Ellingford *et al. Genome Medicine* (2022) 14:73
https://doi.org/10.1186/s13073-022-01073-3

Genome Medicine

GUIDELINE

Open Access

Recommendations for clinical interpretation of variants found in non-coding regions of the genome

Jamie M. Ellingford[1,2,3*†], Joo Wook Ahn[4†], Richard D. Bagnall[5], Diana Baralle[6,7†], Stephanie Barton[2], Chris Campbell[2], Kate Downes[4], Sian Ellard[8,9†], Celia Duff-Farrier[10], David R. FitzPatrick[11†], John M. Greally[12], Jodie Ingles[13,14], Neesha Krishnan[13,14], Jenny Lord[6], Hilary C. Martin[15], William G. Newman[1,2†], Anne O'Donnell-Luria[16,17,18], Simon C. Ramsden[2], Heidi L. Rehm[16,18], Ebony Richardson[13,14], Moriel Singer-Berk[16], Jenny C. Taylor[19,20†], Maggie Williams[10], Jordan C. Wood[16], Caroline F. Wright[8], Steven M. Harrison[16,21†] and Nicola Whiffin[16,20*†]

Abstract

Background: The majority of clinical genetic testing focuses almost exclusively on regions of the genome that directly encode proteins. The important role of variants in non-coding regions in penetrant disease is, however, increasingly being demonstrated, and the use of whole genome sequencing in clinical diagnostic settings is rising across a large range of genetic disorders. Despite this, there is no existing guidance on how current guidelines designed primarily for variants in protein-coding regions should be adapted for variants identified in other genomic contexts.

Methods: We convened a panel of nine clinical and research scientists with wide-ranging expertise in clinical variant interpretation, with specific experience in variants within non-coding regions. This panel discussed and refined an initial draft of the guidelines which were then extensively tested and reviewed by external groups.

Results: We discuss considerations specifically for variants in non-coding regions of the genome. We outline how to define candidate regulatory elements, highlight examples of mechanisms through which non-coding region variants can lead to penetrant monogenic disease, and outline how existing guidelines can be adapted for the interpretation of these variants.

Conclusions: These recommendations aim to increase the number and range of non-coding region variants that can be clinically interpreted, which, together with a compatible phenotype, can lead to new diagnoses and catalyse the discovery of novel disease mechanisms.

Keywords: Variant interpretation, Non-coding variation, Gene regulation

Background

Genomic sequencing is commonplace in the diagnosis of disorders with suspected genetic cause. Traditionally, sequencing and analysis has focussed primarily on variants that fall within regions of the genome that code directly for protein, or that are within canonical splice sites of genes with a confirmed role in disease. With these approaches, however, many rare disease cases remain genetically unexplained [1, 2].

†The following authors formed the expert panel: Jamie M. Ellingford; Joo Wook Ahn; Diana Baralle; Sian Ellard; David R. FitzPatrick; William G. Newman; Jenny C. Taylor; Steven M. Harrison; Nicola Whiffin

*Correspondence: jamie.ellingford@manchester.ac.uk; nwhiffin@well.ox.ac.uk

[1] Division of Evolution, Infection and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicines and Health, University of Manchester, Manchester M13 9PT, UK

[20] Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK

Full list of author information is available at the end of the article

Increasingly, whole genome sequencing (WGS) is being performed on individuals in which a genetic cause is not identified through gene panel or exome sequencing. For some disease subsets and in specific healthcare settings, WGS has become a first-line diagnostic test (for example, for selected rare developmental disorders in the UK National Health Service) [3]. WGS has been shown to have the potential to increase diagnostic yield [2, 4–7] and includes detection of variants in a wide range of regulatory regions (Box 1, Fig. 1) as well as variants in genes encoding non-coding RNAs (e.g. micro RNAs (miRNA), small nuclear RNAs (snRNA) and long non-coding RNAs). Analysis of WGS data, however, often excludes variants that fall in non-coding regions of the genome or classifies them as variants of uncertain significance (VUS), primarily due to difficulties in predicting or determining their impact. Whilst the triplet amino acid code allows us to predict the effect of variants within protein-coding regions with reasonable accuracy, the absence of a regulatory equivalent means that the impacts of non-coding region variants are usually much harder to predict. This is further confounded by these variants often having gene-specific effects. For example, binding sites for the zinc finger protein CCCTC-binding factor (CTCF) are enriched at topologically associated domain (TAD) boundaries and have been suggested as a mechanism to ensure appropriate genome regulation and chromatin structure. However, it is unclear why in some instances disrupting CTCF binding sites significantly impacts gene expression [8], and in others it does not [9, 10]. Such differences may be dependent on the surrounding genomic context and the temporal/spatial activity of other cis-regulatory elements (CREs) [11].

Box 1 Regulatory elements controlling gene and protein expression

Gene and protein expression are tightly controlled processes mediated by a multitude of regulatory elements (Fig. 1). Transcription of a gene into RNA is mediated by a promoter element directly upstream of the gene [12], along with more distal enhancer and repressor elements (collectively referred to as cis-regulatory elements, or CREs) to which transcription factors bind [13]. CREs may be within their target gene or in other intragenic or intergenic space either 3′ or 5′ of the transcription unit they influence. Within the gene itself, intronic regions contain specific sequences that control their removal through splicing to form the mature RNA (mRNA) transcript, and untranslated regions (UTRs) regulate RNA stability, trafficking, and the rate at which it is translated into protein [14]. Each gene also sits within a wider regulatory context, or topologically associated domain (TAD), flanked by boundary/insulator elements which restrict the action of CREs to within specific TADs [13, 15].

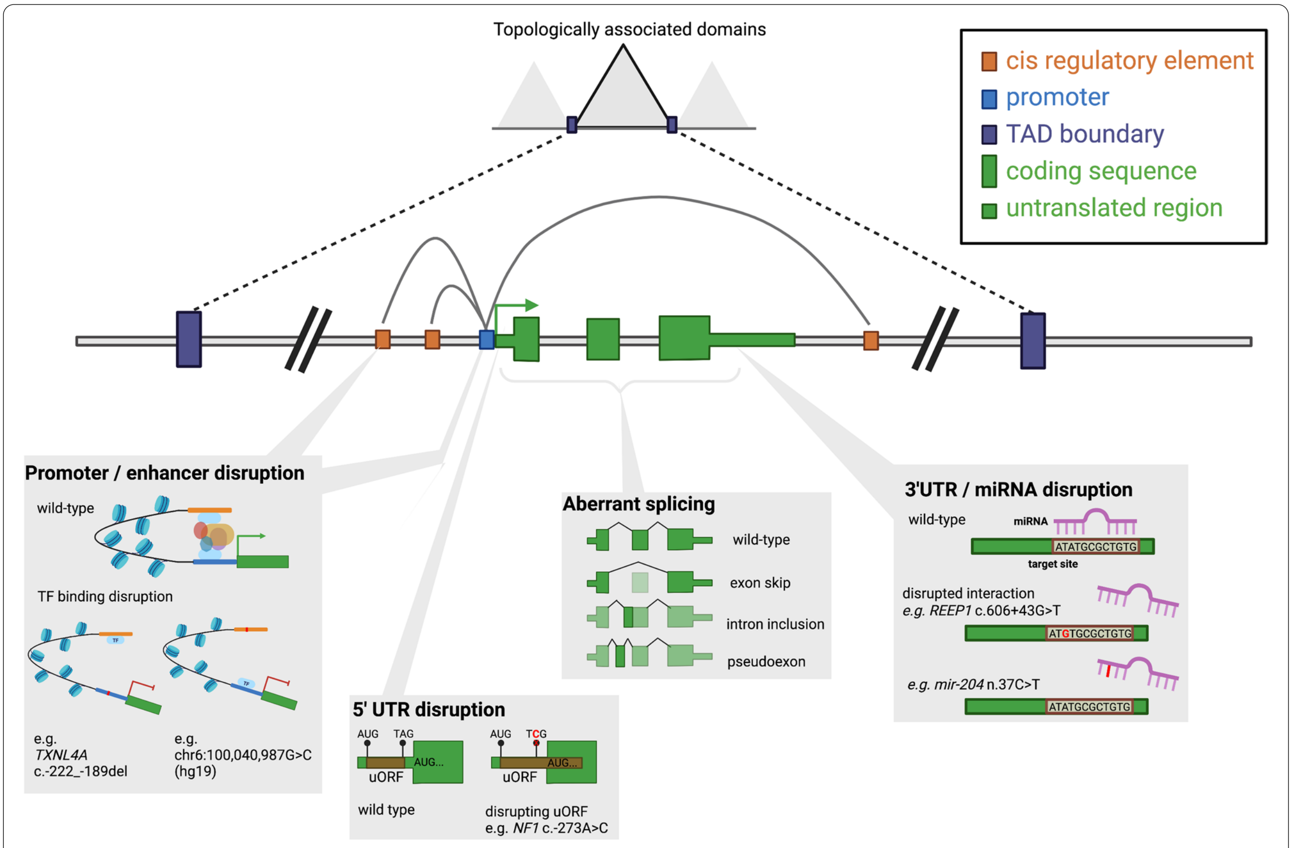


Fig. 1 Schematic of regulatory elements within and around a gene and examples of disruptions that can lead to disease

An important role for a range of non-coding variants in rare disease is increasingly being demonstrated [16–18]. For example, variants in upstream non-coding regions that cause loss-of-function of *MEF2C* comprise almost one quarter of all likely diagnoses impacting *MEF2C* in the Deciphering Developmental Disorders (DDD) dataset [19], and RNA sequencing can be used to identify likely disease-causing splicing variants in 35% of previously undiagnosed rare muscle disease probands, many in deep intronic regions [20]. A number of disease-causing variants for X-linked Charcot-Marie-Tooth disease in *GJB1* [21] and ABCA4-associated disease [22, 23] are also known to be in non-coding regions. These are just a few examples, with many more existing across a range of different genes and diseases.

There are a multitude of documented mechanisms through which non-coding region variants which disrupt non-coding elements have been demonstrated to cause severe disease (Table 1). These include acting through affecting splicing [20, 24, 25], transcription [26, 27], translation [19], RNA processing and stability [16, 28] and chromatin interactions [29]. Detecting and classifying these variants accurately for a likely disease-causing role is important to increase diagnostic yield and enable a robust genetic diagnosis for more individuals.

The American College of Medical Genetics and Genomics and Association for Molecular Pathology (ACMG/AMP) released a set of guidelines in 2015 that have become the global standard for interpreting the pathogenicity of short sequence variants (single-nucleotide variants (SNVs) and indels <50 bps) identified in individuals with rare disease [30]. These guidelines outline a set of rules that should be assessed for each identified variant. Many of these rules pertain specifically to variants in protein-coding regions and there is no existing guidance on how they should be adapted for variants found in other genomic contexts. Here, we provide guidance on how to apply these standards to variants identified in non-coding regions of the genome. Our recommendations will enable consistent interpretation and reporting of these understudied variant types which will in turn enable us to learn more about the diverse mechanisms through which non-coding region variants can lead to disease.

