



# Identification of Genetic Variants and de novo Mutations Based on NGS

# 10

Patricia Basurto-Lozada, Carolina Castañeda-Garcia, Raúl Ossio,  
and Carla Daniela Robles-Espinoza

## Contents

|        |   |     |
|--------|---|-----|
| 10.1   | Introduction: Quick Recap of a Sequencing Experiment Design .....                                       | 124 |
| 10.2   | How Are Novel Genetic Variants Identified? .....  | 124 |
| 10.2.1 | Naive Variant Calling .....   | 125 |
| 10.2.2 | Bayesian Variant Calling .....  | 126 |
| 10.2.3 | Heuristic Variant Calling .....   | 128 |
| 10.2.4 | Other Factors to Take into Account When Performing Variant Calling .....                                | 129 |
| 10.2.5 | How to Choose an Appropriate Algorithm for Variant Calling? .....                                       | 130 |
| 10.3   | Working with Variants .....   | 132 |
| 10.4   | Applying Post-variant Calling Filters .....   | 132 |
| 10.5   | <i>De novo</i> Genetic Variants: Population-Level Studies and Analyses Using Pedigree Information ..... | 133 |
| 10.6   | Filtering Genetic Variants to Identify Those Associated to Phenotypes .....                             | 134 |
| 10.6.1 | Variant Annotation .....  | 135 |
| 10.6.2 | Evaluating the Evidence Linking Variants Causally to Phenotypes .....                                   | 135 |

---

P. Basurto-Lozada · C. Castañeda-Garcia · R. Ossio  
Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional  
Autónoma de México, Campus Juriquilla, Santiago de Querétaro, México

C. D. Robles-Espinoza (✉)  
Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional  
Autónoma de México, Campus Juriquilla, Santiago de Querétaro, México

Experimental Cancer Genetics, Wellcome Sanger Institute, Cambridge, UK  
e-mail: [drobles@liigh.unam.mx](mailto:drobles@liigh.unam.mx)

---

|   |     |
|---|-----|
| 10.6.3 Variant Filtering and Visualization Programs ..... | 136 |
| 10.7 A Practical Example Workflow .....                   | 139 |
| References .....  | 141 |

### What You Will Learn in This Chapter

In this chapter, we will discuss an overview of the bioinformatic process for the identification of genetic variants and *de novo* mutations in data recovered from NGS applications. We will pinpoint critical steps, describe the theoretical basis of different variant calling algorithms, describe data formats, and review the different filtering criteria that can be undertaken to obtain a set of high-confidence mutations. We will also go over crucial issues to take into account when analyzing NGS data, such as tissue source or the choice of sequencing machine. We also discuss different methodologies for analyzing these variants depending on study context, considering population-wide and family-focused analyses. Finally, we also do an overview of available software for variant filtering and genetic data visualization.

---

## 10.1 Introduction: Quick Recap of a Sequencing Experiment Design

As we have seen throughout this book, NGS applications give researchers an all-access pass to the building information of all biological organisms. After establishing the biological question to be pursued and once the organism of interest has been sequenced, the first step is to align this information against a reference genome (see Chap. 8). This reference genome should be one that is as biologically close as it can be to the subject of interest—if there is no reference genome or the one available is not reliable, then a possible option is to attempt to build one (See Box: Genome assembly). After read mapping and alignment, and quality control, one or several variant callers will need to be run to identify the variants present in the query sequence. Finally, depending on the original aim, different post-processing and filtering steps may also need to be deployed to extract meaningful information out of the experiment.

---

## 10.2 How Are Novel Genetic Variants Identified?

The correct identification of variants depends on having accurately performed base calling, and read mapping and alignment previously. “Base calling” refers to the determination of the identity of a nucleotide from the fluorescence intensity information outputted by the sequencing instrument. Read mapping is the process of determining where a read originates from, using the reference genome, and read alignment is the process of finding the exact differences between the sequences. These topics have already been reviewed throughout this

book (Chaps. 4, 8 and 9). Base calling and read alignment results rely on the sequencing instrument and algorithm used; therefore, it is important to state the confidence we have on the assignment of each base. This is expressed by standard Phred quality scores [1].

$$Q_{\text{Phred}} = -10 \log_{10} P(\text{error})$$

This measurement is an intuitive number that tells us the probability of the base call or the alignment being wrong, and the higher  $Q$  is, the more confidence we have that there has not been an error. For example, if  $Q = 20$ , then that means there is a 1 in 100 chance of the call or alignment being wrong, whereas if it is 30 then there is a 1 in 1000 chance of a mistake. Subsequently, steps such as duplicate read marking and base call quality score recalibration can be performed (See Chap. 7, Sect. 7.2.3).

---

**Review Question 1**

What would be the value of  $Q$  for a variant that has a 1 in 3000 chance of being wrong?

After the previous steps have taken place, and an alignment file has been produced (usually in the BAM and CRAM file formats), the next step is to identify differences between the reference genome and the genome that has been sequenced. To this effect, there are different strategies that a researcher can use depending on their experiment, for example, for germline analyses they might use algorithms that assume that the organism of interest is diploid (or another, fixed ploidy) and for cancer genomes they may need to use more flexible programs due to the presence of polyploidy and aneuploidy. In this Chapter, we will focus on the former analyses, but the reader is referred to the publications on somatic variant callers in the “Further Reading” section below if they want to learn more.

