Published in final edited form as:
Nat Rev Genet. 2014 August ; 15(8): 556–570. doi:10.1038/nrg3767.

Expanding the computational toolbox for mining cancer genomes

Li Ding[1,2,3,4,#], **Michael C. Wendl**[1,3,5], **Joshua F. McMichael**[1], and **Benjamin J. Raphael**[6]

[1]The Genome Institute, Washington University in St. Louis, 4444 Forest Park Ave, St. Louis, MO 63108, USA

[2]Department of Medicine, Washington University in St. Louis, 660 S. Euclid Ave., St. Louis, MO 63110, USA

[3]Department of Genetics, Washington University in St. Louis, 660 S. Euclid Ave., St. Louis, MO 63110, USA

[4]Siteman Cancer Center, Washington University in St. Louis, 4921 Parkview Place, St. Louis, MO 63110, USA

[5]Department of Mathematics, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130, USA

[6]Department of Computer Science and Center for Computational Molecular Biology, Brown University, 115 Waterman Street, Providence, RI 02912, USA

Abstract

High-throughput DNA sequencing has revolutionized cancer genomics with numerous discoveries relevant to cancer diagnosis and treatment. The latest sequencing and analysis methods have successfully identified somatic alterations including single nucleotide variants (SNVs), insertions and deletions (indels), structural aberrations, and gene fusions. Additional computational techniques have proved useful to define those mutations, genes, and molecular networks that drive diverse cancer phenotypes as well as determine clonal architectures in tumour samples. Collectively, these tools have advanced the study of genomic, transcriptomic, epigenomic

[#]Corresponding Author: Li Ding, Ph.D, The Genome Institute, Division of Oncology, Department of Medicine, Washington University School of Medicine, St. Louis, MO 63108, lding@genome.wustl.edu.

Online links
Ensembl - http://www.ensembl.org/Homo_sapiens/Info/Index
UCSC - http://genome.ucsc.edu/cgi-bin/hgGateway
GENCODE - http://www.gencodegenes.org/data.html
RefSeq - http://www.ncbi.nlm.nih.gov/refseq/
ENCODE - http://www.genome.gov/Encode/
TransFac - http://www.gene-regulation.com/pub/databases.html
RegulomeDB - http://www.regulomedb.org/
Noncode - http://www.noncode.org/
BodyMap - http://www.illumina.com/science/data_library.ilmn
http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/
miRBase - http://www.mirbase.org/
Pfam - http://pfam.sanger.ac.uk/
Interpro - https://www.ebi.ac.uk/interpro/

alterations and their association to clinical properties. Here, we review cancer genomics software and the insights that have been gained from their application.

Introduction

Fred Sanger and colleagues jump-started the nascent field of genomics in 1977 with their development of chain-termination DNA sequencing[1,2]. It founded a series of commercial instruments that helped produce numerous early milestones, including the sequence of the first human genome[3]. Work was slow and expensive (the Human Genome Project rang-up about 1 billion dollars) and enormous gains in economy and speed would be needed before the approach could be applied widely. Enter 'next generation sequencing', the generic name for a raft of advanced techniques, including pyrosequencing**[G]**, sequencing-by-ligation**[G]** and sequencing-by-synthesis**[G]**. State-of-the-art instruments now process a whole genome in less than a week and for nominally less than ten-thousand dollars. Many thousands of genomes and exomes have since been sequenced and their data have had an enormous impact on cancer research. Cancer genomics is a now-recognized sub-specialty that grew out of adapting sequencing for cancer research. It broadly seeks to characterize germline variants and somatic mutations in the individual, to use such data from cohorts to identify driver mutations**[G]**, germline predispositions and environmental factors related to cancer and, ultimately, to synthesize such information into mechanistic theories and to develop information systems to assist clinicians with diagnosis and treatment decisions.

Aside from instrument advancements, cancer genomics owes a considerable debt to computing hardware and software. Biology has been steadily absorbing the knowledge, techniques and analytical culture of computer science and mathematics, and this has enabled the development of workhorse algorithms for sequence alignment, detection of somatic events and identification of significantly mutated genes**[G]** (SMGs). However, expansion in computing power is no longer pacing increases in instrument throughput, meaning the bottleneck is quickly shifting from data generation to data analysis. Taken with newer high-throughput streams, like RNA and protein sequences, as well as incorporation of data-intensive diagnostics like imaging, and the scope of the problem is clear; As the gap between the investigator's abilities to generate and analyze data grows, genomics will increasingly experience the kinds of "Big Data" pains already familiar to other data-centric disciplines like particle physics. One of the foremost issues will be integrating the grand corpus of these many data types to open new frontiers in research.

The field has advanced substantially since the first cancer genome was sequenced, a mere 5 years ago[4]. Whole-genome, exome and RNA-sequencing are now routinely used in cancer studies and tools continue to be deployed for even more sophisticated analysis, for example combining genome and RNA-seq data for detecting fusion genes and interpreting cancer genomes across multiple patients to discover driver mutations and pathways. Such analyses have led to discovery of new cancer genes and cancer-causing mutations and have demonstrated how environmental exposure leads to characteristic mutational spectra. In this review, we discuss state-of-the-art data generation in cancer genomics, as well as current methods for pre-processing the raw data to detect signals and higher-level analysis of

individuals (Level I) and cohorts (Level II) for research questions and clinical application (Figure 1). Moreover, we remark on some important open problems, and speculate on where the field is moving in the next several years.

Sequencing strategies

"Sequencing" is a broad term for interrogating a variety of molecular entities, including an entire static genome (whole genome sequencing)[5], strictly the coding genomic regions (exome sequencing)[6], the transcriptome[7] as a snapshot of mRNA presence at a given time and tissue location, genomic methylation patterns[8], and peptides (protein sequence). Because coding genomic sequences comprise only 1–2% of the genome, the cost for exome sequencing is still appreciably lower than for whole genome sequencing. However, differences are gradually becoming less important, as technology improvements continue to decrease overall sequencing costs. Despite its higher cost, whole-genome sequencing might be preferable, as it provides information on structural and non-coding variants, which cannot be captured from exome-only data. Whole genome data are therefore considered to be the unbiased "gold standard"[9] and the field is likely to shift increasingly towards this form of data.

Traditional sequencing analysis

For the individual cancer patient, the immediate goal of any sequencing procedure is to identify germline and somatic variants linked to the cancer phenotype. Typically, tumour and normal tissue samples are collected, sequenced, aligned to the reference genome, and compared against each other to identify genomic differences (Figure 1). Many of the reported differences represent bona fide somatic aberrations, but such findings are ideally validated by more comprehensive data from an independent platform. There are many different kinds and sizes of mutational events, for example single nucleotide variations (SNVs), copy number aberrations (CNAs) and small insertions and deletions (collectively referred to as indels), each of which is detected through specific algorithmic methods. In actual practice, detection of all germline and somatic aberrations is a formidable challenge, due to limitations in current analysis algorithms (discussed below) and the quantity/quality of sequence data. Important events may be missed when sequence coverage is too low, or when repetitive or complex genomic regions complicate the alignment and assembly of sequence variants. Sequence coverage theory **[G]** has co-evolved with sequencing technologies and predicts that data must increase as we seek to identify increasingly subtle genomic signals (**see** Box 1).

Box 1

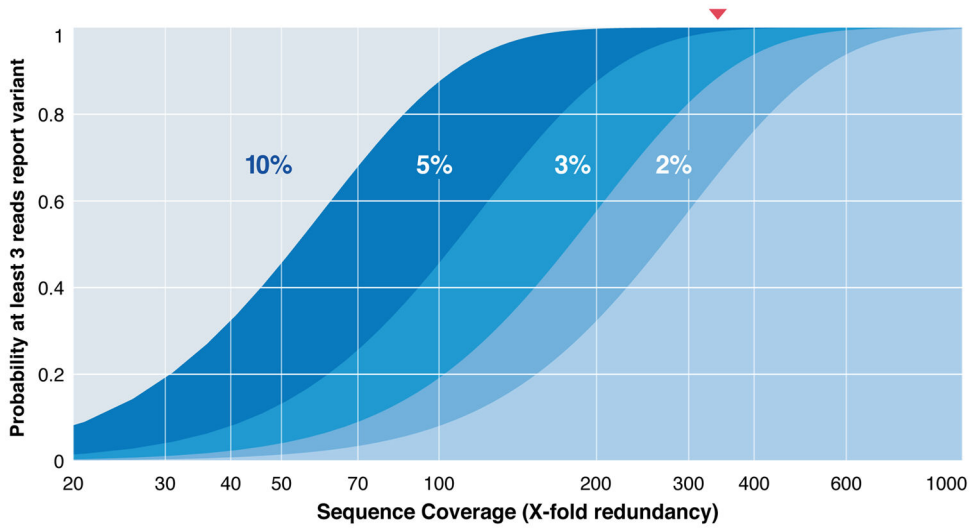
Coverage Considerations

Early sequencing projects were based on Lander-Waterman theory[155] for haploid coverage, which recommended a redundancy factor, ρ, of around 10X. Absent biases, this implies that loci are spanned by an average of 10 sequence reads and that >99.99% of the genome is represented by the data[155]. However, for medical applications, *both* alleles must be reliably identified. If the minimum condition is ≥3 spanning reads per allele, a figure of $\rho \approx 30X$ is then required for attaining roughly the same 99.99% standard[156].

This has been the *de facto* redundancy for cancer sequencing projects with respect to detecting SNVs, although it does not speak to other types of events. For example, presence of an indel is suggested whenever the reference-aligned average length of spanning fragments is significantly different from the average fragment length of the originating library.[52] Application of this principle[49,157] is tricky, because a genuine event has to be distinguished from cases in which predominantly shorter or longer fragments were sampled merely by chance. At 30X, the size range of insertions for which indels can be detected with low Type I error is rather narrow, suggesting more data are needed[158]. Data requirements elevate further if somatic events within subclones of a tumor are to be identified. For a subclone whose mass is a fraction μ of the total tumor, the probability of at least 3 reads reporting a heterozygous variant is

$$P=1-e^{-R}\left(1+R+R^2/2\right) \qquad (1)$$

where $R=\mu\rho/2$. In a 5% mass subclone, the remarkable figure of 340X data is required if 99% of subclonal variants should meet the detection conditions (Box 1 Figure). Smaller subclones require even more data, readily topping 500X in certain instances. The biomedical relevance of subclones, coupled with the growing throughput of instruments means that amounts of data generated for cancer genomics projects will continue to increase.



Box 1 Figure. Data requirements for capturing heterozygous variants

Identifying a single-nucleotide variant (SNV) requires its observation in multiple reads, usually at least 3, but accrual of these reads is governed by the random dynamics of sampling and coverage, quantified in the ideal case (pure samples, perfect data, and no sequence bias) by Eq. (1) for various tumour mass fractions. Data requirements are pushed appreciably higher by subclones that comprise smaller fractions of the entire tumour mass. Red triangle indicates redundancy of 340X for 99% probability of observing ≥3 reads in a 5% subclone.

Subclonal analysis

As sequencing costs continue to decline, researchers are able to sequence tumor samples more deeply, enabling new analyses. For example, cancer progression has long been known to be a fundamentally clonal process[10] and sequence coverages are now becoming sufficiently large to permit detection of the low prevalence events routinely associated with tumour subclones[11]. In recent years, multi-site and/or stage sequencing and tumour sectioning experiments have begun to identify founding clones and subclones contributing to cancer progression[11–14] (Figure 1A). Mutations in subclones are typically mapped at low variant allele fraction (VAF) and often occur against a background of impure tumour and/or normal sample collection. Their identification is extremely difficult, an observation that is partially quantified by coverage theory (Box 1).

