Supplementary Methods

Experimental validation of TR genotypes

The following loci were amplified using separate PCR conditions than those described in **Methods**:

- SCA1 (ATXN1): 95°C for 5 mins, 28 cycles of 95°C for 1 min, 61°C for 1 min, 72°C for 1.5 min; 8 cycles of 95°C for 1 min, 53°C for 1 min, 72° C for 1.5 mins, and a final extension at 72°C for 10 min.
- ATN1: 95°C for 15 min, 27 cycles at 94°C for 20 s, 65° C for 30 s, 72°C for 1.5 min; 8 cycles of 94°C for 20 s, 53°C for 30 s, 72°C for 1.5 mins and a final extension at 72°C for 10 min.

Several additional loci were initially analyzed by capillary electrophoresis but were filtered from downstream analysis since we determined they produced unreliable genotypes:

- Known pathogenic repeats at *HOXD13* (F:tgtaaaacgacggccagtAAGGAGAGGAGGGAGGAGG, R:GACATACGGCAGCTGTAGTAGC and *PAPBN1* (F:tgtaaaacgacggccagtCCAGTGCCCCGCCTTAGA, R: ACAAGATGGCGCCGCCGCCCCGGC) were excluded since all samples were called as homozygous reference by both EnsembleTR and capillary electrophoresis.
- A known pathogenic repeat in FXN was genotyped but excluded from comparisons since it was not genotyped by EnsembleTR.
- A known pathogenic repeat for SCA2 (F:6FAM-GGGCCCCTCACCATGTCG, R:CGGGCTTGCGGACATTGG) was excluded since many capillary calls had evidence of three alleles in a single sample, and we could not determine the correct diploid call.
- A known pathogenic repeat in *PHOX2B* (F:tgtaaaacgacggccagtCCAGGTCCCAATCCCAAC, R:GAGCCCAGCCTTGTCCAG) was excluded since all capillary calls were heterozygous for the same alleles, which is unlikely to be correct.
- We additionally excluded known pathogenic loci on chrX, since our calls focus on autosomal TRs. Results for the FMR1 repeat, which is on chrX, are reported from Asuragen but are not included in the comparison.

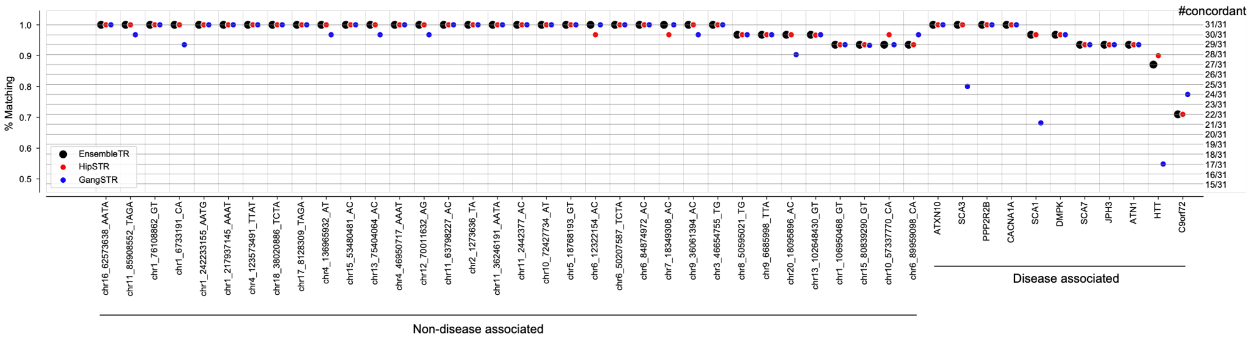
Fragment analysis was performed in separate batches for Platinum Genomes (17 samples; “PG”) and the remaining 31 samples (“Coriell”). Product sizes were converted to repeat differences from reference to be comparable to our WGS genotypes. In some cases, product sizes were offset by a constant value from the expected values. We binned product sizes to allele lengths, with binsets manually set for each locus in each batch. For each locus, we set the minimum (minval) and maximum (maxval) product size, the repeat unit length (period), and allele assignment of the minimum product size (start). We then determined the allele for a given product size (psize) using the following rule:

1. Set allele=start and i=minval
2. While i<maxval: if psize>=i and psize<i+period return allele. Else allele+=1, i+=period
3. If no matching allele was found, return NA

The bin definitions for each locus/batch are provided in **Supplementary Data 7**.

