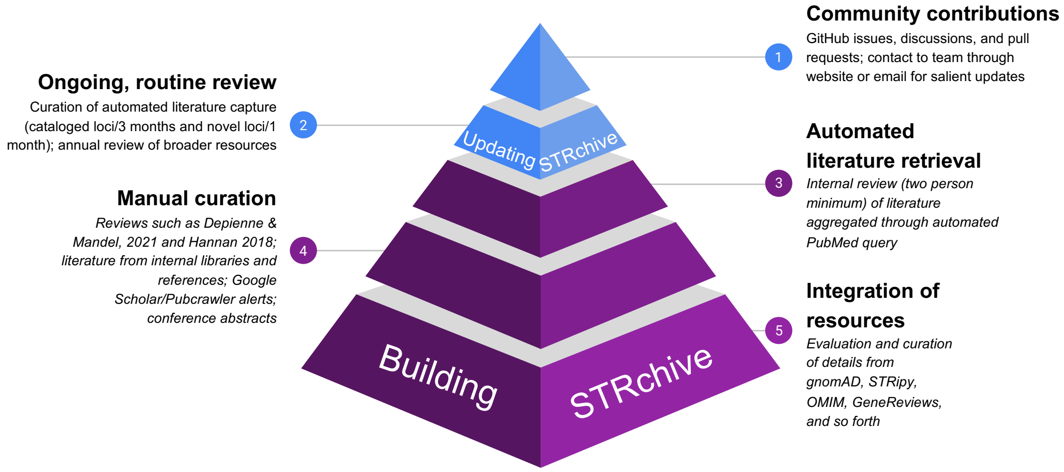
Supplementary Figures



Community contributions
GitHub issues, discussions, and pull requests; contact to team through website or email for salient updates

Ongoing, routine review
Curation of automated literature capture (cataloged loci/3 months and novel loci/1 month); annual review of broader resources

Automated literature retrieval
Internal review (two person minimum) of literature aggregated through automated PubMed query

Manual curation
Reviews such as Depienne & Mandel, 2021 and Hannan 2018; literature from internal libraries and references; Google Scholar/Pubcrawler alerts; conference abstracts

Integration of resources
Evaluation and curation of details from gnomAD, STRipy, OMIM, GeneReviews, and so forth

Fig. S1: Process figure showing the construction (steps 3-5) and maintenance (1-5) of STRchive as a resource.