Single cell sequencing

Pioneering work on assessing CNAs in multiple tumor subpopulations[13] was followed by single-cell sequencing[15] using whole genome amplification (WGA) of DNA extracted from flow sorted nuclei (Figure 1A). Single-cell DNA and RNA sequencing are now routinely used for revealing cellular diversities within a tumour. However, there are still important challenges, such as amplification biases from degenerate oligonucleotide-primed WGA[15] and multiple displacement amplification techniques[16,17]. Biases lead to uneven coverage and consequent difficulties for identifying somatic alterations, including SNVs, CNAs and structural aberrations. Sensitivity is most affected by allele dropout, owing to the preferential amplification of one of two alleles, with rates of 8 to 40%[16,18] reported. Large CNAs can still be examined with low genome coverage (e.g., 5–6%) by computing read counts in variable-sized bins[19], whereas unequal coverage renders analysis of smaller copy number and structural variants extremely difficult. Despite these challenges, recent advances, such as assembly algorithms that handle uneven sequence coverage[20], point to widespread application of single-cell sequencing in the future.

Dissecting genomic changes in cancer

Somatic aberrations contain crucial information about the mechanisms of tumour development, progression, and metastasis/relapse. In addition, a subset that are “clinically actionable” have important implications in inferring prognosis and guiding decisions about treatment. The need for accurately identifying these events has spawned a wide collection of algorithms and software (Figure 1B). Most tools use either a composite statistical “score” or a formal probability test, though simple heuristic thresholds persist too. Table 1 lists some of the more widely used algorithms for detection of SNVs, indels, structural variants, fusions, as well as for additional analyses, including driver gene identification.

Despite impressive progress, the variant calling problem remains unresolved and we believe there is yet appreciable room for improvement in algorithm sophistication and accuracy. Numerous ad-hoc procedures have been developed to wring more performance out of existing tools. For example, it is known empirically that a candidate event called by several independent algorithms is significantly less likely to be a false positive than if it were called by any single one alone. Consequently, multi-caller strategies have now become more

common, where several detectors[21] are used under a "majority rules" aegis. Of course, sensitivity can suffer, as the overall discovery power now depends jointly on the individual tools' powers. Although preliminary work[22] has been reported, there are no conclusive studies that recommend specific tool combinations for optimally balancing Type I **[G]** versus Type II errors **[G]**, perhaps because such studies require enormous computation. As there are more than two dozen published SNV-specific detectors alone, a 3-caller approach would demand evaluating more than 2000 possible combinations. Knowing the best algorithm combinations for various types of somatic events would be tremendously valuable to the community.

Single nucleotide variant detection

SNVs are the most frequent alterations in cancer genomes. Numerous SNV detection algorithms have been developed, including GATK[23] [This is a broad and widely-used toolkit for variant discovery and data processing.], VarScan[24,25] [VarScan is one of the early programs for single-nucleotide somatic detection and has since added additional capability for germline, copy number, and indel events.], SAMtools[26] [SAMtools is a broad set of utilities for processing sequence data in the standardized SAM/BAM format, including variant calling.], SomaticSniper[27], Mutect[28] [MuTect is a widely used program for identifying single nucleotide somatic events in tumor-normal pair sequencing data.], Strelka[29], and JointSNVMix/SNVMix[30,31] (Figure 1B and Table 1). The first three handle both germline and somatic variants, whereas the others were designed for calling somatic mutations using tumour and matched normal genomic sequences. Although heterozygous variant allele frequencies of 50% are expected in germline samples, this number often does not hold for somatic sites in tumours, mainly owing to normal contamination and/or tumour heterogeneity. Algorithm development is now focusing on handling somatic mutations over a wide range of variant allele fractions. One example is the Bassovac algorithm that considers dependence upon bi-directional impurities and tumour subclonal structures (heterogeneity) at the read level, a necessary condition for avoiding *ad hoc* modeling and heuristics (Wendl *et al.*, unpublished observations). Preliminary findings show improved performance, especially for events having low allelic fractions.

Indel detection

Indel detection is still challenging, largely owing to both their lower frequencies compared with those of SNVs[32,33] and to mapping difficulties[34]. Although existing alignment tools are adequate for mapping reads containing SNVs, they lack the necessary accuracy and sensitivity for reads that overlap with indels or structural variants. Most tools by default allow for only two mismatches and no gaps in 'seeded' regions (that is in the first 28 bp in a read), which prohibits indel-containing reads from aligning to the reference. Paired-end mapping **[G]** is tremendously helpful in identifying larger indels when the ends occur in flanking regions, enabling inference of altered intervening sequences (Box 1). Gapped alignment **[G]**, split read **[G]**, and *de novo* assembly **[G]** are common approaches for detecting indels. VarScan[25] and GATK Unified Genotyper[23] are based on heuristics for indel calling using raw statistics, such as coverage, numbers of indel-supporting reads, read mapping qualities and mismatch counts.

Many existing tools[23,25,26] work well for detecting short indels (<5–8 bp), but suffer from lack of precision **[G]** (Figure 1B and Table 1). Further, they often cannot detect medium-size indels, including some known druggable and/or prognostic events, using short read data. For example, internal tandem duplications of *FLT3* (*FLT3 ITD)*, present in 20% of patients with acute myeloid leukaemia (AML) and associated with poor prognosis[35], are often overlooked because of mapping difficulties[36]. Finally, detection around low-complexity regions (such as homopolymers) is particularly challenging. SAMtools[26] finds short indels by correcting for the effect of flanking tandem repeats, usually producing a large number of indel calls in the process. Dindel[37] applies a Bayesian approach for calling small (<50 nucleotides) indels by realigning previously-mapped reads to generate candidate haplotypes and computes a posterior probability for each haplotype for downstream analysis. Conversely, Pindel[38] [Pindel is focused on identifying breakpoints at single base resolution of indels, inversions, and tandem duplications.] takes a pattern growth approach borrowed from protein data analysis[39] to detect indel breakpoints using both split reads and paired-end reads. A similar approach is employed in Delly[40]. Pindel achieves high precision and its sensitivity has been improved by allowing mismatches during the pattern matching process. The recent application of BWA-MEM[41] alignment allows better mapping around long indels and structural variants. Moreover, local *de novo* assembly or multiple alignments around candidate indel sites (for example, using GATK haplotype caller and TIGRA local assembly[42]) reduce the number of false-positive indels. This process is utilized in many pipelines for indel detection.

Copy number aberration and structural variant detection

Unlike SNVs or small indels, CNAs typically affect more than one gene. Traditionally, single nucleotide polymorphism (SNP) genotyping data have been utilized for studying CNAs in cancer, and the CNA landscape across multiple cancer types has been reported[43,44]. Accurate inference of copy number from sequence data requires normalization procedures that consider certain biases inherent to short read sequencing methods (such as GC content and library biases). Approaches have been implemented for both GC-based coverage normalization and mapping bias[45,46]. GISTIC[47] [GISTIC is one of the standard tools for finding genes affected by copy number changes that have a bearing on cancer initiation or progression.] and CMDS[48] have been developed for the identification of recurrent CNVs.

Structural changes in chromosomes, such as chromosome deletions, insertions, inversions and translocations represent another major source of somatic variation in cancer genomes. The majority of known cancer genes are affected in varying degrees by rearrangements that result in either a fusion transcript or transcriptional dysregulation. Cytogenetics, spectral karyotyping, and fluorescent *in situ* hybridization have previously identified large chromosomal events in multiple cancer types (such as the BCR–ABL translocation in chronic myelogeneous leukaemia (CML)). Early end-sequencing profiling of bacterial artificial chromosome (BAC) or Fosmid libraries revealed complex chromosomal architectures in several human cancers[49–53]. In recent years, high-throughput whole-genome sequencing of tumour samples has further improved the ability to detect somatic rearrangements and to characterize their breakpoints with base pair resolution. The

identification and analysis of read pairs that do not align as anticipated enable the detection of a wide range of structural alterations, including deletions, tandem or inverted duplications, inversions, insertions and translocations in many cancer genomes. BreakDancer[54] [BreakDancer is a general tool for identifying structural variations, including insertions, deletions, inversions, and translocations using the concept of discordant read pairs.], CREST[55], VariationHunter[56], GASV-Pro[57,58] and GenomeSTRIP[59] are among the pioneering and most popular algorithms for such analysis (Figure 1B and Table 1). BreakDancer performs *de novo* prediction of deletions, insertions, inversions and translocations based on a Poisson model for the number of supporting reads, size of anchoring regions and overall genome coverage. CREST utilizes the soft-clipping performed by the software package Burrows Wheeler aligner (BWA) and similar aligners to predict diverse structural events. GASV analyzes structural variants, improving breakpoint identification using a geometric bounding algorithm; GASV-Pro extends this approach incorporating read depth to further improve variant calls. GenomeSTRIP characterizes genome deletion polymorphism using population-level concepts to reinterpret the technical features of sequence data that often reflect structural variation. Although these approaches are quite sensitive, the paired-end strategy tends to yield many false positives owing to sequencing errors or read mis-alignments, especially within repetitive sequences. As for indel detection, local assembly is also widely considered to be a reliable supplement for reducing false positives and improving breakpoint resolution of structural variants.

Fusion detection

The expression of gene fusions that arise through genomic structural rearrangements is a major mechanism for tumour initiation and progression. *BCR–ABL1* in CML[60], *PML–RARa* in acute promyelocytic leukemia[61,62] and *TMPRSS2–ERG* in prostate cancer[63] are among the most recurrent, functional gene fusions identified to date. Recently, algorithms such as Tophat-fusion[64], deFuse[65], MapSplice[66], ChimeraScan[67] and BreakFusion[68] have been developed to detect fusions from RNA sequencing data (Figure 1B and Table 1). These tools are algorithmically similar to their genomic counterparts, although they focus primarily on mapping and ascertaining novel sequence junctions produced by mRNA-splicing and depend more on genome annotations. It is increasingly clear that fusions can arise from both simple translocations involving only two distal genomic loci[60] and complex rearrangements consisting of multiple distal loci[69,70]. Therefore, concurrent detection of gene fusions and the originating rearrangements using systematic approaches can improve the accuracy of predictions, as well as help to delineate the underlying mechanistic aspects of gene fusion products. Two tools, Comrad[71] and nFuse[72], were developed to address this challenge. Both align raw whole-genome and RNA sequencing reads, while simultaneously corroborating fusions and genomic breakpoint discovery. Comrad, which was the first to be developed, only maps a single fusion breakpoint to a single genomic breakpoint through the application of a set of *ad hoc* rules. An extension, nFuse maps fusion breakpoints to complex structural rearrangements using a graph-theoretic approach. Their advantage is that they account for ambiguous read alignment and therefore minimize errors caused by misalignments. We have recently developed BreakTrans[73], which jointly analyzes whole-genome and transcriptome sequencing data to test hypotheses produced by other tools, such as Tophat-fusion, MapSplice, BreakDancer and CREST, for further delineating the mechanistic components of

gene fusions. Variants of various types and sizes described above require sophisticated tools for annotating and interpreting their effects and significance.

Variant Annotation and Prediction of Driver Mutations and Pathways

Following the identification of somatic alterations, the next challenge is to distinguish driver mutations from passenger mutations **[G]**. Because of the ease of assessing the recurrence and frequency of somatic mutations relative to the efforts necessary to validate their function, many computational and statistical techniques have been introduced to predict driver mutations and genes. These techniques can be divided into three general categories based on their underlying strategies: variant effect prediction; recurrence/frequency assessment; and pathway/network analysis.