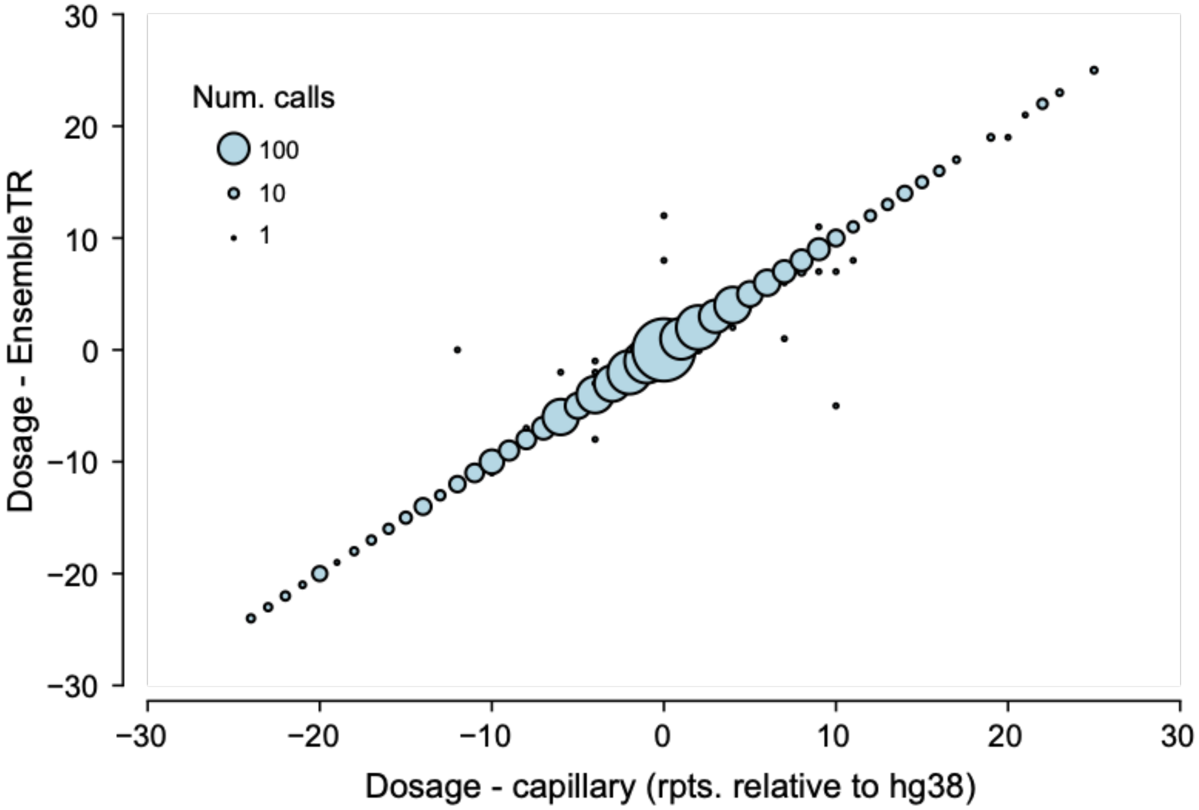
Supplementary Figures

Supplementary Fig. 1



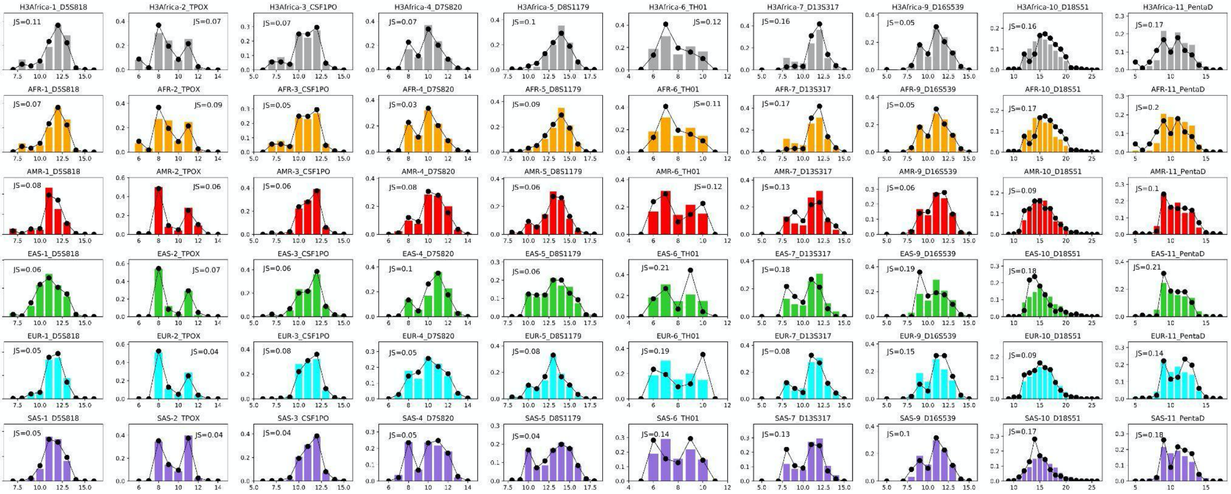
Per-locus comparison between WGS-based TR genotypes and capillary electrophoresis results. Each point gives the percent of genotypes that were concordant (y-axis) between capillary electrophoresis and WGS (black=EnsembleTR; red=HipSTR; blue=GangSTR) for each TR (x-axis). TRs are arranged by category (left=non-disease associated TRs; right=disease-associated TRs). Within each category, TRs are sorted by EnsembleTR concordance. Horizontal lines denote the % matching corresponding to the different possible numbers of concordant genotypes (out of a maximum of 31 samples). Full genotype comparison results are provided in **Supplementary Data 6**. Note, for HTT, GangSTR does not distinguish between the CAG_n and CCG_n repeats, whereas the Asuragen test used for PCR counts only CAG copies. Similarly, SCA1 consists of an imperfect repeat, and GangSTR only considers the length of perfect repeat stretches. This likely accounts for lower concordance of GangSTR compared to EnsembleTR and HipSTR at those loci.

Supplementary Fig. 2



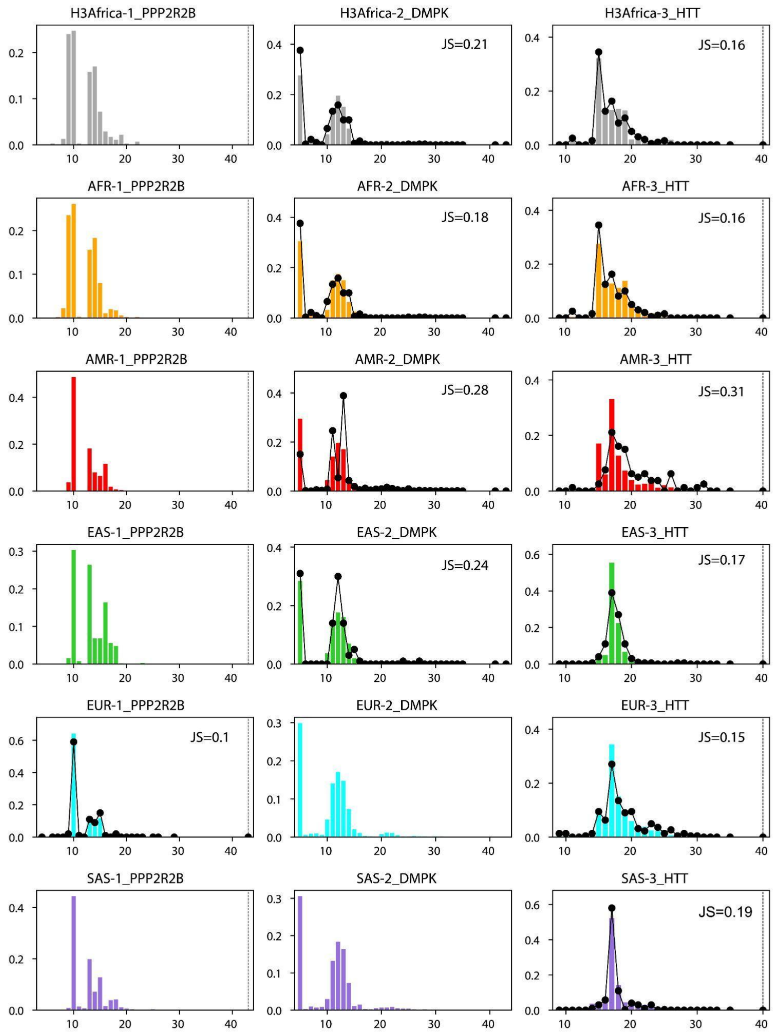
Comparison of genotypes called by capillary electrophoresis vs. EnsembleTR. The length of each allele relative to the hg38 reference was computed and lengths of the two alleles were summed for each call. The x-axis gives the capillary result, and the y-axis gives the result called by EnsembleTR. Bubble size scales with the number of genotype calls at each coordinate.

Supplementary Fig. 3



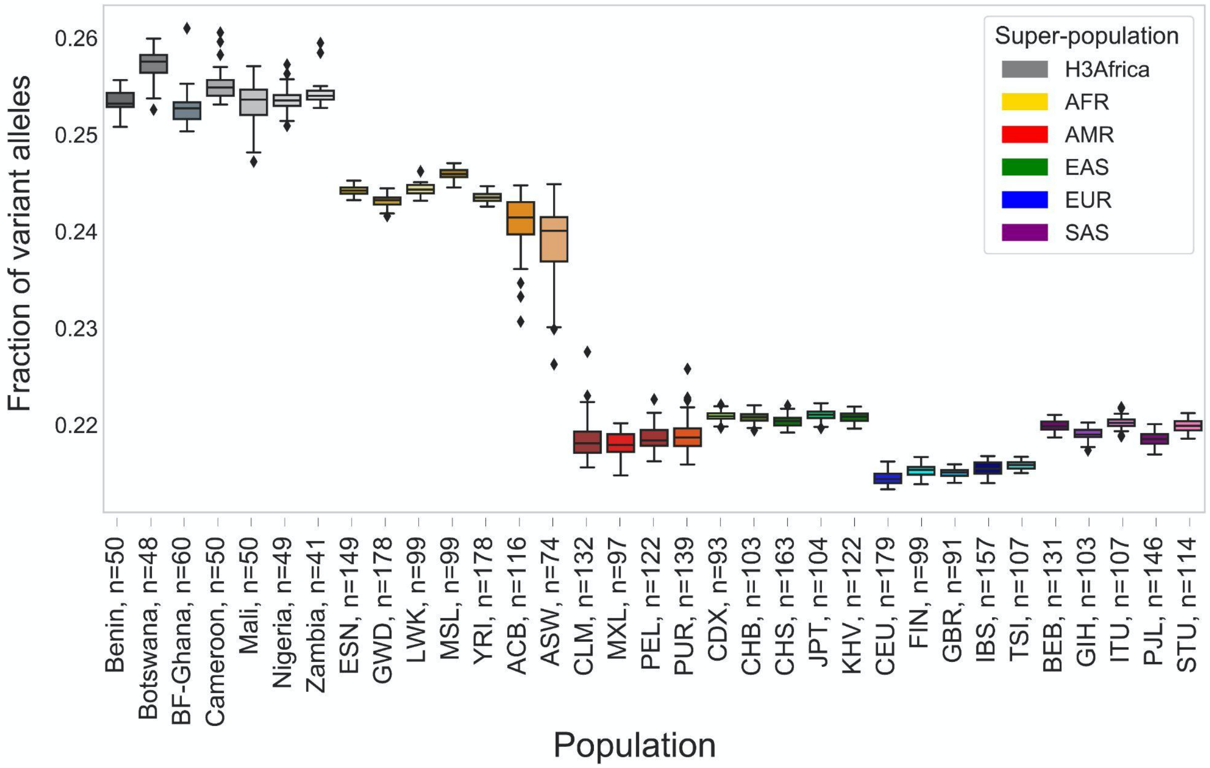
Comparison of population-specific allele length distributions called by EnsembleTR vs. published frequencies at the CODIS forensic loci. Rows represent 1000GP superpopulations and H3Africa (gray=H3Africa, blue=EUR, green=EAS, orange=AFR, red=AMR, purple=SAS). Each column represents a different CODIS locus. Colored bars give the frequency (y-axis) of each allele length (x-axis) observed in EnsembleTR results in terms of the number of repeat units. EnsembleTR alleles at the TH01 locus, a tetranucleotide repeat with a common 3bp interruption, were rounded down to the nearest repeat unit. All EnsembleTR alleles at D5S818 were adjusted down by one repeat unit since the EnsembleTR call contains an extra imperfect repeat unit not counted in published frequency data. Black dots give the expected frequency of each length based on previous literature (see **Methods**). For each locus/population pair, the Jensen-Shannon (JS) divergence between expected frequencies and EnsembleTR frequencies is annotated at the top of the plot.

