Knowledge multiplies when shared — when calling things by their right name: improving the validation and exchange of genetic data in research and diagnostics

Fokkema, I.F.A.C.

CHAPTER 2

Genetic variation and related standards

Ivo F.A.C. Fokkema[1], Johan T. den Dunnen[1,2]

1 - Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands.
2 - Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands.

Adapted from: *General considerations: terminology and standards.*
Chapter in: *Clinical DNA variant interpretation: Theory and Practice.* (2021) Elsevier.
DOI: 10.1016/C2019-0-01138-7.

2.1 Abstract

The human genome, over three billion nucleotides in size, collects changes over time that become part of the naturally existing variation in the population. In the case of a genetic disorder, of all these variants, only one or two are usually associated with the presented condition. Finding these variants in such large data sets is truly like searching for a needle in a haystack. This chapter describes the types of genetic variation and their possible consequences, as well as various standards and their importance related to describing, interpreting, and reporting genetic variants and phenotypes. Some important points to consider when classifying variants are discussed, as well as general challenges and considerations to keep in mind when performing sequencing analysis.

2.2 Introduction

The human genome, the collection of our entire DNA sequence, is often referred to as "the book of life". Decades of research have brought us releases of the human genome complete enough to allow for various applications, including clinical diagnoses. The latest of such releases, the Genome Reference Consortium's GRCh38 reference genome build, was completed in December 2013 and consists of over 3 billion letters (nucleotides). In our body, almost all DNA is present in two copies: one maternal and one paternal copy. Every time a cell in our body divides, both copies need to be duplicated (replicated), a process that is very precise but not without errors. These errors will be copied and passed on to the next generation, and over time, the human DNA sequence slowly changes. The speed at which this happens, the mutation rate, is estimated to be around 1.5 nucleotides per year,[1] with estimates ranging between 36 and 63 changes (variants) passed on to the next generation.[2,3] Most of these variants do not cause disease but become part of the naturally existing variation in the population.

Nowadays, we are able to determine the sequence of a human individual within a few days. Since, especially on a global scale, the natural variation in the human DNA is high, comparing a person's DNA to a standard reference sequence is not without problems. Compared to the reference, an average human genome contains about 4 million variants, while an average exome analysis (i.e., analysis of all protein-coding sequences) returns some 40,000 variants. In the case of a genetic disorder, of these variants, only one or two are usually associated with the presented condition. Finding these variants in such large datasets is truly like searching for a needle in a haystack. Being able to see the difference between the vast majority of benign variants and the few disease-causing (pathogenic) variants, requires a good understanding of the different types of variants and the possible consequences these variants have on the function of the genes they affect.

This chapter describes the types of genetic variation and their possible consequences, as well as various standards and their importance related to describing, interpreting, and reporting genetic variants and phenotypes. Some important points to consider when classifying variants are discussed, as well as general challenges and considerations to keep in mind when performing sequencing analysis.

2.3 Genetic variation

2.3.1 Types of DNA sequence changes

DNA variants can be characterized by the type of variation that occurs on the DNA level, as well as their consequences on either RNA or protein level. To prevent those consequences from getting mixed, it is best to strictly separate and report each level individually (DNA, RNA, and protein). As variant screening is mainly based on DNA analysis, variants detected are primarily described on the DNA level. In addition, the (predicted or confirmed) consequences on the RNA and protein level can be given.

In general, current short-read, high-throughput sequencing technologies cannot easily detect all different DNA variant types. To detect all variant types, either special analysis pipelines are required or long-read sequencing technologies need to be applied. Table 2.1 lists the basic DNA sequence variant types.

Table 2.1: **DNA sequence variant types.** Each type is explained by an example sequence; the original DNA sequence and the changed DNA sequence, in which the variant occurred. The nucleotides that are part of the variant have been highlighted in red.

Variant type	Sequence	Description
Substitution	`AACGTT` `AACCTT`	One nucleotide has been replaced by another.
Deletion	`AACGTT` `AAC-TT`	One or more nucleotides have been removed.
Insertion	`AAC-GTT` `AACAGTT`	One or more nucleotides have been inserted.
Duplication	`AAC-GTT` `AACCGTT`	One or more nucleotides have been duplicated in tandem.
Deletion-Insertion	`AACGTT` `AATATT`	One or more nucleotides have been removed and replaced by one or more other nucleotides, other than a substitution.
Inversion	`AACGTT` `ACGTTT`	More than one nucleotide has been inverted into their reverse complement sequence.
Structural variation		A variation where large parts of chromosomes have rearranged.

- **Substitutions** are variants where one single DNA nucleotide is replaced by another single DNA nucleotide. This is by far the most common type of DNA sequence variant, taking up ~80% of all reported DNA variation.
- **Deletions** are variants where one or more nucleotides have been removed from the original DNA sequence. This is the next most common variant type. When a deletion

spans one or more exons of a gene or more than 1000 nucleotides, it is referred to as a copy number variant (CNV).

- **Insertions** are the reverse of deletions and occur when one or more nucleotides are added to the original sequence. When the inserted sequence is a tandem copy of the original DNA sequence, it is called a **duplication**. Both duplications and deletions frequently occur where the DNA contains repeated copies of a small sequence. When a duplication spans one or more exons of a gene or more than 1000 nucleotides, it is referred to as a CNV.
- **Deletion-insertions** are a combination of a deletion and an insertion in the same location in the DNA (excluding substitutions). One or more nucleotides are replaced by one or more other nucleotides.
- **Inversions** are variants where a stretch of DNA turns around (inverts); the inserted sequence is the exact reverse complement of the deleted sequence. Inversions have a minimum length of two nucleotides; one-nucleotide inversions are classified as simple substitutions.
- **Structural variation** is a term for various large chromosomal changes such as translocations and transpositions. Note that these are usually not picked up by short-read sequencing methods and require additional tests to be detected. If the structural changes are large enough, they can be seen using optical mapping technologies or microscopy (karyotyping).

2.3.2 Types of RNA sequence changes

Variants on RNA level include those detected on the DNA level, i.e., substitution, deletion, insertion, duplication, deletion-insertion, and inversion, as well as variants affecting RNA processing. RNA processing includes transcription initiation (promoter and locus control regions), RNA capping, RNA splicing, RNA modification (editing), polyadenylation, and transcription termination. In addition, variants may indirectly influence the RNA, altering its folding, stability, and degradation, and thereby its quantity in the cell. An exceptional case is RNA fusion transcripts, where parts of two different genes get fused into one transcript. RNA fusion transcripts usually occur after a translocation or deletion removing the 3′ end of a gene.

2.3.3 Types of protein sequence changes

Variants on protein level include those detected on DNA and RNA level, i.e., substitution, deletion, insertion, duplication, and deletion-insertion. Although effectively part of these categories, some variants affect protein translation and are treated separately, i.e., frameshift and extension variants. Just like for RNA, variants may affect protein processing, translation initiation, translation termination, protein modification, protein folding, and protein-protein

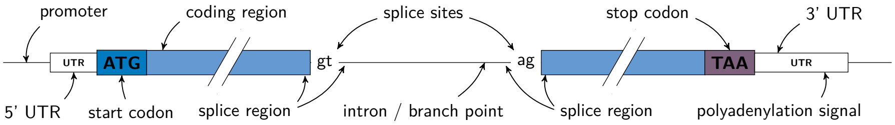

Figure 2.1: **Locations in which variants can influence a gene's function.**

interaction. In addition, variants may influence the stability and degradation of the protein molecule and, thereby, its quantity in the cell. An exceptional case, like for RNA, is fusion proteins translated from RNA fusion transcripts.