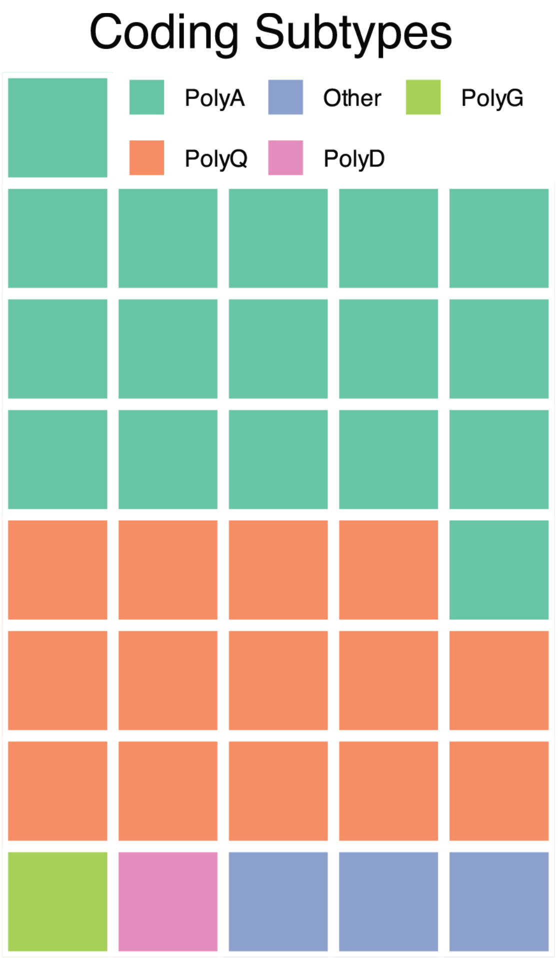


Fig. S2: Most coding TRs result in polyalanine and polyglutamine tracts. A waffle plot of the coding types' specific coding consequence, where each block represents a locus in STRchive. Script for generation (with further details of the other categories) is available in Figure 1 script.
PolyA: polyalanine; PolyG: polyglycine; PolyQ: polyglutamine; Other: amino acid patterns that do not fall under the previously listed categories. *MUC1* is not included as it is a homopolymer repeat of variable location within a VNTR cluster.

