3 Somatic and Germline Variant Calling from Next-Generation Sequencing Data

Ti-Cheng Chang, Ke Xu, Zhongshan Cheng, and Gang Wu

Abstract

Re-sequencing of the human genome by next-generation sequencing (NGS) has been widely applied to discover pathogenic genetic variants and/or causative genes accounting for various types of diseases including cancers. The advances in NGS have allowed the sequencing of the entire genome of patients and identification of disease-associated variants in a reasonable timeframe and cost. The core of the variant identification relies on accurate variant calling and annotation. Numerous algorithms have been developed to elucidate the repertoire of somatic and germline variants. Each algorithm has its own distinct strengths, weaknesses, and limitations due to the difference in the statistical modeling approach adopted and read information utilized. Accurate variant calling remains challenging due to the presence of sequencing artifacts and read misalignments. All of these can lead to the discordance of the variant calling results and even misinterpretation of the discovery. For somatic variant detection, multiple factors including chromosomal abnormalities, tumor heterogeneity, tumor-normal cross contaminations, unbalanced tumor/normal sample coverage, and variants with low allele frequencies add even more layers of complexity to accurate variant identification. Given the discordances and difficulties, ensemble approaches have emerged by harmonizing information from different algorithms to improve variant calling performance. In this chapter, we first introduce the general scheme of variant calling algorithms and potential challenges at distinct stages. We next review the existing workflows of variant calling and annotation, and finally explore the strategies deployed by different callers as well as their strengths and caveats. Overall, NGS-based variant identification with careful consideration allows reliable detection of pathogenic variant and candidate variant selection for precision medicine.

Introduction

Germline variants are nucleotide changes in a germ or egg cells and can be passed to a child from parents during conception. Since the variants are in reproductive cells, they are hereditary mutations and can be passed to future genera-

Ti-Cheng Chang and Ke Xu contributed equally

T.-C. Chang (✉) · K. Xu · Z. Cheng · G. Wu
Center for Applied Bioinformatics, St Jude Children's Research Hospital, Memphis, TN, USA
e-mail: ti-cheng.chang@stjude.org

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_3

tions. Germline mutations account for ~5–10% of cancers [1]. Somatic variants are variants that arose in any cells except germline cells, i.e., sperm and egg, and cannot be transmitted to progeny. Somatic variants include mosaicisms in different subsets of somatic cells including clonal hematopoiesis of indeterminant potential (CHIP). Somatic variants are of particular interests because they are associated with various human diseases, including cancers.

Traditional germline/somatic genetic testing relied on a "panel" of gene testing with a focus on hotspot variants in a number of well-characterized driver genes, such as BRCA1 and BRCA2 [2]. With the advances and reduced cost of the next-generation sequencing (NGS) technology, whole exome/genome sequencing (WES/WGS) and targeted sequencing have become an option for detecting variants on a much larger scale and higher definition. A major challenge of WGS/WES analysis is the accuracy of mutation calling analyses on single nucleotide variants (SNVs) and small *in*sertions and *del*etions (indels).

Development of SNV/Indel Variant Calling in the Past Years

NGS workflow usually starts with the fragmentation of the genome or targeted regions of genomes into small fragments, followed by alignments to reference genomes or genome re-assembly. The aligned/piled-up segments are used subsequently for variant detection. In early studies, the variant calling was performed by counting alleles at each site with simple cutoff rules to determine a variant call, which oftentimes lacks sensitivity to detect heterozygous alleles and does not provide confidence level of the genotype calls [3].

Uncertainties of variant calls arise when a sample's coverage is shallow, sequencing read quality is poor, or a variant site has low allele count support [4]. After variant calling, layers of filters are therefore suggested to be applied to filter the variant calls to reduce the likelihood of sequencing artifacts in the call sets and increase the confidence of variant calls. An in-depth overview of filters that can be considered is described in section "Contributing Factors for Bogus Somatic Variant Calling" of this chapter.

Germline and somatic variant calling algorithms differ in the assumption of expected allele frequency. Germline variants are expected to have 50% or 100% allele frequencies to differentiate three basic genotypes harbor at each variant site, e.g., homozygous allele A (AA), heterozygous (AB), or homozygous allele B (BB). On the contrary, for somatic variant calling, the allele frequency displays a larger spectrum of variations symbolizing distinct stages of cell development. An increasing number of algorithms have been developed in the past decades to enhance the calling accuracy by incorporating error rate estimation and probability frameworks to model the genotyping and phasing likelihoods. Given the complexity of genomes, local re-assembly was also placed into the calling scheme to increase the confidence of variant calling. Table 3.1 provides a summary of available tools for somatic and/or germline variant calling to date. In the following section, we will introduce the algorithms implemented in a few popular variant callers.

Algorithm Basis of Germline SNV/Indel Variant Calling

Samtools mpileup [5] deployed the approach of read coverage depth counting to identify coverage characteristics of potential SNVs/indel sites. The coverage information was then fed into BCFtools [6] for variant calling based on general Bayesian likelihood. This approach is usually used for germline variant calling.

GATK HaplotypeCaller [7] is a widely used germline variant caller. An advantage of GATK is that the algorithm can be applied for the joint calling of a group of samples at the same time to control the false discovery rate and increase the sensitivity of low-frequency variant detection. In addition, GATK allows the re-assembly of reads to re-construct the real allelic segment or haplotype, which will be realigned to the reference genome to identify the variant sites. GATK HaplotypeCaller begins with defining active

Table 3.1 List of publicly available tools for variant calling in chronological order