Annotations and functional predictions

Recent years have seen consolidation of various genome annotation databases into centralized sources, with great improvement in quality and comprehensiveness. Ensembl and UCSC have emerged as leading repositories of genes and transcripts from GENCODE and Refseq; regulatory elements identified by ENCODE, TransFac and RegulomeDB; noncoding RNAs from Noncode, BodyMap and MiRBase; and protein annotations from Pfam and Interpro (**see online links**). There has been a concurrent emergence of software that leverages these resources to perform genome-wide annotation of variants in coding and non-coding regions. Annovar[74] [Annovar is a versatile and widely-used tool for functional annotation of variants. It is often accessed through its web interface wAnnovar] and SNPeff[75] provide annotation of transcript variants, SKIPPY[76] predicts cryptic splice effectors, and Ensembl VEP, FunSeq[77] and SNPnexus[78] all extend support to include annotation of noncoding elements and regulatory features (Figure 1B and Table 1). Further, VAAST[79] and GEMINI[80] allow for comprehensive analysis and integration of coding variants, noncoding variants, regulatory elements and phenotype. In cancer studies, PolyPhen[81] [A concatenation of “polymorphism phenotyping”, PolyPhen predicts the impact of amino acid changes on proteins and is often used in conjunction with SIFT.], SIFT[82] [SIFT (sorting intolerant from tolerant) infers whether amino acid substitution has an effect on subsequent functioning of protein and is often used in conjunction with PolyPhen], MutationAssessor[83] and Condel[84] are commonly used to predict deleterious mutations. In addition, CHASM[85,86] [CHASM is a popular tool for assessing functional impact of somatic missense mutations based on whether they furnish selective advantage to cancerous cells.], TransFIC[87] and OncodriveFM[88] use features learned from known cancer mutations for highlighting potential driver mutations. Finally, tools such as ActiveDriver[89] have been developed to predict effects related to protein aggregation, protein stability and alterations of residues targeted by post-translational modification.

Significantly mutated genes

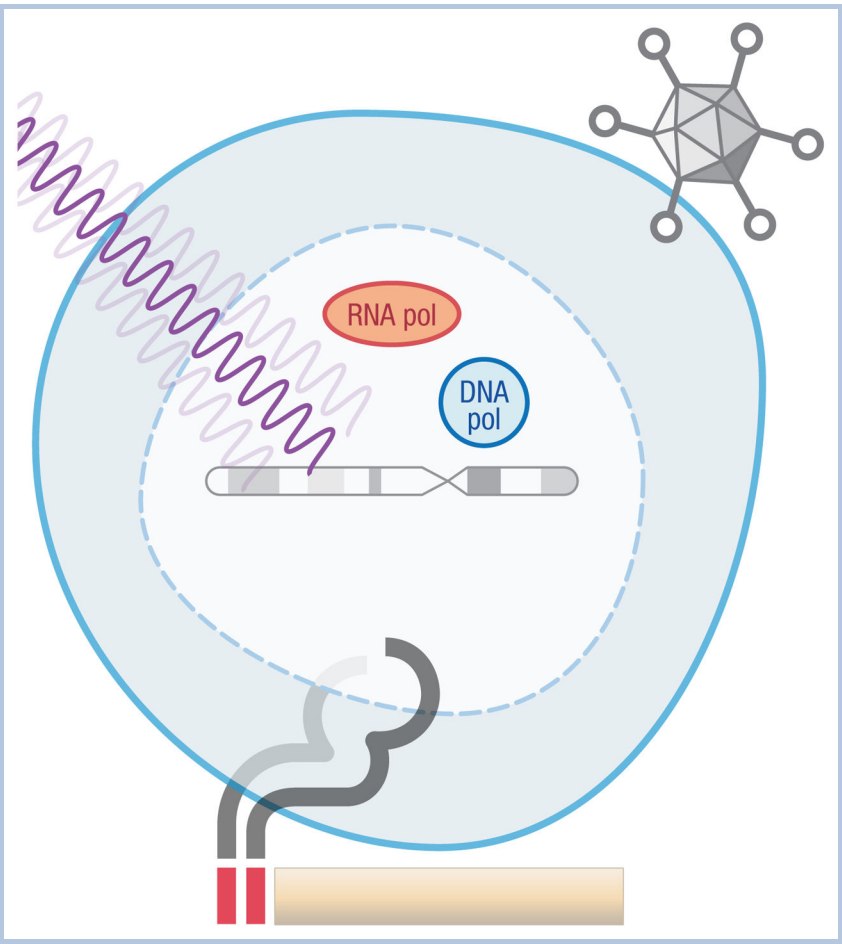
The most widely used approach to distinguish driver mutations from passenger mutations is to identify those mutations that occur more often than expected by chance. This approach is generally applicable across cancer types and is especially well suited for mutagenic phenomena associated with specific kinds of cancers, for example viral disruption in ovarian

and cervical cancers, smoking and tobacco-induced mutations in lung and oral cancers, and ultraviolet (UVA, UVB or UVC) radiation-induced mutations in melanoma (Figure 2 and Box 2). In the simplest case, one assumes that the background mutation rate **[G]** (BMR) of a gene is known and evaluates the probability of passenger mutations in a given number of samples using a statistical test[90,91].

Box 2

Detection of environmental impact on cancer genomes

Healthy cells are subjected to various external insults that promote mutagenesis, well-known examples being cigarette smoke, asbestos, and ultraviolet radiation[159] (Box 2 Figure). Viral infection and age also play roles[160]. These factors leave their marks on the cancer genome. For example, comparison of the mutation profiles across 12 common cancer types reveals that lung tumours contain higher proportions of C→A transversions[131], which are classical signatures of exposure to cigarette smoke. Mutation dynamics are compliant with circumstances[133], such as by ultraviolet exposure in melanomas, mismatch repair defects in colon cancers, or viral infections in head and neck tumors[133,161,162]. There is also growing appreciation that viral sequences, both episomal and those integrated into a genome, are more important in cancer than previously thought. Several oncoviruses have already been confirmed, including human papilloma virus (HPV), hepatitis B virus (HBV), hepatitis C virus (HCV), Epstein-Barr virus (EBV), human T-lymphotropic virus and Merkel cell polyomavirus, but there are undoubtedly more and they affect 15% to 20% of all human cancers[163]. Efforts to systematically characterize viruses in cancer are forthcoming and screening cancer genomes for viral sequences will likely be routine in the future. Despite their propensity for rapid evolution, it is likely that viral sequences will be reasonably detectable owing to their size, for example using homology-based read alignment and comparison with viral and bacterial databases. PathSeq[164] and RINS[165] investigate microbial sequences using the traditional subtraction and intersection approaches, respectively and research is now underway for developing additional tools for viral discovery.



Box 2 Figure. Environmental factor contributing to cancer risk

Smoking, viruses, and radiation can strongly affect mutation rates across the cancer genome and mutation profiles across cancer types and human populations. Signatures of these effects can often be detected in tumour genome sequences.

The primary difficulty is to obtain good BMR estimates, as inaccuracies can lead to incorrect association of a gene with cancer. Many factors are known to affect the BMR of a gene (Figure 2), including covariates, variation among samples, and errors in upstream analysis. Covariates include differences in gene length, expression level, and replication timing. Mutation frequencies can differ not only across patients within a cancer type, but also because of diverse mutation spectra across cancer types that are possibly associated with environmental factors and viral signatures. Finally, incorrect or biased annotation of mutations can contribute markedly to potential false positives in cancer gene analysis. For example, multiple open reading frames in genes like *TTN* or incomplete description of pseudogenes in olfactory receptors can lead to incorrect assignment and annotation of mutations resulting in false predictions. Inadequate sequence coverage of a gene exacerbates these problems. Software that accounts for these contingencies includes MuSiC[92] and MutSig[93], which have been extensively used in many large-scale cancer studies[94–98]. Both tools integrate heterogeneities using convolution to obtain probability tails. There are

additional covariates not accounted for and it is likely that frequency methods will continue to be developed.

Another method that has been used to distinguish between driver and passenger mutations is to examine whether mutations cluster in specific residues of the protein sequence. The so-called '20/20 Rule'[99] advises that a gene be classified as an oncogene if at least 20% of its missense mutations (or identical in-frame indels) are located at a particular residue. Conversely, a gene is classified as a tumour suppressor if at least 20% of the mutations are inactivating (nonsense, frame-shift, splice site, or stop codon read-through). This heuristic is applicable to many well-known cancer genes, but is also somewhat arbitrary in the use of a fixed 20% threshold. It is now being supplemented by algorithms that assess patterns of mutational signatures[100] and clustering of mutations in protein sequence[101] or 3D protein structure[102] using more rigorous statistical scores. Recent methods have shown that combining different signals of positive selection holds great potential for finding reliable lists of driver genes[103].

Pathway and network analysis

Enhanced understanding of somatic mutations can be gained by examining collections of mutations in signaling, regulatory, or metabolic pathways (Figure 3). It is well established that functional somatic mutations deregulate these pathways, and researchers have used a variety of approaches to assess the clustering of mutations in known pathways and interaction networks. These approaches can be divided into two classes: those that analyze known (curated) pathways, represented as gene sets, and those that analyze interaction networks to implicitly build pathways *de novo*.

A straightforward approach to evaluate combinations of mutated genes is to examine the overlap between lists of mutated genes and pre-defined gene sets having known biological function. This technique has been used for over a decade in gene expression analysis to evaluate lists of differentially-expressed genes. Databases that record functional annotations of human genes include KEGG[104], GO[105], MSigDB[106] and others. For example, suppose we have a list ***M*** of mutated genes, and we aim to see whether this list contains genes involved in regulation of cell cycle. Using the KEGG database, we find the list ***L*** of over two dozen cell cycle genes. There are two statistical tests that can be used to test whether ***M*** and ***L*** have significant overlap. First, if the list ***M*** of mutated genes is ranked (for example, using one of the mutation significance scores described above), gene-set enrichment analysis (GSEA)[106] can be used to determine whether the genes in ***L*** are near the top of the ranked list ***M***[107]. Second, if the list ***M*** is unranked, then the overlap between the lists ***M*** and ***L*** can be assessed using a hypergeometric test[108]. More recently, specialized tests for SMG sets have been introduced. The most direct approach is to adapt one of the SMG tests (e.g., MuSiC and MutSig) described above. More sophisticated approaches such as PathScan[109] and the method of Boca, *et al.*[110] allow for varying BMR across annotated genes.

Examination of gene sets overcomes some of the limitations of single-gene tests of recurrence; in particular, these tests can assign significance to rarely mutated genes, when these genes appear in the same pathway. However, these tests also have some limitations. Human gene annotations and pathway databases remain incomplete and there is extensive

crosstalk between pathways, meaning that decisions regarding which genes form the boundary of a pathway are somewhat arbitrary. The crosstalk is represented in gene set and pathway databases by the presence of multiple, overlapping gene sets, thus complicating the interpretation of reported enrichments. Finally, signaling and regulatory pathways have a rich topology of activating and inhibitory interactions, and this information is not represented in the list of genes genes/proteins that are members of the pathway.

To overcome these limitations, a second approach to analyzing combinations of mutations is to utilize biological interaction networks. A variety of genome-scale protein–protein interaction networks have been constructed in the past few years. For example, HPRD[111], KEGG[104] and Reactome[112] summarize experimentally validated protein–protein interactions, whereas other databases, such as BioGrid[113], STRING[114], HINT[115] and iRefIndex[116] integrate interaction information from multiple data sources including protein–protein interactions derived from high-throughput experiments. The resulting protein–protein interaction networks contain over 10,000 proteins and 50,000 interactions. More recently, protein–DNA interactions from the ENCODE project[117] have been integrated into these networks[118].