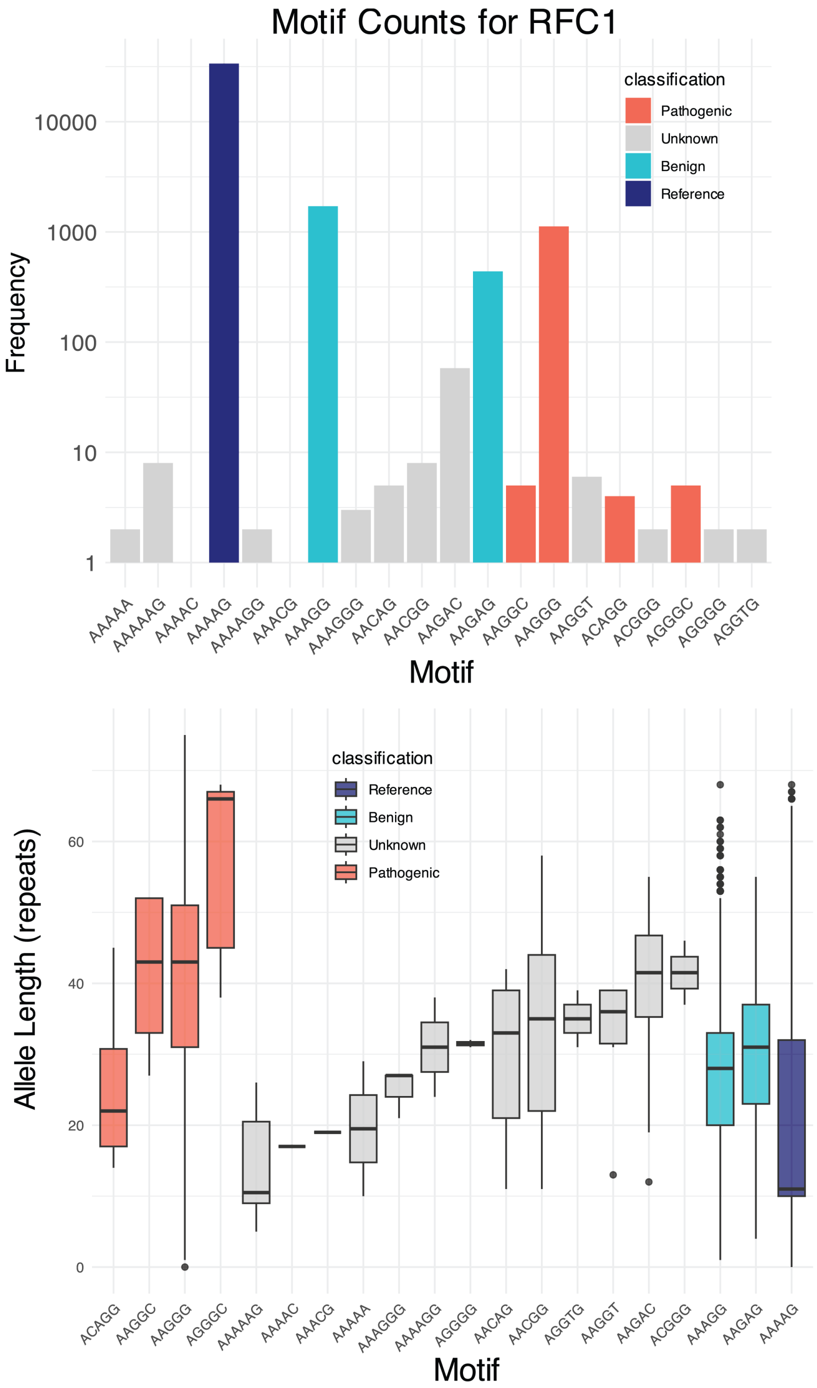


Fig. S3: RFC1 has the highest motif diversity of the gnomAD dataset, with motifs of all classifications. RFC1 motif counts from the gnomAD data for all motifs colored by clinical classification, which show a relatively high proportion of pathogenic motifs as well as a collectively high proportion of unknown motifs. **A.** Motif frequency for the gnomAD dataset. **B.** Motifs plotted by allele length and ordered by median allele size. Plot scripts are available in script for Figure 5 generation.

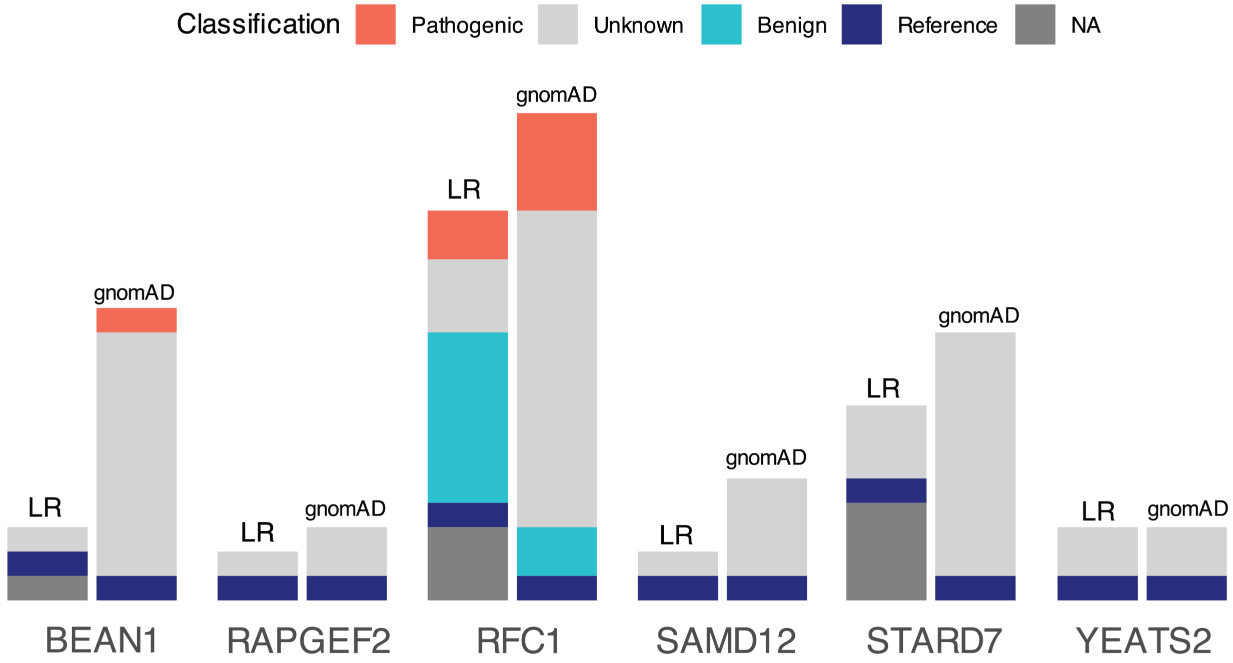


Fig. S4: Six TR disease loci show motif heterogeneity across gnomAD and HPRC cohorts. Unique motifs taken from the HPRC long-read (LR) data compared to the gnomAD data at overlapping STRchive loci. NA indicates “Not Available” in STRchive; other “Unknown” motifs have been documented in the literature before despite not being associated with a phenotype. Script is available in script for Figure 5 generation.

