





# Clinical exome sequencing—Mistakes and caveats

Jordi Corominas<sup>1</sup>  | Sanne P. Smeeckens<sup>1</sup> | Marcel R. Nelen<sup>1</sup> |  
Helger G. Yntema<sup>1,2</sup> | Erik-Jan Kamsteeg<sup>1,2</sup>  | Rolph Pfundt<sup>1,2</sup>  |  
Christian Gilissen<sup>3</sup> 

<sup>1</sup>Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands

<sup>2</sup>Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, the Netherlands

<sup>3</sup>Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, the Netherlands

## Correspondence

Christian Gilissen, Department of Human Genetics, Radboud University Medical Center, Geert Grooteplein 10 Nijmegen, GA 6525, the Netherlands.

Email: [christian.gilissen@radboudumc.nl](mailto:christian.gilissen@radboudumc.nl)

## Funding information

Horizon 2020 Framework Programme, Grant/Award Number: 779257; Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 917-17-353

## Abstract

Massive parallel sequencing technology has become the predominant technique for genetic diagnostics and research. Many genetic laboratories have wrestled with the challenges of setting up genetic testing workflows based on a completely new technology. The learning curve we went through as a laboratory was accompanied by growing pains while we gained new knowledge and expertise. Here we discuss some important mistakes that have been made in our laboratory through 10 years of clinical exome sequencing but that have given us important new insights on how to adapt our working methods. We provide these examples and the lessons that we learned to help other laboratories avoid to make the same mistakes.

## KEYWORDS

clinical exome, clinical variant interpretation, genetic diagnostics, next generation sequencing, NGS data analysis, whole exome sequencing

## 1 | INTRODUCTION

Massively parallel sequencing technology, or next-generation sequencing (NGS), has become the standard technique for genetic diagnostics and research. Especially exome and genome sequencing are now applied worldwide to molecularly diagnose patients (Hartman et al., 2019; Marshall et al., 2020; Matthijs et al., 2016). Over the last few years many laboratories have wrestled with the challenges of setting up genetic testing workflows based on a completely new technology. These challenges have been amplified by the fact that sequencing technology has been evolving ever since its introduction with novel instruments, chemistry and analysis methods.

Throughout the past decade, new sequencing technologies have come to market, whereas others have disappeared, and all of them have undergone rapid changes and upgrades (Giani et al., 2020; Heather & Chain, 2016). The same holds true for exome capture kits (Zhou et al., 2021), concomitant equipment and consumables. In this continuously changing field, laboratories have strived to consistently generate high-quality sequencing data. Various studies have reported how biases in sequencing data may result in either reduced sensitivity or false positive variants for exome and genome sequencing. For example, with NGS, high sequencing error rates and polymerase chain reaction (PCR) duplicates will result in potential false-positive calls whereas nonuniform sequence coverage or lack of coverage may lead to reduced sensitivity (Barbitoff et al., 2020; Lelieveld et al., 2015).

Jordi Corominas, Sanne P. Smeeckens, Jan Kamsteeg, Rolph Pfundt, and Christian Gilissen are contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Human Mutation* published by Wiley Periodicals LLC

Other issues such as strand bias and insert size distribution may also adversely affect sequencing results (Y. Guo, Li, et al., 2012).

NGS technology is also much more data-intensive than traditional genetic testing approaches and requires expertise in information technology (IT) and bioinformatics which was initially scarce in many laboratories. Bioinformatics has dealt with the difficulty of setting up rigorous quality control for sequencing data, but also with the challenge of reliable variant identification from sequencing data (Y. Guo et al., 2014). For example, it is relatively difficult to detect insertions and deletions, to identify variants in repeat-rich or low coverage regions (Jiang et al., 2015; Weißbach et al., 2021), or to distinguish single nucleotide variants (SNVs) from sequencing errors and mapping artifacts. Additionally, the detection of copy number variants (CNVs) from exome data has become a standard procedure giving rise to its own specific challenges (Hong et al., 2016). Similarly, as with sequencing instruments, bioinformatics needs to handle continuous updates from software tools, gene panels and other annotation resources to ensure that molecular geneticists have the latest information available for up-to-date data interpretation (Lelieveld et al., 2016). This in turn requires laboratories to implement strategies for automated testing of their analyses as well as systematic approaches for the reanalysis of existing data (Fung et al., 2020; Liu et al., 2019).

Driven by the new sequencing possibilities and the genetic and phenotypic variability of many diseases, clinical genetic testing has changed drastically in the last decade. From targeted gene testing where only one or a few genes would be sequenced based on the clinical phenotype, genetic requests now often concern the analysis of large panels of disease genes. Compared to single gene analysis, the interpretation of the large number of variants from exome or genome sequencing is obviously quite different. This requires not only in-depth knowledge of the technique to assess the quality of data and identified variants, but also new approaches for variant interpretation. Initial reporting of NGS variants was sometimes too stringent whereby variants that did not exactly match the patient phenotype were omitted, or too lenient, giving rise to many variants of uncertain significance (VUS) (Brownstein et al., 2014; Richards et al., 2015). Over time the quality of the sequencing data has greatly improved and the development of large publicly available databases with variant frequencies, such as the GnomAD database (Karczewski et al., 2020), have helped greatly in the development of more efficient variant filtering options. Moreover, in the last few years various recommendations and quality assessment schemes have been developed that guide the interpretation, classification and reporting of NGS variants (MacArthur et al., 2014; Miller et al., 2021; Rehm et al., 2015; Richards et al., 2015).

There are now several guidelines available on NGS testing, including concrete instructions from the College of American Pathologists (CAP) that can aid in the design, optimization, validation, quality management and bioinformatic aspects of NGS testing (Santani et al., 2019). Nevertheless, many challenges remain and mistakes are bound to happen, even in regulated clinical genetic testing laboratories where quality is of foremost importance. Here we

show examples of some of the mistakes that were made in our laboratory throughout 10 years of clinical exome sequencing and the lessons we learned from these mistakes (Table S1). Whereas the wet-lab has its own particular challenges, here we focus mostly on the issues related to data analysis and variant interpretation. We hope that by sharing these examples other laboratories are safeguarded from making the same mistakes.

## 2 | DATA ANALYSIS

Data management and the development of analysis pipelines for sequencing data have become important for many diagnostic laboratories. Building a complete, efficient and robust NGS analysis pipeline is an elaborate task that includes multiple delicate steps from alignment of NGS reads to calling and annotation of different types of genetic variation, such as SNVs, small insertions and deletions, CNVs and short tandem repeats (STRs). Because of the many different processing steps that need to be carried out and the large amount of data, it is relatively easy to make a small mistake with a large but nonobvious impact on the final results. Here we show five examples of mistakes that we made throughout the process of data analysis and that have so far not been abundantly highlighted in literature.

### 2.1 | Sequence quality

“Garbage in, garbage out” is a well-known saying in computer science that captures the concept that flawed input data produces flawed output or “garbage.” The same applies to sequencing data. Our laboratory encountered many issues with sequencing results that were not due to mistakes in the processing of the data, but rather due to the fact that there were issues with the initial data generation itself. Identifying the underlying cause of downstream issues can be a challenging task because subtle quality issues in the sequencing data can have large effects on subsequent variant calling. A relatively common issue is data with many spurious variant calls. This happened on occasion due to an unexpected high sequencing error rate, sample contamination, or due to incorrect trimming of adapter sequences (Figure S1). Most of these quality issues can be recognized by inspecting the raw sequencing data or by the observation that called variants have low-quality scores and deviate from the expected allele fraction of 50% for heterozygous calls. The opposite, a reduced number of variant calls is in most cases due to low sequence coverage. However, there may also be other reasons for reduced sensitivity. In two batches of exome sequencing samples, we noticed a lower number of variant calls only because we performed a trend analysis across several batches of samples. Initially, we expected this to be due to lower sequence coverage of the samples (Figure S2). However, the sequence coverage for these samples was not different from that of other samples. Eventually we discovered that this problem was due to a

10%–20% increase of the fraction of duplicate reads. Because duplicate reads could be due to PCR amplification, and potentially introduce false-positive variant calls, most variant callers will not consider them for variant calling. Therefore, the effective coverage for many regions was 10%–20% lower than what it appeared in these two batches (Figure S2).

Many quality issues can be readily identified by using tools such as Qualimap that compute quality statistics for sequencing experiments, such as coverage statistics, sequencing error rate, and the percentage of duplicate reads (García-Alcalde et al., 2012). Therefore, we strongly recommend to embed extensive quality control at all steps of the bioinformatic pipeline and follow trends of quality parameters such as percentage of duplicate reads, coverage distribution, overall number of variants called and percentage of rare variants not found in gnomAD. Deviations from expected values should be investigated closely. Establishing quality thresholds during development and testing will help to identify quality issues later on (Roy et al., 2018; Santani et al., 2019). These thresholds may need to be updated when laboratory protocols are changed, for example with the introduction of new sequencing instruments. A comprehensive quality control analysis on sequencing data can prevent many downstream issues with data interpretation.

## 2.2 | Sequence alignment: Alternate contigs

The most primary processing step in NGS data is the alignment of reads to a reference genome. The genome structure of particular regions may, however, vary considerably between different individuals and populations. To properly represent these loci the reference genome makes use of alternate contigs, that is, different reference sequences for particular regions in the genome. These alternate contigs contain regions in the genome that vary in such complex ways that they cannot be represented as a single reference sequence. In our initial analysis workflow we attempted to be as comprehensive as possible and included the largest possible reference genome, that included alternative contigs. However, most read mapping algorithms will, by default, assign a poor mapping quality score to reads that align equally well to multiple regions in the reference genome. These reads with mapping quality (MAPQ) equal to zero are typically indicated in the Integrated Genomics Viewer (IGV; Robinson et al., 2011) with blank reads (Figure 1a). Variant calling algorithms, in turn, will ignore such reads and will not identify variants in regions where reads have low MAPQ scores. Variants in such regions, although visible by manual inspection, will not be called. This mistake was identified with the help of laboratory specialists that looked into the aligned sequencing data to identify whether there was a potential second mutation in a recessive gene (see Section 3.3). We found that by including alternative contigs the number of coding bases where reads cannot be unambiguously aligned will triple. The same issue was recently reported for data from the UK Biobank where the introduction of an important number of alternate contigs in GRCh38 reference genome caused the absence of thousands of variants (Jia et al., 2020). There are two ways in which this

