**Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies**

**Eleanor G. Seaby1,2,Sarah Ennis1**

1. Genomic Informatics Group, University Hospital Southampton, Tremona Road, Southampton, SO16 6YD, UK

2. Translational Genomics Group, Broad Institute of MIT and Harvard, 105 Broadway, Cambridge, MA 02142, USA

**Corresponding Author**

Dr Eleanor Seaby BMBS(hons) MMedSc(hons). Genomic Informatics Group, University Hospital Southampton, Tremona Road, Southampton, Hampshire, SO16 6YD, UK; Tel: +44(0)2381 206536; Email: [e.seaby@soton.ac.uk](mailto:e.seaby@soton.ac.uk).

**Authors**

* Dr Eleanor Seaby is a junior clinician and MD/PhD student at the University of Southampton. She additionally holds a position as a genomic analyst at the Broad Institute.
* Professor Sarah Ennis is Professor of Genomics at the University of Southampton and leads the Genomic Informatics Group.

**Abstract**

Next generation sequencing has revolutionised rare disease diagnostics. Concomitant with advancing technologies has been a rise in the number of new gene disorders discovered and diagnoses made for patients and their families. However, despite the trend towards whole exome and whole genome sequencing, diagnostic rates remain suboptimal. On average, only ~30% of patients receive a molecular diagnosis. National sequencing projects launched in the last five years are integrating clinical diagnostic testing with research avenues to widen the spectrum of known genetic disorders. Consequently, efforts to diagnose genetic disorders in a clinical setting are now often shared with efforts to prioritise candidate variants for the detection of new disease genes.

Herein we discuss some of the biggest obstacles precluding molecular diagnosis and discovery of new gene disorders. We consider bioinformatic and analytical challenges faced when interpreting next generation sequencing data and showcase some of the newest tools available to mitigate these issues. We consider how incomplete penetrance, non-coding variation and structural variants are likely to impact diagnostic rates and we further discuss methods for uplifting novel gene discovery by adopting a gene-to-patient based approach.

**Introduction**

In the modern genomics era, there has been an exponential rise in exome and genome sequencing for the 400 million patients worldwide with rare genetic diseases (RGD). These next-generation sequencing (NGS) technologies are thriving in the diagnosis of RGD where traditional methods e.g. cytogenetics, Sanger sequencing, and linkage analysis have been limited by low resolution testing, single test throughput, and the requirement to test multiple affected families, respectively. Whole exome sequencing (WES) and whole genome sequencing (WGS) can objectively test the genome for potential disease-causing variants without an *a priori* candidate gene, which has facilitated a rapid expansion of research into the discovery of new disease genes.

WES and WGS have now transitioned from a research environment into routine clinical practice as recognised diagnostic tests.[1] However, despite their promise, current diagnostic rates are suboptimal; on average only ~30%[2, 3] of patients offered WES or WGS receive a molecular diagnosis, although diagnostic rates vary by disease[3] with developmental delay phenotypes reporting higher solve rates at ~40%[4]. Regardless, most patients sequenced remain without a prognosis or treatment option. This is in part because NGS diagnostic testing is limited to known disease genes. Therefore, causal variants in novel genes or those outside of targeted (virtual) gene panels are ignored, at least until iterative analyses in light of new gene discoveries become commonplace. However, the genomics field is still an active research arena and national projects such as the UK’s 100,000 Genomes Project[5] and the USA’s All of Us Research Program[6] are utilising large-scale sequencing efforts to diagnose patients and discover new RGD. As a result, the boundary between clinical diagnostics and novel gene discovery is increasingly blurred, with many analysts adopting a dual diagnostic and research role. Patients consented only for clinical diagnostics will have their genome tested against known disease genes, whereas patients consented additionally for research may receive a diagnosis following functional studies on candidate variants undertaken in a research setting. This approach is steadily increasing the number of new gene disorders identified, however there is a still much to be discovered about the human genome. In fact, >75% of the phenotypical consequences for variation in all ~20,000 human genes is unaccounted for, thus rendering the exome far from obsolete[7]. Novel gene discovery is still hindered by vast quantities of variants of uncertain significance found in genes of undetermined biological function. A sequenced genome will output millions of variants; yet just a single nucleotide change may cause disease. Therefore, deciphering benign- from disease-causing variation is limiting disease-gene discoveries. Moreover, current analytical and annotation methods are imperfect[8, 9] and variation is not restricted to coding regions, which has necessitated research into structural variants and non-coding DNA. With the physiology of most genes unknown, and much still to be learnt about the non-coding genome, many genetic diseases will preclude detection until a greater proportion of the human genome has been biologically elucidated, or new methods are developed to uplift novel gene discovery.

This article will consider some of the biggest opportunities and challenges in the diagnosis and discovery of RGD with a focus on variant analysis from NGS data. Herein we consider patient selection, technological challenges, variant prioritisation, incomplete penetrance, transcript- and tissue-specific annotation, in addition to non-coding and structural variation. We further discuss current and future avenues to ameliorate these factors and propose a new approach for novel gene discovery to tackle the notorious diagnostic odyssey of rare disease.

**Application of NGS technologies**

WES and WGS are usually applied for suspected Mendelian disease of putative complete penetrance. With many patients being contemporaneously consented for both clinical diagnostic and research studies, the selection criteria are largely overlapping. These usually include: a severe phenotype that is unexplained by acquired or environmental factors; high suspicion for a monogenic disorder and; previous negative gene panel and/or microarray screening. Perhaps the main difference is where research studies are not limited to monogenic disease and may use NGS to elucidate the genetic architecture of complex diseases e.g. asthma, diabetes and inflammatory bowel disease. However, the success of WES/WGS diminishes rapidly for these diseases, therefore it is imperative to be clear on the uses and limitations of NGS prior to offering genetic testing.

**Pedigree selection**

With clear evidence of a Mendelian disorder, it is crucial to sequence the most appropriate individuals where applicable. Trio exomes (parent/offspring) are most useful for *de novo* inheritance[10], however if multiple children are affected then sequencing additional siblings increases diagnostic power where funds permit[11]. For dominant inheritance, sequencing the most distantly related affected individuals, e.g. cousins, increases power to discriminate pathogenic variants by reducing DNA shared between relatives by chance[12]. Where insufficient tissue is available for sequencing, e.g. prenatal death, parents can be sequenced and tested for rare, recessive variants in the absence of the proband; this is particularly useful in cases of recurrent fetal death[13].

**Clinical phenotyping**

With the most appropriate cases selected, it is vital that all clinical information (deemed relevant or not) is available for variant analysis. These data are usually collected and summarised before being assessed alongside biological evidence to support or refute the relevance of a given variant[14]. However, phenotyping only represents a snapshot in time; clinical data require updating longitudinally to accurately catalogue the phenotypic spectrum of disease, or to facilitate re-analysis when NGS proves negative.

When the predominant phenotype is restricted to a particular organ system it is tempting to assume that the genetic focus is within genes associated with that organ alone; often clinicians refer patients under the guise of an organ-specific phenotype and perhaps omit other clinical features that may prove critical to diagnosis. In 2017, we reported on a child referred for WES to resolve the aetiology of his renal disease with atypical features[15]. Initial analysis, restricted to nephrological genes, proved equivocal and only when the clinical phenotype was reviewed independently later was a wealth of additional dysmorphic features identified, explained by a pathogenic variant in *CBL*, causing a Noonan-like phenotype. This was aided by having direct access to the clinical notes, which many diagnostic laboratories do not have permission to view. Therefore, it is essential that all salient information is shared with the analysis team or that avenues are available to contact referring clinicians where applicable.

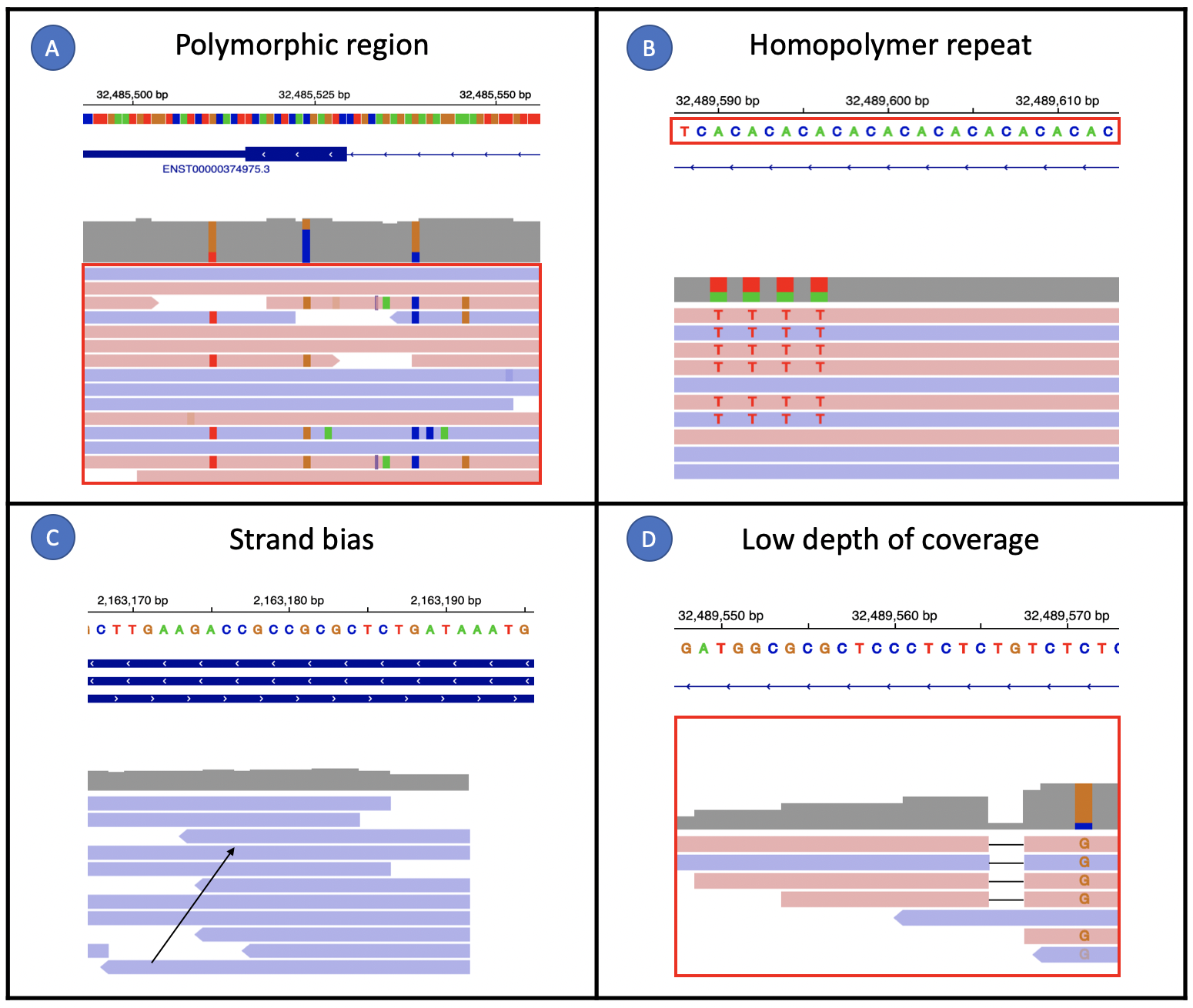
**Technological challenges**

NGS technologies are prone to technical errors. Detailed discussion on short-read sequencing (SRS), competing chemistries, sequencing platforms, and quality control pipelines have been extensively reviewed elsewhere[12, 16, 17].