Software	Algorithm detail	Type of variant	Single-sample mode	Year published	References
GATK	Haplotype analysis and Joint genotype analysis	SNV/indel	Yes	2010	[7]
SAMtools	Joint genotype analysis	SNV/indel	Yes	2011	[6]
SomaticSniper	Joint genotype analysis	SNV	No	2011	[14]
MutationSeq	Machine learning	SNV	No	2012	[33]
JointSNVMix2	Joint genotype analysis	SNV	No	2012	[79]
VarScan2	Heuristic threshold	SNV/indel	Yes	2012	[16]
deepSNV	Allele frequency analysis	SNV	No	2012	[80]
LoFreq	Allele frequency analysis	SNV/indel	Yes	2012	[81]
FreeBayes	Haplotype analysis	SNV/indel	Yes	2012	[9]
EBCall	Allele frequency analysis	SNV/indel	No	2013	[82]
Shimmer	Heuristic threshold improved for highly contaminated or heterogeneous samples	SNV/indel	No	2013	[83]
Seurat	Bayesian-based analysis of sequenced genome pairs	SNV/indel, SV	No	2013	[84]
Virmid	Joint genotype analysis improved by inferring sample impurity	SNV	No	2013	[85]
qSNP	Heuristic threshold with low tumor content	SNV	No	2013	[86]
MuTect2	Allele frequency analysis	SNV/indel	Yes	2013	[13]
BAYSIC	Machine learning (ensemble caller)	SNV	No	2014	[87]
FaSD-somatic	Joint genotype analysis	SNV	Yes	2014	[88]
Platypus	Haplotype analysis	SNV/indel, SV	Yes	2014	[89]
HapMuC	Haplotype analysis	SNV/indel	Yes	2014	[90]
RADIA	Heuristic threshold with RNA and DNA integrated analysis	SNV	No	2014	[91]
SOAPsnv	An integrated tool for somatic single-nucleotide variants detection with or without normal tissues in cancer genome	SNV	No	2014	[92]
SomaticSeq	Machine learning (an ensemble approach to detect somatic mutations)	SNV	No	2015	[34]
LocHap	Haplotype analysis	SNV/indel	No	2015	[93]
VarDict	Heuristic threshold	SNV/indel, SV	Yes	2016	[94]
SNVSniffer	An integrated caller for germline and somatic single-nucleotide and indel mutations	SNV/indel	Yes	2016	[95]
MuSE	Markov chain model	SNV	No	2016	[17]
SNooPer	Machine learning for low-pass next-generation sequencing	SNV/indel	Yes	2016	[96]

(continued)

Table 3.1 (continued)

Software	Algorithm detail	Type of variant	Single-sample mode	Year published	References
CaVEMan	Joint genotype analysis	SNV	No	2016	[97]
LoLoPicker	Allele frequency analysis	SNV	No	2017	[98]
Strelka2	Mixture-model-based estimation for calling of germline and somatic variants	SNV/indel	No	2018	[18]
Cerebro	Machine learning (random forest)	SNV/indel	No	2018	[35]
DeepVariant	Deep convolutional neural network (CNN) to call germline SNV/indel	SNV/indel	No	2018	[12]
NeuSomatic	Convolutional neural network	SNV/indel	No	2019	[99]
NeoMutate	An ensemble machine learning framework	SNV/indel	No	2019	[29]
SMuRF	Machine learning	SNV/indel	No	2019	[31]
DeepSSV	Convolutional neural network	SNV/indel	No	2020	[100]

regions where abundant evidence has shown the presence of variants. Only the active region is used for variant calling to reduce the time on the assembly. With the assembly step, the variant calling is not only dependent on the read alignment against the reference genome but also the reconstructed haplotype. The overall GATK algorithm takes a divide-and-conquer concept by shredding the sequencing data into small chunks for parallel processing; however, its efficiency is still a concern when processing a large collection of samples for joint calling. Approaches have been proposed to address the performance issue when dealing with a large number of samples [8].

FreeBayes [9] applied a Bayesian framework to relate the likelihood of sequencing errors of the reads and the prior likelihood of a particular genotype. Also, the phase of haplotypes was inferred from the reads, and the non-uniform copy number of samples was taken into consideration. FreeBayes is usually used for germline variant calling, while it has been expanded for somatic calling [10]. FreeBayes shows good performance across sequencing platforms for SNV calling, but it tends to have a higher false-positive rate for indel sites [11].

DeepVariant [12] performs variant detection using a convolutional neural network (CNN) learning model implemented via the python TensorFlow library. DeepVariant identifies variants through learning the features in images of pileup reads surrounding putative variants and true genotypes. A version of DeepVariant for somatic calling is still under development.

Algorithm Basis of Somatic SNV/Indel Variant Calling

Mutect2 [13] as a part of the GATK toolkit shares a similar process of variant calling with GATK and is mainly used for somatic calling with matched, paired tumor-normal samples. Mutect2 also allows tumor only calling (see section "SNV/Indel Variant Calling"). Mutect2 calls SNVs and indels simultaneously via the local de novo assembly of haplotypes in an active region as described previously. Mutect2 reassembles the reads present in the active regions to candidate variant haplotypes. Each read is then aligned to each haplotype via the Pair-HMM algorithm to obtain a matrix of likelihoods. Finally, log odds were derived to distinguish somatic variants from sequencing errors by a Bayesian somatic likelihood model.

SomaticSniper [14] is another somatic variant caller. SomaticSniper determines the somatic status of a variant site by comparing the site's genotyping likelihood between normal and tumor derived from the MAQ tool [15] using a Bayesian approach. SomaticSniper implemented internal filters to exclude the sites with poor read/base quality or with low read support to reduce calling artifacts.

VarScan2 [16] relies on the results from SAMtools pileup or mpileup for somatic variant calling. At each variant site, VarScan2 compares the genotypes and supporting read counts between tumor and normal to determine the somatic status, and the call-set is refined with post-calling filters including the variant position in a read, strand bias, read coverage depth, variant frequency, homopolymer, mapping quality, and so on [16]. Of note, VarScan2 also allows the germline variant calling and detection of somatic copy number abnormality (SCNA).

MuSE [17] somatic calling starts with matched tumor-normal alignment BAM files. The alignment is first filtered for sequencing artifacts. The evolutionary F81 Markov substitution model of DNA is applied to describe the changes from reference to tumor allele compositions with estimates of equilibrium frequencies for all alleles and evolutionary distance. With the frequencies, MuSE derived a sample-specific error model and five-tier-based cutoffs to address the variations present in the frequency distribution in tumor and normal samples. The tier-based approach allows the MuSE to retain variants with low variant allele frequency to achieve a higher sensitivity.

Strelka2 [18] is an open-source somatic/germline variant caller developed by Illumina®. The somatic calling algorithm of Strelka2 is enhanced based on the original Strelka [19] method to account for tumor-in-normal contamination that is essential for liquid tumor variant analyses.

Strelka first identifies indel regions and performs realignment. After realignments, Strelka derives a somatic variant probability using the tumor and normal samples and deduces the somatic status of a site after accounting for the status of loss of heterozygosity (LOH) or copy number change regions. Strelka applied a two-tier-based filtering strategy with distinct filters and sensitivity. Similar to other tools, post-filtering is applied by Strelka2 to handle different types of potential calling errors.

The variant calling is usually computationally intensive, particularly when the sample number is large. To improve efficiency, Illumina® has released a Dynamic Read Analysis for GENomics (DRAGEN) platform using a highly configurable field-programmable gate arrays (FPGAs) hardware to accelerate the analysis processes [20]. DRAGEN first identifies callable regions and assembles the haplotypes using *De Bruijn* graph method. The reassembly is aligned to the reference genome to identify the variants. The probability of all read alignments to the haplotype is calculated via the pair hidden Markov model that is speeded up using the FPGA and summed up for each read. In the end, the diploid genotype is calculated to determine the variant calls.