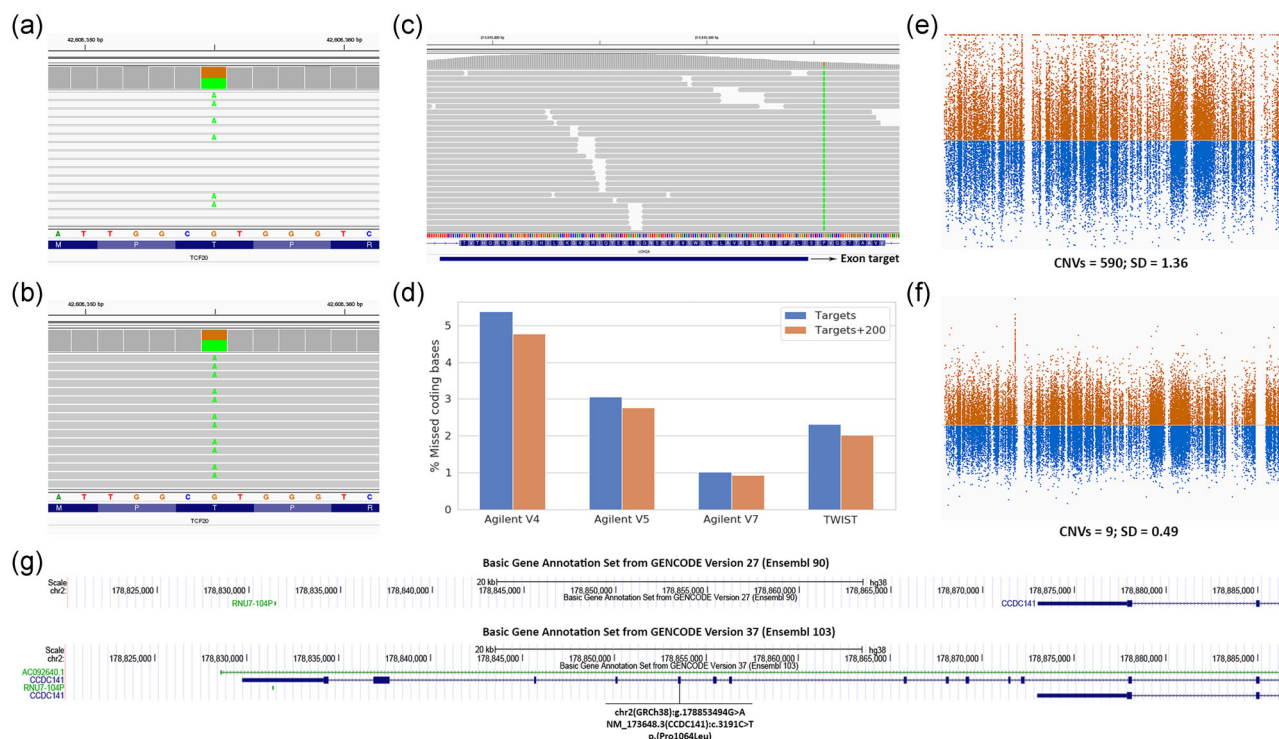
problem can be circumvented. The straightforward solution is to simply exclude alternative-contigs in the analysis, which is currently what is done in our own analysis for exomes on GRCh37. Analyzing the data without alternate contigs will properly align reads in the primary assembly of the human reference genome (Figure 1b). A more sophisticated solution is to apply alignment algorithms that can handle alternate contigs using the corresponding index file which we now do for genomes analyzed using the GRCh38 build of the reference genome (Jia et al., 2020). Considering that GRCh38 greatly expands the repertoire of alternative contigs (among other improvements), it would be advantageous for the clinical community to start transitioning towards GRCh38 to be able to properly detect and analyze genomic variation in population-specific haplotypes.

## 2.3 | Variant calling: Capture target file

There are many different exome kits available that all use their own definition of “regions of interest” (Lelieveld et al., 2015; Pengelly et al., 2020). The naive approach for calling variants from an exome would be to call genome-wide, without considering capture targets or coding regions. However, this is computationally burdensome, and the resulting data will contain many low-quality variant calls from off-target reads in regions that are not of interest. Therefore, it seems reasonable to restrict the analysis to regions where sufficient coverage for reliable variant calling can reasonably be expected. Although the original exome kits tried to exactly target the coding regions, many manufacturers started to move capture probes such that they would be partly overlapping or close to the exon of interest, to optimize the enrichment efficiency. The idea behind this is that a combination of the length of the sequence reads (typically 100–150 bp) and the enrichment of genomic DNA fragments extending beyond but overlapping the targets, will result in sufficient coverage not just for the capture target itself but also for the 100–150 adjacent bases. This indeed improves the capture efficiency for many “difficult” exons but makes it more difficult to decide in which regions variants should be called.

With our initial implementation of a new exome capture design we made the mistake of calling variants only in the exome capture targets, not realizing that a proportion of exons was not directly covered by any capture target, and thereby missing relevant coding variants (Figure 1c). Although we performed several quality checks when testing the exome kit, we did not immediately realize that we were missing as much as 5.4% (1897 kb) of all coding bases (Agilent SureSelect version 4). Again, this mistake was observed when variants that were visible in the sequence alignment through IGV were not present in the variant call files. In more recent exome kits the number of coding bases adjacent to a capture target is less but still considerable (Figure 1d).

Most manufacturers and studies guarantee sufficient coverage 100 bp adjacent to a capture target (Pengelly et al., 2020), but we currently extend our targets with 200 bp, balancing the additional compute time and additional variants called in coding regions.



**FIGURE 1** Issues that were encountered in data analysis. (a) IGV screenshot of sequence alignment for a pathogenic coding variant in the gene *TCF20* that was initially not detected because the sequencing reads align to multiple locations in the reference genome due to the inclusion of alternate contigs. (b) Reanalysis of the same sample while excluding alternate contigs led to unique alignments of the sequence reads and detection of this variant. (c) Example of a coding exon where a variant may be missed because the capture target (Agilent SureSelect v5) does not fully overlap with the exonic region. (d) Overview of the percentage of coding bases (Gencode Basic v.34lift37) that is not exactly within the capture targets, and within 200 bp vicinity of the capture targets, for different enrichment methods. (e) Normalized coverage of capture targets (Agilent v5) for an exome sample when using a heterogeneous reference cohort for CNV calling (CoNIFER). Information related to the number of CNVs and autosomal standard deviation (SD) is added to capture the effects of using a heterogeneous reference cohort. (f) The same sample analyzed with a more homogenous reference cohort showing a reduced variation and less CNV calls. (g) UCSC genome browser gene view showing gene structure and transcripts for the gene *CCDC141* for two different GENCODE versions highlighting how additional coding exons may be added that can change variant interpretation. An exome variant is indicated in exon five of the transcript present only in GENCODE version 37. CNV, copy number variants; IGV, Integrated Genomics Viewer

Obviously calling variants genome-wide will circumvent these issues, but we have judged that the additional compute time and increase in low-quality variant calls does not make this sufficiently worthwhile. We estimated that calling variants genome-wide would double analysis time and would yield many more variants called, of which an important number would be artifacts.

When implementing a new exome capture design it is highly recommended to define the clinical targets or regions of interest beforehand and then determine completeness of coverage for these intervals (Matthijs et al., 2016).

## 2.4 | Exome CNV calling: Reference pools

Early onwards it became clear that WES can also be used to infer CNVs, based on deviations in sequence coverage between samples (Marchuk et al., 2018). Comparison of coverage between exomes is hampered by coverage biases of individual targets due to sequence capture and GC content (Fromer et al., 2012). Most tools for the

detection of CNVs from exome data rely on the creation of a reference pool to standardize the depth of coverage per region and overcome coverage biases in the data (Krumm et al., 2012; Plagnol et al., 2012; Sathirapongsasuti et al., 2011). We discovered that the size and quality of the reference pool has a large impact on the quality of CNV calls. Reference pools with small numbers of samples or a mix of samples with different sequencing characteristics, will lead to increased variability on expected coverage for sequencing targets (Figure 1e). This will result in many spurious calls, making the interpretation much more laborious. In 2016 we accidentally combined samples of which reads were aligned using two different methods in the same reference pool. Unexpectedly this resulted not only in spurious CNV calls but also in large CNVs that were missed but that had already been detected in a previous CNV analysis. Currently, our CNV reference pools are continuously updated using the latest samples, to have minimal technical variability due to changes in sequencing chemistry and protocols (Figure 1f). Besides this continuous updating several separate reference pools are used that match samples based on sequencing platform, enrichment

platform, and sex for calling CNVs on chromosome X. To pick-up on potential quality issues we monitor the number of CNV calls per sample and sequencing batch as well as the average variability of the normalized target coverage per sample in our trend analysis. Based on our experience we would recommend to use a reference cohort for CNV calling that is matched for the capture kit, sequencing instrument and chemistry, and sex.

## 2.5 | Annotation: Gene definitions

Whereas we perform regular updates of reference datasets such as population frequencies, OMIM information and HGMD/ClinVar classification, we initially did not regularly update our gene definitions, naively expecting that all genes and transcripts in the human genome have been thoroughly charted. Gene definitions are the most basic resource for the interpretation of genetic variants. Several publicly available resources for gene definition exist, such as RefSeq (developed by the National Center for Biotechnology Information (NCBI)) (Pruitt et al., 2014) and GENCODE that combines a manual annotation by the HAVANA group with computational annotation by Ensembl (Harrow et al., 2012).

When we updated our 2017 GENCODE BASIC gene definitions to a more recent version, somewhat to our surprise we encountered several variants that were initially annotated as noncoding, but that turned out to be in a newly annotated exon, thereby potentially completely changing the interpretation, for example as with the gene *CCDC141* (Figure 1g). There are still regular updates from RefSeq and GENCODE that change known gene definitions and that can have a profound impact on the interpretation of variants for WES. Especially for WGS using more extensive gene definitions can be worthwhile since variants are detected genome-wide and are not limited to predefined regions as with WES. These ongoing improvements are nicely illustrated by the regular GENCODE updates. GENCODE was updated four times in the last 12 months, and the latest Gencode V38, May 2021 update includes more than 2500 new protein-coding transcripts, together with several modifications in the list of protein-coding genes compared to version V33 from January 2020 (Table S2). Regular updates (e.g., every 6 months) for all annotations including gene definitions and periodic re-annotation of existing samples will likely result in additional diagnoses.