2.4 Variant consequences by location

In the literature, rather than the genomic DNA change, variants are often described by their location in the gene or by the effect of the variant on the protein level. Such an effect can be deleterious in two ways: loss of function or gain of function. A single-nucleotide substitution in a crucial part of the gene results in a far more devastating effect than a deletion or insertion in, for instance, an intron. Figure 2.1 shows the different locations where a DNA variant can influence the function of a gene. Below is a brief overview of the main categories, based on the basic genetic unit - a gene, and the RNA and protein it encodes.

2.4.1 Promoter region

The **promoter region** regulates the gene's transcription. Variants in this region may affect the transcription factor binding sites that are present and may increase, lower, or even abolish the gene's expression. As such, variants in this region may render the gene dysfunctional. Since it is quite difficult to determine the exact location of the promoter for each gene, it is a challenge to distinguish variants in the promoter region from intergenic variants that do not affect the gene's expression. Furthermore, for many genes, the timing of expression (when during development and in which tissue the gene is expressed) is controlled by far distant sequences, such as locus control regions, enhancers, topologically associating domains, etc.

2.4.2 5′ untranslated region

The 5′ untranslated region (**5′ UTR**), before the initiation (start) codon, contains regulatory elements such as internal ribosome entry sites (IRESs) and upstream open reading frames (uORFs) that control the translation.[4] Variants in these sequences mainly influence translation initiation and affect translation levels. As with the promoter region, the annotation of relevant active sites within the 5′ UTR is usually lacking, and functionally relevant variants cannot easily be distinguished from nonrelevant 5′ UTR variants.

2.4.3 Start codon

Variants in the **start codon** that alter the ATG sequence block translation initiation and usually have serious consequences. When the ATG is affected, initiation may move to another initiation site, either up- or downstream, and only when that site is in frame with the normal protein sequence can the translation product (partially) be functional. The consequences of duplications involving the ATG motif are difficult to predict. Although they leave a normal sequence, at the same time, they introduce a new competing upstream initiation site. The sequence surrounding the start codon, coined the Kozak sequence, also shows conservation and is sensitive to variants that can change the level of translation.[5]

2.4.4 Protein-coding region

The consequences of variants in the **protein-coding region** are, in general, more severe when a large segment of the protein is altered. When a deletion removes the entire gene or the start of a gene, no protein is made, and whether this loss can be compensated depends on the activity of the copy on the other chromosome. Note that deletions on the X-chromosome in males cannot be compensated since there is no second gene copy. Duplications of an entire gene may increase the amount of RNA/protein produced, and it depends on the gene whether this has harmful consequences. Some genes are "dosage-sensitive" while others are not (see `clinicalgenome.org/curation-activities/dosage-sensitivity/` for a list). Furthermore, in many cases, missing one copy is often tolerated better than having one copy with an altered protein sequence disturbing normal cellular processes.

Most variants in the protein-coding region lead to the production of an altered protein. The resulting protein may not be functional at all, function only partially, or even give an additional or completely new function. The consequences of **missense** variants, which replace one amino acid for another amino acid, vary depending on the change in size, charge, and hydrophilicity of the affected amino acid, as well as its position relative to the functional domains of the protein.

Nonsense variants replace an amino acid for a translation stop codon, causing premature protein translation termination. This usually has deleterious consequences. The same applies for **frameshift** variants, which not only truncate normal translation but also add a completely new C-terminal tail to the protein. This new protein tail may be either shorter or longer than the original and can be of considerable size. It may have undesired functional consequences (gain of function), interfering with and disturbing normal cellular processes. When nonsense or frameshift variants occur near the end of the protein, and the length of the C-terminal tail is small, normal protein function may be unaffected.

In-frame variants (deletions, insertions, and duplications) do not disturb the reading frame and may have less severe consequences. A well-known example is the DMD gene, where in-frame deletions or duplications, even when spanning many exons, cause a relatively mild phenotype compared to truncating variants (nonsense or frameshift variants).[6] The effect of in-frame variants mainly depends on the function of the protein and the size of the segment of the protein affected. In general, variants affecting larger stretches have a more significant impact. Still, like the `p.Phe508del` variant in the *CFTR* gene causing cystic fibrosis, even the deletion of a single amino acid may already have seriously harmful consequences.[7]

When DNA variants in the coding region do not lead to a (predicted) change in the amino acid sequence, they are referred to as **silent** or **synonymous** variants. It should be noted, however, that such variants may still influence RNA stability or alter binding sites and RNA processing, particularly splicing, which will then significantly affect the gene's function.

2.4.5 Splice region, splice sites, and introns

After transcription, the RNA molecule undergoes a range of steps before the mature RNA is ready. The 5′ end of the transcript is capped, an important step to protect the RNA from degradation. Most genes are spliced, a process whereby some parts of a gene (the exons, mostly protein-coding) are fused together after removing other sequences (the introns). Finally, many transcripts are processed at the 3′ end by cleavage and the addition of a polyA tail, again protecting the RNA from degradation.

The splicing process is rather complex and involves many sequences. While there is a clear and almost entirely invariable DNA sequence motif spanning the first and last two nucleotides of the intron (GT and AG, respectively, see Figure 2.1), the surrounding sequence is also important yet less well conserved, making predictions of the effect of variants in this region difficult. Changes in the first and last two nucleotides of the intron nearly always disrupt normal splicing. Additionally, on the 5′ side, the splice donor site, variants in the last nucleotide of the exon and nucleotides +3 to +6 often affect splicing. On the 3′ side, the splice acceptor site, especially variants creating a close-by AG dinucleotide cause problems.[8] Some variant effect prediction tools consider a more cautious approach and extend the region that possibly affects splicing to the first and last eight nucleotides of the intron and the first and last three nucleotides of the exon.[9] The intron also contains the branch point — a small region close to the 3′ end of the intron, containing a single strongly conserved adenine nucleotide. The branch point initiates the formation of the loop structure (lariat) that is formed when the intron is spliced out. A variant in the branch point will disrupt the intron's splicing completely.[10] Finally, variants in intronic and exonic splice enhancer and silencer motifs (ISE, ISS, ESE, and ESS) also influence splicing. However, since their sequence is less conserved, their position is rarely known, and their involvement is not considered.

Disruption of splicing can also occur through the creation of a new or activation of a cryptic splice site (CSS). CSSs are normally dormant sites that are silenced (suppressed) by stronger, nearby canonical splice sites. Activation occurs when a sequence change strengthens the cryptic site or weakens the canonical site. Upon activation of the CSS, the canonical splice site is no longer or not fully used, and normal splicing is wholly or partially disrupted.

Disruption of splicing has a range of different consequences on the RNA level. Note that RNA analysis is essential to determine the effect of a variant on splicing. Frequently, variants affecting splicing lead to multiple transcripts being produced, with the overall effect depending on the relative abundance of each of these transcripts. Possible effects include:

- The deletion of an exon or part of an exon. When a splice site is damaged, an exon might not be recognized at all (deleted), or splicing may shift to a new site in the exon. The resulting deletion can be in-frame or out-of-frame. Out-of-frame deletions result in a frameshift and have a more devastating effect on the resulting protein than in-frame deletions.
- The insertion of intronic sequences. When a splice site is nonfunctional, an intron may not be removed at all (intron retention), splicing may shift to a new site in the intron (thus elongating the exon), or a new exon (pseudoexon) may be inserted. The inserted sequence may contain a translation stop codon or contain an open reading frame that fuses in-frame or out-of-frame with the remainder of the encoded protein sequence. Truncating insertions have a stronger negative effect on the resulting protein than in-frame insertions.

2.4.6 Stop codon

Changes in the **stop codon** that prevent the stop codon from being recognized lead to the elongation of the protein sequence. The effect of the additional C-terminal tail on the protein's function is difficult to predict. In general, a longer tail will have more severe consequences, and most extensions negatively influence protein folding, function, and stability.

