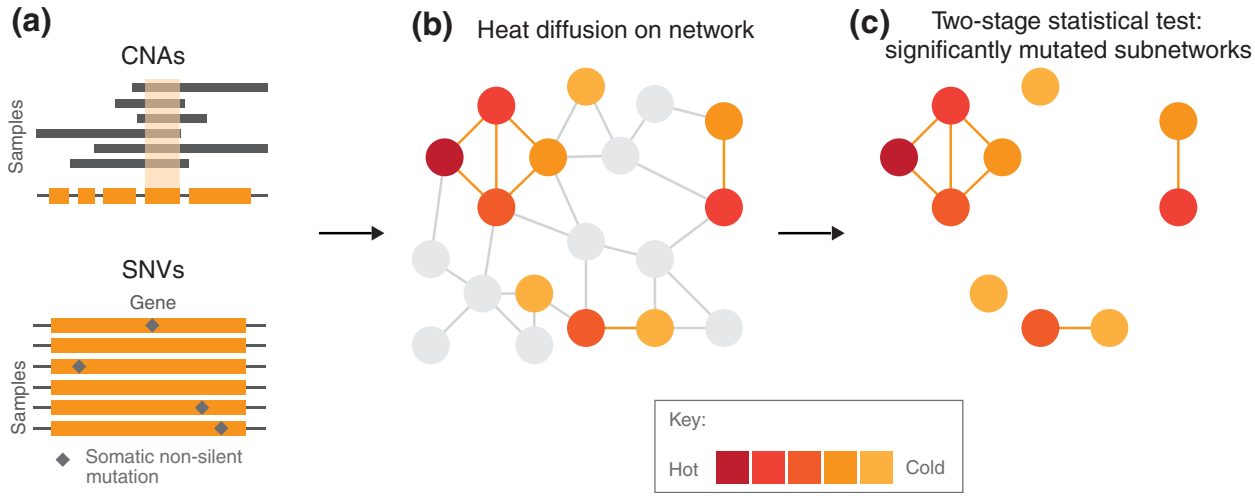

Figure 4 Overview of the HotNet algorithm. HotNet [102] uses a heat-diffusion process to identify significantly mutated subnetworks within an interaction network. **(a)** Heat is assigned to each gene according to the proportion of samples containing a single-nucleotide variant (SNV) or copy number aberration (CNA) in the gene. **(b)** The initial heat then spreads on the edges of the network for a fixed amount of time. Removing cold edges connecting genes that do not exchange large amounts of heat breaks the network into smaller subnetworks. **(c)** HotNet assesses the number and size of the resulting subnetworks using a two-stage statistical test.

by the quality and coverage of the interaction network. High-quality interaction networks derived from well-characterized experimental interactions remain relatively scarce. Thus, to increase the coverage of the network, most interaction networks are constructed using data from high-throughput screens (such as yeast two-hybrid screens or mass spectrometry), thereby increasing the number of false positives. In addition, interaction networks may suffer from ascertainment bias, as genes whose roles in cancer are well-documented are likely to have been extensively tested for interactions, whereas novel cancer genes may not have been characterized at all. Finally, nearly all currently available interaction networks are the superposition of interactions between proteins that occur in different tissues, in different cellular locations, or at different developmental time points or cell-cycle phases. Such limitations will need to be overcome in order to improve the identification of combinations of driver mutations using interaction networks.

De novo approaches

To identify novel combinations of mutations or mutated genes, it would be ideal to test all possible combinations for recurrent mutations across a cohort of cancer patients, but such a *de novo* approach is impractical. For example, there are more than 10^{29} possible sets of eight genes in the human genome, which is both too many to evaluate computationally and too many hypotheses to test while retaining statistical power. One promising approach is to restrict the possible combinations of mutations that are evaluated by focusing on those combinations that exhibit particular patterns of occurrence. One such pattern is mutual exclusivity between driver mutations. Under the hypotheses that each tumor has relatively few driver mutations [1] and these driver mutations perturb multiple cellular functions in different pathways [87], one can conclude that a tumor rarely possesses more than one driver mutation per pathway. Thus, when examining data across cancer samples, driver pathways (pathways with driver mutations) correspond to mutually exclusive sets of genes (with mutual exclusivity in individual samples). Mutually exclusive pairs of interacting proteins [106] and sets of interacting proteins [107] in the same pathway have previously been reported in many cancer types. Examples include BRAF and KRAS [108] (in the RAS-RAF signaling pathway) and APC and CTNNB1 (in the β-catenin signaling pathway), both in colorectal cancer [109], and TP53 and MDM2 in ovarian cancer [7].