In the past few years, GPU-based read alignment and variant calling solutions have also been developed to reduce the WGS data processing time to a couple of hours. For example, NVIDIA Clara Parabricks pipelines include a somatic variant calling workflow that integrates GPU-based alignments by BWA-MEM and downstream somatic variant calling by Mutect2 [13] or DeepVariant [12]. Parabrick also allows germline calling using GATK HaplotypeCaller [7]. The pipeline reduces the time taken for a typical 30× WGS data by over an order of magnitude.

SNV/Indel Variant Calling Workflows

Variant calling workflow can be compartmentalized into four steps: data preprocessing, variant calling, variant filtering, and variant annotation. Each step has its challenges and strategies. We detail these steps as follows.

Data Preprocessing

The raw read quality can be examined using FastQC [21]. FastQC identifies the potential read issues before mapping. A good WGS/WES read library usually has an average read base quality >20 and a low level of duplicated or overrepresented sequences.

Selection of the reference genome is the first step for correct variant calling. The latest version of the human reference genome GRCh38 (Hg38) with improved resolution [22] is suggested for human variant analyses. Also, the reference is recommended to include decoy genome sequences for the alignment purpose to reduce misalignments, as well as virus sequences that are known in human to attract the viral reads. In addition, the alternative contigs from highly complex loci, such as the human HLA allele region, should be included to reduce SNV/indel calling artifacts. For read alignments, frequently used aligners are BWA [5], Bowtie2 [23], and Novoalign (http://www.novocraft.com/products/novoalign/). Benchmarks of short-read aligners indicated that the MEM algorithm implemented in BWA achieved a better balance between specificity and sensitivity [24, 25]. BWA-MEM is suggested to use when read length is greater than 70, while BWA-ALN for shorter reads [26].

Following alignments, duplicate reads generated from PCR artifacts are flagged using tools such as GATK MarkDuplicates to prevent downstream variant calling errors. Incorrect read alignment surrounding the indel regions frequently causes inaccurate substitution calls. These alignment artifacts can be reduced through indel realignments by GATK IndelRealigner or similar tools. Furthermore, the base quality produced by different library preparation protocols and sequencing instruments would have different levels of technical or chemistry errors. GATK toolkits comprised two tools, BaseRecalibrator and ApplyBQSR, to facilitate the correction of these systematic errors. These tools implemented

machine learning approaches to model errors and adjust base qualities to obtain a more accurate overall base quality profile. Figure 3.1a shows a general workflow for the data preprocessing.

SNV/Indel Variant Calling

The next step is to choose appropriate variant callers. The GATK tool suite is well performed for the germline SNV/indel calling. A number of best practices for variant callings have been provided by GATK (https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows). For somatic variant calling, accurate identification of a somatic variant is still not trivial due to varied caller performance and tumor heterogeneity. Below we describe three common scenarios in somatic and germline variant calling as well as variant prioritization in cancer genomics.

Somatic Mutation Calling on Matched Tumor-Normal Pairs

Variant calling with matched tumor-normal sample pairs is the most common scenario for the identification of somatic variants (Fig. 3.1b). Most of the callers use the aligned BAM files of paired tumor and normal samples as the standard inputs. To identify low-frequency variants, a caller that can model the allele frequency is suggested, such as Mutect2, MuSE, and Strelka2 as detailed in the Introduction. Due to the differences of underlying algorithms and statistic modeling, the somatic variant callers differ in sensitivity and specificity when detecting variants at different levels of variant allele frequencies (VAF) [27]. Compared with Strelka and Mutect, SomaticSniper has a lower sensitivity and specificity when calling the variants with VAF <8%. However, the performance of SomaticSniper is comparable with Strelka and Mutect for variants with VAF >18%. The sensitivity of VarScan2 was increased with lower minimum allele fraction thresholds, which was however compromised with reduced specificity [28]. Therefore, a careful setting of thresholds to

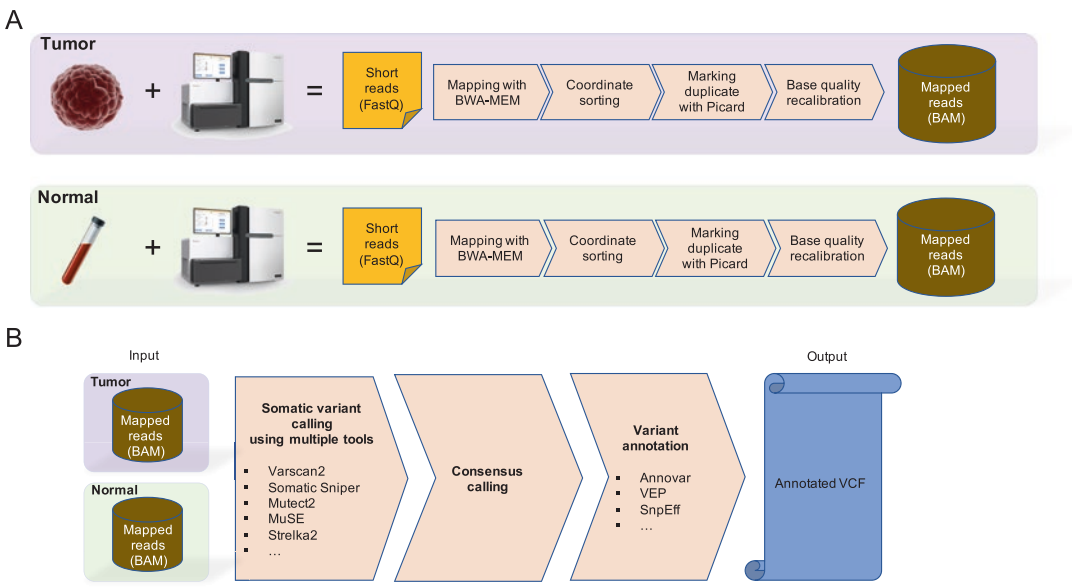

Fig. 3.1 The workflow of the somatic variant calling of paired tumor-normal samples. (**a**) Data preprocessing steps from sample preparation to short reads mapping and calibration into binary version of Sequence Alignment/Map (BAM) files for paired tumor and normal samples. (**b**) Variant calling and annotation steps from paired tumor-normal BAM files to annotated somatic variants in VCF format

achieve a balance between sensitivity and specificity for each caller and a well-considered post-calling filtering strategy play important roles to assure the validity of final call sets.