2.4.7 3′ untranslated region and the polyadenylation signal

The 3′ untranslated region (**3′ UTR**), the sequence after the translation termination (stop) codon, contains several regulatory elements, such as binding sites for miRNAs and RNA-binding proteins, and the polyadenylation signal and addition site. Together, these directly or indirectly influence RNA stability, folding, transport, localization, and translation efficiency, and consequently, RNA and protein levels.[4] As the functional annotation of these elements (except for the polyadenylation signal) is largely lacking, variants in this region are rarely considered as having deleterious consequences.

2.4.8 Other variation

A specific type of disease is caused by repeat expansions. In these disorders, a short repetitive sequence may increase in length to up to many kilobases. When this sequence is translated (e.g., the CAG repeat in Huntington's disease), it directly affects protein function. When it is located in an intron or the UTR of a transcript, RNA processing (splicing) and stability are affected, and transcription may be silenced (e.g., methylation in the fragile X syndrome).

Epigenetic variation, i.e., variation in the methylation of the DNA sequence, may influence the density of the chromatin structure, thereby affecting transcription. Methylation changes can cause disease by inappropriately silencing or activating gene expression.[11] Most sequencing protocols cannot measure methylation unless specific sample preparation steps are included. Also, RNA editing, the process where the RNA sequence is altered and becomes different from the genomic DNA template, is not detected by standard sequencing protocols but can be involved in disease.[12] Although disease through these mechanisms is rare, they should not be overlooked.

2.5 Standards on describing genetic variation

Universal standards are essential when implementing DNA sequence variant analysis in a clinical setting. These standards are required to remove ambiguity, prevent false-negative or false-positive results, and ensure there is no misunderstanding of what has been found and the associated consequences for the health of the individual. The standards required include naming genes, accepted reference sequences for the human genome and the encoded transcripts, the file formats to exchange sequence information, the description of sequence variants identified, the description of the phenotype of the individual studied, standards to classify the variants detected, and standards to store the information in gene variant and phenotype databases.

Even when there is a universal standard, this does not mean it is applied correctly. An analysis published in 2016 of how diagnostic laboratories applied the same standard, the HGVS variant nomenclature (see also Chapter 5), showed that 31% of the reports checked described the variant incorrectly and in such a way that it could not be reconstructed to the variant observed.[13] Other reports also described the variant identified incorrectly, but in a way that could be recognized and corrected. In both situations, however, incorrect variant descriptions cause inconsistencies and mismatches when comparing reports or searching databases and the literature. Querying external sources for variants identified is essential to variant interpretation and classification. Data from large population studies yield allele frequencies that provide evidence of a variant's pathogenicity, with increasing frequencies making it less likely a variant has serious deleterious consequences. Gene variant databases contain data from the literature and from unpublished cases and provide detailed information on variants and phenotypes and the likelihood they are causally linked, i.e., variant pathogenicity. Given the multiple steps in the process, to maximize efficacy and reduce the chance of mistakes during variant interpretation, it is essential the same standards are used by everyone involved.

2.5.1 Gene symbols

Genes are widely identified using their symbols — abbreviations mainly based on the gene's function. The HUGO Gene Nomenclature Committee (HGNC) is the official international organization approving gene names and symbols.[14] Although genes can be identified by their unique numerical ID that does not change, e.g., HGNC:1100, people prefer symbols that are more easily remembered, e.g., *BRCA1*. Gene symbols rarely change, but when they do, the HGNC keeps track of a gene's previously used symbols such that expired gene symbols referenced in reports can still be identified. The HGNC is currently actively renaming genes that do not refer to their function (e.g., *FAM#*, *KIAA#*, *c#orf#*, etc.).

2.5.2 Reference sequences

As variants are defined as differences between the sample DNA sequence determined and a reference sequence, the main requirement for any variant description is to clearly define which reference sequence was used. Reference sequences have unique identifiers referring to their respective entries in the reference sequence databases. Reference sequences can be obtained from several databases, including GenBank at the National Center for Biotechnology Information (NCBI) in the United States,[15] Ensembl at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) in the United Kingdom,[16] and the shared NCBI/EBI Locus Reference Genomic (LRG) project.[17] A reference sequence identifier should be stable, and its contained sequence should not change over time. When the sequence does change, this is indicated by the addition of a version number in the identifier. There is no version number for LRG sequences, and a new ID should be issued. More information can be found on the HGVS Nomenclature website (`hgvs-nomenclature.org/background/refseq/`).

The reference sequence for genomic variants detected by next-generation sequencing will most likely be the human genome reference sequence. The first human genome reference sequence was published in 2000. Over many years, this reference sequence has improved, and the human genome's latest version is build 38. Each genome build contains reference sequences per chromosome — chromosomes 1–22, X, Y, and the mitochondrial genome (mtDNA), together representing an entire genome. Only the NCBI's GenBank has identifiers available for whole-chromosome genomic reference sequences. Examples are `NC_000015.9` for chromosome 15 of genome build GRCh37 (also known as hg19), and `NC_000023.11` for chromosome X on genome build GRCh38 (also known as hg38).

2.5.3 Describing variants

After sequence analysis and variant calling, variants are generally stored in the Variant Call Format (VCF) file format, developed for the 1000 Genomes Project and since then adopted by many other large-scale sequencing projects.[18] The VCF file normally does not contain information on the reference genome used. The file has a tabular format indicating the chromosome, a genomic position, the reference sequence at that position, the variant (alternate) sequence identified in the sample(s), and optionally, various details on the sequencing quality, such as coverage, genotype quality, etc.

In publications and databases, variants are described using the "Recommendations for the description of sequence variants" from the Human Genome Variation Society (HGVS) — the HGVS Nomenclature. These standards were first published in 2000 and are regularly updated[19] (**`hgvs-nomenclature.org`**). The HGVS Nomenclature facilitates variant descriptions based on a genomic reference sequence (g. descriptions), based on a transcript (c./n. descriptions), and descriptions for variants on the RNA and protein level (r. and p. descriptions). Any reference sequence can be used as long as the residues altered (nucleotides, amino acids) are located within the reference sequence.

Table 2.2 shows an overview of the main features of both formats.

2.5.3.1 The Variant Call Format

The Variant Call Format was developed for the 1000 Genomes Project[18] and has been designed to be machine-readable for faster processing of large genomic variant datasets. It has become the most often used file format for storing and exchanging large-scale genomic variant data. It supports multiple samples within one file, rich annotation including mapping on transcripts and predicted protein change, and a method to indicate the absence of variants on a certain region (gVCF). Structural variation can also be stored within a VCF file. The VCF standard is currently maintained by the GA4GH (**`ga4gh.org`**) and is currently at version 4.4 (**`github.com/samtools/hts-specs`**).

Table 2.2: **Main features of the HGVS Nomenclature and VCF files.**

	HGVS Nomenclature	VCF files
In use since	2000	2010
Intended use	Human-readable format; databases, literature, reporting.	Machine-readable format; large-scale genome sequencing projects.
Type of variation	Any type of variant, on any reference sequence.	Any genomic variant, excluding complex genomic rearrangements. It can contain transcript variant descriptions in HGVS format in its annotation.
Variant disambiguation	3′ (right) aligned; in principle, each variant has only one correct description.[a] However, the normalization rules are not always followed by its users.	5′ (left) aligned; has a nearly infinite number of possible descriptions for each variant. Requires specialized tools to compare variant sets.

[a] Even though HGVS aims for each variant to have only one correct description, there are currently no unambiguous rules defined regarding when two neighboring variants should be considered as one. These are currently under development. Also, some deletions, duplications, or insertions can be described using an alternative repeat-syntax.