Supplementary Fig. 4



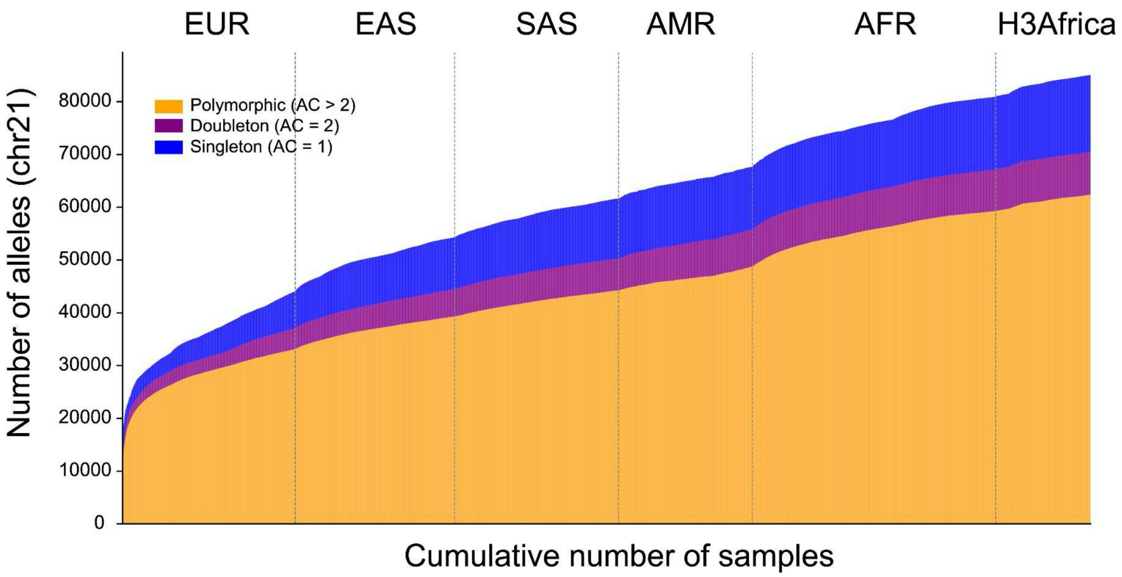
Comparison of population-specific allele length distributions called by EnsembleTR vs. published frequencies at disease-associated loci. Rows represent 1000GP superpopulations and H3Africa (gray=H3Africa, blue=EUR, green=EAS, orange=AFR, red=AMR, purple=SAS). Each column represents a different locus implicated in a TR expansion disorder. Colored bars give the frequency (y-axis) of each allele length (x-axis) observed in EnsembleTR results in terms of the number of repeat units. Black dots give the expected frequency of each length based on previous literature (see **Methods**). For each locus/population pair, the Jensen-Shannon (JS) divergence between expected frequencies and EnsembleTR frequencies is annotated at the top of the plot. For HTT, we counted only the number of CAG repeats in alleles returned by EnsembleTR to be consistent with counts obtained in the literature, which do not count the adjacent CCG repeat.

Supplementary Fig. 5



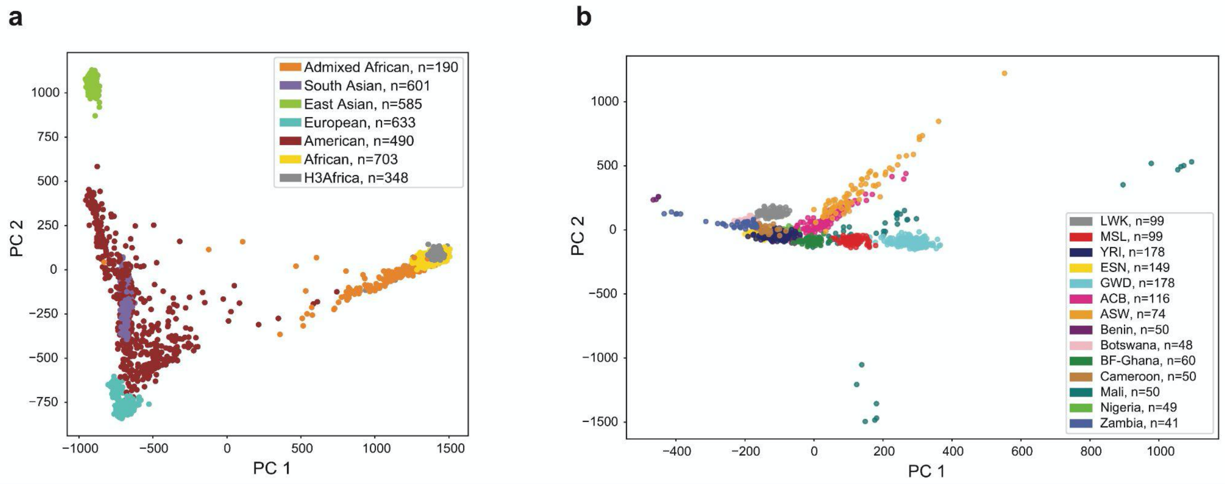
Distribution of the fraction of non-reference alleles in individuals by population. Boxplots summarize the distribution of the fraction of variant alleles in each sample. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1–1.5*IQR (bottom) and Q3+1.5*IQR (top), where IQR gives the interquartile range (Q3-Q1). Data is the same as in **Fig. 1d** except homopolymer TRs are included. Gray denotes H3Africa. Other colors denote 1000 Genomes superpopulations.

Supplementary Fig. 6



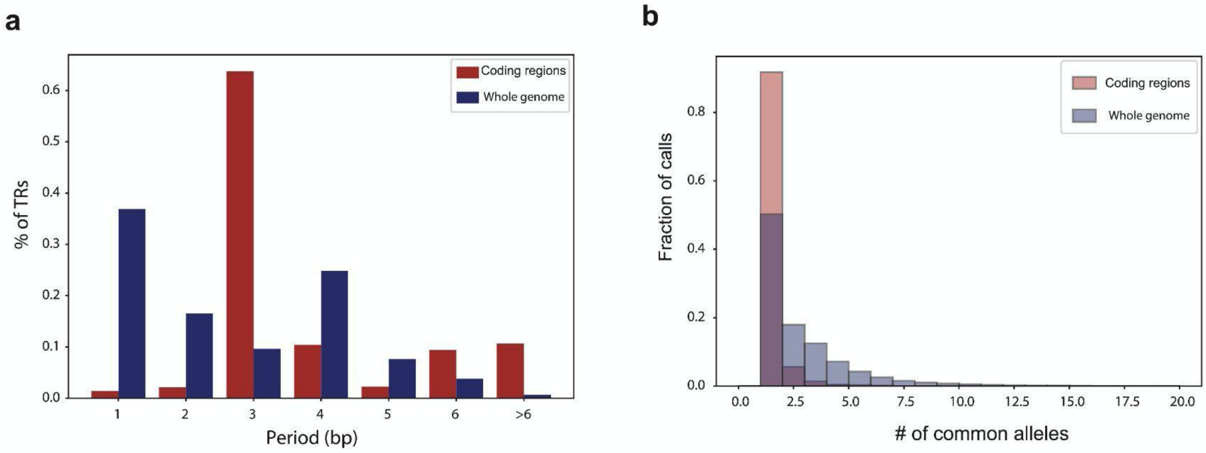
Cumulative number of TR alleles with each additional sample. The y-axis gives the number of unique TR alleles discovered with each additional sample. Samples on the x-axis are sorted by superpopulation. Orange=polymorphic alleles (allele count > 2), purple=doubleton (observed twice), blue=singleton (observed once). Data is based on chromosome 21.