Methods

Process for drafting and refining the recommendations

We convened a panel of nine clinical and research scientists with wide-ranging expertise in clinical variant interpretation, with specific interests and experience in variants within non-coding regions. The initial document was drafted and circulated before a series of online calls to discuss and refine the guidance. Subsequently, we asked a range of clinical scientists and those actively involved in clinical variant interpretation to assess the usability of the guidelines on both a common list of 30 diverse variants (Additional file 1: Table S1) and in-house identified variants. Feedback from testers was used to further refine the guidance.

At the Association for Clinical Genomic Science (ACGS) meeting in September 2021, we presented an overview of the new guidelines to the UK Clinical Science community. At this meeting, we polled opinions on current best practice, the need for specific guidelines for non-coding region variants, and the appetite for training workshops/seminars (Additional file 1: Table S2). The attendees of the workshop were overwhelmingly in support of this effort with 98% (59/60) agreeing that additional guidance is needed to support interpretation of non-coding region variants.

Assessing the under-ascertainment of non-coding region variants in ClinVar

To identify non-coding region variants in ClinVar [31], all variants (n = 789,941) from the ClinVar GRCh38 VCF dated 01/31/2021 were annotated with respect to Matched Annotation from NCBI and EMBL-EBI (MANE) Select [32] v0.93 transcripts. Variants were assigned as falling within the coding sequence (n = 597,408), 5′UTR (n = 10,820), 3′UTR (n = 53,988), intronic regions (n = 110,618), or in the 2-kb upstream of the transcription start site (annotated as promoter; n = 4348). All remaining variants (n = 12,759) were assigned as 'other'. The majority of these 'other' variants were coding sequence variants in genes without designated MANE Select transcripts.

High-confidence pathogenic variants were designated as those with a review status of 'criteria_provided,_multiple_submitters,_no_conflicts', 'reviewed_by_expert_panel', or 'practice_guideline'. Pathogenic variants were taken as those with significance of 'Pathogenic', 'Likely_pathogenic', or 'Pathogenic/Likely_pathogenic'.

Identifying *in trans* non-coding region variants in 100,000 Genomes Project to inform PM3

To determine the frequency of variants observed *in trans* with potentially pathogenic variants, we used the Genomics England (GEL) 100,000 Genomes dataset (version 7) [7]. We identified all probands recruited as full trios (i.e. an affected proband and both unaffected parents) without variants classified as either tier 1 or tier 2 in the GEL clinical filtering pipeline [33]. We next identified all remaining probands with a single heterozygous predicted loss-of-function (pLoF) variant in one of 794 genes catalogued as biallelic loss-of-function (LoF) genes

Table 1 Categories of small variants in non-coding regions previously implicated in penetrant Mendelian disease

Region	Mechanism	Example (variant/gene)	Example (disease)	ClinVar Var ID	Reference	VEP categories	In silico tools to predict effect
Promoter	Altering transcription factor binding	GATA1:−113A>G	Hereditary persistence of foetal haemoglobin		1	Upstream gene variant / regulatory region / TF binding site	TF binding site disruption prediction tools e.g. motifbreakR / SEMpl / QBiC-Pred
Promoter	Altering transcription	CHM c.-98C>A, c.-98C>T	Choroideremia		2		
Promoter/5′UTR	Altering methylation patterns	BRCA1:c.-107A>T	Breast and ovarian cancer		3	Upstream gene variant / 5 prime UTR variant	
5′UTR	Creating upstream start site (uAUG)	NF1:c.-272G>A	Neurofibromatosis type 1	1013130	4	5 prime UTR variant	e.g. UTRannotator
5′UTR	Perturbing upstream open reading frames	NF2:−66-65insT	Neurofibromatosis type 2		5		e.g. UTRannotator
5′UTR	Disrupting internal ribosome entry sites (IRES)	GJB1:c.-103C>T	Charcot-Marie-Tooth disease	217166	6		e.g. IRESpy / IRESfinder / IRESite
5′UTR	Disrupting splicing	SHOX:c.-19G>A	SHOX haploinsufficiency	933226	7		Splicing prediction tools e.g. SpliceAI
5′UTR	Altering Kozak consensus of start site	GATA4:c.-6G>C	Atrial septal defect		8		e.g. utR.annotation
5′UTR	N-terminal transcript elongation	MEF2C:c.−8C>T	Developmental disorder		9		e.g. UTRannotator
Intron	Disrupting canonical splice sites	MYBPC3:c.3490+1G>A	Hypertrophic cardiomyopathy	42715	10	Splice donor variant / splice acceptor variant / splice region variant / splice_donor_5th_base_variant / splice_polypyrimidine_tract_variant	Splicing prediction tools e.g. SpliceAI
Intron	Disrupting splicing branch point	HNF4A:c.264-21A>G	Maturity-onset diabetes of the young		11	Intron variant	Splicing prediction tools e.g. SpliceAI
Intron	Pseudo-exon activation	DMD: c.7310-19A>G	Muscular dystrophy		12		Splicing prediction tools e.g. SpliceAI
Intron	Poison-exon inclusion	SCN1A:c.4002+2165C>T	Dravet Syndrome		13		Splicing prediction tools e.g. SpliceAI
Intron	Branchpoint mutation	BBS1:c.592-21A>T	Retinitis pigmentosa		14		Splicing prediction tools e.g. SpliceAI
Intron	Indels & spacing of splicing motifs	DOK7:c.54+8_54+17del			15		Splicing prediction tools e.g. SpliceAI
Intron	Cryptic exon	VHL: c.3401770T.C	Erythrocytosis		16		

Table 1 (continued)

Region	Mechanism	Example (variant/gene)	Example (disease)	ClinVar Var ID	Reference	VEP categories	In silico tools to predict effect
3′UTR	Disrupting polyA signal motif	NAA10:c.^{a}43A>G	Microphthalmia	617463	17	3 prime UTR variant	PolyA signal motif prediction tools e.g. Omni-PolyA
3′UTR	Disrupting miRNA interactions	REEP1	Hereditary spastic paraplegia		18		miRNA binding site prediction e.g. miRTarBase
3′UTR	Disrupting splicing	LHFPL5:c.^{a}16+1G>A	Hearing impairment		19		Splicing prediction tools e.g. SpliceAI
CRE	Altering transcription factor binding	chr7:155754267:C>T; NCBI build 36.3	Holoprosencephaly		20	Upstream gene variant / downstream gene variant / regulatory region variant / TF binding site variant / intergenic variant / TFBS ablation / TFBS amplification / regulatory region ablation / regulatory region amplification	TF binding site disruption prediction tools e.g. motifbreakR / SEMpl / QBiC-Pred
CRE	Abolishing enhancer activity	PTF1A - 6 variants	Isolated pancreatic agenesis		21		
CRE	Disrupting enhancer activity	SOX9 - deletion (chr17:67,628,756–67,634,155)	Pierre Robin sequence		22		
Intergenic	Creating new regulatory element	chr16:209,709 T>C	α-thalassaemia		23	Intergenic variant / upstream gene variant / downstream gene variant	
miRNA	Disrupting seed region	miR-204:n.37C>T	Retinal dystrophy		24	non-coding transcript exon variant / non-coding transcript variant / mature_miRNA_variant	
snRNA	Altering structure	RNU12	Cerebellar ataxia		25	Non-coding transcript exon variant / non-coding transcript variant	
snRNA	Abnormal splicing, accumulation of minor intron retained transcripts	RNU4ATAC	Roifman Syndrome		26		Splicing prediction tools e.g. SpliceAI
snRNA	Affecting expression, processing and protein binding	SNORD118	Cerebral microangiopathy leukoencephalopathy		27		
TAD boundary	Disrupting chromatin looping leading to enhancer loss or adoption	WNT6/IHH/EPHA4/PAX3 locus	Limb phenotypes		28	Intergenic variant	

This is not intended as an exhaustive list. Reference DOIs: 1. 10.1182/blood-2018-07-863951. 2. 10.1002/humu.23212. 3. 10.1016/j.ajhg.2018.07.002. 4. 10.1016/j.ebiom.2016.04.005. 5. 10.1038/s41467-019-10717-9. 6. 10.1074/jbc.M005199200. 7. 10.1038/s41431-020-0676-y. 8. 10.1002/ajmg.a.36703. 9. 10.1016/j.ajhg.2021.04.025. 10. 10.1172/JCI119555. 11. 10.2337/db07-1657. 12. 10.3390/genes11101180. 13. 10.1016/j.ajhg.2018.10.023. 14. 10.1136/jmedgenet-2020-107626. 15. 10.1016/j.ajhg.2019.07.013. 16. 10.1182/blood-2018-03-838235. 17. 10.1136/jmedgenet-2018-105836. 18. 10.1086/505361. 19. 10.1038/s10038-018-0502-3. 20. 10.1038/ng.230. 21. 10.1038/ng.2826. 22. 10.1038/ng.329. 23. 10.1038/s41467-021-23980-6. 24. 10.1073/pnas.1401464112. 25. 10.1002/ana.24826. 26. 10.1038/ncomms9718. 27. 10.1038/ng.3661. 28. 10.1016/j.cell.2015.04.004

in the Developmental Disorders Gene to Phenotype (DDG2P) database [34] (downloaded on 02/04/2019). Variants were filtered to only those classified as high-confidence by LOFTEE [35], with allele frequency (AF) <0.5% across the GEL rare disease cohort and/or in gnomAD v2.1.1 [35], and with >25% but <75% of reads containing the variant.

Each DDG2P biallelic LoF gene was annotated with a minimal set of non-coding regulatory regions comprising all intronic regions, the 5′UTR and 3′UTRs, and a core promoter region comprising the first 200 bps directly upstream of the transcription start site. Regions were identified using the MANE Select v0.9 transcript where available [32], and otherwise, the canonical transcript as defined by UCSC [36]. Variants transmitted by the alternative parent to the pLoF variant were identified in the non-coding regions of the same gene and filtered to only those with filtering allele frequency [37] <0.5% and no observed homozygotes in gnomAD v3.1.