Interaction networks have been used in place of gene sets to determine combinations of mutations that should be further evaluated. However, most biological networks have a non-uniform topology that is characterized by the presence of hubs or nodes. This topology must be taken into account when defining mutated subnetworks. HotNet[119] is a method to find subnetworks of a large interaction network that are mutated in more samples than expected by chance. HotNet employs a heat diffusion model to simultaneously encode both the topology of the network and the significance of the observed frequencies of each mutated gene. Genes (or their corresponding proteins) are assigned an initial heat according to their mutation frequency or significance. This heat then diffuses over the edges of a network. Thus, significantly mutated subnetworks correspond to hotspots on the network. The number and size of such subnetworks is then tested for statistical significance. HotNet has been used to determine subnetworks in multiple cancer types analyzed in the context of TCGA [95,97,120], and has, for example, implicated mutations in the Notch signaling pathway in ovarian carcinoma[95].

Recently, network-based stratification (NBS)[121] used a similar heat diffusion model to define subtypes of tumour samples by clustering smoothed mutation profiles. MeMo[122] is another approach to find mutated subnetworks, using the observation that driver mutations in interacting proteins are often mutually exclusive across patients[123,124](see also below). MeMo first defines modules of highly-connected nodes in the network, and then assesses whether these network modules exhibit mutually exclusive mutations. MeMo has been used in several cancer types reported in the TCGA [96,120]. Another approach used in TCGA studies[120] is TieDIE[125], which employs a network diffusion approach to connect genetic abnormalities (e.g. somatic mutations) to transcriptional changes. Many other methods have been introduced to examine networks using gene expression[126], which are not discussed in detail here.

The third approach that has been used to analyze combinations of mutations is the identification of mutually-exclusive sets of mutations. For example, *PIK3CA* mutations and *PTEN* deletion are mutually exclusive in breast cancer[127]. Inverting this idea, one might find combinations of driver mutations by identifying mutually exclusive sets of mutations. MeMo[122] uses this idea to examine genes with known interactions, as noted above. Alternatively, one may attempt to discover sets of mutually exclusive genes *de novo*, with no prior restrictions on the sets of genes. This idea is the basis of the De Novo Driver Exclusivity (Dendrix) algorithm[128], as well as the Multi-Dendrix[129] and RME[130] algorithms. The Dendrix algorithm was used in the TCGA acute myeloid leukemia project[97] and in Pan-Cancer TCGA analysis of 12 cancer types[131].

Today, a substantial number of significant genes and pathways have been identified in individual cancer types as well as across cancer types. The next challenge is to better understand how these genes and pathways interact and function in concert in individual cancer patient.

Genome integrity and clonal architectures

Accumulation of somatic mutations in a population of tumor cells is the foundation of the clonal theory of cancer, as described by Peter Nowell in 1976[10]. High-throughput sequencing has led to new insights into this process, including the discovery of novel mutational processes and the quantification of the clonal architecture of tumors.

Kataegis, chromothripsis and chromoplexy

One of the more fascinating observations from cancer-genome sequencing studies are genomes with extreme numbers and types on mutations. Kataegis **[G]** is the occurrence of an unusually large number of single nucleotide mutations clustered in a single locus, and was first reported in breast tumors[132] and other cancer types[133]. Kataegisis identified from “rainfall plots” that illustrate the frequency of single nucleotide mutations across the genome.

The analogous phenomenon of many genome rearrangement breakpoints clustered at a single locus has long been observed from lower resolution microarray and cytogenetic studies[134]. However, genome sequencing has revealed a different phenomenon of chromosome shattering, or chromothripsis **[G],** where one or more loci undergo a catastrophic event of simultaneous breakage and aberrant repair at multiple breakpoints in a single cell division[70]. Chromothripsis was originally reported in ~2–3% of all cancers, but was shown to be particularly common in bone cancers (~25%). It was later reported in pediatric medulloblastomas[135], and associated with *TP53* mutations, suggesting a possible mechanism for its appearance[136,137]. A related process called chromoplexy **[G]** has now been observed in prostate cancers[138].

Distinguishing chromothripsis/chromplexy from sequential accumulation of chromosomal rearrangements over multiple cellular generations is a challenge. Secondary rearrangements often obscure the signatures of chromothripsis and chromoplexy. The distinction between simultaneous and sequential rearrangement is typically made via simulations[70,135,139],

although there have been criticisms of these approaches[140]. In lieu of simulation, putative signatures of chromothripsis have been proposed[141]. Tools, such as PREGO[142], nFuse[72] and extensions of Hydra[139] that simultaneously analyze multiple rearrangement breakpoints facilitate the evaluation of these signatures. However, more work is needed to find quantitative measures that distinguish chromothripsis and chromoplexy from sequential accumulation of rearrangements.

Defining clonal architecture in heterogeneous tumours

All genomic alterations discussed above have a role in clonal evolution **[G]**. Tumour clones are subject to changing selective pressures and continually accumulating mutations (Figure 4A). Genomic alterations collectively reflect this evolutionary history and can be used to reconstruct the subclonal architectures and progression processes that might have led to relapse or metastasis. Such information is enormously important, as clonality has already been implicated in numerous aspects of cancer, including clinical outcome[143], increased progression and malignancy[144] and drug resistance[145].

Clonal inference can be challenging. The number and positioning of clones within a tumour is often unknown, so uniform sampling is routinely presumed. Dot-plots are often used to obtain visual estimates of clones. For example; each heterozygous SNV event can be represented by a dot positioned on orthogonal axes of VAF versus frequency or total reads representing the event. Because the process is stochastic, such plots cluster into “dot clouds” (Figure 4) that are suggestive of clones. If the collective distribution is non-Gaussian (as determined by tests like Shapiro-Wilk or D’Agostino K-squared), multiple clones are presumed to be present. The process of discerning clones individually then encounters more confounders from both experimental contingencies (such as mutual impurities of the tumor and normal samples owing to suffusion or insufficient margin) and biological complications (such as copy number variations within the tumor genome). There are also subtle statistical factors, including differences in variances of clonal VAF distributions. Specifically, mutations that exist in all tumour cells, namely those present in the most recent common ancestor, have a variance (σ^2) proportional to unity ($\sigma^2 \propto 1$). Conversely, mutations present in a minor subclone whose mass is a fraction μ of the total tumor have $\sigma^2 \propto \mu$, meaning its distribution is “flattened” in this dimension.

Some clonal discovery methods center around the mathematical concept of density estimation, a process through which a probability density function (PDF) that best describes the observed data is constructed, for example using the Parzen-Rosenblatt[146,147] “window” method. If clusters are sufficiently separated by VAF, the PDF readily identifies the tumour clones. There are a variety of more recent and sophisticated methods. For example ABSOLUTE[148] adds an optimally-fitted copy number alteration model and karyotype likelihood model. Conversely, PyClone[149] and the methods of Nik-Zainal *et al*. identify clones using hierarchical Bayesian clustering[11,150]. We have developed a method called SciClone (Miller *et al*. unpublished observations) (Figure 4B) that uses Bayesian mixture modeling to examine multiple samples from a patient over time (initial and relapsing tumour samples) or space (multiple biopsies). The THetA algorithm[151] accounts for the presence of copy number aberrations, which can confound analysis of VAFs. Like the variant calling

problem, progress has been significant, but substantial improvement is still needed. Not only will better variant detection improve clonal analysis, but also additional classes of information, including cancer-specific and pan-cancer population data, as well as information from other affected family members, will help to better define tumour architecture. Finally, direct integration with phylogenetic analysis algorithms may help to arbitrate among certain kinds of multiple alternatives that are currently undecidable[148].

Conclusion: basic and clinical applications

In the short time since cancer genomics burst onto the biomedical scene it has made numerous fundamental contributions: 1) cancer-associated genes and pathways have been identified; 2) germline predispositions have been established; 3) technologies and algorithms have been improved; 4) vast datasets have been organized and recorded; and 5) knowledge has been classified into new databases. These accomplishments can be attributed to many individual research lab driven projects as well as large scale collaborative projects conducted by The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) using cutting edge computational approaches[152,153]. TCGA has completed nearly 10,000 cancer cases across 20 cancer types and ICGC will be sequencing approximately 25,000 additional genomes across 50 cancer types over the next several years. Further, efforts by the Cancer Cell Line Encyclopedia[154] and Genomics of Drugs sensitivity in Cancer (http://www.cancerrxgene.org/) will help establish genomic determinants of resistance or sensitivity to drugs. The information and knowledge that will pour out of such projects are expected to have enormous implications for understanding cancer broadly, as well as for diagnosing and treating tumours at the individual patient level. This will be a tangible step towards personalized medicine.

Widespread clinical application of cancer genome and transcriptome sequencing is a certainty, although the timing remains unclear because of several outstanding issues related to both cost and reliability. First, the “data spectrum” and associated analysis tools are not yet complete. A significant portion of driver events in cancer are DNA or RNA alterations that affect protein expression, but proteomics has not yet ramped to the same high-throughput rates and sample census that genomic sequencing has. In our view, proteomic data are increasingly important in ascertaining driver genes and pathways, especially in terms of winnowing false positives from the large lists of hypotheses generated by pathway, network and significant gene mutation algorithms. However, it is clear that the proteomic gap is starting to close. For example, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) launched by the National Cancer Institute (NCI) will further many goals, including characterizing tumour protein inventories, integrating genomics with proteomics and developing biomarker assays for high-priority proteins. Associated bioinformatics tools will be further developed, as well. This will be an increasingly fertile area of research. The second factor is the reality of cost. The sequence of an individual genome has dropped about 5 orders of magnitude, from about $1B for the first human genome to around $10K today. Technology development continues apace, but the overall cost for an entire “package” (DNA, RNA, and proteomic sequencing and companion systematic analysis) will likely have to drop yet another order of magnitude before sequencing can become anything like a

routine clinical test. There will probably be some form of certification process for analysis software, as well.

There have even been a few early clinical victories, like the amazing case of Dr. Lukas Wartman, where comprehensive genome, exome and RNA analysis implicated *FLT3* over-expression in his particular form of leukaemia (In Treatment for Leukemia, Glimpses of the Future, The New York Times, July 7th, 2012). This analysis led to the decision to administer Sutent, an FDA-approved tyrosine kinase inhibitor targeting *FLT3* expression that quickly put Dr. Wartman's disease into remission, which continues today. The next chapter of cancer research will undoubtedly see further pushes toward clinical application, as well as increased involvement of big pharmain developing new therapeutic agents. The cancer landscape will look vastly different from today in a decade and we will be at the threshold, if not well into the era of finding cures (or means of conferring long-term remission) for some cancers. Stay tuned.

Acknowledgments

This work was supported by the National Human Genome Research Institute grants U01HG006517 to L.D. and R01HG005690 and R01HG007069 to B.J.R. and the National Cancer Institute grant R01CA180006 to L.D. We would like to thank Kai Ye and Michael D. McLellan for helpful comments.