A few algorithms have been developed to identify putative driver pathways by finding sets of genes that exhibit a statistically significant pattern of mutual exclusivity. Note that because many recurrently mutated genes are present in a minority of samples, mutually exclusive sets of genes will also be present just by chance; it is therefore necessary to determine the statistical significance of mutual exclusivity. The Recurrent and Mutually Exclusive (RME) [110] algorithm identifies modules with exclusive patterns of mutations using an information theoretic measure to test for the significance of the observed exclusivity. RME starts from scores that measure the exclusivity of pairs of genes, and includes only genes mutated with relatively high frequency (≥ 10% in [110]), limiting its effectiveness in identifying rare driver mutations. The *De novo* Driver Exclusivity (Dendrix) algorithm [102] identifies sets of genes that are mutated across a large number of samples and whose mutations are mutually exclusive by determining the statistical significance of the optimal set of genes of a fixed size. In data from the TCGA glioblastoma study, Dendrix identified significant exclusivity between mutations in three sets of genes that are part of the Rb pathway, the p53 pathway, and the RTK pathway, respectively [111]. Multi-Dendrix [112] simultaneously identifies multiple mutually exclusive sets of genes. In data from the TCGA breast cancer study, Multi-Dendrix identified significant exclusivity of mutations in pathways involved in p53 signaling, PI3K/AKT signaling, cell-cycle checkpoints, and p38-JNK1 signaling. Finally, the MEMo algorithm described earlier also examines pairs of genes with mutually exclusive mutations, but these are restricted to those pairs that share multiple interacting partners in an interaction network.

Approaches based on mutual exclusivity provide a strategy for assessing combinations of mutations that is less biased by prior information, but they do not consider all possible combinations of mutations. Moreover, there are examples of co-occurring driver mutations in cancer [106]. The hypothesis pertaining to mutual exclusivity is only for mutations in the same pathway, therefore co-occurring mutations do not violate this hypothesis if they are in different pathways. There are, however, examples of co-occurring mutations in genes that directly interact, such as *KRAS* and *PIK3CA* in colorectal tumors [113]. Thus, the pattern of mutual exclusivity is not enough to characterize all functional combinations of mutations.

Conclusions and future perspective

This review focused on some of the challenges in the sequencing and identification of driver mutations and driver genes in cancer genomes using high-throughput DNA sequencing. We highlighted several computational approaches that are used to detect somatic mutations and to prioritize these mutations for further experimental validation. These and other approaches are increasingly being translated from the research laboratory into the clinical setting. Several academic medical centers

have begun targeted or whole-exome sequencing of cancer patients [114-117] in order to guide clinical treatment. Such precision medicine approaches have begun to bear fruit in clinical trials in which the drug regime is tailored to the mutational landscape of the individual patient [118]. Consortiums like TCGA and ICGC are continually expanding the number of sequenced cancer genomes or exomes. Given the dividends that these and other studies have returned in only a few years, the rapid, precise computational identification of driver mutations is likely to be a key step in determining patient prognosis and treatment.

The past 5 years has witnessed a revolution in cancer genome sequencing, but additional challenges remain if the promise of high-throughput DNA sequencing for cancer diagnosis and treatment is to be fully exploited. First, non-coding somatic mutations have not yet received the same amount of scrutiny as coding variants. Huang *et al.* [119] recently discovered a mutation in the promoter region of the *TERT* gene (which encodes telomerase reverse transcriptase) that increased the transcription of *TERT* in melanoma. This observation, coupled with recent ENCODE reports that provide functional annotations for many non-coding regions of the human genome [120], indicates that the identification of intergenic driver mutations will also prove useful for understanding tumorigenesis. Second, certain cancers exhibit different subtypes, and a mixture of these subtypes can complicate the identification of recurrent mutations or combinations of mutations. Recently, Hofree *et al.* [121] introduced the Network-based stratification approach to predict subtypes with different clinical outcomes directly from mutation data, a useful step in addressing this issue. In addition, the Pan-Cancer project within TCGA showed that, in some cases, combining different cancer types improved rather than complicated the analysis [58,122-125]. Third, more work is needed to determine the extent to which different cancer types or subtypes can be analyzed together. Finally, the interpretation of somatic mutations is informed by other types of genomic and epigenomic data including RNA sequencing, DNA methylation, and chromatin modifications. Some methods have been designed to integrate across these different types of sequencing data [69,126,127], but more work is required to fully integrate the various types of information. Finally, the translation of genomic, epigenomic or transcriptomic discoveries into practical cancer treatment faces numerous hurdles in functional validation and drug design. For some patients, precision oncology is a reality now, but for many other patients, difficult but important work remains.

