

REVIEW

Open Access



# Whole genome sequencing in clinical practice

Frederik Otzen Bagger<sup>1†</sup>, Line Borgwardt<sup>1†</sup>, Andreas Sand Jespersen<sup>1</sup>, Anna Reimer Hansen<sup>1</sup>, Birgitte Bertelsen<sup>1</sup>, Miyako Kodama<sup>1</sup> and Finn Cilius Nielsen<sup>1\*</sup>

## Abstract

Whole genome sequencing (WGS) is becoming the preferred method for molecular genetic diagnosis of rare and unknown diseases and for identification of actionable cancer drivers. Compared to other molecular genetic methods, WGS captures most genomic variation and eliminates the need for sequential genetic testing. Whereas, the laboratory requirements are similar to conventional molecular genetics, the amount of data is large and WGS requires a comprehensive computational and storage infrastructure in order to facilitate data processing within a clinically relevant timeframe. The output of a single WGS analyses is roughly 5 MIO variants and data interpretation involves specialized staff collaborating with the clinical specialists in order to provide standard of care reports. Although the field is continuously refining the standards for variant classification, there are still unresolved issues associated with the clinical application. The review provides an overview of WGS in clinical practice - describing the technology and current applications as well as challenges connected with data processing, interpretation and clinical reporting.

**Keywords** Whole genome sequencing, Clinical bioinformatics infrastructure, Variant filtering and interpretation, Functional variant testing

## Background

The human genome project was a ground-breaking scientific endeavour that not only gave us a near complete map of our genetic code but also paved the way for new innovative sequencing technologies and computational methods that have enabled the clinical application of genomics [1–4]. While DNA sequencing dates back to the late 1970s [5], it was not until the beginning of the 90s that sequencing, with advent of semi-automized four-color dye sequencing [6], became available for routine clinical

use. Since then, the development of Next Generation Sequencing (NGS), has revolutionized the field, enabling the analysis of entire genomes in a fast and cost-effective manner [7, 8]. At this stage the last hard-to-sequence bits of the human genome have been mapped, and hundreds of thousands of people have had their entire genome sequenced [9].

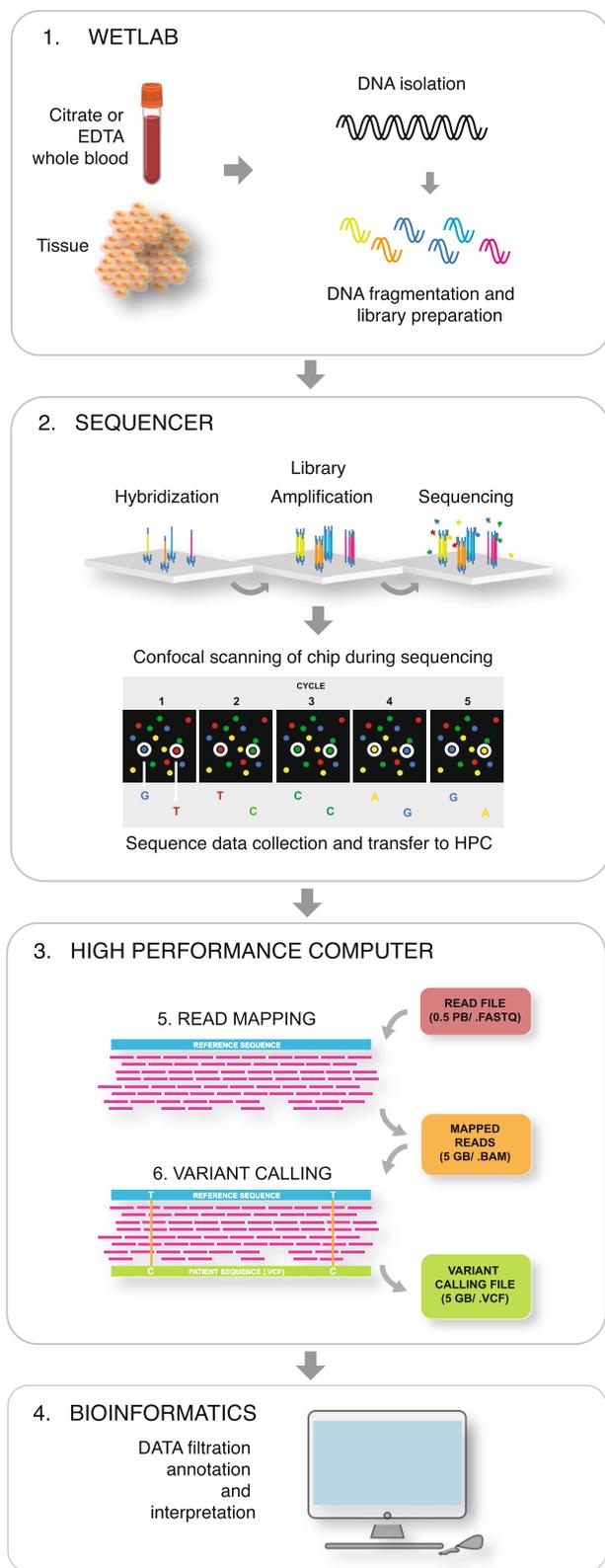
The capacity of NGS has steadily increased and with the latest generation of sequencing platforms, an entire human genome can be sequenced within 2 days at the price of a few hundred dollars. The relatively modest costs per analysis, combined with excellent data quality [10], make whole genome sequencing (WGS) a valuable source of information in many clinical situations. Compared to other genomic analysis, archived WGS data moreover have the potential to serve as a lifelong companion for patients that can be reanalysed and reinterpreted several times along the patient journey.

<sup>†</sup>Frederik Otzen Bagger and Line Borgwardt these two authors contributed equally.

\*Correspondence:  
Finn Cilius Nielsen  
finn.cilius.nielsen@regionh.dk

<sup>1</sup> Center for Genomic Medicine, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark





**Fig. 1** Schematic representation of the WGS laboratory and bioinformatics flow. Short-read WGS protocols can in general be divided into four separate steps: 1. Sample preparation, 2. Library preparation, 3. Cluster generation, and 4. Sequencing. Panel 1, WGS is routinely performed with DNA from EDTA or citrate stabilized whole blood or surgically removed or biopsy tissue. DNA is isolated by conventional methods, but to facilitate CNV detection high molecular DNA is preferred. Historically, WGS required a DNA amplification step, but with newer protocols this step is no longer needed. Omission of the amplification step eliminates the PCR-bias and provides a more uniform coverage and quality [12]. The library is generated by fragmenting the high molecular DNA followed by ligation of adapters that will bind to the linker DNA on the chip surface. Moreover, barcodes allowing pooling of samples from different patients on the same chip may be attached. Panel 2, The libraries are subsequently loaded onto a flow cell and placed on the sequencer, after which the individual DNA fragments are clonally amplified by a polymerase, generating small single-stranded clusters of the particular fragments. The sequencing is in principle a conventional Sanger sequencing [5], where elongation is initiated by the addition of a sequence primer and polymerase and the nucleotide sequence is determined by the incorporation of complementary fluorescent-tagged nucleotide terminators. The fluorescent signal from the incorporated terminators is detected by scanning the chip and the individual clusters with a high-resolution confocal fluorescence laser detector after every round of nucleotide incorporation. Panel 3, Data are compiled in a fastq file that is being transferred to the high performance computer (HPC). In the HPC the reads are mapped and compiled in a .BAM file before variants are called listed in a .VCF file. Panel 4, The VCF is finally uploaded to the interpreters in the genomic laboratory for filtration, annotation and prioritization

Similar to other medical developments, the clinical implementation of WGS requires that we closely consider advantages compared to the current practice, as well as the limitations and ethical issues of the technology. In this review, we describe the elements and concerns of WGS in clinical practice. Following the trail of the patient sample, we explain the technological platforms and the data infrastructure as well as the processing and interpretation of the results. Finally, we outline and discuss the clinical applications, guidelines and clinical reporting.

**Whole genome sequencing**

NGS was originally referred to as massive parallel sequencing (MPS) [11] describing the parallel processing and sequencing of millions of DNA fragments in small vesicles or on a solid phase and the subsequent alignment of the sequence reads to a reference genome. The output of NGS has steadily increased since 2005 [8], where it was suitable for sequencing of smaller selected parts of the genome, to WGS that became possible around 2010 and was FDA approved in 2018. The laboratory procedures are relatively simple and can be performed in any conventional molecular biology laboratory. The general WGS workflow is outlined in Fig. 1.

The major difference between WGS and other types of NGS analyses is basically that there is no sequence capture and the amount of data generated. Until a few years ago, the cost of WGS was relatively high, but with the advent of second-generation chips and improved chemistry, the pricing has become comparable to the majority of other clinical diagnostic procedures. There exists a number of different NGS platforms. Each has its particular virtues but from a user perspective, it is meaningful to distinguish between short- [7] and long-read sequencing [13]. Short-read protocols generate reads of <300 base pairs (bp), whereas long-read sequencing can provide uninterrupted reads ranging from 10 kbp to several megabases depending on the technology [13]. Long-read sequencing improves the sequence phasing and it is the preferred method for solving larger haplotypes and detection of complex structural variants and repeats. In comparison short-read sequencing is the most widely applied method for detection of smaller variations because it is fast and provides high -accuracy and -sequencing depth for smaller, as well as, larger variants [14] at a low cost per base. Short reads can also be employed for applications aimed at counting the abundance of specific reads and expression analysis. Whereas, short read instruments are far more common, both platforms are appreciated and, in many laboratories, they supplement each other. Procedures are being developed that will facilitate the generation of long reads on short-read instruments, underscoring the complementarity of the methods. Nowadays short-read WGS protocols routinely provide 10 times (10X) coverage of more than 95% of the human genome and a median coverage of 30X in a single analysis, and this is generally considered sufficient for germline analysis. In order to identify minority clones, tumour analysis requires about 90X coverage. WGS is normally performed as paired-end sequencing, which enables more accurate read alignment and detection of structural rearrangements. Current, WGS protocols take approximately four working days and they are less labour-intensive than panel or exome sequencing due to the absence of the capture and amplification step.