## 3 | VARIANT INTERPRETATION

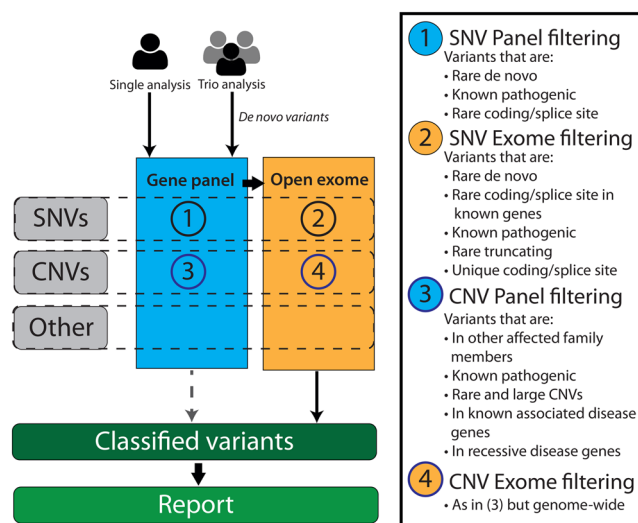
Next to data analysis, variant interpretation for NGS differs greatly from traditional practices and has come with its own challenges for molecular and clinical geneticists. Here we describe issues that we have encountered and the lessons that we have learned for clinical exome variant interpretation and illustrate these using real-life examples. These lessons are tentatively in order of importance, starting with what in our experience are the most valuable lessons that we have learned. In all the provided examples, variants were

initially interpreted according to our standard protocol which is depicted in Figure 2. We note that in practice these lessons are usually applied in combination, and some examples that we provide could have been used for multiple lessons.

### 3.1 | Visually inspect the data

Variant calling algorithms need to balance sensitivity, specificity and performance and will therefore not always provide perfect results (Kumaran et al., 2019). Hence it is good practice to visually inspect sequence alignment data (BAM/CRAM files) to manually filter-out false positive calls. False-positive calls often occur in repeat-rich regions and are readily visible upon inspection of the sequence alignment data. On the other hand, variants and especially insertion/deletion variants may be missed or inaccurately called.

In a patient with a neurodevelopmental disorder we identified two separately called de novo variants (NM\_001271.4:c.4592+37del and NM\_001271.4:c.4592+38C>G) in the gene *CHD2* (Figure S3). Individually each of these variants is predicted to have a benign or modest effect on splicing, and both variants were initially disregarded. However, after inspection of the alignment data it was clear that this represents a single variant, Chr15(GRCh37):g.93552590\_93552591delinsG NM\_001271.4:c.4592+37\_4592+38delinsG that introduces a new donor splice site predicted to lead to partial intron retention and a premature nonsense variant. Similarly, through visual inspection of the alignment data we found that a 13 base pair heterozygous deletion in *GPSM2* was actually



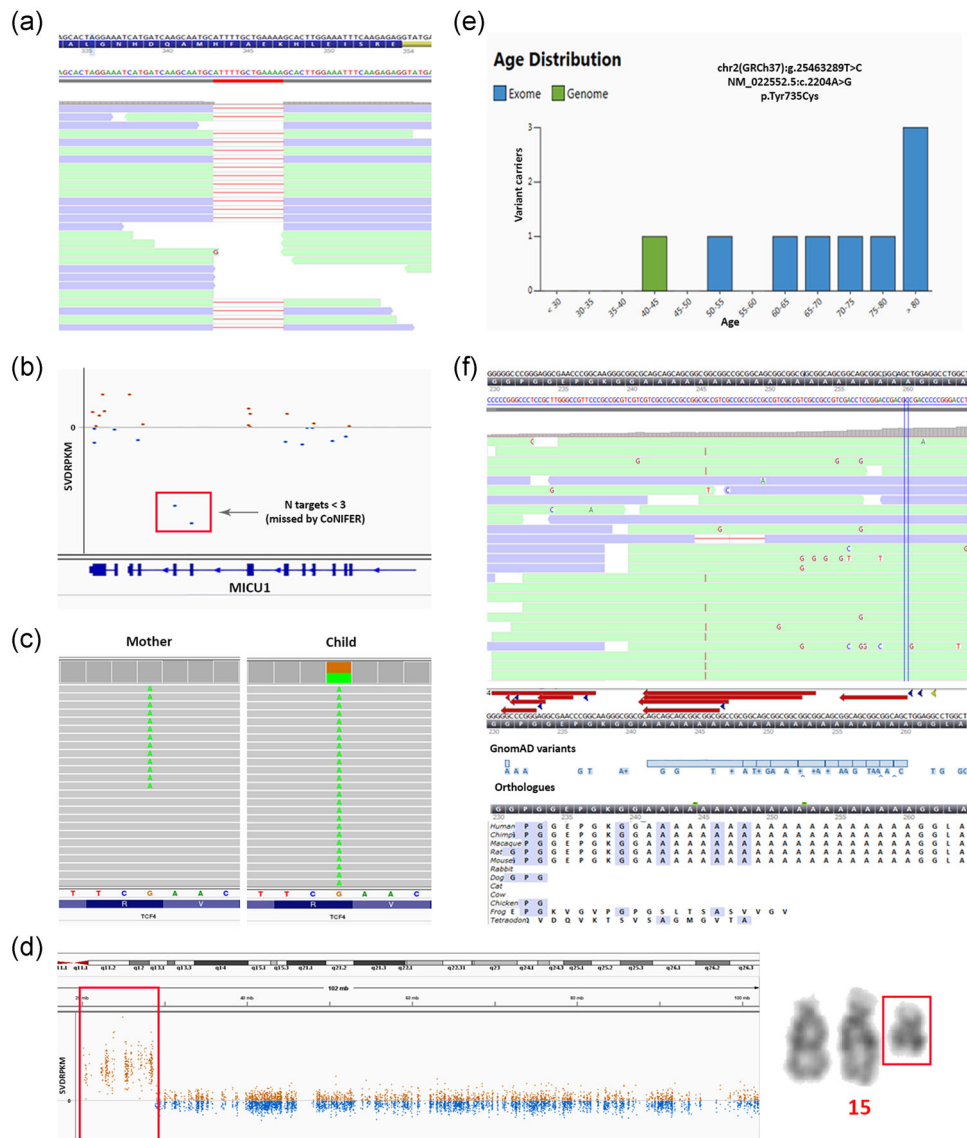
**FIGURE 2** Schematic representation of our interpretation workflow. Gray boxes on the left indicate the analysis of single nucleotide variants (SNVs; GATK calling), copy number variants (CNVs; CoNIFER; Krumm et al., 2012), and ExomeDepth (Plagnol et al., 2012) and “Other,” which includes interpretation of regions of homozygosity (Magi et al., 2014), Uniparental disomy (Yauy et al., 2020), and repeat expansions in coding regions (Dolzhenko et al., 2017)



present in a homozygous state (Figure 3a), and was inherited from both parents who were heterozygous for the variant.

Especially in the case of CNVs detected in WES data, visual inspection (of normalized coverage profiles and BAM files) is crucial. For example, a duplication event in the *MTMR2* gene can be recognized as a retrotransposon, that is, the insertion of copy-DNA in the genome (Huang et al., 2010), by the fact that multiple reads end exactly at the exon-intron boundaries (Figure S4). Similarly, visual inspection is especially important in the case of mosaic deletions, duplications and uniparental disomy, which otherwise could be missed.

In a patient with multiple congenital anomalies (left palate, ectopic anus, micropenis, and short proximal limbs) no genetic cause could be discovered with exome sequencing in 2015. However, upon CNV reanalysis of the same data in 2016, we discovered several small copy number gains of which only a few were visible within the restrictions of the requested gene panel. Visual inspection of the normalized coverage profile instantly revealed a gain of the whole short arm of chromosome 12 (Figure S5). The patient was eventually diagnosed with a mosaic quadruplication of the short arm of chromosome 12, causative of Pallister Killian syndrome (OMIM #601803).



**FIGURE 3** Issues that were encountered in data interpretation. (a) An obvious homozygous deletion in *GPM2* was incorrectly called by GATK as being heterozygous, indicating that visual inspection of the data is crucial for the correct interpretation of variants. (b) A pathogenic variant in the AR *MICU1* gene was detected together with a two exon deletion in the same gene. (c) A nonsense variant in *TCF4* was not detected when filtering on de novo variants because the mother was a mosaic carrier (30%) of this variant (alignments sorted by base in IGV). (d) An isodicentric marker chromosome (q13.1) was detected in WES data as an ~8.4 Mb terminal gain on 15q11.1q13.1, indicating that it is important to keep the chromosome structure in mind when analyzing WES CNV data. (e) Pathogenic variants in control databases, like the p.(Tyr735Cys) variant in *DNMT3A*, can be recognized by their overrepresentation in older individuals. (f) An 18bp duplication in *PHOX2B* was not called by GATK, but was detected after reanalysis prompted by the distinctive phenotype, emphasizing the power of hypothesis-driven diagnostics. IGV, Integrated Genomics Viewer

Visual inspection of the data is an essential aspect of variant interpretation. There are several tools to do this, including the Integrative Genomics Viewer (IGV) (Robinson et al., 2011). However, visual inspection of data is time-consuming and should be limited to variants with a higher likelihood of being called incorrectly. Such variants include CNVs, frameshift variants, variants with allele ratios that deviate from the ideal Mendelian ratios (i.e., not clearly heterozygous or homozygous) and multiple adjacent variants in a single gene. In addition, visual inspection should be performed for all variants that the laboratory intends to report.

### 3.2 | Variants other than nonsynonymous single nucleotide variants are easily missed

Exome sequencing was originally intended to detect single or multiple nucleotide replacements, or small deletions and duplications (~1–25 bp) within the coding regions and splice sites. In recent years, multiple studies have shown that other types of variants can also, to some degree, be detected in exome sequencing data. This includes among others CNVs (Pfundt et al., 2017), intronic variants (Y. Guo, Long, et al., 2012), uniparental disomies (UPDs) (Yauy et al., 2020), mitochondrial variants (Schlegel et al., 1992), repeat expansions (van der Sanden et al., 2021), and mobile element insertions (Torene et al., 2020). Whereas all of these resolve the cause of disease in only a relatively small number of patients compared to coding single nucleotide variants, together this bycatch can contribute substantially to the diagnostic yield.

For example, routine WES analysis of the coding regions and ±20bp splice site regions did not provide a diagnosis for a leukodystrophy patient with spastic hemiplegia and anarthria. As part of a comprehensive reanalysis by the Solve-RD consortium (Zurek et al., 2021) a homozygous known pathogenic deep intronic c.1969+115\_1969+116del variant in the *CSF1R* gene (Figure S6) was identified that leads to the inclusion of a pseudo-exon in the *CSF1R* transcripts (L. Guo et al., 2019). Although there was no specific capture target for this region, sequence coverage turned out to be sufficient at this position to call this particular variant.