Abbreviations

AML: acute myeloid leukemia; BMR: background mutation rate; CaMP: Cancer Mutation Prevalence; CNA: copy number aberration; Dendrix: *De novo* Driver Exclusivity; GSEA: Gene Set Enrichment Analysis; ICGC: International Cancer Genome Consortium; nsSNV: non-silent SNV; PCR: polymerase chain reaction; RME: Recurrent and Mutually Exclusive; SCNA: somatic copy number aberration; SNP: single-nucleotide polymorphism; SNV: single-nucleotide variant; SV: structural variant; TCGA: The Cancer Genome Atlas; THetA: Tumor Heterogeneity Analysis.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Jason Hu for assistance with the figures. BJR is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P. Sloan Research Fellowship, a grant from the National Human Genome Research Institute (R01HG005690), an NSF CAREER Award (CCF-1053753) and NSF grant IIS-1016648. LO is supported by NSF Graduate Research Fellowship DGE 0228243. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

[1]Department of Computer Science, Brown University, 115 Waterman Street, Providence, RI 02912, USA. [2]Center for Computational Molecular Biology, Brown University, 115 Waterman Street, Providence, RI 02912, USA. [3]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, 185 Meeting Street, Providence, RI 02912, USA.

Published: 30 January 2014

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW: **Cancer genome landscapes.** *Science* 2013, **339:**1546–1558.
2. Garraway LA, Lander ES: **Lessons from the cancer genome.** *Cell* 2013, **153:**17–37.
3. Goldman JM, Melo JV: **Chronic myeloid leukemia – advances in biology and new approaches to treatment.** *N Engl J Med* 2003, **349:**1451–1464.
4. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M: **EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.** *Science* 2004, **304:**1497–1500.
5. Network CGA: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487:**330–337.
6. Cancer Genome Atlas Research Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455:**1061–1068.
7. Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474:**609–615.
8. Cancer Genome Atlas Research Network: **Comprehensive genomic characterization of squamous cell lung cancers.** *Nature* 2012, **489:**519–525.
9. Cancer Genome Atlas Research Network: **Comprehensive molecular characterization of clear cell renal cell carcinoma.** *Nature* 2013, **499:**43–49.
10. Cancer Genome Atlas Research Network: **Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia.** *N Engl J Med* 2013, **368:**2059–2074.
11. Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R, Benz CC, Yau C, Laird PW, Ding L, Zhang W, Mills GB, Kucherlapati R, Mardis ER, Levine DA: **Integrated genomic characterization of endometrial carcinoma.** *Nature* 2013, **497:**67–73.
12. The Cancer Genome Atlas Network: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490:**61–70.
13. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C, Godfrey AL, Rapado I, Cvejic A, Rance R, McGee C, Ellis P, Mudie LJ, Stephens PJ, McLaren S, Massie CE, Tarpey PS, Varela I, Nik-Zainal S, Davies HR, Shlien A, Jones D, Raine K, Hinton J, Butler AP, Teague JW, *et al*: **Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts.** *N Engl J Med* 2011, **365:**1384–1395.
14. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J, Bignell G, Butler A, Cho J, Dalgliesh GL, Galappaththige D,

Greenman C, Hardy C, Jia M, Latimer C, Lau KW, Marshall J, McLaren S, Menzies A, Mudie L, Stebbings L, Largaespada DA, Wessels LFA, Richard S, Kahnoski RJ, Anema J, *et al*: **Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma.** *Nature* 2011, **469:**539–542.

15. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR, Yates LR, Papaemmanuil E, Beare D, Butler A, Cheverton A, Gamble J, Hinton J, Jia M, Jayakumar A, Jones D, Latimer C, Lau KW, McLaren S, McBride DJ, Menzies A, Mudie L, Raine K, Rad R, Chapman MS, Teague J, *et al*: **The landscape of cancer genes and mutational processes in breast cancer.** *Nature* 2012, **486:**400–404.
16. Garraway LA: **Genomics-driven oncology: framework for an emerging paradigm.** *J Clin Oncol* 2013, **31:**1806–1814.
17. Soon WW, Hariharan M, Snyder MP: **High-throughput sequencing for biology and medicine.** *Mol Syst Biol* 2013, **9:**640.
18. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, *et al*: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.** *Nature* 2008, **456:**66–72.
19. Mardis ER: **Next-generation DNA sequencing methods.** *Annu Rev Genomics Hum Genet* 2008, **9:**387–402.
20. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11:**31–46.
21. Ding J, Bashashati A, Roth A, Oloumi A, Tse K, Zeng T, Haffari G, Hirst M, Marra MA, Condon A, Aparicio S, Shah SP: **Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data.** *Bioinformatics* 2012, **28:**167–175.
22. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: **Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.** *Nat Biotechnol* 2013, **31:**213–219.
23. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK: **Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs.** *Bioinformatics* 2012, **28:**1811–1817.
24. 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.
25. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, Dahlman KB, Pao W, Zhao Z: **Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers.** *Genome Med* 2013, **5:**91.
26. Ding L, Raphael BJ, Chen F, Wendl MC: **Advances for studying clonal evolution in cancer.** *Cancer Lett* 2013, **340:**212–219.
27. Nowell PC: **The clonal evolution of tumor cell populations.** *Science* 1976, **194:**23–28.
28. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G: **Absolute quantification of somatic DNA alterations in human cancer.** *Nat Biotechnol* 2012, **30:**413–421.
29. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, Perou CM, Børresen-Dale AL, Kristensen VN: **Allele-specific copy number analysis of tumors.** *Proc Natl Acad Sci U S A* 2010, **107:**16910–16915.
30. Oesper L, Mahmoody A, Raphael BJ: **THetA: inferring intra-tumor heterogeneity from high- throughput DNA sequencing data.** *Genome Biol* 2013, **14:**R80.
31. Chen M, Gunel M, Zhao H: **SomatiCA: identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data.** *PLoS One* 2013, **8:**e78143.
32. Yuan Y, Failmezger H, Rueda OM, Ali HR, Gr¨af S, Chin SF, Schwarz RF, Curtis C, Dunning MJ, Bardwell H, Johnson N, Doyle S, Turashvili G, Provenzano E, Aparicio S, Caldas C, Markowetz F: **Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling.** *Sci Transl Med* 2012, **4:**157ra143.
33. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L, Wan Y, Zhang W, Shukla SA, Vartanov A, Fernandes SM, Saksena G, Cibulskis K, Tesar B, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, Getz G, Wu CJ: **Evolution and impact of subclonal mutations in chronic lymphocytic leukemia.** *Cell* 2013, **152:**714–726.
34. Mullighan CG, Phillips LA, Su X, Ma J, Miller CB, Shurtleff SA, Downing JR: **Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia.** *Science* 2008, **322:**1377–1380.
35. Shibata D: **Cancer. Heterogeneity and tumor history.** *Science* 2012, **336:**304–305.
36. Greaves M, Maley CC: **Clonal evolution in cancer.** *Nature* 2012, **481:**306–313.
37. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, *et al*: **Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing.** *Nature* 2012, **481:**506–510.
38. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, Bashashati A, Prentice LM, Khattra J, Burleigh A, Yap D, Bernard V, McPherson A, Shumansky K, Crisan A, Giuliany R, Heravi-Moussavi A, Rosner J, Lai D, Birol I, Varhol R, Tam A, Dhalla N, Zeng T, Ma K, Chan SK, *et al*: **The clonal and mutational evolution spectrum of primary triple-negative breast cancers.** *Nature* 2012, **486:**395–399.
39. 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.
40. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, Riggs M, Eberling Y, Troge J, Grubor V, Levy D, Lundin P, Månér S, Zetterberg A, Hicks J, Wigler M: **Inferring tumor progression from genomic heterogeneity.** *Genome Res* 2010, **20:**68–80.
41. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M: **Tumour evolution inferred by single-cell sequencing.** *Nature* 2011, **472:**90–94.
42. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A, Stebbings LA, Leroy C, Jia M, Rance R, Mudie LJ, Gamble SJ, Stephens PJ, McLaren S, Tarpey PS, Papaemmanuil E, Davies HR, Varela I, McBride DJ, Bignell GR, Leung K, Butler AP, *et al*: **The life history of 21 breast cancers.** *Cell* 2012, **149:**994–1007.
43. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, Li F, Wu K, Liang J, Shao D, Wu H, Ye X, Ye C, Wu R, Jian M, Chen Y, Xie W, Zhang R, Chen L, Liu X, Yao X, Zheng H, Yu C, Li Q, Gong Z, Mao M, Yang X, Yang L, Li J, Wang W, *et al*: **Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm.** *Cell* 2012, **148:**873–885.
44. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, Li F, Tsang S, Wu K, Wu H, He W, Zeng L, Xing M, Wu R, Jiang H, Liu X, Cao D, Guo G, Hu X, Gui Y, Li Z, Xie W, Sun X, Shi M, Cai Z, Wang B, Zhong M, Li J, Lu Z, Gu N, *et al*: **Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor.** *Cell* 2012, **148:**886–895.
45. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B, Wigler M, Navin N, Hicks J: **Genome-wide copy number analysis of single cells.** *Nat Protoc* 2012, **7:**1024–1241.
46. Voet T, Kumar P, Van Loo P, Cooke SL, Marshall J, Lin ML, Zamani Esteki M, Van der Aa N, Mateiu L, McBride DJ, Bignell GR, McLaren S, Teague J, Butler A, Raine K, Stebbings LA, Quail MA, D'Hooghe T, Moreau Y, Futreal PA, Stratton MR, Vermeesch JR, Campbell PJ: **Single-cell paired-end genome sequencing reveals structural variation per cell cycle.** *Nucleic Acids Res* 2013, **41:**6119–6138.
47. Zong C, Lu S, Chapman AR, Xie XS: **Genome-wide detection of single-nucleotide and copy-number variations of a single human cell.** *Science* 2012, **338:**1622–1626.
48. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, *et al*: **Mutational heterogeneity in**