Due to the impressive technical performance of the many commercial solutions and the defined laboratory procedures, clinical WGS workflows can be accredited according to ISO 15189. Great efforts are made to automate procedures, since sample exchange is a significant source of error. Because WGS is unlikely to be repeated, and may be reanalysed if new clinical insights or causes of a particular disease are discovered, it is crucial to reduce the risk of sample exchange. The frequency of

sample exchange is incompletely documented, but based on our experience from panel sequencing, we estimate that it occurs in approximately 1 out of every 3000 samples. To mitigate the risk of sample exchange, we recommend that single nucleotide polymorphism (SNP\_ID) surveillance is included for all WGS samples. This means that an independent patient sample undergoes panel analysis of a small number of highly polymorphic SNPs in parallel with the WGS sample, and that WGS data are only released for interpretation if the IDs match, and only match, the same individual. Additionally, manual pipetting steps may be video monitored to enable the tracking of sample mixing. These measures have not only improved the detection of sample exchanges in the laboratory, but also prior to arrival at the facility. Moreover, they provide an additional check for the correct family identification of trio samples.

### Bioinformatics

WGS requires a robust computational infrastructure to ensure fast and reliable data processing [15]. While the turn-around-time for patients with stable conditions may not be critical, neonates or patients in unstable and severe conditions may require prompt analysis. Also, tumour analysis should also be swift in order to begin treatment as soon as possible [16]. Consequently, clinical WGS pipelines must fulfil a set of requirements concerning both the physical computational and the software application infrastructure. The challenge is illustrated by the amount of data produced by WGS compared to large gene panels or exomes. Whereas, panel and exome analyses generate about 0.15GB and 5GB raw data, the output of a WGS analysis is about 30GB. The corresponding variant files (.vcf) from gene panels or exomes are about 7E-05GB and 0.04GB, whereas, WGS come near 1GB which corresponds to an increase in data of 13.000- and 24-fold, respectively.

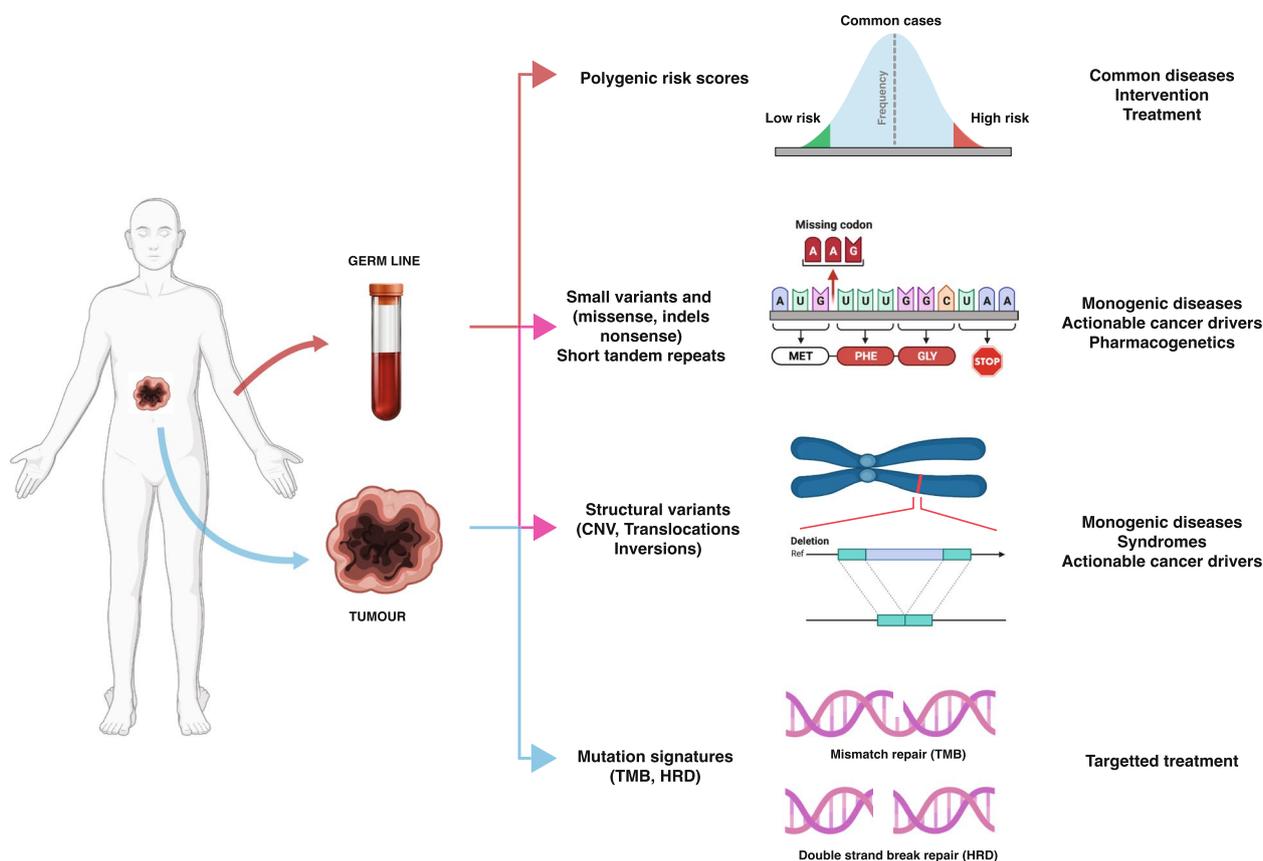
Figure 1 depicts the three most important steps in the data analysis pipeline: 1. mapping, 2. calling and 3. Interpretation. Interpretation, is in principle independent of the variant calling and is performed by dedicated staff using third-party software with a graphical interface that enables interactive and flexible sorting annotation and filtering of the data. The creation of standardised end-to-end variant calling workflows was pioneered by the open-source Genome Analysis Tool Kit (GATK) [17], which forms the basis for many clinical, academic, and national WGS centres. However, a number of commercial hardware-accelerated solutions such as DRAGEN™ and Sentieon® [18], as well as prediction-based approaches

are also available [19, 20]. None of these solutions are plug-and-play, and centres performing large-scale WGS analysis should be prepared to participate in pipeline development and maintenance to provide a safe, reliable and updated analytic environment.

In a production environment considerable engineering effort is dedicated to data handling, such as book-keeping of IDs and linking clinical metadata. From these, at times complex, sources of information it is possible to automate a specific pipeline run, and transfer a tailored set of output files to their proper destination. The data management includes renaming files, generating delivery notifications, logs, archives and clean-up of hundreds of intermediary files. In a clinical environment the system integration needed for the correct information flow often crosses multiple firewalls, domains and databases, and daily operation depends on support from a clinical production grade IT-organisation. Pipeline managers like *snakemake* [21] or *nextflow* [22] are important to orchestrate jobs and processes in the pipelines which may consist of several hundred steps - each with distinct resource requirements and parallelisation potential. In this environment commercial hardware-accelerated solutions that runs each sample serially can sometimes experience problems and tools that can run in parallel based on generic computers may be faster for the last finished sample on a high-performance computer cluster (HPC). More recent sequencing machines with build-in data processing hardware and closed end-to-end workflows may also bring limitations on how to reprocess samples and integrate historic data to advance diagnostics. Since the bottleneck in processing and variant calling from short-read sequencing often is the data-transfer times it is worthwhile to consider the design of the data storage system and the connection to the compute units, as well as cost-efficient storage tiers for active and archived data, respectively. Cloud solutions can be difficult to engineer for fast WGS, because the data is physically generated, and sometimes also physically stored, far from the computation units. Taken together, the initial and very general tasks of demultiplexing pooling barcodes, read alignment and marking of duplicate reads can be performed close to - or inside - the sequencing machine and will result in considerably less data transfer needs, but for more specialised tasks that are impacted by local optimisation and historic background data an HPC or cloud solution is needed.

Test, validation and accreditation is equally critical for bioinformatics production as it is for laboratory. For germline variant calling, initiatives like the Genome in a Bottle project have made it possible to benchmark and

optimize tools, and there are even competitions from the American Food and Drug Administration (“FDA challenges”) in place to encourage such optimization. However, there is still no established reference for somatic variant calling. While the 1+ Million Genomes initiative [23] and the Somatic Mutation Working Group of the Sequencing Quality Control Phase II Consortium [24] have begun to address this building a community standard truth set of somatic variants remains a challenging task. Instead, in-house data comprising hundreds of manually curated somatic mutations must be reanalysed each time a new modality is implemented. A similar need of standard exists for detection of copy number alterations and inversions, and it is still a major challenge to call these in bioinformatic pipelines. Current tools are unable to detect all CNVs [25, 26], and because each algorithm has a specific recall bias so the only viable solution is to combine tools with different strategies. Since the output contains thousands of called variants, most of which could be correct but are not clinically relevant, it is also necessary to employ a large background panel from uniformly processed historic in-house samples to remove irrelevant calls. Correspondingly, somatic variant callers like Mutect2 (<https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2>) and GATK-gCNV [27] also rely on pre-processed background cohorts in a panel of normals, and it is recommended to avoid using public data because it may have a different noise modality. Most clinical bioinformatic units therefore relies on the access to a large harmonized in-house database of historic patient data. As described below polygenic scores and somatic mutations signatures are also expected to become part of the WGS pipelines. Regardless of the computational method the calculations are highly dependent on the sequencing platform, library preparation, sequencing depth and variant calling and filtering pipeline. Consequently, computations of mutational signatures [28] and polygenic scores [29] (see below) should be interpreted with great caution - and always - in relation to a scale of historic cases potentially blinded as quartiles if per-sample information cannot be displayed. In the very last step of the bioinformatics pipeline, it should also be recalled that most interpretation softwares do not require filtering before uploading and e.g., filtering on genome frequencies [30] should only be applied in the analysis software by the clinical interpreter as a conscious decision. Finally, as always - it is important to underscore that clinical data are sensitive and data privacy and safety should be highly prioritized in the WGS bioinformatics solutions.



