

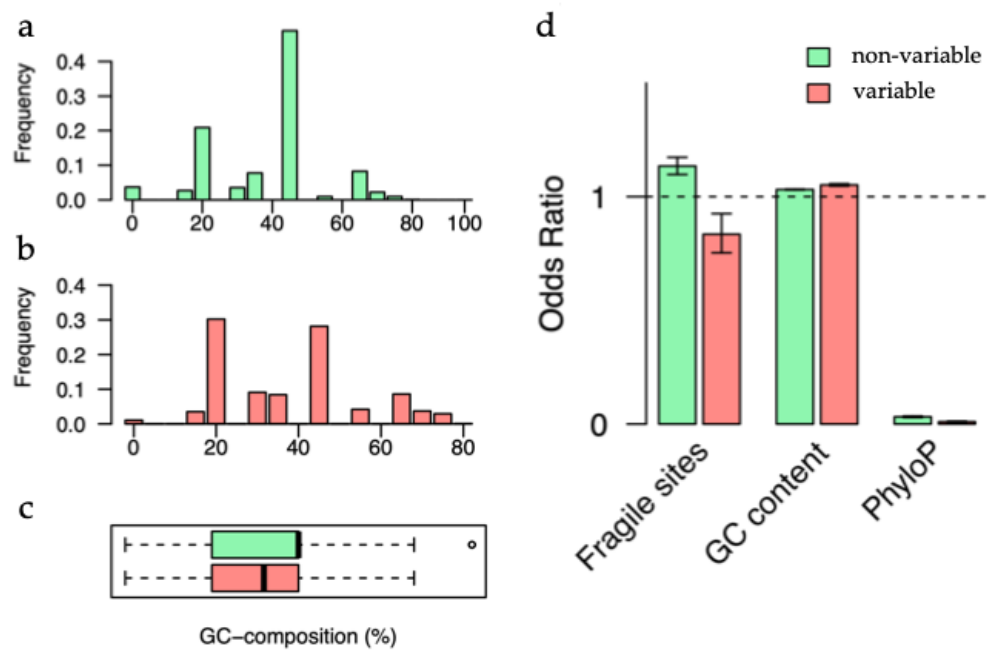


Fig S1. Characteristics of STRs in healthy individuals in GTEx cohort. **a,b** Distribution of the GC-composition of the motif in non-variable (a) and variable (b) STRs. The x-axis represents GC-composition, and the y-axis represents frequency. **c** Boxplot showing the GC-composition in non-variable (green) and variable (red) STRs. **d** Association analysis between the number of non-variable and variable STRs and different genomic features (x-axis) based on 1 kb segments. Odds ratios (y-axis) were derived from logistic regression coefficients of the genomic features with 95% confidence intervals.

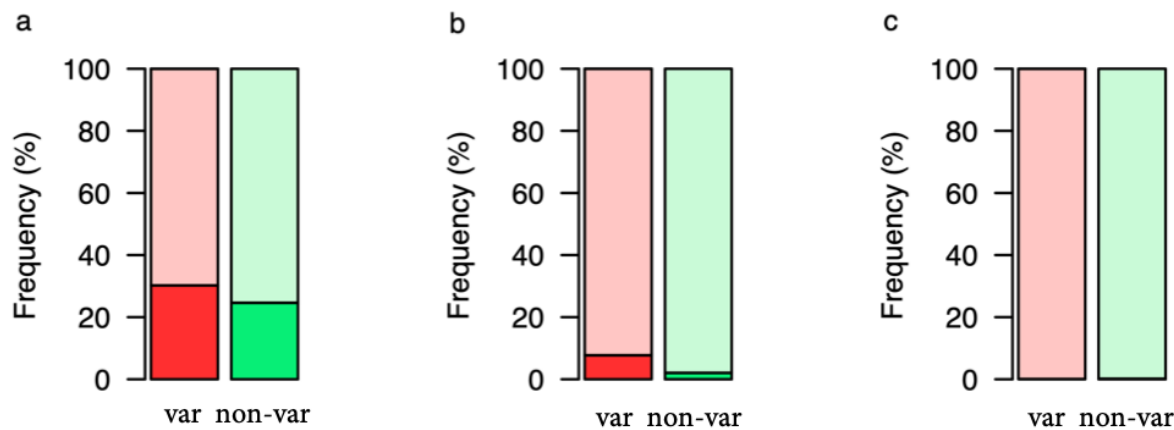


Fig S2. Comparison of non-variable and variable STRs in GTEx cohort. **a,b,c** Proportion of known STRs (a), expanded STRs (b) and known disease-causing regions (c) among variable STRs (red bars) and non-variable STRs (green bars).