For a patient with a clinical diagnosis of Stargardt disease, WES analysis of the vision-disorders panel genes and special focus on the Stargardt genes (*ABCA4* and *ELOVL4*) did not result in a molecular diagnosis. A reanalysis aimed at uniparental disomies detected a paternal isodisomy of chromosome 1 in this patient (Figure S7) (Magi et al., 2014; Yauy et al., 2020). Subsequent Sanger sequencing of the *ABCA4* Stargardt disease gene, located on chromosome 1, uncovered a homozygous pathogenic deep intronic variant (Chr1(GRCh37):g.94546780C>G NM\_000350.2(*ABCA4*):c.859-506G>C) leading to a pseudo-exon in a substantial proportion of the *ABCA4* transcripts (Khan et al., 2020).

In a deceased child that suffered from frontal pachygyria, breathing pattern disorders and brachycardia, whole exome analysis was performed. Two rare homozygous variants were detected in the *PLAA* gene, a missense and a synonymous variant. Although initially

we focused on the missense mutation it remained a VUS after interpretation. For the synonymous variant, splice prediction tools suggested that it might create an alternative splice donor site in exon 6 of this gene. Since the clinical phenotype of the patient fit with mutations in the *PLAA* gene, subsequent analysis of this predicted splice site effect was requested. Sequencing analysis of complementary DNA that was generated from lymphoblastoid cells from the carrier's parents indeed confirmed the usage of an alternative splice donor site that lead to an out-of-frame deletion of 11 nucleotides in the transcript that is encoded by the mutated allele (Figure S8). Instead of "just" being a silent variant, this variant leads to a loss-of-function of this allele.

Therefore we would recommend to consider all types of variants within genes that are clinically relevant to the patient's phenotype, and to highlight known pathogenic variants of all types (i.e., independent of their location or frequencies) from databases such as HGMD and ClinVar, during interpretation.

### 3.3 | Compound heterozygous variants are easily missed when one of the two is "hiding"

We found that in many cases where recessive inheritance was expected we could initially only identify a single heterozygous (pathogenic) variant in a recessive disease gene, which would be a very good matching gene for the patient's disorder if a second pathogenic variant were to be present. In these cases, the second variant may be a different type of mutation (see Section 3.2), may not meet the quality standards, or may seem less likely to be pathogenic. For example, a heterozygous loss-of-function variant, p.(Lys440\*), in the *MICU1* gene was detected using standard filtering in a child with a suspicion of a myopathy, based on increased creatine kinase (CK) levels and motor retardation. Only after visual inspection of the CNV data, the second variant, a heterozygous two exon deletion in *MICU1*, was detected (Figure 3b). This CNV was not called by the CNV algorithm (CoNIFER) that was used at the time, since the cut-off values for calling variants is three or more exons (Krumm et al., 2012).

Another example is the identification of heterozygous loss-of-function mutations in the *POLR3A* gene in four unrelated individuals with a movement disorder. Whereas initially these patients received no diagnosis, upon examination we identified an additional intronic variant (NM\_007055.4:c.1909+22G>A) in all four patients. The effect of this variant was uncertain, since it is predicted to enhance a cryptic donor splice site, while leaving the original donor splice site intact. This mutation was later shown to be a common hypomorphic variant (i.e., resulting in a milder *POLR3A* phenotype) that results in retention of 19 base pairs in a tissue- and stage- of development-specific manner (Minnerop et al., 2017).

These examples demonstrate that when a single heterozygous variant is detected in a recessive disease gene, which could be a good explanation of the patient's phenotype, one should be triggered to take extra efforts to identify a second variant (Kamphans et al., 2013).

### 3.4 | Remember mosaicism

Another challenge in the analysis of NGS data that was already alluded to (see Section 3.1) is the occurrence of mosaic SNVs and CNVs. Mosaic SNVs have been shown to be relevant for many disorders. In fact, ~3.5% of variants detected in patients with epilepsy-related neurodevelopmental disorder were present in a mosaic form (Stosser et al., 2018).

A common practice to remove sequencing and analysis artifacts is to exclude variants with a lower than expected variant allele frequency (VAF). However, such filtering will also remove mosaic SNVs. For example, initial filtering discarded a mosaic (~16%) variant in *PIK3CA* as an artifact in a fetus at 33 weeks of gestation. This pathogenic variant (Chr3(GRCh37):g.178916854G>A NM\_006218.4:c.241G>A p.(Glu81Lys)) causes abnormality of the cardiovascular system morphology, which could very well explain the ultrasound abnormalities seen in this fetus. Mosaicism for this variant was confirmed using targeted deep sequencing, revealing ~30% mosaicism allele fraction in the fetus and absence in both parents.

Another challenge arises when a pathogenic variant is also present in a mosaic state in an unaffected parent (Palomares-Bralo et al., 2017). When performing a trio analysis, the main focus is on the detection of *de novo* variants in dominant genes. As such, variants that occur in an unaffected (mosaic) parent are not labeled as *de novo* in the child. Therefore, variants inherited from a mosaic parent, will not be detected when solely looking for *de novo* variants. For instance, we initially missed a nonsense variant in *TCF4*, Chr18(GRCh37):g.53017619G>A NM\_001083962.1:c.520C>T p.(Arg174\*), when filtering for *de novo* variants, because 9% of the reads of the mother also contained this variant (Figure 3c). Ideally, such variants would be detected as a separate category when performing a *de novo* analysis. Alternatively, an inherited variant may be misinterpreted as sporadic due to the low level of mosaicism in a carrier parent, resulting in wrongly estimating the recurrence risk for the parents.

Overall, mosaic variants are not extremely rare. Mosaic variants in genes linked to autosomal dominant-, autosomal recessive-, and X-linked disorders are estimated to occur in 3.3% of individuals whereas parental mosaicism is estimated to be as high as 17.5% of apparently *de novo* mutations Qin et al. (2016). Whenever considering a potential pathogenic variant relevant to the patient's phenotype, it is worthwhile to also consider the possibility of mosaicism in either the patient or the parents.

### 3.5 | Think chromosomes

WES was initially aimed at detecting SNVs and although CNVs can be called from WES data it is important to keep in mind the limitations of WES when interpreting variants. For example, the CoNIFER algorithm does not detect aneuploidies because it normalizes the target coverage per chromosome (Krumm et al., 2012). We initially missed a case of isodisomy X Klinefelter syndrome (XXY) because

there were no CNV calls with CoNIFER (the only CNV calling tool used in our lab at that time). Since these were two identical X-chromosomes, there were Regions of Homozygosity (ROH) calls all over the X-chromosome, as you would expect in unaffected males. This isodisomy X Klinefelter was discovered using QF-PCR analysis, but could have been detected sooner by looking at the Y/X coverage ratio in the WES data.

A relatively common copy number finding from WES is the detection of a terminal duplication on one chromosome coinciding with a terminal deletion on another chromosome. This combination is a clear indication of an unbalanced translocation and should be followed up by regular karyotyping. A similar event, a ~265 kb terminal deletion on chromosome 22q13.3, was identified in a patient with severe intellectual disability, developmental delay, absent speech and language, hypotonia and reflux. Since chromosome 22 is an acrocentric chromosome, there were no calls on the short arm of this chromosome. Such a terminal deletion on the long and short arm of the same chromosome is indicative of a ring chromosome. Follow-up karyotyping revealed that this was indeed a *de novo* ring chromosome 22 (Figure S9). It is essential to differentiate a ring chromosome from a "regular" terminal aberration since instability during mitosis is a well-known characteristic of ring chromosomes (Nikitina et al., 2021). Subsequent secondary aberrations, like expansion of the deleted region or even monosomy of the affected chromosome, can have relevant clinical consequences for the affected individual. For chromosome 22 this risk has been described with respect to neurofibromatosis type 2 (NF2; OMIM #607379) where subsequent lifelong routine screening for features of NF2 in these patients is strongly advised (Zirn et al., 2012).

Another example was the identification of a ~8.4 Mb terminal gain on 15q11.1q13.1 in WES data from a patient with intellectual disability and epilepsy. Based on the WES data alone it was not clear whether this gain was caused by an interstitial duplication or by an extra numerical marker chromosome. Upon follow-up karyotyping this event turned out to be an isodicentric marker chromosome (q13.1) (Figure 3d) and thus in fact was a quadruplication of the q11q13.1 region. This is a clinically relevant finding because tetrasomy 15q gives rise to many nonspecific characteristics including intellectual disability, behavioral disorders, ataxia and epilepsy Finucane et al. (1993).

These examples demonstrate that it is necessary to also have cytogenetics expertise for WES interpretation. Existing guidelines on the interpretation of copy number variants from microarray data can provide guidance for the interpretation and follow-up of CNVs from exome sequencing data (Shao et al., 2021; Silva et al., 2019).

### 3.6 | Genuine disease-causing variants may be prevalent in population databases

Eliminating common variants is an essential step in exome data filtering (Gilissen et al., 2012). Publicly available databases such as gnomAD that provide aggregated variant information from large populations cohorts are of great help (Karczewski et al., 2020).



Commonly used thresholds for such filtering eliminate all data with an allele frequency >1% or based on the frequency and inheritance patterns of the disease (Whiffin et al., 2017). When applying such allele frequency filtering there are a number of reasons why clinically relevant variants may be wrongly discarded.

In a patient with intellectual disability we detected a missense variant in *DNMT3A* (c.2204A>G, p.(Tyr735Cys); NM\_022552.5). However, this variant also occurs in 11 individuals in the GnomAD database and therefore was initially considered likely benign. Several studies have now pointed out that particular variants may occur somatically in healthy individuals as a result of clonal hematopoiesis (Acuna-Hidalgo et al., 2017; Shlush, 2018). Therefore these (somatic) variants occur relatively frequently in control databases where they can be recognized by the fact that they are overrepresented in older individuals (Figure 3e) and have low variant allele fractions Carlston et al. (2017). It is useful to flag such genes that are involved in clonal hematopoiesis. When in doubt, targeted mutation analysis of alternative tissues can help to distinguish between constitutional and somatic variants.

Seemingly frequent pathogenic variants may also be due to homopolymeric stretches. Homopolymeric stretches in genes are regions that are prone to polymerase slippage that can result in the insertion or deletion of a number of nucleotides. These variants may be present in control databases as artifacts, but also may be genuine causative variants in the sequencing data being analyzed. An intriguing example is a deletion or duplication of a single cytosine from a homopolymer stretch of nine nucleotides in the *PRRT2* gene (NM\_145239.3:c.641\_649) (Figure S10). The subsequent c.649del and c.649dup (rs587778771) variants are present in the gnomAD database with an allele frequency of 0.96% and 0.47%, respectively. These high frequencies initially led us to not consider these variants as a likely cause. However, both events are considered pathogenic, since they lead to frameshifts in the *PRRT2* gene, where haploinsufficiency causes epilepsy, episodic kinesigenic dyskinesia or both. The penetrance of the *PRRT2* related disorders is estimated to be 60% or higher (van Vliet et al., 2012), suggesting that the high allele frequencies of the homopolymer changes in public databases may be due to sequencing artefacts. Indeed, limited alignment data present in gnomAD shows an unequal distribution of the mutant allele in some. It is therefore important to confirm such variants with another test if relevant to the case before reporting.

