Evaluation of whole genome sequencing data

Daniel Hübschmann, Matthias Schlesner

Angaben zur Veröffentlichung / Publication details:

Hübschmann, Daniel, and Matthias Schlesner. 2019. "Evaluation of whole genome sequencing data." In *Lymphoma: methods and protocols*, edited by Ralf Küppers, 321–36. New York, NY: Humana Press. https://doi.org/10.1007/978-1-4939-9151-8_15.

Evaluation of Whole Genome Sequencing Data

Daniel Hübschmann and Matthias Schlesner

Abstract

Whole genome sequencing (WGS) can provide comprehensive insights into the genetic makeup of lymphomas. Here we describe a selection of methods for the analysis of WGS data, including alignment, identification of different classes of genomic variants, the identification of driver mutations, and the identification of mutational signatures. We further outline design considerations for WGS studies and provide a variety of quality control measures to detect common quality problems in the data.

Key words Next-generation sequencing, Quality control, Variant calling, Mutational signatures

1 Introduction

Compared to traditional sequencing techniques, next-generation sequencing (NGS) offers a tremendous increase in throughput combined with a drastic decrease in cost per sequenced base. This makes it possible to perform whole genome sequencing (WGS) for large cohorts of samples and also to use it as a diagnostic tool in clinical settings.

Especially in cancer research, genome sequencing has been quickly adopted as genome alterations play a central role in malignant transformation [1]. Since the first publication of a cancer genome in 2008 [2], NGS formed the basis for substantial advancements in the understanding of the mechanisms underlying initiation and progression of several cancers including lymphomas.

Cancer genomes can either be sequenced in total (WGS), or sequencing is restricted to protein-coding regions (whole exome sequencing, WES) or selected genes or regions (panel sequencing). Only WGS covers the entire genome, including all exons, introns, and intergenic regions (but excluding highly repetitive parts which are not accessible with current short-read sequencing techniques), reveals all classes of alterations, and thus provides a comprehensive characterization of the cancer genome.

Four classes of variants can be identified from WGS data: single-nucleotide variants (SNVs), small insertions and deletions (indels), structural variants (SVs), and copy number aberrations (CNAs) [3]. In cancer sequencing studies, it is important to distinguish between germline variants, i.e., variants which the patient has inherited or which arose de novo in the very early stages of embryonic development and which are hence present in (almost) all cells of the patient, and somatic variants, which were acquired later during lifetime. Since a tumor stems from clonal expansion of a single cell, all variants which this cell has acquired before or in the early stages of tumor development appear as somatic variants in all cells of the tumor. Genomic variants which are acquired later during tumor evolution are only present in a subset of tumor cells and are commonly referred to as subclonal variants.

Although numerous computational tools dedicated to specific aspects of NGS data analysis have been developed in the past few years, most have project-specific features, and their functionality and parameterization are complicated. Furthermore, even when analyzing the same input data, the use of different processing pipelines results in different results [4]. In this chapter, we will introduce selected methods for processing and analysis of cancer WGS data. The current de facto standard for human genome sequencing experiments is paired-end sequencing on the Illumina platform. The analysis methods presented here assume short-read (100–150 bp) paired-end WGS data generated on Illumina sequencers. To enable reliable detection of genomic variants and to a certain extent also subclonal variants, the average depth of coverage should be at least 30×. The presented methods require availability of a Unix or Linux computing system with sufficient memory and storage (a BAM file of a human genome sequenced at 30× coverage is almost 100 GByte in size; the minimum RAM requirement of some of the processing steps is 50 GByte). It is recommended to perform WGS data analysis on a high-performance compute cluster or in a cloud environment to enable parallel processing and reduce analysis time. It should be noted that human WGS data cannot be de-identified without drastically reducing its utility and hence always bears the risk that the sample donor can be identified. Therefore, all infrastructure involved in storage, processing, and transfer of such data has to be adequately secured to minimize the risk of a privacy breach. The exact regulations differ between countries, and the patient's informed consent might impose additional restrictions on the allowed usage of the data. It is thus necessary to check for each individual scenario that the planned data generation, transfer, storage, and sharing are in agreement with all relevant regulations.

2 Methods

2.1 Experimental Design Considerations

2.1.1 Use of Matched Normal Controls

In addition to the tumor tissue, sequencing of matched normal tissue from the same patient is necessary to discriminate between germline and somatic variants (*see* **Note 1**). Any non-clonal tissue from the same patient can be used as matched normal tissue; due to good accessibility, most often white blood cells are chosen. Care has to be taken that the matched normal sample is not contaminated with tumor cells, which would lead to a misclassification of somatic variants as germline.

