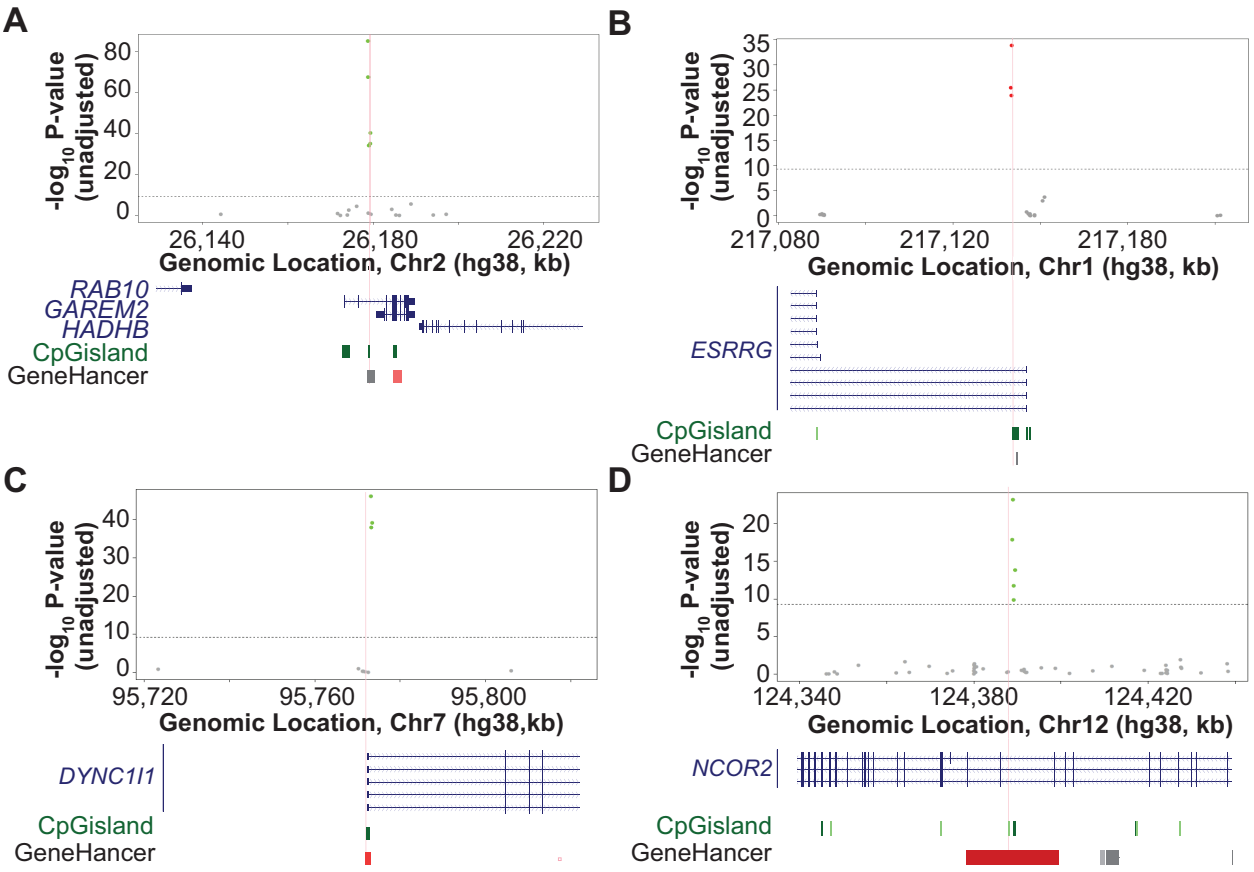
SUPPLEMENTAL INFORMATION

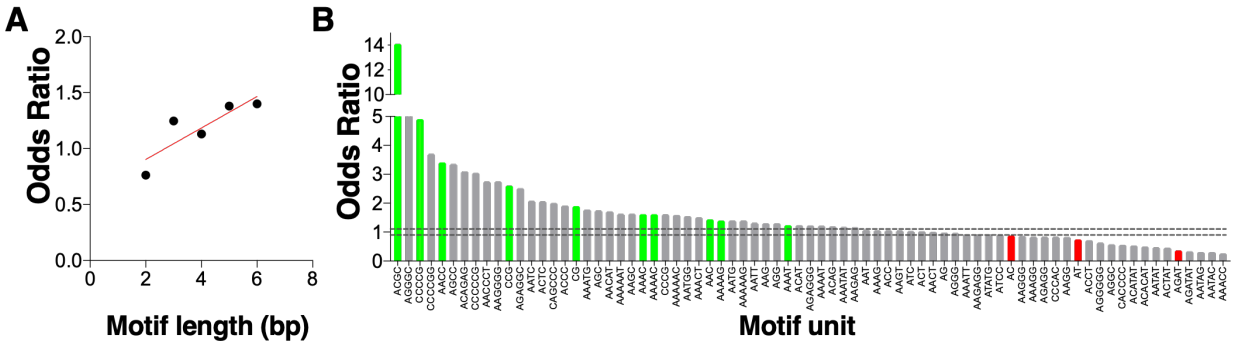
Table of contents

Supplemental Figures

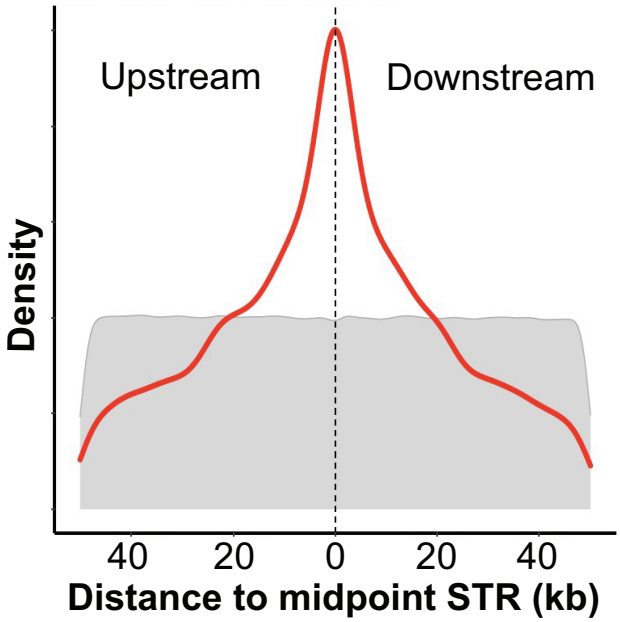
Supplemental Tables



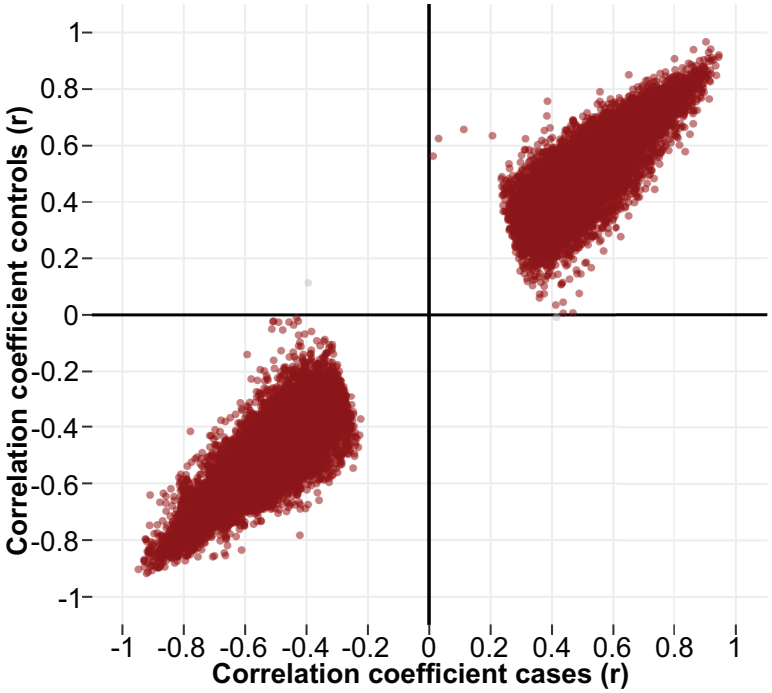
Supplemental Fig S1. Additional examples of *cis* associations between STR lengths and local CpG methylation levels. Each Manhattan plot shows the association signals with CpG methylation ±50 kb of a genotyped STR (**A**, Chr2:26,178,879−26,178,913; **B**, Chr1:217,133,602-217,133,684; **C,** Chr7:95,772,112-95,772,171 and **D,** Chr12:124,388,866−124,388,911). Each circle represents a CpG, with significant