Results

Non-coding region variants are under-ascertained in clinical variant databases

Non-coding regions are not regularly captured in clinical sequencing pipelines, where they are most often excluded from capture regions, or removed during bioinformatic processing of the data. Consequently, even variants within non-coding regions known to regulate established disease genes are under-reported. In the ClinVar database [31], only 1 in 294 (0.34%) high-confidence pathogenic variants are in UTRs or immediately upstream regions (within 2 kb; Fig. 2). This is despite UTRs having approximately the same genomic footprint as protein-coding regions and important regulatory roles [28]. Whilst this is in part due to the lower likelihood of any single non-coding region variant being pathogenic, it also reflects under-ascertainment.

Regulatory variants are also more likely to be categorised as variants of uncertain significance (VUS), with 63.4% of all UTR variants in ClinVar categorised as VUS, compared to 44.2% of coding sequence variants (Fig. 2b), highlighting the need for clearer guidelines for interpretation and strategies for functional validation.

Defining and filtering candidate non-coding regions

The vast majority of the thousands of non-coding region variants identified in each individual will have very little or no effect. For example, whilst 43% of all assessed common variants (minor allele frequency ≥0.01) are significantly associated with expression of at least one gene in at least one human tissue, 78% of these show <2-fold changes [38]. Furthermore, many of these variants may simply tag the effect of true causal variants and themselves have no effect on gene expression. To avoid both a huge burden of interpretation and many variants being reported as variants of uncertain significance (VUS), it is important to only clinically interpret variants that (1) fall into regulatory elements that have well-established or

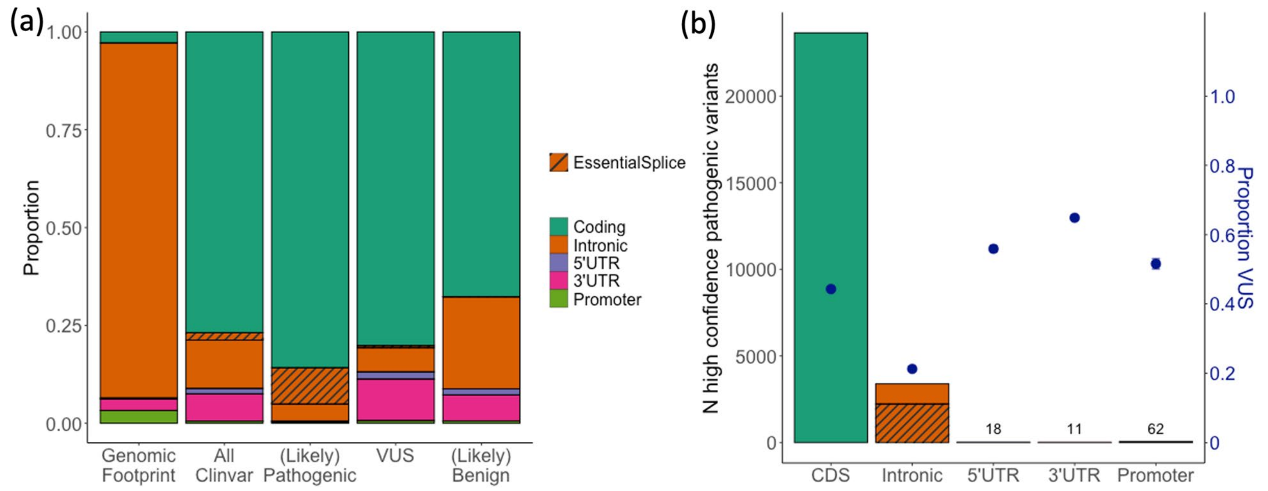


Fig. 2 Non-coding region variants are under-ascertained in ClinVar and are more likely to be classified as variants of uncertain significance (VUS) when compared to protein-coding variants. **a** The proportion of the genomic footprint of MANE transcripts that fall into each of five region categories and the proportion of variants in ClinVar (all, likely pathogenic or pathogenic, likely benign or benign, and VUS) within those regions. **b** The number of high-confidence pathogenic variants in ClinVar (see '2') that fall into each of the five region categories plotted as bars, with the proportion of variants in each region classified as VUS as blue points

functionally validated links to target genes, and (2) those genes have documented associations to the phenotype of interest (i.e. at definitive, strong or moderate level using the ClinGen classification approach [39], or green for the phenotype of interest in PanelApp [40]). This should be used as a filtering approach rather than evidence towards pathogenicity. We therefore only recommend the use of ACMG/AMP rule PP4 for non-coding region variants where the gene is the only, or one of very few genes associated with a discriminative set of phenotypic features, in accordance with existing guidance [30]. For variants within candidate CREs or non-coding genes without proven gene-disease validity, we recommend that they are treated as research variants, and not interpreted or reported, until meaningful functional experiments prove a direct effect of the CRE on the target gene and/or a role for the non-coding gene in disease is established.

We note that identifying CREs and linking them to specific genes is a very active area of research [13, 41]. However, based on current knowledge, we recommend that regions of interest within which variants should be interpreted should be defined using the following parameters. Possible sources of data to support these definitions are listed in Table 2.

- *Introns and UTRs*: The definition of intronic and UTR regions is transcript dependent. In general, these should be defined using well-validated clinically relevant transcripts. Even if a well-validated transcript exists for the coding regions of a gene of interest, the UTRs may not be well defined. We therefore recommend using transcripts defined by the MANE project which has integrated multiple diverse datasets to accurately define these elements [32]. Each gene has a single 'Select' transcript (98% of genes), and some genes have additional 'Plus Clinical' transcripts.
- *Promoters*: Promoters are marked by a region of open chromatin surrounding the transcription start site (TSS) that is described as a nucleosome-free region. Whilst the exact size and composition of the promoter region may be cell type and temporally specific, for many genes a minimal core region is visible in epigenetic [42] data across most cell types [43]. To define the promoter region, we (1) first need to determine the location of the TSS, before (2) defining the region around the TSS corresponding to the promoter. (1) The TSS should be defined in a relevant tissue or cell type as the most 5′ position of the 5′UTR of a MANE Select or other well supported transcript, or using a peak defined by CAGE (Cap Analysis Gene Expression). (2) Following this, the promoter should be defined as the region of open chromatin surrounding the TSS (e.g. using ATAC-seq or DNase-seq data). Ideally, this should be based on epigenetic data from a relevant tissue or cell type; however, in the absence of this, a minimal promoter should be defined using the consensus data across all available cell types. These data are summarised by the ENCODE Candidate Regulatory Elements [44] and Ensembl Regulatory Build datasets [45, 46]. We note that other data types can support the definition of the promoter region, including histone modifications that mark active promoter elements (H3K27Ac and H3K4Me3), overlapping transcription factor binding sites identified by ChIP-seq, and poised RNA Pol II identified by ChIP-seq. Where the existence of a promoter is not supported by epigenetic data, potentially because it only acts in a cell type or at a developmental time-point not represented by published data, and hence a region of open chromatin is not known to exist around the TSS, a minimum promoter region can be defined as the 250 bps immediately up- and downstream of the TSS.
- *CREs*: There may be multiple 'candidate' CREs in the region surrounding a gene of interest. As noted above, these must have a known or functionally validated link to the gene of interest for variants within them to be interpreted clinically. We therefore outline a two-step process to identify CREs. Firstly, candidate CREs can be defined as regions of open chromatin (defined by ATAC-seq or DNase-seq), marked by histone modifications that mark active enhancer elements (H3K27Ac and H3K4Me1), and/or with evidence of multiple overlapping transcription factor binding sites identified by ChIP-seq. Candidate CREs should then be filtered to only those with experimental evidence of a link to the gene of interest, for example through chromatin interaction data (from promoter capture Hi-C), functional perturbation showing an effect on gene expression, or the presence of one or more expression quantitative trait loci (eQTLs) for the gene.

It merits repeating that enhancer and promoter usage can vary across tissues and also temporally, for example throughout development [47]. It is therefore important that when defining both promoters and CREs the above datasets should be derived from a relevant tissue or cell type for the phenotype of interest and where possible, from a relevant developmental time-point.

We appreciate that the definition of candidate regulatory elements using the above guidance will vary depending on data availability and access. For equitable and consistent variant interpretation, these 'interpretable'

Table 2 Data types used for identification of candidate non-coding regions

Evidence	Description	Region(s)	Possible source	Extra considerations[b]
ATAC-seq or DNase	Flags regions of open chromatin	Promoter / CRE	ENCODE; ROADMAP epigenomics	
Promoter capture Hi-C	Links enhancer regions to promoters of target genes	CRE	Published datasets; 3D genome browser[a]	Transient interactions that may not be needed for enhancer function
Hi-C	Define topologically associated domains (TADs)	TAD boundaries	Published datasets; 3D genome browser[a]	
Cap analysis gene expression (CAGE)	Marks the 5′ cap of mRNA	Promoter / 5′UTR	FANTOM5	Alternative transcripts
CTCF ChIP-seq	Identifies regions bound by the insulator protein CTCF	TAD boundaries	ENCODE; ROADMAP epigenomics	
Transcription factor ChIP-seq	Identifies regions bound by specific transcription factors	Promoter / CRE	ENCODE; ROADMAP epigenomics	
H3K4Me	Histone modification found near enhancers	CRE	ENCODE; ROADMAP epigenomics	
H3K4Me3	Histone modification found near promoters	Promoter	ENCODE; ROADMAP epigenomics	
H3K27Ac	Histone modification found at active regulatory elements	Promoter / CRE	ENCODE; ROADMAP epigenomics	
Expression quantitative trait loci (eQTLs)	Identifies variants that are associated with changes in gene expression	Promoter / CRE	GTEx; eqtlgen.org	
Experimental perturbation	Demonstrates an impact of altering/deleting all or part an element on gene expression	Promoter / CRE	Published data	Assays may not be representative of endogenous situation
Northern blot, RT-qPCR, RNA sequencing and microarrays	Detection of miRNAs	miRNAs	miRBase; published datasets	
Multiple approaches	Experimentally validated miRNA–target interactions	miRNA targets	miRTarBase; published datasets	
Bisulphite sequencing	Detects methylated DNA and allows identification of differentially methylated regions	Promoter / upstream gene regions	ENCODE; ROADMAP epigenomics	
RNA Pol II ChIP-seq	Detection of poised polymerase II	Promoter	ENCODE; ROADMAP epigenomics	

[a] https://doi.org/10.1186/s13059-018-1519-9

[b] Tissue specificity and temporal specificity (e.g. specific to a developmental time-point) should be considerations for all

regions need to be standardised for specific genes and diseases, and regularly updated as new data and knowledge become available. We call on the community to generate and make openly accessible relevant datasets, and to work together to produce resources that define interpretable promoters and CREs across all known disease genes.

Predicting the impact of variants identified in candidate non-coding regions

Predicting the effect of any individual variant may not be straightforward. In Table 1, we have listed many of the mechanisms through which non-coding region variants have previously been shown to cause disease. This list is not exhaustive, and new mechanisms will be identified as more variants are identified and comprehensively studied. Often, the only way to fully determine a variant's impact will be through functional studies (see section '20' below). Where in silico tools exist to predict the effect of certain classes of non-coding region variants, we have noted examples of these in Table 1.