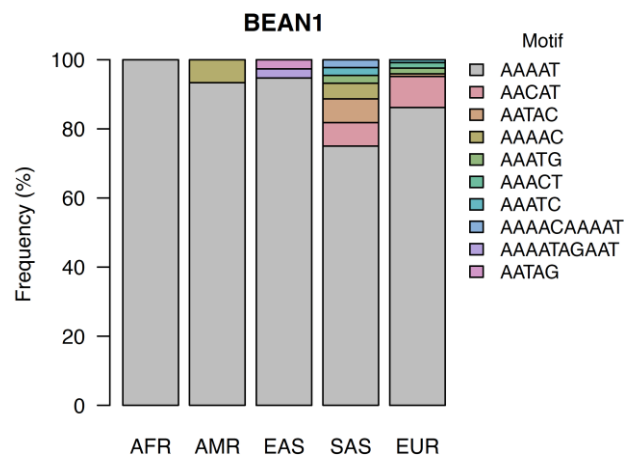


Fig S3. Motif frequency distribution in STR region in *BEAN1* gene across ethnic groups in 1000G samples.

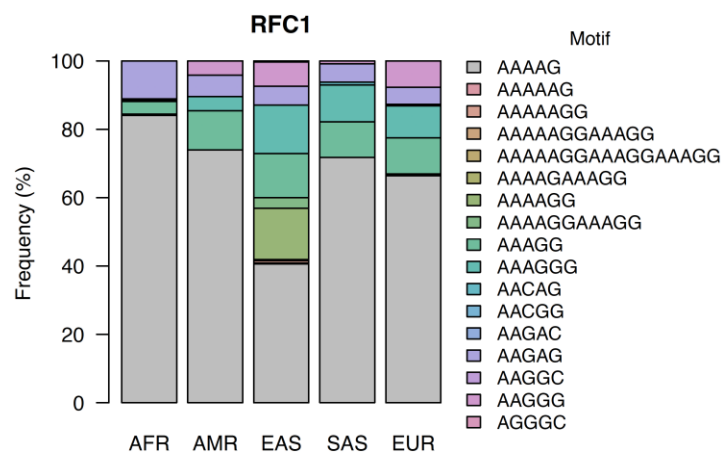


Fig S4. Motif frequency distribution in STR region in *RFC1* gene across ethnic groups in 1000G samples.

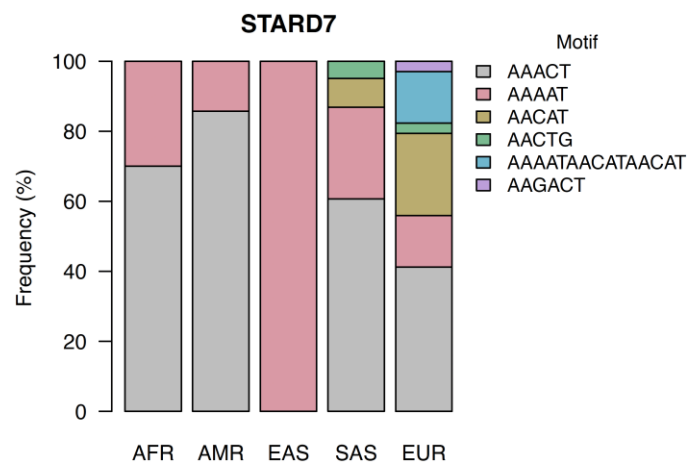


Fig S5. Motif frequency distribution in STR region in *STARD7* gene across ethnic groups in 1000G samples.