associations shown in either red (positive correlations) or green (negative correlations). The location of the STR is shown by the vertical red line. The horizontal grey dashed line indicates the genome-wide significance threshold (Bonferroni adjusted p<0.001). Manhattan plots are annotated with genes (blue), CpG islands (green), enhancers (grey bars) and promoters (red bars) using ncbiRefSeq, cpgIslandExt and geneHancer tracks from the UCSC Genome Browser, respectively.



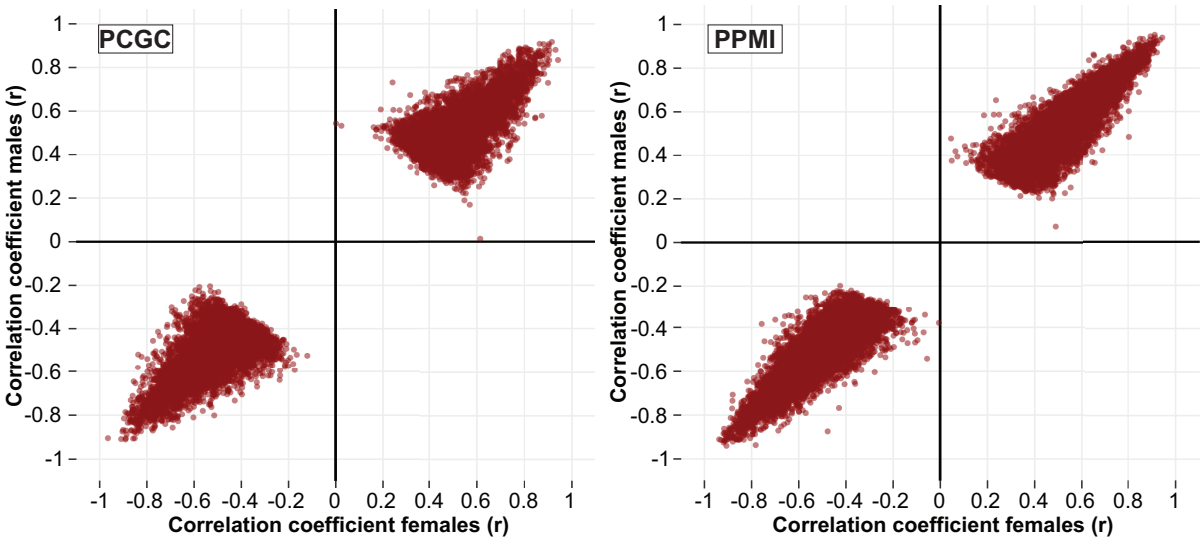
Supplemental Fig S2. Sequence and motif size features of mSTRs. After grouping STRs by motif size or sequence and subsequent annotation with mQTL status (mSTR, STR), enrichment analyses were computed using Fisher's exact test. **(A)** Scatter plot showing a positive correlation between motif size and their relative abundance in our set of mSTRs when compared to the whole set of tested STRs. **(B)** Bar plot showing enrichment for specific motifs (n=75). Significant enrichments (OR>1.1 and Bonferroni-adjusted p<0.05) or depletions (OR<0.9 and Bonferroni-adjusted p<0.05) are shown in green and red colored bars, respectively. Grey bars represent non-significant (Bonferroni-adjusted p<0.05) biases according to the motif sequence. Gray dashed lines indicate OR thresholds for enrichment/depletion.