General considerations

Variant types covered by this guidance

This guidance is intended to cover short sequence variants (SNVs and insertions and deletions (indels) <50 bps in size) to mirror the original ACMG/AMP guidelines. We do not explicitly consider larger copy number (CNV) and structural variants (SVs), which are discussed in separate existing guidelines [48]. We note, however, that multiple principles of our recommendations will apply to CNVs and SVs that do not overlap protein-coding sequence. The change in disease risk associated with variants identified through genome-wide association studies are very small and outside the scope of these guidelines.

We intend these recommendations to cover all variants identified outside of protein-coding exons, including UTRs, intronic sequence, promoters and more distal regulatory elements. We note that canonical splice site variants (altering the GT in the first two bases of the intron or AG in the last two bases) are generally considered to be loss-of-function and are well covered by existing guidelines [49]. We caution, however, that this is not always the case and we will discuss specific scenarios where exceptions may apply.

Terminology

Referring to variants as either 'coding' or 'non-coding' based on where they are in genomic sequence can be an unhelpful distinction. It is much more informative to instead refer to the predicted (or possible) downstream effect of the variant, which may be to alter protein sequence, and/or to change the abundance of expressed protein. For example, both coding loss-of-function (LoF) variants and regulatory variants in non-coding regions that abolish protein expression can have equivalent downstream effects. Conversely, a UTR variant that extends the coding sequence at either the N- or C-terminus could exert a pathogenic impact through changes to the protein sequence rather than changing protein levels, and hence is best described primarily by its mechanism of pathogenesis.

Defining relevant tissues and cell types

Both in our above discussion of defining regulatory regions and below when we consider functional studies, we mention the need for assays to be performed in 'relevant' tissues or cell types. We deliberately do not give solid guidance on how to decide what constitutes a 'relevant' cell type or tissue, or an appropriate developmental time-point, as these should be defined on a gene and/or disease-specific basis. We call on the community to work together to determine and standardise these.

Applying the ACMG/AMP guidelines to non-coding region variants

Whilst the primary consideration for the ACMG/AMP guidelines was interpretation of variants in protein-coding regions, they were intended as all-encompassing and can be applied to interpretation of variants genome wide. These recommendations are therefore designed to sit alongside this existing guidance, noting adaptations to these rules rather than replacements.

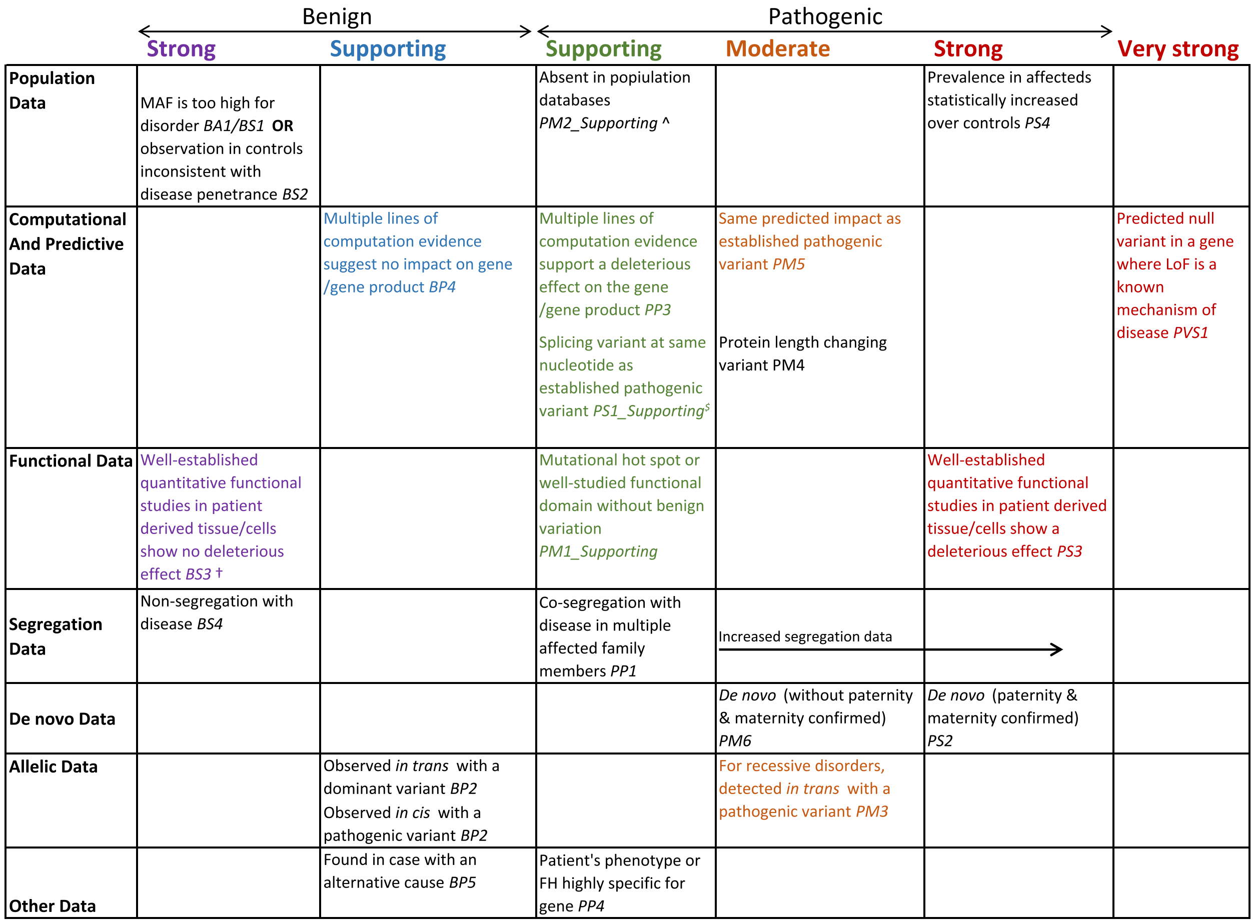
Many of the rules from Richards et al. [30] can be directly applied to variants in non-coding regions, without requiring extra considerations (Fig. 3; Additional file 1: Table S3). These include the use of frequency information (BA1, BS1, BS2 and PM2), upweighting of confirmed de novo variants (PM6 and PS2), and incorporation of co-segregation evidence (PP1 and BS4). Conversely, some rules are not applicable to non-coding region variants, for example those that refer specifically to missense variants and are not further adapted here (PP2 and BP1; Additional file 1: Table S3).

For nine of the specific ACMG/AMP rules, we recommend some modification in the way in which they are used (for example by reducing the strength to reflect lower certainty) or note extra considerations for when they are activated (Fig. 3; Additional file 1: Table S3). Each of these is discussed in detail below.

PVS1

In the ACMG/AMP guidelines, PVS1 is defined as 'predicted null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where loss-of-function (LoF) is a known mechanism of disease' [30]. Given the extreme caution that is required when applying this criterion given its 'very strong' weighting, the Clinical Genome Resource (ClinGen) subsequently released further guidance on its application, including recommendations to decrease the strength applied to this rule under situations where the confidence in a variant being true or complete LoF is reduced [49]. Neither the original ACMG/AMP guidelines nor the updated guidance refer to non-coding region variants other than canonical splice site variants, or splicing defects that would delete one or more exons, as most variation in non-coding regions cannot be confidently predicted to lead to a null effect in the absence of experimental data. Given this, we do not recommend the use of PVS1 for variant types not covered by existing guidance. We also caution that PVS1 should not be applied to canonical splice sites within UTRs where the downstream impact is not clearly loss-of-function. Additionally, PVS1 should not be used in combination with in silico prediction tools (rule PP3), as specified for canonical splice variants in previous guidance [49].

Although PVS1 is often applied to canonical splice site variants, we caution that these do not always have a null effect; factors such as alternative splicing may mediate the pathogenic impact of these variants despite the variant causing changes to the spliced transcript [51]. Moreover, variants which disrupt splicing can have multiple impacts [52] and/or only partial effects (see guidance below), with some aberrantly spliced transcripts resulting in a loss-of-function and others



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*		Absent in popiulation databases *PM2_Supporting ^*		Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computation evidence suggest no impact on gene /gene product *BP4*	Multiple lines of computation evidence support a deleterious effect on the gene /gene product *PP3* Splicing variant at same nucleotide as established pathogenic variant *PS1_Supporting$*	Same predicted impact as established pathogenic variant *PM5* Protein length changing variant PM4		Predicted null variant in a gene where LoF is a known mechanism of disease *PVS1*
Functional Data	Well-established quantitative functional studies in patient derived tissue/cells show no deleterious effect *BS3* †		Mutational hot spot or well-studied functional domain without benign variation *PM1_Supporting*		Well-established quantitative functional studies in patient derived tissue/cells show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed *in trans* with a dominant variant *BP2* Observed *in cis* with a pathogenic variant *BP2*		For recessive disorders, detected *in trans* with a pathogenic variant *PM3*		
Other Data		Found in case with an alternative cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Fig. 3 ACMG evidence framework for non-coding region variants. An adapted version of the figure from Richards et al. [30] (permission granted). Rules that require no extra guidance for non-coding region variants are written in black, with those requiring extra considerations or adaptation in colour. †Should not be applied if the assay only assessed one of multiple possible mechanisms. ^Reduced to supporting following guidance from ClinGen SVI [50]. $Variant must have at least as great an impact predicted by in silico tools

showing no discernible change or creating a functional transcript. In such situations, these consequences would receive different interpretations, and the relative dosage of each transcript is an important consideration.

PM1

Rule PM1, 'Located in a mutational hot spot and/or critical and well-established functional domain (e.g. active site of an enzyme) without benign variation', was designed initially to capture variants within important protein domains that are critical to function. It is important to note that sequence constraint and variant effect in non-coding regions has been shown to likely be base-specific rather than consistent across larger regions [17, 53]. It is therefore not appropriate to activate PM1 for a variant within a region (e.g. a UTR, or cis-regulatory element) just because multiple previous variants within that region have been shown to be pathogenic. There are, however, occasions where activating PM1 may be appropriate; for example, when a variant disrupts the binding motif of a transcription factor, perturbation of which has repeatedly been shown to be pathogenic, or where multiple known pathogenic variants are clustered within the same well-defined enhancer region, such as sub-regions of the ZPA regulatory sequence (ZRS) that controls expression of *SHH* [54]. In these instances, however, we would recommend always lowering the strength of PM1 to supporting.