cancer and the search for new cancer-associated genes. *Nature* 2013, **499:**214–218.
49. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L: **MuSiC: identifying mutational significance in cancer genomes.** *Genome Res* 2012, **22:**1589–1598.
50. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, *et al*: **The genomic complexity of primary human prostate cancer.** *Nature* 2011, **470:**214–220.
51. Youn A, Simon R: **Identifying cancer driver genes in tumor genome sequencing studies.** *Bioinformatics* 2011, **27:**175–181.
52. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JKV, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314:**268–274.
53. Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF: **Statistical analysis of pathogenicity of somatic mutations in cancer.** *Genetics* 2006, **173:**2187–2198.
54. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, *et al*: **Initial genome sequencing and analysis of multiple myeloma.** *Nature* 2011, **471:**467–472.
55. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR: **Human mutation rate associated with DNA replication timing.** *Nat Genet* 2009, **41:**393–395.
56. Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, D'Aubenton Carafa Y, Arneodo A, Hyrien O, Thermes C: **Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes.** *Genome Res* 2010, **20:**447–457.
57. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, *et al*: **Mutational processes molding the genomes of 21 breast cancers.** *Cell* 2012, **149:**979–993.
58. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N: **Comprehensive identification of mutational cancer driver genes across 12 tumor types.** *Sci Rep* 2013, **3:**2650.
59. De S, Michor F: **DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes.** *Nat Biotechnol* 2011, **29:**1103–1108.
60. Fudenberg G, Getz G, Meyerson M, Mirny LA: **High order chromatin architecture shapes the landscape of chromosomal alterations in cancer.** *Nat Biotechnol* 2011, **29:**1109–1113.
61. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.** *Genome Biol* 2011, **12:**R41.
62. Zhang Q, Ding L, Larson DE, Koboldt DC, McLellan MD, Chen K, Shi X, Kraja A, Mardis ER, Wilson RK, Borecki IB, Province MA: **CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data.** *Bioinformatics* 2010, **26:**464–469.
63. Sanchez-Garcia F, Akavia UD, Mozes E, Pe'er D: **JISTIC: identification of significant targets in cancer.** *BMC Bioinformatics* 2010, **11:**189.
64. Walter V, Nobel AB, Wright FA: **DiNAMIC: a method to identify recurrent DNA copy number aberrations in tumors.** *Bioinformatics* 2011, **27:**678–685.
65. van Dyk E, Reinders MJT, Wessels LFA: **A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control.** *Nucleic Acids Res* 2013, **41:**e100.
66. Ritz A, Paris PL, Ittmann MM, Collins C, Raphael BJ: **Detection of recurrent rearrangement breakpoints from copy number data.** *BMC Bioinformatics* 2011, **12:**114.
67. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, *et al*: **Massive genomic rearrangement acquired in a single catastrophic event during cancer development.** *Cell* 2011, **144:**27–40.
68. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, Van Allen E, Kryukov GV, Sboner A, Theurillat JP, Soong TD, Nickerson E, Auclair D, Tewari A, Beltran H, Onofrio RC, Boysen G, Guiducci C, Barbieri CE, Cibulskis K, Sivachenko A, Carter SL, Saksena G, Voet D, Ramos AH, Winckler W, *et al*: **Punctuated evolution of prostate cancer genomes.** *Cell* 2013, **153:**666–677.
69. McPherson A, Wu C, Wyatt AW, Shah S, Collins C, Sahinalp SC: **nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing.** *Genome Res* 2012, **22:**2250–2061.
70. Malhotra A, Lindberg M, Faust GG, Leibowitz ML, Clark RA, Layer RM, Quinlan AR, Hall IM: **Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms.** *Genome Res* 2013, **23:**762–776.
71. Venables JP: **Aberrant and alternative splicing in cancer.** *Cancer Res* 2004, **64:**7647–7654.
72. 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.
73. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP: **Predicting the functional effect of amino acid substitutions and indels.** *PloS One* 2012, **7:**e46688.
74. 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.
75. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics.** *Nucleic Acids Res* 2011, **39:**e118.
76. Fischer A, Greenman C, Mustonen V: **Germline fitness-based scoring of cancer mutations.** *Genetics* 2011, **188:**383–393.
77. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R: **Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.** *Cancer Res* 2009, **69:**6660–6667.
78. Li Y, Wen Z, Xiao J, Yin H, Yu L, Yang L, Li M: **Predicting disease-associated substitution of a single amino acid by analyzing residue interactions.** *BMC Bioinformatics* 2011, **12:**14.
79. Gonzalez-Perez A, Lopez-Bigas N: **Functional impact bias reveals cancer drivers.** *Nucleic Acids Res* 2012, **40:**e169.
80. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, *et al*: **Mutations of the BRAF gene in human cancer.** *Nature* 2002, **417:**949–954.
81. Bos JL: **The ras gene family and human carcinogenesis.** *Mutat Res* 1988, **195:**255–271.
82. Olivier M, Hollstein M, Hainaut P: **TP53 mutations in human cancers: origins, consequences, and clinical use.** *Cold Spring Harb Perspect Biol* 2010, **2:**a001008.
83. Ye J, Pavlicek A, Lunney EA, Rejto PA, Teng CH: **Statistical method on nonrandom clustering with application to somatic mutations in cancer.** *BMC Bioinformatics* 2010, **11:**11.
84. Ryslik GA, Cheng Y, Cheung KH, Modis Y, Zhao H: **Utilizing protein structure to identify non-random somatic mutations.** *BMC Bioinformatics* 2013, **14:**190.
85. Ubersax JA, Ferrell JE: **Mechanisms of specificity in protein phosphorylation.** *Nat Rev Mol Cell Biol* 2007, **8:**530–541.
86. Bader AG, Kang S, Zhao L, Vogt PK: **Oncogenic PI3K deregulates transcription and translation.** *Nature Rev Cancer* 2005, **5:**921–929.
87. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144:**646–674.
88. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44–57.

89. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37:**1–13.
90. Al-Shahrour F, Díaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes.** *Bioinformatics* 2004, **20:**578–580.
91. Beissbarth T, Speed TP: **GOstat: find statistically overrepresented gene ontologies within a group of genes.** *Bioinformatics* 2004, **20:**1464–1465.
92. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545–15550.
93. Lin J, Gan CM, Zhang X, Jones S, Sjöblom T, Wood LD, Parsons DW, Papadopoulos N, Kinzler KW, Vogelstein B, Parmigiani G, Velculescu VE: **A multidimensional analysis of genes mutated in breast and colorectal cancers.** *Genome Res* 2007, **17:**1304–1318.
94. Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G: **Patient-oriented gene set analysis for cancer mutation data.** *Genome Biol* 2010, **11:**R112.
95. Wendl MC, Wallis JW, Lin L, Kandoth C, Mardis ER, Wilson RK, Ding L: **PathScan: a tool for discerning mutational significance in groups of putative cancer genes.** *Bioinformatics* 2011, **27:**1595–1602.
96. McCormick F: **Signalling networks that cause cancer.** *Trends Cell Biol* 1999, **9:**M53–M56.
97. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, *et al*: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Res* 2003, **13:**12363–12371.
98. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34:**D535–D539.
99. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P: **Reactome knowledgebase of human biological pathways and processes.** *Nucleic Acids Res* 2009, **37:**D619–D622.
100. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ: **STRING v9.1: protein-protein interaction networks, with increased coverage and integration.** *Nucleic Acids Res* 2013, **41:**D808–D815.
101. Turner B, Razick S, Turinsky AL, Vlasblom J, Crowdy EK, Cho E, Morrison K, Donaldson IM, Wodak SJ: **iRe-fWeb: interactive analysis of consolidated protein interaction data and their supporting evidence.** *Database* 2010, **2010:**baq023.
102. Vandin F, Upfal E, Raphael BJ: **Algorithms for detecting significantly mutated pathways in cancer.** *J Comput Biol* 2011, **18:**507–522.
103. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, Quist MJ, Jing X, Lonigro RJ, Brenner JC, Asangani IA, Ateeq B, Chun SY, Siddiqui J, Sam L, Anstett M, Mehra R, Prensner JR, Palanisamy N, Ryslik GA, Vandin F, Raphael BJ, Kunju LP, Rhodes DR, Pienta KJ, Chinnaiyan AM, Tomlins SA: **The mutational landscape of lethal castration-resistant prostate cancer.** *Nature* 2012, **487:**239–243.
104. Ciriello G, Cerami E, Sander C, Schultz N: **Mutual exclusivity analysis identifies oncogenic network modules.** *Genome Res* 2012, **22:**398–406.
105. Ciriello G, Cerami E, Aksoy BA, Sander C, Schultz N: **Using MEMo to discover mutual exclusivity modules in cancer.** *Curr Protoc Bioinformatics* 2013, **:**. Mar; Chapter 8:Unit 8.17.
106. Yeang CH, McCormick F, Levine A: **Combinatorial patterns of somatic gene mutations in cancer.** *FASEB J* 2008, **22:**2605–2622.
107. Thomas RK, Baker AC, Debiasi RM, Winckler W, Laframboise T, Lin WM, Wang M, Feng W, Zander T, MacConaill L, Macconnaill LE, Lee JC, Nicoletti R, Hatton C, Goyette M, Girard L, Majmudar K, Ziaugra L, Wong KK, Gabriel S, Beroukhim R, Peyton M, Barretina J, Dutt A, Emery C, Greulich H, Shah K, Sasaki H, Gazdar A, Minna J, *et al*: **High-throughput oncogene mutation profiling in human cancer.** *Nat Genet* 2007, **39:**347–351.
108. Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE: **Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status.** *Nature* 2002, **418:**934.