Given the complex heterogeneity and structural rearrangements of tumor tissue, finding an appropriate somatic variant caller along with parameter fine-tuning and development of a solid calling strategy remain a major challenge for cancer genomics. To tackle this complexity and exploit each caller's strength, a consensus voting to determine a valid variant call by multiple callers has gradually become a prevalent strategy in studies [29–33]. In addition to a simple voting strategy, machine learning has been incorporated into the consensus calling steps to improve calling performance. MutationSeq incorporated multiple sequence quality features derived from normal data based on Samtools and GATK, along with several sequence artifacts and low-frequency variant features to build classifiers to determine the somatic variants [33]. SomaticSeq [34] integrated five somatic callers from which feature sets were identified for each candidate variant position to build a classifier using a stochastic boosting machine-learning algorithm. Cerebro [35] applied a random forest classification model to generate a confidence score for each candidate variant derived from whole-exome sequencing data, which is limited to the coding region with >150× coverage. These approaches generally lack portability, i.e., users are required to obtain appropriate training data and have knowledge about the machine learning to re-train the models. In light of these issues, SMuRF [31] was developed and generalized for either WGS or WES data. SMuRF implemented a supervised machine learning using features derived from four variant callers along with mapping auxiliary features. NeoMutate [29], as another machine learning based caller, profiled a collection of seven distinct classifiers based on a training dataset of >3000 cancer variants from the Catalogue of Somatic Mutations in Cancer (COSMIC) database [36].

Machine learning–based callers determine the somatic status of a variant through different features of a variant harbors and therefore offer a higher level of flexibility than rule-based filtering strategy, especially for the tumor samples with intra-heterogeneity and normal tissue admixtures. However, a detailed curation of a set of ground-truth training data including both true-positive and true-negative variants is the key to optimize and refine the training models.

Mutation Calling and Prioritization on Tumor Sample Without Matched Normal Sample

In large-scale cancer genomic projects, it is common to have tumor samples without matched normal samples or with tumor-contaminated adjacent normal samples, due to the difficulties to collect patients' blood samples. In these cases, the somatic variant calling oftentimes has a high rate of false positives, because it is almost impossible to confidently determine whether a called variant is of germline origin or somatically acquired. Mutect2 can call somatic mutations in tumor-only mode; however, the calling results require careful filtering for false positives due to the deficiency of corresponding germline information. Common germline SNPs can be eliminated by filtering against appropriate human genome variation databases such as Genome Aggregation Database (gnomAD). To date, limited number of studies have compared the performance of Mutect2 tumor-only and tumor/normal calling modes when both tumor/normal WGS/WES data are available. A tool designed specifically for somatic mutation calling on tumor-only WES samples is ISOWN [37], which utilizes a family of supervised learning classifications to distinguish somatic SNVs in NGS data from SNPs in the absence of normal samples. In terms of performance, the F1-measure of ISOWN is between 75.9% and 98.6% across different cancer types, cell lines, fresh frozen tissues, and formalin-fixed paraffin-embedded tissues. Calling somatic variants in tumor only WGS/WES data still warrants further improvement.

Due to these challenges, one can consider focusing on identifying putatively pathogenic variants in a set of genes of interest to specific tumors, irrespective of their germline or somatic origin (Fig. 3.2). Specifically, after basic variant

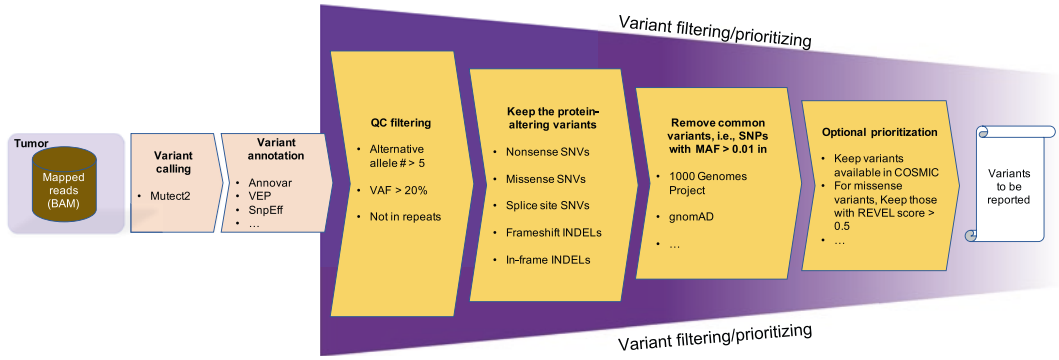

Fig. 3.2 The workflow of the variant calling of tumor sample without a matched normal sample. The workflow focuses on reporting potentially pathogenic variants regardless of their tumor or germline origin

quality filtering such as keeping variants with higher alternative allele count (>5) and VAF (>20%), and excluding those located in regions of low complexity or regions with extreme GC content, additional filters can be applied for the variant class and population frequency filter, i.e., only keeping protein-altering variants with minor allele frequency <0.01 in population frequency databases such as 1000 Genomes [38] and gnomAD [39]. In addition, optional filters can be added to increase the calling confidence such as keeping any variants that are available in the COSMIC catalog of somatic mutations or missense variants with a REVEL score >0.5 [36, 40].

Germline Mutation Calling and Prioritization

Identifying germline mutations in cancer predisposition genes has important implications in understanding tumorigenesis and guiding clinical practice. A common germline mutation calling workflow is illustrated in Fig. 3.3a. The recommended germline variant calling follows the GATK best practices including read mapping, alignment sorting, duplicated reads marking, and variant calling by GATK HaplotypeCaller [7]. Also, joint variant calling in multiple germline samples is recommended whenever possible because the genotype information at the population level can be leveraged to rescue the variant at a site with low coverage or with lower quality in a sample. The efficiency of GATK calling can be enhanced by a divide-and-conquer strategy, i.e., splitting the genomes into multiple small chunks for parallel variant calling followed by merging the output variant files (VCFs). After variant calling, the GATK Variant Quality Score Recalibration (VQSR) method is the suggested approach to filter the germline variants. VQSR relies on a deep learning method and therefore requires a sufficient amount of the variant sites to establish a reliable training model. The variant number for a single-sample WGS is usually sufficient for VQSR; however, for WES data, at least 30 samples are required to perform VQSR. When the sample size is limited, the variant call set can be filtered by the GATK VariantFiltration tool.

