

Supplementary Materials

Construction of MNVAnno

Principle of MNV identification: The identification of MNVs comprises two main steps: i) Filter phased VCF: considering the sensitivity or accuracy of phasing and the range of regulatory elements (e.g., miRBS and TFBS), MNVAnno filters out variants that exist alone within a certain window (default 10 bp). In this study, to construct a relatively comprehensive MNV resource, we did not filter out MNVs located in low-complexity regions (LCRs). Meanwhile, MNVAnno only retains core information including chromosome, position, ID, reference alleles, alternative alleles and genotypes of samples. The above two operations greatly improve the computational efficiency of identification and reduce memory usage (Supplementary Fig. 16); ii) Identify MNVs: For each individual, MNVAnno scans the processed file with a certain window (default 10 bp) to detect MNVs in the same haplotype. To address MNVs containing >2 variants, MNVAnno prioritizes high-joint MNVs (e.g., 3-joint) and iteratively checks lower-joint combinations. This approach corrects misclassified low-joint MNVs and enhances MNV type diversity (Supplementary Fig. 17). Each MNV requires the real allele count (AC) value to be greater than 0. The calculation formula of the real AC value for a j -joint MNV is as follows:

$$\text{real AC} = \sum_{n=1}^N a_{jn} - \sum_{k=j+1}^K \sum_{n=1}^N a_{kn}$$

Here, N represents the number of samples, K is the maximum value of joint set by MNVAnno, j and k are the joint numbers of MNVs. The former calculates the AC of the j -joint MNV, and the latter calculates the real AC of the higher joint MNVs containing the j -joint MNV.

Annotation module of MNVAnno: Referring to two classic SNV annotation tools, ANNOVAR¹ and SnpEff², we developed an annotation module tailored for MNVs. In addition to Ensembl's GTF annotations, we integrated third-party databases, including: lncRNA from NONCODEV6³, miRNA from miRBase⁴, circRNA from circBase⁵, piRNA from piRBase⁶, enhancer from FANTOM5⁷, ATAC from ATACdb⁸, miRBS from TargetScan⁹, TFBS from UCSC¹⁰ and conserved element (CE)

from UCSC¹⁰. All datasets were lifted to a consistent genome build (GRCh38) using LiftOver¹¹. Users can also customize libraries for annotation as needed. The annotation module consists of three types: gene-based annotation (focusing on protein-coding genes), noncoding-based annotation, and region-based annotation. For gene-based annotation, MNVAnno provides abundant information such as variant types, variant effects and risk classifications (Supplementary Table 2 and 3). The risk classification must be used with care, as it was designed solely to streamline filtering and cannot predict their functional impact (high/low) on phenotypes of interest. Furthermore, MNVAnno offers several useful functions, including format conversion, extraction of regions of interest, and genotype retrieval for identified MNVs.

Factors influencing the performance of MNV analysis tools

Further careful analysis of the comparison results reveals that the differences in MNV identification results among various methods stem primarily from three aspects: supported MNV types and identification scope, input data requirements, and sub-joint variant merging capability.

(1) Supported MNV types and identification scope

The most prominent discrepancies across tools arise from their supported MNV types and genomic coverage. VarDict only detects continuous MNVs; Wang's tool mainly supports 2-joint MNVs and a few 3-joint MNVs within coding regions; MACARON is restricted to single codons; BCFtools focuses on exonic MNVs in intron-free General Feature Format (GFF) regions. MAC handles multi-joint MNVs genome-wide but identifies far fewer than others. In contrast, MNVAnno can handle both continuous and discontinuous ones (~80% of all MNVs), and supports 3-5 joint MNVs (~10% of our results) across the entire genome.

(2) Input data requirements

BAM-based tools (VarDict): VarDict relies on raw sequencing reads (BAM files) to detect MNVs by detecting adjacent variants in the same read. However, it cannot verify haplotype phase (cis/trans status) of variants. This leads to false positives (e.g., incorrect read mapping and haplotype phase error) and false negatives (cannot identify non-continuous MNVs).

Phased VCF-based tools (MNVA_{nno}, Wang's tool, BCFtools): These tools use phased VCFs to avoid phase-related errors. However, their performance depends on phasing quality—rare errors (mean error rate < 0.005) still introduce minor inaccuracies.

MAC requires both a SNP list from VCF files and BAM files for its operation. It infers haplotypes from reads spanning multiple SNVs, defines connected components as “Block of Mutation (BM)”, and then identifies MNVs based on BMs. MAC was published in 2015 and identifies far fewer MNVs than other tools. However, the authors have not provided specific details regarding its identification criteria or filtering thresholds, which makes it difficult to pinpoint the exact reasons for its discrepancies compared to other methods.

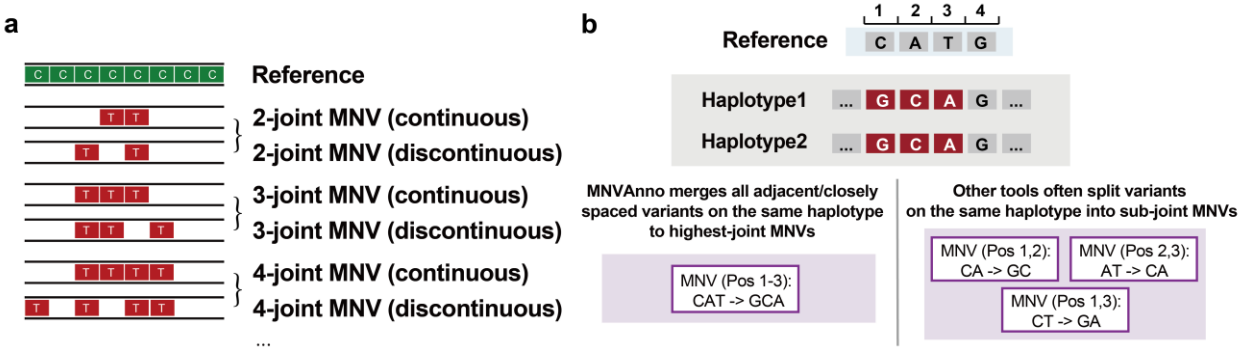
(3) Sub-joint MNV merging capability

MNVA_{nno} employs a "high-joint priority" identification strategy (Supplementary Fig. 1b). For example, if three adjacent or closely spaced variants are detected on the same haplotype, MNVA_{nno} directly classifies them as a single 3-joint MNV instead of splitting them into multiple 2-joint subsets (e.g., variants at positions 100, 102, 103 are annotated as one 3-joint MNV, not three separate 2-joint MNVs: 100+102, 102+103 and 100+103). This merging step prevents misclassification of sub-joint variants as independent MNVs and preserves the biological authenticity. In contrast, other tools lack this merging step, leading to redundant calls of overlapping sub-joint MNVs that reduce precision and recall.