When identifying variants, and particularly if a researcher is performing whole exome or genome sequencing, the main objective is to determine the genotype of the sample under study at each position of the genome. For each variant position there will be a reference ( $R$ ) and an alternate ( $A$ ) allele, the former refers to the sequence present in the reference genome. Therefore, it follows that in the case of diploid organisms, there will be three different possible genotypes:  $RR$  (homozygous reference),  $RA$  (heterozygous), and  $AA$  (homozygous alternative).

### 10.2.1 Naive Variant Calling

A naive approach to determining these genotypes from a pile of sequencing reads mapped to a site in the genome may be to count the number of reads with the reference and alternate alleles and to establish hard genotype thresholds; for example, if more than 75% of reads are  $R$ , then the genotype is called as  $RR$ ; if these are less than 25%, then the genotype is called as  $AA$ ; and anything in between is deemed  $RA$ . However, even if careful steps are taken to ensure that only high-quality bases and reads are counted in the calculation, this method is still prone to under-calling variants in low-coverage data, as the counts from



as the product of a prior genotype probability and the genotype likelihood, divided by a constant:

$$P(G|D) = \frac{P(D|G) P(G)}{P(D)}$$

where

- $P(G|D)$  is the posterior probability of the genotype given the sequencing data
- $P(D|G)$  is the genotype likelihood
- $P(G)$  is the prior genotype probability
- $P(D)$  is a factor to normalize the sum of all posterior probabilities to 1, it is constant throughout all possible genotypes

As this is the most commonly used method for variant calling, has been for a number of years and is unlikely to change, special attention should be given to it to understand its basics. For deciding what the genotype is at a particular site in an individual, a variant calling algorithm would calculate the posterior probability  $P(G|D)$  for each possible genotype, and pick the genotype with the highest one. As the sum of all posterior genotype probabilities must be equal to 1, the number of different possible genotypes that the algorithm (referred to as “genotype partitioning”) considers is crucial. Different algorithms will calculate these differently, for example, some algorithms may only consider three genotype classes: homozygous reference, heterozygous reference, and all others, whereas others may consider all possible genotypes [6]. The choice of algorithm would depend on the original biological question: For example, if tumors are being analyzed, which can be aneuploid or polyploid, an algorithm that only considers three possible genotypes may be inadequate.

The prior genotype probability,  $P(G)$ , can be calculated taking into account the results of previous sequencing projects. For example, if a researcher is sequencing humans, the prior probability of a genotype being found at a particular site could depend on previously reported allele frequencies at that site as well as the Hardy–Weinberg Equilibrium principle [2]. Information from linkage disequilibrium calculations can also be incorporated. Otherwise, if no information is available, then the prior probability of a variant occurring at a site may be set as a constant for all loci. Algorithms also differ in the way they calculate these priors, and the information they take into account. The denominator of the equation,  $P(D)$ , remains constant throughout all genotypes being considered and serves to normalize all posterior probabilities so they sum up to 1. Therefore, it is equal to the sum of all numerators,  $P(D) = \sum P(D|G_i) P(G_i)$ , where  $G_i$  is the  $i$ th genotype being considered.

The last part of the equation is the genotype likelihood,  $P(D|G)$ . This can be interpreted as the probability of obtaining the sequencing reads we have given a particular genotype. It can be calculated from the quality scores associated with each read at the site being considered, and then multiplying these across all existing reads, assuming all reads are independent [2]. For example, in what is perhaps the most commonly used variant calling

algorithm, GATK HaplotypeCaller [4], a number of steps are followed to determine genotype likelihoods: First, regions of the genome where there is evidence of a variant are defined (“SNP calling”), then the sequencing reads are used to identify the most likely genotypes supported by these data, and each read is re-aligned to all these most likely haplotypes in order to obtain the likelihoods per haplotype and per variant given the read data. These are then input into Bayes’ formula to identify the most likely genotype for a sample [4].

### 10.2.3 Heuristic Variant Calling

Other methods that do not rely on naive or Bayesian approaches have been developed; these methods rely on heuristic quantities to call a variant site, such as a minimum coverage and alignment quality thresholds, and stringent cut-offs like a minimum number of reads supporting a variant allele. If performing somatic variant calling, a statistical test such as a Fisher’s exact comparing the number of reference and alternate alleles in the tumor and normal samples is then performed in order to determine the genotype at a site [7]. Parameters used for variant calling can be tuned, and generally this method will work well with high-coverage sequencing data, but may not achieve an optimal equilibrium between high specificity and high sensitivity at low to medium sequencing depths, or when searching for low-frequency variants in a population [8].

The following GATK commands depict an example workflow for calling variants in NGS data. The installation instruction is covered in Chap. 5.

First you can check all available GATK tools by typing `gatk-list`. If not already done, you also have to install all required software packages (<http://gatkforums.broadinstitute.org/gatk/discussion/7098/howto-install-software-for-gatk-workshops>) for GATK analyses workflows. Moreover, be sure to set the PATH in your `.bashrc` to your GATK executable PATH.