VCF files begin with a header section in which metadata is stored. For instance, the values and contents of some configurable fields found in the VCF file's body are defined and explained in the header. After the header, a single line defines the order of fields and the sample names that are stored in the file. What follows is the file's body, consisting of a list of genomic positions and the relevant sequences. The basic format of each line in the file's body is the chromosome (CHROM), position (POS), external database ID (ID), the sequence of the reference (REF), the sequence found at this position (ALT), and several fields containing annotations. Variant annotations stored in the VCF file can be sample-dependent or sample-independent. Sample-dependent annotations include genotype, genotype quality, and read depth. Examples of sample-independent annotations are gene symbols, mappings on transcripts, and protein change predictions. See Table 2.4 for examples of how variants can be represented in a VCF file.

The VCF standard recommends describing variants using the "simplest representation possible". However, this is not a requirement and such an open recommendation is bound to give different implementations. Also, users are encouraged to use the lowest coordinate for a variant, therefore shifting the variant as far 5′ as possible. Unfortunately, this is in contrast with the existing HGVS standard (see below), which requires variants to be shifted as far 3′ as possible. These issues make a direct comparison of variants between different VCF files or a VCF file and a list of HGVS formatted variants quite difficult. Also, it is common for variants around the same location to be merged into one line in VCF files. In this case, the ALT column will contain multiple values. With a POS of 1 and a REF value of "AG", it is possible to have "TG,AGG" as an ALT value, meaning an A to T substitution on position 1 and a duplication of the G on position 2. This can also be the case for larger variants, causing a large sequence overlap between the REF and ALT columns. In cases like these, the variant description differs quite significantly from the simplest form of describing the variant, and a simple comparison of variants is not possible.

The heterogeneity of variant descriptions in a VCF file is known to cause problems, even within individual diagnostic laboratories.[20] However, several tools exist to normalize and compare VCF files or convert VCF files into lists of HGVS formatted variants, which can then be compared more easily. Examples of VCF normalization and comparison tools are vcfeval, part of RTG Tools,[21] vt normalize,[22] and Best Alignment Normalisation (BAN).[23] Tools that can convert VCF files into HGVS variant descriptions include the hgvs python module[24] and VariantValidator.[25] It should be noted that none of these tools are perfect, partly also because current HGVS recommendations are not unequivocal for describing more complex variants.

In the VCF format, there are no strict rules regarding when to describe variants independently and when as one variant. As variant callers in use in NGS software pipelines rarely call deletion-insertion events but instead prefer calling multiple consecutive variants, VCF files often contain variants found directly next to each other or in very close proximity to each other (also see HGVS below).

2.5.3.2 The Human Genome Variation Society Nomenclature

Since the HGVS Nomenclature was first described in 2000, it has been widely adopted as the human-readable standard for genetic variation. The HGVS Nomenclature aims to remove ambiguity in variant descriptions to improve variant reporting in databases, literature, and genetic test reports. The nomenclature defines detailed rules for describing variants on DNA level (genomic and transcript), RNA level, and protein level. Recommendations include complex cases such as RNA fusion transcripts, chromosome translocations, and how to describe variants that have not been determined exactly down to the sequence level. The latest version of the HGVS Nomenclature can be found online at `hgvs-nomenclature.org`.

The basic structure of an HGVS variant description is (reference sequence):(numbering scheme).(variant). Reference sequence identifiers should always include version numbers where available, and the numbering scheme indicates the type of reference sequence used (e.g., "g." or "c."). For example, `NC_000015.9:g.40699841G>C` describes a substitution of a G to C at genomic position 40699841 of reference sequence `NC_000015.9`. The numbering schemes that are allowed depend on the reference sequence given. For an overview of the HGVS numbering schemes, see Table 2.3. For an overview of the most common variant type descriptions, see Table 2.4. Note that we will not go into detail here, like how to describe variants relative to a coding DNA reference sequence in 5′ or 3′ UTRs, exons and introns, or on the protein level. For this, please consult the HGVS variant nomenclature website.

Table 2.3: **Overview of the HGVS numbering schemes.**

Indicator	Usage	Example
g.	Genomic, non-circular reference sequences. Counting starts at the first nucleotide.	`NC_000015.9:g.40699841G>C`
m.	Genomic, circular reference sequences. Counting starts at the first nucleotide.	`NC_012920.1:m.3243A>G`
n.	Non-coding transcript reference sequences. Counting starts at the first nucleotide.	`NR_002725.2:n.1725_1726insA`
c.	Coding transcript reference sequences. Counting starts at the first nucleotide of the translation initiation codon (ATG).	`NM_002225.3:c.158G>C`
r.	RNA reference sequences. Counting starts at the first nucleotide for non-coding RNA, and at the first nucleotide of the translation initiation codon (AUG) for coding RNA.	`NM_002225.3:r.154_243del`
p.	Protein reference sequences. Counting starts at the first amino acid.	`NP_002216.2:p.Leu52_Arg81del`

Using the HGVS Nomenclature, there is, in principle, only one correct way to describe a variant. To remove ambiguity in the description of variants in repeated sequences, the HGVS Nomenclature uses the so-called 3′ rule, defining that any variant should be described by its most 3′ position possible. If a stretch of nucleotides is shortened by one, the 3′ rule states that the variant is described as if the last nucleotide has been deleted. In addition, HGVS Nomenclature uses strict definitions per variant type as well as prioritization rules when several options would be possible. For instance, prioritization defines a T to A change as a substitution, not an inversion or a deletion-insertion.

Unfortunately, the HGVS Nomenclature guidelines are not used without error. Frequently observed errors include not applying the 3′ rule and incorrectly describing duplications as insertions.[13] Both partially derive from NGS pipelines where deletions and insertions are mostly 5′ aligned, and duplications cannot be defined in the most commonly used file format. Fortunately, several computational tools exist that can help describe variants correctly: the hgvs software package[24] for direct integration into bioinformatics projects written in the Python programming language, and the online Mutalyzer[26] and VariantValidator[25] tools,

Table 2.4: **Examples of the most common DNA variant descriptions using the HGVS Nomenclature and the VCF file format.** Note that the HGVS rules for variants on DNA level are different from the rules for variants on RNA and protein levels.

Variant type	Sequence	HGVS description	VCF file format[a]		
Substitution	`AACGTT` `AACCTT`	`g.4G>C`	`4`	`G`	`C`
Deletion	`AACGTT` `AAC-TT`	`g.4del`	`3`	`CG`	`C`
Insertion	`AAC-GTT` `AACAGTT`	`g.3_4insA`	`3`	`C`	`CA`
Duplication	`AAC-GTT` `AACCGTT`	`g.3dup`	`3`	`C`	`CC`
Deletion-Insertion	`AACGTT` `AATATT`	`g.3_4delinsTA`	`3`	`CG`	`TA`
Inversion	`AACGTT` `ACGTTT`	`g.2_4inv`	`2`	`ACG`	`CGT`

[a] This column shows the POS, REF, and ALT fields respectively. Other fields like the CHROM field (storing the chromosome) and the ID field (where a possible dbSNP identifier can be stored) are removed for simplification. Also, the most simple representation of the variant is chosen; in reality, the VCF file format can describe the same variant in many ways.

which both can also be installed locally. The latter two tools provide a website interface for verifying variants one by one, a batch interface to verify a file with variants, and online Application Programming Interfaces (APIs), online interfaces allowing software to communicate with these online tools.

Although the HGVS Nomenclature is comprehensive, with mostly one valid description for each variant, not all areas have yet been covered in great detail. For instance, changing the sequence ACTG to TC can, following the current recommendations, be described as a deletion-insertion event of the entire sequence (`1_4delinsTC`) or as a substitution of the first base followed by a deletion of the last two bases (`[1A>T;3_4del]`). Although describing this change as one variant seems obvious, variant callers used in NGS software pipelines often choose the latter and define two single variants. When data are then shared without allelic information, it is no longer clear whether these two variants are in *cis* or *trans*, and they can no longer be merged into one variant. In addition, when the consequences of such variants on protein level are reported, serious errors may occur.[27] When encountering two closely spaced variants, checking whether they are on the same allele is recommended. If so, check for the combination in external sources like population frequency and gene variant databases.