The vast quantity of raw sequencing data produced necessitates sophisticated informatics pipelines for data handling, analysis and interpretation. The GATK best practices guideline[18] is now considered the gold-standard tool for variant-calling and has provided some analytical consensus[19], however with a large suite of competing tools available and varying accepted standards for genotype quality, read depth and mapping scores, many diagnostic centres vary in their approaches.

Understanding the computational steps of a variant-calling pipeline is important for downstream variant interpretation. The inclusion/exclusion of variants will depend on pre-programmed user-specific parameters selected in the pipeline. Interrogating variant quality is paramount to minimise type 1 errors and is aided by visualising read data and utilising genome browsers[20, 21]. Variant quality will depend on variant type, genomic region, depth of sequencing coverage, and analytical pipeline used. Particular attention should be paid to: indels and loss-of-function variants; mapping issues around repetitive and polymorphic regions; poor quality scores; strand bias; homopolymer tracts; high GC content; and insufficient or low coverage (**Figure 1**). The punctate nature of WES and its imperfect probe design mean that some regions escape capture altogether. These errors are minimised by the uniform coverage (but at the expense of lower depth) of WGS. However, WGS does not capture the full range of genomic variation, missing ~5% due to mapping errors[22].

**Figure 1 |** Common examples where variant calling may prove erroneous



*Forward reads are coloured red and point to the right and reverse reads are coloured lilac and point to the left. (A)* ***Polymorphic regions*** *are prone to common variation. Here, forward and reverse reads align over a short exon on a highly polymorphic HLA gene as visualised in the integrative genomics viewer. Multiple variants are scattered throughout the region, otherwise known as a ‘confetti effect’. The accuracy and performance of variant calling is challenged when reads are highly divergent from the reference sequence. (B)* ***Homopolymer repeats*** *are tracts of repeated nucleotide sequences; this may be single nucleotide repeats, double, triple etc. Here, the red box outlines a repeated sequence of CA. Careful attention should be paid to variants in repeat tracts longer than 7 repeats. Variant calling is prone to errors in homopolymer tracts, particularly when a variant is an expansion of a repeat i.e. if the reference sequence is CCCCCCC****T*** *and the variant called is CCCCCCC****C****. (C)* ***Strand bias*** *occurs when reads aligning to a reference are biased towards the forward or reverse strands. Here, all reads aligning to the region are on the reverse strand as shown by the arrow (they are all lilac and point towards the left). Strand bias is common around exon boundaries, particularly for whole exome sequencing. This challenges variant calling when it is not possible to accurately assess whether variants are evenly distributed on both the forward and reverse strands. False positive calls are made when a given variant is skewed towards either the forward or reverse strands. (D)* ***Low depth of coverage*** *is when the number of reads covering a region are few – many pipelines will filter variants with a read depth of <4. Here, the forward and reverse reads covering the small deletion have a read depth of only 5, (four with the indel and one without). Variant accuracy and allele balance cannot be accurately assessed without a greater depth of coverage.*

**Virtual gene panels and clinical diagnostics**

For well characterised RGD, a rapid diagnosis can be made if a variant is found in a gene already associated with that disease. This is the premise behind *in silico* (virtual) gene panels applied in clinical diagnostic testing. If an overlapping phenotype and formal assessment by the application of curation guidelines, most typically the American College of Medical Genetics and Genomics (ACMG), support a pathogenic variant then there is potential for diagnosis. Yet, despite a drive towards standardised guidelines, many diagnostic labs vary in their variant interpretation[23, 24]. Further, gene panels are problematic as they are subjectively curated and outdated at the point that new discoveries are published. This necessitates iterative re-analysis, which can boost diagnostic rates by 10-15%[25, 26]; however this is seldom done due to constraints on workload and time[27]. Nevertheless, since ~50% of haploinsufficient disease-genes remain to be discovered[7], gene panels are rarely successful for the majority of exomes and genomes analysed. Research is therefore required to resolve pathogenic variation in the remaining disease exome, which has been facilitated by patients recruited into research studies. Challenges and considerations specifically related to prioritising variants for novel gene discovery are reviewed below.

**Data volume**

An average genome sequenced has ~3-4 million variants and sifting through these manually is impossible. WES produces significantly fewer variants at 20,000-22,000[28], however this number is still too large for individual variant-level scrutiny and it misses non-coding variation. Bioinformatic methods can restrict data to a more manageable number, however this still leaves an overwhelmingly large number (tens to thousands) of variants to consider in more detail. This range is highly dependent on whether analysis is conducted on an exome or genome and whether trio sequencing is available.

**Variant analysis and prioritisation**

Most variant analysts have a filtering framework to whittle down vast quantities of genomic data to a more manageable list of candidates. Although general filtering strategies are recommended, there is no universal framework governing this process. Invariably, analysis is tailored to individual cases by considering an aggregate of information.

***De novo* analysis**

Many RGD occur as one-off (*de novo*) events i.e. present in the proband but absent from unaffected parents[29]. The rate of *de novo* mutation is between 44-82 events per genome with 1-2 affecting coding regions, and this number increases with paternal age[28, 30]. Restricting analysis to *de novo* variants can reduce a genome significantly[28]. This is achieved using trio data (parents/offspring), which is not always available due to cost restrictions or inability to sequence related individuals due to parental death, loss of contact, unknown paternity and parental refusal. Consequently, some exomes and genomes have to be analysed as singletons (proband-only) at the expense of analytical noise. Duo studies (one parent/proband) are preferable to proband-only, however many novel variants in the proband will be inherited by the other (unsequenced) parent. Overwhelmingly, singleton analysis is too noisy to identify candidates without arbitrary and restrictive filtering. Further, in rare cases, parental germline mosaicism can result in multiple affected offspring with assumed *de novo* inheritance[31].

**Recessive analysis**

Multiple affected siblings to unaffected parents increases the probability for recessive inheritance. Where consanguinity is declared or implied through relatedness checks, regions of autozygosity (where regions of DNA are identical by descent) can be targeted as hotspots for rare homozygous variation. Biallelic inheritance of compound heterozygotes *in trans* are best identified with trio analysis, unless the variants are close enough together in the proband that visualising overlapping read data can determine phase. However, assigning pathogenicity becomes more challenging when one variant is a known pathogenic variant and a second is of low quality or has low *in silico* predictors. That said, there are cases of known pathogenic variants causing disease when co-inherited with a second more common variant (or hypomorph)[32]. The tradeoff here is ensuring that allele frequencies are not filtered too restrictively especially when hypomorphs can have allele frequencies as high as 7%[33].

**Allele frequency**

Since population allele frequency (AF) data have been made publicly available, detecting rare variation has become far easier[5, 34]. The largest repository of WES/WGS data for ostensibly healthy adults, depleted for rare childhood disease, is the genome aggregation database, or gnomAD ([http://gnomad.broadinstitute.org](http://gnomad.broadinstitute.org/)), hosting 125,748 exomes and 71,702 genomes. With AF differing between ethnicities, it is important that rare variation is compared with the AF of the patient’s ethnic group; i.e. what may be rare in Caucasians could be common in Latinos. Historically, global repositories have been biased towards European ancestry, and whilst this bias still exists, many more ancestries are now included with updated databases.

Undoubtedly global repositories are immensely powerful, however most individuals harbour a number of private variants absent from these repositories. Even at 197,450 samples, gnomAD is underpowered to accurately catalogue rare variation on a population level[35]. That said, AF still remains a fundamental filtering tool in rare disease, but frequency cutoffs are often arbitrary and overly lenient. Whiffin *et al*.[36] developed a statistical tool (<http://cardiodb.org/allelefrequencyapp/>) for the frequency-based filtering of candidate variants accounting for disease prevalence, inheritance mode, allelic heterogeneity, penetrance and sampling variance. The tool was shown to reduce candidate variants by two-thirds whilst maintaining sensitivity.

**Variant type**

Variant type is an important consideration in variant analysis (**Table 1**). Coding variants are usually considered first, although splicing variants and indels may impact both coding and non-coding regions. Loss-of-function variants (frameshift, stop-gain and essential splice-site) are regarded as the most pathogenic coding variants and are rarer compared to all other variant types. However, they are enriched for technical artefacts and many of these error types are not accounted for by current curation guidelines[35, 37]. Missense variants are scattered across disease genes in healthy populations, challenging their interpretation. Regional missense constraint metrics are helping to prioritise candidates[38, 39]. Splicing variants (exonic and intronic) can augment regulatory domains within mRNA, particularly those controlling splicing and may also influence translation[40, 41]. However, these are difficult to assess beyond the canonical splice site and many *in silico* tools perform poorly[42, 43]. That said, recent splicing libraries are widening the biological understanding of splicing variants and these can be used to train splicing models[44, 45]. Synonymous variants are often discarded but are not benign by default and show a degree of selective pressure. These variants can alter the consensus sequence around splice sites and may also augment translation[40]. Short indels spanning coding regions may result in loss-of-function, sequence truncation/elongation, or splicing events. Yet these are often poorly resolved and aligned[46]. Non-coding variation is considered later in this review.

**Table 1 |** Variant classes including occurrence, consequence and considerations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variant class | description | Occurrence | Consequence | Considerations |
| Synonymous | Single nucleotide variant | ~11,800 per genome | Do not alter the amino acid sequence | * Not always benign * Can affect transcription and splicing |
| Missense | Single nucleotide variant | ~10,600 per genome | Single amino acid substitution | * Scattered across disease genes in healthy populations * Metrics available that map missense constraint at a regional gene level and highlight critical regions that may be implicated in disease |
| Splicing | Single nucleotide variant (essential or non-essential splice site) or indel | ~2,300 per genome | Alter the sequence around splice junctions and may result in aberrant mRNA transcription | * Variants may reduce or strengthen the natural splice site, disrupt regulatory sequences (e.g. silencers and enhancers), introduce new splice sites, or activate cryptic splice sites * Difficult to assess beyond the canonical splice site due to limitations in annotation software, particularly for indels and deep intronic variants |
| loss of function | Frameshift, nonsense or essential splice site | ~100 per genome | May result in protein truncation or protein loss due to nonsense medicated decay | * Predicted most pathogenic compared to all other variant types * Enriched for annotation, technical and transcript artefacts * Single exon genes escape nonsense medicated decay |
| Short indels | Small insertions or deletions <50 base-pairs | ~420,000 per genome | - In-frame indels may elongate or truncate the protein product  - Out of frame indels will result in loss-of-function  - Indels across splice sites may disrupt splicing | * May have regulatory roles on DNA structure and function * Often poorly resolved and aligned particularly from short read exome data * WGS improves indel calling * Long read sequencing outperforms short read sequencing |

*Rates of variant occurrence were collated from 1000 genomes data, ExAC data, and a cohort of 44 Caucasian genomes[47-49]. Note that rates of variation are likely to be an underestimate as cohort sizes are still underpowered to detect all rare variation[35]. Additionally, indel rates are likely grossly underestimated due to alignment and mapping issues[50].*

***In silico* tools**

*In silico* tools score variants according to functional consequence (**Table 2**). A wealth of competing software is available, yet many informatics pipelines are incompatible with these tools, nor do they keep up with updated versions.Discordant evidence lowers the confidence of relying on any one score. Composite scores, which utilise statistical methods, machine-learning, and deep neural networks help to resolve this issue[51, 52]. Yet, no single method emerges superior and therefore a consensus approach is still recommended[53].