To narrow down from the vast amount of germline variants reported by germline variant caller, usually only rare, non-silent coding variants in cancer-related genes, such as autosomal dominant or autosomal recessive cancer-predisposition genes, or genes that are recurrently mutated in tumors, are considered. For example, Zhang et al. evaluated germline mutations in a cohort of pediatric cancers in a curated list of 565 cancer-related genes based on expert reviews of the genes from American College of Medical Genetics and Genomics (ACMG) and genes from related literatures [41]. Specifically, after germline variant calling, QC-passed variants are shortlisted based on their frequencies in human populations such that only novel variants or the variants with minor allele frequency <0.001 in NHLBI Exome Sequencing Project (ESP) are kept [42]. These shortlisted variants

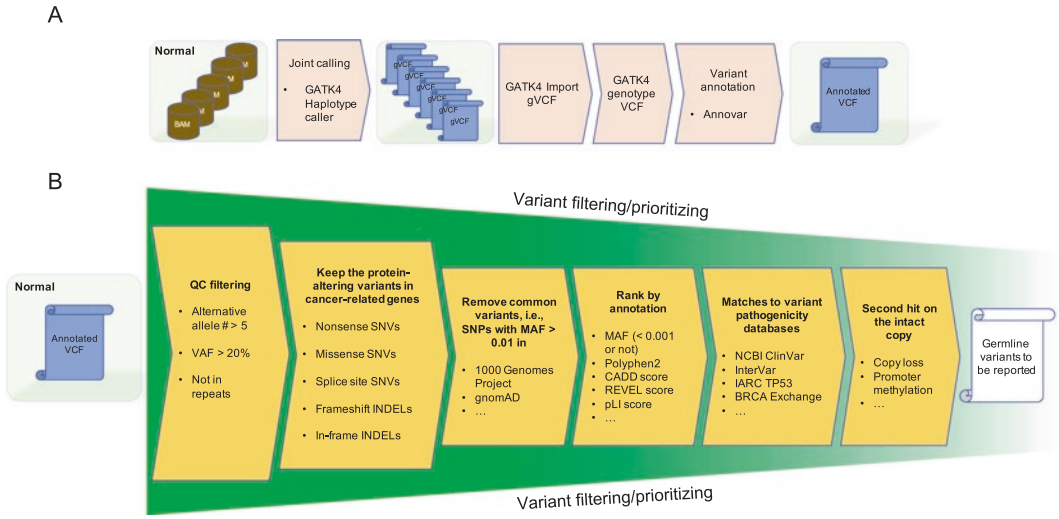

Fig. 3.3 The workflow of germline variant calling and prioritization. (**a**) Steps of the joint calling of germline variants from pooled germline BAM files. (**b**) Steps of filtering and prioritizing potentially pathogenic germline variants or variants of unknown significance

can be then ranked based on (1) mutational class such as nonsense SNVs, missense SNVs, splice site SNVs, frameshift indels, or in-frame indels; (2) functional annotation databases such as PolyPhen2 and MutationAssessor [43, 44], (3) matches to curated variant pathogenicity databases such as NCBI ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), locus-specific databases such as IARC TP53 (https://p53.iarc.fr/) and BRCA Exchange (https://brcaexchange.org/); and (4) second hit on the intact copy in the tumor genome due to one copy loss or promoter methylation of the intact copy. Other popular databases for germline variant classification and prioritization include pLI and LOFTEE scores for loss-of-function variant prioritization [39, 45]; REVEL and CADD scores for missense variant prioritization [40, 46]; and dbscSNV scores for splice variant prioritization [47]. In addition, InterVar, an automatic interpretation of variants based on dozens of criteria laid out by ACMG and Association for Molecular Pathology (AMP), can be included to aid manual review of clinical significance [48]. Figure 3.3b summarizes the filtering steps to prioritize germline variants to be reported. The final ranked list of putatively pathogenic germline variants will then need to be manually reviewed and validated based on phenotype data, RNA-seq, and literature review. The whole prioritization process before manual reviews can be automated. For example, St. Jude Pediatric Cancer Variant Pathogenicity Information Exchange (PeCan PIE, https://pecan.stjude.cloud/pie), a free cloud service for non-commercial use, offer variant annotation and ranking service based on MedalCeremony pipeline to triage the germline variants into three categories, including Gold, Silver, and Bronze [41, 49].

Variant Annotation

To understand the context of the germline variants and somatic mutations, several tools are available to perform variant annotation on the called variants. Typically, the genomic locations of the variants are compared against a gene-based annotation database such as a GENCODE release (https://www.gencodegenes.org/pages/data_access.html) to determine if a variant is exonic, intronic, or intergenic [50]. Variants in exonic regions are further classified as missense

variants, nonsense variants, silent variants, splice acceptor variants, splice donor variants, splice region variants, in-frame indels, and frameshift indels. Some annotation tools such as ANNOVAR [51], VEP [52], and SnpEff [53] also add population allele frequency from 1000 Genomes Project [38], NHLBI ESP [42], Exome Aggregation Consortium (ExAC) [45], and gnomAD [39]; and provide comparative genomics-based scores such as GERP++ [54], SIFT [55], PolyPhen2 [43]; and include machine learning–based pathogenicity scores such as CADD [46, 56] and REVEL [40].

ANNOVAR [51] is an annotation pipeline to functionally annotate variants. The workflow can be performed for either gene-based coding change annotations or region-based non-genic genomic element annotations. Moreover, ANNOVAR has extended functionality to identify and filter variants documented in specific databases, which can be used for enriching causal variants in diseases. ANNOVAR allows the annotation of SNVs and structural variants from a standard VCF. A web interface is available via wANNOVAR (http://wannovar.wglab.org/).

VEP [52] is another popular toolkit for variant annotation. Compared to ANNOVAR, VEP provides cell-line-based annotation. VEP generates transcript-level annotations, while ANNOVAR gives gene-level annotations. LOFTEE (Loss-Of-Function Transcript Effect Estimator, https://github.com/konradjk/loftee) is a very useful VEP plugin to evaluate the loss of function of splice variant [39]. VEP also allows the variant annotation of species other than human and mouse. In addition to local installation, users can perform annotations through the VWP web interface (https://uswest.ensembl.org/info/docs/tools/vep/online/index.html) or cloud virtual machine.

SnpEff [53] implements an interval forest algorithm to efficiently query, annotate, and predict the effect of the variants. SnpEff can run locally or via a Galaxy instance. Similar to VEP, SnpEff also provides a cloud VM for users. SnpEff allows the assessment of nonsense mediated decay (NMD), a functionality absent from ANNOVAR and VEP.

Contributing Factors for Bogus Somatic Variant Calling

Somatic variants generated from the variant callers oftentimes include false positives due to various types of contributing factors. Below we describe four common scenarios that cause bogus somatic variants calling and need to be considered in postprocessing.

Strand Bias

Strand bias is observed when reads are favorably sequenced for one strand over the other; only one strand of the DNA has reads covered in extreme cases. The sources of this type of artifact remain elusive but may be relevant to library preparation of analytic procedures [57]. This bias raises the concerns of variant call accuracy. GATK and Samtools both implement functionality to calculate strand bias scores.