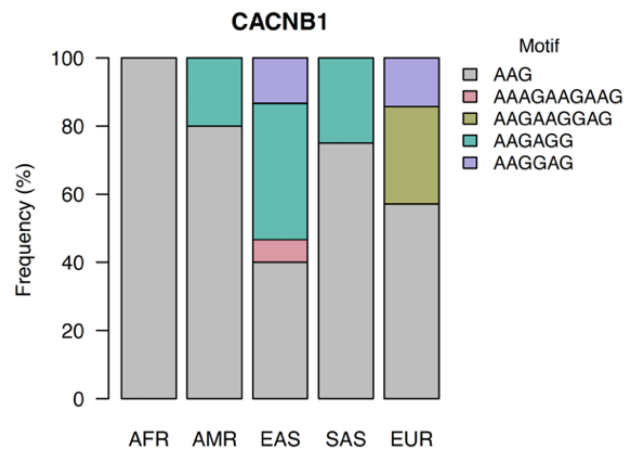


Fig S6. Motif frequency distribution in STR region in *CACNB1* gene across ethnic groups in 1000G samples.

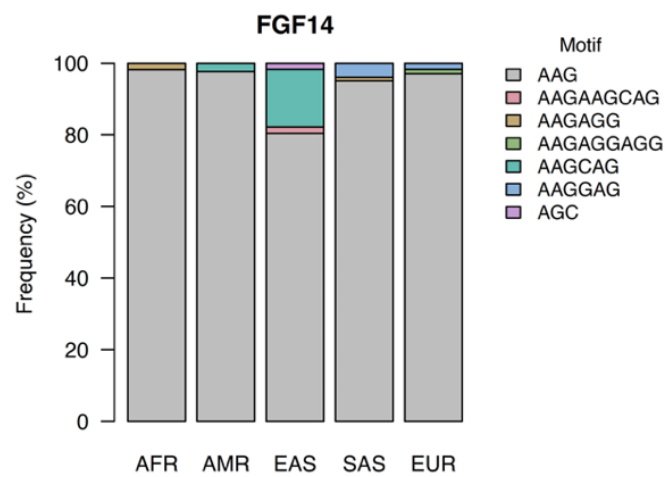
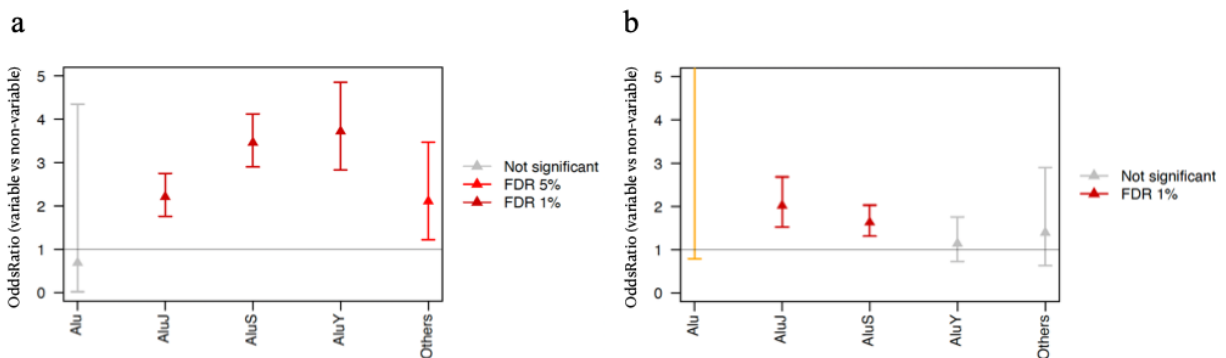


Fig S7. Motif frequency distribution in STR region in *FGF14* gene across ethnic groups in 1000G samples.