**Table** **2 |** Selection of commonly applied in silico prediction tools available for variant prioritisation.

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Algorithm | Source | Principle |
| Non-synonymous SNV prediction | SIFT[54] | [http://sift.jcvi.org](http://sift.jcvi.org/) | Evolutionary conservation |
| PolyPhen-2[55] | <http://genetics.bwh.harvard.edu/pph2> | Evolutionary conservation and protein structure/function |
| MutationTaster[56] | [http://www.mutationtaster.org](http://www.mutationtaster.org/) | Evolutionary conservation and protein structure/function |
| Grantham[57] | <https://gist.github.com/danielecook/> | Biological consequence of amino-acid change |
| PROVEAN[58] | <http://provean.jcvi.org> | Delta score model |
| ClinPred[59] | <https://bio.tools/ClinPred> | Machine learning |
| FunSAV[60] | <http://sunflower.kuicr.kyoto-u.ac.jp/sjn/FunSAV> | Random forest model |
| Synonymous snv prediction | FATHMM-MKL[61] | <http://fathmm.biocompute.org.uk/fathmmMKL.htm> | Sequence conservation within hidden Markov models |
| GWAVA[62] | <https://www.sanger.ac.uk/sanger/StatGen_Gwava> | Integration of various genomic and epigenomic annotations |
| TraP[63] | <http://trap-score.org> | Integrates sequence motif changes and GERP++ |
| INDEL PREDICTION | IndelMINER[64] | <http://omicstools.com/indelminer-tool> | Heuristic model based on split read data |
| ABRA[65] | <https://github.com/mozack/abra> | *De novo* assembly |
| VarScan[66] | <http://varscan.sourceforge.net/> | Integration of read coverage, base quality and number of strands |
| DINDEL[67] | <https://omictools.com/dindel-tool> | Bayesian method |
| GATK Haplotype Caller[18] | <http://software.broadinstitute.org/gatk/> | Bayesian method |
| DDIG-in[68] | <http://sparks-lab.org/ddig> | Machine learning method |
| DeepVariant[69] | <https://github.com/google/deepvariant/> | Deep neural network |
| Splicing prediction | GeneSplicer[70] | <https://ccb.jhu.edu/software/genesplicer/> | Markov model |
| MaxEntScan[71] | <http://genes.mit.edu/burgelab/maxent/> | Maximum entropy model |
| Human Splicing Finder[72] | <http://www.umd.be/HSF/> | Position dependent logic |
| MutPred Splice[73] | <http://mutdb.org/mutpredsplice/about.htm> | Machine-learning prediction of exonic variants |
| SpliceAi[74] | <https://pypi.org/project/spliceai/> | Deep neural network |
| Conservation prediction | PhyloP[75] | <http://compgen.bscb.cornell.edu/phast/> | Statistical phylogenetic tests |
| GERP++[76] | <http://mendel.stanford.edu/SidowLab/downloads/gerp/> | Conservation across species |
| PhastCons[77] | <http://compgen.cshl.edu/phast/> | Phylogenetic hidden Markov model |
| Panther-PSEP[78] | <http://pantherdb.org/tools/csnpScoreForm.jsp> | Position-specific evolutionary conservation |
| Non-coding prediction | DeepSEA[79] | <http://deepsea.princeton.edu/> | Deep learning algorithm |
| GenoCanyon[80] | [http://genocanyon.med.yale.edu](http://genocanyon.med.yale.edu/) | Unsupervised learning |
| SInBaD[81] | <http://tingchenlab.cmb.usc.edu/sinbad/> | Sequence information based decision model |
| protein prediction | AGGRESCAN3D[82] | <http://biocomp.chem.uw.edu.pl/A3D/> | Protein aggregation method |
| DUET[83] | <http://structure.bioc.cam.ac.uk/duet> | Integrated computational approach |
| HMMvar-func[84] | <http://bioinformatics.cs.vt.edu/zhanglab/HMMvar/download.php> | Hidden Markov models |
| LS-SNP/PDB[85] | <http://ls-snp.icm.jhu.edu/ls-snp-pdb> | Genome-wide mapping of SNVs onto protein structures |
| NeEMO[86] | <http://protein.bio.unipd.it/neemo/> | Residue interaction networks |
| PMut[87] | <http://mmb.irbbarcelona.org/PMut> | Neural network |
| Composite score | CADD[88] | <http://cadd.gs.washington.edu/> | Multiple genomic annotations |
| M-CAP[51] | <http://bejerano.stanford.edu/mcap/> | Machine-learning composite score |
| Eigen[89] | <http://www.columbia.edu/~ii2135/eigen.html> | Unsupervised spectral model |
| Dann[52] | <http://cbcl.ics.uci.edu/public_data/DANN> | Deep neural network |
| REVEL[90] | <https://sites.google.com/site/revelgenomics/> | Ensemble learning method |
| Alamut | <https://interactive-biosoftware/alamut-visual> | Composite score using splicing, missense and ESE prediction tools |

*Alamut is a commercial product and does not have a corresponding research paper. Updated versions are available from the source links. [ESE - exonic splicing enhancer; SNV – single nucleotide variant].*

Although *in silico* tools can help to prioritise variants, that is not to say they predict disease-causality. Most metrics cannot discriminate by gene; however new methods are shifting data interpretation from the variant to the gene-level. GenePy is a gene-based pathogenicity score that can be applied to both cohorts and individuals. This metric correctly selected pathogenic genes for inflammatory bowel disease[91]. However, no current method can link a specific phenotype to a gene based on biological function - all gene-based metrics reflect the burden of mutation within genes of equally-predicted causality.

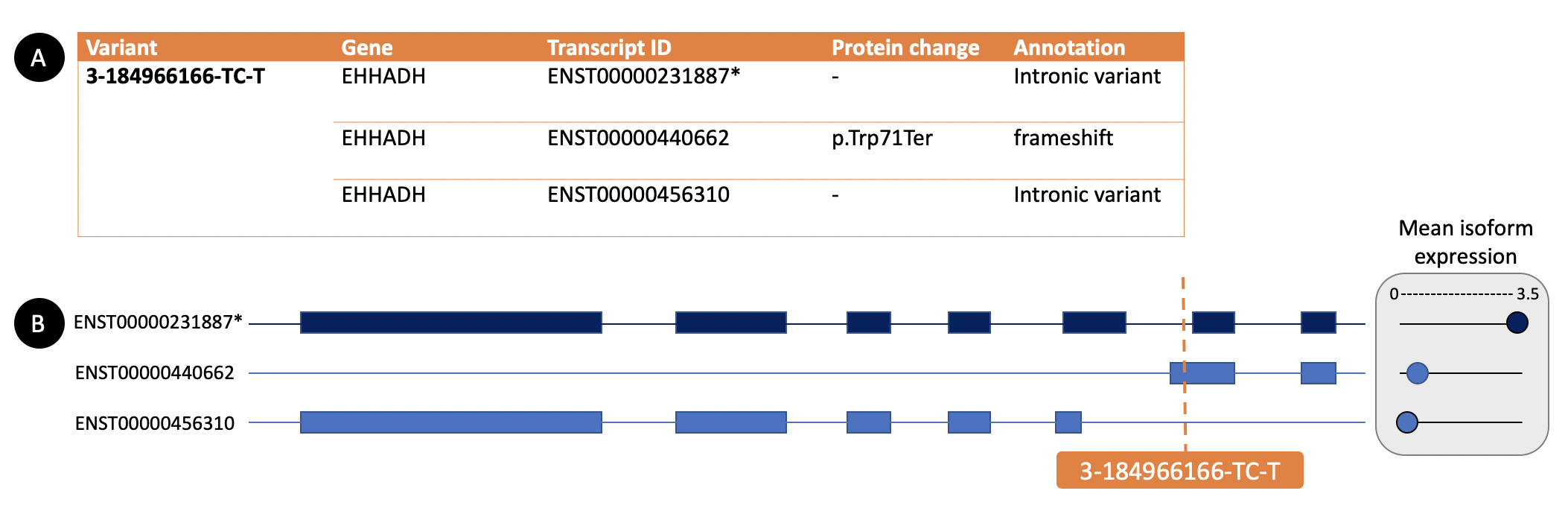
**Tissue-specific annotation**

Tissue-specific annotation has been a neglected area of variant interpretation. The genotype tissue expression (GTEx) portal has facilitated the study of tissue-specific gene expression and regulation important for variant analysis, e.g. a predicted pathogenic variant only expressed in testes in the context of a strong neurodevelopmental phenotype would likely be deprioritised. Equally, a variant strongly expressed in the kidney for a nephrological phenotype would warrant further scrutiny, even if overall tissue expression was low. That said, prudence is advised here; GTEx data were collected from cadavers, which will not necessarily represent *in utero* isoform gene expression when disease pathogenesis may have occurred[92]. Furthermore, some individuals present mosaicism and variants will be missed if the tissue of origin is not sequenced.

**Transcript-specific annotation**

Transcript-specific annotation considers how a variant affects all annotated transcripts of that given gene. Certain disorders are transcript-specific, i.e. only variants disrupting particular transcripts cause disease. Protein truncating variants in titan (*TTN)* - one of the largest genes encoding an essential protein component of striated muscle - cause dilated cardiomyopathy[93], yet loss-of-function variants are found in healthy individuals. In healthy controls, loss-of-function variants occur in exons absent from dominantly expressed isoforms (the resultant spliced RNA as a result of a given transcript)[94]. This is where nomenclature and understanding variant notation is essential. Diagnostic reports often describe variants with coding (*c.)* and protein (*p.)* notation in keeping with HGVS guidance[95], yet often neglect to include which transcript the variant relates to. To confuse matters further, there are different annotation databases (RefSeq, Gencode, Ensembl) and transcripts vary between them. Nevertheless, reporting variants using genome (*g.)* nomenclature e.g. 1-55516888-G-A is preferable as all transcript consequences (from any preferred annotation database) can be assessed, e.g. a variant may be loss-of-function on one transcript yet intronic on all others (**Figure 2**). Furthermore, the same loss-of-function variant may disrupt a poorly expressed transcript in only one (clinically irrelevant) tissue. If this loss-of-function variant was reported without the transcript information e.g. p.Trp71Ter, then an analyst could be forgiven for assuming the variant may be pathogenic if found in a known disease gene. On the other hand, this same variant may be uniquely and moderately expressed in a tissue of interest, yet have low overall tissue expression, risking its exclusion. This is where amalgamating both tissue- and transcript-specific information is most valuable.