In the given example of disrupting a known transcription factor binding site, if this is predicted using an in silico tool, then this should be used to inform PP3 and should not also be used for PM1.

PS1

In the ACMG/AMP guidelines, PS1 is used specifically for missense variants, when a different nucleotide change results in the same amino acid change as an established pathogenic variant. Subsequent guidelines from the UK ACGS stated that PS1 could also be used ‘at a supporting level for splicing variants where a different nucleotide substitution has been classified as (likely) pathogenic and the variant being assessed is predicted by in silico tools to have a similar or greater deleterious impact on the mRNA/protein function’ [55]. Whilst we support this use of PS1, we also caution that different base changes can have different effects on activation of alternative splice sites and hence could have different impacts.

There are other specific occasions where activation of PS1 may also be appropriate, for example, uORF stop-lost variants where disruption of the same stop codon has previously been shown to be pathogenic [17].

PM5

Similarly to PS1, PM5 is also described in the ACMG/AMP guidelines as specific to missense variants, although PM5 refers to variants disrupting the same amino acid residue, but leading to a different alternate residue. Here, we recommend using this evidence code to capture non-coding region variants that are predicted to have exactly the same impact, on the same gene, as established pathogenic variants, but themselves may not have been described before. Examples of this include upstream start codon creating variants that result in out-of-frame overlapping open reading frames (see *NF1* example curation below), and near completely overlapping deletions of the same transcription factor binding site or promoter region.

We further caution on our use of the phrase ‘predicted to have exactly the same impact’. The gene specificity of regulatory elements means that identifying a variant with exactly the same impact is often not possible. For example, a variant that creates an upstream start codon in a 5′UTR may need to be created into the same context and in the same frame with respect to the coding sequence to have an identical effect, and even then, differing distances to the coding sequence may have an impact. Similarly, two variants could disrupt binding of one transcription factor, but result in opposite effects on the target gene, e.g. if one variant creates a novel binding site for a paralogous transcription factor.

In general, PS1 should be used where the same base, or residue, as the previously pathogenic change is impacted, and PM5 should be used for other variants with the same predicted effect, but that are not at the same specific base/residue. If the effect of the variant is predicted using an in silico tool (e.g. splice region variants), then this information should inform PP3 and not PM5.

PM3

The PM3 rule can be used for recessive disorders when a variant is detected in trans with a known pathogenic variant. Non-coding region variants, in particular deeply intronic variants that impact splicing, have been identified in trans with coding variants [52]. Given the increased search area for possible in trans variants when including non-coding regions (particularly intronic regions), we sought to determine the frequency at which we would expect to observe one or more rare variants in trans using rare disease trios from the GEL 100,000 Genomics Project dataset.

We identified 2016 undiagnosed trio probands with 2714 single, rare (AF<0.5%) heterozygous pLoF variants in 794 genes in DDG2P annotated as biallelic (i.e. recessive) with a LoF mechanism. For each sample-pLoF pair, we searched for rare variants in non-coding regions of the same gene that were inherited from the alternative parent to the pLoF variant. These non-coding regions comprised intronic, 5′UTR and 3′UTR regions, and a core promoter region (200 bps immediately upstream). In total, 1027 sample-pLoF pairs (37.8%) had at least one regulatory variant in trans, with a mean of 0.89 (range 0–22) identified per sample-pLoF pair (Additional file 2: Fig S1). As expected, the vast majority (93.9%) of in trans variants mapped to intronic regions; however, only seven of these passed a permissive SpliceAI threshold of 0.2 [56]. If we filter the intronic variants using this SpliceAI threshold, only 2.0% of sample-pLoF pairs had a candidate variant in trans.

Given the low numbers of in trans variants found in this analysis, we believe it is appropriate to apply PM3 as per existing guidelines for coding variants [57]. However, it is especially important to use strict allele frequency cut-offs to limit consideration to only suitably rare variants [37] and only consider variants impacting genes that are a credible cause of an individual’s phenotype. Of note, we identified 22 variants in trans with one pLoF variant in the *WWOX* gene, which is an extremely large gene spanning >1.1 Mbs. We also identified 16 in trans variants in an individual self-reported as ‘Black or Black British: African’. These examples highlight extra care required when considering genes with particularly large intronic regions or where reference datasets do not adequately match an individual’s genetic ancestry. In these instances, it may be appropriate to lower the strength of PM3. One possible approach is to calculate the likelihood of observing a similar variant (e.g. with an equivalent or greater in silico score) in the gene of interest, given the distribution of all scores across the gene. This approach would adjust for both region size and localised mutability. It should also be noted that non-coding region variants and hypomorphic coding variants that appear to be tolerated as homozygotes can be pathogenic when found in trans with a null effect coding variant [58].

PS3/BS3

Functional evidence is extremely important to support either a pathogenic or a benign role for non-coding region variants. Bespoke assays are often required depending on the variant context and its predicted effect (e.g. on splicing, transcription, translation, or chromatin looping). Functional assays should be designed and assessed following existing guidance [59]. Below we discuss in more detail considerations for different categories of functional assays commonly used for non-coding region variants.

- *RNA sequencing*: RNA-seq and/or targeted approaches enable the assessment of a number of characteristics which may be indicative of the functional impact that a variant has on normal gene expression. These include the characterisation and quantification of aberrant transcript isoforms, differential gene expression and allelic expression imbalances [60]. Detection of aberrant splicing isoforms is well-described in the literature, with numerous examples of Mendelian disease causation and often multiple known pathogenic variants per gene. This enables functional assays to be designed that allow application of PS3 at moderate/strong weighting [59]. In contrast, it may be difficult to establish a suitable number of controls for functional assays assessing allelic expression imbalances and therefore more difficult to achieve higher levels of support for PS3. Some genomic variants will cause binary changes in the measured characteristics, for example, the abolition of canonical splice sites, whereas others will cause changes in the relative ratio of normal:aberrant gene expression profiles (see guidance below on partial effects). The discovery of aberrant gene expression through RNA-seq requires comparison to a control cohort (e.g. GTEx [38]) and usually also to individuals from the same sequencing and analysis process. We recommend that software used to detect aberrant splicing events has been benchmarked specifically for their discovery in the context of rare disease [61]. The technical appropriateness of biosamples for the discovery of aberrant events in the gene of interest should be considered (i.e. is the gene normally expressed in this tissue?), including when using bespoke control sets. Strong weighting of PS3 for identified aberrant splicing should only be used when (1) expression profiles in the biosample used match those of the primary disease tissue of the candidate genes [62], and (2) the sequencing data generated is appropriate for the detection of aberrant splicing events (e.g. exceeds the 95% confidence interval recommendations from MRSD (minimum required sequencing depth) using appropriate parameters for the laboratory and bioinformatics approaches applied [63]). In addition, we recommend that BS3 is only used in the absence of aberrant splicing events when there is evidence that both alleles are being expressed (e.g. data supporting heterozygous alt/ref alleles present in roughly equal quantities) and appropriate sequencing coverage has been achieved [62, 63]. Even in such scenarios, caution must be taken when using BS3 due to the known cell-specific impacts of some splicing variants [64] and, in some cases, complex alternative splicing dynamics [65]; we recommend that BS3 is not used in these situations, unless an appropriate system for functional assessment has been used.
- *MAVE approaches*: multiplexed assays of variant effects (MAVEs) that classify variants as functionally normal or functionally abnormal have great potential to aid the interpretation of both protein-coding and non-coding region variants. Whilst the majority of studies to date have focused on protein-coding regions, smaller studies have profiled portions of UTRs [66], and others have used deletion tiling to identify and study enhancers [67, 68], offering insights into the regulatory code. In general, use of MAVE data in variant interpretation should follow existing guidance [69]. Extra care should be taken when interpreting MAVE results for non-coding region variants, however, for the following reasons: (1) assays that only test a short section of a regulatory element for function do not account for regulation mechanisms that rely on neighbouring DNA, such as the formation of secondary structure, binding of co-factors, or presence of internal ribosome entry sites (IRES); (2) if only a single output is assessed, this may not be relevant for the mechanism of the variant of interest (i.e. when an RNA-seq read-out is used, but the variant is predicted to impact translation); (3) experiments may be performed in a cell type or model system where the applicability to the disease of interest is unclear. One clear limitation to current use of MAVEs is the focus of each experiment on a single gene, or even only a single exon within a gene. Collaboration through initiatives such as the Atlas of Variant Effects (AVE; www.varianteffect.org) is essential to achieving comprehensive coverage across both genes and regulatory regions.
- *Chromatin interaction assays*: Chromosome conformation capture (3C) approaches can be used to identify regions of the genome that are co-restricted following chemical cross-linkage, including CREs and the promoters of their target genes. As we advise

above, this information should be used to inform which variants are interpreted using these guidelines rather than being used as functional evidence to inform PS3. An exception to this is where chromatin interaction is shown to be disrupted in individuals with a variant in a candidate enhancer region (which passes the above inclusion criteria) when compared to appropriate controls, which could be used to activate PS3.

– *Reporter gene assays, e.g. luciferase assays*: can be used to assess the impact of variants in promoter and regulatory regions of genes. Using the bioluminescent properties of a gene inserted into a plasmid along with a candidate regulatory region, e.g. downstream of a promoter region, one can enable assessment of the relative quantitative impact of variants in the candidate regulatory region on levels of protein production [19]. Significant disruption between reporter assays containing variant and wild-type regulatory sequences can be used to activate PS3, although with the very important caveats that this is an artificial assay system and must be appropriately validated [59].

Many non-coding region variants may only have a *partial effect*; for example, splicing variants that affect a sub-set of transcripts, or 5′UTR variants that only partially reduce downstream coding sequence translation. For splicing variants, assays can be quantitative as described above; however, for many other assays, quantifying the precise level of an effect is difficult. Even when an effect can be quantified, whether a variant with a partial effect can cause disease is very gene dependent; for some genes, only a partial reduction in functional protein can be severely detrimental, but others will tolerate partial dosage changes. Benchmarking assays across the full range of effects will therefore be important to determine gene-specific thresholds for activation of PS3.

When considering BS3, if a functional assay has not shown an effect on the gene product or its expression, care should be taken when the assay was not performed in a relevant tissue or cell type as regulation and transcript usage can be very tissue specific. In addition, BS3 should not be applied if an assay only assesses a single output, but there are multiple possible underlying mechanisms (e.g. for 5′UTR variants).

Regulatory variants can act to either increase or decrease expression of a target gene. On some occasions, a functionally tested variant may cause a gain-of-function when the known mechanism for the gene is loss-of-function. BS3 could be applied when this direction of effect is inconsistent with the known gene mechanism (although we note that many dosage-sensitive genes may be both haploinsufficient and triplo-sensitive [70]).