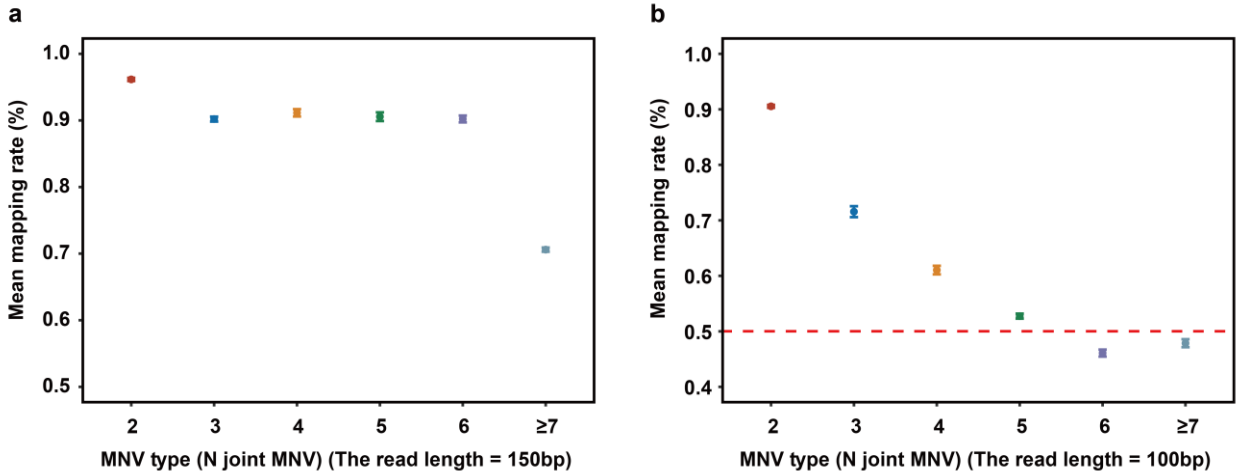
However, this strategy introduces a minor trade-off in comparative evaluations across tools, and it is the main reason MNVA_{nno} fails to achieve 100% precision/recall in gold-standard dataset comparisons. MNVA_{nno}'s merging strategy may lead to occasional false positives and false negatives due to rare phasing errors. For instance, if a 3-joint MNV identified by MNVA_{nno} includes one variant with phasing errors (e.g., an artifactual SNV), the entire 3-joint MNV results in a false positive classification. Meanwhile, this 3-joint MNV is split into multiple sub-joint MNVs by other tools (e.g., Wang's tool), and if those sub-joints have sufficient read support, MNVA_{nno} may be marked with false negatives for failing to detect those sub-joints independently. We carefully examined the comparison results of MNVA_{nno}. There were 391 false positives and 1,436 false negatives, and most of them were caused by high-joint MNV combinations. Nevertheless, we

still maintain that reporting the observed haplotype as a single MNV remains biologically more faithful than enumerating redundant sub-sets, and this principle aligns with our emphasis on analyzing MNV rather than individual constituent SNVs.

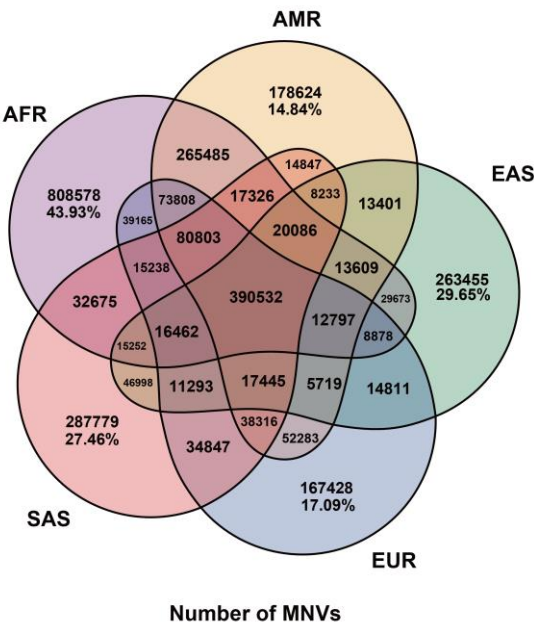
Supplementary Figures



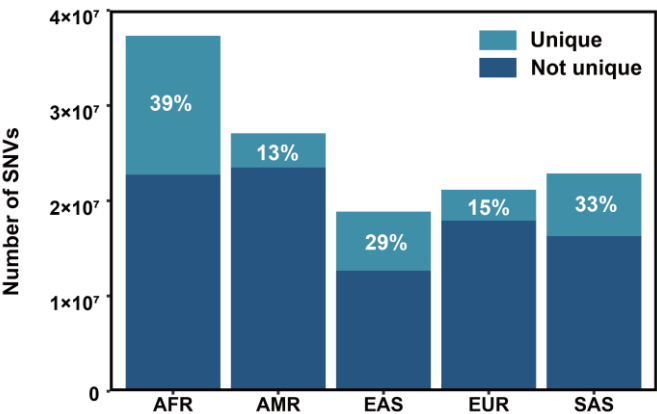
Supplementary Fig. 1: Types and identification strategy of MNVs. a Illustrative examples of the different MNV types; **b** Sub-joint MNV merging strategy.