Repetitive DNA Sequences

Repetitive DNA sequences are sequences that are identical or similar across the genome. They vary in sizes and frequencies and cause mapping ambiguities. RepeatMasker [58] can be used to mark or mask the repetitive sequences in the genome to reduce such ambiguities. The error rate of short reads sequencing has been shown to increase in genomic regions with high- and low-GC content or with long homopolymer runs [59]. Also, the GC-rich regions frequently suffered from low coverage issues. Segmental duplication can also cause some reads mapped to multiple places in the genome and give rise to unusual coverage. A BLAT (BLAST-like alignment tool, available at http://genome.ucsc.edu/cgi-bin/hgBlat) search can be used to determine if the flanking sequence of a variant with high coverage is uniquely mapped to a locus or multiple different loci. Those that can be mapped to multiple loci in the genome are recommended to be reviewed manually.

Variants in simple repeats or homopolymer regions, such as CCCCCCCC or ACGACGACGACG ($[ACG]_n$), often lead to false-positive variant calls due to sequencing errors and following read misalignments. Indels

in repetitive regions coupled with low alternative reads count support are usually filtered out. However, frameshift indels in disease-causing genes (e.g., ATRX, PMS2) require careful visual inspection and perhaps validation with an orthogonal sequencing approach to avoid missing important findings.

Low-Frequency Variants

VAF is the number of reads supporting the alternative allele divided by the total number of reads covering the genomic location. For germline samples, a heterozygous germline variant would have an approximately 50% VAF. Germline variants with significantly low VAF and a low number of alternative reads count could be due to sequencing errors. Germline variants with sufficient alternative read count and total read count but with low VAF may indicate mutation mosaicism [60]. If a large number of germline variants have low VAFs, it may suggest that the normal sample is contaminated by the tumor sample, which sometimes happens when the normal sample is collected as tissue adjacent to the tumor or blood after treatment. Paralogous mapping can also lead to VAF ranging from 10% to 25%.

Somatic mutations, on the other hand, exhibit a broader range of VAFs. A heterozygous somatic mutation in a copy-intact region would have an approximately 50% VAF. However, since tumor genomes are frequently subject to copy number alteration, the VAF of a somatic mutation could be around 33% or 67% due to one copy gain and could be close to 100% because of LOH. In addition, since patient tumor samples are rarely 100% pure, low tumor purity may further contribute to the global dilution of VAFs of somatic mutations in a tumor genome. Mutations with significantly lower VAFs than the truncal mutations in a tumor genome but with sufficient mutant read counts may suggest that they are subclonal. Somatic mutations with significantly low VAF and few alternative allele read counts could be due to sequencing error/artifacts and are recommended to be filtered out.

Germline Variant Contamination

A few somatic SNV callers, e.g., Mutect, have implemented specific filters to eliminate the potential germline variant contamination in somatic variants calling. Mutect allows the inclusion of a panel of normal samples (PON) and dbSNP database to exclude germline variants. The germline variant contamination can also be reduced by checking minor allele frequencies of mutations across different population frequency databases such as gnomAD and the 1000 Genome Project database. A recent study [61] reported that there would be one germline SNP among a median somatic SNVs prediction set containing 4325 somatic SNVs; the study also reported a negative correlation between germline SNP contamination and tumor purity.

Concluding Notes

Somatic variant calling from WGS/WES is critical for cancer genomics as it not only depicts the mutational landscape for a tumor sample but also serves as input data for downstream analyses such as mutational signature and clonal evolution. Consequently, there has been great interest in developing fast, accurate, and scalable methodologies and tools for variant calling across academia and industry. In addition to the tools mentioned above, there are also other variant calling tools acting on different data types and different platforms as described below.

Mitochondria Mutation Calling

Variants present in the mitochondria genome (mtDNA) is implicated in a wide spectrum of human disorders and diseases with highly divergent phenotypes and penetrance. The challenges of mtDNA variant calling arise from the circular topology of mtDNA as well as the homology between mtDNA and a part of the nuclear genome with mitochondrial origin (nuMTs). The mtDNA mutation load also varies greatly among tissues

and organs from heteroplasmy (<100%) to homoplasmy (100%). The Human Mitochondrial Genome Database, Mitomap [62], provides a repertoire of reported mtDNA variants. Nuclear genome variant callers such as VarScan and LoFreq have been used for identifying the somatic mtDNA variants [63, 64]. MitoCaller [65] of the MitoAnalyzer toolkit was designed specifically to infer the mutation status of each position of the mitochondria genome using likelihood-based models and adapted an iterative alignment strategy to account for the circularity of the mtDNA genome. Importantly, discrepancies of mtDNA variant calling have been reported when using different reference genome and enrichment strategies [64], which should be taken into consideration when performing mtDNA variant calling and interpretation.

Long-Read Variant Calling

While short reads from paired-end sequencing were used by most state-of-the-art SNV callers to accurately detect variations in diploid genomes, they provide limited haplotype information that is required by some SNV callers, such as GATK HaplotyperCaller and FreeBayes. In addition, the accurate calling of SNVs in repetitive regions of the human genome is another challenge. Third-generation sequencing (TGS) technologies, including Pacific Biosciences and Oxford Nanopore (ONT), have the potential to overcome the limitations of short-read sequencing. Nevertheless, compared to short-read sequencing, long-read sequencing usually costs more and generates less-accurate long reads (e.g., sporadic indels in ONT data), posing challenges for accurate variant detection [66]. Current SNV callers using TGS data are mostly designed for germline variants calling and usually optimized based on the publicly available data from the Genome in a Bottle (GIAB) Consortium. Somatic SNV calling based on long reads technology is still underdeveloped.

NGS-based mapping tool such as BWA-mem is not suitable for long reads mapping. Instead, new mapping tools such as Minimap2 [67] and NGMLR [68] have been developed specifically for long reads mapping. Similarly, NGS-based SNV calling tools such as GATK HaplotyperCaller and FreeBayes are not recommended for variant calling on long-reads sequencing data. Instead, several variant callers have been developed specifically for long-reads data to leverage haplotype information available in long reads to improve the accuracy to call and phase SNVs in diploid genomes, as well as mapping variants in duplicated regions of the genome that are not possibly mapped using short reads. For example, Longshot [66] takes advantage of the haplotype information present in PacBio long reads to improve the SNV calling accuracy [69]. WhatsApp [69] introduces a novel statistical framework for the joint inference of haplotypes and genotypes from noisy long reads, which takes full advantage of linkage information provided by PacBio long reads. Clairvoyante [70] uses a multi-task five-layer convolutional neural network model to predict variants. Other tools include DeepVariant for variant calling on PacBio data [12] and MarginPhase (https://github.com/benedictpaten/marginPhase) for simultaneous haplotyping and genotyping on Oxford Nanopore data.