2.1.2 Whole Genome, Whole Exome, or Panel Sequencing

Only WGS enables the comprehensive identification of variants in the nonprotein-coding part of the genome. However, many other downstream analyses also benefit tremendously from the availability of broader data as delivered by WGS. These include the identification of structural variants (SVs) and copy number aberrations (CNAs) as well as the determination of tumor cell content and ploidy. Furthermore, WGS information results in a much higher power to identify genomic patterns including mutational signatures. The identification of mutational signatures and similar patterns might even have therapeutic implications, for example, in the case of "BRCAness" which indicates potential sensitivity of the tumor to PARP inhibitors [5]. A disadvantage of WGS is the higher cost (for data generation and for data analysis), especially if a high sequencing coverage is needed to detect variants with low allele frequencies (e.g., to analyze samples with low tumor cell content). In such settings exome sequencing or gene panel sequencing might be a more suitable choice.

2.1.3 Coverage

In the first large-scale cancer WGS projects as, for example, in the International Cancer Genome Consortium (ICGC) [6], the required minimum coverage was 30×. This depth should be considered as absolute minimum for cancer WGS analyses. If a study aims at deciphering the subclonal composition of a tumor or analyzes tumors with low tumor cell content (*see* below), considerably higher coverage might be required. It is discussed controversially whether the matched control tissue should be sequenced to the same coverage depth as the tumor or whether a fixed coverage of 30× in the matched normal control is sufficient even if the tumor is sequenced much deeper. In our experience, the analysis of samples with higher coverage in the tumor than in the matched normal requires additional filtering steps like filtering against a panel of normals to remove artifacts from the somatic variant calls [4].

2.1.4 Tumor Cell Content

Low tumor cell content of a sample causes reduced sensitivity in somatic variant identification [4] and especially impairs the identification of subclonal variants. The ICGC requires a tumor cell

content of at least 60%. It should be noted that the tumor cell content as determined from histological analysis is often considerably higher than the tumor cell content as determined from sequencing data analysis, so that samples which appear to be "still good" in histopathology might result in hardly usable data. To a certain extent, a lower tumor cell content can be compensated by higher sequencing depth and the use of improved bioinformatic methods to detect variants with low mutant allele fraction (MAF), but the discrimination between sequencing and library errors and true variants becomes more challenging.

2.1.5 Sample Requirements

WGS should be done on DNA extracted from fresh frozen tissue. Blood samples, for example, for use as matched normal, should be collected in EDTA-coated tubes; heparin-containing tubes should not be used because heparin inhibits PCR reactions and hence interferes with sequencing library preparation. Sequencing of DNA from FFPE tissue can, with some limitations, be used for small variant and CNA identification [7], but the identification of SVs is usually not possible. The depth of coverage in FFPE-derived data often shows strong fluctuations, and considerable parts of the genome might not be sufficiently covered for reliable variant calling. In many cases FFPE-induced DNA damage causes drastically increased sequencing error rates, and hence additional filtering steps are needed to reduce the number of false-positive variant calls.

2.2 Alignment

2.2.1 Choice of the Reference Genome

Next-generation sequencers produce millions of short sequence reads, which are stored in FASTQ files. The first step of WGS data analysis is to align the sequence reads to a reference genome. Two builds of the human reference genome are concurrently used: GRCh37 (hg19) and GRCh38. For both builds several different versions exist, which differ in the patches applied by the Genome Reference Consortium but also in the inclusion (or exclusion) of additional sequences like the "unlocalized sequences," "unplaced sequences," "decoy sequences," and in the case of GRCh38 also "alternate loci" (ALT contigs). For GRCh37, the reference sequence of the 1000 Genomes Project Phase II, called hs37d5 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence), is commonly used. This version differs from the standard build in the sequence of mitochondrial DNA, in masking of the PAR regions in chromosome Y and in the inclusion of decoy sequences, which reduce artifacts in variant calling. For GRCh38, the inclusion of ALT contigs, which represent common variation (e.g., the HLA loci) and multi-placed regions, makes this step of the analysis more challenging. Processing with standard pipelines which are not adapted to handle these new features properly will lead to a loss of variant calls in all regions affected by these issues. Many recent aligners (e.g., the current version of BWA-MEM) are "ALT-

aware" and can hence be used with a reference sequence containing ALT contigs, as, for example, included in the current GATK bundle (https://software.broadinstitute.org/gatk/download/bundle). For other pipelines, versions without ALT contigs have to be used (e.g., ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/README_analysis_sets.txt).