Although filtering variants using frequency databases is a useful approach, it is not perfect. Again, we would recommend to incorporate safeguards that highlight known pathogenic variants during the data interpretation process to not miss variants with higher populations frequencies (see Section 3.2).

### 3.7 | Distinctive clinical features may drive a correct diagnosis

Data analysis may sometimes discard potential variants based on quality criteria. In particular cases, the clinical phenotype can help

prioritize variants without the need of additional filtering steps, or can even suggest detailed analysis of specific genes. A de novo 18 bp duplication event in the *PHOX2B* gene was only identified after visual inspection of the sequencing data, which was prompted by the distinctive phenotype of congenital central hypoventilation syndrome in a newborn. This variant was not called, possibly due to poor alignment of sequencing reads in the GC-rich repetitive sequence of this region (Figure 3f). Interpretation was also a challenge, because the region is not conserved among vertebrates (many lack the repetitive stretch coding for an Alanine repeat) and since many overlapping deletion and duplication events are present in gnomAD. Nevertheless, a duplication event at this position is a recurrent cause of central hypoventilation syndrome.

Another example where a distinct clinical phenotype may help is with identifying highly frequent hypomorphic alleles (see also Section 3.6). We performed a prenatal exome analysis of a fetus with ultrasound anomalies (phocomelia, small chin, prenasal thickness, lower extremities in adducted position) where we at first only detected a paternal 1q21.1 deletion. The fetal phenotype matched with the possible clinical diagnosis of thrombocytopenia-absent-radius (TAR) syndrome (Albers et al., 2012). This syndrome is generally caused by a recurrent microdeletion in 1q21.1 in combination with, for example, a hypomorphic variant in the 5'-untranslated region at position -21 that has an allele frequency of >2% in the gnomAD database. Upon loosening the frequency filtering indeed the variant at position -21 emerged and was of maternal origin.

These examples show how a patient's phenotype may very specifically point to a single gene or a small number of genes. The attention should not only be directed to variants in those genes that may not have been called, but also to other less likely variants, such as silent or deep-intronic variants that may affect splicing (also see Section 3.2). It is thus beneficial to have dedicated specialists interpreting clinical exome sequencing data of specific groups of disorders, as this allows for deeper knowledge of gene etiologies, atypical variant types, or genotype-phenotype correlations within their area of expertise. The ability to reach a correct diagnosis will however always depend on the availability of complete clinical phenotype information, preferably in a standardized format.

### 3.8 | Phenotypic information may be misleading

Whereas phenotypic information is essential for proper genetic testing, it might also hinder the genetic diagnosis by the selection of gene-targeted tests. With the introduction of NGS techniques such as WES and WGS in genetic labs, the diagnostic strategy of referring clinicians changed from a phenotype-first to a genotype-first approach. By more-or-less unbiased sequencing analysis it became clear that pathogenic variants in well-known disease genes can also lead to a very different clinical phenotype depending on the position or type of genetic variation.

Compound heterozygous pathogenic variants in the *IL11RA* gene were detected in a 2-year-old child with neonatal hypotonia, feeding

problems, myoclonic movements, opsoclonus, frontal bossing, and club feet, and a mitochondrial disorder was suspected. The *IL11RA* gene is, however, involved in “craniosynostosis with dental anomalies” (OMIM #614188). In this rare disorder, no hypotonia or movement disorders were described. Prompted by the finding, a computed tomography scan revealed early closure of the sutures in the child and in a 3-year-old sibling. This sibling was then also shown to be compound heterozygous for the *IL11RA* variants. Thus, the frontal bossing, and perhaps the clubfeet, were early indicators of craniosynostosis, while the neurological features may or may not be explained by the *IL11RA* variants.

This kind of phenotypic heterogeneity is of course not new, but NGS implementation has generated many recent examples such as pathogenic *SRCAP* and *CREBBP* variants being causative for Floating Harbor (OMIM #136140) and Rubinstein-Taybi (OMIM #613684) syndrome, respectively. Variants in these genes have also been described causing a separate syndromic entity depending on the location of the (de novo) loss-of-function variant (Menke et al., 2018; Rots et al., 2021). Disease progression, incomplete clinical assessments, or phenotypic heterogeneity may initially be misleading. When detecting obvious pathogenic variants, they should not be set aside as “not compatible with the phenotype” too easily.

### 3.9 | Non-Mendelian inheritance

Most standard filtering strategies for WES data analysis and interpretation are based on classic Mendelian inheritance patterns. Whereas incomplete penetrance is obviously not a new phenomenon in genetic diseases, it does pose a challenge in efficiently filtering large sets of variants from NGS data (Cooper et al., 2013). Especially when handling patient-(healthy) parent trio data, variant filtering can lead to rejecting inherited heterozygous variants in dominant genes, or rejecting heterozygous X-linked variants in females of paternal origin or in X-linked recessive genes.

A trio-based WES analysis for a young woman with severe intellectual disability, autism and epilepsy initially did not result in a diagnosis. When discussing this result with the referring clinician, the possibility of a variant in the *PCDH19* gene was mentioned. *PCDH19* causes a female-restricted X-linked disorder of developmental and epileptic encephalopathy-9 (OMIM #300088). Targeted inspection of the data indeed revealed a paternally inherited pathogenic variant (ChrX (GRCh37):g.99662889G>A NM\_001184880.1:c.707C>T p.Pro236Leu) in the *PCDH19* gene. This missense variant was initially missed because of the inheritance from the healthy hemizygous father. One should thus be aware of heterozygous *PCDH19* variants that may very well be inherited from unaffected hemizygous fathers.

Another challenging group of genes are those that are parentally imprinted, and thus expressed depending on the gender of the parent that passes on the allele. There are approximately 15 well-described disorders already known to be caused by imprinted loci (Monk et al., 2019), but in addition several hundred genes are known or predicted to be subjected to genomic imprinting (<https://www.geneimprint.com/site/home>;

Monk et al., 2019). In a patient with multiple congenital anomalies we detected a *de novo* frameshift variant in the *IGF2* gene, that is, known to be subjected to imprinting and to be exclusively expressed on the paternal allele. Since genomic phasing information could not be extracted from the WES data of this patient, we were not able to determine on which allele the *IGF2* variant was present. Using an informative SNP (rs368743181) located 3.5 kb upstream of the frameshift variant in combination with genomic phased long-read sequencing could confirm that this mutation indeed arose on the paternal allele and could therefore be considered as causative. Had this variant not been *de novo*, but inherited from a healthy parent it would have been much more challenging to identify.

Also relevant here is the detection of uniparental disomy events that occur in one in 500–2000 individuals (Nakka et al., 2019; Yauy et al., 2020). In the case of a UPD, both chromosomes are inherited from the same parent and variants in imprinted genes can be a likely cause of disease. Annotation of genes with information about known disease mechanisms can be very useful for interpretation of WES data.

### 3.10 | Be aware of isoforms, pseudogenes and gene copies

Our concept of a gene's regulation has long been simplified as a single promoter driving the transcription of a gene, followed by the splicing of the pre-messenger RNA deleting all introns. Nowadays, we know that gene expression is controlled in a time-, tissue-, or developmental stage-dependent manner. For example, splicing isoforms may lack one or more exons (natural exon skipping), have additional relevant exons (Bodian et al., 2021), have different translation initiation sites, or genes may have multiple promoters causing the occurrence of different isoforms. The difficulty is to consider which isoform is relevant to disease, how to value a variant that is present in just a subset of isoforms, or, in case the reading frame is different between isoforms, how to ensure not missing the relevant “annotation” (Frankish et al., 2015; Schoch et al., 2020).

For example, we identified the Chr19(GRCh37):g.13339572G>A variant in the *CACNA1A* gene in a patient with episodic ataxia. In only one out of five isoforms of *CACNA1A*, this variant is a nonsense variant, NM\_001127221.1:c.5569C>T p.(Arg1857\*), while it is intronic in the other four (Figure S11). The polyQ expansion track involved in spinocerebellar ataxia type 6 (OMIM #183086) is encoded by two other *CACNA1A* isoforms (NM\_001127222.2 and NM\_023035.3), suggesting these two isoforms to be essential for proper cerebellar function. Thus, the fact that the nonsense variant is only present in the isoform that does not encode the polyQ track, initially led us to consider this variant as likely benign. However, Graves et al. (2008) showed that this isoform uses an alternative exon 37A instead of the original exon 37B and that nonsense variants in this isoform cause episodic ataxia (OMIM #108500).

Alternatively, isoform-specific variants may appear pathogenic, but may be benign since the entire isoform is redundant. Finally, some isoforms have partially different reading frames due to exon skipping, making it particularly difficult to annotate variants in them correctly. For variants with different effects in different isoforms all consequences are usually available, but for convenience the most severe consequence is prioritized (e.g., stop-loss over missense). Nevertheless, this may have consequences for diseases, like Noonan syndrome, with gain-of-function or dominant-negative mechanisms, where missense variants are pathogenic and nonsense variants are not. It is, overall, important to ensure variant calling and annotation in multiple isoforms followed by correct interpretation to not miss the relevant variants.

Also, gene copies and pseudogenes pose a serious problem in WES because of ambiguous sequence alignment of short sequence reads and the subsequent lack of variant calls in such regions. Notorious are copies of complete disease genes, such as *SMN1*, *CYP21A2*, *PKD1*, *STRC*, or parts of genes, such as the invariant triplicate of eight exons within the *NEB* gene (Donner et al., 2004; Mandelker et al., 2016). However, other variants may be called and display aberrant variant allele fractions, that is, heterozygous when homozygous or very low percentages in heterozygotes, or represent false-positive calls from the pseudogene(s) as we found for a nonsense variant in the *STRC* gene (Figure S12). One should be made aware of these genes during interpretation based on existing resources and perform validations of the presence and zygosity of such variants if identified, using independent techniques. Different laboratory approaches, such as NGS-based copy number assessment supplemented with a long-range PCR-base Sanger or MiSeq assay (Mandelker et al., 2014), have been suggested for this. In addition, it is possible to simply exclude segmental duplications from the analysis (Santani et al., 2017).

When based on the patient phenotype the detection of known pathogenic mutations could be difficult because of pseudogenes, patients should also be tested in a targeted fashion.