**Fig. 2** Clinical applications of WGS. Whole Genome Sequencing (WGS) finds its primary clinical applications in diagnosing rare diseases and pinpointing actionable somatic variants within tumors. Beyond these crucial roles, WGS serves to unveil polygenic risk scores (PRS) and pharmacogenetic profiles. The spectrum of rare diseases and somatic variants encompasses both small and structural variations, all discernible through WGS data analysis. WGS also enables the identification of trinucleotide repeat expansions prevalent in neuro-muscular and degenerative diseases. Additionally, it sheds light on polygenic and pharmacogenomic profiles, elucidated by the presence of widespread small common variants. In a comprehensive approach, WGS not only captures the intricate details of genetic makeup but also unveils tumor signatures by deciphering distinctive patterns within somatic variants. Human insert was created with BioRender.com

**Data filtering and interpretation**

Whole-genome sequencing (WGS) is widely employed to diagnose rare [31–35] and undiagnosed diseases [36] and identify actionable cancer drivers and signatures. The different clinical applications and the type of analyses that are implicated in the diagnostics are shown in Fig. 2. There are several reasons why whole-genome sequencing (WGS) is becoming the preferred method for genetic analysis over alternative methods such as panel and exome sequencing. Firstly, WGS detects more variants not only in the large noncoding parts of the genome but also in exons due to a superior mapping quality [31, 36, 37]. Secondly, WGS captures copy number variations and

structural rearrangements as well as mutation signatures and polygenic scores. Finally, WGS can be considered a lifelong investment that may be revisited for different clinical purposes and reanalysed when novel pathogenic variants and disease-causing genes emerge [38]. It has been estimated that about 250 new disease genes are discovered every year, and that up to 6000 Mendelian conditions remain to be discovered [39, 40]. As a direct illustration of the situation, it is worth mentioning that almost half of the variants identified in the recent UK and Ireland rare pediatric disease WGS study [37], were unknown by the time the study was initiated. The relation between human genetic variation and disease is summarised in Text Box 1.

### Text Box 1 Human genetic variation and disease

The human genome is composed of 3.2 billion base pairs of DNA, organized into 23 pairs of chromosomes [2–4]. In addition to the nuclear genome there is also a small amount of maternal DNA located in the mitochondria. Only about 1.5% of the genome sequences consists of protein coding exons [41]. The remaining 98% of the genome is made up of non-coding regions, which include regulatory elements, repetitive DNA sequences, and other functional elements [42]. While there is general consensus that we have about 20,000 protein-coding genes, the size of the proteome is still debated [41, 43]. Moreover, a number of non-coding RNAs such as micro RNAs and long non coding transcripts are also produced but their number and biological significance are with few exceptions uncertain [44].

Like any other species humans are under constant selection and genetic variation is an integral part of the evolution. We continuously acquire both positive adaptive germ cell mutations as well as neutral and disease causing variants [45]. Mutations result from radiation, environmental stress factors and deficient DNA repair [46] and they locate to all parts of the genome [47] albeit with varying frequency. On average a human genome accumulate about 75 mutations per generation [45]. Dominantly inherited variations leading to lactase persistence has for example allowed adult northern Europeans to digest milk [48] and the caspase-12 gene is polymorphic for a stop codon, that makes carriers more resistant to severe sepsis. We can also observe how the Black Death shaped genetic diversity around particular immune loci such as *ERAP2* and *CTLA4*, highlighting how natural selection may have played a role in present-day susceptibility towards chronic inflammatory and autoimmune disease [49]. Finally, it is clear that genes encoding transcription factors and RNA binding proteins which are essential for fetal development are subject to a strong selective pressure as illustrated by their low or entirely absent occurrence of loss of function variants [30]. In genetic terms humans are 99.9% identical to each other. The remaining 0.1% of our genome corresponding to ~ 3,000,000 simple variants distinguish us from another. Among these ~ 45,000 (1.5%) are found in protein coding exons [2]. In addition, numerous structural variations, such as copy number variations (CNVs) and structural variations (SVs) may contribute to our genetic diversity [50, 51]. From a medical perspective, this genetic variation significantly influences individual susceptibility and disease development. The impact extends to pharmaceutical side effects and clinical outcomes, underscoring the integral role of genome sequencing in personalized medicine.

Both rare (< 1% minor allele frequency) and common variants (> 1% minor allele frequency) contribute to the risk of developing a disease, and they can sometimes interact with each other in complex ways. From a diagnostic point of view this is one of the major challenges for the current interpretation of WGS data. Common variants are typically associated with a small increase in disease risk, but because they are so common, they can have a significant impact on the population as a whole. At the individual level the presence of numerous common variants may generate a significant risk for a particular disease and their cumulative effect is captured by the current polygenic risk scores (PRS). Rare disease associated variants, on the other hand, with few exceptions occur at a much lower frequency in the population, often far less than 1% of individuals. Most of the rare variants that are considered in diagnostics locates to the coding exons and alters or reduce the function of the encoded proteins. In families they exhibit a mendelian segregation pattern in the families, but they may also occur as de novo variants. During the past decade genome-wide association studies (GWAS) have associated thousands of common-variants to various diseases and traits, and in the same a series of large-scale sequencing studies have recently started to identify rare-variant associations [52–54]. A surprising finding has been that for a particular trait, common and rare variants appear to be mechanistically convergent [55]. The relative contribution of rare variants to the total genetic burden may be relatively small but rare variants may serve to improve the fundamental understanding of the disease pathogenesis and define possible targets of treatment.

#### Rare monogenic disorders

The output of a single WGS is about 5 million variants and the data interpretation begins by importing variants (.vcf files) into one of the many commercially available or in-house designed software tools that makes it possible to filter and annotate the variants. Filtrations, include exclusion of variants based on their quality, population frequency, functional impact and clinical relevance, in order to focus on variants with a putative causal role for the patient's disease. A number of analytical approaches and filtering schemes have been put forward by various expert groups and initiatives and these may serve as a fine starting points for the interpretation units [56–60]. Figure 3 provides an example of a filtering scheme and how it affects the selection of variants.

In principle the analytic strategy may be genotype-driven or symptom/disease (phenotype)-driven [56]. Genotype-driven analyses are focused on the identification of pathogenic variants loss of function variants, whereas the symptom/disease driven analyses focus on variants that are compatible with the inheritance pattern. There is no strong delineation between the two approaches and they are often combined. In cases where the diseases have a well-defined symptomatology an *in-silico* gene panel of known disease related genes can moreover be applied at an early stage to focus the analysis even further. In this way the exact analytical approach

depends on the clinical presentation and whether the patient represents an isolated case or has a familial predisposition.

For children with healthy parents, a trio examination can be performed to identify pathogenic de novo heterozygous or compound heterozygous variants that are compatible with the clinical diagnosis [61]. The diagnostic success is higher for trios than singletons and usually only 10–30 variants have to be scrutinized [37]. In cases with a familial predisposition, relevant affected and healthy family members can be included to subtract variants from healthy subjects and focus on shared variants in the probands and affected family members. Analysis of singletons is the most variable and challenging. Approximately 90% of variants are common variants with a frequency greater than 2% and these are typically filtered out. Known pathogenic variants should obviously be retained for downstream analyses (Fig. 3C). The remaining ~ 500,000 variants may be further filtered based on minor allele frequency and their location and significance focusing on nonsense, indels, proximal splice-site, and missense variants with a frequency below 1% or 2%. This normally reduces the number of variants to around 2500 or fewer, especially, if combined with relevant gene panels (Fig. 3A). About 40–80 variants normally represents pathogenic loss of function variants (LOF) that may be assessed directly (Fig. 3B). Factors such as ethnicity or

founder effects occasionally warrant changes to the general filtering scheme and it is important to note that the expected frequency of a pathogenic variant in the population depends on the penetrance of the variant or gene.