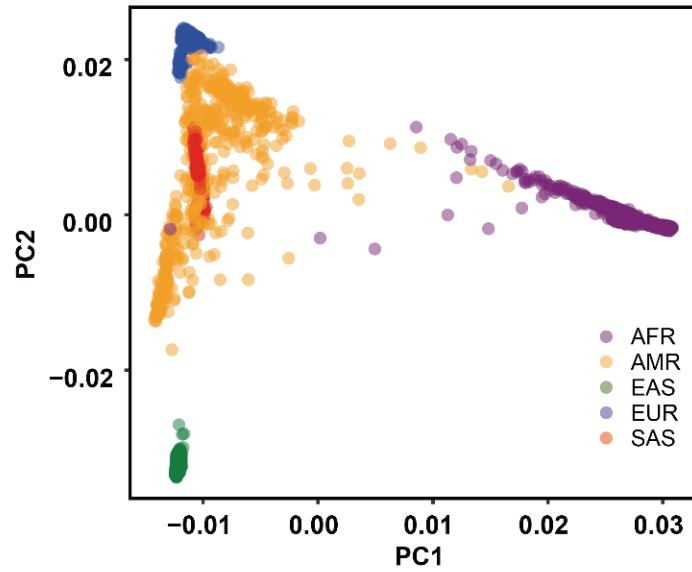
Supplementary Fig. 2: Mean mapping rates across different MNV types with varying read lengths. a Mean mapping rates for different MNV types with a read length of 150 bp; **b** Mean mapping rates with a read length of 100 bp. The red dashed line represents a 50% mapping rate threshold.



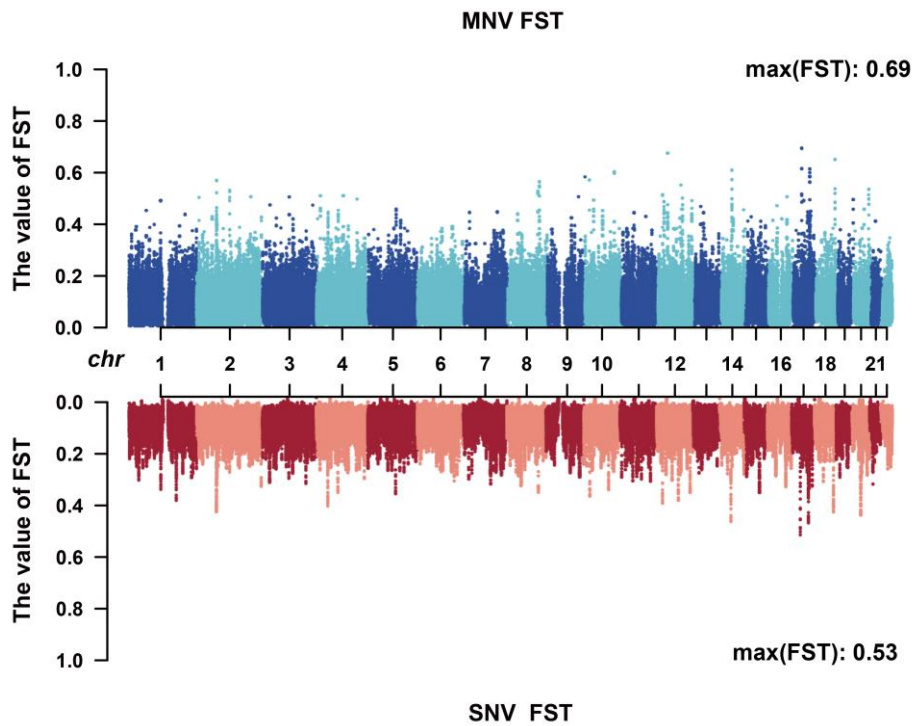
Supplementary Fig. 3: The overlap of MNVs across five populations (1000G).



Supplementary Fig. 4: The number and proportion of unique SNVs in five populations (1000G).



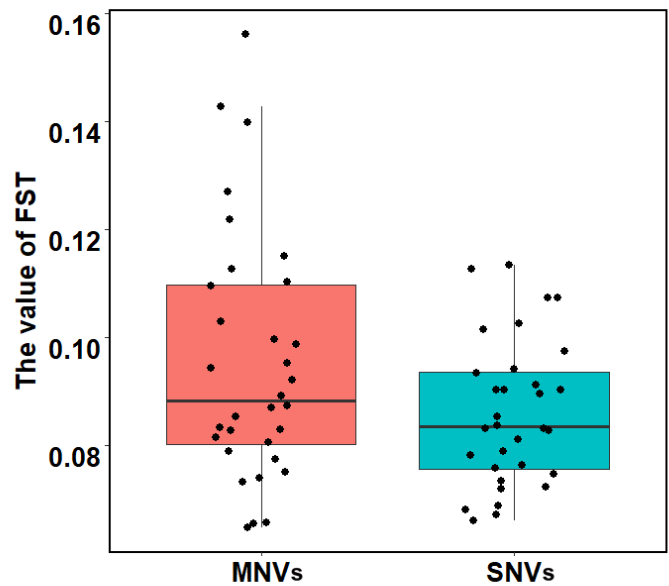
Supplementary Fig. 5: PCA result of SNV genotypes for population stratification.



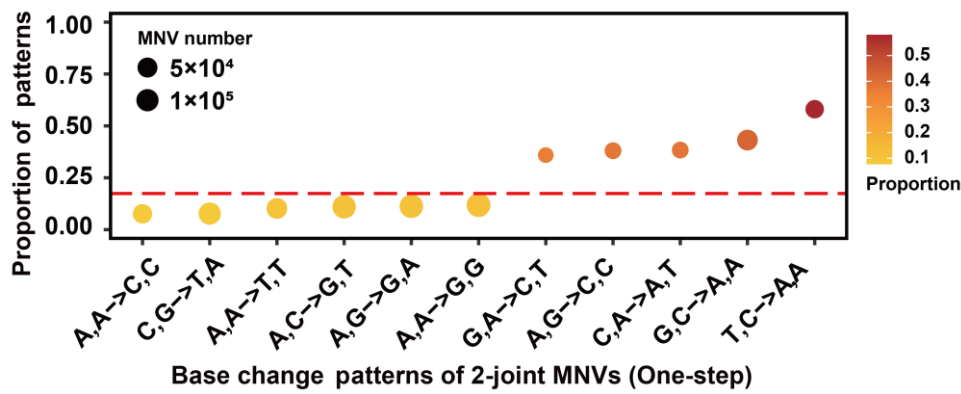
Supplementary Fig. 6: The weighted FST values of SNVs and MNVs in different populations.