The GATK (v4) uses two files to access and safety check access to the reference files: a `.dict` dictionary of the contig names and sizes and a `.fai` fasta index file to allow efficient random access to the reference bases. You have to generate these files in order to be able to use a Fasta file as reference.

```
#Create .fai fasta index file:
samtools faidx /path/to/GenomeIndices/GRCh38_index_100/GRCh38.
primary_assembly.genome.fa
#Create .dict dictionary of the contig names and sizes file:
gatk CreateSequenceDictionary R=GRCh38.primary_assembly.genome.fa
O=GRCh38.primary_assembly.genome.dict
#Add read groups, sort, mark duplicates, and create index
```

```
gatk AddOrReplaceReadGroups I=/path/to/Aligned.sortedByCoord.out.sam
O=sampleName_rg_added_sorted.bam SO=coordinate RGID=id RGLB=library
RGPL=platform RGPU=machine RGSM=sample
gatk MarkDuplicates I=sampleName_rg_added_sorted.bam O=dedup_sampleName.
bam M=metrics.txt
gatk BuildBamIndex I=dedup_sampleName.bam
```

Next, we use a GATK tool called `SplitNCigarReads` developed specially for RNAseq, which splits reads into exon segments (getting rid of Ns but maintaining grouping information) and hard-clip any sequences overhanging into the intronic regions.

```
#Split 'N' Trim and reassign mapping qualities
gatk SplitNCigarReads -R path/to/GenomeIndices/GRCh38_index_100/GRCh38.
primary_assembly.genome.fa -I dedup_sampleName.bam -o
split_dedup_sampleName.bam -rf ReassignOneMappingQuality -RMQF 255 -RMQT 60
-U ALLOW_N_CIGAR_READS
```

In this example we will use `Mutect2` to perform variant calling, which identifies somatic SNVs and indels via local assembly of haplotypes.

```
#Variant calling
gatk Mutect2 -R /path/to/GenomeIndices/GRCh38_index_100/GRCh38.
primary_assembly.genome.fa -I split_sampleName.bam -o
CalledVariants_sampleName.out.vcf
#Variant Filtration
gatk VariantFiltration -R /path/to/GenomeIndices/GRCh38_index_100/GRCh38.
primary_assembly.genome.fa -V CalledVariants_sampleName.out.vcf -window
35 -cluster 3 -filterName FS -filter "FS > 30.0" -filterName QD -filter "QD <
2.0" -o FilteredCalledVariants_sampleName.out.vcf
```

For a more detailed description see <https://github.com/gatk-workflows/gatk4-jupyter-notebook-tutorials/blob/master/notebooks/Day3-Somatic/1-somatic-mutect2-tutorial.ipynb>.

### 10.2.4 Other Factors to Take into Account When Performing Variant Calling

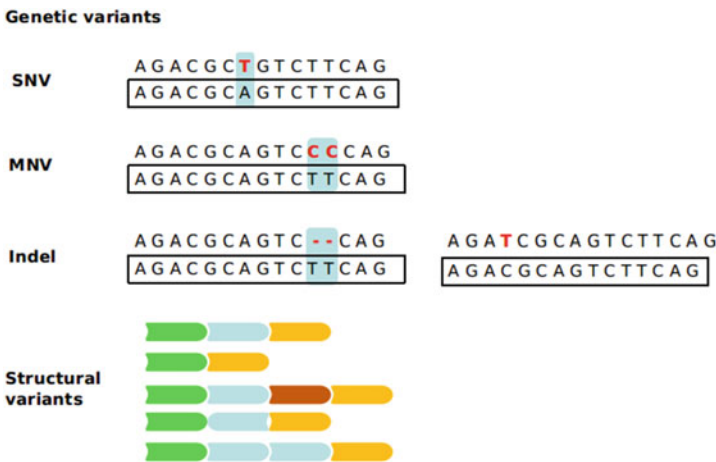
As we have seen, errors can be introduced at every step of the variant calling process. On top of errors brought in during the base calling and read mapping and alignment steps, other factors that can influence data quality are the preparation and storage of samples prior to analysis. For example, it is known that samples stored as formalin-fixed paraffin embedded (FFPE) tissue will have a higher bias toward C>T mutations due to deamination events

triggered by a long fixation time [9]. Although these events are detectable only at a small fraction of the reads aligning to a particular site, they can become important when analyzing a pool of genomes sequenced at low frequency or when studying tumor samples that could have subclonal mutations—furthermore confounded by the tendency of some of these tumor types toward having more real C>T mutations [10]. Another example comes from the observation that DNA oxidation can happen during the shearing step most NGS protocols have implemented, and that this results in artifactual C>A mutations [11]. Ancient DNA and ctDNA can also suffer from these problems [12, 13]. Therefore, a researcher needs to consider their sample origin and preparation protocol and undertake post-processing filtering steps accordingly.

### 10.2.5 How to Choose an Appropriate Algorithm for Variant Calling?

In addition to considering the variant calling method (*e.g.*, naive, probabilistic, or heuristic) that an algorithm implements, a researcher also needs to consider the types of genetic variants that they are interested in analyzing, perhaps having to run several programs at the same time to obtain a comprehensive picture of the genetic variation in their samples.