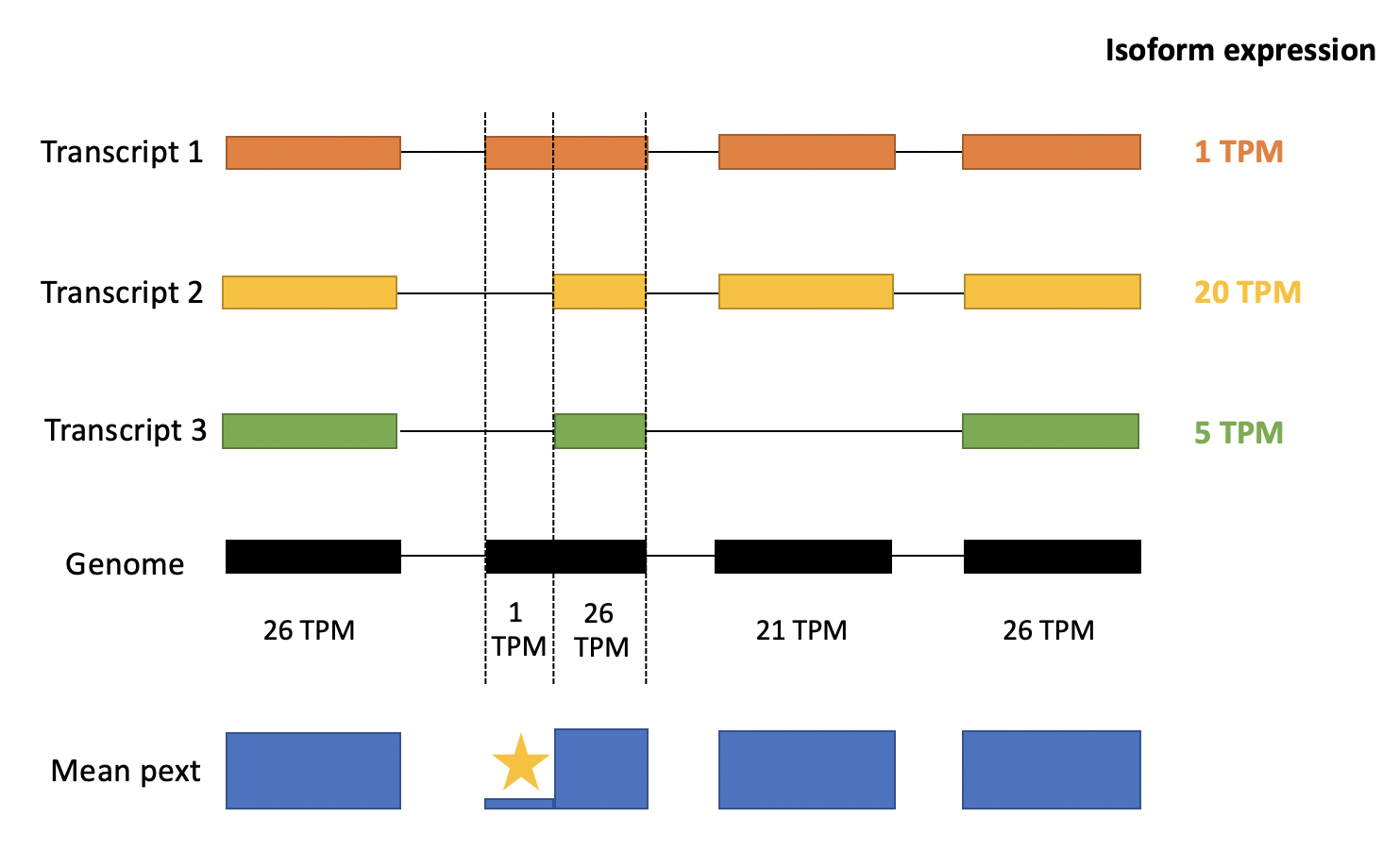
**Figure 2 |** Schematic of the 3-184966166-TC-T variant and its transcript-specific annotations.



***(A)*** *shows how the 3-184966166-TC-T variant causes a frameshift event on one transcript (ENST00000440662) and is intronic on the two remaining transcripts, including the canonical transcript identified by the asterisk (ENST00000231887\*). If the variant were just annotated as EHHADH: p.Trp71Ter, this would neglect important transcript-specific information that would otherwise lower the pathogenicity of the variant.* ***(B)*** *Shows the individual transcripts and their corresponding exons (blue boxes). The canonical transcript is shaded in dark blue. Mean isoform expression of each transcript across GTEx tissues is shown on the right. The dotted orange line shows the position of the variant across all annotated transcripts. The transcript harbouring the p.Trp71Ter variant only disrupts the coding sequence of the ENST00000440662 isoform and has low mean expression across tissues, suggesting this transcript is not biologically important. Contrastingly, the canonical transcript (ENST00000231887\*) is highly expressed approaching 3.5 transcripts per million, yet the variant is intronic for this transcript. Therefore, by using transcript level annotation, variant pathogenicity can be more accurately assessed and avoid providing a false sense of pathogenicity, particularly if the gene has a known disease association.*

To aid in the interpretation of transcript-specific annotation, Cummings *et al*.[37] developed a tool that facilitates the rapid visualisation of isoform expression values across tissues, called the proportion expressed across transcript (pext) score. Pext is capable of differentiating between weakly and highly evolutionarily conserved exons and thus serves as a surrogate for functional importance (**Figure 3**). The integration of pext into the gnomAD browser now provides an opportunity to rapidly exclude variants from weakly expressed exons.

**Figure** **3 |** Schematic of how pext enables visualisation of isoform expression across exonic regions.



*Pext values are integrated from the GTEx dataset by computing the median expression of a transcript for GTEx tissue samples. The expression of a given base is defined as the summed expression of all transcripts that touch that base. This is repeated for every GTEx tissue and then normalised by the whole gene expression in that tissue. Transcript 1 has low tissue expression compared with transcripts 2 and 3. The overhanging region of exon 2 (on transcript 1) has a base pext of 1 TPM and therefore this region is likely to have low conservation or be enriched for annotation errors. Therefore, a variant in the region of the yellow star would be easily identified as a low priority candidate due to the relative drop in mean pext secondary to low tissue expression in GTEx across that region. The caveat to this rule is where transcript 1 is uniquely expressed in the tissue of interest, and therefore expression values should be assessed against tissues of interest, rather than just the mean expression across all tissues. The integration of pext into the gnomAD browser allows the analyst to visualise both mean pext and tissue specific-pext. [GTEx – genotype tissue expression; TPM - transcript per million]. Figure is adapted from Cummings et al.[37].*

**Incomplete penetrance**

Harbouring a disease-causing variant does not always result in expression of disease, a phenomenon known as incomplete penetrance, reviewed elsewhere[96]. To illustrate this, gnomAD (which is depleted for individuals with rare childhood diseases) contains known pathogenic variants in haploinsufficient genes (*RB1*, *APC)*. Indeed, incomplete penetrance is not uncommon; several examples exist whereby pathogenic variants do not always present clinically in the individuals who carry them (**Table 3**)[97].

**Table 3 |** Examples of incompletely penetrant genes.

|  |  |  |
| --- | --- | --- |
| Gene(s) | Phenotype | Paper |
| *CX46* | Congenital cataract | Shawky *et al*.[98] |
| *ASXL1* | Bohring-Opitz syndrome | Ropers *et al.*[99] |
| *NF1* | Neurofibromatosis type 1 | Fahsold *et al*.[100] |
| *CFH, CFI, CFB, C3, MCB* | Atypical haemolytic syndrome | Bresin *et al*.[101] |
| *COL1A1, COL1A2* | Osteogenesis imperfecta | Veitia *et al*.[102] |
| *LRRK2* | Parkinsons disease | Alessi *et al*.[103] |
| *SCN5A* | Brugada syndrome | Gourraud *et al*.[104] |
| *DCC* | Agenesis of the corpus callosum | Marsh *et al*.[105] |

*Examples of genes displaying incomplete or reduced penetrance supported by scientific literature across a diverse range of phenotypes.*

Population sequencing projects have revealed a high burden of potential protein-damaging variants in apparently healthy individuals[35, 47, 96]. Data extracted from the UK biobank, are beginning to evaluate rare disease-causing variants and refine penetrance estimates[106]. This challenges the analysis of NGS data when some genetic diseases may segregate through unaffected parents at allele frequencies higher than expected for a fully penetrant disease variant. On the other hand, some variants may be fully penetrant, yet modifier genes or rescue events may vary phenotype expressivity.

**Non-coding variation**

Until recently, non-coding DNA (comprising 98% of the human genome) was considered ‘junk’. With the falling costs of WGS and advancing technologies (**Table 4**), non-coding variation is proving itself to be important for gene regulation, epigenetics and 3D genome structure[107].

**Table 4 |** Description of technologies for the detection of gene regulation, epigenetics and genome structure.

|  |  |
| --- | --- |
| Type | Description |
| *technologies to assess non-coding variation* | |
| ChIP‐seq | ChIP‐sequencing (ChIP-seq) is a method to analyse protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with DNA sequencing to identify DNA regions bound by proteins of interest. Histone marks indicating regulatory properties include H3K4me3, H3K27me3, and H3K27ac. Transcription factor binding sites can be analysed using ChIP‐seq in a genome‐wide manner. |
| DNase‐seq, ATAC‐seq and FAIRE‐seq | These technologies identify ‘open chromatin’ regions where DNA binding proteins can access target DNA sequences. |
| Hi‐C and ChIA‐PET | These technologies are chromosome conformation capture‐based methods for genome‐wide analysis of 3D chromatin interactions. Hi‐C is capable of detecting all possible interactions between DNA sequences in the genome, such as topological association domains. Chromatin interaction analysis by paired-end tag sequencing (ChIA‐PET) incorporates ChIP-based enrichment and high-throughput sequencing to determine *de novo* long range chromatin interactions across the genome. |
| MeDIP‐seq and whole‐genome bisulfite‐seq | These technologies provide genome‐wide analysis of methylated DNA sequences. MeDIP‐seq combines ChIP and high-throughput sequencing. Whole‐genome bisulfite‐seq uses both treated (with sodium bisulfite) and untreated genomic DNA to identify methylated regions. |

*ATAC-seq – assay for transposase-accessible chromatin using sequencing assay; ChIA-PET – chromatin interaction analysis by paired-end tag sequencing; ChIP-seq – chromatin immunoprecipitation sequencing; DNASE-seq – Dnase I hypersensitive sites sequencing; FAIRE-seq – formaldehyde-assisted isolation of regulatory elements sequencing; MEDIP-seq – methylated DNA immunoprecipitation sequencing.*

Efforts to understand non-coding variation have come through the ENCODE project, which has provided experimental evidence on the function of non-coding regulatory elements and gained new insights into the nature of transcription, chromatin structure and histone modification[108]. Results suggest that 80% of the genome contains elements linked to biochemical function, and that intergenic regions contain *cis*-regulatory elements such as enhancers, promoters, silencers and untranslated RNA transcripts with regulatory function (**Table 5**)[109]. *In silico* methods that exploit ENCODE data and use machine-learning to predict the functional impact of non-coding single nucleotide variants are emerging[61, 62, 89, 110]. However, the routine integration of these metrics is lacking, not least because the same variant may have conflicting effects in different tissues, individuals, and developmental stages and most analytical workflows are ill-equipped to interpret their effects[62].