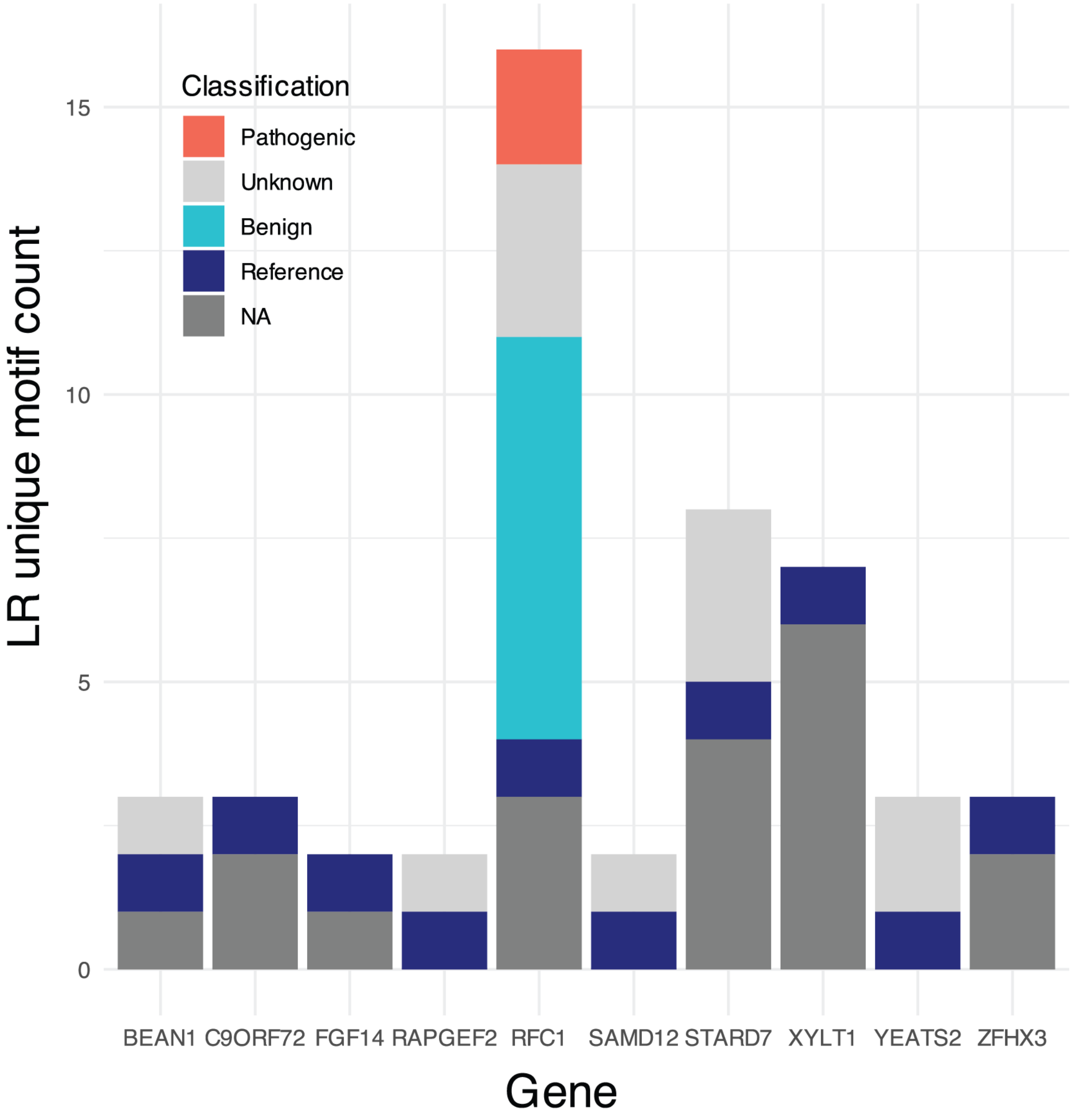


Fig. S5: Nine loci show motif heterogeneity within HPRC cohort, six with motifs previously not documented. Unique motifs taken from the HPRC long-read data, classified by STRchive, for all loci with more than one unique motif. NA indicates “Not Available” in STRchive; other “Unknown” motifs have been documented in the literature before despite not being associated with a phenotype. Script is available in script for Figure 5 generation.