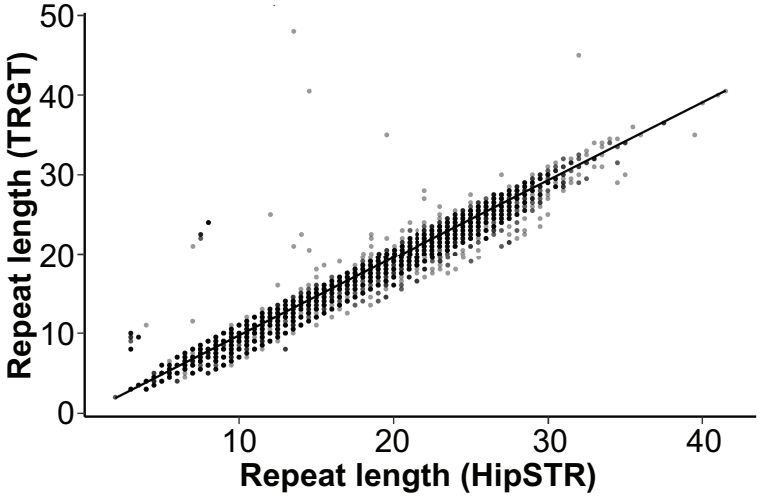
Supplemental Fig S3. Distance between mSTR and targeted CpGs. Density plot showing the distribution of the distance between mSTRs and their associated CpGs (red line) compared to the background distribution of all tested STRs in the mQTL analysis (grey shaded area).



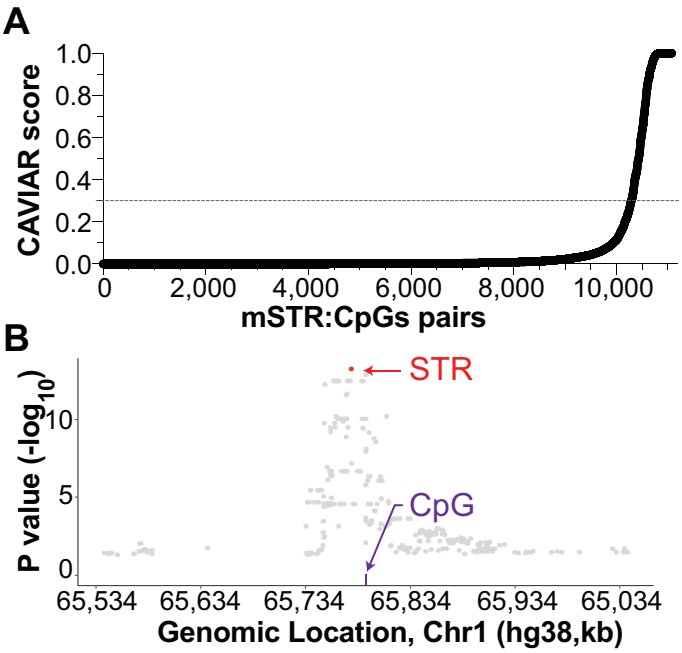
Supplemental Fig S4. Associations between repeat length and DNA methylation are not affected by diagnosis status. Scatter plot showing comparison between the correlation coefficient (*r*) obtained from association analysis using either individuals with PD (n=362) or controls (n=117). Each STR:CpG pair is represented by a single dot, which is color-coded according to the concordance between the direction of the resulting associations (dark red and grey color depicts same and opposite directionality, respectively).