Weighted FST values of SNVs and MNVs were calculated by VCFtools using 100 kb windows.

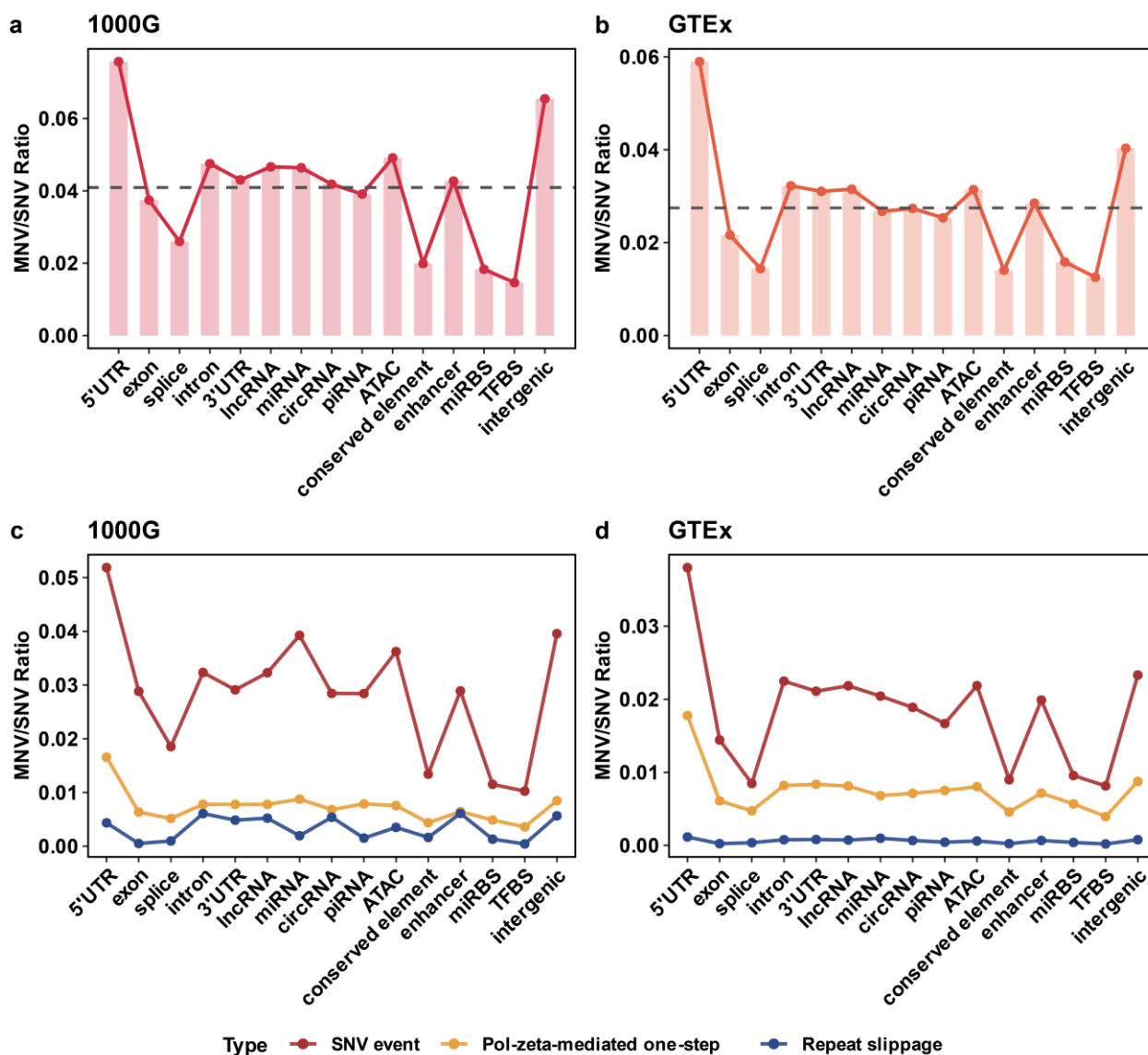
The difference between the largest weighted FST of MNVs and SNVs reaches 0.16.