2.2.2 Preprocessing

Most sequencing facilities remove sequencing barcodes already during generation of the FASTQ files; otherwise they should be removed in the data analysis prior to alignment, for example, using TRIMMOMATIC [8]. Trimming of low-quality parts of the sequencing reads improves alignment quality if methods which require reads to be mapped in full length are used (e.g., BWA-backtrack [9]), but it is not required for methods based on local alignments like BWA-MEM [10].

2.2.3 Alignment Algorithms

While a plethora of algorithms to align short low-divergent sequences to a large reference genome has been developed, only few of them are commonly used for human WGS data, including BWA [9], Bowtie 2 [11], and GEM [12]. Several papers have reviewed the differences between them [13, 14], and we will not review these tools here but focus on the most commonly used algorithm, BWA. BWA is based on backward search with the Burrows-Wheeler transform [15], which enables high alignment accuracy while keeping a small memory footprint also for large genomes [9]. The first two algorithms of the BWA family, BWA-backtrack and BWA-SW [16], have later been complemented by a third algorithm BWA-MEM [10], which is suited to align sequences in the range from 70 bp to a few megabases against reference genomes. Hence BWA-MEM is usually the algorithm of choice when aligning WGS data generated on current Illumina instruments.

2.2.4 BAM File Processing

After the alignment, several postprocessing steps are required to generate BAM files for further analyses. It is recommended to combine several steps with Unix pipes instead of writing every output file to disk in order to reduce disk space requirements and enable fast processing. Several tools are available for these processing steps, including samtools [17], Picard [18], biobambam [19], and sambamba [20]. Essential postprocessing steps after alignment are coordinate sorting of the reads, merging of data (if a sample has been sequenced on more than one lane), and marking or removal of PCR duplicates (*see* **Note 2**). To reduce processing time, it is recommended to choose a tool that can perform merging and duplicate marking in one step (e.g., Picard or biobambam). In addition to these essential steps, other protocols [21, 22] recommend to perform base quality score recalibration and indel

realignment to improve BAM file quality, but in our setting these steps did not result in detectable improvements in the downstream analyses.

2.3 Variant Calling

In cancer sequencing projects, variant calling methods need to distinguish between germline and somatic variants. For somatic variant calling, it is crucial to use dedicated tools. Germline variant callers usually make assumptions about allele frequencies when they determine the most likely genotype as a measure to reduce the number of false-positive calls. For somatic variants, however, these assumptions cannot be made, because due to variable tumor cell content, aneuploidy, copy number aberrations, and the presence of different subclones, virtually all allele frequencies can be present.

2.3.1 Single-Nucleotide Variants (SNVs)

SNVs are the most abundant variant type in most cancer genomes. In general, SNV detection relies on single-nucleotide mismatches in the alignment after mapping the reads to the reference genome. Almost all SNV calling algorithms compute "pileups" (i.e., a section of all bases aligned to the respective position and their quality values through a stack of all reads overlapping this position), which are then used to determine the most likely genotype (germline variants) given the observed data or to determine the presence of a variant and its variant allele frequency (somatic variants). Prior to computation of the pileups, additional steps like local realignment or adjustment of quality scores might be performed.

Widely used tools for germline SNV calling include GATK HaplotypeCaller [23] and Platypus [24]. Dedicated somatic SNV callers are, for example, Mutect2 and our own pipeline based on samtools and a chain of empirical filters (https://dockstore.org/containers/quay.io/pancancer/pcawg-dkfz-workflow). FreeBayes [25] and Strelka2 [26] can be applied for both germline and somatic variant calling. A current review of somatic SNV calling algorithms is, for example, provided by [27].

2.3.2 Insertions and Deletions (Indels)

Like SNVs, small indels (usually up to 15–20 bp) can be directly detected from the alignment. Similar to SNV callers, many tools for small indel identification perform local realignment around indel candidate sites. For this reason, many recent tools combine SNV and small indel detection, like Mutect2 [28], Strelka2 [26], Platypus [24], and others. As described above, Mutect2 is designed to identify somatic variants, while Strelka2 and Platypus can be used for both germline and somatic variant calling. In our hands, Platypus works especially well for somatic indel calling if additional filtering rules are added which rescue variants that did not pass all Platypus internal filters to allow the detection of low allele frequency variants (https://dockstore.org/containers/quay.io/pancancer/pcawg-dkfz-workflow).