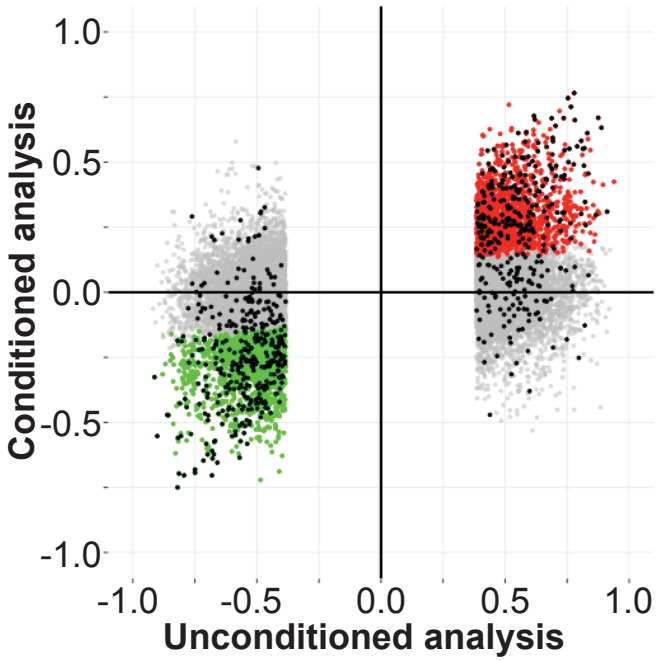
Supplemental Fig S5. Minimal or no effect of sex on the association between STR length and DNA methylation. Scatter plot showing comparison of r correlation coefficient obtained from association analysis using only female (x axis) and male (y axis) individuals in the PCGC (left panel) and PPMI (right panel) cohorts. Briefly, after dividing samples according to their genotypic sex, DNA methylation values were adjusted for the remaining covariates (age, ethnicity, cell type composition), we then tested the resulting residuals for association with STR length. We observed that the vast majority of the associations shared both the same direction and showed similar effect size (Spearman's rank correlation=0.875, $p<2.2x10^{-16}$; PPMI, Spearman's rank correlation=0.960, $p<2.2x10^{-16}$). Each STR:CpG pair is represented by a single dot.



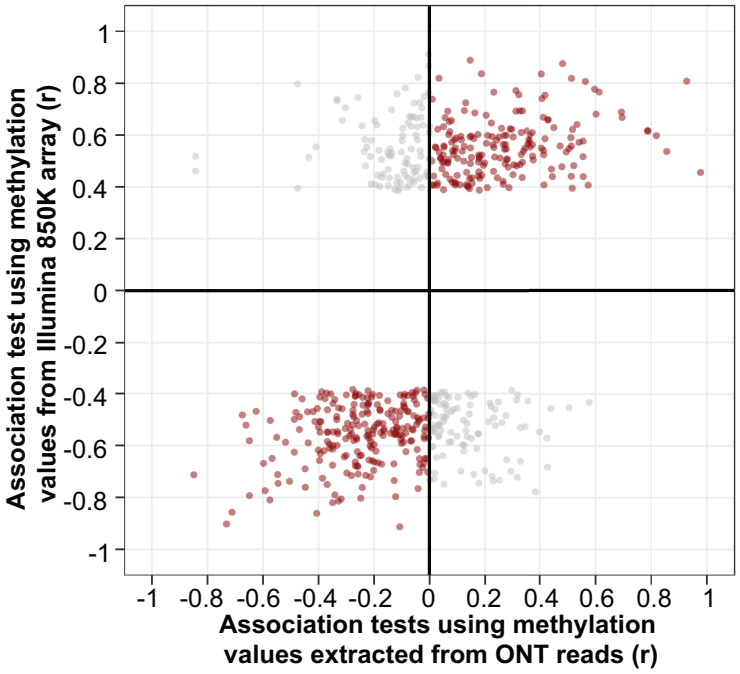
Supplemental Fig S6. HipSTR provides accurate STR genotypes. Scatter plot showing a strong correlation (Spearman's rank correlation coefficient=0.989, $p<0.0001$) between the repeat length of a subset of validated mSTRs (n=4,167) for which both HipSTR and TRGT genotypes derived from PacBio HiFi WGS were available for at least ten individuals. Each dot represents the average allele size of a single STR and sample. Best fit of the data is represented by the black solid line.