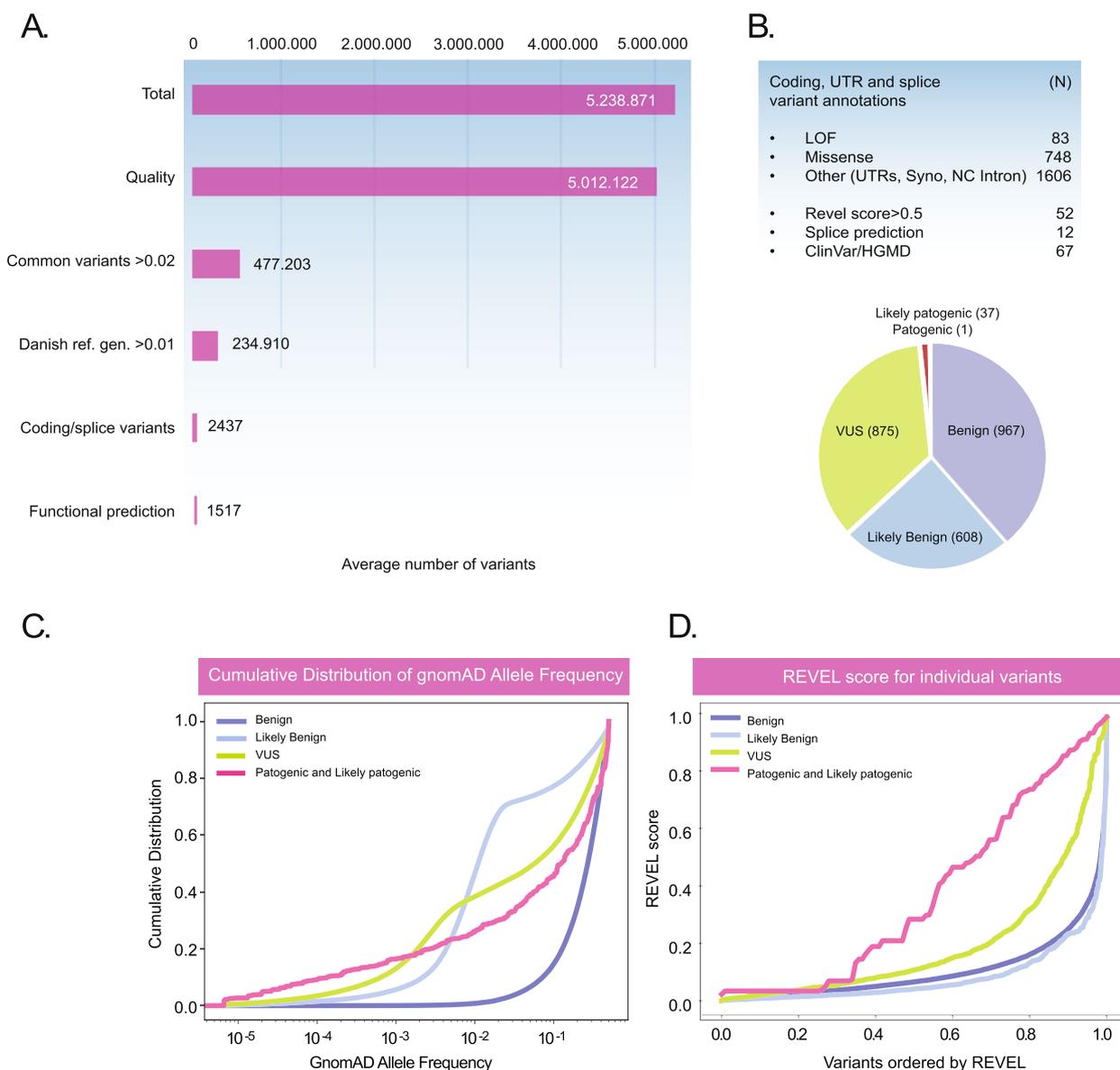
The ACMG/AMP classification criteria [59, 62] are widely used for prioritizing variants based on their pathogenic significance. Based on characteristics such as allele frequency, case data, functional data, and data sources, variants are categorized into five classes: 1. benign, 2. likely benign, 3. variant of uncertain significance (VUS), 4. likely pathogenic, and 5. pathogenic. The prioritization of VUS and putative pathogenic variants involves several considerations. As shown in Fig. 3C, allele frequencies are not very discriminative between VUS and benign variants and a number of other features needs to be considered in order to classify VUS. It is obviously important if the variant has been observed in other patients and whether there is direct evidence linking the variant to the patient's disease or symptoms. This information can sometimes be obtained from databases such as The Human Gene Mutation Database (HGMD) or ClinVar (Table 1), or from the scientific literature. Moreover, the presence of homozygous individuals in population databases such as The Genome Aggregation Database (*gnomAD*) may support that the variant is benign. Additionally, search engines like PubMed, OMIM, and Find Zebra are also useful in establishing the significance of a variant or gene. Many commercial software tools even offer access to knowledge databases, providing more systematic reviews of the literature and databases that allow the interpreter to narrow down genes and variants associated with particular diseases or symptoms. Finally, predictive functional scores such as the REVEL score [63] (Fig. 3D) and the recent AlphaMissense prediction

tool [64] (see below) are likely to play a larger role in the future. Note that the available database solutions are not standardized or accredited, and it is important that the interpreter document the reasons for the classification of a particular variant. If the analysis fails to identify an association between a gene and a disease, the molecular pathway in which the protein functions may eventually be considered. Pathway analysis is still in its early stages, and associations should be confirmed by functional analysis to support that a variant is in fact pathogenic. Finally, it is important to mention that VUS and even clear loss of function variants sometimes are located in genes of unknown significance (GUS). GUS are defined as genes without validated association with a given phenotype [59] and as a result of the uncertainty current guidelines recommend that any variant in GUS is reported as VUS. Rare, predicted damaging variants in GUS are obviously of great interest because they may eventually lead to the discovery of new disease gene. It is important that they are reported to relevant databases such as the Matchmaker Exchange that promote Genomic discovery through the exchange of phenotypic & genotypic profiles [65] ([www.matchmakerexchange.org](http://www.matchmakerexchange.org)) or even for improved functional annotation in MaveDB [66].

From a clinical standpoint, VUS obviously represent a dilemma because their causative role in a particular disease is not fully established. Some argue that VUS should simply be eliminated from the analysis [67, 68] and await further evaluation, while others emphasize the risk of leaving patients without a diagnosis if a clinically relevant VUS is disregarded. The number of VUS will likely decrease over time when databases accumulate more data and our understanding of disease pathogenesis improves. Currently, there are no definitive guidelines for

(See figure on next page.)

**Fig. 3** Variant analysis of patients with rare diseases. Panel **A** Overview of the filtering steps and the number of variants in rare disease patients referred for WGS analysis (means of 6 patients). The total number of variants in each patient is just above 5 MIO. The analysis begins by elimination of ~200,000 low quality variants. Subsequently, common variants with an allele frequency above 2% are excluded, since these are considered unlikely to explain the occurrence of a rare disease. Known pathogenic variants are retained. Since *gnomAD* may not represent all common variants, variants are moreover filtered against a local (Danish) reference genome and this further reduces the number of variants to about 200,000. Thereafter, the analysis is focused on coding and splice site variants and on average this reduces the number of variants to ~2400. Application of additional filters e.g., omitting ACMG/AMP benign variants or those with low REVEL scores further brings the number of variants down to ~1500. Panel **B** On average the patients exhibit 83 loss of function (LOF) variants and 748 missense variants. The remaining variants belonged to other categories such as variants in the UTRs and deep into the intron. Finally, on average 67 variants were previously registered in ClinVar or HGMD and information on these can be readily retrieved and used in the interpretation. The pie chart below shows the ACMG/AMP classification of the variants showing that only a minority are classified as pathogenic and likely pathogenic (<2.5%). On average only a single pathogenic variant is identified. In many cases the variant represents a recessive heterozygote variant with no obvious relevance for the patient's disease. Almost one third of the variants represents variants of unknown significance (VUS). Panels **C** and **D** shows the total cumulative distribution of *gnomAD* allele frequencies and REVEL scores of ACMG/AMP scored variants (from Varseq) among 63 unrelated patients, respectively. Intergenic variants were filtered away and any variant which had conflicting classifications was removed. Moreover, variants with an allele frequency of more than 0.5 or for which an allele frequency could not be found was removed. The results illustrate that allele frequency is relatively effective in excluding benign variants, whereas likely benign and VUS are not effectively separated from the likely pathogenic and pathogenic variants by frequency filtering. The REVEL score combining pathogenicity predictions from 18 individual scores, in contrast, is clearly discriminative and high scores are enriched among pathogenic variants. About 25% of the VUS exhibit REVEL score above 0.5 that may warrant further analysis of these variants. The number and details of variants in the plots is summarized the attached [Supplemental data](#)



**Fig. 3** (See legend on previous page.)

all clinical situations, so common sense and clinical experience are important. In many WGS centers, variants are discussed with the attending physician or in multidisciplinary teams to ascertain their clinical relevance. In general, only pathogenic (class 5) and likely pathogenic (class 4) variants are included in the final clinical report. Finally, it is convenient to include details about the sequencing and analysis method used and the composition of in silico gene panels in the report for future reference. Figure 4 illustrates the general scheme of clinical WGS reporting.