2.6 Variant classification

To get to a clinical classification of a variant, i.e., draw a conclusion regarding the effect of the variant on an individual's health, one has to combine all available knowledge. The available knowledge has two major components: all observations of the variant in individuals with or without the associated phenotype and the interpretation of the variant's (predicted) consequences for the gene's function (functional or molecular classification). The HGVS recommends clearly separating the functional classification from the clinical classification to discriminate between a variant's effect on a gene's function and its consequences for the individual carrying the variant.

2.6.1 Functional classification

Functional classification of a certain variant can only be done in an animal model or by performing a functional assay, where the gene's function with the variant is compared with that of the wild-type form of the gene. A very simple, semi-functional assay which is mostly neglected is the analysis of an RNA sample from the patient. This analysis can provide valuable insights into the possible effects of a variant on RNA processing, especially splicing, and, therefore, the probable effect of the variant on the gene's function. Actual functional assays are often complex and costly to be performed. Firstly, a clear idea of the function of a gene is required. Secondly, cell types must be available where the gene is expressed and the consequences of variants can be measured. Functional assays have been developed for several relatively common diseases, e.g., breast cancer[28] or colon cancer.[29] These assays can be applied when a new variant is encountered, and no data from other studies are available. Although functional assays cannot directly demonstrate the consequences of a patient's variant, they do help provide evidence for a weighted clinical classification.

For functional classifications, there is currently no standard that is broadly followed among different areas of research. Assays measuring the function of genes affected by certain variants commonly use relative efficiency, indicated in a percentage relative to the wild-type gene.[28,30] For pharmacogenetics, a four-class system is now standard, describing the efficiency of drug metabolism: "Poor", "Intermediate", "Extensive/Normal", and "Ultrarapid".[31,32] The HGVS proposed a 5-tier functional classification, implemented by the LOVD databases: "Affects function", "Probably affects function", "Effect unknown", "Probably does not affect function", and "Does not affect function". It should be noted that "affects function" includes "improved function" (e.g., increased enzymatic activity), which may give a protective effect, a feature most classification systems are not able to cope with.

2.6.2 Clinical classification

Genome diagnostic laboratories and researchers have broadly accepted the use of a standard, 5-tier scheme for classifying variants.[33] "Benign", "Likely benign", "Variant of uncertain significance (VUS)", "Likely pathogenic", and "Pathogenic", also described as Class 1 through 5, respectively. Although this system standardized the naming of the different variant classifications, it did not cover the required evidence to get a variant classified in each category. This issue was tackled by the ACMG/AMP classification guidelines, published in 2015,[34] giving detailed recommendations on how to build up the evidence to classify a variant into one of these five categories. The recommendations clearly fulfilled a need and were quickly adopted, greatly improving the comparability of classifications made by laboratories worldwide. Several improvements to the ACMG/AMP guidelines have been published since, as some of the original specifications were open to different interpretations. Additional modifications were sometimes required to apply the guidelines for certain genes or diseases.[35–38] When clinically classifying variants, it is highly recommended to use the ACMG/AMP guidelines.

Although a one-on-one relationship between a functional and a clinical classification seems obvious, many exceptions exist. It is clear that when a variant does not alter a gene's function, the individual's health will not be affected, either. The opposite, however, is not always true. Whether a "pathogenic" variant will cause a disease first depends on the mode of inheritance, including dominant and recessive, autosomal and X/Y-linked, mitochondrial, imprinting, etc. In some cases, variants are "pathogenic" only in a compound heterozygous state, while they have no effect in a homozygous state. A variant's effect on a gene's function is also rarely "all or nothing". In many cases, the function is decreased, but not to zero. In pharmacogenomics, variants are cataloged that increase or decrease enzyme function, thereby affecting the level at which an individual is able to metabolize chemicals (medicine) and, therefore, how effective a certain drug dosage is or whether the drug is effective at all. A variant increasing enzyme activity, e.g., a whole gene duplication, is "protective" (compensating) when combined with a variant that decreases enzyme activity. The same variant should, however, be considered as "deleterious" in a homozygous state when the enzyme has to metabolize the drug to generate the active substance or when it removes the substance from the body too fast.

Another reason for discordance between functional and clinical classifications of a variant is penetrance. A variant may clearly affect the function of a gene but may cause disease only in a subset of the individuals in which the change has been found. One example is variants in the *BRCA1* or *BRCA2* gene, each increasing the risk of developing breast cancer before a certain age, yet some with a much higher risk than others, e.g., the low penetrant *BRCA1* `NP_009225.1:p.Arg1699Gln` variant.[39] A recent study[20] shows that when clearly pathogenic variants with a reduced penetrance are classified by different labs, some will classify them as Class 5 (Pathogenic) while others classify them as Class 3 (VUS). Another example is the many non-disease phenotypes, including eye color, the ability to taste bitter, and blood group types. While any variant in the ABO gene would be classified as benign (Class 1), the variant is clearly of medical relevance when the individual needs a blood transfusion. Another problem is the gray zone between a disease and a trait, where the same variant for a disease would be clinically classified as a Class 5 but for a trait as a Class 1.

There are currently no guidelines for dealing with clinical classifications and decreased penetrance. However, the ClinGen project established a Low Penetrance/Risk Allele working group that has begun investigating researchers' and clinicians' opinions on the matter, and the ENIGMA consortium that classifies breast cancer variants published recommendations[40] to always report variant-associated risks as absolute measurements. To make overall classification more informative, the "Global Variome shared LOVD"[41] has started to work with classifications including "pathogenic (dominant)", "pathogenic (recessive)", "pathogenic (maternal)", "pathogenic (paternal)", "pathogenic (!)", "association", "VUS", "VUS (!)", "benign (dominant)", "benign (recessive)", "benign (!)", etc. "Dominant", "recessive", "maternal", and "paternal" are used to indicate the mode of inheritance. The "!" is used to warn for exceptional features like reduced penetrance, protective variants, pathogenic but not in a homozygous state, etc. "Benign (dominant)" and "benign (recessive)" are used to indicate associations with non-disease phenotypes.

2.7 Standards on reporting disorders and phenotypes

Describing and classifying a variant is only relevant in the context of a particular phenotype (disease). As such, standards for describing phenotypes are equally important as those for describing genetic variants. A clear description of the characteristic features observed in the individual investigated following a standardized ontology is crucial for elucidating gene-phenotype relationships, developing gene panels, and recruiting patients for clinical trials. Most unresolved genetic diseases remaining nowadays derive from rare to ultra-rare cases, with only a few patients known worldwide. A critical component of establishing causative disease-gene links is always to identify more cases in which variants in a gene give a similar phenotype. The phenotype ontology used most widely is the Human Phenotype Ontology (HPO).[42] HPO was specifically developed to facilitate automated phenotype matching. For this, a nested tree structure was defined where deeper terms give more detail on each specific phenotypic feature. HPO is actively updated, and community efforts have been initiated to translate HPO terms into different languages, which is an important step in increasing its value further. One element of the very successful GeneMatcher initiative,[43] built to identify patients with similar gene-phenotype properties, is HPO-based phenotype matching. Finally, HPO allows phenotype matching across species, facilitating correlations between human disease and observations in animal models (e.g., mouse).

Another important standard is provided by OMIM (Online Mendelian Inheritance of Man),[44] providing standardized disease names and the description of the primary disease features observed. While HPO defines individually identifiable phenotypic features, OMIM focuses on disorders (diseases) in which these features are found in specific combinations. For instance, OMIM defines "Duchenne muscular dystrophy" as associated with pathogenic variants in the DMD gene, while HPO defines abnormalities such as "muscle weakness", "difficulty climbing stairs", and "scoliosis", which, among others, make up the complete phenotype of Duchenne muscular dystrophy. As such, OMIM is suitable for *diagnoses* while HPO is suitable for *anamneses* and automated phenotype matching.