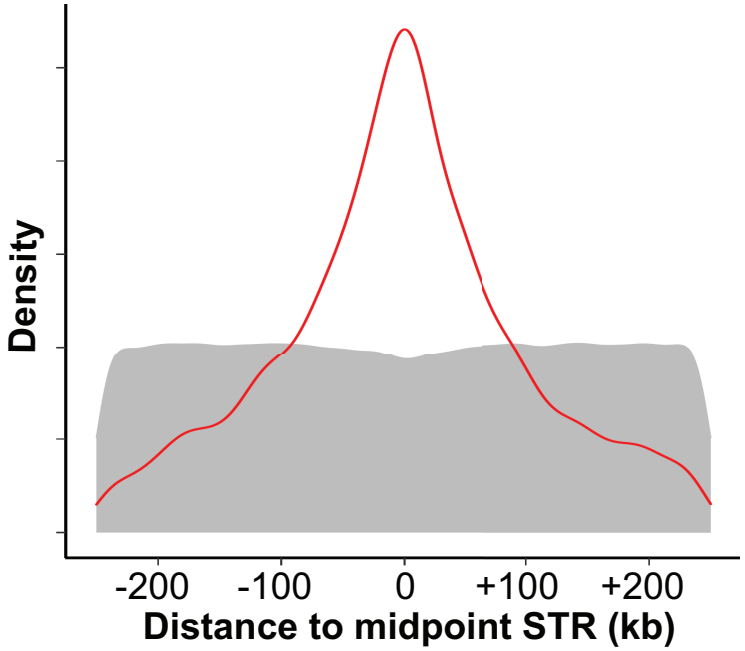
Supplemental Fig S7. Identification of causal STR using CAVIAR. (A) Distribution of posterior probabilities (CAVIAR score, y-axis) across 17,231 replicated mSTR:CpG pairwise associations (x-axis). Individual mSTR:CpG pairs are represented by single black dot and ranked according to CAVIAR score. mSTRs were considered as the causal variant when CAVIAR score is >0.3 (grey dashed line) and show the highest causal probability among all the tested variants at each locus. **(B)** Regional Manhattan plot for associations between genetic variants and β values of the CpG located at Chr1:65,792,363 (Probe ID cg22488256) against genomic location (CpG ±250 kb). According to CAVIAR, the CT repeat (Chr1:65,791,915-65,791,986, hg38) located at the promoter of *PDE4B* is the causal variant responsible for the phenotype, and is also the variant showing the strongest association with CpG methylation. Individual variants are indicated by dots, which are color-coded according to causality (red, causal variant). Location of the CpG is indicated by a purple line.