Glossary

Background mutation rate	Rate at which spontaneous mutations occur due to uncorrected copying errors
Chromoplexy	A mutational event that results in significant, complex rearrangements involving multiple loci, though not as dramatic as chromothripsis and involving less clustering of rearrangement breakpoints
Chromothripsis	A catastrophic mutational event that "shatters" one or more chromosomes, with simultaneous loss and rearrangement of multiple chromosomal segments
Clonal evolution	the emergence of novel clones having improved survival or propagational fitness according to the particular sets of somatic mutations they have accumulated
Driver mutation	A somatic mutation that plays a causal role in initiation, progression, metastasis, or recurrence of cancer
De novo assembly	Reconstruction of a genomic target by assessing consensus sequence from alignments of overlapping reads and clones
Gapped alignment	Alignment process where small gaps are allowed if they support a better fit
Kataegis	The appearance of regions of local hyper-mutation in a tumor genome

Paired-end mapping	Coordinated mapping of both sequenced ends of a fragment to a reference genome, where their approximately known separation furnishes extra information against misalignments
Passenger mutations	Somatic mutations that arise incidentally and play no mechanistic role in cancer initiation or progression
Precision	The fraction of the total number of called events that are true, sometimes called the positive predictive value
Pyrosequencing	Specific sequencing-by-synthesis method where detection is based on chemiluminescent signals from luciferin conversion
Sequence coverage theory	Characterization of sequencing processes mathematically in order to support development of detection methods and analysis and design of sequencing projects
Sequencing-by-ligation	Sequencing based on using the mismatch sensitivity of DNA ligase to detect nucleotides
Sequencing-by-synthesis	Sequential polymerization of nucleotides to a template with each incorporation inferred by an imaging process, usually from a fluorescent dye attached to the added nucleotide
Significantly mutated gene	A gene having a rate of somatic mutation that is higher than what can be attributed to a random background rate, suggesting a role in tumor initiation or progression
Split read	The phenomenon in which a read spans a deleted site, whereby it appears to be split in its alignment to a reference
Type I error	Declaring an effect where none actually exists, which leads to a “false positive”
Type II error	Overlooking an actual effect, which leads to a “false negative”

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977; 74:5463–5467. [PubMed: 271968]
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. 1977. Biotechnology. 1992; 24:104–108. [PubMed: 1422003]
3. Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921.10.1038/35057062 [PubMed: 11237011]
4. Ley TJ, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008; 456:66–72.10.1038/nature07485 [PubMed: 18987736]
5. Shendure J, Lieberman Aiden E. The expanding scope of DNA sequencing. Nat Biotechnol. 2012; 30:1084–1094.10.1038/nbt.2421 [PubMed: 23138308]
6. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011; 48:580–589.10.1136/jmedgenet-2011-100223 [PubMed: 21730106]
7. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011; 12:87–98.10.1038/nrg2934 [PubMed: 21191423]

8. Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. Nat Methods. 2012; 9:145–151.10.1038/nmeth.1828 [PubMed: 22290186]
9. Ding L, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature. 2010; 464:999–1005.10.1038/nature08989 [PubMed: 20393555]
10. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976; 194:23–28. [PubMed: 959840]
11. Ding L, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012; 481:506–510.10.1038/nature10738 [PubMed: 22237025]
12. Gerlinger M, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012; 366:883–892.10.1056/NEJMoa1113205 [PubMed: 22397650]
13. Navin N, et al. Inferring tumor progression from genomic heterogeneity. Genome Res. 2010; 20:68–80.10.1101/gr.099622.109 [PubMed: 19903760]
14. Navin NE, Hicks J. Tracing the tumor lineage. Mol Oncol. 2010; 4:267–283.10.1016/j.molonc.2010.04.010 [PubMed: 20537601]
15. Navin N, et al. Tumour evolution inferred by single-cell sequencing. Nature. 201110.1038/nature09807
16. Hou Y, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012; 148:873–885.10.1016/j.cell.2012.02.028 [PubMed: 22385957]
17. Xu X, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012; 148:886–895.10.1016/j.cell.2012.02.025 [PubMed: 22385958]
18. Gundry M, Li W, Maqbool SB, Vijg J. Direct, genome-wide assessment of DNA mutations in single cells. Nucleic Acids Res. 2012; 40:2032–2040.10.1093/nar/gkr949 [PubMed: 22086961]
19. Baslan T, et al. Genome-wide copy number analysis of single cells. Nat Protoc. 2012; 7:1024–1041.10.1038/nprot.2012.039 [PubMed: 22555242]
20. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19:455–477.10.1089/cmb.2012.0021 [PubMed: 22506599]
21. Kim SY, Speed TP. Comparing somatic mutation-callers: beyond Venn diagrams. BMC Bioinformatics. 2013; 14:189.10.1186/1471-2105-14-189 [PubMed: 23758877]
22. Goode DL, et al. A simple consensus approach improves somatic mutation prediction accuracy. Genome Med. 2013; 5:90.10.1186/gm494 [PubMed: 24073752]
23. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303.10.1101/gr.107524.110 [PubMed: 20644199]
24. Koboldt DC, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009; 25:2283–2285.10.1093/bioinformatics/btp373 [PubMed: 19542151]
25. Koboldt DC, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22:568–576.10.1101/gr.129684.111 [PubMed: 22300766]
26. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079.10.1093/bioinformatics/btp352 [PubMed: 19505943]
27. Larson DE, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012; 28:311–317.10.1093/bioinformatics/btr665 [PubMed: 22155872]
28. Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013; 31:213–219.10.1038/nbt.2514 [PubMed: 23396013]
29. Saunders CT, et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012; 28:1811–1817.10.1093/bioinformatics/bts271 [PubMed: 22581179]
30. Goya R, et al. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics. 2010; 26:730–736.10.1093/bioinformatics/btq040 [PubMed: 20130035]

31. Roth A, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics. 2012; 28:907–913.10.1093/bioinformatics/bts053 [PubMed: 22285562]
32. Lunter G. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. Bioinformatics. 2007; 23:i289–296.10.1093/bioinformatics/btm185 [PubMed: 17646308]
33. Cartwright RA. Problems and solutions for estimating indel rates and length distributions. Molecular biology and evolution. 2009; 26:473–480.10.1093/molbev/msn275 [PubMed: 19042944]
34. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18:1851–1858. gr.078212.108 [pii]. 10.1101/gr.078212.108 [PubMed: 18714091]
35. Smith CC, et al. Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia. Nature. 2012; 485:260–263.10.1038/nature11016 [PubMed: 22504184]
36. Spencer DH, et al. Detection of FLT3 internal tandem duplication in targeted, short-read-length, next-generation sequencing data. J Mol Diagn. 2013; 15:81–93.10.1016/j.jmoldx.2012.08.001 [PubMed: 23159595]
37. Albers CA, et al. Dindel: accurate indel calls from short-read data. Genome Res. 2011; 21:961–973.10.1101/gr.112326.110 [PubMed: 20980555]
38. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25:2865–2871. btp394 [pii]. 10.1093/bioinformatics/btp394 [PubMed: 19561018]
39. Ye K, Kosters WA, Ijzerman AP. An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences. Bioinformatics. 2007; 23:687–693. btl665 [pii]. 10.1093/bioinformatics/btl665 [PubMed: 17237070]
40. Rausch T, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012; 28:i333–i339.10.1093/bioinformatics/bts378 [PubMed: 22962449]
41. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. arXiv:1303.3997 [q-bio.GN]
42. Chen K, et al. TIGRA: A Targeted Iterative Graph Routing Assembler for breakpoint assembly. Genome Res. 201310.1101/gr.162883.113
43. Bignell GR, et al. Signatures of mutation and selection in the cancer genome. Nature. 2010; 463:893–898.10.1038/nature08768 [PubMed: 20164919]
44. Beroukhim R, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463:899–905.10.1038/nature08822 [PubMed: 20164920]
45. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009; 19:1586–1592. gr.092981.109 [pii]. 10.1101/gr.092981.109 [PubMed: 19657104]
46. Campbell PJ, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008; 40:722–729. ng.128 [pii]. 10.1038/ng.128 [PubMed: 18438408]
47. Beroukhim R, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A. 2007; 104:20007–20012. 0710052104 [pii]. 10.1073/pnas.0710052104 [PubMed: 18077431]
48. Zhang Q, et al. CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. Bioinformatics. 2010; 26:464–469.10.1093/bioinformatics/btp708 [PubMed: 20031968]
49. Raphael BJ, Volik S, Collins C, Pevzner PA. Reconstructing tumor genome architectures. Bioinformatics. 2003; 19(Suppl 2):ii162–171. [PubMed: 14534186]
50. Raphael BJ, et al. A sequence-based survey of the complex structural organization of tumor genomes. Genome Biol. 2008; 9:R59. gb-2008-9-3-r59 [pii]. 10.1186/gb-2008-9-3-r59 [PubMed: 18364049]
51. Volik S, et al. Decoding the fine-scale structure of a breast cancer genome and transcriptome. Genome Res. 2006; 16:394–404. gr.4247306 [pii]. 10.1101/gr.4247306 [PubMed: 16461635]

52. Volik S, et al. End-sequence profiling: sequence-based analysis of aberrant genomes. Proc Natl Acad Sci U S A. 2003; 100:7696–7701.10.1073/pnas.1232418100 [PubMed: 12788976]
53. Bignell GR, et al. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Res. 2007; 17:1296–1303.10.1101/gr.6522707 [PubMed: 17675364]
54. Chen K, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009; 6:677–681. nmeth.1363 [pii]. 10.1038/nmeth.1363 [PubMed: 19668202]
55. Wang J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods. 2011; 8:652–654.10.1038/nmeth.1628 [PubMed: 21666668]
56. Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res. 2009; 19:1270–1278.10.1101/gr.088633.108 [PubMed: 19447966]
57. Sindi S, Helman E, Bashir A, Raphael BJ. A geometric approach for classification and comparison of structural variants. Bioinformatics. 2009; 25:i222–230. btp208 [pii]. 10.1093/bioinformatics/btp208 [PubMed: 19477992]
58. 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:R22.10.1186/gb-2012-13-3-r22 [PubMed: 22452995]
59. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet. 2011; 43:269–276.10.1038/ng.768 [PubMed: 21317889]
60. Rowley JD. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature. 1973; 243:290–293. [PubMed: 4126434]
61. Huang ME, et al. Use of all-trans retinoic acid in the treatment of acute promyelocytic leukemia. Blood. 1988; 72:567–572. [PubMed: 3165295]
62. Huang ME. Treatment of acute promyelocytic leukemia with all-trans retinoic acid. Zhonghua yi xue za zhi. 1988; 68:131–133. 110. [PubMed: 3136889]
63. Tomlins SA, et al. Integrative molecular concept modeling of prostate cancer progression. Nat Genet. 2007; 39:41–51.10.1038/ng1935 [PubMed: 17173048]
64. Kim YK, et al. Cooperation of H2O2-mediated ERK activation with Smad pathway in TGF-beta1 induction of p21WAF1/Cip1. Cellular signalling. 2006; 18:236–243.10.1016/j.cellsig.2005.04.008 [PubMed: 15979845]
65. McPherson A, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol. 2011; 7:e1001138.10.1371/journal.pcbi.1001138 [PubMed: 21625565]
66. Wang K, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010; 38:e178.10.1093/nar/gkq622 [PubMed: 20802226]
67. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011; 27:2903–2904.10.1093/bioinformatics/btr467 [PubMed: 21840877]
68. Chen K, et al. BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. Bioinformatics. 2012; 28:1923–1924.10.1093/bioinformatics/bts272 [PubMed: 22563071]
69. Berger MF, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470:214–220.10.1038/nature09744 [PubMed: 21307934]
70. Stephens PJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011; 144:27–40.10.1016/j.cell.2010.11.055 [PubMed: 21215367]
71. McPherson A, et al. Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. Bioinformatics. 2011; 27:1481–1488.10.1093/bioinformatics/btr184 [PubMed: 21478487]
72. McPherson A, et al. nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. Genome Res. 2012; 22:2250–2261.10.1101/gr.136572.111 [PubMed: 22745232]