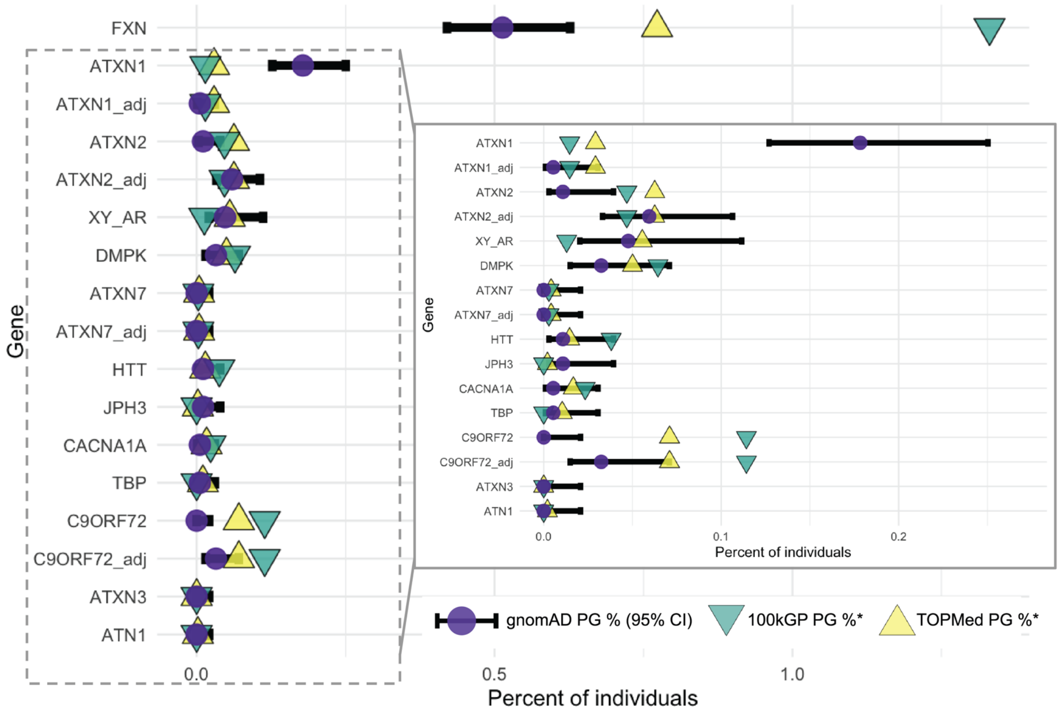
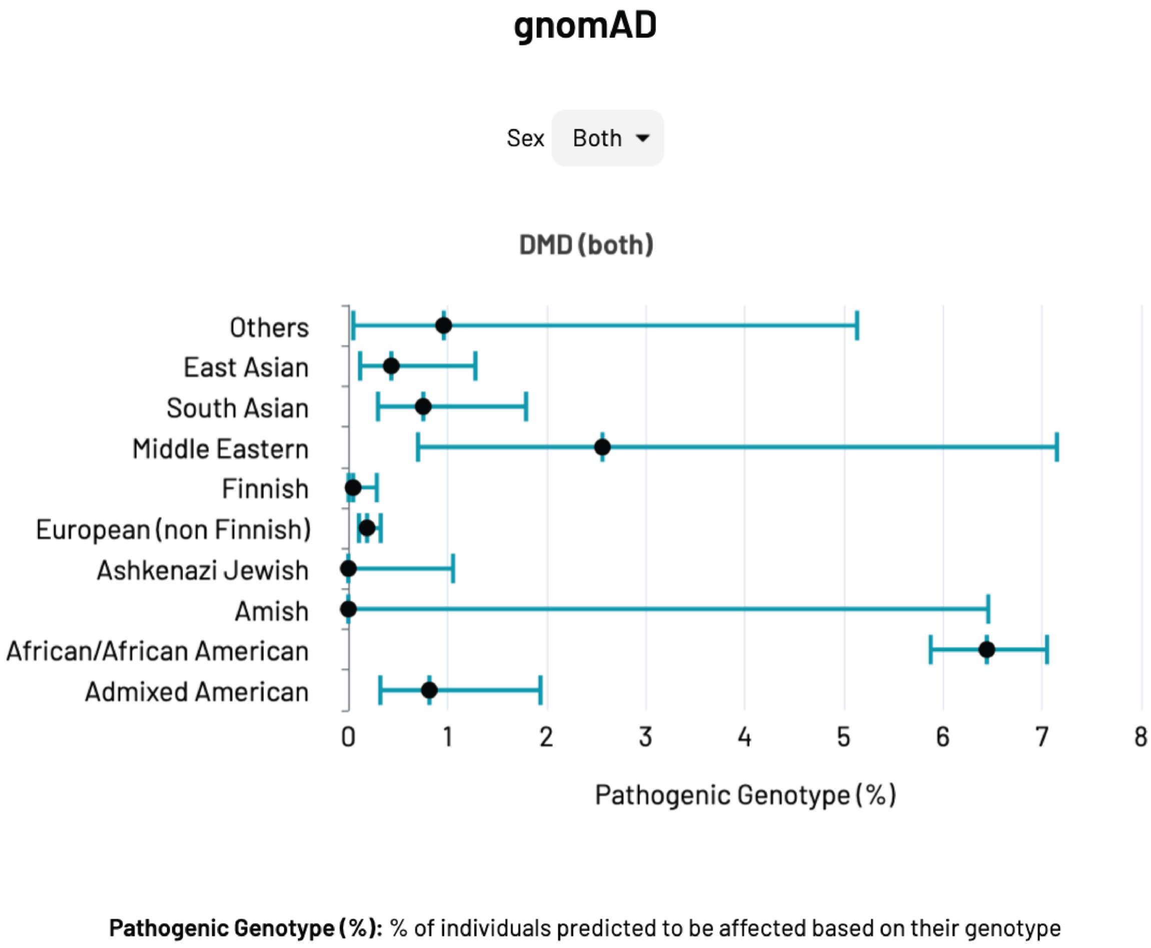