**Somatic variant analysis**

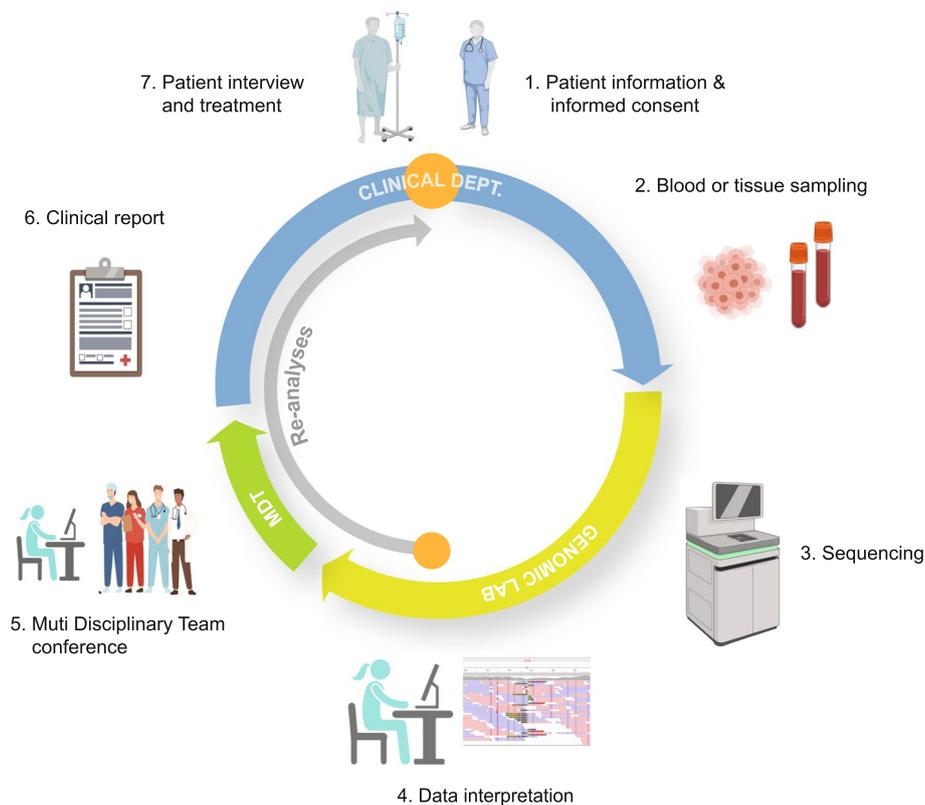
WGS of tumour and germline DNA in combination with RNA sequencing-based expression analysis is widely

used to identify actionable tumour drivers and host factors. WGS is the preferred method for tailored treatment because it potentially uncovers both the small somatic tumor variants, CNVs and facilitate the detection of characteristic mutation signatures such as HRD and TMB. The complete map of somatic mutations and alterations in gene expression patterns provides integrated information for selection of the optimal treatment.

Somatic variant calling requires a whole blood sample for germline variants and a tumour sample for somatic variants and transcriptome analysis. Somatic variants are identified by subtracting germline variants from the tumour sequence. It is not recommended to exchange the blood sample for a panel-of-normals germ-line variant set because of the

**Table 1** Biomedical databases relevant for clinical WGS

Database/Resource	Web address	Content
ClinGen	<a href="https://clinicalgenome.org/">https://clinicalgenome.org/</a>	Clinical relevance of genes and variants
ClinVar	<a href="http://www.ncbi.nlm.nih.gov/clinvar/intro/">www.ncbi.nlm.nih.gov/clinvar/intro/</a>	Database of genomic variants with public submissions of variant interpretations and disease relations.
Cosmic	<a href="https://cancer.sanger.ac.uk/cosmic">https://cancer.sanger.ac.uk/cosmic</a>	Catalogue Of Somatic Mutations In Cancer
dbSNP	<a href="https://www.ncbi.nlm.nih.gov/snp/">https://www.ncbi.nlm.nih.gov/snp/</a>	Contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions
Ensembl	<a href="https://www.ensembl.org/index.html">https://www.ensembl.org/index.html</a>	Genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation
Find Zebra	<a href="https://www.findzebra.com/">https://www.findzebra.com/</a>	Tool for helping diagnosis of rare diseases. It uses freely available high quality curated information on rare diseases
Genomics England	<a href="https://www.genomicsengland.co.uk/">https://www.genomicsengland.co.uk/</a>	Comprehensive site describing the progress of the UK sequencing initiative. Site contains useful overviews over gene panels and diseases.
Geo		Repository supporting MIAME-compliant data submissions. Array- and sequence-based data
gnomAD	<a href="https://gnomad.broadinstitute.org">https://gnomad.broadinstitute.org</a>	Exome and genome sequencing data with allele frequencies from a wide variety of large-scale sequencing projects
GTEX	<a href="https://gtexportal.org/home/">https://gtexportal.org/home/</a>	Comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq.
HGMD	<a href="https://www.hgmd.cf.ac.uk/ac/index.php">https://www.hgmd.cf.ac.uk/ac/index.php</a>	Collate all known (published) gene lesions responsible for human inherited disease
Human Phenotype Ontology (HPO)	<a href="https://hpo.jax.org/app/">https://hpo.jax.org/app/</a>	Provides a standardized vocabulary of phenotypic abnormalities encountered in human disease
Matchmaker Exchange	<a href="https://www.matchmakerexchange.org">https://www.matchmakerexchange.org</a>	Genomic discovery through the exchange of phenotypic & genotypic profiles
MaveDB	<a href="https://www.mavedb.org/">https://www.mavedb.org/</a>	Collection, distribution, and analysis of variant effect maps
MedGen	<a href="https://www.ncbi.nlm.nih.gov/medgen/">https://www.ncbi.nlm.nih.gov/medgen/</a>	Organizes information related to human medical genetics, such as attributes of conditions with a genetic contribution
NCBI	<a href="https://www.ncbi.nlm.nih.gov/">https://www.ncbi.nlm.nih.gov/</a>	The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information
OMIM	<a href="https://www.omim.org/">https://www.omim.org/</a>	Compendium of human genes and genetic phenotypes
RefSeq	<a href="https://www.ncbi.nlm.nih.gov/refseq/">https://www.ncbi.nlm.nih.gov/refseq/</a>	A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.
The Cancer Genome Atlas Program (TCGA)	<a href="https://www.cancer.gov/ccg/research/genome-sequencing/tcga">https://www.cancer.gov/ccg/research/genome-sequencing/tcga</a>	The Cancer Genome Atlas (TCGA) has molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types.
UCSC genome browser	<a href="https://genome.ucsc.edu">https://genome.ucsc.edu</a>	Interactively visualize genomic data
Uniprot	<a href="https://www.uniprot.org/">https://www.uniprot.org/</a>	Comprehensive and freely accessible resource of protein sequence and functional information.



**Fig. 4** WGS from patient to clinical report. The journey of Whole Genome Sequencing (WGS) commences and concludes at the patient's bedside. Upon the attending physician's assessment, a WGS analysis is deemed potentially beneficial for offering crucial clinical insights, either through diagnosis or by presenting alternative treatment options. Following comprehensive patient briefing and obtaining consent, a sample of whole blood or tumor is dispatched to the specialized laboratory equipped for WGS. Within the genomic laboratory, the sequence data undergo meticulous analysis by the skilled staff. Putative disease-associated variants are subsequently deliberated with the attending physician and, if necessary, a multidisciplinary team comprising medical professionals from pertinent specialties, forming a Multidisciplinary Team (MDT). Specialties include pathology, clinical genetics, immunology, and more. This collaboration aims to establish a conclusive diagnosis and assess the clinical relevance of identified variants. The conclusive clinical report is then transmitted to the clinical department, where the attending physician shares the results with the patient. This communication includes a comprehensive discussion of the implications for the patient and their condition, along with recommended actions. In instances where the initial analysis fails to pinpoint disease-causing variants, the stored WGS data undergoes periodic re-analysis (inner grey arrow). This ongoing process ensures the continuous integration of new knowledge, potentially leading to a diagnosis without the need for additional hospitalization and sampling. Furthermore, throughout the treatment course, various clinically relevant information, such as pharmacogenetics, may be extracted to enhance the overall patient care experience. Inserts were created with BioRender.com

higher noise level. Typically, tumours exhibit about 500,000 somatic variants and as described for the germ line analysis the variants undergo a series of filtering's based on their frequency, call quality and read depth as well as their cancer relevance before interpretation. After filtration between 20 and 1500 variants are normally eligible for further evaluation. Based on their significance in cancer, prognosis, and/or therapeutics somatic variants may be classified into four tiers. Tier I, represents variants with strong clinical significance, Tier II variants with potential clinical significance and Tier III variants of unknown clinical significance whereas Tier IV is benign or likely benign variants [60]. Actionable somatic variants are subsequently be queried in relevant databases [69, 70]. Many laboratories also report

the tumour mutation burden (TMB) score that is associated with immune cell infiltration and increased sensitivity to programmed cell death-1 (PD-1) or PD-1 ligand (PD-L1) blockade. Finally, a homologous recombination deficiency (HRD) signature linked to poly(ADP ribose) polymerase (PARP) inhibitor sensitivity [71–73] may also be generated from the WGS data.

#### Polygenic risk scores

Genome-wide association studies have revealed that common disorders such as type 2 diabetes, cardiovascular diseases, and some cancers, are associated with combinations of common variants each providing a small increase in risk for the particular disease [74–78]. The polygenic risk

burden is combined into a polygenic risk score (PRS) that can support diagnosis, screening, and intervention at early stages of disease. The number of variants included in the PRS can range from a few (<10) to thousands of variants, and while the discriminative ability of PRS in the general population has been debated, larger and more diverse studies, as well as refined computational strategies, have revitalized the clinical interest in PRS [29, 76–78]. Cheap chip-based assays are useful for PRS analyses, but WGS may become an appealing alternative because it will identify both common and rare variants that potentially may contribute to the genetic makeup of a diseases. Extraction of data for individual PRS can be integrated into the WGS pipeline and added automatically to the clinical report, providing a comprehensive genetic profile of the patient.