## 4 | DISCUSSION

Here we provide some of the most important lessons that we have learned from performing clinical exome sequencing for over 10 years. As a diagnostic laboratory the focus on quality and robustness does not encourage continuous change, but keeping up with updates and innovations has become an essential process in this fast-evolving field. By providing examples of mistakes that we have made in the development of our diagnostic workflows we hope we can not only create awareness of these specific issues but also of the fact that mistakes *do* occur in diagnostic laboratories. It is essential to be transparent to patients and referring clinicians about the limitations of clinical exome sequencing. These limitations should ideally be mentioned in diagnostic reports Claustres et al. (2014). Although some of the mistakes that were made have required us to recontact patients with a correct diagnosis, we feel that this is partly

unavoidable and that a fear of making mistakes should not hamper innovation and improvements as this would do more harm to patient care in the long term.

For this reason, it is however important to have a comprehensive framework for the timely detection of mistakes and problems at the level of the sequencing, data analysis as well as interpretation. Several initiatives can aid laboratories in this by providing benchmark datasets (Zook et al., 2019), and facilitating comparisons between laboratories Muller and European Molecular Genetics Quality Network (2001). An interesting observation from these examples is that issues that occurred during sequencing were sometimes not identified by the sequencing laboratory itself, but rather by the bioinformaticians who analyzed the data. Similarly, mistakes made in data processing were often picked-up by molecular geneticists during data interpretation. It is therefore essential to have routine procedures for feedback between the members involved in the different parts of the clinical exome sequencing process (i.e., sequencing facility, bioinformatics and data interpretation).

Although it may seem that the examples are very rare exceptions that are unlikely to have much relevance for everyday cases, we would argue that these “exceptions” are alike to rare genetic disorders that may be individually rare, but altogether quite common. It is of course not always feasible to dedicate the amount of time needed to consider all rare possibilities when performing routine exome interpretation. Therefore, data analysis, annotation and procedures should be gradually optimized to increase the automated pickup of such clinically relevant genetic variants. Similarly, it may be a relatively high investment to validate, setup and perform the multitude of possible analyses for WES, such as detection UPDs, mitochondrial variants, repeat expansions, mobile element insertions, and so forth. Data-sharing and reanalysis efforts such as the Solve-RD consortium (Zurek et al., 2021) may then prove beneficial and can leverage the large number of samples to perform analyses that are unlikely to diagnose any individual sample but within a large cohort will identify a handful of cases.

The mistakes that we presented here will probably not be our last ones. We strive to learn from our mistakes to improve diagnostics in the long run, and we hope that others can learn from our mistakes as well.

## ACKNOWLEDGMENTS

We thank all of the laboratory specialists, bioinformaticians, technicians and the Genome Sequencing facility in Radboudumc, department of genetics and Maastricht university medical centre, department of clinical genetics. We would like to thank Conny van Ravenswaaij for help with the *PLAA* gene example. This study was in part supported by the Solve-RD project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 779257 as well as grants from the Netherlands Organization for Scientific Research (917-17-353 to C.G.).

## CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ORCID

Jordi Corominas  <https://orcid.org/0000-0003-1736-1278>

Erik-Jan Kamsteeg  <https://orcid.org/0000-0001-6480-1892>

Rolph Pfundt  <https://orcid.org/0000-0002-0584-4398>

Christian Gilissen  <https://orcid.org/0000-0003-1693-9699>

## REFERENCES

- Acuna-Hidalgo, R., Sengul, H., Steehouwer, M., van de Vorst, M., Vermeulen, S. H., Kiemeny, L., Veltman, J. A., Gilissen, C., & Hoischen, A. (2017). Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *American Journal of Human Genetics*, 101(1), 50–64. <https://doi.org/10.1016/j.ajhg.2017.05.013>
- Albers, C. A., Paul, D. S., Schulze, H., Freson, K., Stephens, J. C., Smethurst, P. A., Jolley, J. D., Cvejic, A., Kostadima, M., Bertone, P., Breuning, M. H., Debili, N., Deloukas, P., Favier, R., Fiedler, J., Hobbs, C. M., Huang, N., Hurler, M. E., Kiddle, G., ... Ghevaert, C. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nature Genetics*, 44, 435–439. <https://doi.org/10.1038/ng.1083>
- Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., Kostareva, A. A., Glotov, O. S., & Predeus, A. V. (2020). Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Scientific Reports*, 10(1), 2057. <https://doi.org/10.1038/s41598-020-59026-y>
- Bodian, D. L., Kothiyal, P., & Hauser, N. S. (2021). Correction to: Pitfalls of clinical exome and gene panel testing: Alternative transcripts. *Genetics in Medicine*, 23, 2229. <https://doi.org/10.1038/s41436-021-01131-y>
- Brownstein, C. A., Beggs, A. H., Homer, N., Merriman, B., Yu, T. W., Flannery, K. C., Dechene, E. T., Towne, M. C., Savage, S. K., Price, E. N., Holm, I. A., Luquette, L. J., Lyon, E., Majzoub, J., Neupert, P., McCallie D. Jr., Szolovits, P., Willard, H. F., Mendelsohn, N. J., ... Caskey, C. T. (2014). An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY challenge. *Genome Biology*, 15(3), R53. <https://doi.org/10.1186/gb-2014-15-3-r53>
- Carlston, C. M., O'Donnell-Luria, A. H., Underhill, H. R., Cummings, B. B., Weisburd, B., Minikel, E. V., Birnbaum, D. P., Exome Aggregation, C., Tvrđik, T., MacArthur, D. G., & Mao, R. (2017). Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for Bohring-Opitz Syndrome. *Human Mutation*, 38(5), 517–523. <https://doi.org/10.1002/humu.23203>
- Claustres, M., Kožich, V., Dequeker, E., Fowler, B., Hehir-Kwa, J. Y., Miller, K., Oosterwijk, C., Peterlin, B., van Ravenswaaij-Arts, C., Zimmermann, U., Zuffardi, O., Hastings, R. J., & Barton, D. E., European Society of Human Genetics. (2014). Recommendations for reporting results of diagnostic genetic testing (biochemical, cytogenetic and molecular genetic). *European Journal of Human Genetics*, 22(2), 160–170. <https://doi.org/10.1038/ejhg.2013.125>
- Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C., & Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics*, 132(10), 1077–1130. <https://doi.org/10.1007/s00439-013-1331-2>
- Dolzhenko, E., van Vugt, J., Shaw, R. J., Bekritsky, M. A., van Blitterswijk, M., Narzisi, G., Ajay, S. S., Rajan, V., Lajoie, B. R., Johnson, N. H., Kingsbury, Z., Humphray, S. J., Schellevis, R. D., Brands, W. J., Baker, M., Rademakers, R., Kooyman, M., Tazelaar, G., van Es, M. A., ... Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research*, 27(11), 1895–1903. <https://doi.org/10.1101/gr.225672.117>
- Donner, K., Sandbacka, M., Lehtokari, V. L., Wallgren-Pettersson, C., & Pelin, K. (2004). Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts. *European Journal of Human Genetics*, 12(9), 744–751. <https://doi.org/10.1038/sj.ejhg.5201242>
- Finucane, B. M., Lusk, L., Arkilo, D., Chamberlain, S., Devinsky, O., Dindot, S., & Cook, E. H. (1993). 15q duplication syndrome and related disorders. In M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, G. Mirzaz, & A. Amemiya (Eds.), *GeneReviews*(R).
- Frankish, A., Uszczynska, B., Ritchie, G. R., Gonzalez, J. M., Pervouchine, D., Petryszak, R., Mudge, J. M., Fonseca, N., Brazma, A., Guigo, R., & Harrow, J. (2015). Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics*, 16(suppl 8), S2. <https://doi.org/10.1186/1471-2164-16-S8-S2>
- Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., Kirov, G., Sullivan, P. F., Hultman, C. M., Sklar, P., & Purcell, S. M. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American Journal of Human Genetics*, 91(4), 597–607. <https://doi.org/10.1016/j.ajhg.2012.08.005>
- Fung, J., Yu, M., Huang, S., Chung, C., Chan, M., Pajusalu, S., Mak, C., Hui, V., Tsang, M., Yeung, K. S., Lek, M., & Chung, B. (2020). A three-year follow-up study evaluating clinical utility of exome sequencing and diagnostic potential of reanalysis. *NPJ Genomic Medicine*, 5, 37. <https://doi.org/10.1038/s41525-020-00144-x>
- García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*, 28(20), 2678–2679. <https://doi.org/10.1093/bioinformatics/bts503>
- Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18, 9–19. <https://doi.org/10.1016/j.csbj.2019.11.002>
- Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. A. (2012). Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, 20(5), 490–497. <https://doi.org/10.1038/ejhg.2011.258>
- Graves, T. D., Imbrici, P., Kors, E. E., Terwindt, G. M., Eunson, L. H., Frants, R. R., Haan, J., Ferrari, M. D., Goadsby, P. J., Hanna, M. G., van den Maagdenberg, A. M., & Kullmann, D. M. (2008). Premature stop codons in a facilitating EF-hand splice variant of CaV2.1 cause episodic ataxia type 2. *Neurobiology of Disease*, 32(1), 10–15. <https://doi.org/10.1016/j.nbd.2008.06.002>
- Guo, L., Bertola, D. R., Takanohashi, A., Saito, A., Segawa, Y., Yokota, T., Ishibashi, S., Nishida, Y., Yamamoto, G. L., Franco, J., Honjo, R. S., Kim, C. A., Musso, C. M., Timmons, M., Pizzino, A., Taft, R. J., Lajoie, B., Knight, M. A., Fischbeck, K. H., ... Ikegawa, S. (2019). Bi-allelic CSF1R mutations cause skeletal dysplasia of dysosteosclerosis-pyle disease spectrum and degenerative encephalopathy with brain malformation. *American Journal of Human Genetics*, 104(5), 925–935. <https://doi.org/10.1016/j.ajhg.2019.03.004>