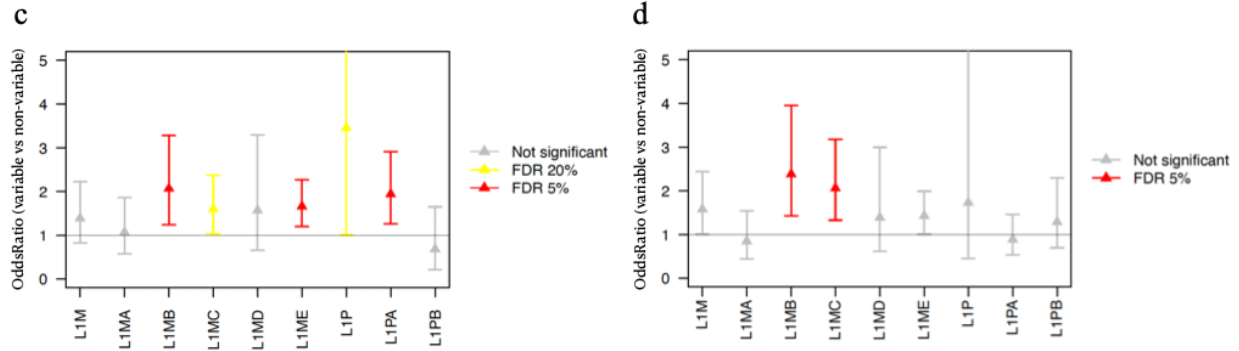


Fig S8. Characteristics of TRs near mobile elements. **a,b** Enrichment of variable over non-variable sequence STRs (a) and variable over non-variable sequence VNTRs (b) in proximity to *Alu* subfamilies. **c,d** Enrichment of variable over non-variable sequence STRs (c) and variable over non-variable sequence VNTRs (d) in proximity to *L1* subfamilies. The x-axis represents the distance from a transposable element, the y-axis represents the Fisher's exact test odds ratio. Triangles represent 1000G cohort samples. The color represents FDR.

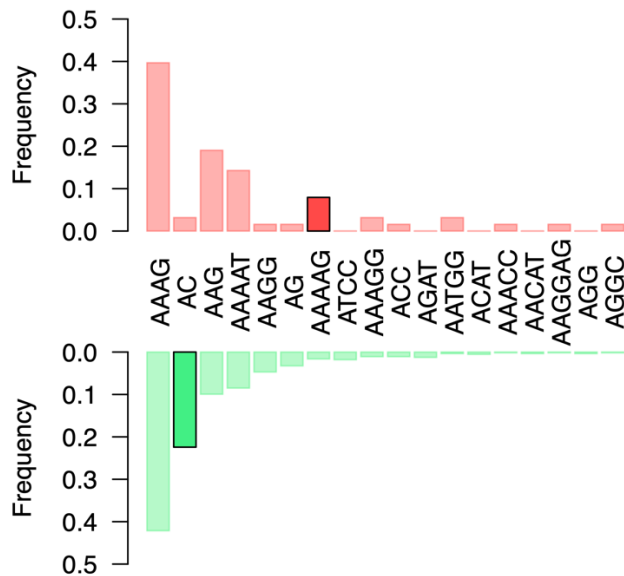


Fig S9. Distribution of the motif for variable sequence STRs (upper plot, red) and non-variable sequence STRs (lower plot, green) near the 3' end of *Alu* elements in 1000G cohort samples. The x-axis represents motif, the y-axis represents the fraction of the motif among STRs of the corresponding group. Bright red bars indicate motifs significantly enriched in variable sequence STRs over non-variable sequence STRs, bright green bars indicate motifs significantly enriched in non-variable sequence STRs over non-variable sequence STRs.

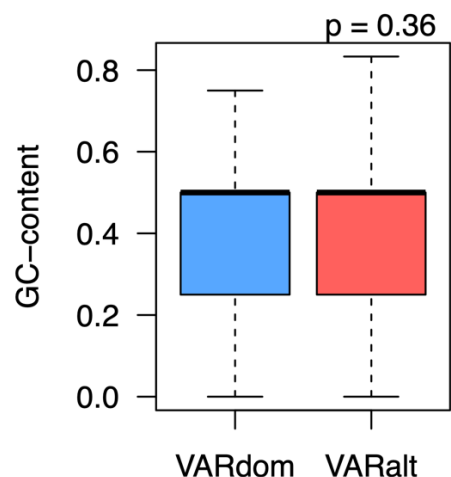


Fig S10. Boxplot for GC-composition in dominant motifs (VARdom) as compared to alternative motifs (VARalt) in the variable STRs that are up-regulated in the presence of alternative motif. The x-axis represents motif type, the y-axis represents GC-composition. Two-sided Wilcoxon's test p-value is labelled on top.