**Table 5 |** Description of non-coding variant classes.

|  |  |
| --- | --- |
| Type | Description |
| *Representative functional non‐coding elements* | |
| Untranslated region | Untranslated regions (UTRs) are transcribed DNA included in the mature mRNA but not translated into protein. The 5’ (upstream) UTR and 3’ (downstream) UTR are important for translational control, mRNA stability and intracellular localisation. The 5’ UTR contains the 5’ cap and the 3’ UTR contains the polyadenylation signal (AAUAAA) and miRNA-binding sites. |
| Promoter | Promoters provide binding for the promoter machinery that initiates gene transcription, typically located upstream of a gene. Epigenetic markers for active and repressive promoters include H3K4me3 and H3K27me3, respectively. |
| Enhancer | Enhancers provide binding sites for proteins that help active transcription and control of expression levels of distal gene(s). H3K27ac is known as the representative mark of active enhancers. |
| Silencer | Silencers provide binding sites for proteins important in the repression of transcription. |
| Transcription factor binding sites | TF binding sites are genomic elements bound by TF proteins that play a crucial role in regulation of gene transcription. TFs bind preferably to specific DNA sequence motifs in *cis*-regulatory regions of promoter and enhancer DNA. |
| Insulators | Insulators are DNA elements that act as a blocker of interactions between regulatory elements (e.g., enhancer-blocker insultators) and/or a barrier to propagation of repressive chromatin (barrier insulators). |
| Transcribed non‐coding regions | Some non-coding DNA is transcribed into functional RNA but not translated into protein. Specific examples include transfer RNA, ribosomal RNA, miRNA, long (intergenic) non‐coding RNA, small nucleolar RNA, and others. |

*[TF – transcription factor; miRNA – microRNA; UTR – untranslated region].*

**Structural variation**

Structural variants (SVs) are genomic rearrangements larger than 50 base-pairs. SVs not only affect gene dosage, but can modulate the basic mechanisms of gene regulation by altering the copy number of regulatory elements and by modifying the 3D genome through topological associating domains. SV disease mechanisms are reviewed elsewhere[107].

SVs can be balanced (inversions and translocations), or unbalanced (deletions, insertions and duplications) and we are only beginning to appreciate their prevalence. A preprint by Collins *et al.*[111] catalogued 433,371 distinct SVs across >14,000 genomes in gnomAD, producing 7,439 SVs on average per individual, contributing to approximately 25-29% of all rare protein truncating events. Therefore, there is likely to be an appreciable burden of rare pathogenic variation secondary to SVs undetected by current clinical testing standards. This has often been suspected, particularly in cases when one pathogenic variant is detected in a recessive gene of interest and there’s suspicion for a ‘second hit’ somewhere else in the genome. Indeed, a recent preprint showed that 4% of undiagnosed genomes were solved using SV calling[112].

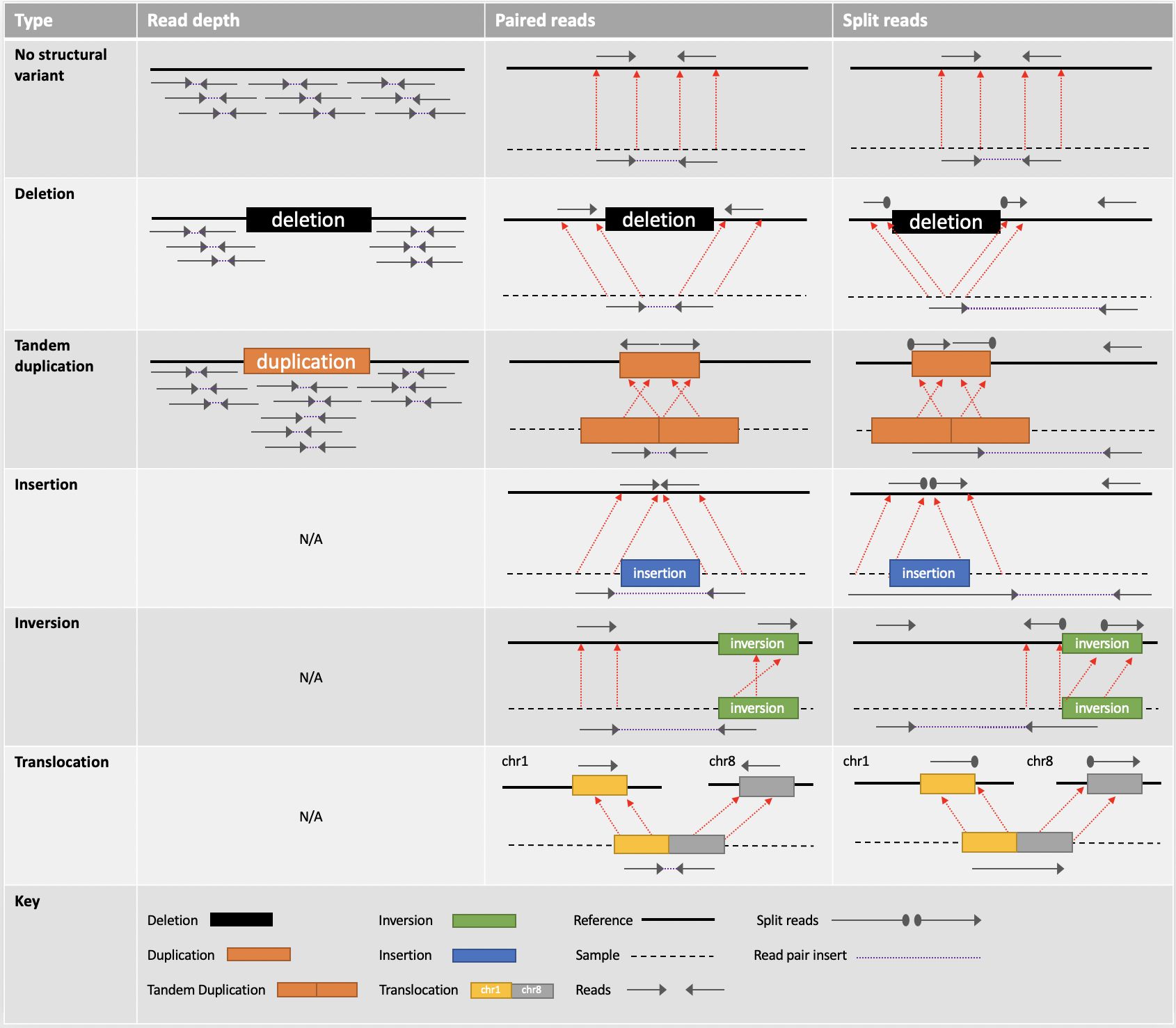
The most well studied SVs are unbalanced SVs, or copy number variants (CNVs). Array comparative genomic hybridisation is the first-line diagnostic CNV test, however its low resolution and inability to detect balanced rearrangements mean that a large proportion of SVs are missed[107]. As technologies have advanced, the detection of SVs has moved away from cytogenetics towards NGS (**Table 6**). Methods to detect SVs from short-read WGS data has been made possible by an abundance of tools that exploit paired-end reads, read depth, and split reads (**Figure 4**). Theoretically, all SV classes can be detected at a finer scale than offered by traditional methods. However, these methods are limited, particularly from short-read data where mapping is notoriously inaccurate round repetitive regions, and most analytical pipelines do not support SV calling when their throughput is predominantly WES. Long-read sequencing (LRS) is emerging as a solution and has already revealed many times more SVs per genome when compared with SRS; however, cost is still a substantial issue[113]. Linked-read technology, provided by 10X Genomics, integrates SRS with long range information. Long molecules of DNA are fragmented and labelled with a barcode; these fragments undergo SRS, producing a set of labelled reads that originate from the same DNA molecule. This technology is cheaper compared with LRS and facilitates simultaneous detection of small and large variants from a single library, however this technology is prone to the same limitations as both SRS and LRS.[114]

**Table** **6 |** Methods for the detection of structural variants and their corresponding advantages and disadvantages.

|  |  |  |
| --- | --- | --- |
| Methods | Advantages | Disadvantages |
| G banded karyotype | * Unbiased * Good for aneuploidy | * Limited resolution (5-7 megabases) |
| FISH | * Uses fixed interphase cells * Assesses balanced translocations * Detects mosaicism and tumour heterogeneity | * Requires multiple probes/assays to test multiple loci * Low resolution (100-200 kilobases) |
| Array CGH | * Good for CNV and loss of heterozygosity * Better resolution than FISH * Low cost per data point | * Resolution still limited (50-100 kilobases) * Cannot detect balanced translocations * Cannot detect copy neutral CNVs * Low throughput |
| MLPA | * Detects small rearrangements up to 50bp * High throughput * Uses up to 40 probes | * Cannot detect copy neutral CNVs * Poor for mosaicism and tumour heterogeneity |
| Short read sequencing (SRS) | * Can detect full range of genetic variation * High throughput * Resolution from 50bp * Greater per base accuracy than LRS * Deeper coverage | * Dependent on coverage * Susceptible to GC bias * Prone to errors particularly around low complexity regions * Requires WGS to assess all SV classes |
| Long read sequencing (LRS) | * Can detect full range of genetic variation * High throughput * Detects more SVs than SRS * More uniform coverage * Greater overall accuracy than SRS | * Expensive * Lower per base accuracy than SRS * Lower depth of coverage * Cannot resolve sequences larger than input DNA |
| Linked-read sequencing | * Can detect full range of genetic variation * Constructs long range haplotypes * Low error rate * Lower cost compared with LRS | * Cannot resolve sequences larger than input DNA * Susceptible to GC bias * Reduced performance around small indel calling |

*[Bp – base pair; CGH – comparative genomic hybridisation; CNV – copy number variant; FISH – fluorescence in situ hybridisation; kbp – kilo-base pair; LRS – long read sequencing; MLPA – multiplex ligation-dependent probe amplification; SRS – short read sequencing; SV – structural variant; WGS – whole genome sequencing].*