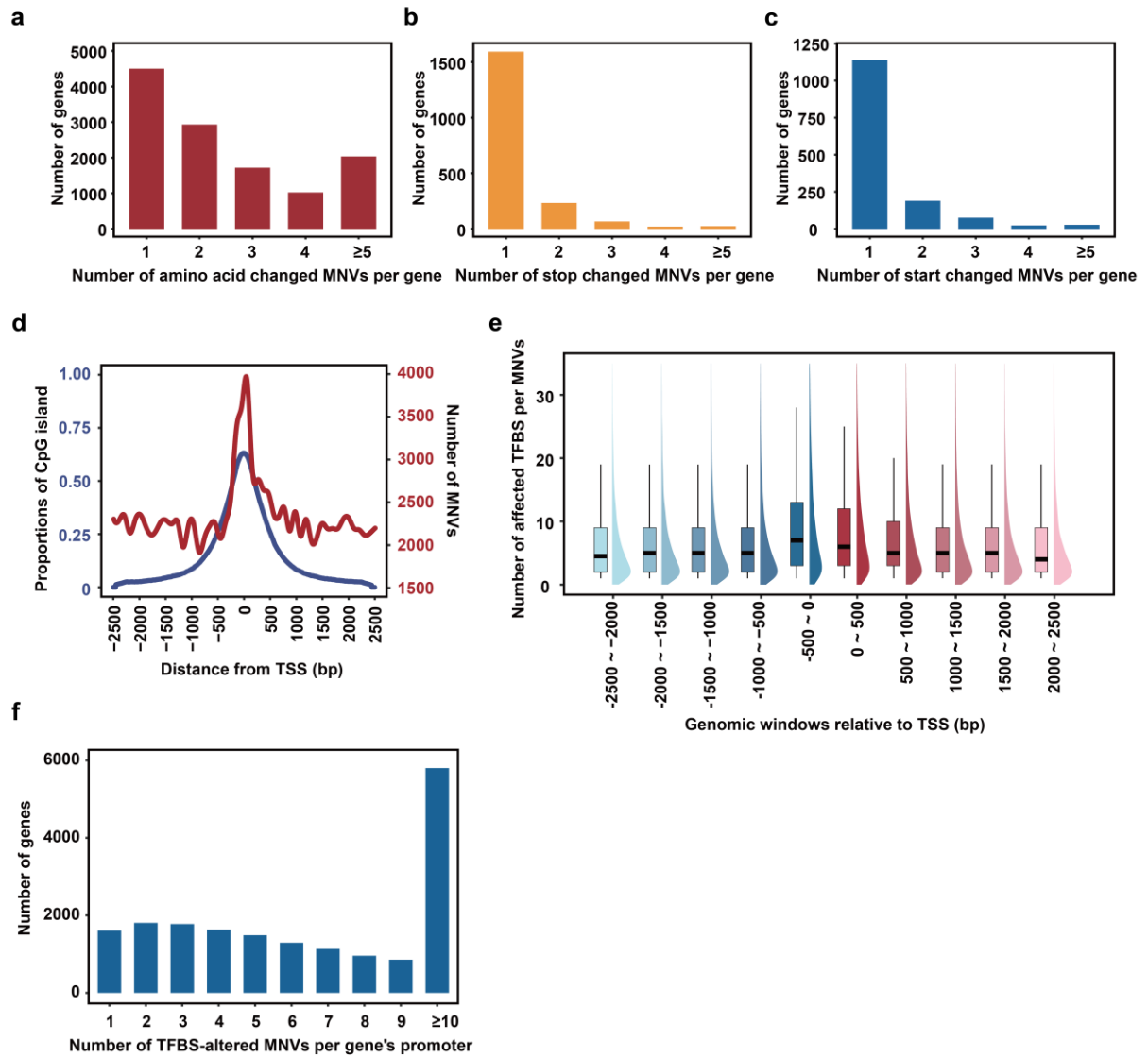
Supplementary Fig. 7: The FST values of MNVs and SNVs in cardiomyopathy-related genes



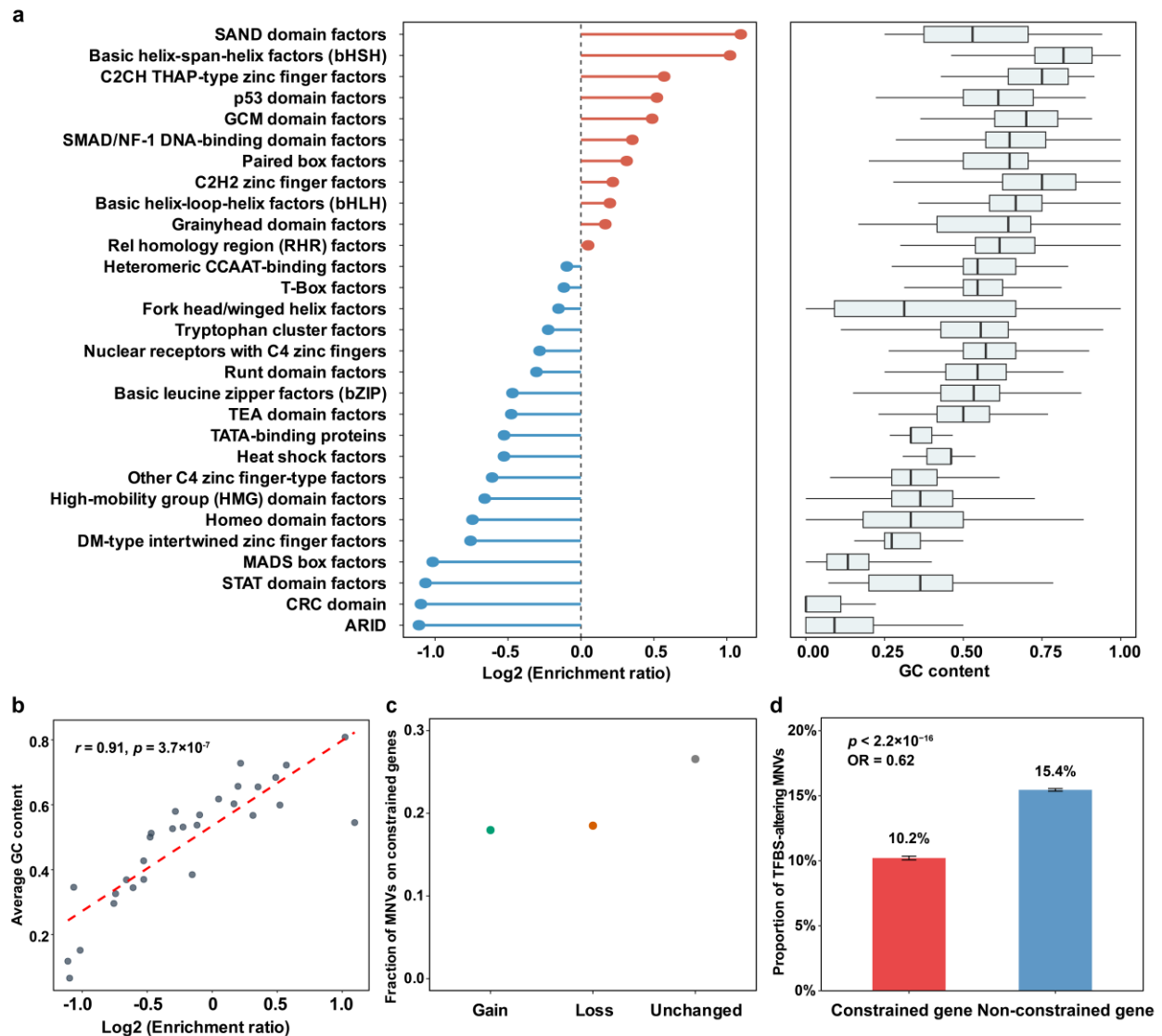
Supplementary Fig. 8: The proportion of different base change patterns in one-step for 2-joint MNVs. The red line represents the background, defined as the number of 2-joint MNVs belonging to the one-step divided by the total number of 2-joint MNVs.