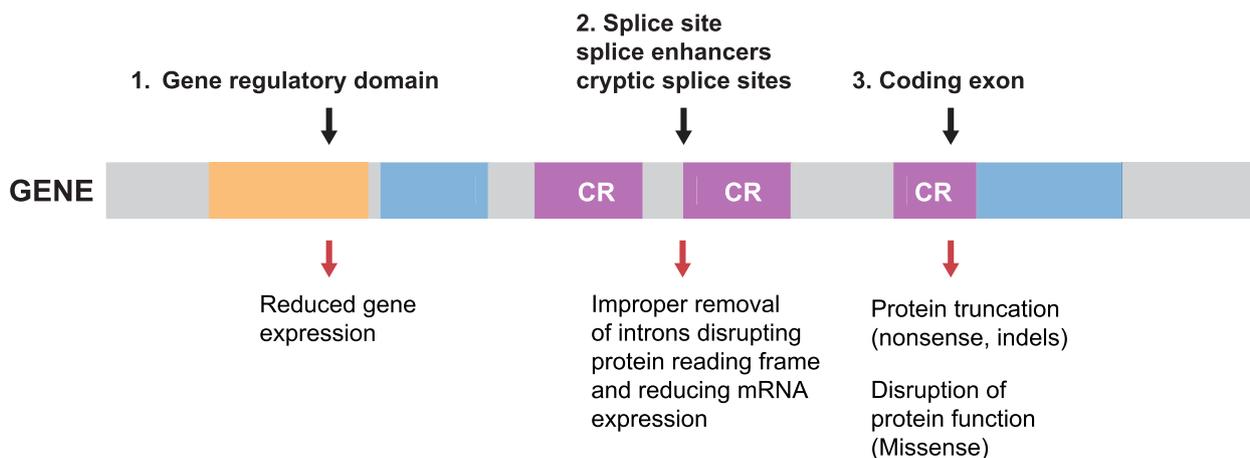
**In-silico prediction and functional testing of variants**

With the increasing diagnostic sequencing and identification of new disease genes, the number of VUS that needs to be considered will increase [68]. Consequently, there is great focus on *in-silico* and *in vivo* analyses to better understand the significance of these variants. Figure 5

provides a schematic representation of the functional consequence of various types of mutations.

**Predictive scores and protein structure**

Missense variants are commonly assessed based on their frequency, conservation, and the location and type of amino acid substitution. Predictive scores that take this information into account are being developed, and among the most widely used are Polyphen [79], SIFT [80], and CADD [81]. The REVEL score, in particular, combines scores from a wide range of tools and provides a relatively high enrichment of pathogenic variants [63]. Precomputed REVEL scores are available for all possible human missense variants and can be integrated into the clinical analyses. With the rapid accumulation of AI-driven protein structures in the AlphaFold Protein Structure Database [82], many hoped that structural predictions could be used for assessment of Variants of Uncertain Significance (VUS). Although, initial attempts were not entirely successful [83, 84], the recent AlphaMissense (AM) algorithm, represents a major leap forward [64]. AM integrates information of evolutionary conservation and protein structure - both of which are intimately linked to protein function - and classifies variants as



**Fig. 5** Genomic localization of variants and their functional consequence. 1. Germ-line variants located in the gene regulatory domains such as promoters or locus control regions will affect the level of gene transcription. In most instances variants in the promoters disrupt the binding of trans-acting factors thereby reducing expression of the gene. The composition of regulatory motifs is in many instances incompletely understood and it is in general difficult to predict the consequence of these variants. A few diseases exhibit unstable trinucleotide repeat sequences in the promoter, that when expanded is known to impair transcription. The functional significance of promoter variants is normally demonstrated by loss of expression (LOE) via RNA sequencing or measurement of the encoded protein. Repeat expansions may also be directly discerned from the WGS data 2. Variants located at the canonical splice donor (GT) or acceptor (AG) sites or at a known A – branch-site are in general pathogenic since these strongly conserved sequences are essential for splicing. Variants located deeper in the intron or in the connecting exons can also disrupt splicing due to disruption of enhancer or silencer motifs but the significance of these variants is more difficult to predict. The evaluation of these variants in general requires minigene analysis and/or RNA sequencing. 3. Coding nonsense or frameshift variants lead to premature translation termination and shortening of the encoded protein. In most cases this can lead to loss of function (LOF). Missense variants and small indels may disrupt protein function in a number of different ways such as reducing enzymatic activity, stability, localization or structure and macromolecular assembly. Consequently, the evaluation of these variants requires deep insight into the proteins function and in many instances various kinds of functional analysis is necessary in order to classify the variants as pathogenic. Since the functional significance of a particular variant may be difficult to predict - even for canonical splice mutations and LOF variants - it recommended that all classes of variants undergo evaluation according to ACMG/AMP criteria in order to determine pathogenicity

likely -pathogenic or -benign. The precision of AM is thus far unmatched and the algorithm holds great potential to facilitate the classification of VUS.

### mRNA expression and splicing

The processing of primary RNA transcripts from transcription to translation and decay involves a series of well-characterized steps that can be affected by both coding and intronic variants (Fig. 5). In addition to the canonical GT-AG donor and acceptor sites, variants may involve exonic splice enhancers and/or intronic silencers or generate novel splice sites. The percentage of variants that affect pre-mRNA splicing varies among diseases ranging from 10 to 50% (reviewed in [85]) and studies have indicated that as many as 25% of exonic mutations may have an effect on splicing [86, 87]. RNA sequencing reveals the expression of individual alleles and the exonic composition of the transcripts and may uncover that coding variants are located in exons failing to be expressed in the relevant tissue. Calling of fusion genes from RNA-seq data is also important. In particular for cancer diagnostics, because the fusion protein may be targeted by drugs. Given the relatively poor accuracy of fusion gene calling [88] it is recommended to use a number of fusion calling tools and rely on a weighted consensus score to prioritise the predictions. For selected clinically relevant fusions a whitelist may even be incorporated in the consensus calling so low frequency targetable fusions are not overlooked. Finally, minigene analysis remains a paradigm for the functional categorization of splice variations [89]. Several *in silico* prediction tools have also been developed to predict whether a particular variant is likely to affect splicing [90–92]. *In silico* prediction cannot stand alone but should prompt further analysis of RNA sequences or minigene splicing.

### Protein function

The classification of a coding VUS should ultimately rely on the characterization of the protein's function. Although, functional testing of an enzyme may be relatively straightforward, complex processes such as homologous recombination requires the assembly and concerted effort of several factors. As a result, there is no one-size-fits-all approach to functional testing and the analysis varies from disease to disease and from protein to protein. Variants implicated in metabolic diseases may e.g., be directly visualized by NMR, whereas disruption of protein assemblies can be examined through conventional pull-down experiments. Dislocations may be visualized through the expression of the factors in suitable cell systems followed by microscopy. Some cell systems, such as induced pluripotent stem cells (iPSc), may even reconcile tissue-specific effects [93]. Many of the

assays are difficult to perform in a routine clinical context, and to solve this problem more systematic screenings of variants are emerging. A recent example of this is the CRISPR-based saturation genome editing screening and classification of over 4000 BRCA1 variants [94, 95], which has facilitated diagnostics of woman with breast ovarian cancer significantly.

### Results from the clinical application of WGS

For rare diseases pediatric- and clinical genetics departments are major requestors, but in principle any medical specialty, may encounter patients with diseases where conventional workup has failed to provide a diagnosis. Large series of patients with rare diseases [31, 32, 35, 36, 96] have demonstrated an average diagnostic yield of ~25% for probands. Somewhat over 10% of these diagnoses were caused by variants in genomic regions that would not have been identified by other methods. Moreover, a few percent involved coding variants in regions of low coverage on exome sequencing [31]. The results are in line with data from screening of undiagnosed patients, where about half of the patients who receive a diagnosis from WGS have previously undergone exome sequencing [36]. The diagnostic yield varies across different patient groups, ranging from a few percent for respiratory and some hematological disorders to 40–50% for hearing and ophthalmologic disorders, intellectual, and neurodevelopmental disorders. For patients with heart disease or immune deficiency, the diagnostic yield is 20–30% [31]. In a recent study of rare paediatric disorders - a diagnosis was made in about 40% of the probands of whom 76% exhibited a pathogenic *de novo* variant [37]. The diagnostic yield is highest among probands analysed in trios and for patients with more pronounced symptoms. On average 2.5 and 1 candidate variant were reported in singletons and probands analysed as part of trios, respectively. Children with intellectual disability, neurodevelopmental disorders, and complex syndromes usually require a complex diagnostic workup, and since the WGS results are positive front-loading of the analysis during the diagnostic work-up have been recommended [97–99]. Another important experience from the use of WGS is moreover that the analysis may uncover unique presentations of known diseases or a completely new disease. In this way WGS may have a significant influence on future disease classification and identification of novel syndromes.

For oncological patient's comprehensive tumour characterization has demonstrated the effectiveness of tumour sequencing in conjunction with transcriptome analysis to support targeted treatment. WGS uncovers actionable tumour variants in approximately two thirds of metastatic tumors but it should be underscored that there is large variation among tumor types [69, 100–103]. In addition,

germ line sequencing has revealed that a significant number of cancer patients carry predisposing mutations in tumour-suppressor genes [104–106]. The combination of tumour and germ line sequencing has significant potential for improving patient outcomes in cancer treatment, although, there is a strong need to improve the prioritization and characterization of variants in order to increase the response rate of the new drugs.

### Ethical concerns

Like any other medical tests, genome sequencing, raises ethical dilemmas for the society and patients. A number of the concerns such as privacy and confidentiality issues, consent, patients psychological stress, involvement of biologic relatives, social stigmatization, insurance and employment issues are shared with genetic testing in general [107–110]. Genome sequencing, however, also presents a few unique challenges due to the vast amount of information generated. We may not be in a position, where we can fully understand the implications of the data and there is moreover greater potential for incidental findings. This demonstrates the necessity of in-depth information to the patient prior to the analysis (Fig. 4). Moreover, the permanent and complete nature of the data makes it difficult to foresee future applications and dilemmas for the patients [111, 112] (CADTH report). Finally, privacy concerns and data-sharing issues are more challenging because data management often involves third parties outside the health-care systems. It is important that health-care providers take responsibility for safe data storage and prevention of unauthorized use of patient data. Since WGS technology is relatively new and is relevant in many medical specialties, we also want to highlight the importance of proper guidelines and education of the staff in general. MDs and nurses close to the patients should be comfortable with the technology in order to inform the patients.