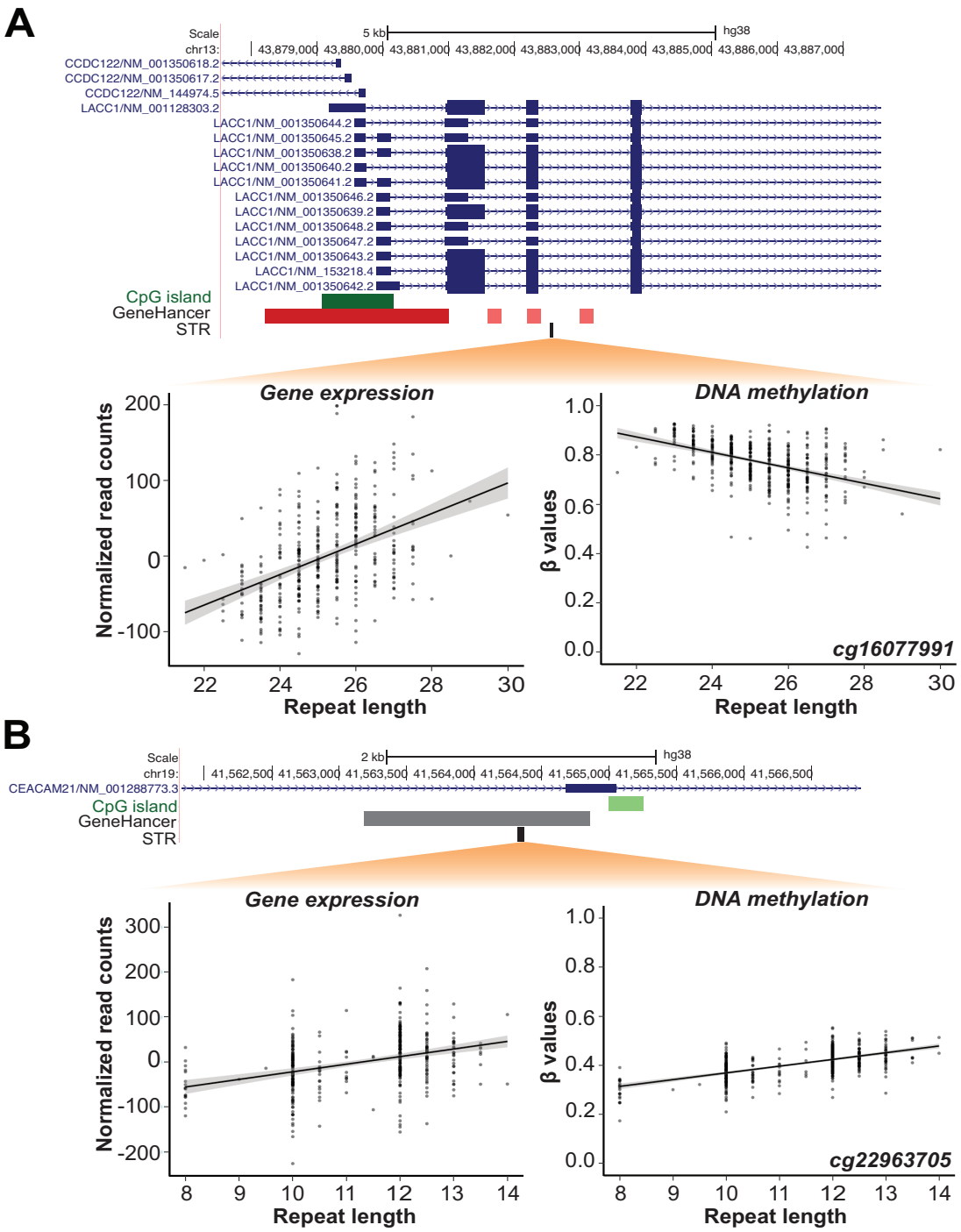
Supplemental Fig S8. Comparison of conditional analysis and statistical fine-mapping. STR:CpG associations where STR is predicted to be the causal variant by CAVIAR (black dots) are projected over Fig. 4A. In 68% of the STR:CpG associations where the STR represents the causal genetic variant according to CAVIAR, conditional analysis indicated that the responsible STR acts as independent signal.



Supplemental Fig S9. Replication of a subset of fm-mSTRs using DNA methylation data obtained from Illumina 850K array and long read Oxford Nanopore Technology (ONT) whole genome sequencing data. Scatter plot showing comparison between the correlation coefficient (*r*) obtained from association analysis using DNA methylation originated from Illumina 850K array and ONT reads (y axis). Each STR:CpG pair is represented by a single dot, which is color-coded accordingly to the concordance between the direction of the resulting associations (dark red and grey color depicts same and opposite directionality, respectively).