Several tools are available for searching and collecting HPO terms, suggesting disorders that match the given terms. Examples are Phenomizer[45] and PhenoTips,[46] both web-based systems. When the underlying disorder is unknown, HPO terms can still be used for identifying genes for gene-panel-based exome data analysis like with PanelApp,[47] and for matching phenotypes when variants identified in patients are also found in external databases. When a diagnosis is possible, adding a disease identifier from OMIM supports submission to external databases such as LOVD or ClinVar and facilitates disease-oriented queries.

2.8 Challenges and considerations

Although NGS has successfully been implemented into the clinical workflow and the technology has since continuously been improved, challenges remain, and there are important caveats to consider. There are both technical and biological reasons for false-negative and false-positive results. It is important to keep these in mind when analyzing NGS data.

The most apparent difference between analyzing data from NGS studies and techniques such as Sanger sequencing is the sheer size of the region in the genome covered. Naturally, technology has to do part of the data analysis for us, as it's not humanly possible to check every sequence read that is generated. However, there is a major difference in how results are presented, which is often overlooked. With Sanger sequencing, not having a proper sequence returned would simply mean a failed analysis. With today's NGS results, a lack of variants found in any genomic region may mean there are no variants, the pipeline is unable to determine whether there are any variants, or the region is missing (deleted) in the sample.

There are several reasons why the sequencing pipeline software may generate a false-negative result. The most common reason for a false-negative result is the **lack of coverage**. Where not enough sequencing reads cover a certain region, the variant caller software can only produce low-confidence genotypes or will not produce genotypes at all. It is good practice to prevent false positives by only selecting good-quality variant calls for analysis and by putting a threshold on the minimum number of reads required to report a certain variant. During this step, it is often overlooked that the number of expected reads on the sex chromosomes in males is only half that in females, requiring flexible thresholds to prevent false negatives. A lack of coverage can also result from how the sequencing was performed, the quality of the sample, an incomplete or incorrect reference sequence, or reads mapping to multiple regions in the genome. Reads are often discarded when they map to multiple regions in the genome, like the telomeres and centromeres, (large) repetitive sequences, and segmental duplications. Finally, variants not present in all cells (mosaic variants), including heteroplasmic mitochondrial variants, are hard to detect because either the tissue sample sequenced does not harbor the variant or the allele fraction may be too low. When a variant is only represented in a small subset of reads, it is considered likely a sequencing error and ignored or called with only low confidence.[48]

Variants too large to be contained in one sequencing read are also often missed. **Large CNVs**, like whole exon or gene deletions and duplication, can be observed by a drop or a rise in the coverage in the affected region. Still, they are often left undetected by common variant calling methods as the reads spanning the variant breakpoints do not align well to the reference genome and end up discarded or are not present in the sample analyzed, as can be the case with whole-exome sequencing. Specialized tools have become available that specifically detect such large changes.[49,50] When using long-range single-molecule sequencing techniques, it is much easier to detect CNVs.

False-positive results can occur due to reads aligning to **pseudogenes or duplicated regions** such as the pseudoautosomal regions (PARs) on the X and Y chromosomes. It is not uncommon for variants to be detected on the Y chromosome in females or heterozygous variants to be detected on the X chromosome in males. Besides rare genetic disorders, a far more common cause is sequencing reads aligning to the wrong chromosome in these PARs. The same holds for pseudogenes; variants detected in a gene could very well be variants belonging to the homologous region in the pseudogene. Variants derived from gene conversion, where the sequence of a pseudogene gets copied to the normal gene, are especially problematic since all variant reads will map to the pseudogene. In such cases, designing gene-specific primers and confirming variants by Sanger sequencing is essential to rule out false positives.[48]

Another source for error is **non-normalized variant calling**. Due to variants not being normalized before annotation is loaded, a common variant with a high frequency in the population might not get annotated as such when it has been described differently. On the other hand, a causative variant might not immediately be recognized as such because its description doesn't match that of its record in gene variant databases. Also, NGS analysis pipelines often split deletion-insertion events into multiple variants, possibly causing similar problems.

2.9 Conclusions

This chapter introduced genetic variation and how NGS analyses measure it. We discussed the different types of variants on the DNA, RNA, and protein levels, their possible locations, and how they can influence a gene's function at many levels, including transcription, RNA processing, translation, protein processing, and protein modification.

We listed the relevant standards for describing, interpreting, and reporting genetic variants and phenotypes, the relationship and differences between these standards, explained their importance and current status of needed improvement, and mentioned some caveats to remember when describing, classifying, and reporting variants and phenotypes.

Finally, we explained which technical and biological factors can lead to false-negative and false-positive results and suggested some solutions to these problems.

2.10 References

[1] Aylwyn Scally. *The mutation rate in human evolution and demographic inference*. Current Opinion in Genetics and Development 2016; 41 36–43.

[2] Augustine Kong, Michael L. Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A. Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Wendy S.W. Wong, Gunnar Sigurdsson, G. Bragi Walters, Stacy Steinberg, Hannes Helgason, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Agnar Helgason, Olafur Th Magnusson, Unnur Thorsteinsdottir, and Kari Stefansson. *Rate of de novo mutations and the importance of father's age to disease risk*. Nature 2012; 488 (7412) 471–475.

[3] Laurent C. Francioli, Androniki Menelaou, Sara L. Pulit, Freerk van Dijk, Pier Francesco Palamara, Clara C. Elbers, Pieter B.T. Neerincx, Kai Ye, Victor Guryev, Wigard P. Kloosterman, Patrick Deelen, Abdel Abdellaoui, Elisabeth M. van Leeuwen, Mannis van Oven, Martijn Vermaat, Mingkun Li, Jeroen F.J. Laros, Lennart C. Karssen, Alexandros Kanterakis, Najaf Amin, Jouke Jan Hottenga, Eric Wubbo Lameijer, Mathijs Kattenberg, Martijn Dijkstra, Heorhiy Byelas, et al. *Whole-genome sequence variation, population structure and demographic history of the Dutch population*. Nature Genetics 2014; 46 (8) 818–825.

[4] Maristella Steri, M. Laura Idda, Michael B. Whalen, and Valeria Orrù. *Genetic variants in mRNA untranslated regions*. Wiley Interdisciplinary Reviews: RNA 2018; 9 (4) e1474.

[5] Maria De Angioletti, Giuseppina Lacerra, Vincenzo Sabato, and Clementina Carestia. *β+45 G $\rightarrow$ C: A novel silent β-thalassaemia mutation, the first in the Kozak sequence*. British Journal of Haematology 2004; 124 (2) 224–231.

[6] Annemieke Aartsma-Rus, Judith C.T. van Deutekom, Ivo F.A.C. Fokkema, Gert Jan B. van Ommen, and Johan T. den Dunnen. *Entries in the Leiden Duchenne muscular dystrophy mutation database: An overview of mutation types and paradoxical cases that confirm the reading-frame rule*. Muscle and Nerve 2006; 34 (2) 135–144.

[7] Franziska M. Gisler, Thomas Von Kanel, Richard Kraemer, André Schaller, and Sabina Gallati. *Identification of SNPs in the cystic fibrosis interactome influencing pulmonary progression in cystic fibrosis*. European Journal of Human Genetics 2013; 21 (4) 397–403.

[8] Anna Abramowicz and Monika Gos. *Splicing mutations in human genetic disorders: examples, detection, and confirmation*. Journal of Applied Genetics 2018; 59 (3) 253–268.

[9] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor*. Bioinformatics 2010; 26 (16) 2069–2070.

[10] Enza Di Leo, Francesca Panico, Patrizia Tarugi, Carla Battisti, Antonio Federico, and Sebastiano Calandra. *A point mutation in the lariat branch point of intron 6 of NPC1 as the cause of abnormal pre-mRNA splicing in Niemann-Pick type C disease*. Human Mutation 2004; 24 (5) 440.