Supplementary Fig. 7



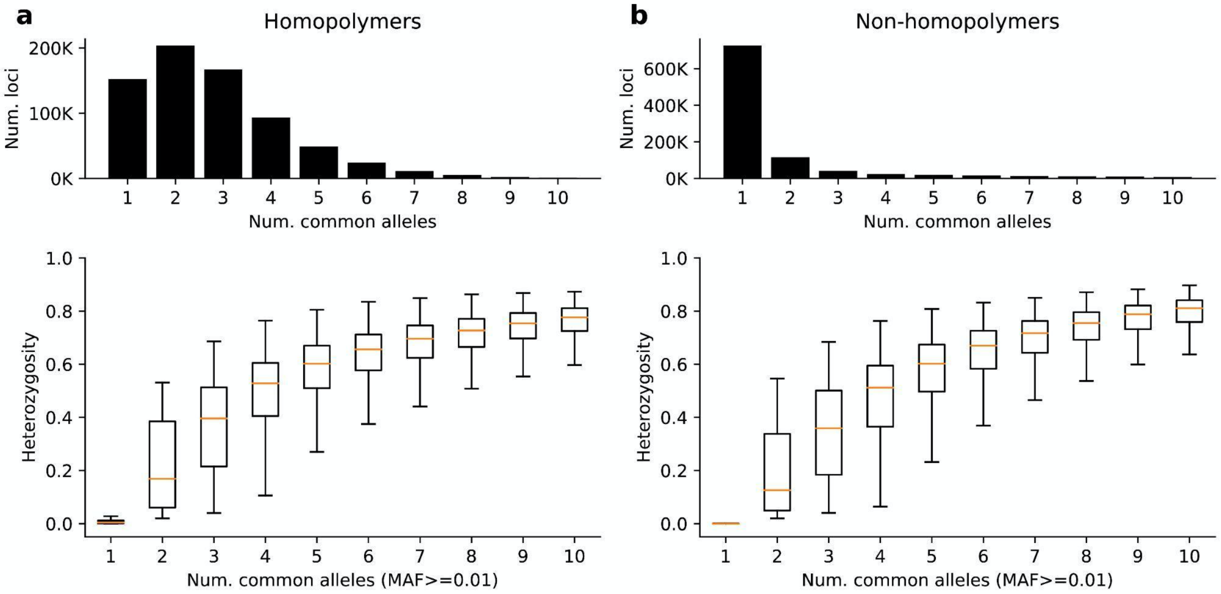
TR variants capture known population structure. a. Principal components analysis of all samples. PCA was performed using TR genotypes on all samples (**Methods**). The projection of each sample along PC1 and PC2 are shown. **b. Principal components analysis of African populations.** A separate PCA was performed on only the H3Africa and 1000GP AFR samples. Samples of similar origin cluster despite being sequenced and genotyped separately (e.g. YRI [Yorubans from Nigeria] from 1000G and Nigeria from H3Africa are clustered together). For **a-b**, each dot represents a single individual. Dots are colored based on the population or superpopulation of origin.

Supplementary Fig. 8



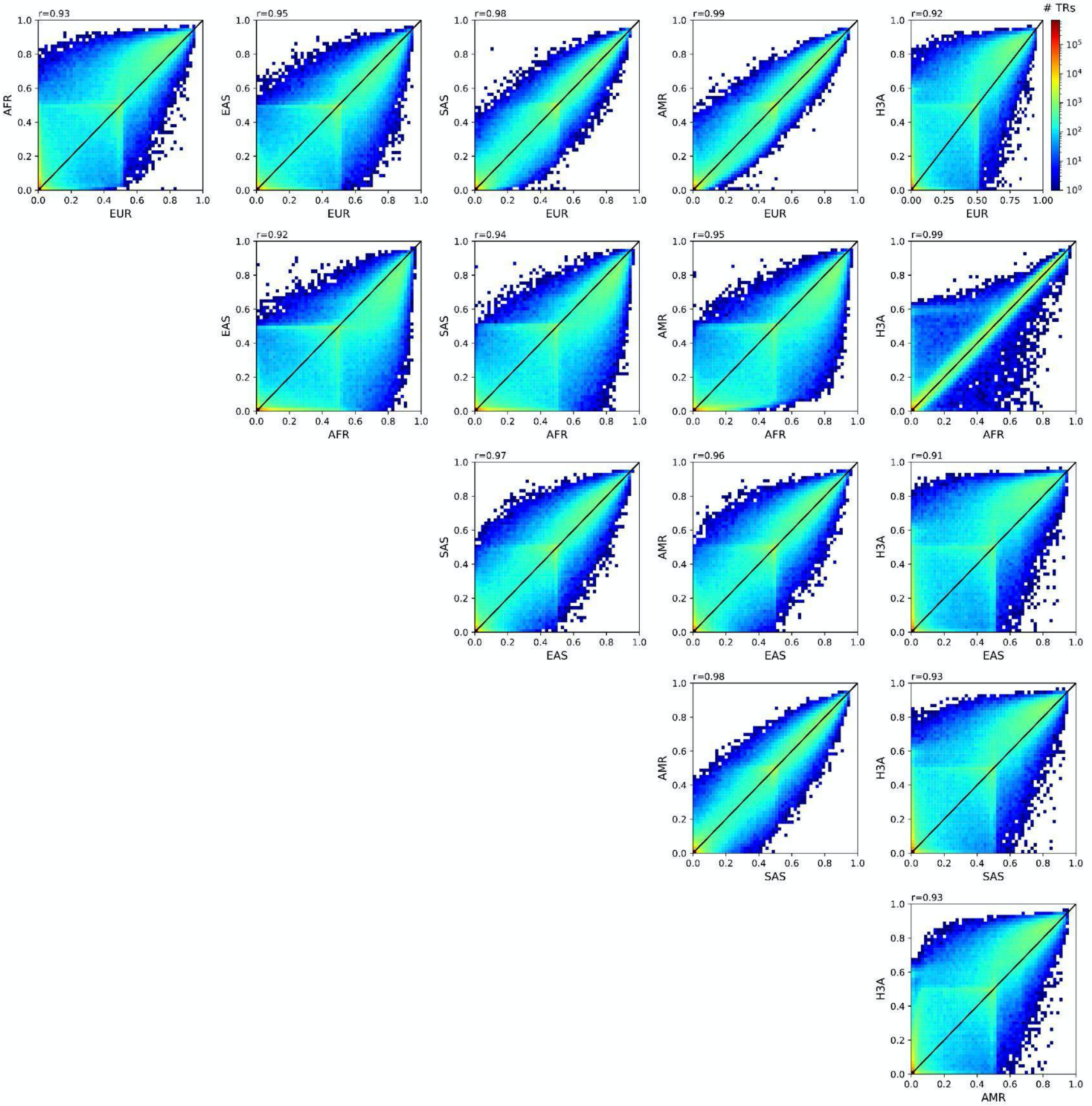
Comparison of TR patterns in coding regions vs. genome-wide. a. Comparison of TR repeat unit length distributions. While homopolymer and tetranucleotides repeats are most prevalent genome-wide, trinucleotide TRs are over-represented in coding regions. **b. Comparison of the distribution of the number of common alleles at each locus.** Common alleles are defined as having a frequency > 1%. As expected, coding TRs are less polymorphic and have fewer common alleles per TR compared to genome-wide TRs. In both panels, red bars denote TRs in coding regions and blue bars denote genome-wide TRs.