PP3/BP4

The majority of widely used computational tools that predict variant deleteriousness were designed to interpret the impact of coding missense variants. These tools cannot be applied to variants in non-coding regions. There are, however, multiple tools that predict the likely impact of variants on splicing, and a few that make predictions along the complete length of a pre-mRNA transcript [71]. Comparisons between available splicing tools have been published recently [52], showing that some tools (e.g. SpliceAI [56]) perform well to prioritise variants with functional evidence of aberrant splicing. These tools are not further reviewed here. We also do not recommend specific tools and thresholds as such guidance is likely to be quickly outdated given the ongoing rapid emergence of new predictive splicing tools showing iterative improvements. We instead encourage use of evidence-based thresholds and highlight recent papers benchmarking the performance of newly developed tools [49, 72, 73]. Additionally, we caution that many tools are trained on existing canonical splice junctions, or known pathogenic/benign variants that are enriched near exon-intron junctions and hence may perform less well for deep intronic splice variants. Conversely, we note that the importance of genomic variation at the canonical donor +5 site has been well established [74], and variants at this position should be treated with a higher prior probability of pathogenicity than other proximal positions.

In silico tools that can be used to predict the deleteriousness of other categories of non-coding region variants have also been recently reviewed (see Table 3 in Rajano et al. [75]). In addition to those mentioned, we note recent tools designed specifically for rare disease: NCBoost [76], ReMM (Genomiser) [77] and GREEN-DB [78].

For genome-wide machine-learning tools that rely on a set of true positive pathogenic variants for training, we caution that accurate datasets for this purpose covering non-coding regions are currently very limited. These data are biassed towards certain subsets of variants, including those very close to the coding sequence and only within a small number of genes (i.e. those causing single-gene disorders). Indeed, a recent paper describing NCBoost demonstrated this regional bias [76]. How well these tools predict the pathogenicity of the full range of non-coding region variants is currently unknown. We therefore recommend extreme caution against over-interpreting the output of any genome-wide predictor.

Whilst a limited selection of in silico scores for non-coding region variants are accessible through widely used annotation tools (e.g. Ensembl VEP), the vast majority must be queried individually and many are only currently available as large file downloads of pre-computed scores or through running software/scripts (Additional file 1: Table S4). This presents a barrier to the use of many of these tools. Of note, the GREEN-VARAN [78] tool returns scores from a group of seven in silico algorithms, although it is not currently available as a web-tool.

Example variant curations

NF1:c.-160C>A hypothetically identified variant in an individual with neurofibromatosis type 1

The NF1:c.-160C>A (ENST00000356175.7; chr17:31095150:C>A GRCh38) variant creates an upstream start codon (uAUG) in the 5′UTR of the NF1 gene. It has not been reported in ClinVar or the literature. uAUG-creating variants in *NF1* have previously been shown to cause neurofibromatosis [17, 79]. This variant has the same predicted impact as these previously identified pathogenic variants: it is created into a strong Kozak consensus, and translation from this uAUG would create an upstream open reading frame (uORF) that overlaps the coding sequence out-of-frame with the canonical start site. We would therefore activate PM5. This variant is also absent from gnomAD (PM2_Supporting), and it would be appropriate to activate PP4 if the variant was identified in an individual with classic neurofibromatosis type 1 (NF1) features given the specificity of the NF1 phenotype for the *NF1* gene. If this variant occurred either de novo (PM6/PS2), or if there was evidence of segregation with disease at a moderate level (PP1_Moderate), it would reach a classification of Likely Pathogenic. Similarly, a functional assay demonstrating a reduction of translation in a validated cell line model, or a reduction in protein levels in an appropriate tissue sample would enable activation of PS3 resulting in a Likely Pathogenic or Pathogenic classification.

CFTR c.3874-4522 A>G identified in a patient with cystic fibrosis

The proband received a late diagnosis of cystic fibrosis (16–20 years old). Previous genetic testing uncovered the common c.1521_1523delCTT (p.Phe508del; chr7:117559590:ATCT>A GRCh38) pathogenic variant in CFTR in a heterozygous state and was proven in trans to c.3874-4522A>G (chr7:117648320:A>G; PM3). c.3874-4522A>G is absent from gnomAD (PM2_Supporting) and at the time of analysis has previously been reported in a single case of cystic fibrosis with no additional functional evidence. There is a strong phenotype-genotype correlation (PP4). MaxEntScan supported the activation of a cryptic splicing site but a number of in silico splicing tools did not support that the variant would impact splicing (SpliceAI, TraP); therefore, neither PP3 nor BP4 were applied. Functional assessment of splicing impact was performed using RNA extracted from whole-cell blood for the proband, showing abnormal splicing of *CFTR* with the introduction of a 125 bp cryptic exon containing a stop codon. PS3_Strong was therefore applied as we were using an established method for splicing variant investigation, with appropriate controls. As we are using PS3, we would have had to exclude PP3 in our final classification had this code been used. The variant was classified as Likely Pathogenic (PS3_Strong, PM3, PM2_Supporting, PP4).

Emergence of new evidence

After initial classification, additional evidence became available from the literature. This included functional evidence from minigenes and patient samples [80]. As the assays performed in the initial classification were appropriate for use of PS3_Strong, there was no alteration due to this evidence. Additional families with cystic fibrosis from multiple ethnicities have also been reported to carry the c.3874-4522A>G variant, and in at least 4 of these families, symptomatic individuals are proven to carry c.3874-4522A>G in trans to a proven pathogenic allele [81]. This evidence allows us to upgrade PM3 to VeryStrong, and our final classification to Pathogenic (PS3_Strong, PM3_VeryStrong, PM2_Supporting, PP4).

PAX6 distal enhancer variant hypothetically identified in a patient with aniridia

The chr11:31664397 C>A (GRCh38) variant is located in a candidate CRE downstream of *PAX6* and intronic in *ELP4*. The region is highly conserved and the element is identified as a 'distal enhancer' by the ENCODE regulatory build (visualised on the UCSC genome browser [36]). Functional experiments show that deletion of the element disrupts maintenance of *PAX6* expression [82]. Multiple in silico scores support a deleterious role (CADD = 17.4; ReMM = 0.985; FATHMM_MKL = 0.993; PP3), and the variant is absent from gnomAD (PM2_Supporting). In silico modelling suggests the variant disrupts a PAX6 binding site which was validated through disruption of reporter expression in the lens (PS3_Moderate). The variant was also identified de novo via trio analysis (PS2) in a patient with a highly specific phenotype (PP4). Taken together, these data result in a Likely Pathogenic classification (PS2, PS3_Moderate, PM2_Supporting, PP3, PP4). We note that the distal enhancer variant for *PAX6* is also intronic within *ELP4* and can be scored for predicted impact on splicing (e.g. SpliceAI = 0.00); we caution that predictive tools should be used in relevant context, and this SpliceAI score cannot be used to support the impact of chr11:31664397 C>A on *PAX6* expression.

Discussion

Here, we have outlined considerations for adaptation of the ACMG/AMP guidelines for variant interpretation to variants identified outside of protein-coding regions of the genome. These recommendations have been carefully reviewed and refined by an expert panel and extensively tested by clinical variant scientists.

It is clear that our knowledge of the impact of non-coding region variants in rare disease is a fast-evolving field, which makes it complex to provide comprehensive guidance that will fit every possible scenario. We have therefore tried to provide general guidance that can be applied to most variant types and have included specific examples of how to apply this guidance in practice. We hope that these recommendations will enable increased interpretation of non-coding region variants and catalyse the discovery of additional examples of disease-causing variant types, which will in turn inform further revisions to this guidance. To enable this continued learning and refinement of these guidelines, we encourage the sharing of variant data, including for variants that are not initially classified as pathogenic, for example by submitting classified variants to ClinVar [31] and enabling access to individual-level data and through DECIPHER [46].

During testing of these guidelines, it became clear that one of the largest barriers to widespread adoption is access to the epigenomic data (for example to define candidate regulatory elements) and in silico scores required to interpret non-coding region variants. We are in desperate need of accessible tools that allow better visualisation of these data by individuals not well-versed in bioinformatics, ideally allowing for queries by cell type/tissue, to allow transparent curation and reproducible interpretation of these data. We also need more research aimed at deciphering the full 'regulatory code' and development of in silico prediction tools for non-coding variants not trained on limited pools of known pathogenic variants. It was also clear from feedback that there is substantial appetite for educational webinars and workshops around various aspects of using the guidelines (Supplementary Table 2). This includes training on finding and evaluating functional data which may use assays that are unfamiliar. We are actively engaged in developing these in the hope they will increase usability and adoption.

A substantial barrier to our understanding of the role of non-coding region variants in rare disease to date has been the lack of statistical power when searching for enrichment of variants across and between different regulatory regions and/or variant classes. This is due to multiple factors, including (1) the majority of non-coding region variants having little or no effect, (2) a lack of clarity on how to subdivide the genome for genome-wide scans, (3) potential opposing actions of different variants and (4) a lack of large-scale whole genome sequenced datasets derived from different genetic ancestry groups with linked phenotypic data. The study of large phenotypically characterised WGS cohorts using a systematic and reproducible analysis approach, including informed region-based variant filtering, is needed to establish the overall contribution of non-coding region variants to diagnostic yield and clinical utility [7]. Conversely, much of our success to date in identifying disease-causing non-coding region variants has been using phenotypically highly selected cohorts, where only one or a very small number of genes are suspected as being involved. It is likely that this approach will continue to be successful at identifying new disease-causing variants in non-coding regions.

The full range of mechanisms through which variants in non-coding regions cause, and contribute to the risk of genetic disease remains unknown. It is likely, however, that many regulatory variants have smaller effects than those impacting protein sequence and that these effects may be highly tissue-specific. Whilst for some extremely dosage-sensitive genes, even partial effect variants will cause severe disease (as has been shown for MEF2C [19]); in others, single variants with only a moderate effect will be insufficiently deleterious or may only cause disease in a single tissue or organ. For example, in *PRPF31* disease-causing variants causing significantly reduced expression of a single allele can be incompletely penetrant [83]. The penetrance of these variants can be modified by the relative expression of the other allele, or regulatory variants could themselves modify the penetrance of damaging protein-coding variants [84], and/or cause disease only in combination with other variants. Further research is needed to fully elucidate the frequency and impact of these different mechanisms.

The majority of non-coding region variants would be expected to either decrease or increase the protein product of an impacted gene, whether through affecting transcription, or post-transcriptional regulation mechanisms. It is therefore expected that these variants would primarily impact genes that are dosage sensitive. There are, however, exceptions, such as uAUG-creating variants in 5′UTRs that elongate the coding sequence at the N-terminus, which could negatively impact non-dosage-sensitive transcripts. Furthermore, there may be occasions where a gene does not appear to be constrained against coding loss-of-function variants (so may not be considered haploinsufficient), but where regulatory variants that decrease protein levels could be deleterious. This could be the case, for example, if a compensatory mechanism relies on protein truncating variants that trigger nonsense-mediated decay [85].