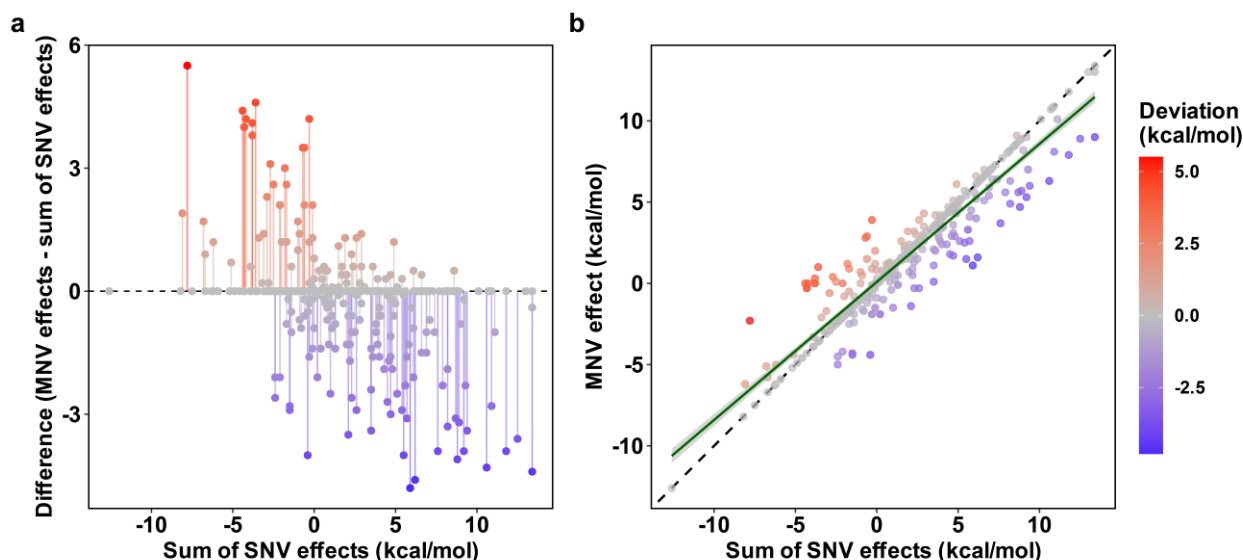
Supplementary Fig. 9: Distribution of the MNV/SNV ratios in different regions of 1000 Genomes Project and GTEx. **a,b** MNV/SNV ratios across annotated genomic regions in the 1000 Genomes Project (1000G; **a**) and GTEx (**b**). The dashed horizontal line represents the genome-wide background (baseline) MNV/SNV ratio. **c,d** MNV/SNV ratios of three mechanistic categories (SNV event, pol-zeta-mediated one-step, and repeat slippage) in 1000G (**c**) and GTEx (**d**).



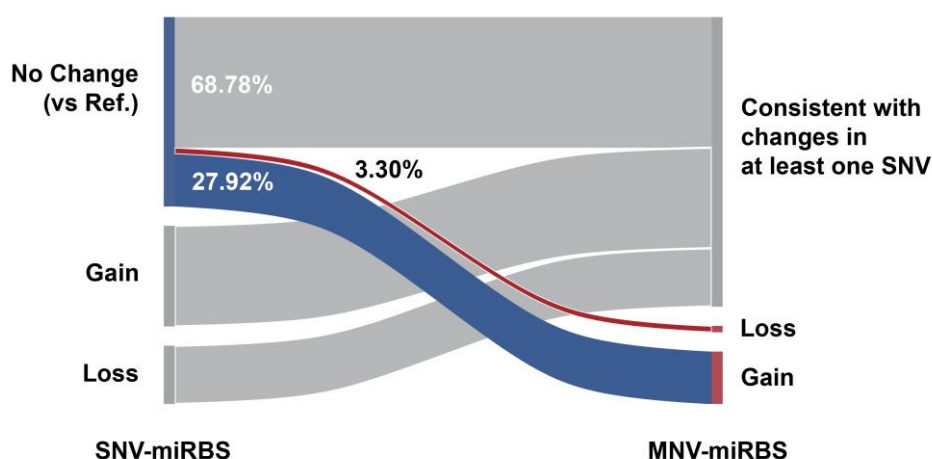
Supplementary Fig. 10: Functional impacts of MNVs. **a** The number of genes that carry amino acid changes caused by MNVs. **b** The number of genes that carry stop changes caused by MNVs. **c** The number of genes that carry start changes caused by MNVs. **d** Proportion of CpG island (blue) and MNV abundance (red) across different distances from the transcription start site (TSS). **e** The number of affected TFBS per MNV varies with distances from TSS. **f** The number of genes that carry differential MNV-TFBS.



Supplementary Fig. 11: Enrichment patterns and evolutionary constraints of TFBS-altering MNVs. **a** Enrichment of MNV-related TFBSs and the GC content across TF families. The enrichment ratio is defined as the observed rate of predicted gain/loss MNV-TFBSs relative to all potential TFBSs for each TF family. **b** Correlation between the GC content of TFBSs and the log2 enrichment ratio of TF families. **c** Fraction of MNVs (stratified by their TFBS functional categories: Creating/Gain, Disrupting/Loss, Unchanged) located in the regulatory regions of functionally constrained genes. **d** Proportion of TFBS-altering MNVs in functionally constrained versus non-constrained genes. Statistical significance was assessed using Fisher's exact test (p -value $< 2.2 \times 10^{-16}$).