**Figure 4** | Methods to detect structural variation using next generation sequencing data.

****

***Read depth*** *is exploited to assess deletions and duplications. Reads aligned to the reference genome will show increased coverage over duplicated regions and decreased coverage in the presence of a deletion.* ***Paired reads*** *are sequenced from either end, leaving an insertion gap in the middle. In a deletion the read pairs align to the reference further apart than expected. Dotted red lines denote change in read position/orientation between sample and reference. In a tandem duplication the orientation of the read pair is reversed in the reference. In an insertion the read pairs are aligned closer together than expected i.e. their insert size is less than expected. For inversions, the paired reads align in the same direction (either in the forward or reverse direction). For a translocation, the reads pairs are mapped to different chromosomes.* ***Split reads*** *occur over breakpoints. For a deletion, reads across the breakpoint will be split in the reference. For a tandem duplication, reads spanning the concatenation of the duplication will be split where the insertion begins and ends. For an insertion, reads will split at the point where the insertion begins or ends. For an inversion, the right read pair is split at the point where the inversion begins. The remaining split read will change orientation and align at the reverse end of the inversion. For a translocation, reads spanning the join between translocated regions will be split and align with their independent chromosomes e.g. chromosome 1 (chr1) and chromosome 8 (chr8).*

**Applying supporting evidence from literature**

When a list of candidates remains following variant prioritisation, scrupulous assessment of the scientific literature is mandatory to evaluate biological significance. This is achieved by reading papers on a gene(s) of interest and accessing databases such as OMIM[115] (http://omim.org/), ClinVar[116] (https://www.ncbi.nlm.nih.gov/clinvar/), and Decipher[117] (<https://decipher.sanger.ac.uk/>), which catalogue genotype/phenotype relationships with degrees of supporting evidence. This part of analysis is perhaps the most time-consuming as it involves evaluating copious amounts of evidence. Often the only genotype/phenotype association is through a genome wide association study; yet rare pathogenic variants may not present the same phenotype.

There is also a risk that published papers make false associations between a gene and disease. This was apparent in historical papers that linked ‘rare’ (at the time) segregating variants as pathogenic for a particular phenotype based on public databases primarily representing Europeans. Some rare variants in European populations may be common in other ethnicities. Only when public repositories started to include a diverse range of ethnicities did these ‘rare variants’ prove to be common and not disease causing at all. For example, the *MYBPC3* (p.G278E) variant was classified as pathogenic for hypertrophic cardiomyopathy in African populations. The variant was later reclassified as benign when found to be common in Africans[118]. Additionally, as public repositories have increased in size, many presumed disease-causing variants have been deemed ‘too common’ to be pathogenic[119]. Removing these variants from the literature is impossible, but the data-sharing ethos of ClinVar ameliorates this issue by encouraging evidence-based submissions which may refute previous claims of pathogenicity. By leveraging knowledge from multiple submissions, clinical significance and assertion criteria, ClinVar can be exploited as an important tool for the accurate interpretation of variant pathogenicity[120].

**Genes of unknown function and functional validation**

Although phenotype-genotype associations are continually expanding, most variants remain variants of unknown significance (VUS). Perhaps the biggest challenge in reaching a diagnosis is the paucity of knowledge into the biological function of all 20,000 genes. It is estimated that 70-75% of human genes still have no known function in human health and 50% of disease genes are yet to be discovered[7, 121, 122]. Regretfully, most WES/WGS does not yield a diagnosis and consequently multiple VUS candidates remain. New diagnoses cannot be made without functional studies on potential candidates; this is an unrealistic and financially unviable method to discover new disease-causing genes and is a major bottleneck. The MatchMaker Exchange[123] is aiding in this regard; clinicians and analysts can share patient phenotype and genotype data in hope of finding a match. Discovering additional affected kindreds promotes novel disease-gene discovery through functional experiments and shared phenotype data. It is noteworthy though that functional validation is not synonymous with causation. For one, rescue events can restore a given genotype and proving a gene is knocked out/down does not mean it is pathogenic. Furthermore, many functional studies use mouse models; these are contentious and at best can only approximate effects in humans. Zebrafish and xenopus serve as alternatives due to their short generation time and high fecundity, yet are less related to humans than mice[124, 125]. There has been a recent trend towards CRISPR/cas9 gene editing of organoids derived from human or embryonic stem cells, however, many syndromes overlap multiple cell types, therefore organisms often serve as a superior method[126].

**Prioritising candidates using constraint metrics**

If it were possible to introduce disruptive mutations into human genes and survey their effects, many genetic

disorders would already be discovered. However, it is not ethical to create human knockouts. Therefore, these are only identified when comparing naturally-occurring knockouts with patient phenotypes[127, 128]. Yet, with ~100 per genome[47], it is difficult to assign which knockout, if any, is responsible.

Instead, the aggregation of large population datasets has begun to reveal the natural variation of inactivating

variants across the human genome, serving as a model for human gene inactivation. We can leverage population data to evaluate the strength of natural selection at a gene level and model constraint (or shift in allele frequency distribution) compared to an expectation. In 2019, Karczewski *et al.*[35] developed a continuous measure of intolerance to predicted loss-of-function variation modelled on 141,000 WES/WGS samples from the gnomAD database, thus placing each gene on a spectrum of loss-of-function intolerance. This method, called the lower observed/expected upper bound fraction (LOEUF) can predict haploinsufficient genes even in the absence of biological function. Furthermore, similar approaches can be applied to detect missense constraint, including at a regional level, and prioritise both dominant and recessive disease genes[38, 129].

**Discussion**

The genomics field is moving at an unprecedented pace in an era of global collaboration. Efforts to help resolve incomplete penetrance, non-coding variation and structural variants are rapidly advancing and the use of LRS shows real promise. The drive towards data sharing and open-access resources are uplifting novel gene discoveries and refining computational methods. Community assessment efforts to evaluate computational methods are helping to resolve genomic issues around variant annotation and variant calling[9]. Population aggregation databases are widening the spectrum of genomic variation across populations and these data can be used to model mutational patterns and predict essentiality of genes[35]. National genomic sequencing projects are accelerating health research and personalised medicine[6, 130]. That said, the data sharing process is complex with issues around data access permissions, firewalls, and data compatibility[131].

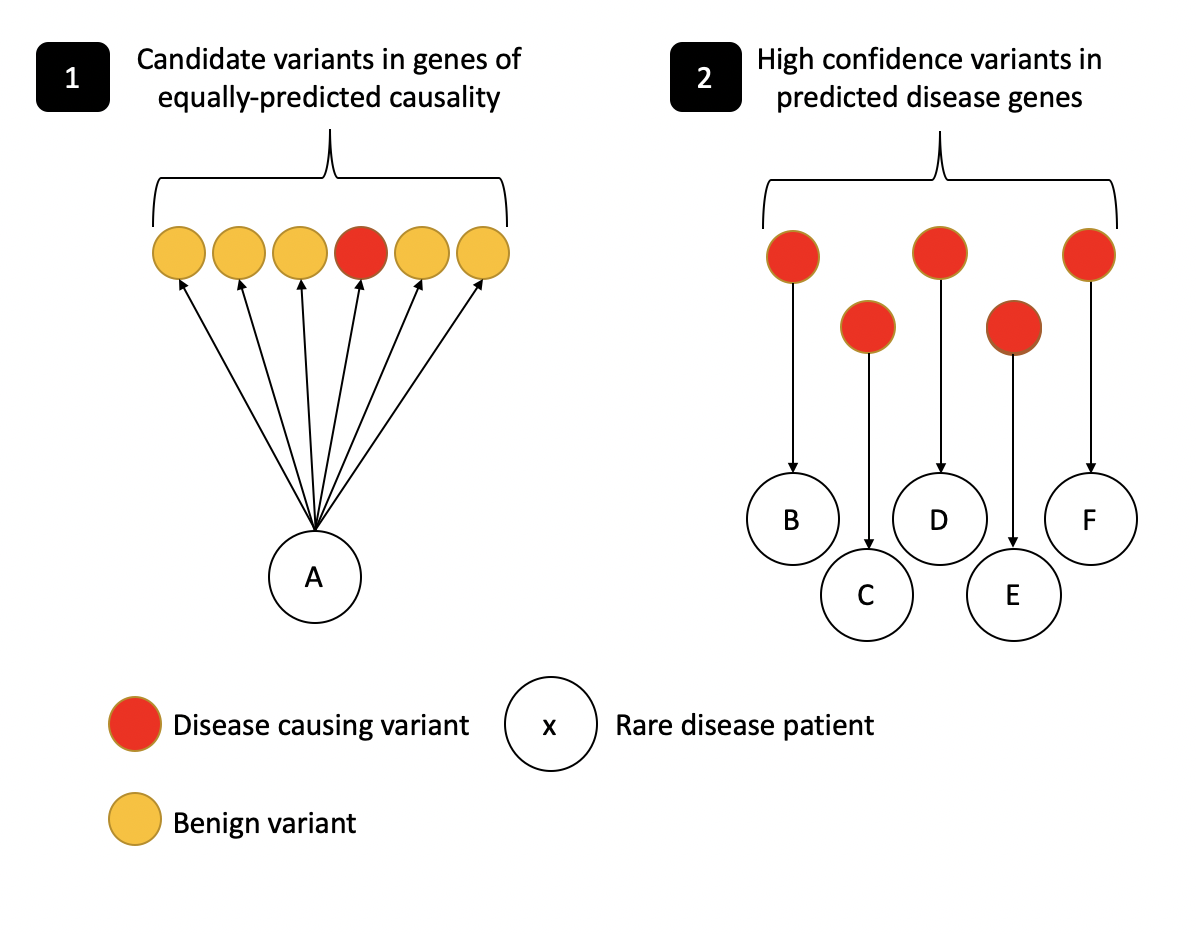
Despite the huge advances in genomics, current approaches to rare disease diagnosis and novel discoveries are still hindered by vast data of uncertain significance found in genes of unknown function. Even with myriad advancing methods, diagnosing rare disease is extremely challenging without a prior correlation between a clinical phenotype and causal gene. This is particularly frustrating when, for many rare disease patients, the causal mutation is present in their sequencing data. New gene discovery is still largely confined by taking single patient cases and bringing the phenotype to the gene; an approach limited by intensive functional experiments on genes of equally-predicted causality. Therefore, there is perhaps need to re-think the current approach to NGS analysis.

**Shifting the current diagnostic paradigm**

Large-scale rare disease sequencing projects are working to end the diagnostic odyssey tainting rare disease[6, 130]. By amassing large cohorts, there are opportunities to improve novel gene discovery using bespoke methodology. The Deciphering Developmental Disorders project has utilised large cohorts to discover novel disease genes by exploiting statistically-driven methods e.g. by computing rare mutational burden in genes versus controls[132].