2.3.3 Structural Variants (SVs)

Structural variants comprise all types of genomic alterations except SNVs and small indels (*see* **Note 3**). This includes insertions, deletions, duplications, inversions, and translocations. Copy number aberrations (CNAs) are also a class of structural variants, but since different methods are used for CNA identification, CNAs are discussed separately in Subheading 2.3.4. Current tools for SV detection typically rely on one or several of the following signals: (a) discordant read pairs in the mapping of paired-end data, (b) split-read mapping, and (c) depth of coverage. Some tools as, for example, Manta [29], novoBreak [30], and SvABA [31] further employ assembly-based sequence reconstruction after initial candidate identification. Other popular tools for SV detection include DELLY [32] and LUMPY [33]. As the abovementioned signals (a–c) are independent from the assignment of tumor-normal pairs, these tools can be used for both germline and somatic variant calling.

2.3.4 Copy Number Aberrations (CNAs)

The detection of CNAs and allelic imbalances relies on changes in the depth of coverage and the B-allele frequencies of heterozygous SNPs. Before WGS became widely available, analogous information (with slightly lower resolution) could be derived from high-density SNP arrays, and indeed several CNA callers for WGS data are based on methods which have initially been developed for SNP array analysis. Most tools employ change-point detection methods to partition the genome into segments of equal coverage and allelic balance.

The depth of coverage in WGS data is often affected by various biases. The most common bias is GC bias, i.e., the dependence of the coverage on the GC content of the respective genomic window [34]. Another coverage bias, which is particularly prominent in fast-replicating tumor cells (e.g., Burkitt lymphomas), is replication timing bias [35, 36]. Samples affected by this bias have a higher coverage in early-replicating regions of the genome than in late-replicating regions. Many CNA callers correct for GC bias [36–38] and some also for replication timing bias [36] prior to genome segmentation to prevent the erroneous introduction of segment borders and thus over-segmentation. The inclusion of previously determined breakpoints (from SV calls) into the segmentation can additionally improve genome segmentation [36]. The segment borders indicate (relative) changes in the total or allele-specific copy numbers, but the absolute copy numbers of the segments cannot be directly inferred from depth of coverage and B-allele frequencies. Absolute allele-specific copy numbers can only be calculated if the tumor ploidy and tumor cell content (TCC) are known. Since this information is usually not known a priori, tools to determine absolute copy numbers are equipped with methods to estimate TCC and ploidy from the WGS data [36, 39, 40] (*see* **Note 4**).

2.4 Quality Control

At various stages of the analysis, thorough quality control (QC) should be performed.

The first QC can already be done at the level of FASTQ files so that samples with severe quality problems are not even processed further. A popular tool for this is FastQC [41]. Issues detected at this stage include problems originating from sequencing but also certain problems originating from library preparation like adapter contamination. The FastQC homepage [41] contains extensive documentation and also example reports for good and bad data.

Next, several parameters which are acquired during BAM file postprocessing and at BAM file level (e.g., through samtools flagstat) should be assessed. These include:

- The duplication rate. For WGS libraries the duplication rate is usually below 15%. A high duplication rate indicates that the library complexity is rather low. However, as long as the desired coverage (without counting duplicates) is still reached, downstream analyses are usually not affected.
- Median insert size and insert size distribution. The median insert size should be considerably larger than two times the read length for paired-end sequencing, as otherwise for many fragments the middle part would be sequenced twice. Such overlap between the mates can result in artifacts during variant calling (if the used tools do not handle the overlap adequately) and reduces the effective coverage. The insert size distribution should have only one mode and no long tail to either side. Libraries with bimodal insert size distribution can lead to drastically increased false-positive SV calls when choosing SV callers which use the paired-end insert size as detection criterion.
- Read pairs mapping to different chromosomes. Usually this value is below 5%; higher values indicate problems originating from library preparation. Samples with slightly increased values are usually not problematic for further analyses, but samples with drastically increased values (>15–20%) can cause errors especially during SV calling. Note that even in highly rearranged tumor genomes and in genomes with chromothripsis, this parameter is only slightly elevated.
- It should be verified that the required coverage has been reached. To determine the usable coverage, PCR duplicates should not be counted. Especially in libraries with small insert size (considerably smaller than two times the read length), also overlapping parts of both mates should only be counted once to avoid overestimation of the effectively usable coverage.