Supplementary Fig. 9



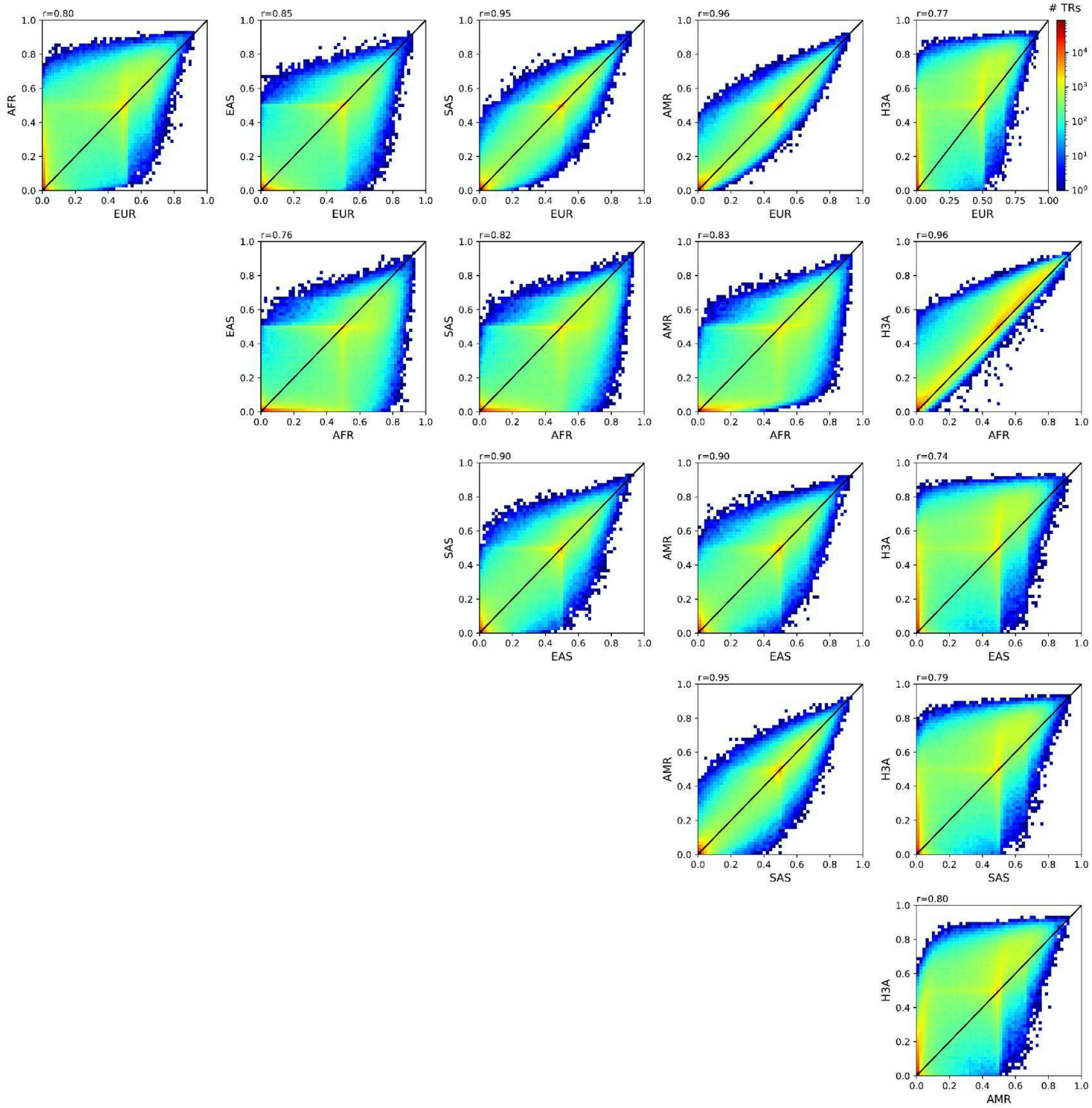
TRs exhibit a range of polymorphism levels across loci. **a.** shows data for only homopolymer TRs, and **b.** shows data for non-homopolymer TRs. Both panels show heterozygosity as a function of the number of common alleles. The y-axis gives the distribution of heterozygosity values for TRs with the number of common alleles specified in the x-axis. Boxplots summarize the distribution of heterozygosity values. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1–1.5*IQR (bottom) and Q3+1.5*IQR (top), where IQR gives the interquartile range (Q3-Q1). In **a-b**, top panels show the number of TRs with each number of common alleles.

Supplementary Fig. 10



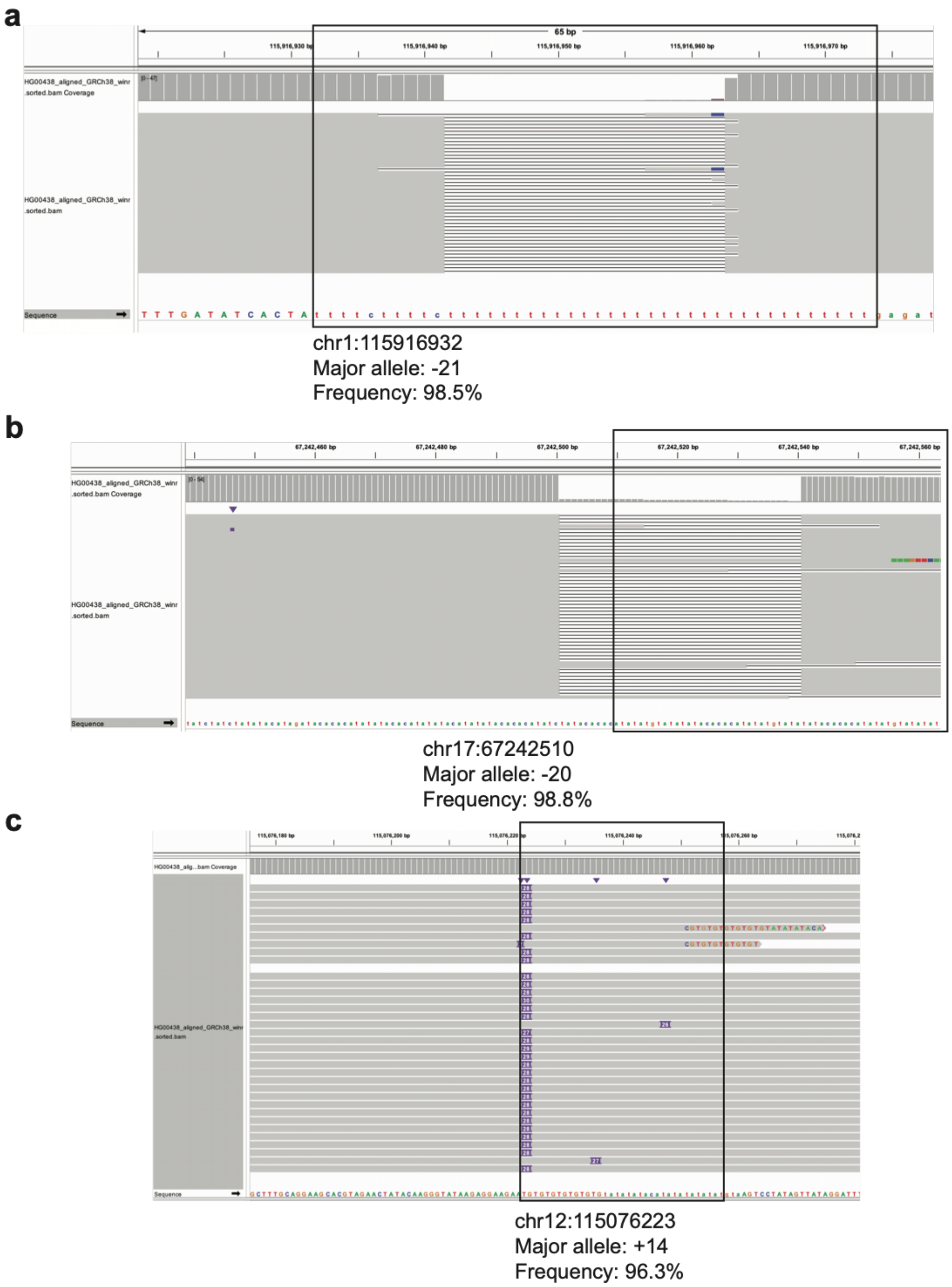
Comparison of TR heterozygosity values across populations. Each scatterplot compares the heterozygosity of each TR across two superpopulations. The black line denotes the diagonal. Heatmap shading indicates the number of TRs falling into each bin. The Pearson correlation comparing heterozygosity values in each pair of superpopulations is annotated above each panel. Data is shown for non-homopolymer TRs.

Supplementary Fig. 11



Comparison of TR heterozygosity values across populations. Each scatterplot compares the heterozygosity of each TR across two superpopulations. The black line denotes the diagonal. Heatmap shading indicates the number of TRs falling into each bin. The Pearson correlation comparing heterozygosity values in each pair of superpopulations is annotated above each panel. Data is shown for homopolymer TRs.