New constraint metrics such as LOEUF provides a continuous spectrum of intolerance to gene inactivation and can be used to discover novel disease genes[35]. Using the extremes of this spectrum, there is an opportunity to identify predicted haploinsufficient disease genes yet to be discovered by accurately focusing genomic analysis to the most damaging (loss-of-function) variants in LOEUF constrained genes. In essence, adopting a gene-to-patient based approach that matches predicted pathogenic variants to rare disease patients, thus reducing most of the analytical noise that currently precludes novel discoveries (**Figure 5**) and building a much-needed phenotypic catalogue of human gene knockouts that has otherwise been unachievable. This will be most valuable if and when phenotype data are accurate, comprehensive and collected longitudinally. Given the rarity of these mutations, this is only achievable by exploiting massive cohort sizes, now possible by recent global efforts. This method is best applied to non-embryonically lethal diseases caused by variants in haploinsufficient genes, but also shows promise for recessive inheritance; genes that falls in the middle of the LOEUF spectrum are enriched for recessive diseases and can be targeted accordingly[35].

**Figure 5 |** A proposed method for novel disease-gene discovery.



***(1)*** *The current ‘patient to gene approach’. Following sequencing of a given patient (A) there is usually a high burden of variants of unknown significance. One of these candidates may be the causal variant (red), however with literature lacking on the function of all 20,000 human genes, all candidates are essentially of equally-predicted causality; at least until a time that functional experiments can validate pathogenicity.* ***(2)*** *A proposed new method that reverses the patient-to-gene approach. Here, genes are ranked based on predicted haploinsufficiency, modelled from population data (LOEUF constraint metric). High confidence inactivating variants are filtered on a curated list of predicted disease-causing genes and are extracted from rare disease cohorts independent of phenotype. These predicted pathogenic variants are then matched to individual patients (B, C, D, E and F).*

**Conclusion**

Diagnosing rare diseases is challenging, however the pace at which genomics has moved from a research setting into mainstream clinical medicine exemplifies its utility. Methods are continually advancing to improve variant analysis and interpretation, particularly by modelling gene constraint from population data, utilising machine-learning and deep neural networks to predict variant consequence, and visualising transcript- and tissue-specific annotation. Furthermore, recent efforts to understand the non-coding genome are revealing a large spectrum of variation implicated in disease.

With a wealth of research into the human genome, it is expected that diagnostic rates will continue to rise. Rethinking current practices by utilising large cohorts and adopting gene-based approaches may accelerate novel gene discovery, where functional experiments on genes of equally-predicted causality have been a major bottleneck. The mainstream integration of SV calling into informatics pipelines is likely to resolve a percentage of undiagnosed rare diseases and it is hoped that as control cohorts increase in size, incomplete penetrance will be more accurately quantified. Genomics has already revolutionised rare disease diagnostics, and as we start the new decade we can likely look ahead to significant diagnostic uplift which will facilitate the development of personalised therapies and begin to end the diagnostic odyssey of rare disease.

**Key points**

1. Genomics has revolutionised rare disease diagnostics. Consequently, whole exome and genome sequencing are now recognised clinical diagnostic tests.
2. Current diagnostic rates for exome and genome sequencing are ~30% on average. This is challenged by clinical diagnostic testing limiting analysis to known disease genes. Therefore, disease variants in genes of unknown function are missed. Novel disease-gene discovery is required to uplift diagnostic rates and is being facilitated by national genome sequencing projects, whereby patients are contemporaneously consented for both diagnostic and research studies.
3. Over 75% of the phenotypical consequences of variants in the exome are unaccounted for. Therefore, there is still much to be understood about the biological function of all ~20,000 human genes before clinical diagnostic testing will significantly uplift diagnostic rates.
4. Many disease variants reside outside of coding regions. Research into structural variants and non-coding variation is rapidly advancing. Long-read sequencing shows real promise in elucidating the role of non-coding DNA and structural variation, however cost is a substantial issue.
5. Variant interpretation is challenging and is hindered by millions of variants called per genome. Utilising segregation, *in silico* tools and allele frequencies can help in this regard, however in many cases, multiple variants of uncertain significance remain. These require functional studies to validate their effects, which is impossible to achieve for multiple candidates per exome/genome. Leveraging large datasets and exploiting constraint metrics provide an opportunity to adopt gene-to-patient based approaches.

**Keywords**

Rare disease, genomics, diagnostics, exome sequencing, novel gene discovery, diagnostic odyssey

**Funding**

EGS is supported by the Kerkut Charitable Trust.

**References:**

ADDIN EN.REFLIST 1. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing, New England Journal of Medicine 2014;370:2418-2425.

2. Adams DR, Eng CM. Next-generation sequencing to diagnose suspected genetic disorders, New England Journal of Medicine 2018;379:1353-1362.

3. Retterer K, Juusola J, Cho MT et al. Clinical application of whole-exome sequencing across clinical indications, Genetics in Medicine 2016;18:696-704.

4. Lee H, Deignan JL, Dorrani N et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders, Jama 2014;312:1880-1887.

5. Siva N. 1000 Genomes project. Nature Publishing Group, 2008.

6. Sankar PL, Parker LS. The Precision Medicine Initiative’s All of Us Research Program: an agenda for research on its ethical, legal, and social issues, Genetics in Medicine 2017;19:743-750.

7. Posey JE, O’Donnell-Luria AH, Chong JX et al. Insights into genetics, human biology and disease gleaned from family based genomic studies, Genetics in Medicine 2019;21:798-812.

8. Saez-Rodriguez J, Costello JC, Friend SH et al. Crowdsourcing biomedical research: leveraging communities as innovation engines, Nature Reviews Genetics 2016;17:470.

9. Hoskins RA, Repo S, Barsky D et al. Reports from CAGI: the critical assessment of genome interpretation, Human mutation 2017;38:1039-1041.

10. Vissers LE, de Ligt J, Gilissen C et al. A de novo paradigm for mental retardation, Nature genetics 2010;42:1109-1112.

11. Bamshad MJ, Ng SB, Bigham AW et al. Exome sequencing as a tool for Mendelian disease gene discovery, Nature Reviews Genetics 2011;12:745-755.

12. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: a practical guide to its clinical application, Briefings in Functional Genomics 2015;15:374-384.

13. Stals KL, Wakeling M, Baptista J et al. Diagnosis of lethal or prenatal‐onset autosomal recessive disorders by parental exome sequencing, Prenatal diagnosis 2018;38:33-43.

14. Robinson PN, Köhler S, Bauer S et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease, The American Journal of Human Genetics 2008;83:610-615.

15. Seaby EG, Gilbert RD, Andreoletti G et al. Unexpected findings in a child with atypical hemolytic uremic syndrome: An example of how genomics is changing the clinical diagnostic paradigm, Frontiers in pediatrics 2017;5:113.

16. Escalona M, Rocha S, Posada D. A comparison of tools for the simulation of genomic next-generation sequencing data, Nature Reviews Genetics 2016;17:459.

17. Roy S, Coldren C, Karunamurthy A et al. 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, The Journal of Molecular Diagnostics 2018;20:4-27.

18. DePristo MA, Banks E, Poplin R et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nature genetics 2011;43:491.

19. Cornish A, Guda C. A comparison of variant calling pipelines using genome in a bottle as a reference, BioMed research international 2015;2015.

20. Robinson JT, Thorvaldsdóttir H, Winckler W et al. Integrative genomics viewer, Nature biotechnology 2011;29:24-26.

21. Fujita PA, Rhead B, Zweig AS et al. The UCSC genome browser database: update 2011, Nucleic acids research 2010;39:D876-D882.

22. Sims D, Sudbery I, Ilott NE et al. Sequencing depth and coverage: key considerations in genomic analyses, Nature Reviews Genetics 2014;15:121.

23. Tavtigian SV, Greenblatt MS, Harrison SM et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework, Genetics in Medicine 2018;20:1054-1060.

24. Amendola LM, Jarvik GP, Leo MC et al. Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the Clinical Sequencing Exploratory Research Consortium, The American Journal of Human Genetics 2016;98:1067-1076.

25. Wenger AM, Guturu H, Bernstein JA et al. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers, Genetics in Medicine 2017;19:209-214.

26. Nambot S, Thevenon J, Kuentz P et al. Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis, Genetics in Medicine 2018;20:645-654.

27. Ewans LJ, Schofield D, Shrestha R et al. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders, Genetics in Medicine 2018;20:1564-1574.

28. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease, Genome Biology 2016;17:241.

29. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders, Nature 2017;542:433-438.

30. Kong A, Frigge ML, Masson G et al. Rate of de novo mutations and the importance of father’s age to disease risk, Nature 2012;488:471-475.

31. Rahbari R, Wuster A, Lindsay SJ et al. Timing, rates and spectra of human germline mutation, Nature genetics 2016;48:126-133.

32. Arnadottir GA, Jensson BO, Marelsson SE et al. Compound heterozygous mutations in UBA5 causing early-onset epileptic encephalopathy in two sisters, BMC Medical Genetics 2017;18:103.

33. Zernant J, Lee W, Collison FT et al. Frequent hypomorphic alleles account for a significant fraction of ABCA4 disease and distinguish it from age-related macular degeneration, Journal of Medical Genetics 2017;54:404-412.

34. Karczewski KJ, Weisburd B, Thomas B et al. The ExAC browser: displaying reference data information from over 60 000 exomes, Nucleic acids research 2017;45:D840-D845.

35. Karczewski KJ, Francioli LC, Tiao G et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes, bioRxiv 2019:531210.

36. Whiffin N, Minikel E, Walsh R et al. Using high-resolution variant frequencies to empower clinical genome interpretation, Genetics in Medicine 2017;19:1151-1158.

37. Cummings BB, Karczewski KJ, Kosmicki JA et al. Transcript expression-aware annotation improves rare variant discovery and interpretation, bioRxiv 2019:554444.

38. Havrilla JM, Pedersen BS, Layer RM et al. A map of constrained coding regions in the human genome, Nature genetics 2019;51:88-95.

39. Samocha KE, Kosmicki JA, Karczewski KJ et al. Regional missense constraint improves variant deleteriousness prediction, bioRxiv 2017:148353.

40. Hurst LD. The sound of silence, Nature 2011;471:582-583.

41. Hunt RC, Simhadri VL, Iandoli M et al. Exposing synonymous mutations, Trends in Genetics 2014;30:308-321.

42. Bao S, Moakley DF, Zhang C. The splicing code goes deep, Cell 2019;176:414-416.

43. Wai HA, Lord J, Lyon M et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance, Genetics in Medicine 2020:1-10.

44. Soemedi R, Cygan KJ, Rhine CL et al. Pathogenic variants that alter protein code often disrupt splicing, Nature genetics 2017;49:848.

45. Adamson SI, Zhan L, Graveley BR. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency, Genome Biology 2018;19:71.

46. Lin M, Whitmire S, Chen J et al. Effects of short indels on protein structure and function in human genomes, Scientific Reports 2017;7:9313.

47. MacArthur DG, Balasubramanian S, Frankish A et al. A systematic survey of loss-of-function variants in human protein-coding genes, Science 2012;335:823-828.

48. Consortium GP. A global reference for human genetic variation, Nature 2015;526:68-74.

49. Shen H, Li J, Zhang J et al. Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians, PloS one 2013;8:e59494-e59494.

50. Jiang Y, Turinsky AL, Brudno M. The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection, Nucleic acids research 2015;43:7217-7228.