### The way forward

After the initial discovery and great expectations there is often a period of debate before the benefits of new technologies become evident. It is sometimes argued that WGS produce too much data that we are unable to interpret. In some way this is correct, but in our opinion, it should be regarded as an opportunity rather than a problem and prompt us to increase our efforts to understand disease pathology and genetics even deeper. One of the most important objectives for the fields is to improve variant interpretation and annotation. This will require integration of clinical data, functional studies, population databases, and extensive data sharing and development of computational tools. WGS data should moreover

be further integrated with transcriptomics, epigenomics, and proteomics, in order to provide a more comprehensive understanding of disease mechanisms in Text box 1. By combining multiple layers of genomic information clinicians will be able to identify functional variants, regulatory elements, and pathways associated with diseases, enabling more accurate diagnoses and targeted treatments. Compared to a number of conventional methods WGS has also been considered expensive and to require huge storage capacity. The need for storage and high-performance computing is a concern but should perhaps be perceived in a broader context and regarded as an investment in precision medicine. Moreover, the high-performance computing infrastructure will facilitate a number of the associated research lines and stimulate the integration between clinical care and research. With respect to the clinical use of WGS, there has been a fast progress in the standards for data analysis due to the initiative of e.g., the Medical Genome Initiative [57] and ACMG/AMP as well as patient focused genome initiatives around the world. These efforts should be supported in order to advance diagnostics. Taken together, we are confident that WGS has the potential to make a difference for patients and we foresee that the clinical use will increase in the coming years.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12920-024-01795-w>.

Additional file 1.

### Acknowledgements

Angels Mateu, Caroline Maria Rossing, Ida Kappel Buhl, Majbrit Busk Madsen, Mette Dandanell Nielsen, Mira Marie Laustsen are thanked for helpful suggestions to the manuscript and help with data annotations. The Danish National Genome Center is thanked for collaboration and financial support during the construction the WGS laboratory pipeline. The authors of this review moreover acknowledge support from several public funding parties – including the NOVO Nordic Foundation, The Danish Cancer Society, Region Hovedstaden and Rigshospitalets Forskningspulje.

### Authors' contributions

FOB, LGB, ARH, BB and FCN contributed to writing the manuscript. ASJ and MK contributed to collecting data. FCN conceived the idea for the manuscript and designed figures. All authors contributed to editing and critical review of the manuscript.

### Funding

Open access funding provided by Copenhagen University This manuscript has been produced without external funding. All authors are employed by the Center for Genomic Medicine, Rigshospitalet, University of Copenhagen.

### Availability of data and materials

According to Danish legislation, WGS files are deposited in the Danish National Genome Center (NGC) where they can be accessed after approval from NGC (<https://ngc.dk>). Variant frequencies and REVEL scores are publically available from gnomAD (<https://gnomad.broadinstitute.org>).

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

Received: 14 August 2023 Accepted: 1 January 2024

Published online: 29 January 2024

## References

- Bodmer WF, McKie R. The book of man: the human genome project and the quest to discover our genetic heritage. New York: Scribner; 1995.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
- Nurk S, Koren S, Rhie A, Rautiainen M, Bizkadez AV, Mikheenko A, et al. The complete sequence of a human genome. *Science*. 2022;376(6588):44–53.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001;291(5507):1304–51.
- Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463–7.
- Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. Fluorescence detection in automated DNA sequence analysis. *Nature*. 1986;321(6071):674–9.
- Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53–9.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437(7057):376–80.
- Kaiser J. 200,000 whole genomes made available for biomedical studies. *Science*. 2021;374(6571):1036.
- Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A*. 2015;112(17):5473–8.
- Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*. 2010;11(1):31–46.
- Zhou G, Zhou M, Zeng F, Zhang N, Sun Y, Qiao Z, et al. Performance characterization of PCR-free whole genome sequencing for clinical diagnosis. *Medicine (Baltimore)*. 2022;101(10):e28972.
- Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. 2020;21(10):597–614.
- Choo ZN, Behr JM, Deshpande A, Hadi K, Yao X, Tian H, et al. Most large structural variants in cancer genomes can be detected without long reads. *Nat Genet*. 2023;55:2139–48.
- Grealey J, Lannelongue L, Saw WY, Marten J, Meric G, Ruiz-Carmona S, et al. The carbon footprint of bioinformatics. *Mol Biol Evol*. 2022;39(3):msc034.
- Meggendorfer M, Jobanputra V, Wrzeszczynski KO, Roepman P, de Bruijn E, Cuppen E, et al. Analytical demands to use whole-genome sequencing in precision oncology. *Semin Cancer Biol*. 2022;34:16–22.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43(1110):11–10.
- Franke KR, Crowgey EL. Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for genome analysis toolkit algorithms. *Genomics Inform*. 2020;18(1):e10.
- Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983–7.
- Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA truth challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom*. 2022;2(5):100129.
- Molder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. *F1000Res*. 2021;10:33.
- Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 2017;35(4):316–9.
- Saunders G, Baudis M, Becker R, Beltran S, Beroud C, Birney E, et al. Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat Rev Genet*. 2019;20(11):693–701.
- Mercer TR, Xu J, Mason CE, Tong W, Consortium MS. The sequencing quality control 2 study: establishing community standards for sequencing in precision medicine. *Genome Biol*. 2021;22(1):306.
- Gabrielaite M, Torp MH, Rasmussen MS, Andreu-Sanchez S, Vieira FG, Pedersen CB, et al. A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers (Basel)*. 2021;13(24):6283.
- Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*. 2019;20(1):117.
- Babadi M, Fu JM, Lee SK, Smirnov AN, Gauthier LD, Walker M, Benjamin DI, Zhao X, Karczewski KJ, Wong I, et al. GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nat Genet*. 2023;55(9):1589–97.
- Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in Cancer. *Nucleic Acids Res*. 2019;47(D1):D941–7.
- Cavazos TB, Witte JS. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Hum Genet Genom Adv*. 2021;2(1):100017.
- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434–43.
- Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, et al. 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. *N Engl J Med*. 2021;385(20):1868–80.
- Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemssen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature*. 2014;511(7509):344–7.
- Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B, et al. Whole genome sequencing expands diagnostic utility and improves clinical Management in Pediatric Medicine. *NPJ Genom Med*. 2016;1:1–9.
- Ostrander BEP, Butterfield RJ, Pedersen BS, Farrell AJ, Layer RM, Ward A, et al. Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. *NPJ Genom Med*. 2018;3:22.
- Stranneheim H, Lagerstedt-Robinson K, Magnusson M, Kvarnung M, Nilsson D, Lesko N, et al. Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med*. 2021;13(1):40.
- Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheattle-Jarvela AM, et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. *N Engl J Med*. 2018;379(22):2131–9.
- Wright CF, Campbell P, Eberhardt RY, Aitken S, Perrett D, Brent S, et al. Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland. *N Engl J Med*. 2023;388(17):1559–71.
- Palmer EE, Sachdev R, Macintosh R, Melo US, Mundlos S, Righetti S, et al. Diagnostic yield of whole genome sequencing after nondiagnostic exome sequencing or gene panel in developmental and epileptic encephalopathies. *Neurology*. 2021;96(13):e1770–82.
- Seaby EG, Rehm HL, O'Donnell-Luria A. Strategies to uplift novel Mendelian gene discovery for improved clinical outcomes. *Front Genet*. 2021;12:674295.
- Bamshad MJ, Nickerson DA, Chong JX. Mendelian gene discovery: fast and furious with no end in sight. *Am J Hum Genet*. 2019;105(3):448–55.
- Piovesan A, Antonaros F, Vitale L, Strippoli P, Pelleri MC, Caracausi M. Human protein-coding genes and gene feature statistics in 2019. *BMC Res Notes*. 2019;12(1):315.