Finally, different variant calling methods themselves provide valuable QC information:

- Number and fraction of somatic SNVs which match an entry in dbSNP [42]. A very high number (greater than one million) of somatic SNVs in dbSNP indicate that tumor and matched normal are not from the same genetic background. This context is even more likely if the peak of the distribution of MAF values is close to 0.5. Such a situation can either be due to a sample swap or when the patient has received an allogenic stem cell transplant. An increased number and fraction of somatic SNVs in dbSNP with a low mutant allele frequency can indicate contamination of the tumor sample with DNA from another individual.
- Fraction of synonymous SNVs among all SNVs in coding regions. This fraction is increased when the tumor sample is contaminated with DNA from another species (e.g., mouse) as a result of cross contamination or if a xenograft is sequenced. Good samples usually have a synonymous fraction <0.35, while contaminated samples have often >0.5. Samples with values between 0.35 and 0.5 should be carefully checked. Samples which are contaminated with DNA from another species can be rescued by alignment against a combined reference genome containing both the human genome sequence and the genome of the respective species.
- Number of "intron deletions." A high number of deletions with breakpoints exactly at exon boundaries can indicate contamination with RNA. In addition, RNA-contaminated samples show an increased fraction of coding SNVs among all SNVs.
- Tumor cell content. Most CNA calling algorithms provide estimates for the tumor cell content, and it should be checked that the tumor cell content is high enough to enable somatic variant calling with adequate sensitivity.
- GC bias. Many sequencing libraries show a coverage bias which depends on the GC content of the sequenced fragment. CNA callers like ACEseq [36] correct the GC bias prior to copy number estimation to prevent false-positive copy number variant calls and can provide a quantitative estimate of the GC bias. Note that while GC bias can be corrected for CNA calling, strong GC bias will still negatively affect the identification of other types of variants due to the reduced coverage in parts of the genome.
- Other coverage fluctuations. Various problems during library preparation or with the template DNA can result in GC-independent coverage fluctuations. Such coverage fluctuations can lead to difficulties in the detection of CNAs and, like GC bias, will lead to reduced power for the detection of other variant types in genomic regions with low coverage.

2.5 Variant Annotation

To increase usability of variant information in biological analyses, the variants need to be annotated with functional information. This includes gene annotation to identify whether the variant affects, e.g., the protein-coding sequence of a gene; variant database information to disclose if a variant is, e.g., a known SNP or a known somatic cancer mutation; and potentially other information tracks, e.g., about sequence conservation or regulatory elements. Commonly used tools for variant annotation include ANNOVAR [43], SnpEff [44], variant effect predictor [45], and Rbbt [46]. Note that both the choice of the variant annotation software and the choice of the gene and transcript database can have a large impact on variant annotation and variant consequence prediction [47–50]. It is not possible to make a general recommendation which gene annotation should be used, but the choice should be made taking into consideration the aim of the study. For example, it should be evaluated whether a more comprehensive annotation is of higher importance or rather a more simple and reliable annotation. Furthermore, when it comes to the detection of high-impact variants, it may be beneficial to assess *a priori* the respective importance of sensitivity and specificity.

2.6 Identification of Driver Mutations

Driver genes are genes whose deregulation confers a selective advantage for the tumor. Mutations which cause such deregulation are called driver mutations. In a typical tumor genome, the vast majority of mutations are passenger mutations, i.e., mutations which have not been selected and have not conferred a clonal growth advantage. To understand the mechanisms of tumor development, it is important to delineate the driver mutations in a tumor genome. Methods to identify driver mutations search for signals of selection. This can be a higher mutation rate in a gene than expected by chance (e.g., MuSiC [51] or MutsigCV [52]), a bias toward high functional impact of mutations (e.g., Oncodrive-fm [53]), or clustering of mutations in certain parts of a protein (OncodriveCLUST [54]). Recent approaches have generalized these techniques to also enable the detection of driver mutations in noncoding regions of the genome; examples include LARVA [55] and OncodriveFML [56].

2.7 Mutational Signature Analysis

The set of mutations in a cancer genome (including both driver and passenger mutations) is the imprint of the activity of multiple mutational processes. Several mutational processes have specific preferences with respect to the caused mutations. For SNVs, this means that a specific mutational process has a certain probability to introduce each of the possible transitions or transversions. Furthermore, this specificity does not only result in different nucleotide exchanges but also extends over the nucleotides flanking the mutated position. Each mutational process thus leaves a certain footprint in the genome, which is referred to as a mutational

signature [57]. An analysis of mutational signatures can therefore provide insights into the mutational processes that have been active during the life history of a tumor and its precursor cell.