109. Sparks AB, Morin PJ, Vogelstein B, Kinzler KW: **Mutational analysis of the APC/beta-catenin/Tcf pathway in colorectal cancer.** *Cancer Res* 1998, **58:**1130–1134.
110. Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A: **Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors.** *BMC Med Genomics* 2011, **4:**34.
111. Vandin F, Upfal E, Raphael BJ: ***De novo* discovery of mutated driver pathways in cancer.** *Genome Res* 2012, **22:**375–385.
112. Leiserson MDM, Blokh D, Sharan R, Raphael BJ: **Simultaneous identification of multiple driver pathways in cancer.** *PLoS Comput Biol* 2013, **9:**e1003054.
113. Parsons DW, Wang TL, Samuels Y, Bardelli A, Cummins JM, DeLong L, Silliman N, Ptak J, Szabo S, Willson JKV, Markowitz S, Kinzler KW, Vogelstein B, Lengauer C, Velculescu VE: **Colorectal cancer: mutations in a signalling pathway.** *Nature* 2005, **436:**792.
114. Dias-Santagata D, Akhavanfard S, David SS, Vernovsky K, Kuhlmann G, Boisvert SL, Stubbs H, McDermott U, Settleman J, Kwak EL, Clark JW, Isakoff SJ, Sequist LV, Engelman JA, Lynch TJ, Haber DA, Louis DN, Ellisen LW, Borger DR, Iafrate AJ: **Rapid targeted mutational analysis of human tumours: a clinical platform to guide personalized cancer medicine.** *EMBO Mol Med* 2010, **2:**146–158.
115. Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, Ducar M, Van Hummelen P, Macconaill LE, Hahn WC, Meyerson M, Gabriel SB, Garraway LA: **High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing.** *Cancer Discov* 2012, **2:**82–93.
116. MacConaill LE, Campbell CD, Kehoe SM, Bass AJ, Hatton C, Niu L, Davis M, Yao K, Hanna M, Mondal C, Luongo L, Emery CM, Baker AC, Philips J, Goff DJ, Fiorentino M, Rubin MA, Polyak K, Chan J, Wang Y, Fletcher JA, Santagata S, Corso G, Roviello F, Shivdasani R, Kieran MW, Ligon KL, Stiles CD, Hahn WC, Meyerson ML, Garraway LA: **Profiling critical cancer gene mutations in clinical tumor samples.** *PLoS One* 2009, **4:**e7887.
117. Thomas RK, Baker AC, Debiasi RM, Winckler W, Laframboise T, Lin WM, Wang M, Feng W, Zander T, MacConaill L, MacConnaill LE, Lee JC, Nicoletti R, Hatton C, Goyette M, Girard L, Majmudar K, Ziaugra L, Wong KK, Gabriel S, Beroukhim R, Peyton M, Barretina J, Dutt A, Emery C, Greulich H, Shah K, Sasaki H, Gazdar A, Minna J, *et al*: **High-throughput oncogene mutation profiling in human cancer.** *Nature Genetics* 2007, **39:**347–351.
118. Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, Stewart DJ, Hicks ME, Erasmus J, Gupta S, Alden CM, Liu S, Tang X, Khuri FR, Tran HT, Johnson BE, Heymach JV, Mao L, Fossella F, Kies MS, Papadimitrakopoulou V, Davis SE, Lippman SM, Hong WK: **The BATTLE trial: personalizing therapy for lung cancer.** *Cancer Discov* 2011, **1:**44–53.
119. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA: **Highly recurrent TERT promoter mutations in human melanoma.** *Science* 2013, **339:**957–959.
120. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Ernst J, *et al*: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57–74.
121. Hofree M, Shen JP, Carter H, Gross A, Ideker T: **Network-based stratification of tumor mutations.** *Nat Methods* 2013, **10:**1108–1115.
122. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nat Genet* 2013, **45:**1113–1120.
123. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L: **Mutational landscape and significance across 12 major cancer types.** *Nature* 2013, **502:**333–339.
124. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C: **Emerging landscape of oncogenic signatures across human cancers.** *Nat Genet* 2013, **45:**1127–1133.
125. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, Getz G, Meyerson M, Beroukhim R:

Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 2013, **45:**1134–1140.
126. 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.
127. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM: **Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE).** *Bioinformatics* 2013, **29:**2757–2764.
128. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany 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.
129. Xi R, Luquette J, Hadjipanayis A, Kim TM, Park PJ: **BIC-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data.** *Genome Biol* 2010, **11:**O10.
130. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, Zhang C, Ren X, Protopopov A, Chin L, Kucherlapati R, Lee C, Park PJ: **Diverse mechanisms of somatic structural variations in human cancer genomes.** *Cell* 2013, **153:**919–929.
131. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, NHLBI Exome Sequencing Project, Quinlan AR, Nickerson DA, Eichler EE: **Copy number variation detection and genotyping from exome sequence data.** *Genome Res* 2012, **22:**1525–1532.
132. 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.
133. Hormozdiari F, Hajirasouliha I, McPherson A, Eichler EE, Sahinalp SC: **Simultaneous structural variation discovery among multiple paired-end sequenced genomes.** *Genome Res* 2011, **21:**2203–2212.
134. Sindi S, Helman E, Bashir A, Raphael BJ: **A geometric approach for classification and comparison of structural variants.** *Bioinformatics* 2009, **25:**i222–i230.
135. Sindi SS, Onal S, Peng LC, Wu HT, Raphael BJ: **An integrative probabilistic model for identification of structural variation in sequencing data.** *Genome Biol* 2012, **13:**R22.
136. Escaramís G, Tornador C, Bassaganyas L, Rabionet R, Tubio JMC, Martínez-Fundichely A, Cáceres M, Gut M, Ossowski S, Estivill X: **PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data.** *PLoS One* 2013, **8:**e63377.
137. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO: **DELLY: structural variant discovery by integrated paired-end and split-read analysis.** *Bioinformatics* 2012, **28:**i333–i339.
138. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, Getz G: **Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability.** *Genome Res* 2013, **23:**228–235.
139. Hua X, Xu H, Yang Y, Zhu J, Liu P, Lu Y: **DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies.** *Am J Hum Genet* 2013, **93:**439–451.
140. Cerami E, Demir E, Schultz N, Taylor BS, Sander C: **Automated network analysis identifies core pathways in glioblastoma.** *PLoS One* 2010, **5:**e8918.

doi:10.1186/gm524
Cite this article as: Raphael *et al.*: **Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine.** *Genome Medicine* 2014 **6**:5.