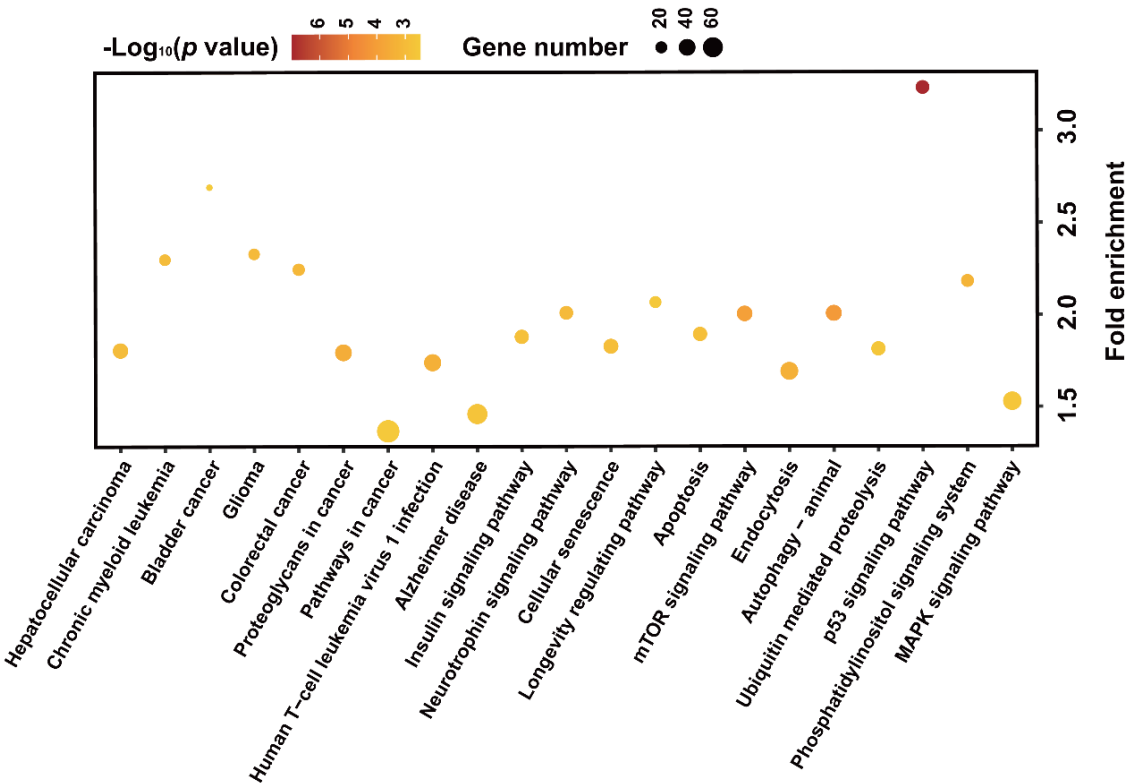


Supplementary Fig. 12: Non-additive and mitigation effects of MNVs on ncRNA secondary structures. **a** Divergence in Δ MFE between MNV effects and cumulative SNV effects. **b** Comparison of MNV effects versus the additive sum of constituent SNV effects on RNA secondary structure stability. The black dashed line ($y=x$) denotes the perfect additive model, while the solid green line with grey shading indicates the linear regression fit and 95% confidence interval. Points are colored by deviation magnitude.



Supplementary Fig. 13: Alterations from SNV-miRBS pairs (left) to MNV-miRBS pairs (right) for lncRNAs.

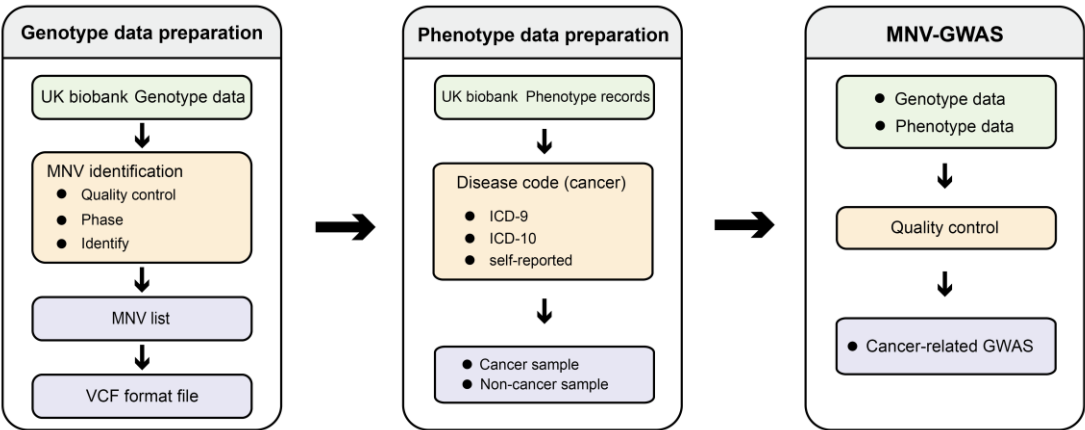
161



162

163 **Supplementary Fig. 14: KEGG pathway analysis for differential genes in hsa-miR-585-3p-**
164 **MNV group.**