2.7.1 Unsupervised Analysis of Mutational Signatures

A mutational catalogue contains the frequencies of SNVs in their trinucleotide context per sample. It is possible to run a de novo extraction of mutational signatures from the mutational catalogue of a sufficiently large number of cancer genomes. To this end, the mutational catalogue is decomposed in a mutational signature matrix and an exposure matrix (which contains the activity of each identified signature in each genome) using nonnegative matrix factorization (NMF). In a large-scale analysis of more than 500 WGS and more than 6,500 exome samples from 30 different tumor entities, Alexandrov et al. identified 21 mutational signatures, of which around one-half could be associated to known mutational processes [57]. Later this analysis has been extended to more than 1,000 WGS samples and more than 10,000 exomes from 40 entities, and 30 validated mutational signatures have been identified. This set of signatures is available from the COSMIC database [58].

Different frameworks are available to perform unsupervised analysis of mutational signatures. The original framework by Alexandrov et al. [57] is implemented in MATLAB. As alternative, an R package for mutational signature analysis is available from the Bioconductor Project [59]. Finally, the R package Bratwurst [60] can be used for different types of NMF analyses, including unsupervised mutational signature analysis. It provides wrapper functions for NMF solvers on graphical processing units (GPUs) using the Compute Unified Device Architecture (CUDA) 8 framework and the CUDAMat library [61] and therefore enables much faster NMF calculations due to massive parallelization.

2.7.2 Supervised Analysis of Mutational Signatures

Unsupervised mutational signature analysis enables the detection of novel signatures. However, unsupervised signature analysis requires the availability of large cohorts of samples. Especially for smaller studies, a supervised analysis of mutational signatures is therefore the better option. In a supervised analysis, the contributions of known mutational signatures to a given mutational catalogue are determined. This approach requires much less statistical power and can hence be applied to small cohorts or even to single cases. Supervised mutational signature analysis can be performed, e.g., with the R packages deconstructSigs [62] or YAPSA [63].

3 Notes

1. Analysis of samples without matched control may be performed if databases (e.g., dbSNP [42] or more recent resources

like ExAC or gnomAD [64]) are used to remove common variants. However, with such an approach, a complete removal of germline variants from the somatic set is not possible since every individual has private germline variants. It should be noted that certain databases like dbSNP do also contain well-studied somatic variants, and filtering against these databases without additional selection criteria should thus be avoided. In case of deep sequencing (usually >100×) of tumors of relatively low tumor cell content (<70%), the mutant allele fraction can be employed for a better discrimination between germline and somatic variants. Here the allele-specific copy numbers of the genomic segment where the variant is located have to be taken into account, and it is recommended to filter against a panel of independent normal control samples processed with the same workflows as the tumor samples to remove pipeline-specific artifacts. Some dedicated tools for the analysis of tumor samples without matched controls integrate various sources of information to predict somatic mutations [65–67].

2. Although current analyses typically make no use of duplicate reads, we recommend to just mark them instead of removing them from the BAM file to keep the full information in the BAM file. With this strategy, the FASTQ files can be reconstructed from the BAM files and might be discarded to save disk space. If read trimming algorithms were applied before alignment, however, loss of information might occur, and FASTQ files should be kept regardless of whether duplicates were only marked or removed.
3. Many SV callers have a reduced sensitivity for the identification of SVs of short length (20–300 bp) [29, 31], and hence there might be a gap between the events reported by small indel callers and SV callers. If SVs in this size range are in the focus of the analysis, a careful choice of tools and possibly the use of multiple tools is recommended.
4. The determination of TCC and ploidy from WGS data is nontrivial, and often multiple combinations of TCC and ploidy lead to good fits to the data. Furthermore, all existing methods make assumptions about the tumor sample (e.g., that the majority of aberrations is present in all tumor cells or that there is only a low number of subclones in the sample) which are not necessarily true for the analyzed samples. Certain conditions like the presence of tumor cells of different ploidy (i.e., diploid and tetraploid tumor cells after genome duplication as late event in tumor evolution) are to our experience not handled well by any existing tool. It might therefore be necessary to employ complementary techniques like FISH or karyotyping to determine the tumor cell ploidy and reliably estimate absolute copy numbers.

Acknowledgments

This work has been supported by the German Ministry of Science and Education (BMBF) in the framework of the ICGC MMML-Seq (01KU1002A-J) and the ICGC DE-MINING (01KU1505E) projects and the Heidelberg Center for Human Bioinformatics (HD-HuB) within the German Network for Bioinformatics Infrastructure (de.NBI) (#031A537A, #031A537C). We are grateful to all present and previous members of the Division of Theoretical Bioinformatics, the DKFZ-HIPO bioinformatics team, the Omics IT and Data Management Core Facility, and the Bioinformatics and Omics Data Analytics group of the German Cancer Research Center (DKFZ, Heidelberg) as well as coworkers in the ICGC MMML-seq and PedBrain projects who were involved in the establishment of the procedures described here.