[11] Huda Y. Zoghbi and Arthur L. Beaudet. *Epigenetics and Human Disease*. Cold Spring Harbor Perspectives in Biology 2016; 8 (2) a019497.

[12] Che-Pei Kung, Leonard B. Maggi, and Jason D. Weber. *The Role of RNA Editing in Cancer Development and Metabolic Disorders*. Frontiers in Endocrinology 2018; 9.

[13] Zandra C. Deans, Jennifer A. Fairley, Johan T. den Dunnen, and Caroline Clark. *HGVS Nomenclature in Practice: An Example from the United Kingdom National External Quality Assessment Scheme*. Human Mutation 2016; 37 (6) 576–578.

[14] Bryony Braschi, Paul Denny, Kristian Gray, Tamsin Jones, Ruth Seal, Susan Tweedie, Bethan Yates, and Elspeth Bruford. *Genenames.org: The HGNC and VGNC resources in 2019*. Nucleic Acids Research 2019; 47 (D1) D786–D792.

[15] Eric W. Sayers, Mark Cavanaugh, Karen Clark, James Ostell, Kim D. Pruitt, and Ilene Karsch-Mizrachi. *GenBank*. Nucleic Acids Research 2019; 47 (D1) D94–D99.

[16] Andrew D. Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Andrey G. Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, José Carlos Marugán, Carla Cummins, Claire Davidson, Kamalkumar Dodiya, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, et al. *Ensembl 2020*. Nucleic Acids Research 2020; 48 (D1) D682–D688.

[17] Jacqueline A.L. MacArthur, Joannella Morales, Ray E. Tully, Alex Astashyn, Laurent Gil, Elspeth A. Bruford, Pontus Larsson, Paul Flicek, Raymond Dalgleish, Donna R. Maglott, and Fiona Cunningham. *Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants*. Nucleic Acids Research 2014; 42 (Database issue) D873–D878.

[18] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. *The variant call format and VCFtools*. Bioinformatics 2011; 27 (15) 2156–2158.

[19] Johan T. den Dunnen, Raymond Dalgleish, Donna R. Maglott, Reece K. Hart, Marc S. Greenblatt, Jean McGowan-Jordan, Anne-Francoise Roux, Timothy Smith, Stylianos E. Antonarakis, and Peter E.M. Taschner. *HGVS Recommendations for the Description of Sequence Variants: 2016 Update*. Human Mutation 2016; 37 (6) 564–569.

[20] Ivo F.A.C. Fokkema, Kasper J. van der Velde, Mariska K. Slofstra, Claudia A.L. Ruivenkamp, Maartje J. Vogel, Rolph Pfundt, Marinus J. Blok, Ronald H. Lekanne Deprez, Quinten Waisfisz, Kristin M. Abbott, Richard J. Sinke, Rubayte Rahman, Isaäc J. Nijman, Bart de Koning, Gert Thijs, Nienke Wieskamp, Ruben J.G. Moritz, Bart Charbon, Jasper J. Saris, Johan T. den Dunnen, Jeroen F.J. Laros, Morris A. Swertz, and Marielle E. van Gijn. *Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data*. Human Mutation 2019; 40 (12) 2230–2238.

[21] John G. Cleary, Ross Braithwaite, Kurt Gaastra, Brian S. Hilbush, Stuart Inglis, Sean A. Irvine, Alan Jackson, Richard Littin, Mehul Rathod, David Ware, Justin M. Zook, Len Trigg, and Francisco M.M. De La Vega. *Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines*. bioRxiv 2015; 023754.

[22] Adrian Tan, Gonçalo R. Abecasis, and Hyun Min Kang. *Unified representation of genetic variants*. Bioinformatics 2015; 31 (13) 2202–2204.

[23] Arash Bayat, Bruno Gaëta, Aleksandar Ignjatovic, and Sri Parameswaran. *Improved VCF normalization for accurate VCF comparison*. Bioinformatics 2017; 33 (7) 964–970.

[24] Meng Wang, Keith M. Callenberg, Raymond Dalgleish, Alexandre Fedtsov, Naomi K. Fox, Peter J. Freeman, Kevin B. Jacobs, Piotr Kaleta, Andrew J. McMurry, Andreas Prlić, Veena Rajaraman, and Reece K. Hart. *hgvs: A Python package for manipulating sequence variants using HGVS nomenclature: 2018 Update*. Human Mutation 2018; 39 (12) 1803–1813.

[25] Peter J. Freeman, Reece K. Hart, Liam J. Gretton, Anthony J. Brookes, and Raymond Dalgleish. *VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions*. Human Mutation 2018; 39 (1) 61–68.

[26] Martin Wildeman, Ernest van Ophuizen, Johan T. den Dunnen, and Peter E.M. Taschner. *Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker*. Human Mutation 2008; 29 (1) 6–13.

[27] William Spooner, William McLaren, Timothy Slidel, Donna K. Finch, Robin Butler, Jamie Campbell, Laura Eghobamien, David Rider, Christine Mione Kiefer, Matthew J. Robinson, Colin Hardman, Fiona Cunningham, Tristan Vaughan, Paul Flicek, and Catherine Chaillan Huntington. *Haplosaurus computes protein haplotypes for use in precision drug design*. Nature Communications 2018; 9 (1) 4128.

[28] Gaël A. Millot, Marcelo A. Carvalho, Sandrine M. Caputo, Maaike P.G. Vreeswijk, Melissa A. Brown, Michelle Webb, Etienne Rouleau, Susan L. Neuhausen, Thomas V.O. Hansen, Alvaro Galli, Rita D. Brandão, Marinus J. Blok, Aneliya Velkova, Fergus J. Couch, and Alvaro N.A. Monteiro. *A guide for functional analysis of BRCA1 variants of uncertain significance*. Human Mutation 2012; 33 (11) 1526–1537.

[29] Mark Drost, Yvonne Tiersma, Bryony A. Thompson, Jane H. Frederiksen, Guido Keijzers, Dylan Glubb, Scott Kathe, Jan Osinga, Helga Westers, Lisa Pappas, Kenneth M. Boucher, Siska Molenkamp, José B. Zonneveld, Christi J. van Asperen, David E. Goldgar, Susan S. Wallace, Rolf H. Sijmons, Amanda B. Spurdle, Lene J. Rasmussen, Marc S. Greenblatt, Niels de Wind, and Sean V. Tavtigian. *A functional assay–based procedure to classify mismatch repair gene variants in Lynch syndrome*. Genetics in Medicine 2019; 21 (7) 1486–1496.

[30] Rick A.C.M. Boonen, Amélie Rodrigue, Chantal Stoepker, Wouter W. Wiegant, Bas Vroling, Milan Sharma, Magdalena B. Rother, Nandi Celosse, Maaike P.G. Vreeswijk, Fergus Couch, Jacques Simard, Peter Devilee, Jean Yves Masson, and Haico van Attikum. *Functional analysis of genetic variants in the high-risk breast cancer susceptibility gene PALB2*. Nature Communications 2019; 10 (1) 1–15.

[31] Ann K. Daly. *Pharmacogenetics: a general review on progress to date*. British Medical Bulletin 2017; 124 (1) 65–79.

[32] Frank R. Wendt, Antti Sajantila, Rodrigo S. Moura-Neto, August E. Woerner, and Bruce Budowle. *Full-gene haplotypes refine CYP2D6 metabolizer phenotype inferences*. International Journal of Legal Medicine 2018; 132 (4) 1007–1024.

[33] Sharon E. Plon, Diana M. Eccles, Douglas Easton, William D. Foulkes, Maurizio Genuardi, Marc S. Greenblatt, Frans B.L. Hogervorst, Nicoline Hoogerbrugge, Amanda B. Spurdle, and Sean V. Tavtigian. *Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results*. Human Mutation 2008; 29 (11) 1282–1291.