- Guo, Y., Li, J., Li, C. I., Long, J., Samuels, D. C., & Shyr, Y. (2012). The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, 13, 666. <https://doi.org/10.1186/1471-2164-13-666>
- Guo, Y., Long, J., He, J., Li, C. I., Cai, Q., Shu, X. O., Zheng, W., & Li, C. (2012). Exome sequencing generates high-quality data in non-target regions. *BMC Genomics*, 13, 194. <https://doi.org/10.1186/1471-2164-13-194>
- Guo, Y., Ye, F., Sheng, Q., Clark, T., & Samuels, D. C. (2014). Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics*, 15(6), 879–889. <https://doi.org/10.1093/bib/bbt069>
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., ... Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9), 1760–1774. <https://doi.org/10.1101/gr.135350.111>
- Hartman, P., Beckman, K., Silverstein, K., Yohe, S., Schomaker, M., Henzler, C., Onsongo, G., Lam, H. C., Munro, S., Daniel, J., Billstein, B., Deshpande, A., Hauge, A., Mroz, P., Lee, W., Holle, J., Wiens, K., Karnuth, K., Kemmer, T., ... Thyagarajan, B. (2019). Next generation sequencing for clinical diagnostics: Five year experience of an academic laboratory. *Molecular Genetics and Metabolism Reports*, 19, 100464. <https://doi.org/10.1016/j.ymgmr.2019.100464>
- Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1–8. <https://doi.org/10.1016/j.ygeno.2015.11.003>
- Hong, C. S., Singh, L. N., Mullikin, J. C., & Biesecker, L. G. (2016). Assessing the reproducibility of exome copy number variations predictions. *Genome Medicine*, 8(1), 82. <https://doi.org/10.1186/s13073-016-0336-6>
- Huang, C. R., Schneider, A. M., Lu, Y., Niranjan, T., Shen, P., Robinson, M. A., Steranka, J. P., Valle, D., Civin, C. I., Wang, T., Wheelan, S. J., Ji, H., Boeke, J. D., & Burns, K. H. (2010). Mobile interspersed repeats are major structural variants in the human genome. *Cell*, 141(7), 1171–1182. <https://doi.org/10.1016/j.cell.2010.05.026>
- Jia, T., Munson, B., Lango Allen, H., Ideker, T., & Majithia, A. R. (2020). Thousands of missing variants in the UK Biobank are recoverable by genome realignment. *Annals of Human Genetics*, 84(3), 214–220. <https://doi.org/10.1111/ahg.12383>
- Jiang, Y., Turinsky, A. L., & Brudno, M. (2015). The missing indels: An estimate of indel variation in a human genome and analysis of factors that impede detection. *Nucleic Acids Research*, 43(15), 7217–7228. <https://doi.org/10.1093/nar/gkv677>
- Kamphans, T., Sabri, P., Zhu, N., Heinrich, V., Mundlos, S., Robinson, P. N., Parkhomchuk, D., & Krawitz, P. M. (2013). Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. *PLOS One*, 8(8), e70151. <https://doi.org/10.1371/journal.pone.0070151>
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443. <https://doi.org/10.1038/s41586-020-2308-7>
- Khan, M., Cornelis, S. S., Pozo-Valero, M. D., Whelan, L., Runhart, E. H., Mishra, K., Bults, F., AlSwaiti, Y., AlTalibshi, A., De Baere, E., Banfi, S., Banin, E., Bauwens, M., Ben-Yosef, T., Boon, C., van den Born, L. I., Defoort, S., Devos, A., Dockery, A., ... Cremers, F. (2020). Resolving the dark matter of ABCA4 for 1054 Stargardt disease probands through integrated genomics and transcriptomics. *Genetics in Medicine*, 22(7), 1235–1246. <https://doi.org/10.1038/s41436-020-0787-4>
- Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., NHLBI Exome Sequencing, P., Quinlan, A. R., Nickerson, D. A., & Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research*, 22(8), 1525–1532. <https://doi.org/10.1101/gr.138115.112>
- Kumaran, M., Subramanian, U., & Devarajan, B. (2019). Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. *BMC Bioinformatics*, 20(1), 342. <https://doi.org/10.1186/s12859-019-2928-9>
- Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A., & Gilissen, C. (2015). Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Human Mutation*, 36(8), 815–822. <https://doi.org/10.1002/humu.22813>
- Lelieveld, S. H., Veltman, J. A., & Gilissen, C. (2016). Novel bioinformatic developments for exome sequencing. *Human Genetics*, 135(6), 603–614. <https://doi.org/10.1007/s00439-016-1658-6>
- Liu, P., Meng, L., Normand, E. A., Xia, F., Song, X., Ghazi, A., Rosenfeld, J., Magoulas, P. L., Braxton, A., Ward, P., Dai, H., Yuan, B., Bi, W., Xiao, R., Wang, X., Chiang, T., Vetrini, F., He, W., Cheng, H., ... Yang, Y. (2019). Reanalysis of clinical exome sequencing data. *New England Journal of Medicine*, 380(25), 2478–2480. <https://doi.org/10.1056/NEJMc1812033>
- MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehms, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., Altman, R. B., Antonarakis, S. E., Ashley, E. A., Barrett, J. C., Biesecker, L. G., Conrad, D. F., Cooper, G. M., Cox, N. J., Daly, M. J., Gerstein, M. B., Goldstein, D. B., Hirschhorn, J. N., ... Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497), 469–476. <https://doi.org/10.1038/nature13127>
- Magi, A., Tattini, L., Palombo, F., Benelli, M., Gialluisi, A., Giusti, B., Abbate, R., Seri, M., Gensini, G. F., Romeo, G., & Pippucci, T. (2014). H3M2: Detection of runs of homozygosity from whole-exome sequencing data. *Bioinformatics*, 30(20), 2852–2859. <https://doi.org/10.1093/bioinformatics/btu401>
- Mandelker, D., Amr, S. S., Pugh, T., Gowrisankar, S., Shakhbatyan, R., Duffy, E., Bowser, M., Harrison, B., Lafferty, K., Mahanta, L., Rehms, H. L., & Funke, B. H. (2014). Comprehensive diagnostic testing for stereocilin: An approach for analyzing medically important genes with high homology. *Journal of Molecular Diagnostics*, 16(6), 639–647. <https://doi.org/10.1016/j.jmoldx.2014.06.003>
- Mandelker, D., Schmidt, R. J., Ankala, A., McDonald Gibson, K., Bowser, M., Sharma, H., Duffy, E., Hegde, M., Santani, A., Lebo, M., & Funke, B. (2016). Navigating highly homologous genes in a molecular diagnostic setting: A resource for clinical next-generation sequencing. *Genetics in Medicine*, 18(12), 1282–1289. <https://doi.org/10.1038/gim.2016.58>
- Marchuk, D. S., Crooks, K., Strande, N., Kaiser-Rogers, K., Milko, L. V., Brandt, A., Arreola, A., Tilley, C. R., Bizon, C., Vora, N. L., Wilhelmsen, K. C., Evans, J. P., & Berg, J. S. (2018). Increasing the diagnostic yield of exome sequencing by copy number variant analysis. *PLOS One*, 13(12), e0209185. <https://doi.org/10.1371/journal.pone.0209185>
- Marshall, C. R., Chowdhury, S., Taft, R. J., Lebo, M. S., Buchan, J. G., Harrison, S. M., Rowsey, R., Klee, E. W., Liu, P., Worthey, E. A., Jobanputra, V., Dimmock, D., Kearney, H. M., Bick, D., Kulkarni, S., Taylor, S. L., Belmont, J. W., Stavropoulos, D. J., Lennon, N. J., & Medical Genome, I. (2020). Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. *NPJ Genomic Medicine*, 5, 47. <https://doi.org/10.1038/s41525-020-00154-9>
- Matthijs, G., Souche, E., Alders, M., Corveleyn, A., Eck, S., Feenstra, I., Race, V., Sistermans, E., Sturm, M., Weiss, M., Yntema, H., Bakker, E., Scheffer, H., & Bauer, P. (2016). Guidelines for diagnostic