Supplementary Fig. 12


a
65 bp
115,916,930 bp
115,916,940 bp
115,916,950 bp
115,916,960 bp
115,916,970 bp
HG00438_aligned_GRCh38_winr.sorted.bam Coverage
HG00438_aligned_GRCh38_winr.sorted.bam
Sequence
b
67,242,460 bp
67,242,480 bp
67,242,500 bp
67,242,520 bp
67,242,540 bp
67,242,560 bp
HG00438_aligned_GRCh38_winr.sorted.bam Coverage
HG00438_aligned_GRCh38_winr.sorted.bam
Sequence
c
HG00438_alig...bam Coverage
HG00438_aligned_GRCh38_winr.sorted.bam
Sequence

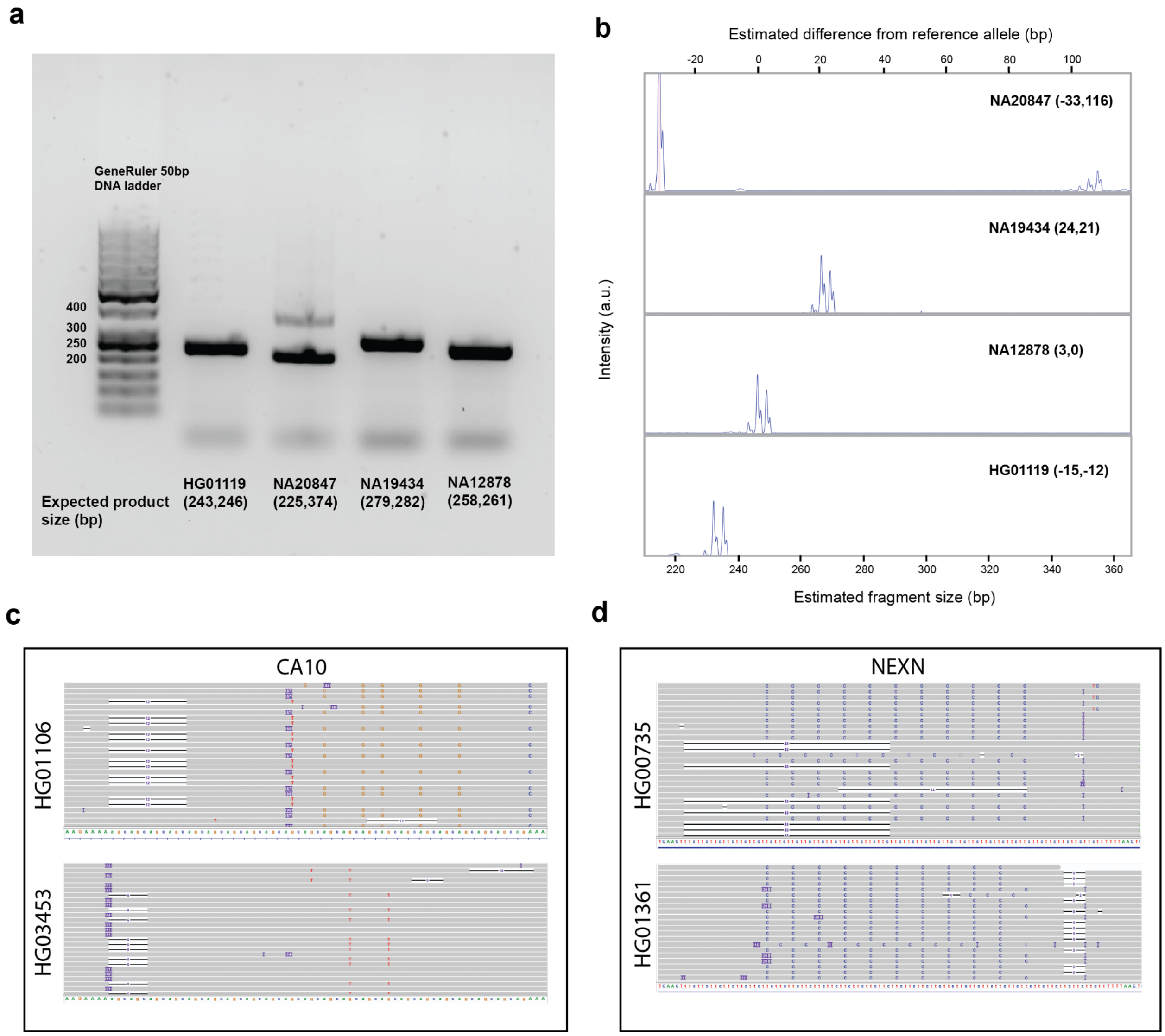
chr1:115916932
Major allele: -21
Frequency: 98.5%

chr17:67242510
Major allele: -20
Frequency: 98.8%

chr12:115076223
Major allele: +14
Frequency: 96.3%

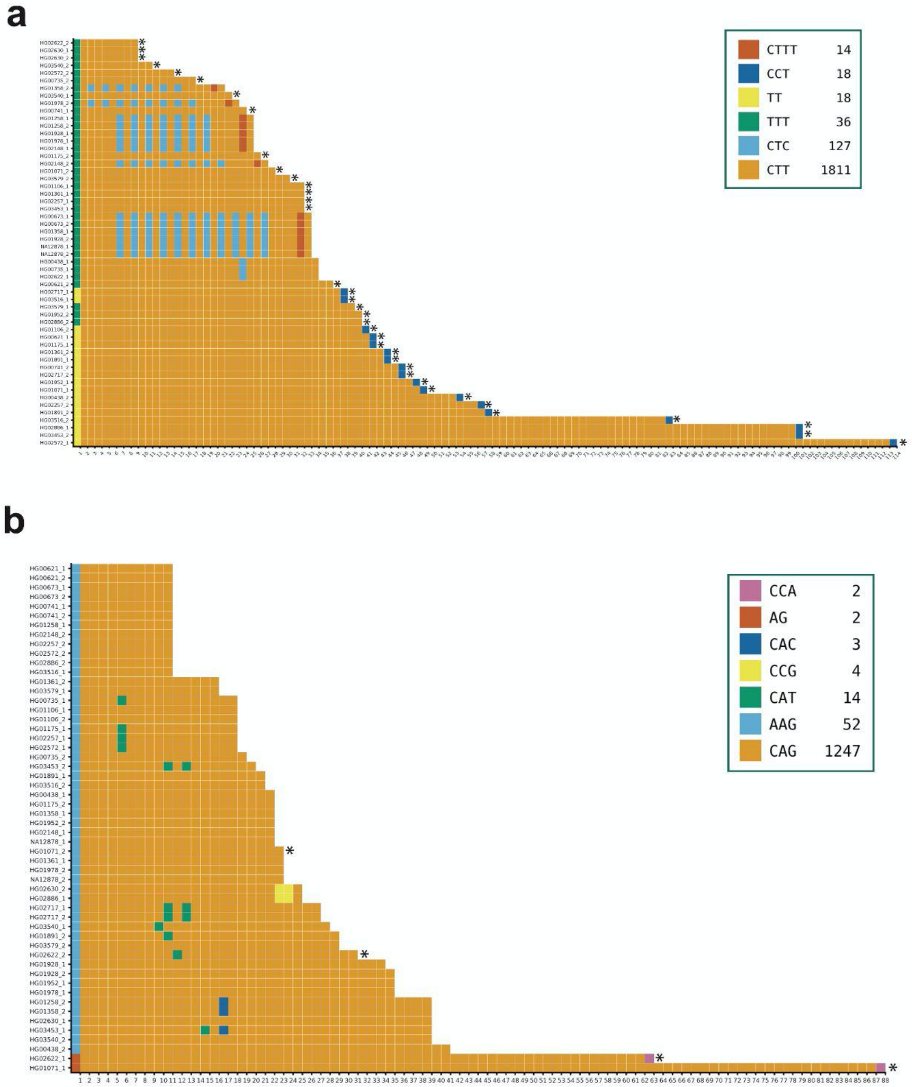
Example TRs for which the majority of observed alleles show dramatic deviations from the hg38 reference. In each panel, the TR region is denoted by a black box. Aligned Pacbio HiFi reads from HPRC for sample HG00438 are visualized using the Integrative Genomics Viewer (IGV). The major allele length (repeat units relative to hg38) and frequency are annotated in each panel.