Genetic variants are usually classified into several groups according to their characteristics (Fig. 10.2):



**Fig. 10.2** Classes of genetic variants. Genetic variants ranging from a single base change, to the insertion or deletion of several bases can occur in a genome. Structural variants are more complex and encompass larger sections of a genome: At the top, a reference sequence, in the second row, a large deletion (blue region), in the third row, a large insertion (red section), in the fourth row, an inversion, and in the fifth row, a duplication. This figure is based on one drawn by Petr Danecek for a teaching presentation



- *SNVs (single nucleotide variants)*, also known as single base substitutions, are the simplest type of variation as they only involve the change of one base for another in a DNA sequence. These can be subcategorized into transitions (Ti) and transversions (Tv); the former are changes between two purines or between two pyrimidines, whereas the latter involve a change from a purine to a pyrimidine or vice versa. An example of a transition would be a G > A variant. If the SNV is common in a population (usually with an allele frequency > 1%), then it is referred to as a SNP (single nucleotide polymorphism). A common post-calling analysis involves looking at the Ti/Tv ratio, which can vary between 2 and 3 depending on the genomic region under analysis [14]. If this ratio is far from the expected, it may indicate a large proportion of false positive calls.
- *MNVs (multi-nucleotide variants)*, which are sequence variants that involve the consecutive change of two or more bases. An example would be one of the types of mutations caused by UV irradiation, CC>TT. Similarly to SNVs, there are some MNVs that are found at higher frequencies in the population, which are referred to as MNPs [15].
- *Indels (portmanteau of insertions and deletions)*, which involve the gain or loss of one or more bases in a sequence. Usually, what is referred to as indel tends to be only a few bases in length. An example of a deletion would be CTGGT > C and an insertion would be represented as T > TGGAT.
- *Structural variants*, which are genomic variations that involve larger segments of the genome. These can involve inversions, which is when a certain sequence in the genome gets reversed end to end, and copy number variants including amplifications, when a fraction of genome gets duplicated one or more times, and larger deletions, when large segments of the genome get lost. There is not a strict rule defining the number of base pairs that make the difference between an indel and a structural variant, but usually, a gain or loss of DNA would be called a structural variant if it involved more than one kilobase of sequence.

Most variant callers identify SNVs, but there are only some variant callers that will report indels or structural variation [3]. This is because usually the algorithms underlying the detection of these types of variants tend to be quite different: SNV, MNV, and short indel detection comprise the comparison of a pile of sequencing reads and their alignments to the reference genome (as has been discussed throughout Chap. 9), whereas larger indels and structural variant calling require calculating a distribution of insert sizes and detecting those read pairs that fall outside it, as well as the direction of alignment of both mate pairs [16].

It is also important to consider the type of sequencing that was used for the experiment. For example, whole genome sequencing and whole exome sequencing have different amounts of coverage, depth, and sequencing uniformity. Some variant callers such as MuTect2 and Strelka2 show better performance in sequencing with higher average sequencing depth and lower coverage [17].

### 10.3 Working with Variants

Variant calling usually outputs VCF files (See Sect. 7.2 File Formats) [18]. To recap, VCF files are plain text files that contain genotype information about all samples in a sequencing project. A VCF file is arranged like a matrix, with chromosome positions in rows and variant and sample information in columns. The sample information contains the genotype called by the algorithm along with a wealth of information such as (depending on the algorithm) genotype likelihoods and sequencing depth supporting each possible allele, among others. For each variant position, the file also contains information outputted by the variant caller such as the reference and alternate alleles, the Phred-based variant quality score, whether overlapping variants in other sequencing or genotype projects have been found, etc. Crucially, this file also contains a column called “FILTER,” where information about whether further quality filters have been applied to the calls and which ones. We will review here some of the most common filters that researchers should consider applying to their data once it has already been called.

---

#### Review Question 2

How do you think a researcher can deal with the uncertainty about false negatives, *i.e.* sites where a variant has not been called? How can they be sure there is no variation there and it is not, let us say, a lack of sequence coverage?

### 10.4 Applying Post-variant Calling Filters

So far, we have seen a number of steps where a researcher must be careful to increase both the sensitivity and specificity of their set of calls in order to have an accurate view of the amount and types of sequencing variation present in their samples. However, there are also a number of post-calling filtering steps that should be applied in the majority of cases in order to further minimize the amount of false positive calls. Here we will review the main such filters, but the reader is referred to [18] if they wish to delve deeper into this topic.

**Strand Bias Filter** . Sometimes a phenomenon, referred to as strand bias, can be observed where one base is called only in reads in one direction, whereas it is absent in the reads in the other direction. This is evidently an error introduced during the preceding steps, and can be detected through a strand bias filter. This filter applies a Fisher’s exact test comparing the number of forward and reverse reads with reference and alternate alleles, and if the  $P$ -value is sufficiently small as determined through a pre-chosen threshold, then the variant is deemed an artifact.