- next-generation sequencing. *European Journal of Human Genetics*, 24(10), 1515. <https://doi.org/10.1038/ejhg.2016.63>
- Menke, L. A., DDD Study, Gardeitchik, T., Hammond, P., Heimdal, K. R., Houge, G., Hufnagel, S. B., Ji, J., Johansson, S., Kant, S. G., Kinning, E., Leon, E. L., Newbury-Ecob, R., Paolacci, S., Pfundt, R., Ragge, N. K., Rinne, T., Ruivenkamp, C., Saitta, S. C., ... Hennekam, R. C. (2018). Further delineation of an entity caused by CREBBP and EP300 mutations but not resembling Rubinstein-Taybi syndrome. *American Journal of Medical Genetics Part A*, 176(4), 862–876. <https://doi.org/10.1002/ajmg.a.38626>
- Miller, D. T., Lee, K., Gordon, A. S., Amendola, L. M., Adelman, K., Bale, S. J., Chung, W. K., Gollob, M. H., Harrison, S. M., Herman, G. E., Hershberger, R. E., Klein, T. E., McKelvey, K., Richards, C. S., Vlangos, C. N., Stewart, D. R., Watson, M. S., & Martin, C. L. (2021). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2021 update: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*, 23, 1391–1398. <https://doi.org/10.1038/s41436-021-01171-4>
- Minnerop, M., Kurzwelly, D., Wagner, H., Soehn, A. S., Reichbauer, J., Tao, F., & Schule, R. (2017). Hypomorphic mutations in POLR3A are a frequent cause of sporadic and recessive spastic ataxia. *Brain*, 140(6), 1561–1578. <https://doi.org/10.1093/brain/awx095>
- Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R., & Riccio, A. (2019). Genomic imprinting disorders: Lessons on how genome, epigenome and environment interact. *Nature Reviews Genetics*, 20(4), 235–248. <https://doi.org/10.1038/s41576-018-0092-0>
- Muller, C. R., & European Molecular Genetics Quality Network. (2001). Quality control in mutation analysis: The European Molecular Genetics Quality Network (EMQN). *European Journal of Pediatrics*, 160(8), 464–467. <https://doi.org/10.1007/s004310100767>
- Nakka, P., Pattillo Smith, S., O'Donnell-Luria, A. H., McManus, K. F., Research, T., Mountain, J. L., Ramachandran, S., & Sathirapongsasuti, J. F. (2019). Characterization of prevalence and health consequences of uniparental disomy in four million individuals from the general population. *American Journal of Human Genetics*, 105(5), 921–932. <https://doi.org/10.1016/j.ajhg.2019.09.016>
- Nikitina, T. V., Kashevarova, A. A., Gridina, M. M., Lopatkina, M. E., Khabarova, A. A., Yakovleva, Y. S., Menzorov, A. G., Minina, Y. A., Pristyazhnyuk, I. E., Vasilyev, S. A., Fedotov, D. A., Serov, O. L., & Lebedev, I. N. (2021). Complex biology of constitutional ring chromosomes structure and (in)stability revealed by somatic cell reprogramming. *Scientific Reports*, 11(1), 4325. <https://doi.org/10.1038/s41598-021-83399-3>
- Palomares-Bralo, M., Vallespín, E., Del Pozo, Á., Ibañez, K., Silla, J. C., Galán, E., Gordo, G., Martínez-Glez, V., Alba-Valdivia, L. I., Heath, K. E., García-Miñaur, S., Lapunzina, P., & Santos-Simarro, F. (2017). Pitfalls of trio-based exome sequencing: Imprinted genes and parental mosaicism-MAGEL2 as an example. *Genetics in Medicine*, 19(11), 1285–1286. <https://doi.org/10.1038/gim.2017.42>
- Pengelly, R. J., Ward, D., Hunt, D., Mattocks, C., & Ennis, S. (2020). Comparison of Mendeliome exome capture kits for use in clinical diagnostics. *Scientific Reports*, 10(1), 3235. <https://doi.org/10.1038/s41598-020-60215-y>
- Pfundt, R., Del Rosario, M., Vissers, L., Kwint, M. P., Janssen, I. M., de Leeuw, N., Yntema, H. G., Nelen, M. R., Lugtenberg, D., Kamsteeg, E. J., Wieskamp, N., Stegmann, A., Stevens, S., Rodenburg, R., Simons, A., Mensenkamp, A. R., Rinne, T., Gilissen, C., Scheffer, H., ... Hehir-Kwa, J. Y. (2017). Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genetics in Medicine*, 19(6), 667–675. <https://doi.org/10.1038/gim.2016.163>
- Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., Kumararatne, D., Doffinger, R., & Nejentsev, S. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21), 2747–2754. <https://doi.org/10.1093/bioinformatics/bts526>
- Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., ... Ostell, J. M. (2014). RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research*, 42(Database issue), D756–D763. <https://doi.org/10.1093/nar/gkt1114>
- Qin, L., Wang, J., Tian, X., Yu, H., Truong, C., Mitchell, J. J., Wierenga, K. J., Craigen, W. J., Zhang, V. W., & Wong, L. C. (2016). Detection and quantification of mosaic mutations in disease genes by next-generation sequencing. *The Journal of Molecular Diagnostics*, 18, 446–453. <https://doi.org/10.1016/j.jmoldx.2016.01.002>
- Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L., Plon, S. E., Ramos, E. M., Sherry, S. T., Watson, M. S., & ClinGen. (2015). ClinGen—The clinical genome resource. *New England Journal of Medicine*, 372(23), 2235–2242. <https://doi.org/10.1056/NEJMs1406261>
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & ACMG Laboratory Quality Assurance Committee. (2015). 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*, 17(5), 405–424. <https://doi.org/10.1038/gim.2015.30>
- Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26. <https://doi.org/10.1038/nbt.1754>
- Rots, D., Chater-Diehl, E., Dingemans, A. J. M., Goodman, S. J., Siu, M. T., Cyttrynbaum, C., Choufani, S., Hoang, N., Walker, S., Awamleh, Z., Charkow, J., Meyn, S., Pfundt, R., Rinne, T., Gardeitchik, T., de Vries, B. B. A., Deden, A. C., Leenders, E., Kwint, M., ... Weksberg, R. (2021). Truncating SRCAP variants outside the Floating-Harbor syndrome locus cause a distinct neurodevelopmental disorder with a specific DNA methylation signature. *American Journal of Human Genetics*, 108(6), 1053–1068. <https://doi.org/10.1016/j.ajhg.2021.04.008>
- Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., Leon, A., Pullambhatla, M., Temple-Smolkin, R. L., Voelkerding, K. V., Wang, C., & Carter, A. B. (2018). Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *Journal of Molecular Diagnostics*, 20(1), 4–27. <https://doi.org/10.1016/j.jmoldx.2017.11.003>
- Santani, A., Murrell, J., Funke, B., Yu, Z., Hegde, M., Mao, R., Ferreira-Gonzalez, A., Voelkerding, K. V., & Weck, K. E. (2017). Development and validation of targeted next-generation sequencing panels for detection of germline variants in inherited diseases. *Archives of Pathology and Laboratory Medicine*, 141(6), 787–797. <https://doi.org/10.5858/arpa.2016-0517-RA>
- Santani, A., Simen, B. B., Briggs, M., Lebo, M., Merker, J. D., Nikiforova, M., Vasalos, P., Voelkerding, K., Pfeifer, J., & Funke, B. (2019). Designing and implementing NGS tests for inherited disorders: A practical framework with step-by-step guidance for clinical laboratories. *Journal of Molecular Diagnostics*, 21(3), 369–374. <https://doi.org/10.1016/j.jmoldx.2018.11.004>
- Sathirapongsasuti, J. F., Lee, H., Horst, B. A., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J., & Nelson, S. F. (2011). Exome

- sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, 27(19), 2648–2654. <https://doi.org/10.1093/bioinformatics/btr462>
- Schlegel, J., Schweizer, M., & Richter, C. (1992). "Pore" formation is not required for the hydroperoxide-induced Ca<sup>2+</sup> release from rat liver mitochondria. *Biochemical Journal*, 285(Pt 1), 65–69. <https://doi.org/10.1042/bj2850065>
- Schoch, K., Tan, Q. K., Stong, N., Deak, K. L., McConkie-Rosell, A., McDonald, M. T., Undiagnosed Diseases, N., Goldstein, D. B., Jiang, Y. H., & Shashi, V. (2020). Alternative transcripts in variant interpretation: The potential for missed diagnoses and misdiagnoses. *Genetics in Medicine*, 22(7), 1269–1275. <https://doi.org/10.1038/s41436-020-0781-x>
- Shao, L., Akkari, Y., Cooley, L. D., Miller, D. T., Seifert, B. A., Wolff, D. J., & Mikhail, F. M. (2021). Chromosomal microarray analysis, including constitutional and neoplastic disease applications, 2021 revision: A technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*, 23(10), 1818–1829. <https://doi.org/10.1038/s41436-021-01214-w>
- Shlush, L. I. (2018). Age-related clonal hematopoiesis. *Blood*, 131(5), 496–504. <https://doi.org/10.1182/blood-2017-07-746453>
- Silva, M., de Leeuw, N., Mann, K., Schuring-Blom, H., Morgan, S., Giardino, D., Rack, K., & Hastings, R. (2019). European guidelines for constitutional cytogenomic analysis. *European Journal of Human Genetics*, 27(1), 1–16. <https://doi.org/10.1038/s41431-018-0244-x>
- Stosser, M. B., Lindy, A. S., Butler, E., Retterer, K., Piccirillo-Stosser, C. M., Richard, G., & McKnight, D. A. (2018). High frequency of mosaic pathogenic variants in genes causing epilepsy-related neurodevelopmental disorders. *Genetics in Medicine*, 20(4), 403–410. <https://doi.org/10.1038/gim.2017.114>
- Torene, R. I., Galens, K., Liu, S., Arvai, K., Borroto, C., Scuffins, J., Zhang, Z., Friedman, B., Sroka, H., Heeley, J., Beaver, E., Clarke, L., Neil, S., Walla, J., Hull, D., Juusola, J., & Retterer, K. (2020). Mobile element insertion detection in 89,874 clinical exomes. *Genetics in Medicine*, 22(5), 974–978. <https://doi.org/10.1038/s41436-020-0749-x>
- van der Sanden, B., Corominas, J., de Groot, M., Pennings, M., Meijer, R., Verbeek, N., van de Warrenburg, B., Schouten, M., Yntema, H. G., Vissers, L., Kamsteeg, E. J., & Gilissen, C. (2021). Systematic analysis of short tandem repeats in 38,095 exomes provides an additional diagnostic yield. *Genetics in Medicine*, 23, 1569–1573. <https://doi.org/10.1038/s41436-021-01174-1>
- van Vliet, R., Breedveld, G., de Rijk-van Andel, J., Brilstra, E., Verbeek, N., Verschuuren-Bemelmans, C., Boon, M., Samijn, J., Diderich, K., van de Laar, I., Oostra, B., Bonifati, V., & Maat-Kievit, A. (2012). PRRT2 phenotypes and penetrance of paroxysmal kinesigenic dyskinesia and infantile convulsions. *Neurology*, 79(8), 777–784. <https://doi.org/10.1212/WNL.0b013e3182661fe3>
- Weißbach, S., Sys, S., Hewel, C., Todorov, H., Schweiger, S., Winter, J., Pfenninger, M., Torkamani, A., Evans, D., Burger, J., Everschor-Sitte, K., May-Simera, H. L., & Gerber, S. (2021). Reliability of genomic variants across different next-generation sequencing platforms and bioinformatic processing pipelines. *BMC Genomics*, 22(1), 62. <https://doi.org/10.1186/s12864-020-07362-8>
- Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A. H., Karczewski, K., Ing, A. Y., Barton, P., Funke, B., Cook, S. A., MacArthur, D., & Ware, J. S. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine*, 19(10), 1151–1158. <https://doi.org/10.1038/gim.2017.26>
- Yauy, K., de Leeuw, N., Yntema, H. G., Pfundt, R., & Gilissen, C. (2020). Accurate detection of clinically relevant uniparental disomy from exome sequencing data. *Genetics in Medicine*, 22(4), 803–808. <https://doi.org/10.1038/s41436-019-0704-x>
- Zhou, J., Zhang, M., Li, X., Wang, Z., Pan, D., & Shi, Y. (2021). Performance comparison of four types of target enrichment baits for exome DNA sequencing. *Heredity*, 158(1), 10. <https://doi.org/10.1186/s41065-021-00171-3>
- Zirn, B., Arning, L., Bartels, I., Shoukier, M., Höffjan, S., Neubauer, B., & Hahn, A. (2012). Ring chromosome 22 and neurofibromatosis type II: Proof of two-hit model for the loss of the NF2 gene in the development of meningioma. *Clinical Genetics*, 81(1), 82–87. <https://doi.org/10.1111/j.1399-0004.2010.01598.x>
- Zook, J. M., McDaniel, J., Olson, N. D., Wagner, J., Parikh, H., Heaton, H., Irvine, S. A., Trigg, L., Truty, R., McLean, C. Y., De La Vega, F. M., Xiao, C., Sherry, S., & Salit, M. (2019). An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, 37(5), 561–566. <https://doi.org/10.1038/s41587-019-0074-6>
- Zurek, B., Ellwanger, K., Vissers, L., Schüle, R., Synofzik, M., Töpf, A., de Voer, R. M., Laurie, S., Matalonga, L., Gilissen, C., Ossowski, S., 't Hoen, P., Vitobello, A., Schulze-Hentrich, J. M., Riess, O., Brunner, H. G., Brookes, A. J., Rath, A., Bonne, G., ... Solve-RD Consortium. (2021). Solve-RD: Systematic pan-European data sharing and collaborative analysis to solve rare diseases. *European Journal of Human Genetics*, 29, 1325–1331. <https://doi.org/10.1038/s41431-021-00859-0>

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Corominas, J., Smeekens, S. P., Nelen, M. R., Yntema, H. G., Kamsteeg, E.-J., Pfundt, R., & Gilissen, C. (2022). Clinical exome sequencing—Mistakes and caveats. *Human Mutation*, 43, 1041–1055. <https://doi.org/10.1002/humu.24360>