Different tools differ in their precision and recall rate. In a benchmark study using PacBio data from GIAB, three callers, including Longshot, WhatsApp, and Clairvoyante, demonstrating very similar performance [66]. Compared to the previous three tools, MarginPhase performed moderately when focused on GIAB high confidence regions [69]. Another software, HELLO [71], has been created to integrate the short read and long read data to improve the robustness of SNV calling by leveraging the Mixture of Experts paradigm that uses an ensemble of deep neural networks (DNNs).

Variant Calling in Single-Cell Data

Single-cell sequencing has been the hotspot of functional genomics to elucidate the heterogeneity of cell compositions. Variant calling of single-cell data can aid the inference of the lineage relationship of cells. Although challenges remain

for large-scale single-cell WGS/WES in terms of experimental design complexity and sequencing cost currently, single-cell RNA sequencing (scRNA) has been applied broadly to examine cell population dynamics and track the development of cell lineages. The preprocessing steps for scRNA data are relatively similar to the usual practice of WGS/WES calling. However, splicing-aware aligners, e.g., STAR [72] or GSNAP [73], are suggested for the read alignment. There are still not many callers designed specifically for single-cell data [74]. Trinity Cancer Transcriptome Analysis Toolkit (CTAT) is one caller with extended functionality for scRNA-seq SNV detection. SCIΦ is another tool that can perform jointly calling of mutations in individual cells followed by an estimation of the tumor phylogeny [75]. SSrGE [76] is an integrative workflow to connect genotype and phenotype in single-cell data which implemented GATK best practice and FreeBayes for variant inference. A few other studies used SAMtools mpileup approach for variant identification [77, 78]. Solid variant calling strategies in single-cell data will be of great needs in the following years.

References

1. Hampel H, Bennett RL, Buchanan A, Pearlman R, Wiesner GL, Guideline Development Group, American College of Medical Genetics and Genomics Professional Practice and Guidelines Committee and National Society of Genetic Counselors Practice Guidelines Committee. A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: referral indications for cancer predisposition assessment. Genet Med. 2015;17(1):70–87.
2. Velazquez C, Lastra E, Avila Cobos F, Abella L, de la Cruz V, Hernando BA, Hernandez L, Martinez N, Infante M, Duran M. A comprehensive custom panel evaluation for routine hereditary cancer testing: improving the yield of germline mutation detection. J Transl Med. 2020;18(1):232.
3. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12(6):443–51.
4. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843–51.
5. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
6. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
7. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
8. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018;2018:201178.
9. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907. 2012.
10. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12(10):966–8.
11. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:17875.
12. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–7.
13. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213–9.
14. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28(3):311–7.
15. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851–8.
16. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76.
17. Fan Y, Xi L, Hughes DS, Zhang J, Zhang J, Futreal PA, Wheeler DA, Wang W. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol. 2016;17(1):178.
18. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Kallberg M, Chen X, Kim Y, Beyter D,

Krusche P, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15(8):591–4.
19. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811–7.
20. Miller NA, Farrow EG, Gibson M, Willig LK, Twist G, Yoo B, Marrs T, Corder S, Krivohlavek L, Walter A, et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. Genome Med. 2015;7:100.
21. Andrews S. FastQC: a quality control tool for high throughput sequence data [Online]. Available online at: http://www.bioinformaticsbabrahamacuk/projects/fastqc/ 2010.
22. Pan B, Kusko R, Xiao W, Zheng Y, Liu Z, Xiao C, Sakkiah S, Guo W, Gong P, Zhang C, et al. Similarities and differences between variants called with human reference genome HG19 or HG38. BMC Bioinformatics. 2019;20(Suppl 2):101.
23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
24. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, Mittelman D. An analytical framework for optimizing variant discovery from personal genomes. Nat Commun. 2015;6:6275.
25. Thankaswamy-Kosalai S, Sen P, Nookaew I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics. 2017;109(3–4):186–91.
26. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.
27. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J. 2018;16:15–24.
28. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014;15:244.
29. Anzar I, Sverchkova A, Stratford R, Clancy T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. BMC Med Genet. 2019;12(1):63.
30. Shin HT, Choi YL, Yun JW, Kim NKD, Kim SY, Jeon HJ, Nam JY, Lee C, Ryu D, Kim SC, et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. Nat Commun. 2017;8(1):1377.
31. Huang W, Guo YA, Muthukumar K, Baruah P, Chang MM, Jacobsen Skanderup A. SMuRF: portable and accurate ensemble prediction of somatic mutations. Bioinformatics. 2019;35(17):3157–9.
32. Wang M, Luo W, Jones K, Bian X, Williams R, Higson H, Wu D, Hicks B, Yeager M, Zhu B. SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. Sci Rep. 2020;10(1):12898.
33. Ding J, Bashashati A, Roth A, Oloumi A, Tse K, Zeng T, Haffari G, Hirst M, Marra MA, Condon A, et al. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. Bioinformatics. 2012;28(2):167–75.
34. Fang LT, Afshar PT, Chhibber A, Mohiyuddin M, Fan Y, Mu JC, Gibeling G, Barr S, Asadi NB, Gerstein MB, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biol. 2015;16:197.
35. Wood DE, White JR, Georgiadis A, Van Emburgh B, Parpart-Li S, Mitchell J, Anagnostou V, Niknafs N, Karchin R, Papp E, et al. A machine learning approach for somatic mutation discovery. Sci Transl Med. 2018;10(457):eaar7939.
36. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91(2):355–8.
37. Kalatskaya I, Trinh QM, Spears M, McPherson JD, Bartlett JMS, Stein L. ISOWN: accurate somatic mutation identification in the absence of normal tissue controls. Genome Med. 2017;9(1):59.
38. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
39. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
40. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85.
41. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA, et al. Germline mutations in predisposition genes in pediatric cancer. N Engl J Med. 2015;373(24):2336–46.
42. Auer PL, Reiner AP, Wang G, Kang HM, Abecasis GR, Altshuler D, Bamshad MJ, Nickerson DA, Tracy RP, Rich SS, et al. Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI exome sequencing project. Am J Hum Genet. 2016;99(4):791–801.
43. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
44. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.
45. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. Analysis of protein-

coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
46. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
47. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42(22):13534–44.
48. Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am J Hum Genet. 2017;100(2):267–80.
49. Edmonson MN, Patel AN, Hedges DJ, Wang Z, Rampersaud E, Kesserwan CA, Zhou X, Liu Y, Newman S, Rusch MC, et al. Pediatric Cancer Variant Pathogenicity Information Exchange (PeCanPIE): a cloud-based platform for curating and classifying germline variants. Genome Res. 2019;29(9):1555–65.
50. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–73.
51. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
52. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
53. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 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 (Austin). 2012;6(2):80–92.
54. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025.
55. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012;40(Web Server issue):W452–7.
56. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–94.
57. Guo Y, Li J, Li CI, Long J, Samuels DC, Shyr Y. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics. 2012;13:666.
58. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. http://www.repeatmaskerorg 2013–2015.
59. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51.
60. Dou Y, Kwon M, Rodin RE, Cortes-Ciriano I, Doan R, Luquette LJ, Galor A, Bohrson C, Walsh CA, Park PJ. Accurate detection of mosaic variants in sequencing data without matched controls. Nat Biotechnol. 2020;38(3):314–9.
61. Sendorek DH, Caloian C, Ellrott K, Bare JC, Yamaguchi TN, Ewing AD, Houlahan KE, Norman TC, Margolin AA, Stuart JM, et al. Germline contamination and leakage in whole genome somatic single nucleotide variant detection. BMC Bioinformatics. 2018;19(1):28.
62. Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, Yi C, Kreuziger J, Baldi P, Wallace DC. An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res. 2007;35(Database issue):D823–8.
63. Payne BA, Wilson IJ, Hateley CA, Horvath R, Santibanez-Koref M, Samuels DC, Price DA, Chinnery PF. Mitochondrial aging is accelerated by anti-retroviral therapy through the clonal expansion of mtDNA mutations. Nat Genet. 2011;43(8):806–10.
64. Santibanez-Koref M, Griffin H, Turnbull DM, Chinnery PF, Herbert M, Hudson G. Assessing mitochondrial heteroplasmy using next generation sequencing: a note of caution. Mitochondrion. 2019;46:302–6.
65. Ding J, Sidore C, Butler TJ, Wing MK, Qian Y, Meirelles O, Busonero F, Tsoi LC, Maschio A, Angius A, et al. Assessing mitochondrial DNA variation and copy number in lymphocytes of ~2,000 Sardinians using tailored sequencing analysis tools. PLoS Genet. 2015;11(7):e1005306.
66. Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. Nat Commun. 2019;10(1):4660.
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
68. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461–8.
69. Ebler J, Haukness M, Pesout T, Marschall T, Paten B. Haplotype-aware diplotyping from noisy long reads. Genome Biol. 2019;20(1):116.
70. Luo R, Sedlazeck FJ, Lam TW, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun. 2019;10(1):998.
71. Ramachandran A, Lumetta SS, Klee E, Chen D. HELLO: a hybrid variant calling approach. bioRxiv. 2003;2020(2020):2023.004473.
72. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

73. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. Methods Mol Biol. 2016;1418:283–334.
74. Liu F, Zhang Y, Zhang L, Li Z, Fang Q, Gao R, Zhang Z. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. Genome Biol. 2019;20(1):242.
75. Singer J, Kuipers J, Jahn K, Beerenwinkel N. Single-cell mutation identification via phylogenetic inference. Nat Commun. 2018;9(1):5144.
76. Poirion O, Zhu X, Ching T, Garmire LX. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. Nat Commun. 2018;9(1):4892.
77. Rodriguez-Meira A, Buck G, Clark SA, Povinelli BJ, Alcolea V, Louka E, McGowan S, Hamblin A, Sousos N, Barkas N, et al. Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. Mol Cell. 2019;73(6):1292–305, e1298.
78. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016;539(7628):309–13.
79. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics. 2012;28(7):907–13.
80. Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, Beerenwinkel N. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nat Commun. 2012;3:811.
81. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40(22):11189–201.
82. Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Res. 2013;41(7):e89.
83. Hansen NF, Gartner JJ, Mei L, Samuels Y, Mullikin JC. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics. 2013;29(12):1498–503.
84. Christoforides A, Carpten JD, Weiss GJ, Demeure MJ, Von Hoff DD, Craig DW. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. BMC Genomics. 2013;14:302.
85. Kim S, Jeong K, Bhutani K, Lee J, Patel A, Scott E, Nam H, Lee H, Gleeson JG, Bafna V. Virmid: accurate detection of somatic mutations with sample impurity inference. Genome Biol. 2013;14(8):R90.
86. Kassahn KS, Holmes O, Nones K, Patch AM, Miller DK, Christ AN, Harliwong I, Bruxner TJ, Xu Q, Anderson M, et al. Somatic point mutation calling in low cellularity tumors. PLoS One. 2013;8(11):e74380.
87. Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinformatics. 2014;15:104.
88. Wang W, Wang P, Xu F, Luo R, Wong MP, Lam TW, Wang J. FaSD-somatic: a fast and accurate somatic SNV detection algorithm for cancer genome sequencing data. Bioinformatics. 2014;30(17):2498–500.
89. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, WGS500 Consortium, Wilkie AOM, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46(8):912–8.
90. Usuyama N, Shiraishi Y, Sato Y, Kume H, Homma Y, Ogawa S, Miyano S, Imoto S. HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations. Bioinformatics. 2014;30(23):3302–9.
91. Radenbaugh AJ, Ma S, Ewing A, Stuart JM, Collisson EA, Zhu J, Haussler D. RADIA: RNA and DNA integrated analysis for somatic mutation detection. PLoS One. 2014;9(11):e111516.
92. Shi Y. SOAPsnv: an integrated tool for somatic single-nucleotide variants detection with or without normal tissues in cancer genome. J Clin Oncol. 2014;32(15_suppl):e22086.
93. Sengupta S, Gulukota K, Zhu Y, Ober C, Naughton K, Wentworth-Sheilds W, Ji Y. Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. Nucleic Acids Res. 2016;44(3):e25.
94. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11):e108.
95. Liu Y, Loewer M, Aluru S, Schmidt B. SNVSniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations. BMC Syst Biol. 2016;10(Suppl 2):47.
96. Spinella JF, Mehanna P, Vidal R, Saillour V, Cassart P, Richer C, Ouimet M, Healy J, Sinnett D. SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. BMC Genomics. 2016;17(1):912.
97. Jones D, Raine KM, Davies H, Tarpey PS, Butler AP, Teague JW, Nik-Zainal S, Campbell PJ. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. Curr Protoc Bioinformatics. 2016;56:15.

98. Carrot-Zhang J, Majewski J. LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples. Oncotarget. 2017;8(23):37032–40.
99. Sahraeian SME, Liu R, Lau B, Podesta K, Mohiyuddin M, Lam HYK. Deep convolutional neural networks for accurate somatic mutation detection. Nat Commun. 2019;10(1):1041.
100. Meng J, Victor B, He Z, Liu H, Jiang T. DeepSSV: detecting somatic small variants in paired tumor and normal sequencing data with convolutional neural network. Brief Bioinform. 2020;22(4):bbaa272.