**Variant Distance and End-Distance Bias Filters** These filters were primarily developed to deal with RNA sequencing data when aligned to a reference genome [19]. Therefore, if a variant is mostly or only supported by differences in the last bases of each read, the call may

be a false positive resulting from a portion of a read coming from a mRNA being aligned to an intron adjacent to a splice junction. For end-distance bias, a *t*-test is performed in order to determine whether the variants occur at a fixed distance from read ends, whereas for the variant distance bias tests whether or not variant bases occur at random positions in the aligned reads [19].

**Indel and SNP Gap Filters** These filters are designed to flag variants that are too close to an indel or each other, respectively, as these may stem from alignment artifacts and therefore be false positives. A value of 5 for these filters would mean that SNPs that are five or fewer bp from an indel call would be discarded. The same would apply to clusters of SNPs that are 5 or fewer bases apart from each other.

---

### Review Question 3

Can you think of other possible filters that would need to be applied to the data post-variant calling to reduce the number of false positive calls?

These and other post-variant calling filters can be applied to a VCF file. Programs such as bcftools [5] or GATK VariantFiltration [4] can be used for this purpose, and their parameters can be tweaked to suit the researcher's needs. Typically, after these are run, the "FILTER" field in the VCF will be annotated as "PASS" if the variant has passed all specified filters, or will have specified the filters that it has failed. Usually, the subsequent analyses would be carried out only with those variants that passed all filters.

Now that we have seen the factors that researchers take into account when performing an NGS study, we will next discuss the types of analyses they can follow to identify *de novo* DNA variants or those that are likely to increase the risk of a disease.

---

## 10.5 *De novo* Genetic Variants: Population-Level Studies and Analyses Using Pedigree Information

*De novo* mutations are crucial to the evolution of species and play an important role in disease. *De novo* genetic variants are defined as those somatically arising during the formation of gametes (oocytes, sperm) or that occur postzygotically. Only the mutations present in germ cells can be transmitted to the next generation. Usually, when searching for *de novo* variants in children (usually affected by developmental disorders), researchers study trios, *i.e.* both parents and the child. The task is simplified because at  $1.0 \times 1.10^{-8}$  to  $1.8 \times 1.10^{-8}$  per nucleotide mutation rate, only a few *de novo* mutations are expected in the germline of the child (a range of 44–88 according to [20]). If more than one variant fulfills these criteria, bioinformatic methodologies such as examination of the extent of conservation throughout evolution, consequence prediction, and gene prioritization are then used to pinpoint the most likely gene variants underlying the phenotype. Functional studies such as

cell growth experiments or luciferase assays can then be performed to demonstrate the biological consequences of the variant.

This type of filtering methodology has been extensively applied by projects such as the Deciphering Developmental Disorders (DDD) study [21]. This Consortium applied microarray and exome sequencing technologies to 1,133 trios (affected children and their parents) and was able to increase by 10% the number of children that could be diagnosed, as well as identifying 12 novel causative genes [22]. It may also be useful for the detection of causal genetic variation for neurodevelopmental disorders [23]. However, while tremendously useful in the cases where the three sequences are available, and where the variant is present in the child and not the parents, this strategy is not that useful in those cases where the causal variant may also be present in the parents or where it has a lower penetrance.

Another definition of a “*de novo*” variant may be one that has never before been seen in a population, which is identified through comparisons against population variation databases such as gnomAD [15] and dbSNP [24]. Sometimes, researchers assume that a rare variant, because it is rare (and perhaps because it falls in a biologically relevant gene) then it must underlie their phenotype of interest. However, this is nearly always not true: Depending on ancestry, estimates are that humans can carry up to 20,000 “singletons” (this is, genetic variants only observed once in a dataset) [25] and can carry more than 50 genetic variants that have been classified as disease-causing [15]. This point has been beautifully illustrated by Goldstein and colleagues [26]: They analyzed sequencing data from a control sample and reported finding genetic variants falling in highly conserved regions from protein-coding genes, that have a low allelic frequency in population databases, that have a strong predicted effect on protein function and in genes that can be connected to specific phenotypes in disease databases. However, even if fulfilling all these criteria, these variants clearly do not have a phenotype. They call the tendency of these kind of variants to be assumed as causal as “the narrative potential,” which is unfortunately common in the literature [27]. Therefore, in the next section we will summarize the aspects that need to be taken into account in order to confidently assign a genetic variant as causative for a phenotype.

---

## 10.6 Filtering Genetic Variants to Identify Those Associated to Phenotypes

Given the huge number of genetic variants usually identified in NGS studies (12,000 in exomes, ~5 million in genomes) [28], filtering and post-processing to pinpoint candidates may be the most labor-intensive tasks out of the whole analysis pipeline. Depending on the researcher’s biological question, they may need to tune these parameters to better answer it. For example, if they are searching for rare variation in pedigrees that may predispose to a disease, they may want to set quite permissive quality thresholds so as to not lose any potential candidates, but making sure that any potential variants are confirmed through re-

sequencing by another orthogonal methodology such as capillary sequencing. If, on the other hand, they are analyzing a large cohort of individuals in order to describe patterns of variation, then they will need to be much stricter quality filters.

### 10.6.1 Variant Annotation

Variant annotation can help researchers filter and prioritize functionally important variants for further study. Several tools for functional annotation have been developed; some of them are based on public databases and are limited to known variants, while others have been developed for the annotation of novel SNPs.