42. Elkon R, Agami R. Characterization of noncoding regulatory DNA in the human genome. *Nat Biotechnol.* 2017;35(8):732–46.
43. Tress ML, Abascal F, Valencia A. Alternative splicing may not be the key to proteome complexity. *Trends Biochem Sci.* 2017;42(2):98–110.
44. Ponting CP, Haerty W. Genome-wide analysis of human long noncoding RNAs: a provocative review. *Annu Rev Genomics Hum Genet.* 2022;23:153–72.
45. Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nat Rev Genet.* 2012;13(8):565–75.
46. Cairns J, Overbaugh J, Miller S. The origin of mutants. *Nature.* 1988;335(6186):142–5.
47. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature.* 2021;590(7845):290–9.
48. Swallow DM. Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet.* 2003;37:197–219.
49. Klunk J, Vilgalys TP, Demeure CE, Cheng X, Shiratori M, Madej J, et al. Evolution of immune genes is associated with the black death. *Nature.* 2022;611(7935):312–9.
50. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med.* 2010;61:437–55.
51. Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. *Nature.* 2020;581(7809):444–51.
52. Sun BB, Kurki MI, Foley CN, Mechakra A, Chen CY, Marshall E, et al. Genetic associations of protein-coding variants in human disease. *Nature.* 2022;603(7899):95–102.
53. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK biobank exomes. *Nature.* 2021;597(7877):527–32.
54. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK biobank participants. *Nature.* 2021;599(7886):628–34.
55. Weiner DJ, Nadig A, Jagadeesh KA, Dey KK, Neale BM, Robinson EB, et al. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature.* 2023;614(7948):492–9.
56. Austin-Tse CA, Jobanputra V, Perry DL, Bick D, Taft RJ, Venner E, et al. Best practices for the interpretation and reporting of clinical whole genome sequencing. *NPJ Genom Med.* 2022;7(1):27.
57. Marshall CR, Bick D, Belmont JW, Taylor SL, Ashley E, Dimmock D, et al. The medical genome initiative: moving whole-genome sequencing for rare disease diagnosis to the clinic. *Genome Med.* 2020;12(1):48.
58. Marshall CR, Chowdhury S, Taft RJ, Lebo MS, Buchan JG, Harrison SM, et al. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. *NPJ Genom Med.* 2020;5:47.
59. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. *Genet Med.* 2015;17(5):405–24.
60. Horak P, Griffith M, Danos AM, Pitel BA, Madhavan S, Liu X, et al. Standards for the classification of pathogenicity of somatic variants in cancer (oncogenicity): joint recommendations of clinical genome resource (ClinGen), Cancer genomics consortium (CGC), and variant interpretation for Cancer consortium (VICC). *Genet Med.* 2022;24(5):986–98.
61. Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom Med.* 2018;3:16.
62. Masson E, Zou W-B, Génin E, Cooper DN, Le Gac G, Fichou Y, et al. Expanding ACMG variant classification guidelines into a general framework. *Human Genomics.* 2022;16(1):1–15.
63. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet.* 2016;99(4):877–85.
64. Cheng J, Novati G, Pan J, Bycroft C, Zemgulyte A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science.* 2023;381(6664):eadg7492.
65. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The matchmaker exchange: a platform for rare disease gene discovery. *Hum Mutat.* 2015;36(10):915–21.
66. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* 2019;20(1):223.
67. Vears DF, Niemiec E, Howard HC, Borry P. Analysis of VUS reporting, variant reinterpretation and recontact policies in clinical genomic sequencing consent forms. *Eur J Hum Genet.* 2018;26(12):1743–51.
68. Rehm HL, Alaimo JT, Aradhya S, Bayrak-Toydemir P, Best H, Brandon R, et al. The landscape of reported VUS in multi-gene panel and genomic testing: time for a change. *Genet Med.* 2023;25(12):100947.
69. Zhao EY, Jones M, Jones SJM. Whole-genome sequencing in Cancer. *Cold Spring Harb Perspect Med.* 2019;9(3):a034579.
70. Yang X, Fu H, Ivanov AA. Online informatics resources to facilitate cancer target and chemical probe discovery. *RSC Med Chem.* 2020;11(6):611–24.
71. Fusco MJ, West HJ, Walko CM. Tumor mutation burden and Cancer treatment. *JAMA Oncol.* 2021;7(2):316.
72. Stewart MD, Merino Vega D, Arend RC, Baden JF, Barbash O, Beaubier N, et al. Homologous recombination deficiency: concepts, definitions, and assays. *Oncologist.* 2022;27(3):167–74.
73. Wei C, Li M, Lin S, Xiao J. Characterization of tumor mutation burden-based gene signature and molecular subtypes to assist precision treatment in gastric Cancer. *Biomed Res Int.* 2022;2022:4006507.
74. O'Sullivan JW, Raghavan S, Marquez-Luna C, Luzum JA, Damrauer SM, Ashley EA, et al. Polygenic risk scores for cardiovascular disease: a scientific statement from the American Heart Association. *Circulation.* 2022;146(8):e93–e118.
75. Hahn SJ, Kim S, Choi YS, Lee J, Kang J. Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: a machine learning analysis of population-based 10-year prospective cohort study. *EBioMedicine.* 2022;86:104383.
76. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* 2020;12(1):44.
77. Marston NA, Pirruccello JP, Melloni GEM, Koyama S, Kamanu FK, Weng LC, et al. Predictive utility of a coronary artery disease polygenic risk score in primary prevention. *JAMA Cardiol.* 2023;8(2):130–7.
78. Hao L, Kraft P, Berriz GF, Hynes ED, Koch C, Korategere VKP, et al. Development of a clinical polygenic risk score assay and reporting workflow. *Nat Med.* 2022;28(5):1006–13.
79. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7(4):248–9.
80. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 2012;40(Web Server issue):W452–7.
81. Rentsch P, Schubach M, Shendure J, Kircher M. CADD-splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* 2021;13(1):31.
82. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Zidek A, et al. Highly accurate protein structure prediction for the human proteome. *Nature.* 2021;596(7873):590–6.
83. Keskin Karakoyun H, Yuksel SK, Amanoglu I, Naserikhojasteh L, Yesilyurt A, Yalciner C, et al. Evaluation of AlphaFold structure-based protein stability prediction on missense variations in cancer. *Front Genet.* 2023;14:1052383.
84. Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, Maksimova ES, et al. Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLoS One.* 2023;18(3):e0282689.
85. Lord J, Baralle D. Splicing in the diagnosis of rare disease: advances and challenges. *Front Genet.* 2021;12:689892.
86. Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A.* 2011;108(27):11093–8.
87. Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* 2011;21(10):1563–71.
88. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and

- de novo fusion transcript assembly-based methods. *Genome Biol.* 2019;20(1):213.
89. Breathnach R, Benoist C, O'Hare K, Gannon F, Chambon P. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proc Natl Acad Sci U S A.* 1978;75(10):4853–7.
  90. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol.* 2004;11(2–3):377–94.
  91. Scalzitti N, Kress A, Orhand R, Weber T, Moulinier L, Jeannin-Girardon A, et al. Spliceator: multi-species splice site prediction using convolutional neural networks. *BMC Bioinformatics.* 2021;22(1):561.
  92. Perteau M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* 2001;29(5):1185–90.
  93. Wadmoe K, Azad AJ, Gehmlich K. The role of Z-disc proteins in myopathy and cardiomyopathy. *Int J Mol Sci.* 2021;22(6):3058.
  94. Kweon J, Jang AH, Shin HR, See JE, Lee W, Lee JW, et al. A CRISPR-based base-editing screen for the functional assessment of BRCA1 variants. *Oncogene.* 2020;39(1):30–5.
  95. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature.* 2018;562(7726):217–22.
  96. Ibanez K, Polke J, Hagelstrom RT, Dolzhenko E, Pasko D, Thomas ERA, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol.* 2022;21(3):234–45.
  97. Sanford Kobayashi E, Waldman B, Engorn BM, Perofsky K, Allred E, Briggs B, et al. Cost efficacy of rapid whole genome sequencing in the pediatric intensive care unit. *Front Pediatr.* 2021;9:809536.
  98. Lowther C, Valkanas E, Giordano JL, Wang HZ, Currall BB, O'Keefe K et al. Systematic evaluation of genome sequencing for the diagnostic assessment of autism spectrum disorder and fetal structural anomalies. *Am J Hum Genet.* 2023;110(9):1454–69.
  99. Lindstrand A, Eisfeldt J, Pettersson M, Carvalho CMB, Kvarnung M, Grigoriou G, et al. From cytogenetics to cytogenomics: whole-genome sequencing as a first-line test comprehensively captures the diverse spectrum of disease-causing genetic variation underlying intellectual disability. *Genome Med.* 2019;11(1):68.
  100. Tuxen IV, Rohrberg KS, Oestrup O, Ahlborn LB, Schmidt AY, Spanggaard I, et al. Copenhagen prospective personalized oncology (CoPPO)-clinical utility of using molecular profiling to select patients to phase I trials. *Clin Cancer Res.* 2019;25(4):1239–47.
  101. Pleasance E, Bohm A, Williamson LM, Nelson JMT, Shen Y, Bonakdar M, et al. Whole-genome and transcriptome analysis enhances precision cancer treatment options. *Ann Oncol.* 2022;33(9):939–49.
  102. Ramarao-Milne KP, Patch AM, Nones K, Koufariotis R, Newell F, Addala VR, et al. Detection of actionable variants in various cancer types reveals value of whole-genome sequencing over in-silico whole-exome and hotspot panel sequencing. *Ann Oncol.* 2019;30:vii33.
  103. Bailey MH, Meyerson WU, Dursi LJ, Wang LB, Dong G, Liang WW, et al. Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples. *Nat Commun.* 2020;11(1):4748.
  104. Bertelsen B, Tuxen IV, Yde CW, Gabrielite M, Torp MH, Kinalis S, et al. High frequency of pathogenic germline variants within homologous recombination repair in patients with advanced cancer. *NPJ Genom Med.* 2019;4:13.
  105. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, et al. Pathogenic germline variants in 10,389 adult cancers. *Cell.* 2018;173(2):355–370 e314.
  106. Mandelker D, Zhang L, Kemel Y, Stadler ZK, Joseph V, Zehir A, et al. Mutation detection in patients with advanced Cancer by universal sequencing of Cancer-related genes in tumor and Normal DNA vs guideline-based germline testing. *JAMA.* 2017;318(9):825–35.
  107. McLean N, Delatycki MB, Macciocia I, Duncan RE. Ethical dilemmas associated with genetic testing: which are most commonly seen and how are they managed? *Genet Med.* 2013;15(5):345–53.
  108. Fulda KG, Lykens K. Ethical issues in predictive genetic testing: a public health perspective. *J Med Ethics.* 2006;32(3):143–7.
  109. Ascencio-Carbajal T, Saruwatari-Zavala G, Navarro-Garcia F, Frixione E. Genetic/genomic testing: defining the parameters for ethical, legal and social implications (ELSI). *BMC Med Ethics.* 2021;22(1):156.
  110. Johnson SB, Slade I, Giubilini A, Graham M. Rethinking the ethical principles of genomic medicine services. *Eur J Hum Genet.* 2020;28(2):147–54.
  111. Lantos JD. Ethical and psychosocial issues in whole genome sequencing (WGS) for newborns. *Pediatrics.* 2019;143(Suppl 1):S1–5.
  112. Bell SG. Ethical implications of rapid whole-genome sequencing in neonates. *Neonatal Netw.* 2018;37(1):42–4.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.