Fig. S6: Comparisons between gnomAD, TOPMed, and 100kGP show general concordance in PG percentage (inset range 0-0.27). An asterisk for PG % is used because FXN is an autosomal recessive locus, and so carrier percentage is used across cohorts rather than pathogenic genotypes. Matching the pathogenic thresholds in gnomAD to what was used by Ibañez et al. for four non-equivalent loci (*ATXN1, ATXN7, ATXN2, C9ORF72*) is indicated by *_adj. This adjustment altered the PG confidence interval to span the TOPMed/100kGP PG percentages for the first three loci while adjusting the pathogenic threshold for *C9ORF72* brought the TOPMed within 0.00018798% of the upper limit of the 95% confidence interval. The TOPMed estimate for *JPH3* was within 0.000879498% of the lower limit of the 95% confidence interval. Curiously, the gnomAD *FXN* PG percentage is comparable to disease prevalence in the literature despite being 37 to 67-fold lower than TOPMed/100kGP estimates. The *FXN* analysis within TOPMed and 100kGP shows ancestry-based PG variation, which aligns with variation in prevalence estimates across different populations reported previously for *FXN*. Consequently, the extreme variation in *FXN* PG percentages across studies may be due to cohort-specific ancestries.



gnomAD

Sex: Both

Pathogenic Genotype (%): % of individuals predicted to be affected based on their genotype

Fig. S7: *DMD* locus shows variation in PG percentage by ancestry. Figure is taken from STRchive.org *DMD* locus page with both sexes represented. The black circle indicates the pathogenic genotype percentage, and the green range shows the 95% confidence interval.