[34] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. *Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*. Genetics in Medicine 2015; 17 (5) 405–424.

[35] Keith Nykamp, Michael Anderson, Martin Powers, John Garcia, Blanca Herrera, Yuan Yuan Ho, Yuya Kobayashi, Nila Patil, Janita Thusberg, Marjorie Westbrook, and Scott Topper. *Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria*. Genetics in Medicine 2017; 19 (10) 1105–1117.

[36] Bruce D. Gelb, Helene Cave, Mitchell W. Dillon, Karen W. Gripp, Jennifer A. Lee, Heather Mason-Suares, Katherine A. Rauen, Bradley Williams, Martin Zenker, and Lisa M. Vincent. *ClinGen's RASopathy Expert Panel consensus methods for variant interpretation*. Genetics in Medicine 2018; 20 (11) 1334–1345.

[37] Melissa A. Kelly, Colleen Caleshu, Ana Morales, Jillian Buchan, Zena Wolf, Steven M. Harrison, Stuart Cook, Mitchell W. Dillon, John Garcia, Eden Haverfield, Jan D.H. Jongbloed, Daniela Macaya, Arjun Manrai, Kate Orland, Gabriele Richard, Katherine Spoonamore, Matthew Thomas, Kate Thomson, Lisa M. Vincent, Roddy Walsh, Hugh Watkins, Nicola Whiffin, Jodie Ingles, J. Peter van Tintelen, Christopher Semsarian, et al. *Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel*. Genetics In Medicine 2018; 20 (3) 351–359.

[38] Pauline Romanet, Marie-Francoise Odou, Marie-Odile North, Alexandru Saveanu, Lucie Coppin, Eric Pasmant, Amira Mohamed, Pierre Goudet, Francoise Borson-Chazot, Alain Calender, Christophe Beroud, Nicolas Levy, Sophie Giraud, and Anne Barlier. *Proposition of adjustments to the ACMG-AMP framework for the interpretation of MEN1 missense variants*. Human Mutation 2019; 40 (6) 661–674.

[39] Setareh Moghadasi, Huong D. Meeks, Maaike P.G. Vreeswijk, Linda A.M. Janssen, Åke Borg, Hans Ehrencrona, Ylva Paulsson-Karlsson, Barbara Wappenschmidt, Christoph Engel, Andrea Gehrig, Norbert Arnold, Thomas van Overeem Hansen, Mads Thomassen, Uffe Birk Jensen, Torben A. Kruse, Bent Ejlertsen, Anne Marie Gerdes, Inge Søkilde Pedersen, Sandrine M. Caputo, Fergus Couch, Emily J. Hallberg, Ans M.W. van den Ouweland, Margriet J. Collée, Erik Teugels, Muriel A. Adank, et al. *The BRCA1 c.5096G>A p.Arg1699Gln (R1699Q) intermediate risk variant: Breast and ovarian cancer risk estimation and recommendations for clinical management from the ENIGMA consortium*. Journal of Medical Genetics 2018; 55 (1) 15–20.

[40] Amanda B. Spurdle, Stephanie Greville-Heygate, Antonis C. Antoniou, Melissa Brown, Leslie Burke, Miguel de la Hoya, Susan Domchek, Thilo Dörk, Helen V. Firth, Alvaro N. Monteiro, Arjen Mensenkamp, Michael T. Parsons, Paolo Radice, Mark Robson, Marc Tischkowitz, Emma Tudini, Clare Turnbull, Maaike P.G. Vreeswijk, Logan C. Walker, Sean Tavtigian, and Diana M. Eccles. *Towards controlled terminology for reporting germline cancer susceptibility variants: an ENIGMA report*. Journal of Medical Genetics 2019; 56 (6) 347–357.

[41] Ivo F.A.C. Fokkema, Peter E.M. Taschner, Gerard C.P. Schaafsma, J. Celli, Jeroen F.J. Laros, and Johan T. den Dunnen. *LOVD v.2.0: the next generation in gene variant databases*. Human Mutation 2011; 32 (5) 557–563.

[42] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O.B. Jacobsen, Daniel Danis, Jean Philippe Gourdine, Michael Gargano, Nomi L. Harris, Nicolas Matentzoglu, Julie A. McMurry, David Osumi-Sutherland, Valentina Cipriani, James P. Balhoff, Tom Conlin, Hannah Blau, Gareth Baynam, Richard Palmer, Dylan Gratian, Hugh Dawkins, Michael Segal, Anna C. Jansen, Ahmed Muaz, Willie H. Chang, Jenna Bergerson, Stanley J.F. Laulederkind, et al. *Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources*. Nucleic Acids Research 2019; 47 (D1) D1018–D1027.

[43] Ange Line Bruel, Antonio Vitobello, Frédéric Tran Mau-Them, Sophie Nambot, Yannis Duffourd, Virginie Quéré, Paul Kuentz, Philippine Garret, Julien Thevenon, Sébastien Moutton, Daphné Lehalle, Nolwenn Jean-Marçais, Aurore Garde, Julian Delanne, Mathilde Lefebvre, François Lecoquierre, Detlef Trost, Megan Cho, Amber Begtrup, Aida Telegrafi, Pierre Vabres, Anne Laure Mosca-Boidron, Patrick Callier, Christophe Philippe, Laurence Faivre, et al. *2.5 years' experience of GeneMatcher data-*

sharing: a powerful tool for identifying new genes responsible for rare diseases. Genetics in Medicine 2019; 21 (7) 1657–1661.

[44] Joanna S. Amberger, Carol A. Bocchini, Alan F. Scott, and Ada Hamosh. *OMIM.org: leveraging knowledge across phenotype–gene relationships*. Nucleic Acids Research 2019; 47 (D1) D1038–D1043.

[45] Sebastian Köhler, Marcel H. Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E. Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N. Robinson. *Clinical diagnostics in human genetics with semantic similarity searches in ontologies*. American Journal of Human Genetics 2009; 85 (4) 457–464.

[46] Marta Girdea, Sergiu Dumitriu, Marc Fiume, Sarah Bowdin, Kym M. Boycott, Sébastien Chénier, David Chitayat, Hanna Faghfoury, M. Stephen Meyn, Peter N. Ray, Joyce So, Dimitri J. Stavropoulos, and Michael Brudno. *PhenoTips: Patient Phenotyping Software for Clinical and Research Use*. Human Mutation 2013; 34 (8) 1057–1065.

[47] Antonio Rueda Martin, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh, Louise C. Daugherty, Olivia Niblock, Ivone U.S. Leong, Katherine R. Smith, Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma Baple, Arianna Tucci, Helen Brittain, Anna de Burca, Kristina Ibañez, Dalia Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M. McDonagh. *PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels*. Nature Genetics 2019; 51 (11) 1560–1565.

[48] Avni Santani, Jill Murrell, Birgit Funke, Zhenming Yu, Madhuri Hegde, Rong Mao, Andrea Ferreira-Gonzalez, Karl V. Voelkerding, and Karen E. Weck. *Development and Validation of Targeted Next-Generation Sequencing Panels for Detection of Germline Variants in Inherited Diseases*. Archives of Pathology and Laboratory Medicine 2017; 141 (6) 787–797.

[49] Jayne Y. Hehir-Kwa, Bastiaan B.J. Tops, and Patrick Kemmeren. *The clinical implementation of copy number detection in the age of next-generation sequencing*. Expert Review of Molecular Diagnostics 2018; 18 (10) 907–915.

[50] Iria Roca, Lorena González-Castro, Helena Fernández, Mª Luz Couce, and Ana Fernández-Marmiesse. *Free-access copy-number variant detection tools for targeted next-generation sequencing data*. Mutation Research / Reviews in Mutation Research 2019; 779 114–125.