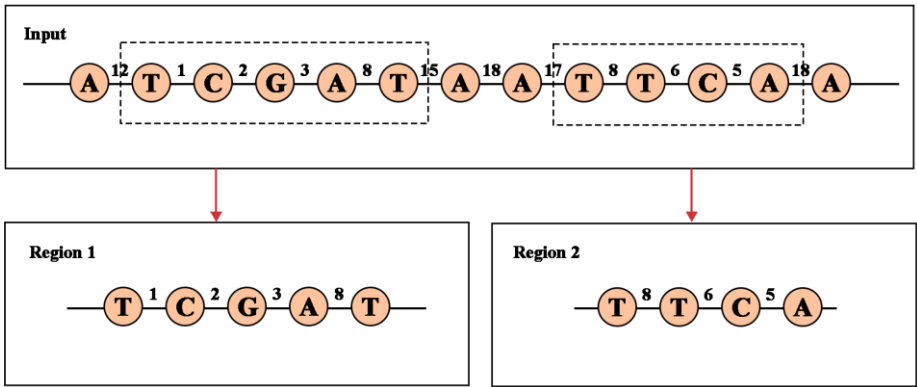
165



166

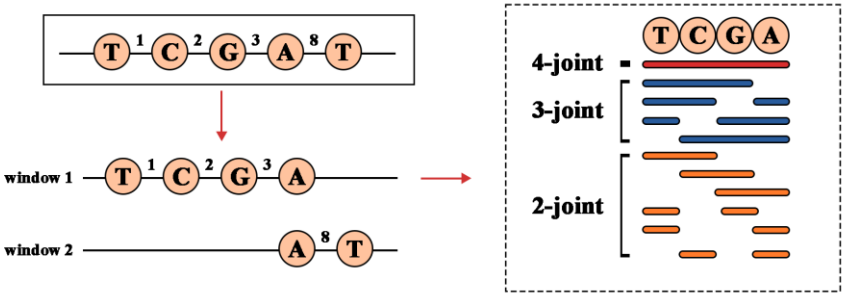
167 **Supplementary Fig. 15: Schematic workflow of the MNV GWAS pipeline. The illustration**
168 **was generated using Adobe Illustrator.**

169
170



171
172
173
174
175

Supplementary Fig. 16: The first step in processing the input VCF file. The values between the circles represent the distances between variants. MNVAnno filters out variants that exist alone within a certain window (default 10 bp).



176
177
178
179
180
181

Supplementary Fig. 17: An example of MNV identification. For the example 'TCGAT', MNVAnno splits it based on the given window (default 10 bp) and divides it into two consecutive subsets. Next, focusing on the subset 'TCGA', MNVAnno enumerates all combinations and prioritizes detection of higher-joint MNVs (e.g., 2-joint, 3-joint).

Supplementary Table 1: Intersection results of MNVs detected by MAC, MNVAnno, Wang's tool and VarDict.

Tool combination	Intersection count of MNVs
Unique to VarDict	13,662
Unique to Wang's tool	1,664
Unique to MNVAnno	495
Unique to MAC	555
MAC&MNVAnno	41
MAC&Wang's tool	0
MAC&VarDict	98
MNVAnno&Wang's tool	5,004
MNVAnno&VarDict	16
Wang's tool&VarDict	89
MAC&MNVAnno&Wang's tool	604
MAC&MNVAnno&VarDict	5
MAC&Wang's tool&VarDict	0
MNVAnno&Wang's tool&VarDict	1,116
MAC&MNVAnno&Wang's tool&VarDict	204

Supplementary Table 2: Sequence ontology (SO) in MNVAnno enables the standardization of terminology for annotating sequence variants and their functional impacts

Seq. Ontology Term	MNVAnno effect type	MNVAnno annotation
downstream_gene_variant	downstream_variant	Downstream of a gene (default length: 1 kb).
intergenic_region	intergenic_variant	The variant is in an intergenic region.
intron_variant	intron_variant	Variant is located within an intron; technically, it does not overlap with any exon in the transcript.
missense_variant	missense_variant	Variant causes a codon that produces a different amino acid.
5_prime_UTR_premature_start_codon_gain_variant	start_gain	A variant in 5'UTR region produces a three base sequence that can be a start codon.
start_loss	start_loss	Variant causes start codon to be mutated into a non-start codon.
stop_loss	stop_loss	Variant causes stop codon to be mutated into a non-stop codon.
stop_gain	stop_gain	Variant causes a stop codon.
splice_acceptor_variant	splice_acceptor_variant	The variant overlaps a splice acceptor site (defined as the two bases immediately upstream of an exon start, excluding the first exon).
splice_donor_variant	splice_donor_variant	The variant overlaps a splice donor site (defined as the two bases immediately downstream of a coding exon end, excluding the last exon).
splice_region_variant	splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, 3-5 bases of the intron.
synonymous_variant	synonymous_variant	Variant results in a codon encoding the identical amino acid.
upstream_gene_variant	upstream_variant	Upstream of a gene (default length: 2 kb).
3_prime_UTR_variant	UTR3_variant	Variant is located within the 3' UTR region.
5_prime_UTR_variant	UTR5_variant	Variant is located within the 5' UTR region.

Supplementary Table 3: Simple risk classification for the putative impact of the variants.

Risk	Note	Effect type
High	Variants in this category are predicted to exert a high (disruptive) impact on protein function, potentially causing protein truncation, loss of function, or inducing nonsense-mediated decay.	start_loss, stop_loss, stop_gain, splice_acceptor_variant, splice_donor_variant
Moderate	Non-disruptive variants that may change protein sequence and function.	missense_variant
Low	Assumed to be mostly harmless or unlikely to change protein behavior.	start_gain, splice_region_variant, synonymous_variant
Modifier	Typically includes non-coding variants or variants affecting non-coding genes, for which functional impact predictions are either challenging or lack supporting evidence	downstream_variant, intergenic_variant, intron_variant, upstream_variant, UTR3_variant, UTR5_variant

NOTE: This risk classification system was formulated with reference to the annotation frameworks of SnpEff and ANNOVAR. It must be used with caution, as it was developed solely to facilitate and streamline the variant filtering process. It should be emphasized that it is impossible to predict whether high-impact or low-impact variants underlie the phenotype of interest.

Supplementary References

1. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
2. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
3. Zhao, L. *et al.* NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res.* **49**, D165–D171 (2021).
4. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, D152–157 (2011).
5. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20**, 1666–1670 (2014).
6. Wang, J. *et al.* piRBase: integrating piRNA annotation in all aspects. *Nucleic Acids Res.* **50**, D265–D272 (2022).
7. Abugessaisa, I. *et al.* FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Res.* **49**, D892–D898 (2021).
8. Wang, F. *et al.* ATACdb: a comprehensive human chromatin accessibility database. *Nucleic Acids Res.* **49**, D55–D64 (2021).
9. Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19**, 92–105 (2009).
10. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).
11. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).