Given the uncertainty in predicting the effect and downstream impact of the majority of non-coding region variants, these recommendations are deliberately conservative, often downgrading the strength of evidence applied to individual rules (e.g. PM1). We acknowledge that it can therefore be difficult for newly identified variants to reach a Likely Pathogenic classification. Our example variant curations (Additional file 1: Table S1), however, demonstrate that multiple different sources of evidence can support a (Likely) Pathogenic classification, including segregation (PP1), functional data (PS3), de novo (PS2/PM6) or in trans occurrence (PM3). Our modification of the PM5 missense rule to capture variants with exactly the same predicted impact as an established pathogenic variant also enables variants to be upgraded to Likely Pathogenic (see *TH* promoter variant in Additional file 1: Table S1). We believe that these recommendations strike an appropriate balance between caution and the ability to classify variants as (Likely) Pathogenic, when appropriate.

In these guidelines, we have primarily discussed variants that impact existing regulatory regions; however, there are examples of disease-causing variants that act through creating novel regulatory elements. For example, a recent paper discussed a SNV that created a new promoter leading to dysregulation of genes in the human α-globin locus [86]. We have also not discussed variants in non-coding genes in detail; however, we note examples of identified pathogenic variants in this area of active research (Table 1). The importance of variants in both creating novel regulatory elements and impacting non-coding genes will increasingly be recognised and ensuring that these guidelines continue to be appropriate for these variant types will be an important consideration in future revisions.

It should be recognised that non-coding regions can be very large. Whilst 5′UTRs average only 197 bps, 3′UTRs are on average approximately the same size as the protein-coding sequence (1775 bps for 3′UTRs vs 1745 bps for protein-coding sequence). The average combined size of intronic regions, by contrast, is 34-fold larger than protein-coding sequence at 59,220 bps (region lengths calculated across all MANE Select v0.95 transcripts [32]). Including non-coding regions in the search for likely disease-causing variants therefore dramatically increases the genomic search space. We have not here recommended decreasing the strength of evidence applied to de novo variants for those found in non-coding regions, but more research is required into the relative rates of de novo occurrence across these different regions to determine whether this should be accounted for in future revisions of these guidelines.

Conclusions

As our knowledge of disease-causing mechanisms, and the size of both available phenotype-linked sequencing data and MAVE datasets profiling non-coding regions expand, our ability to fully interpret variants in non-coding regions for a role in both rare and common diseases will continue to increase. As we gain the ability to reliably understand and interpret these variants, it may well be appropriate to rethink our standard ‘coding first’ strategy for genetic testing of many genes and conditions, not only through using WGS, but also by expanding the regions captured by targeted panels to include standardised community-defined regulatory elements, where these remain more appropriate. Given our continuously increasing knowledge, it may also be necessary to revisit and re-interpret variants initially designated as ‘research only’. This in turn will, however, catalyse the return of a definitive genetic diagnosis to ever increasing numbers of individuals with rare diseases.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-022-01073-3.

Additional file 1: Table S1. Variants used to test usability of guidelines. **Table S2.** Questions and responses from the UK Association for Clinical Genomic Science (ACGS) meeting. **Table S3.** Overview of recommendations for all ACMG/AMP rules. **Table S4.** Accessibility of in silico tools for genome-wide variant prioritisation.

Additional file 2: Fig S1. Identifying regulatory variants in trans with pLoF variants in GEL.

Acknowledgements

We would like to thank the following people and groups for their useful input, feedback and guidance: Helen Firth, Adrian Dockery, Jonathan Edgerley, Konstantinos Varvagiannis, the Clinical Genome Resource (ClinGen) Sequence Variant Interpretation (SVI) working group and attendees at the ACGS Annual Scientific Meeting 2021. Figure 1 was created with BioRender.com.

Authors’ contributions

NW, JME and SMH conceived the project and wrote the initial draft recommendations. JWA, DB, SE, DRF, WGN and JCT formed the expert review panel. JL and HCM analysed and interpreted data. RDB, SB, CC, KD, SE, CDF, JMG, JI, NK, AODL, SCR, HLR, ER, MSB, MW, JCW and CFW tested the guidelines and provided feedback. The author(s) read and approved the final manuscript.

Funding

NW is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 220134/Z/20/Z) and by grant funding from The Rosetrees Trust (Grant Number H5R01320). JME is supported by a postdoctoral research fellowship from the Health Education England Genomics Education Programme (HEE GEP). DB and JL are supported by DB’s NIHR Research Professorship (RP-2016-07-011). WGN is supported by the Manchester NIHR Biomedical Research Centre (IS-BRC-1215-20007) and JCT is supported by the Oxford NIHR Biomedical Research Centre. DRF is supported by the MRC University Unit award to the University of Edinburgh for the MRC Human Genetics Unit and by the Wellcome Strategic Award ‘Transforming Genomic Medicine Initiative’ (200990/Z/16/Z). JI is supported by a National Health and Medical Research Council Career Development Fellowship (#1162929). AODL is supported by the National Human Genome Research Institute grant U01HG011755. The ClinGen SVI as well as SMH and

HLR are supported by the National Human Genome Research Institute grant U24HG006834.

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. This research was funded in whole, or in part, by the Wellcome Trust. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Availability of data and materials

Data from the Genomics England 100,000 Genomes Project are accessible to researchers who sign up for access to the Genomics England Research Embassy (https://www.genomicsengland.co.uk/research/academic/join-gecip) but are access-restricted and cannot be shared openly with this article. All other data generated or analysed during this study are included in this published article and its supplementary information files.

Declarations

Ethics approval and consent to participate

This work used the Genomics England 100,000 Genomes Project dataset (East of England - Cambridge South Research Ethics Committee: 14/EE/1112). All patients included in this study provided written consent for genomic analysis and all investigations were conducted in accordance with the tenets of the Declaration of Helsinki.

Consent for publication

Written consent was obtained from the patient with the ***CTFR*** variant to include this in the example variant curation section of the manuscript. No other patient-specific information is included.

Competing interests

AODL is a paid member of the Scientific Advisory Board of Congenica. SMH is a paid employee of Ambry Genetics. All other authors declare that they have no competing interests.

Author details

[1]Division of Evolution, Infection and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicines and Health, University of Manchester, Manchester M13 9PT, UK. [2]Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Manchester M13 9WL, UK. [3]Genomics England, London, UK. [4]Cambridge Genomics Laboratory, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, UK. [5]Agnes Ginges Centre for Molecular Cardiology at Centenary Institute, University of Sydney, Sydney, Australia. [6]School of Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK. [7]Wessex Clinical Genetics Service, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [8]Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, UK. [9]South West Genomic Laboratory Hub, Exeter Genomic Laboratory, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK. [10]South West NHS Genomic Laboratory Hub, Bristol Genetics Laboratory, North Bristol NHS Trust, Bristol, UK. [11]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Edinburgh, UK. [12]Department of Pediatrics, Division of Pediatric Genetic, Medicine, Children's Hospital at Montefiore/Montefiore Medical Center/Albert, Einstein College of Medicine, Bronx, NY, USA. [13]Centre for Population Genomics, Garvan Institute of Medical Research, and UNSW Sydney, Sydney, Australia. [14]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia. [15]Human Genetics Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [16]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [17]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [18]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [19]National Institute for Health Research Oxford Biomedical Research Centre, Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [20]Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [21]Ambry Genetics, Aliso Viejo, CA, USA.

Received: 9 February 2022 Accepted: 16 June 2022
Published online: 19 July 2022

References

1. Rehm HL. Evolving health care through personal genomics. Nat Rev Genet. 2017;18:259–67.
2. Mattick JS, Dinger M, Schonrock N, Cowley M. Whole genome sequencing provides better diagnostic yield and future value than whole exome sequencing. Med J Aust. 2018;209:197–9.
3. Caulfield M, Davies J, Dennys M, Elbahy L, Fowler T, Hill S, et al. The national genomics research and healthcare knowledgebase: figshare; 2019. Available from: https://figshare.com/articles/dataset/GenomicEnglandProtocol_pdf/4530893/5
4. Ellingford JM, Barton S, Bhaskar S, Williams SG, Sergouniotis PI, O'Sullivan J, et al. Whole genome sequencing increases molecular diagnostic yield compared with current diagnostic testing for inherited retinal disease. Ophthalmology. 2016;123:1143–50.
5. Taylor JC, Martin HC, Lise S, Broxholme J, Cazier J-B, Rimmer A, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47:717–26.
6. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014;511:344–7.
7. 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, et al. 100,000 Genomes pilot on rare-disease diagnosis in health care - preliminary report. N Engl J Med. 2021;385:1868–80.
8. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.
9. Williamson I, Kane L, Devenney PS, Flyamer IM, Anderson E, Kilanowski F, et al. Developmentally regulated Shh expression is robust to TAD perturbations. Development. 2019;146(19):dev179523. https://doi.org/10.1242/dev.179523.
10. Akdemir KC, Le VT, Chandran S, Li Y, Verhaak RG, Beroukhim R, et al. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. Nat Genet. 2020;52:294–305.
11. Kai Y, Andricovich J, Zeng Z, Zhu J, Tzatsos A, Peng W. Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. Nat Commun. 2018;9:4221.
12. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018;19:621–37.
13. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat Rev Genet. 2020;21:292–310.
14. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. Genome Biol. 2002;3:REVIEWS0004.
15. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019;5:eaaw1668.
16. Johnston JJ, Williamson KA, Chou CM, Sapp JC, Ansari M, Chapman HM, et al. NAA10 polyadenylation signal variants cause syndromic microphthalmia. J Med Genet. 2019;56:444–52.
17. Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, et al. Characterising the loss-of-function impact of 5'untranslated region variants in 15,708 individuals. Nat Commun. 2020;11:1–12 Nature Publishing Group.
18. Borck G, Zarhrate M, Cluzeau C, Bal E, Bonnefont J-P, Munnich A, et al. Father-to-daughter transmission of Cornelia de Lange syndrome caused by a mutation in the 5' untranslated region of the NIPBL Gene. Hum Mutat. 2006;27:731–5.
19. Wright CF, Quaife NM, Ramos-Hernández L, Danecek P, Ferla MP, Samocha KE, et al. Non-coding region variants upstream of MEF2C cause severe