73. Chen K, et al. BreakTrans: uncovering the genomic architecture of gene fusions. Genome Biol. 2013; 14:R87.10.1186/gb-2013-14-8-r87 [PubMed: 23972288]
74. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38:e164.10.1093/nar/gkq603 [PubMed: 20601685]
75. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012; 6:80–92.10.4161/fly.19695 [PubMed: 22728672]
76. Woolfe A, Mullikin JC, Elnitski L. Genomic features defining exonic variants that modulate splicing. Genome Biol. 2010; 11:R20.10.1186/gb-2010-11-2-r20 [PubMed: 20158892]
77. Khurana E, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013; 342:1235587.10.1126/science.1235587 [PubMed: 24092746]
78. Chelala C, Khan A, Lemoine NR. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics. 2009; 25:655–661.10.1093/bioinformatics/btn653 [PubMed: 19098027]
79. Yandell M, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. 2011; 21:1529–1542.10.1101/gr.123158.111 [PubMed: 21700766]
80. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol. 2013; 9:e1003153.10.1371/journal.pcbi.1003153 [PubMed: 23874191]
81. Nakken S, Alseth I, Rognes T. Computational prediction of the effects of non-synonymous single nucleotide polymorphisms in human DNA repair genes. Neuroscience. 2007; 145:1273–1279. [PubMed: 17055652]
82. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003; 31:3812–3814. [PubMed: 12824425]
83. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011; 39:e118.10.1093/nar/gkr407 [PubMed: 21727090]
84. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011; 88:440–449.10.1016/j.ajhg.2011.03.004 [PubMed: 21457909]
85. Wong WC, et al. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011; 27:2147–2148.10.1093/bioinformatics/btr357 [PubMed: 21685053]
86. Carter H, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009; 69:6660–6667.10.1158/0008-5472.CAN-09-1133 [PubMed: 19654296]
87. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med. 2012; 4:89.10.1186/gm390 [PubMed: 23181723]
88. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012; 40:e169.10.1093/nar/gks743 [PubMed: 22904074]
89. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Molecular systems biology. 2013; 9:637.10.1038/msb.2012.68 [PubMed: 23340843]
90. Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF. Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics. 2006; 173:2187–2198.10.1534/genetics.105.044677 [PubMed: 16783027]
91. Getz G, et al. Comment on “The consensus coding sequences of human breast and colorectal cancers”. Science. 2007; 317:1500.10.1126/science.1138764 [PubMed: 17872428]
92. Dees ND, et al. MuSiC: Identifying mutational significance in cancer genomes. Genome Res. 2012; 22:1589–1598.10.1101/gr.134635.111 [PubMed: 22759861]
93. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499:214–218.10.1038/nature12213 [PubMed: 23770567]

94. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. nature07385 [pii]. 10.1038/nature07385 [PubMed: 18772890]
95. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615.10.1038/nature10166 [PubMed: 21720365]
96. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70.10.1038/nature11412 [PubMed: 23000897]
97. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N Engl J Med. 201310.1056/NEJMoa1301689
98. Ding L, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008; 455:1069–1075. nature07423 [pii]. 10.1038/nature07423 [PubMed: 18948947]
99. Vogelstein B, et al. Cancer genome landscapes. Science. 2013; 339:1546–1558.10.1126/science.1235122 [PubMed: 23539594]
100. Davoli T, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013; 155:948–962.10.1016/j.cell.2013.10.011 [PubMed: 24183448]
101. 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; 11:11.10.1186/1471-2105-11-11 [PubMed: 20053295]
102. Ryslik GA, Cheng Y, Cheung KH, Modis Y, Zhao H. Utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics. 2013; 14:190.10.1186/1471-2105-14-190 [PubMed: 23758891]
103. Lawrence MS, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014; 505:495–501.10.1038/nature12912 [PubMed: 24390350]
104. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010; 38:D355–360.10.1093/nar/gkp896 [PubMed: 19880382]
105. Ashburner M, et al. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet. 2000; 25:25–29.10.1038/75556 [PubMed: 10802651]
106. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102:15545–15550.10.1073/pnas.0506580102 [PubMed: 16199517]
107. Lin J, et al. A multidimensional analysis of genes mutated in breast and colorectal cancers. Genome Res. 2007; 17:1304–1318.10.1101/gr.6431107 [PubMed: 17693572]
108. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37:1–13.10.1093/nar/gkn923 [PubMed: 19033363]
109. Wendl MC, et al. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics. 2011; 27:1595–1602.10.1093/bioinformatics/btr193 [PubMed: 21498403]
110. Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene set analysis for cancer mutation data. Genome Biol. 2010; 11:R112.10.1186/gb-2010-11-11-r112 [PubMed: 21092299]
111. Peri S, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003; 13:2363–2371.10.1101/gr.1680803 [PubMed: 14525934]
112. Croft D, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011; 39:D691–697.10.1093/nar/gkq1018 [PubMed: 21067998]
113. Chatr-Aryamontri A, et al. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 2013; 41:D816–823.10.1093/nar/gks1158 [PubMed: 23203989]
114. Franceschini A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013; 41:D808–815.10.1093/nar/gks1094 [PubMed: 23203871]

115. Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC systems biology. 2012; 6:92.10.1186/1752-0509-6-92 [PubMed: 22846459]
116. Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics. 2008; 9:405.10.1186/1471-2105-9-405 [PubMed: 18823568]
117. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74.10.1038/nature11247 [PubMed: 22955616]
118. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol. 2013; 9:e1002886.10.1371/journal.pcbi.1002886 [PubMed: 23505346]
119. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011; 18:507–522.10.1089/cmb.2010.0265 [PubMed: 21385051]
120. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013; 499:43–49.10.1038/nature12222 [PubMed: 23792563]
121. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013; 10:1108–1115.10.1038/nmeth.2651 [PubMed: 24037242]
122. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012; 22:398–406.10.1101/gr.125567.111 [PubMed: 21908773]
123. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004; 10:789–799.10.1038/nm1087 [PubMed: 15286780]
124. Yeang CH, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in cancer. Faseb J. 2008; 22:2605–2622.10.1096/fj.08-108985 [PubMed: 18434431]
125. Paull EO, et al. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics. 2013; 29:2757–2764.10.1093/bioinformatics/btt471 [PubMed: 23986566]
126. Vaske CJ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010; 26:i237–245.10.1093/bioinformatics/btq182 [PubMed: 20529912]
127. Saal LH, et al. PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. Cancer Res. 2005; 65:2554–2559.10.1158/0008-5472-CAN-04-3913 [PubMed: 15805248]
128. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res. 2012; 22:375–385.10.1101/gr.120477.111 [PubMed: 21653252]
129. Leiserson MD, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol. 2013; 9:e1003054.10.1371/journal.pcbi.1003054 [PubMed: 23717195]
130. 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; 4:34.10.1186/1755-8794-4-34 [PubMed: 21489305]
131. Kandoth C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013; 502:333–339.10.1038/nature12634 [PubMed: 24132290]
132. Nik-Zainal S, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012; 149:979–993.10.1016/j.cell.2012.04.024 [PubMed: 22608084]
133. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–421.10.1038/nature12477 [PubMed: 23945592]
134. Albertson DG, Collins C, McCormick F, Gray JW. Chromosome aberrations in solid tumors. Nat Genet. 2003; 34:369–376.10.1038/ng1215 [PubMed: 12923544]
135. Rausch T, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell. 2012; 148:59–71.10.1016/j.cell.2011.12.013 [PubMed: 22265402]
136. Maher CA, Wilson RK. Chromothripsis and human disease: piecing together the shattering process. Cell. 2012; 148:29–32.10.1016/j.cell.2012.01.006 [PubMed: 22265399]

137. Forment JV, Kaidi A, Jackson SP. Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer. 2012; 12:663–670.10.1038/nrc3352 [PubMed: 22972457]
138. Baca SC, et al. Punctuated evolution of prostate cancer genomes. Cell. 2013; 153:666–677.10.1016/j.cell.2013.03.021 [PubMed: 23622249]
139. Malhotra A, et al. Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. Genome Res. 2013; 23:762–776.10.1101/gr.143677.112 [PubMed: 23410887]
140. Sorzano CO, Pascual-Montano A, Sanchez de Diego A, Martinez AC, van Wely KH. Chromothripsis: breakage-fusion-bridge over and over again. Cell Cycle. 2013; 12:2016–2023.10.4161/cc.25266 [PubMed: 23759584]
141. Korbel JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. Cell. 2013; 152:1226–1236.10.1016/j.cell.2013.02.023 [PubMed: 23498933]
142. Oesper L, Ritz A, Aerni SJ, Drebin R, Raphael BJ. Reconstructing cancer genomes from paired-end sequencing data. BMC Bioinformatics. 2012; 13(Suppl 6):S10.10.1186/1471-2105-13-S6-S10 [PubMed: 22537039]
143. Landau DA, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013; 152:714–726.10.1016/j.cell.2013.01.019 [PubMed: 23415222]
144. Keats JJ, et al. Clonal competition with alternating dominance in multiple myeloma. Blood. 2012; 120:1067–1076.10.1182/blood-2012-01-405985 [PubMed: 22498740]
145. Turke AB, et al. Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC. Cancer Cell. 2010; 17:77–88.10.1016/j.ccr.2009.11.022 [PubMed: 20129249]
146. Parzen E. On estimation of a probability density function and mode. Annals of Mathematical Statistics. 1962; 33:1065–1076.
147. Rosenblatt M. Remarks on some non-parametric estimates of a density function. Annals of Mathematical Statistics. 1956; 27:832–837.
148. Carter SL, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012; 30:413–421.10.1038/nbt.2203 [PubMed: 22544022]
149. Shah SP, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012; 486:395–399.10.1038/nature10933 [PubMed: 22495314]
150. Nik-Zainal S, et al. The life history of 21 breast cancers. Cell. 2012; 149:994–1007.10.1016/j.cell.2012.04.023 [PubMed: 22608083]
151. Oesper L, Mahmoody A, Raphael BJ. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. 2013; 14:R80.10.1186/gb-2013-14-7-r80 [PubMed: 23895164]
152. Gonzalez-Perez A, et al. Computational approaches to identify functional genetic variants in cancer genomes. Nat Methods. 2013; 10:723–729.10.1038/nmeth.2562 [PubMed: 23900255]
153. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med. 2014; 6:5.10.1186/gm524 [PubMed: 24479672]
154. Barretina J, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483:603–607.10.1038/nature11003 [PubMed: 22460905]
155. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 1988; 2:231–239. [PubMed: 3294162]
156. Wendl MC, Wilson RK. Aspects of coverage in medical DNA sequencing. BMC Bioinformatics. 2008; 9:239.10.1186/1471-2105-9-239 [PubMed: 18485222]
157. Bashir A, Volik S, Collins C, Bafna V, Raphael BJ. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. PLoS Comput Biol. 2008; 4:e1000051.10.1371/journal.pcbi.1000051 [PubMed: 18404202]
158. Wendl MC, Wilson RK. Statistical aspects of discerning indel-type structural variation via DNA sequence alignment. BMC Genomics. 2009; 10:359.10.1186/1471-2164-10-359 [PubMed: 19656394]

159. Boffetta P, Nyberg F. Contribution of environmental factors to cancer risk. British medical bulletin. 2003; 68:71–94. [PubMed: 14757710]