References

1. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458:719–724. https://doi.org/10.1038/nature07943
2. Ley TJ, Mardis ER, Ding L et al (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456:66–72. https://doi.org/10.1038/nature07485
3. Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11:685–696
4. Alioto TS, Buchhalter I, Derdak S et al (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun 6:10001. https://doi.org/10.1038/ncomms10001
5. Davies H, Glodzik D, Morganella S et al (2017) HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med 23:517–525. https://doi.org/10.1038/nm.4292
6. Hudson TJ, Anderson W, Aretz A et al (2010) International network of cancer genome projects. Nature 464:993–998. https://doi.org/10.1038/nature08987
7. Robbe P, Popitsch N, Knight SJL et al (2018) Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. Genet Med 20(10):1196–1205. https://doi.org/10.1038/gim.2017.241
8. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
9. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324
10. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://www.arxiv.org/abs/1303.3997
11. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
12. Marco-Sola S, Sammeth M, Guigó R, Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 9:1185–1188. https://doi.org/10.1038/nmeth.2221
13. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–483
14. Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36–46. https://doi.org/10.1038/nrg3117
15. Lippert RA (2005) Space-efficient whole genome comparisons with burrows–wheeler transforms. J Comput Biol 12:407–415. https://doi.org/10.1089/cmb.2005.12.407
16. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595. https://doi.org/10.1093/bioinformatics/btp698

17. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
18. BroadInstitute (2016) Picard Tools—By Broad Institute. http://broadinstitute.github.io/picard/. Accessed 6 May 2018
19. Tischler G, Leonard S (2014) Biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol Med 9:13
20. Tarasov A, Vilella AJ, Cuppen E et al (2015) Sambamba: fast processing of NGS alignment formats. Bioinformatics 31:2032–2034. https://doi.org/10.1093/bioinformatics/btv098
21. Van der Auwera GA, Carneiro MO, Hartl C et al (2013) From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10.1–11.1033. https://doi.org/10.1002/0471250953.bi1110s43
22. Depristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–501. https://doi.org/10.1038/ng.806
23. Poplin R, Ruano-Rubio V, DePristo MA, et al (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 201178. https://doi.org/10.1101/201178
24. Rimmer A, Phan H, Mathieson I et al (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet 46:912–918. https://doi.org/10.1038/ng.3036
25. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907. https://arxiv.org/abs/1207.3907
26. Kim S, Scheffler K, Halpern AL, et al (2017) Strelka2: Fast and accurate variant calling for clinical sequencing applications. bioRxiv 192872. https://doi.org/10.1101/192872
27. Xu C (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J 16:15–24
28. Cibulskis K, Lawrence MS, Carter SL et al (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31:213–219. https://doi.org/10.1038/nbt.2514
29. Chen X, Schulz-Trieglaff O, Shaw R et al (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220–1222. https://doi.org/10.1093/bioinformatics/btv710
30. Chong Z, Ruan J, Gao M et al (2017) novoBreak: local assembly for breakpoint detection in cancer genomes. Nat Methods 14:65–67. https://doi.org/10.1038/nmeth.4084
31. Wala JA, Bandopadhayay P, Greenwald NF et al (2018) SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res 28:581–591. https://doi.org/10.1101/gr.221028.117
32. Rausch T, Zichner T, Schlattl A et al (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339. https://doi.org/10.1093/bioinformatics/bts378
33. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15:R84. https://doi.org/10.1186/gb-2014-15-6-r84
34. Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40:e72. https://doi.org/10.1093/nar/gks001
35. Koren A, Handsaker RE, Kamitaki N et al (2014) Genetic variation in human DNA replication timing. Cell 159:1015–1026. https://doi.org/10.1016/j.cell.2014.10.025
36. Kleinheinz K, Bludau I, Huebschmann D, et al (2017) ACEseq—allele specific copy number estimation from whole genome sequencing. bioRxiv 210807. https://doi.org/10.1101/210807
37. Boeva V, Popova T, Bleakley K et al (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28:423–425. https://doi.org/10.1093/bioinformatics/btr670
38. Favero F, Joshi T, Marquard AM et al (2015) Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol 26:64–70. https://doi.org/10.1093/annonc/mdu479
39. Van Loo P, Nordgard SH, Lingjærde OC et al (2010) Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A 107:16910–16915. https://doi.org/10.1073/pnas.1009843107
40. Carter SL, Cibulskis K, Helman E et al (2012) Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 30:413–421. https://doi.org/10.1038/nbt.2203