Supplementary Fig. 13



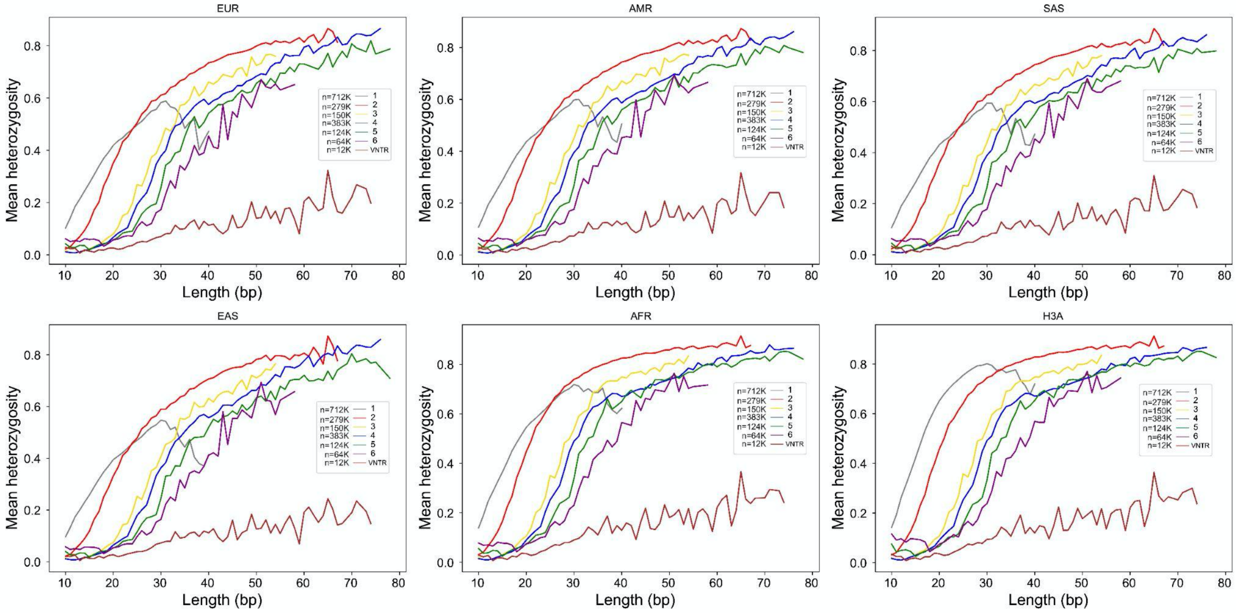
Experimental validation of genotypes at the CAG repeat in an intron of *CA10*. In **a.**, PCR was used to amplify the TR region in four 1000GP samples. Sample names and EnsembleTR calls (expected product size in bp) are annotated below each lane. We identified an expansion in NA20847, which is validated by PCR. Panel **b.** Shows results of fragment analysis on the same PCR products using the GeneMapper software. Samples are annotated with EnsembleTR calls (bp difference from hg38). Panels **c-d** show IGV screenshots of Pacbio Hifi reads aligned to the TRs in CA10 and NEXN for which Africa-specific common repeat expansions were identified.

Supplementary Fig. 14



Allele sequence visualization with TRviz. We used TRviz to visualize allele sequences for 27 1000Genomes samples. **a**. Allele sequences for the repeat upstream of NEXN described in the text are shown. Although it is annotated as a CTT repeat, many observed alleles consist of repeats of the hexamer sequence CTTCTC. For longer alleles, ExpansionHunter calls were chosen by EnsembleTR, therefore sequence imperfections are not reported. **b.** Allele sequences for the intronic repeat in CA10 described in the text are shown. While the repeat is annotated as CAG, other motif sequences such as CAT and CAC are frequently seen. Similar to the NEXN repeat, long alleles are genotyped by ExpansionHunter where the sequence imperfections are not reported. A “*” next to a haplotype indicates it was obtained from ExpansionHunter, and therefore no imperfections could be detected. Colors denote motif sequence.

Supplementary Fig. 15



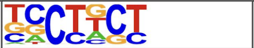
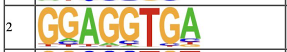
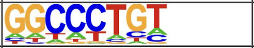
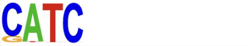
Heterozygosity is correlated with total TR length. The x-axis denotes the length of each TR in the reference genome (in total bp of the longest uninterrupted repeat) and the y-axis shows the mean heterozygosity for TRs with each length. Figures are the same as **Fig. 3a** but plotted separately based on heterozygosity values in each superpopulation. Colors denote different repeat unit lengths. Superpopulation labels are annotated on the top of each panel.

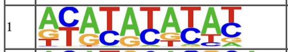
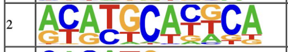
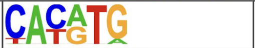
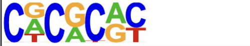
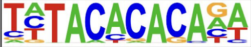
Supplementary Fig. 16

Stable

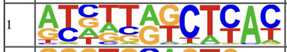
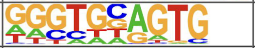
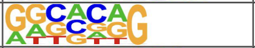
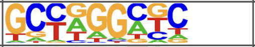
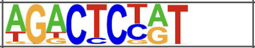
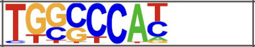
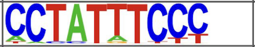
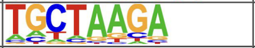
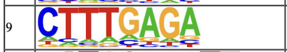
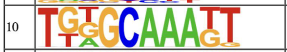
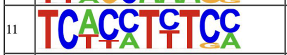
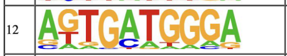
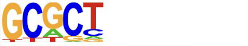
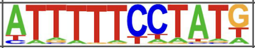
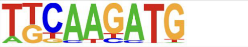
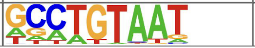
Polymorphic

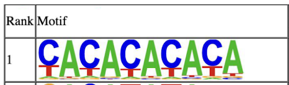
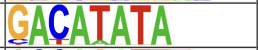
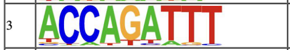
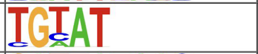
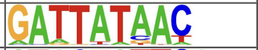
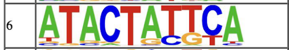
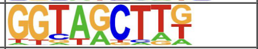
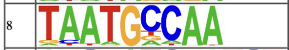
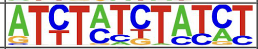
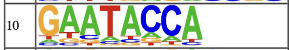
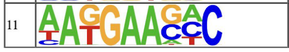
AC

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-19	-4.394e+01	46.41%	36.25%
2		1e-15	-3.482e+01	5.14%	1.76%
3		1e-13	-3.136e+01	19.62%	13.19%
4		1e-13	-3.078e+01	73.16%	65.36%

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-45	-1.044e+02	28.33%	15.28%
2		1e-28	-6.451e+01	21.08%	11.93%
3		1e-21	-4.969e+01	48.39%	37.55%
4		1e-21	-4.968e+01	20.07%	12.15%
5		1e-14	-3.273e+01	6.13%	2.66%

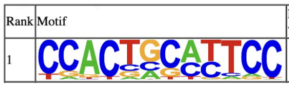
AT

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-23	-5.523e+01	10.90%	5.27%
2		1e-23	-5.431e+01	15.69%	8.93%
3		1e-23	-5.372e+01	39.62%	29.78%
4		1e-22	-5.217e+01	12.67%	6.72%
5		1e-22	-5.180e+01	6.78%	2.60%
6		1e-20	-4.678e+01	6.36%	2.50%
7		1e-18	-4.190e+01	1.15%	0.00%
8		1e-18	-4.150e+01	23.83%	16.60%
9		1e-16	-3.795e+01	19.40%	13.08%
10		1e-15	-3.597e+01	1.93%	0.30%
11		1e-15	-3.590e+01	1.15%	0.03%
12		1e-15	-3.577e+01	4.85%	1.91%
13		1e-15	-3.560e+01	21.53%	15.12%
14		1e-14	-3.448e+01	1.88%	0.30%
15		1e-14	-3.253e+01	2.82%	0.79%
16		1e-14	-3.248e+01	8.39%	4.55%