160. Cerwenka A, Lanier LL. Natural killer cells, viruses and cancer. Nature reviews. Immunology. 2001; 1:41–49.10.1038/35095564

161. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337.10.1038/nature11252 [PubMed: 22810696]

162. Stransky N, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011; 333:1157–1160.10.1126/science.1208130 [PubMed: 21798893]

163. Parkin DM. The global health burden of infection-associated cancers in the year 2002. Int J Cancer. 2006; 118:3030–3044.10.1002/ijc.21731 [PubMed: 16404738]

164. Kostic AD, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol. 2011; 29:393–396.10.1038/nbt.1868 [PubMed: 21552235]

165. Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA. Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics. 2012; 28:1174–1175.10.1093/bioinformatics/bts100 [PubMed: 22377895]

166. Van Loo P, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010; 107:16910–16915.10.1073/pnas.1009843107 [PubMed: 20837533]

167. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249.10.1038/nmeth0410-248 [PubMed: 20354512]

168. McLaren W, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010; 26:2069–2070.10.1093/bioinformatics/btq330 [PubMed: 20562413]

169. Tamborero D, Lopez-Bigas N, Gonzalez-Perez A. Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. PLoS One. 2013; 8:e55489.10.1371/journal.pone.0055489 [PubMed: 23408991]

170. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013; 29:2238–2244.10.1093/bioinformatics/btt395 [PubMed: 23884480]

Biographies

Li Ding has concentrated her research on understanding somatic/germline genetic changes relevant to cancer initiation and progression as well as drug response. Her recent efforts include the discovery of 127 cancer genes across over 3,000 tumors from 12 major cancer types. She is the principle investigator for the NHGRI sponsored genome sequencing informatics (GS-IT) center at Washington University, an Assistant Director at the Genome Institute and an Assistant Professor of Medicine and Genetics at Washington University in St. Louis.

Michael Wendl focuses on applying mathematics and computational methods to pressing problems in the biomedical sciences. He developed much of DNA sequencing theory and co-wrote the PHRED trace analyzer used for processing Sanger sequencing data, including in the Human Genome Project. He now concentrates on problems in cancer genomics, including somatic detection, pathway analysis, and clonal evolution modeling.

Joshua McMichael creates user interfaces and data visualizations for bioinformatics, specializing in cancer genomics. He worked on the Genome Modeling System for high throughput sequencing data analysis and has produced many of the visualizations for cancer genomics discoveries including clonal evolution in acute myeloid leukemia. He currently

works as a software developer at the Genome Institute at Washington University in St. Louis.

Ben Raphael develops novel combinatorial and statistical algorithms for the interpretation of genomes. Recent work focuses on structural variation in human and cancer genomes and on network/pathway analysis of somatic mutations in cancer. He is an Associate Professor in the Department of Computer Science and Director of the Center for Computational Molecular Biology at Brown University.

Online summary

- High-throughput sequencing of cancer genomes, exomes, and transcriptomes has enabled the identification of many novelsomatic aberrations, providing new insights into cancer biology and new therapeutic targets.
- Computational and statistical tools are necessary to interpret the large and complex datasets that result from high-throughput sequencing approaches.
- Mature software for detecting single-nucleotide variants, indels, copy number aberrations, structural aberrations, and gene fusions in cancer genomes are now available. Additional challenges remain in increasing the sensitivity and specificity of these algorithms.
- Computational techniques are essential to prioritize somatic aberrations that are likely to be functional for further experimental validation. Two common approaches are to predict functional impact of individual mutations using prior biological knowledge, and to identify recurrently mutated genes, pathways, and networks across many samples.
- Algorithms to infer the clonal structure and evolutionary history of a tumor from ultra-deep sequencing data have recently been introduced. Applications of these techniques have shown that minority mutations in primary tumors may rise to majority in relapse/metastasis.
- Sequencing of cancer genomes has shown wide range of specialized mutational processes including features like kategis, chromothripsis, and chromoplexy that result in rapid genomic change and punctuated tumor evolution.

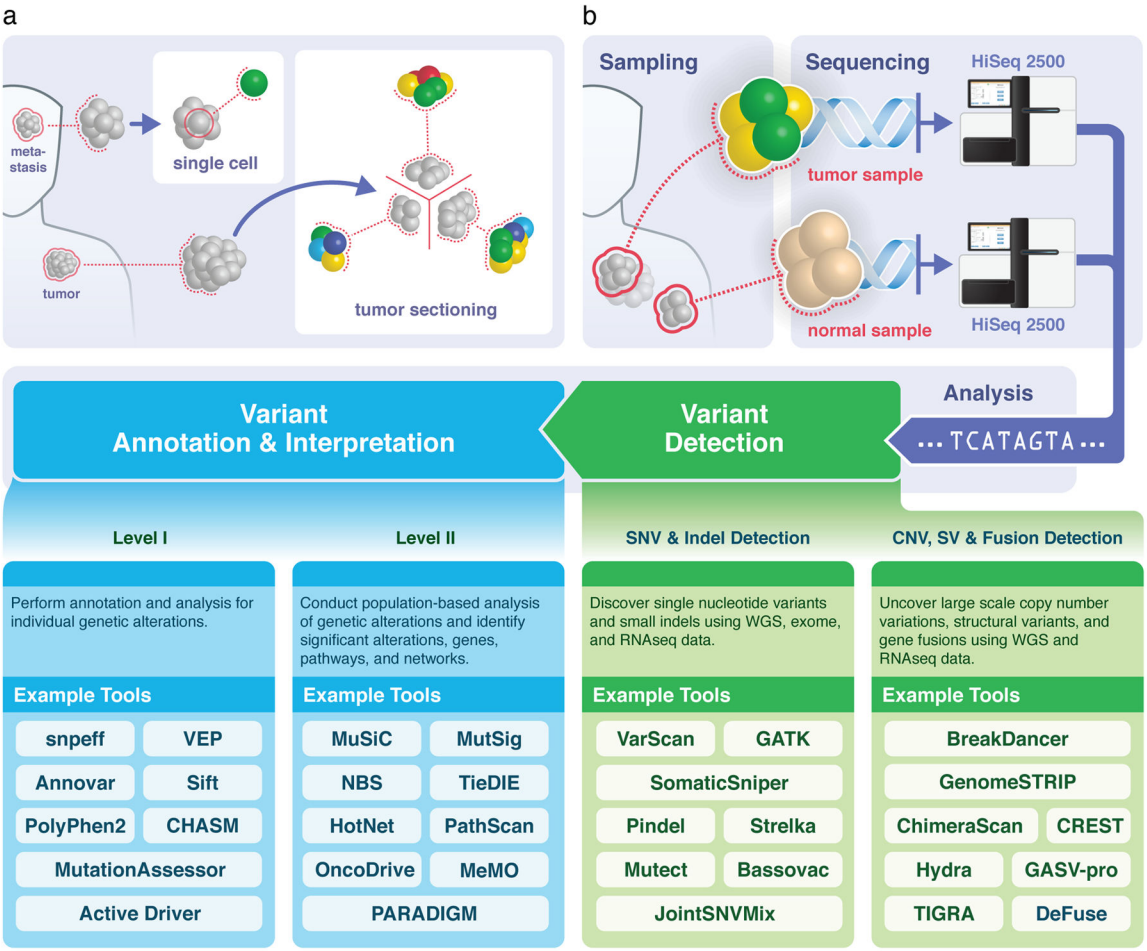


Figure 1. Sample procurement, sequencing, and analysis roadmap

(A) Sequencing strategy: Most cancer genomics investigations sequence the genome of a tumour sample from primary or metastatic lesion, starting with a non-specific 'global' sample pooled from biopsy or resection. Because the spatial distribution of any resident subclones is not known *a priori*, it will become increasingly common to sequence specific regions from a tumor section separately. In the limit, single-cell sequencing can also be performed on flow-sorted nuclei to assess cellular diversity (B) Overview of the sequencing and analysis process: tumour and adjacent healthy tissue samples are sequenced using high-throughput instruments to obtain genome, exome, RNA and other types of data. After alignment, a battery of detection tools identifies both small (SNV, indel) and large (copy number, structural variation, gene fusions) alterations, which are then annotated and analyzed individually (Level I) —for example, for likely functional implications — and collectively (Level II) —for example, to identify relevant gene pathways and networks

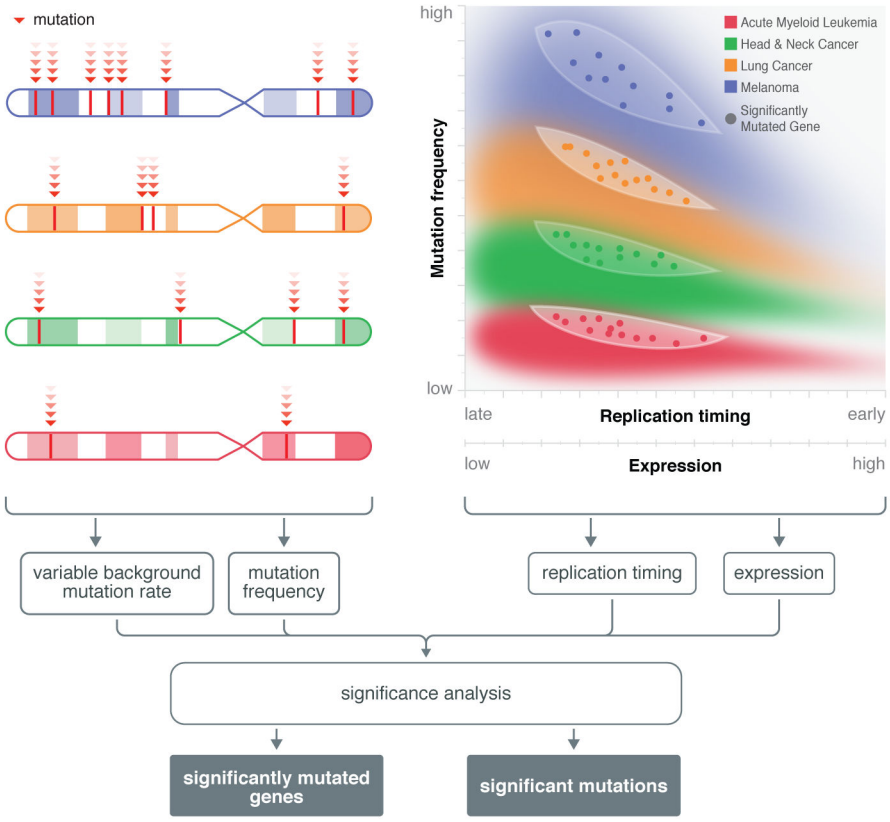


Figure 2. Biological factors relevant to assessing significant genes in cancer

Genomic analysis establishes mutation frequencies of genes and helps characterize background mutation rates. Specific mutation hot spots have been found in the various cancer types. Other factors have also been shown to affect the background mutation rate of a gene, including gene length, expression level, and replication timing. State-of-the-art tools, such as MuSiC and MutSig give proper consideration to these and many other factors, for example transition versus transversion frequency, in determining the significantly mutated genes that contribute substantively to cancer initiation and progression.