Functional prediction of variants can be done through different approaches, from sequence-based analysis to structural impact on proteins. Predicted effects of identified variants can be assessed through tools such as Ensembl-VEP [29] and SnpEff [30]. On top of the predicted consequences on protein function (*e.g.*, whether a variant is missense, stop-gain, frameshift-inducing, etc.), these tools can also perform annotations at the level of allele frequency against public databases such as 1000 Genomes and GnomAD, whether the variant has been seen before either in populations or somatically in cancer (dbSNP and COSMIC annotations), whether it falls in an evolutionarily conserved site (GERP and PolyPhen-2 scores), and whether it has been found to have clinical relevance (ClinVar annotations), among others. Genomic region-based annotations can also be performed, referring to genomic elements other than genes, such as predicted transcription factor binding sites, predicted microRNA target sites, and predicted stable RNA secondary structures [31]. All these annotations can aid a researcher to focus on those variants predicted to be associated to their phenotype of interest.

However, these steps to identify variant candidates are only part of the story. As we mentioned above, even if the variants are real and seem to have an effect on gene function, this alone is not enough evidence to link the variant causally to a phenotype [32]. Researchers should be wary of any potential positive associations and should consider alternate hypotheses before reporting their identified variants as causal (or they may be publicly challenged, see, for example, [33, 34]).

### 10.6.2 Evaluating the Evidence Linking Variants Causally to Phenotypes

After these essential filtering and annotation steps have been performed, a researcher then needs to assess the amount of evidence supporting the potential causality of a genetic variant. The first line of evidence needs to be statistical: Assuming a candidate variant exists, the first question would be, how likely would it be to obtain an equivalent result by chance if any other gene were to be considered? For example, a 2007 study by Chiu and collaborators assumed that two novel missense genetic variants in the *CARD3* gene were causal of familial hypertrophic cardiomyopathy [35]. They assumed causality based on

four criteria: If the variant had been seen in other cardiomyopathy patients, if it was absent from 200 alleles from controls, if it was conserved among species and isoforms and if it cosegregated with the disease in affected families. However, the chance of all these criteria being fulfilled by chance alone if any other genes had been considered is high—as a study subsequently found by assessing a larger gene panel and calculating the expected number of variants in the gene [34]. Additionally, both positive and negative evidence for the hypothesis should be carefully evaluated, for example, in the same cardiomyopathy study some of the “potentially causal” variants predicted by bioinformatics algorithms did not cosegregate with the phenotype [34]. The increasing availability of sequencing data in large cohorts such as gnomAD should help establishing causality as more accurate allele frequencies are reported per population [15]. This is an important point—allele frequencies should be matched by ancestry as closely as possible, as it is known that they can vary greatly among different populations [25].

Another important set of criteria, highlighted by MacArthur et al [32], argues that when analyzing potentially monogenic diseases, genes that have previously been confidently linked to similar phenotypes should be analyzed as the first potential candidates before proceeding to explore novel genes, and that if a researcher does proceed to analyzing further genes, then multiple independent carrier individuals must present with similar clinical phenotypes. Additionally, it is desirable that the distribution of variants in a suitable control population is examined, for example, if a researcher has identified a novel stop-gained variant in a candidate gene, how many other stop-gained variants are found in population-level variation catalogues?

Finally, statistical evidence and multiple computational approaches may strongly suggest that a variant is disease-causing. However, whenever possible, researchers should perform functional studies that indicate this is the case, whether by using tissue derived from patients themselves, cell lines, or model organisms. The comprehensive view provided by statistical, computational, and functional studies then may be enough for a researcher to report a potential causal variant. In doing so, it is recommended that all available evidence is detailed, clear and uncertain associations are reported and that all genetic data is released whenever possible [32].

### 10.6.3 Variant Filtering and Visualization Programs

Finally, visual representation of genomic data can be highly useful for the interpretation of results [28]. Visualization tools can help users browse mapped experimental data along with annotations, visualize structural variants, and compare sequences. These programs can be available as stand-alone tools or as web applications, and vary in the amount of bioinformatics knowledge necessary to operate them. Here we will review some of the most popular and that we consider useful, but there are many others suited for different purposes and with a range of functionalities.

- *Integrative Genomics Viewer (IGV)* [36] A very popular, highly interactive tool that is able to process large amounts of sequencing data in different formats and display read alignments, read- and variant-level annotations, and information from other databases. Website: <http://software.broadinstitute.org/software/igv/>
- *Galaxy* [37] Another highly popular, web-based platform that allows researchers to perform reproducible analyses through a graphical interphase. Users can load files in the FASTA, BAM, and VCF formats, among others, and perform data analysis and variant filtering in an intuitive way. Website: <https://usegalaxy.org/>
- *VCF/Plotein* [38] This web-based, interactive tool allows researchers to load files in the VCF format and interactively visualize and filter variants in protein-coding genes. It incorporates annotations from other external databases. Website: <https://vcfplotein.liigh.unam.mx/>