51. Jagadeesh KA, Wenger AM, Berger MJ et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity, Nature genetics 2016;48:1581-1586.

52. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants, Bioinformatics 2015;31:761-763.

53. Li J, Shi L, Zhang K et al. VarCards: an integrated genetic and clinical database for coding variants in the human genome, Nucleic acids research 2018;46:D1039-D1048.

54. Sim N-L, Kumar P, Hu J et al. SIFT web server: predicting effects of amino acid substitutions on proteins, Nucleic acids research 2012;40:W452-W457.

55. Adzhubei IA, Schmidt S, Peshkin L et al. A method and server for predicting damaging missense mutations, Nature methods 2010;7:248-249.

56. Schwarz JM, Cooper DN, Schuelke M et al. MutationTaster2: mutation prediction for the deep-sequencing age, Nature methods 2014;11:361-362.

57. Grantham R. Amino acid difference formula to help explain protein evolution, Science 1974;185:862-864.

58. Choi Y, Sims GE, Murphy S et al. Predicting the functional effect of amino acid substitutions and indels, PloS one 2012;7.

59. Alirezaie N, Kernohan KD, Hartley T et al. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants, The American Journal of Human Genetics 2018;103:474-483.

60. Wang M, Zhao X-M, Takemoto K et al. FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model, PloS one 2012;7.

61. Shihab HA, Rogers MF, Gough J et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation, Bioinformatics 2015;31:1536-1543.

62. Ritchie GR, Dunham I, Zeggini E et al. Functional annotation of noncoding sequence variants, Nature methods 2014;11:294.

63. Gelfman S, Wang Q, McSweeney KM et al. Annotating pathogenic non-coding variants in genic regions, Nature Communications 2017;8:236.

64. Ratan A, Olson TL, Loughran TP et al. Identification of indels in next-generation sequencing data, BMC Bioinformatics 2015;16:42.

65. Mose LE, Wilkerson MD, Hayes DN et al. ABRA: improved coding indel detection via assembly-based realignment, Bioinformatics 2014;30:2813-2815.

66. Koboldt DC, Chen K, Wylie T et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples, Bioinformatics 2009;25:2283-2285.

67. Albers CA, Lunter G, MacArthur DG et al. Dindel: accurate indel calls from short-read data, Genome research 2011;21:961-973.

68. Folkman L, Yang Y, Li Z et al. DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels, Bioinformatics 2015;31:1599-1606.

69. Poplin R, Chang P-C, Alexander D et al. A universal SNP and small-indel variant caller using deep neural networks, Nature biotechnology 2018;36:983-987.

70. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction, Nucleic acids research 2001;29:1185-1190.

71. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, Journal of computational biology 2004;11:377-394.

72. Desmet F-O, Hamroun D, Lalande M et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals, Nucleic acids research 2009;37:e67-e67.

73. Mort M, Sterne-Weiler T, Li B et al. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing, Genome Biology 2014;15:R19.

74. Jaganathan K, Panagiotopoulou SK, McRae JF et al. Predicting splicing from primary sequence with deep learning, Cell 2019;176:535-548. e524.

75. Pollard KS, Hubisz MJ, Rosenbloom KR et al. Detection of nonneutral substitution rates on mammalian phylogenies, Genome research 2010;20:110-121.

76. Cooper GM, Stone EA, Asimenos G et al. Distribution and intensity of constraint in mammalian genomic sequence, Genome research 2005;15:901-913.

77. Siepel A, Bejerano G, Pedersen JS et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, Genome research 2005;15:1034-1050.

78. Tang H, Thomas PD. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation, Bioinformatics 2016;32:2230-2232.

79. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model, Nature methods 2015;12:931-934.

80. Lu Q, Hu Y, Sun J et al. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data, Scientific Reports 2015;5:10576.

81. Lehmann K-V, Chen T. Exploring functional variant discovery in non-coding regions with SInBaD, Nucleic acids research 2012;41:e7-e7.

82. Zambrano R, Jamroz M, Szczasiuk A et al. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures, Nucleic acids research 2015;43:W306-W313.

83. Pires DEV, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach, Nucleic acids research 2014;42:W314-W319.

84. Liu M, Watson LT, Zhang L. HMMvar-func: a new method for predicting the functional outcome of genetic variants, BMC Bioinformatics 2015;16:351.

85. Ryan M, Diekhans M, Lien S et al. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures, Bioinformatics 2009;25:1431-1432.

86. Giollo M, Martin AJ, Walsh I et al. NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation, BMC genomics 2014;15:S7.

87. López-Ferrando V, Gazzo A, de la Cruz X et al. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update, Nucleic acids research 2017;45:W222-W228.

88. Kircher M, Witten DM, Jain P et al. A general framework for estimating the relative pathogenicity of human genetic variants, Nature genetics 2014;46:310.

89. Ionita-Laza I, McCallum K, Xu B et al. A spectral approach integrating functional genomic annotations for coding and noncoding variants, Nature genetics 2016;48:214.

90. Ioannidis NM, Rothstein JH, Pejaver V et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants, The American Journal of Human Genetics 2016;99:877-885.

91. Mossotto E, Ashton JJ, O’Gorman L et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data, BMC Bioinformatics 2019;20:254.

92. Lonsdale J, Thomas J, Salvatore M et al. The genotype-tissue expression (GTEx) project, Nature genetics 2013;45:580.

93. Ware JS, Li J, Mazaika E et al. Shared genetic predisposition in peripartum and dilated cardiomyopathies, New England Journal of Medicine 2016;374:233-241.

94. Roberts AM, Ware JS, Herman DS et al. Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease, Science translational medicine 2015;7:270ra276-270ra276.

95. den Dunnen JT, Dalgleish R, Maglott DR et al. HGVS recommendations for the description of sequence variants: 2016 update, Human mutation 2016;37:564-569.

96. Cooper DN, Krawczak M, Polychronakos C et al. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease, Human genetics 2013;132:1077-1130.

97. Chen R, Shi L, Hakenberg J et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases, Nature biotechnology 2016;34:531.

98. Shawky RM. Reduced penetrance in human inherited disease, Egyptian Journal of Medical Human Genetics 2014;15:103-111.

99. Ropers HH, Wienker T. Penetrance of pathogenic mutations in haploinsufficient genes for intellectual disability and related disorders, European journal of medical genetics 2015;58:715-718.

100. Fahsold R, Hoffmeyer S, Mischung C et al. Minor lesion mutational spectrum of the entire NF1 gene does not explain its high mutability but points to a functional domain upstream of the GAP-related domain, The American Journal of Human Genetics 2000;66:790-818.

101. Bresin E, Rurali E, Caprioli J et al. Combined complement gene mutations in atypical hemolytic uremic syndrome influence clinical phenotype, Journal of the American Society of Nephrology 2013;24:475-486.

102. Veitia RA, Caburet S, Birchler JA. Mechanisms of Mendelian dominance, Clinical Genetics 2018;93:419-428.

103. Alessi DR, Sammler E. LRRK2 kinase in Parkinson's disease, Science 2018;360:36-37.

104. Gourraud J-B, Barc J, Thollet A et al. The Brugada syndrome: a rare arrhythmia disorder with complex inheritance, Frontiers in cardiovascular medicine 2016;3:9.

105. Marsh AP, Heron D, Edwards TJ et al. Mutations in DCC cause isolated agenesis of the corpus callosum with incomplete penetrance, Nature genetics 2017;49:511-514.

106. Wright CF, West B, Tuke M et al. Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting, The American Journal of Human Genetics 2019;104:275-286.

107. Spielmann M, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome, Nature Reviews Genetics 2018;19:453-467.

108. Zhang F, Lupski JR. Non-coding genetic variants in human disease, Human molecular genetics 2015;24:R102-R110.

109. Spielmann M, Mundlos S. Looking beyond the genes: the role of non-coding variants in human disease, Human molecular genetics 2016;25:R157-R165.

110. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data, Nature genetics 2017;49:618.

111. Collins RL, Brand H, Karczewski KJ et al. An open resource of structural variation for medical and population genetics, bioRxiv 2019:578674.

112. Holt JM, Birch CL, Brown DM et al. Identification of Pathogenic Structural Variants in Rare Disease Patients through Genome Sequencing, bioRxiv 2019:627661.

113. Merker JD, Wenger AM, Sneddon T et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease, Genetics in Medicine 2018;20:159-163.

114. Marks P, Garcia S, Barrio AM et al. Resolving the full spectrum of human genome variation using Linked-Reads, Genome research 2019;29:635-645.

115. Hamosh A, Scott AF, Amberger JS et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic acids research 2005;33:D514-D517.

116. Landrum MJ, Lee JM, Riley GR et al. ClinVar: public archive of relationships among sequence variation and human phenotype, Nucleic acids research 2014;42:D980-D985.

117. Firth HV, Richards SM, Bevan AP et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources, The American Journal of Human Genetics 2009;84:524-533.

118. Manrai AK, Funke BH, Rehm HL et al. Genetic Misdiagnoses and the Potential for Health Disparities, New England Journal of Medicine 2016;375:655-665.

119. Lek M, Karczewski KJ, Minikel EV et al. Analysis of protein-coding genetic variation in 60,706 humans, Nature 2016;536:285-291.

120. Shah N, Hou Y-CC, Yu H-C et al. Identification of misclassified ClinVar variants via disease population prevalence, The American Journal of Human Genetics 2018;102:609-619.

121. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children, Nature Reviews Genetics 2018;19:253.

122. Pengelly RJ, Vergara-Lope A, Alyousfi D et al. Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation, Briefings in Bioinformatics 2017;20:267-273.

123. Philippakis AA, Azzariti DR, Beltran S et al. The Matchmaker Exchange: a platform for rare disease gene discovery, Human mutation 2015;36:915-921.

124. Stainier DYR, Raz E, Lawson ND et al. Guidelines for morpholino use in zebrafish, PLoS genetics 2017;13:e1007000-e1007000.

125. Nenni MJ, Fisher ME, James-Zorn C et al. Xenbase: facilitating the use of Xenopus to model human disease, Frontiers in physiology 2019;10:154.

126. Fujii M, Clevers H, Sato T. Modeling Human Digestive Diseases With CRISPR-Cas9–Modified Organoids, Gastroenterology 2019;156:562-576.

127. Perdigoto C. Mutations: dawn of the human knockout project, Nature Reviews Genetics 2017;18:328.

128. Saleheen D, Natarajan P, Armean IM et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity, Nature 2017;544:235-239.

129. Abramovs N, Brass A, Tassabehji M. GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes, Nature genetics 2020;52:35-39.

130. Turnbull C, Scott RH, Thomas E et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS, Bmj 2018;361:k1687.

131. Kaye J. The tension between data sharing and the protection of privacy in genomics research, Annual review of genomics and human genetics 2012;13:415-431.

132. Study TDDD, Fitzgerald T, Gerety S et al. Large-scale discovery of novel genetic causes of developmental disorders, Nature 2015;519:223-228.