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-35	-8.107e+01	13.31%	5.74%
2		1e-30	-6.949e+01	22.36%	13.24%
3		1e-17	-4.068e+01	1.58%	0.05%
4		1e-17	-3.977e+01	54.68%	45.83%
5		1e-17	-3.938e+01	1.32%	0.00%
6		1e-16	-3.902e+01	2.60%	0.47%
7		1e-13	-3.048e+01	1.02%	0.00%
8		1e-12	-2.949e+01	0.99%	0.00%
9		1e-12	-2.948e+01	1.98%	0.36%
10		1e-12	-2.817e+01	9.62%	5.68%
11		1e-12	-2.773e+01	2.01%	0.42%

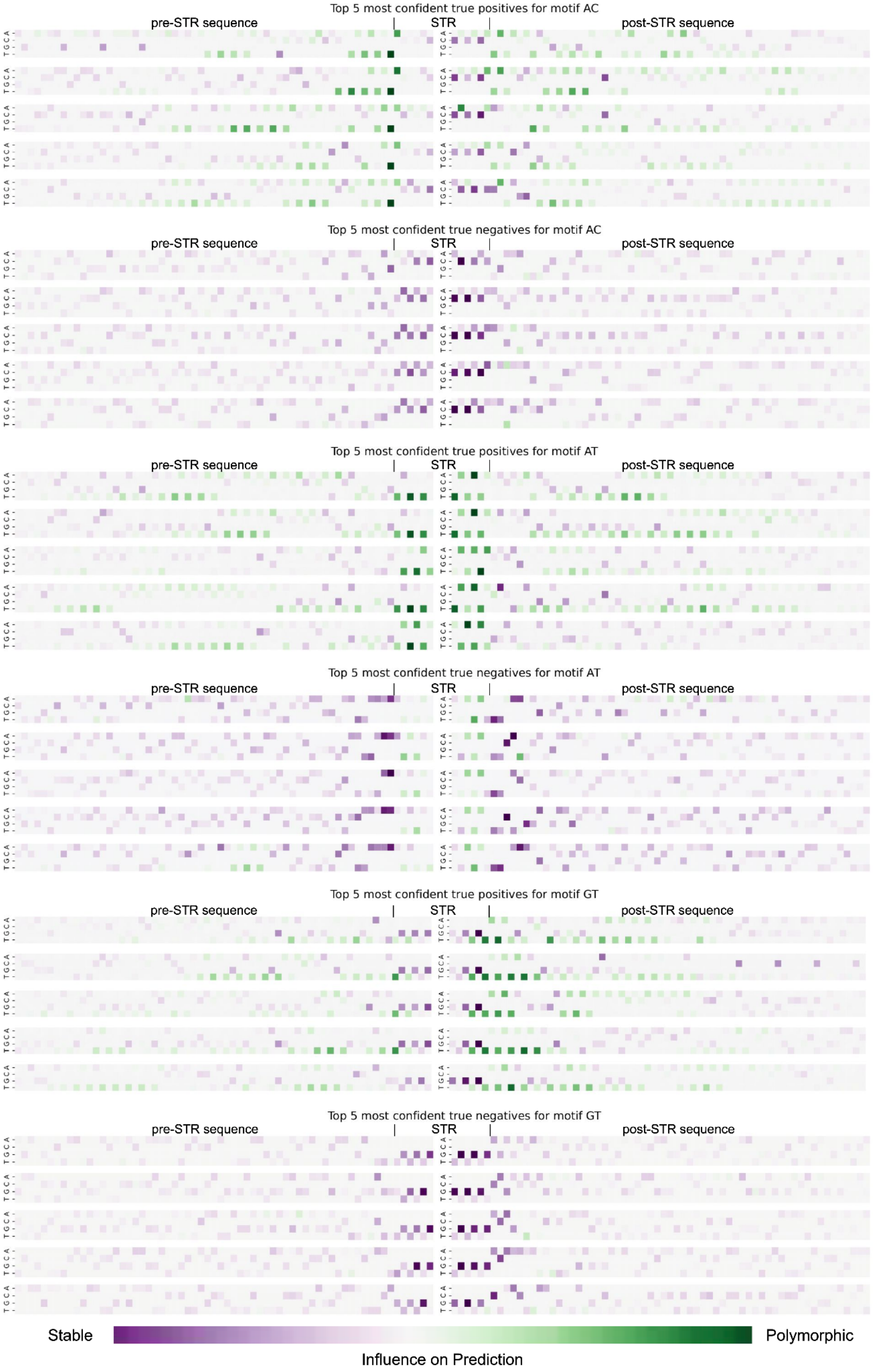
AG

None

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-13	-3.206e+01	5.61%	1.44%

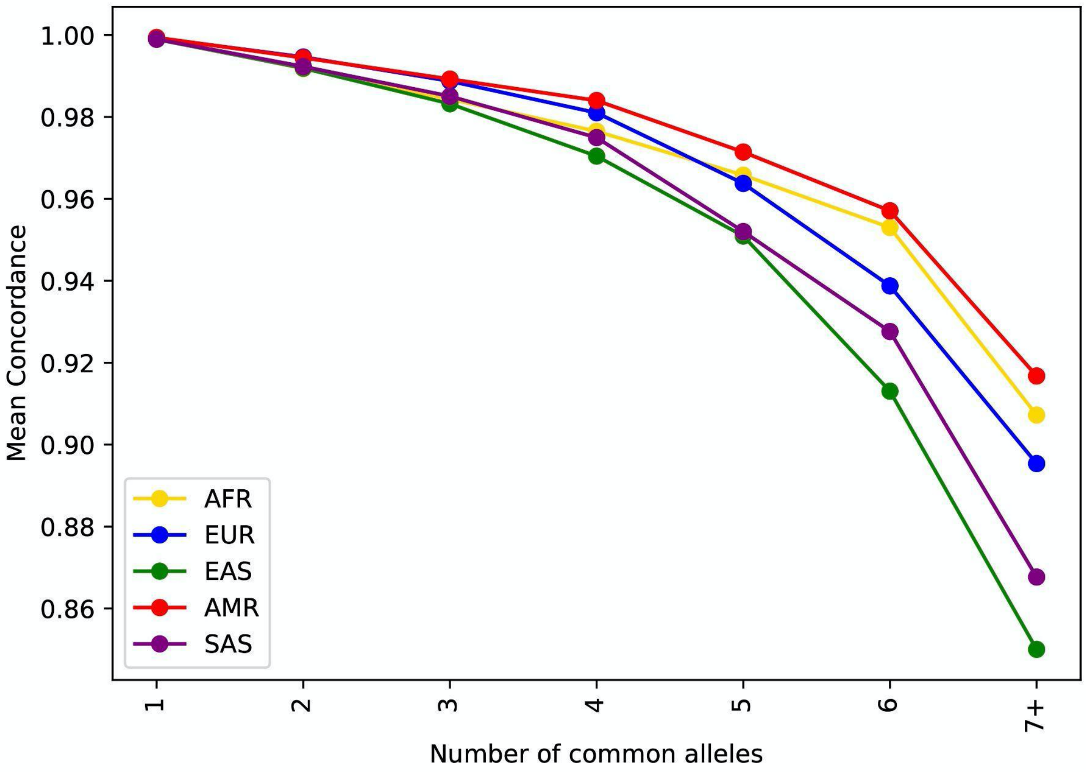
Significant HOMER results not marked as likely false positives. Motifs enriched in the context regions for polymorphic vs. stable were identified using HOMER. Results marked as significant by HOMER are shown. Reverse complements were used such that the AC motif category includes information from all AC and GT motif STRs on the forward strand. P-values are computed using cumulative binomial distributions statistics, are one-sided and are not adjusted for multiple comparisons.

Supplementary Fig. 17



Integrated Gradients attribution scores for most confidently predicted true positive (correctly predicted polymorphic) and true negative (correctly predicted stable) STRs by the CNN model. Each row denotes a different TR. Within each row, the matrix has a row for each nucleotide (A, C, G, T) and a column for each position (centered on the TR). Color denotes the attribution score of each base in each position, where green indicates a base positively contributed towards the model predicting polymorphic and purple indicates contributing towards the model predicting stable.

Supplementary Fig. 18



Imputation accuracy decreases with the number of common alleles. The x-axis denotes the number of common alleles (frequency > 1%) for each TR. The y-axis denotes the mean imputation concordance for TRs in each bin based on a Leave-One-Out analysis on chromosome 21. Colors denote super-populations.