developmental disorder through three distinct loss-of-function mechanisms. Am J Hum Genet. 2021;108:1083–94.
20. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9. https://doi.org/10.1126/scitranslmed.aal5209.
21. Tomaselli PJ, Rossor AM, Horga A, Jaunmuktane Z, Carr A, Saveri P, et al. Mutations in noncoding regions of GJB1 are a major cause of X-linked CMT. Neurology. 2017;88:1445–53.
22. Bauwens M, Garanto A, Sangermano R, Naessens S, Weisschuh N, De Zaeytijd J, et al. ABCA4-associated disease as a model for missing heritability in autosomal recessive disorders: novel noncoding splice, cis-regulatory, structural, and recurrent hypomorphic variants. Genet Med. 2019;21:1761–71.
23. Sangermano R, Garanto A, Khan M, Runhart EH, Bauwens M, Bax NM, et al. Deep-intronic ABCA4 variants explain missing heritability in Stargardt disease and allow correction of splice defects by antisense oligonucleotides. Genet Med. 2019;21:1751–60.
24. Weisschuh N, Buena-Atienza E, Wissinger B. Splicing mutations in inherited retinal diseases. Prog Retin Eye Res. 2021;80:100874.
25. Wai HA, Lord J, Lyon M, Gunning A, Kelly H, Cibin P, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genet Med. 2020;22:1005–14.
26. Jones SA, Cantsilieris S, Fan H, Cheng Q, Russ BE, Tucker EJ, et al. Rare variants in non-coding regulatory regions of the genome that affect gene expression in systemic lupus erythematosus. Sci Rep. 2019;9:15433.
27. Radziwon A, Arno G, Wheaton DK, McDonagh EM, Baple EL, Webb-Jones K, et al. Single-base substitutions in the CHM promoter as a cause of choroideremia. Hum Mutat. 2017;38:704–15.
28. Chatterjee S, Pal JK. Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. Biol Cell. 2009:251–62. https://doi.org/10.1042/bc20080104.
29. Spielmann M, Mundlos S. Looking beyond the genes: the role of non-coding variants in human disease. Hum Mol Genet. 2016;25:R157–65.
30. 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;17:405–24.
31. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46:D1062–7.
32. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022. https://doi.org/10.1038/s41586-022-04558-8.
33. Odhams C. Tiering (rare disease) - genomics England research environment - genomics England research environment. Available from: https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=38046769. Cited 2022 Jun 14.
34. Thormann A, Halachev M, McLaren W, Moore DJ, Svinti V, Campbell A, et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. Nat Commun. 2019;10:2373.
35. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
36. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
37. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med. 2017;19:1151–8.
38. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
39. Strande NT, Riggs ER, Buchanan AH, Ceyhan-Birsoy O, DiStefano M, Dwight SS, et al. Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the clinical genome resource. Am J Hum Genet. 2017;100:895–906.
40. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51:1560–5.
41. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9.
42. Greally JM. A user's guide to the ambiguous word "epigenetics". Nat Rev Mol Cell Biol. 2018;19:207–8 Springer Science and Business Media LLC.
43. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019;20:207–20.
44. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710.
45. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. Genome Biol. 2015;16:56.
46. Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Paul Bevan A, et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res. 2014:D993–1000. https://doi.org/10.1093/nar/gkt937.
47. Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature. 2018;555:538–42.
48. Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet Med. 2020;22:245–57.
49. Abou Tayoun AN, Pesaran T, DiStefano MT, Oza A, Rehm HL, Biesecker LG, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum Mutat. 2018;39:1517–24.
50. SVI Recommendation for Absence/Rarity (PM2) - Version 1.0. Available from: https://www.clinicalgenome.org/site/assets/files/5182/pm2_-_svi_recommendation_-_approved_sept2020.pdf.
51. Mesman RLS, Calléja FMGR, de la Hoya M, Devilee P, van Asperen CJ, Vrieling H, et al. Alternative mRNA splicing can attenuate the pathogenicity of presumed loss-of-function variants in BRCA2. Genet Med. 2020;22:1355–65.
52. Rowlands C, Thomas H, Lord J, Wai H, Arno G, Beaman G, et al. Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. Authorea Preprints: Authorea; 2020. Available from: https://www.authorea.com/users/363411/articles/484210-comparison-of-in-silico-strategies-to-prioritize-rare-genomic-variants-impacting-rna-splicing-for-the-diagnosis-of-genomic-disorders?commit=5b7eba62771fdff57de0b1086ce25ace18383163. Cited 2021 Apr 21
53. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature. 2018;555:611–6.
54. VanderMeer JE, Ahituv N. cis-regulatory mutations are a genetic cause of human limb malformations. Dev Dyn. 2011;240:920–30.
55. ACGS. ACGS best practice guidelines for variant classification in rare disease 2020. 2020.
56. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176:535–48.e24.
57. SVI Recommendation for in trans Criterion (PM3) - version 1.0. Available from: https://clinicalgenome.org/site/assets/files/3717/svi_proposal_for_pm3_criterion_-_version_1.pdf. Accessed 16 June 2022.
58. Wood KA, Ellingford JM, Thomas HB, Genomics England Research Consortium, Douzgou S, Beaman GM, et al. Expanding the genotypic spectrum of TXNL4A variants in Burn-McKeown syndrome. Clin Genet. 2021. https://doi.org/10.1111/cge.14082.
59. Brnich SE, Abou Tayoun AN, Couch FJ, Cutting GR, Greenblatt MS, Heinen CD, et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med. 2019;12:3.
60. Yépez VA, Mertes C, Müller MF, Klaproth-Andrade D, Wachutka L, Frésard L, et al. Detection of aberrant gene expression events in RNA sequencing data. Nat Protoc. 2021;16:1276–96.
61. Mertes C, Scheller IF, Yépez VA, Çelik MH, Liang Y, Kremer LS, et al. Detection of aberrant splicing events in RNA-seq data using FRASER. Nat Commun. 2021;12:529.

62. Aicher JK, Jewell P, Vaquero-Garcia J, Barash Y, Bhoj EJ. Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. Genet Med. 2020;22:1181–90.
63. Rowlands CF, Taylor A, Rice G, Whiffin N, Hall HN, Newman WG, et al. MRSD: a novel quantitative approach for assessing suitability of RNA-seq in the clinical investigation of mis-splicing in Mendelian disease. bioRxiv. medRxiv; 2021. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.03.19.21253973.
64. Cheng J, Çelik MH, Kundaje A, Gagneur J. MTSplice predicts effects of genetic variants on tissue-specific splicing. Genome Biol. 2021;22:94.
65. Vig A, Poulter JA, Ottaviani D, Tavares E, Toropova K, Tracewska AM, et al. DYNC2H1 hypomorphic or retina-predominant variants cause nonsyndromic retinal degeneration. Genet Med. 2020;22:2041–51.
66. Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, et al. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. Nat Biotechnol. 2019;37:803–9 Nature Publishing Group.
67. Perenthaler E, Yousefi S, Niggl E, Barakat TS. Beyond the exome: the non-coding genome and enhancers in neurodevelopmental disorders and malformations of cortical development. Front Cell Neurosci. 2019;13:352.
68. Gasperini M, Findlay GM, McKenna A, Milbank JH, Lee C, Zhang MD, et al. CRISPR/Cas9-mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. Am J Hum Genet. 2017;101:192–205.
69. Gelman H, Dines JN, Berg J, Berger AH, Brnich S, Hisama FM, et al. Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. Genome Med. 2019;11:85.
70. Collins RL, Glessner JT, Porcu E, Niestroj L-M, Ulirsch J, Kellaris G, et al. A cross-disorder dosage sensitivity map of the human genome: bioRxiv. medRxiv; 2021. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.01.26.21250098
71. Rowlands CF, Baralle D, Ellingford JM. Machine learning approaches for the prioritization of genomic variants impacting pre-mRNA splicing. Cells. 2019;8. https://doi.org/10.3390/cells8121513.
72. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021;13:31.
73. Danis D, Jacobsen JOB, Carmody L, Gargano M, McMurry JA, Hegde A, et al. Interpretable prioritization of splice variants in diagnostic next-generation sequencing: bioRxiv; 2021. p. 2021.01.28.428499. Available from: https://www.biorxiv.org/content/10.1101/2021.01.28.428499v1. Cited 2021 Jul 15
74. Lord J, Gallone G, Short PJ, McRae JF, Ironfield H, Wynn EH, et al. Pathogenicity and selective constraint on variation near splice sites. Genome Res. 2019;29:159–70.
75. Rojano E, Seoane P, Ranea JAG, Perkins JR. Regulatory variants: from detection to predicting impact. Brief Bioinform. 2019;20:1639–54.
76. Caron B, Luo Y, Rausell A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. Genome Biol. 2019;20:32.
77. Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. Am J Hum Genet. 2016;99:595–606.
78. Giacopuzzi E, Popitsch N, Taylor JC. GREEN-DB: a framework for the annotation and prioritization of non-coding regulatory variants in whole-genome sequencing: bioRxiv; 2020. p. 2020.09.17.301960. Available from: https://www.biorxiv.org/content/10.1101/2020.09.17.301960v1. Cited 2021 Jul 22
79. Evans DG, Bowers N, Burkitt-Wright E, Miles E, Garg S, Scott-Kitching V, et al. Comprehensive RNA analysis of the NF1 gene in classically affected NF1 affected individuals meeting NIH Criteria has high sensitivity and mutation negative testing is reassuring in isolated cases with pigmentary features only. EBioMedicine. 2016;7:212–20.
80. Bergougnoux A, Délétang K, Pommier A, Varilh J, Houriez F, Altieri JP, et al. Functional characterization and phenotypic spectrum of three recurrent disease-causing deep intronic variants of the CFTR gene. J Cyst Fibros. 2019;18:468–75.
81. Morris-Rosendahl DJ, Edwards M, McDonnell MJ, John S, Alton EWFW, Davies JC, et al. Whole-gene sequencing of reveals a high prevalence of the intronic variant c.3874-4522A>G in cystic fibrosis. Am J Respir Crit Care Med. 2020;201:1438–41.
82. Bhatia S, Bengani H, Fish M, Brown A, Divizia MT, de Marco R, et al. Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. Am J Hum Genet. 2013;93:1126–34.
83. Wheway G, Douglas A, Baralle D, Guillot E. Mutation spectrum of PRPF31, genotype-phenotype correlation in retinitis pigmentosa, and opportunities for therapy. Exp Eye Res. 2020;192:107950.
84. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat Genet. 2018;50:1327–34.
85. El-Brolosy MA, Kontarakis Z, Rossi A, Kuenne C, Günther S, Fukuda N, et al. Genetic compensation triggered by mutant mRNA degradation. Nature. 2019;568:193–7.
86. Bozhilov YK, Downes DJ, Telenius J, Marieke Oudelaar A, Olivier EN, Mountford JC, et al. A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker. Nat Commun. 2021;12:3806.

Publisher's Note