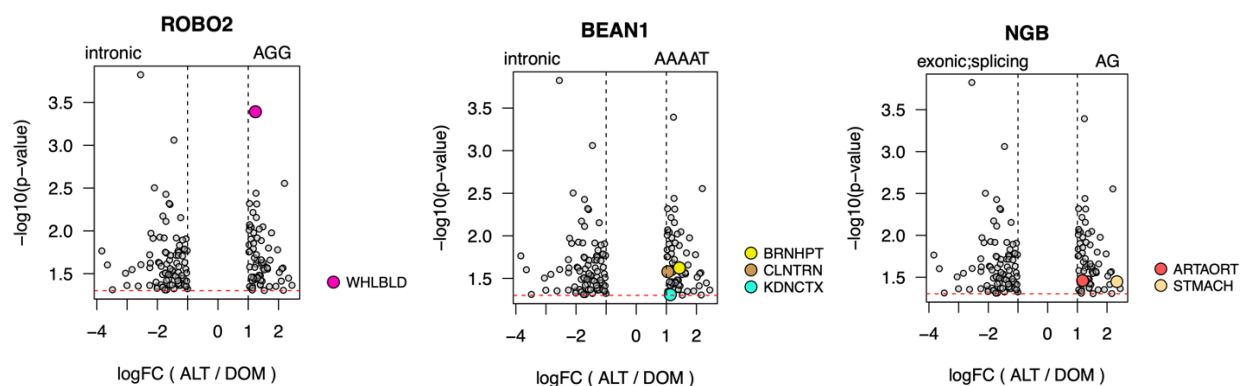


Fig S11. Genes upregulated in the presence of an alternative motif. The x-axis represents log fold change (logFC) for expression level between alternative and dominant motif-containing samples. The y-axis represents $-\log_{10}(\text{p-value})$. Grey dots represent all other genes. Only genes with $|\log\text{FC}| > 1$, and $\text{p-value} < 0.05$ are shown. Large colored dots represent corresponding gene (*ROBO2*, *BEAN1*, or *NGB*). Color represents tissue. Gene name and the dominant motif are labelled on top. Red horizontal dashed line represents significance level = 0.05. Vertical dashed lines represent $\log\text{FC} = \pm 1$

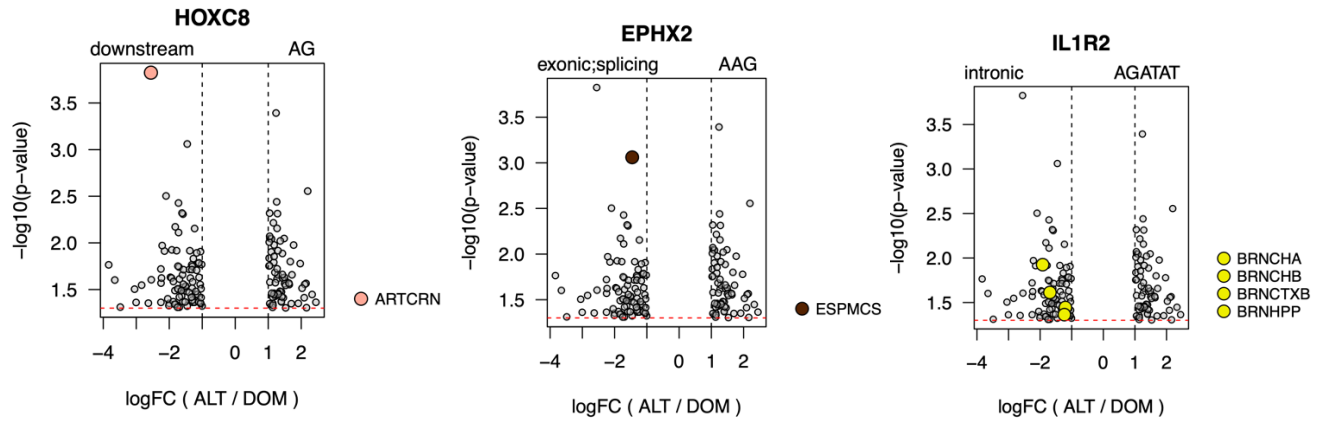


Fig S12. Genes downregulated in the presence of an alternative motif. The x-axis represents log fold change (logFC) for expression level between alternative and dominant motif-containing samples. The y-axis represents $-\log_{10}(\text{p-value})$. Grey dots represent all other genes. Only genes with $|\log\text{FC}| > 1$, and $p\text{-value} < 0.05$ are shown. Large colored dots represent corresponding gene (*HOXC8*, *EPHX2*, or *IL1R2*). Color represents tissue. Gene name and the dominant motif are labelled on top. Red horizontal dashed line represents significance level = 0.05. Vertical dashed lines represent $\log\text{FC} = \pm 1$