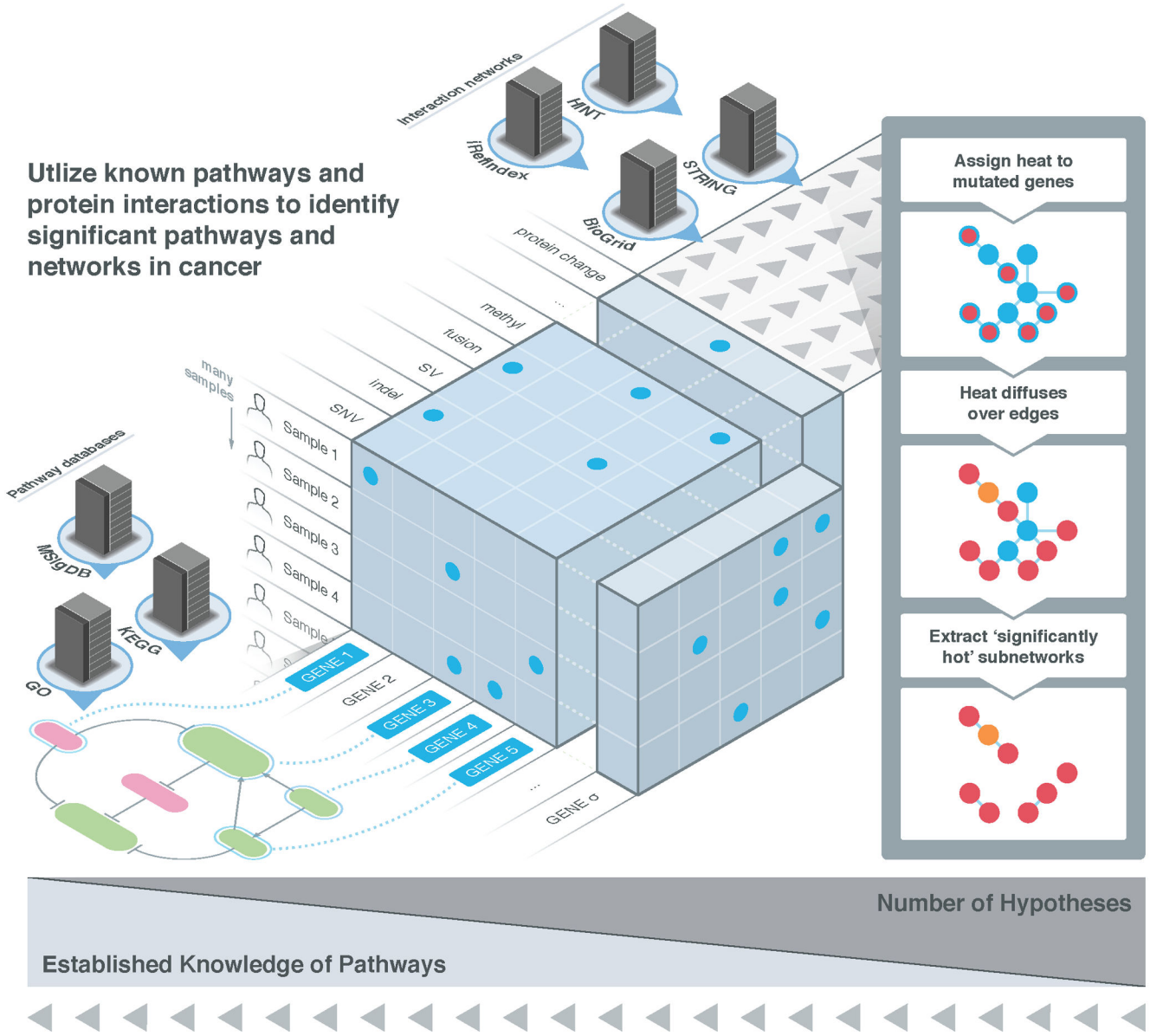


Figure 3. Significantly mutated genes, pathways and networks

Given the mutational status of genes across multiple patients, one can distinguish driver from passenger mutations using several strategies. Single-gene tests determine whether the observed number of samples having a mutation in the gene is significantly greater than what is expected under an appropriate null model. Pathway or gene set approaches examine whether multiple genes in pre-defined sets, as obtained for example from a curated database like KEGG, GO, or MSigDB, have more mutations than expected. These tests are biased to the prior knowledge of gene cascades residing in these databases, but the numbers of tests are relatively small, so the risks associated with Type I error **[G]** tend to be manageable. Conversely, network approaches rely only on knowledge of known protein-protein or protein-DNA interactions in examining combinations of mutations on whole-genome interaction networks, for example using the analog of heat diffusion. Because these

approaches are unbiased, they furnish the possibility of inferring novel combinations of genes relevant to cancer, but larger numbers of hypothesis tests imply that greater care must be taken for multiple testing correction.

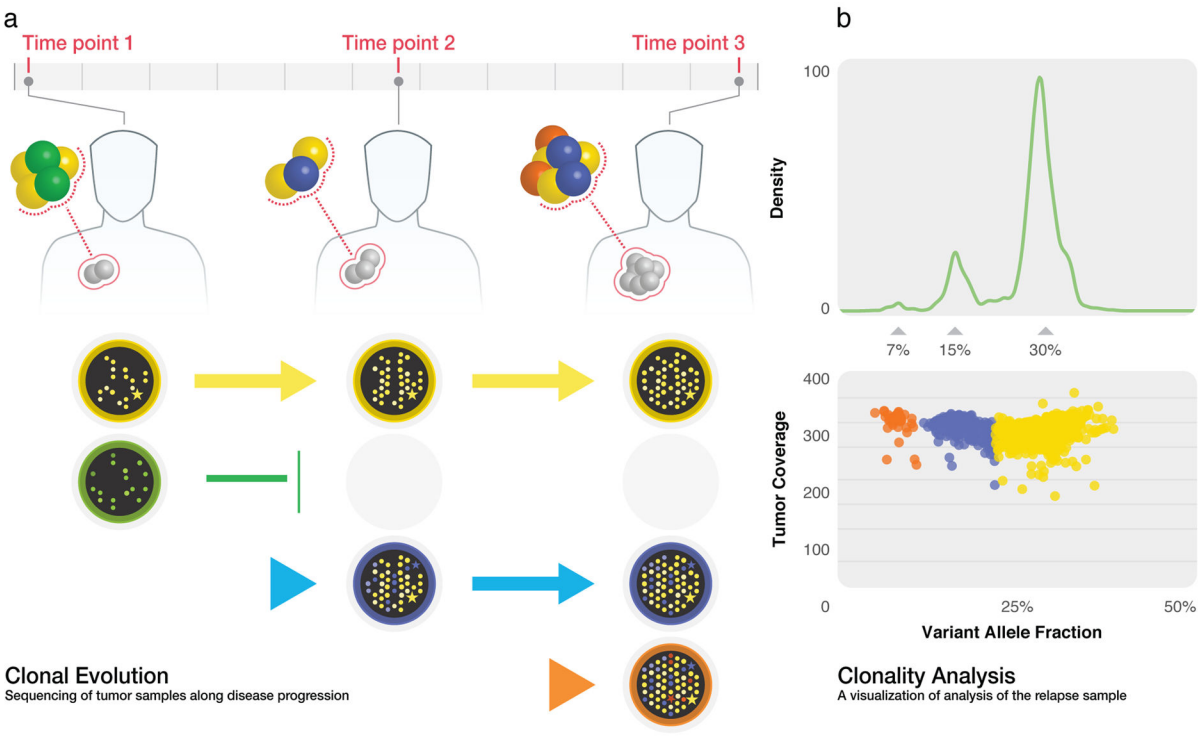


Figure 4. Conceptual example of clonal evolution model and clonality analysis

(A) The founding clone (yellow) persists during the course of the disease. Another clone (green) present at time point 1 faces extinction before time point 2, but new subclones (blue/time point 2 and orange/time point 3) emerge during disease progression. (B) SciClone algorithm detects the three mutation clusters present at time point 3.

Table 1

Computational tools for detecting and interpreting cancer genome alterations.

SV / Indel Detection		
Program	***Analysis***	***Synopsis***
Bassovac (unpublished, Wendl, C et al.)	SNV/indel detection	Bayesian with tumor/normal impurity and clonality
GATK[23]	SNV/indel detection	analysis framework using MapReduce
JointSNVMix[31]	SNV detection	binomial/multinomial probability with pre-filtering
Mutect[28]	SNV/indel detection	Bayesian probability with pre- and post-filtering
Pindel[38]	Indel detection	pattern growth learning method
SomaticSniper[27]	SNV/indel detection	Bayesian probability with posterior filtering
Strelka[29]	SNV/indel detection	Bayesian probability with posterior filtering
SNVMix[30]	SNV detection	Binomial mixture model
VarScan[24,25]	SNV/indel detection	Fisher exact test, fitering, and FDR correction

Copy Number / SV / Fusion Detection		
Program	**Analysis**	**Synopsis**
BreakDancer[54]	SV/indel detection	Kolmogorov-Smirnov test on discordant reads
BreakFusion[68]	fusion detection	alignment-based pipeline for transcriptome data
BreakTrans[73]	fusion mapping	integrate fusion discovery and breakpoint tools
ChimeraScan[67]	chimeric transcription	discordant read pairs with posterior filtering
CREST[55]	SV detection	heuristics/binomial test on soft-clipped reads
DeFuse[65]	fusion detection	dynamic programming split and discordant reads
Delly[40]	SV detection	integrated method of discordant and split reads
GASV-pro[57]	SV detection	plane sweep for segment intersection
GenomeStrip[59]	SV detection	depth and split/discordant reads on populations
Hydra[139]	SV detection	discordant reads with assembly validation
Lumpy (unpublished, Layer, RM et al.)	SV detection	integrated method of discordant and split reads
TIGRA[42]	SV detection	DeBruijn graph-based assembly

Level I Annotation & Interpretation		
Program	***Analysis***	***Synopsis***
Absolute[148]	purity/ploidy/clonality	optimization of log scores
Annovar[74]	functional prediction	annotation-based prediction
ASCAT[166]	purity/ploidy/clonality	goodness of fit ranking of candidate solutions
TUSON Explorer[100]	Gene classification	Oncogene or suppressor using mutation signatures
Classy (unpublished, Bharadwaj, M et al.)	gene classification	oncogene or suppressor using nearest neighbor
CHASM[84,85]	functional preduction	random forest classifier
MutationAssessor[83]	functional prediction	conservation-based prediction (entropy score)
PolyPhen2[81,167]	functional prediction	structure and alignment based probability model
SciClone (unpublished, Miller, C et al.)	tumor clonality	Bayesian mixture model
Sift[82]	functional prediction	conservation-based prediction

Level I Annotation & Interpretation

Program	*Analysis*	*Synopsis*
Snpeff[75]	functional prediction	annotation and coding effect prediction
THetA[151]	purity/ploidy/clonality	maximum likelihood of mixture composition
VEP[168]	functional prediction	annotation-based prediction

Level II Annotation & Interpretation

Program	*Analysis*	*Synopsis*
Dendrix[128]	mutation analysis	de novo discovery of mutual exclusive mutations
HotNet[119]	network analysis	diffusion model for significant networks
MeMO[122]	network analysis	network modules with mutual exclusivity
MuSiC[92]	mutation analysis	framework for significance analysis of mutations
Multi-Dendrix[129]	mutation analysis	de novo discovery of multiple sets of exclusive mutations
MutSigCV[93]	mutation analysis	gene significance with variable background rate
NBS[121]	network analysis	clustering using non-negative matrix factorization
OncoDrive-CIS[169]/-CLUST[170]	mutation analysis	z-statistics for copy numbers of driver genes
Paradigm[126]	gene expression	network analysis of gene expression
PathScan[109]	pathway analysis	probability model for mutation-enriched pathways
TieDIE[125]	network analysis	network diffusion model linking mutations to gene expression