41. Simon A (2010) FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
42. Sherry ST, Ward MH, Kholodov M et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311
43. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164–e164. https://doi.org/10.1093/nar/gkq603
44. Cingolani P, Platts A, Wang LL et al (2012) 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) 6:80–92. https://doi.org/10.4161/fly.19695
45. McLaren W, Gil L, Hunt SE et al (2016) The Ensembl variant effect predictor. Genome Biol 17:122. https://doi.org/10.1186/s13059-016-0974-4
46. Vazquez M, Nogales R, Carmona P et al (2010) Rbbt: a framework for fast bioinformatics development with ruby. Springer, Berlin, Heidelberg
47. McCarthy DJ, Humburg P, Kanapin A et al (2014) Choice of transcripts and software has a large effect on variant annotation. Genome Med 6:26. https://doi.org/10.1186/gm543
48. Frankish A, Uszczynska B, Ritchie GR et al (2015) Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics 16:S2. https://doi.org/10.1186/1471-2164-16-S8-S2
49. Wu PY, Phan JH, Wang MD (2013) Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinformatics 14(Suppl 1):S8. https://doi.org/10.1186/1471-2105-14-S11-S8
50. Zhao S, Zhang B (2015) A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16:97. https://doi.org/10.1186/s12864-015-1308-8
51. Dees ND, Zhang Q, Kandoth C et al (2012) MuSiC: Identifying mutational significance in cancer genomes. Genome Res 22:1589–1598. https://doi.org/10.1101/gr.134635.111
52. Lawrence MS, Stojanov P, Polak P et al (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499:214–218. https://doi.org/10.1038/nature12213
53. Gonzalez-Perez A, Lopez-Bigas N (2012) Functional impact bias reveals cancer drivers. Nucleic Acids Res 40:e169. https://doi.org/10.1093/nar/gks743
54. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 29:2238–2244. https://doi.org/10.1093/bioinformatics/btt395
55. Lochovsky L, Zhang J, Fu Y et al (2015) LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res 43:8123–8134. https://doi.org/10.1093/nar/gkv803
56. Mularoni L, Sabarinathan R, Deu-Pons J et al (2016) OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol 17:128. https://doi.org/10.1186/s13059-016-0994-0
57. Alexandrov LB, Nik-Zainal S, Wedge DC et al (2013) Signatures of mutational processes in human cancer. Nature. https://doi.org/10.1038/nature12477
58. COSMIC—signatures of mutational processes in human cancer. https://cancer.sanger.ac.uk/cosmic/signatures. Accessed 9 May 2018
59. Gehring JS, Fischer B, Lawrence M, Huber W (2015) SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics 31:3673–3675. https://doi.org/10.1093/bioinformatics/btv408
60. Huebschmann D, Kurzawa N, Steinhauser S, et al (2017) Deciphering programs of transcriptional regulation by combined deconvolution of multiple omics layers. bioRxiv 199547. https://doi.org/10.1101/199547
61. Mnih V (2009) CUDAMat: a CUDA-based matrix class for Python. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.232.4776&rep=rep1&type=pdf
62. Rosenthal R, McGranahan N, Herrero J et al (2016) deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol 17:31. https://doi.org/10.1186/s13059-016-0893-4
63. Huebschmann D, Gu Z, Schlesner M (2015) YAPSA: yet another package for signature analysis. R package. http://bioconductor.org/packages/release/bioc/html/YAPSA.html
64. Lek M, Karczewski KJ, Minikel EV et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291. https://doi.org/10.1038/nature19057

65. Kalatskaya I, Trinh QM, Spears M et al (2017) ISOWN: accurate somatic mutation identification in the absence of normal tissue controls. Genome Med 9:59. https://doi.org/10.1186/s13073-017-0446-9
66. Smith KS, Yadav VK, Pei S et al (2016) SomVarIUS: somatic variant identification from unpaired tissue samples. Bioinformatics 32:808–813. https://doi.org/10.1093/bioinformatics/btv685
67. Madubata CJ, Roshan-Ghias A, Chu T et al (2017) Identification of potentially oncogenic alterations from tumor-only samples reveals Fanconi anemia pathway mutations in bladder carcinomas. NPJ Genomic Med 2:29. https://doi.org/10.1038/s41525-017-0032-5