### Take Home Message

- There are a number of different methods for performing variant calling, these can be naive, probabilistic, and heuristic. Probabilistic methods are the most widely used and implement a form of Bayes' Theorem. However, algorithm choice will depend on the researcher's study design.
- Sample storage and preparation methods may introduce errors that increase false positive calls and therefore should be considered when designing an analysis pipeline.
- Post-variant calling filters that analyze the distribution of variants across all sequencing reads will usually need to be applied to data in order to reduce false positive calls.
- True *de novo* genetic variants can be identified by analyzing trios with an affected child, in other scenarios a number of annotations and filtering steps need to be applied to identify candidate variants.
- For a researcher to ascribe phenotype causality to a genetic variant, the result of gene- and variant-level annotations are not enough; a number of further statistical, bioinformatic, and functional considerations need to be taken into account.
- Variant filtering and visualization tools can aid a researcher to perform the above mentioned steps in an easy and intuitive way.

### Answers to Review Questions

*Answer to Question 1:*  $Q = 34.77$ .

*Answer to Question 2:* The logical option is for the researcher to go back and analyze, through tools such as *Samtools depth*, whether indeed there is enough coverage at every assessed site, and to mark it as "no call" otherwise. A novel VCF format, called gVCF and outputted by GATK, can now give reference call confidence scores.

*Answer to Review Question 3:* There are a number of filters already implemented in variant filtering tools, some of these are a threshold for Phred-scaled variant quality,

minimum depth, whether the variant is in a low-complexity region (*i.e.*, a highly repetitive region that may increase alignment errors), and more.

### **Box: Genome Assembly**

If there is no reference genome available for our species of interest, it may be worth trying to create one from scratch. To do this, DNA fragments of the targeted species are sequenced in high quantity, resulting in sequenced reads that theoretically cover the entire genome. Reads are aligned and merged based on their overlapping nucleotides, assembling long DNA sequences. When the order of bases is known to a high-confidence level, this genomic sequence will be named a “contig.” Multiple contigs can be assembled together to form a scaffold based on paired read information. A scaffold is a portion of the genome sequences composed of contigs but which might contain gaps in between them.

There are various tools to close gaps between scaffolds. Scaffolds can then be joined together to form a chromosome. Despite how easy this may sound, genome assembly has its difficulties and it can vary between one organism to another (for example, an uneven representation of the genome due to sequencing sensitivity to GC bias, which can cause gaps between scaffolds) [39].

### **Further Reading**

The GATK blog: <https://software.broadinstitute.org/gatk/documentation/article.php?id=4148>

- Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, et al. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet.* 2013 Jul;14(7):460–70.
- Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature.* 2015 Mar 12;519(7542):223–8.
- MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature.* 2014 Apr 24;508(7497):469–76.

#### *Somatic variant callers*

*Mutect2*. (Bayesian) Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. “Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples.” *Nature Biotechnology.* 2013;31(3):213–19.

*GATK HaplotypeCaller* (Bayesian). McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. “The



Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research*. 2010;20(9):1297–1303.

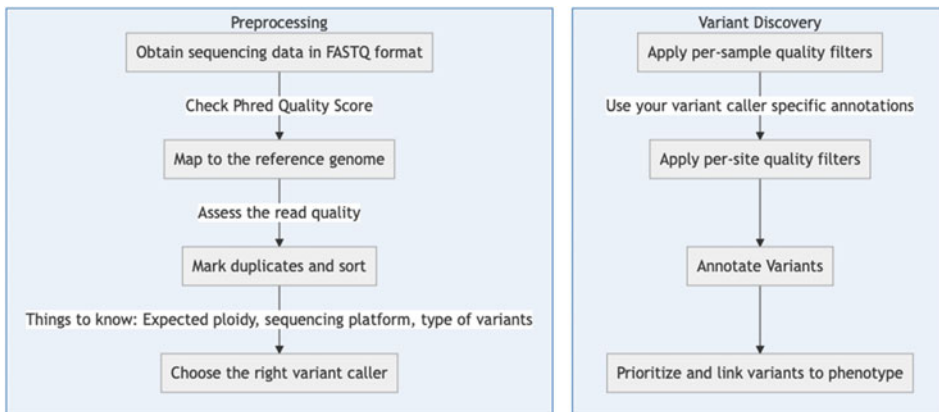
*Varscan2*. (Heuristic). Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. “VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing.” *Genome Research*. 2012;22(3):568–76.

## 10.7 A Practical Example Workflow

A workflow chart for a typical variant calling analysis is shown in Fig. 10.3.

### 1. Preprocessing

- Check that the base calling and read alignment are accurate using the standard Phred quality score.
- Know the expected ploidy in your experiment.
- Know the type of variants you want to identify.
- Know your sequencing platform.
- Calculate the sequencing depth.
- Check the coverage ratio between X and Y chromosome to determine sample sex concordance.
- Identify duplicated or related samples.
- Apply a filter on low-complexity regions according to your research interest.
- Select the appropriate variant caller.



**Fig. 10.3** Workflow-Chart for identification of genetic variants and de novo mutations

## 2. Once the VCF file is generated:

- Apply per-sample filters that can relate to your cohort, some examples are:

```
#Check the missingness on a per-individual basis
vcftools -vcf *.vcf -missing-indv -out *
Check the heterozygosity rate of each sample
vcftools -vcf *.vcf -het -out *
# Check the p-value for each site from a Hardy-Weinberg Equilibrium test
vcftools -vcf *.vcf -hardy -out
```

After all these metrics are calculated, we suggest you graph each of them to easily identify outliers and define a threshold for further filtering. These metrics should also be calculated per variant site and filters should be applied under that dimension.

- Apply per-site filters that can relate to your variant calling method, for example, check the strand bias (identified by performing a Fisher test) “FS” and/or the strand OR “SOR” values.

## 3. To identify *de novo* variants

Annotate your VCF file with the previously known information for each variant using tools like Ensembl-VEP [29] or SnpEff [30].

```
ensembl-vep/vep -i *.vcf -o *.vep -everything
SnpEff eff -v * -stats * refgenome *.vcf
```

Check for the allele frequency of your variants in the population that your samples came from in the different available data bases, is it significantly different from the allele frequency you observed in your experiment? How can you explain this?

## 4. Link your candidate variants to a phenotype

Follow the advice by MacArthur et al 2014 [32] for identifying causality of genetic variants, in particular, identify whether your result is statistically significant or whether it may have arisen by chance. Perform functional experiments that can explain the mechanism by which your variant affects the phenotype in the specific context of the background your samples carry. Search for literature that support your findings.

**Acknowledgements** We thank Dr. Stefan Fischer (Biochemist at the Faculty of Applied Informatics, Deggendorf Institute of Technology, Germany), and Dr. Petr Danecek (Wellcome Sanger Institute, United Kingdom) for reviewing this chapter and suggesting extremely relevant enhancements to the original manuscript.

## References

1. Ewing B, Green P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* 1998;8(3):186–94.
2. 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.
3. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J.* 2018;16:15–24.
4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303.
5. 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.
6. You N, Murillo G, Su X, Zeng X, Xu J, Ning K, et al. SNP calling using genotype model selection on high-throughput sequencing data. *Bioinformatics.* 2012;28(5):643–50.
7. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22(3):568–76.
8. Cai L, Yuan W, Zhang Z, He L, Chou K-C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci Rep.* 2016;6:36540.
9. Prentice LM, Miller RR, Knaggs J, Mazloomian A, Aguirre Hernandez R, Franchini P, et al. Formalin fixation increases deamination mutation signature but should not lead to false positive mutations in clinical practice. *PLoS One.* 2018;13(4):e0196434.
10. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. *Nature.* 2017;545(7653):175–80.
11. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* 2013;41(6):e67.
12. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* 2010;38(6):e87.
13. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol.* 2016;34(5):547–55.
14. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics.* 2015;31(3):318–23.
15. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285–91.
16. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 2019;20(1):117.
17. Bohannon ZS, Mitrofanova A. Calling variants in the clinic: informed variant calling decisions based on biological, clinical, and laboratory variables. *Comput Struct Biotechnol J.* 2019;17:561–9.
18. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27(15):2156–8.
19. Danecek P, Nellåker C, McIntyre RE, Buendia-Buendia JE, Bumpstead S, Ponting CP, et al. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biol.* 2012;13(4):26.

20. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* 2016;17(1):241.
21. Firth HV, Wright CF, Study DDD. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol.* 2011;53(8):702–3.
22. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature.* 2015;519(7542):223–8.
23. Carneiro TN, Krepischi AC, Costa SS, Tojal da Silva I, Vianna-Morgante AM, Valieris R, et al. Utility of trio-based exome sequencing in the elucidation of the genetic basis of isolated syndromic intellectual disability: illustrative cases. *Appl Clin Genet.* 2018;11:93–8.
24. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308–11.
25. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68–74.
26. Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, et al. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet.* 2013;14(7):460–70.
27. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med.* 2011;3(65):65ra4.
28. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efreanova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform.* 2014;15(2):256–78.
29. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 2016;17(1):122.
30. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly.* 2012;6(2):80–92.
31. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164.
32. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature.* 2014;508(7497):469–76.
33. Minikel EV, MacArthur DG. Publicly available data provide evidence against NR1H3 R415Q causing multiple sclerosis. *Neuron.* 2016;92(2):336–8.
34. Verhagen JMA, Veldman JH, van der Zwaag PA, von der Thüsen JH, Brosens E, Christiaans I, et al. Lack of evidence for a causal role of CALR3 in monogenic cardiomyopathy. *Eur J Hum Genet.* 2018;26(11):1603–10.
35. Chiu C, Tebo M, Ingles J, Yeates L, Arthur JW, Lind JM, et al. Genetic screening of calcium regulation genes in familial hypertrophic cardiomyopathy. *J Mol Cell Cardiol.* 2007;43(3):337–43.
36. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24–6.
37. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 2016;44(W1):W3–10.
38. Ossio R, Garcia-Salinas OI, Anaya-Mancilla DS, Garcia-Sotelo JS, Aguilar LA, Adams DJ, et al. VCF/Plotein: visualization and prioritization of genomic variants from human exome sequencing projects. *Bioinformatics.* 2019;35(22):4803–5.
39. Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinform.* 2009;10(4):354–66.