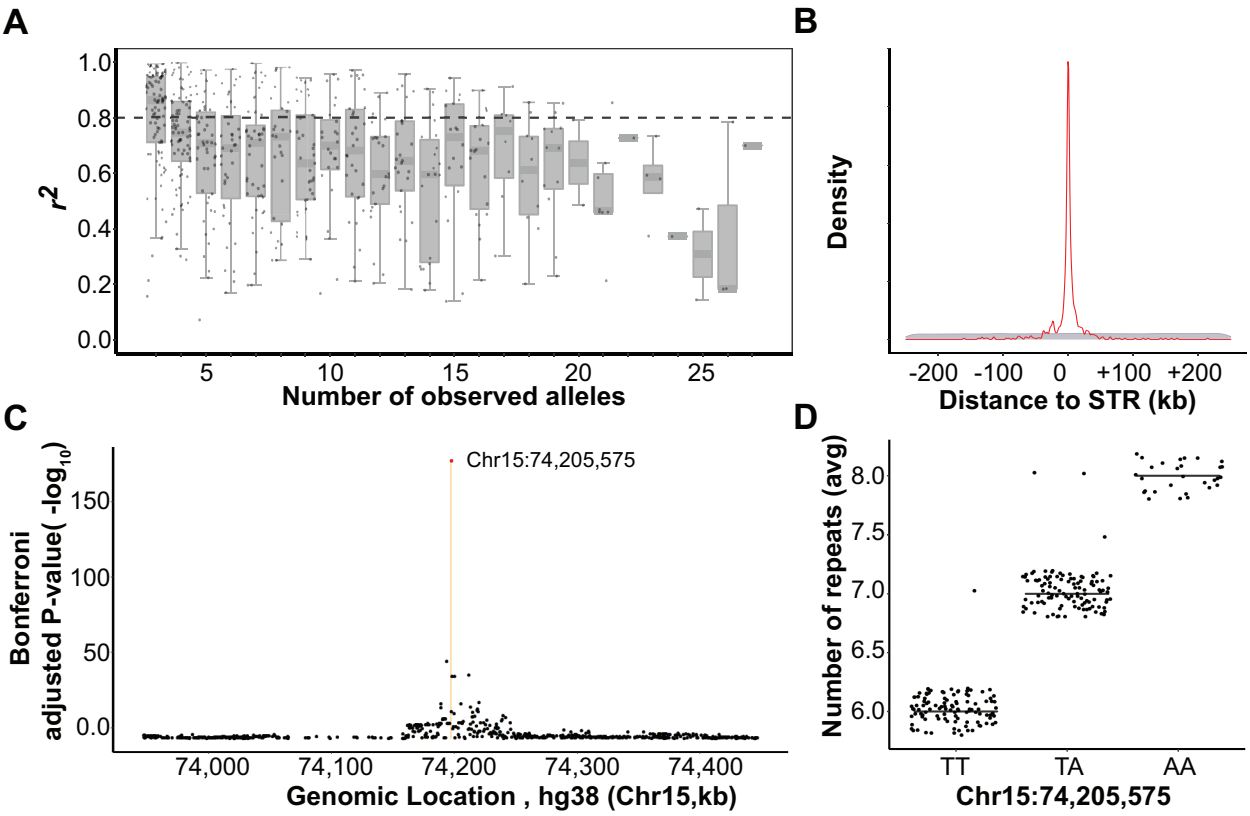


Supplemental Fig S10. Distance between eSTR and TSS of target gene. Density plot showing the distribution of the distance between eSTRs (Bonferroni adjusted <0. 001) and their associated gene (red line) compared to the background distribution of all tested STRs in the eQTL analysis (grey shaded area).



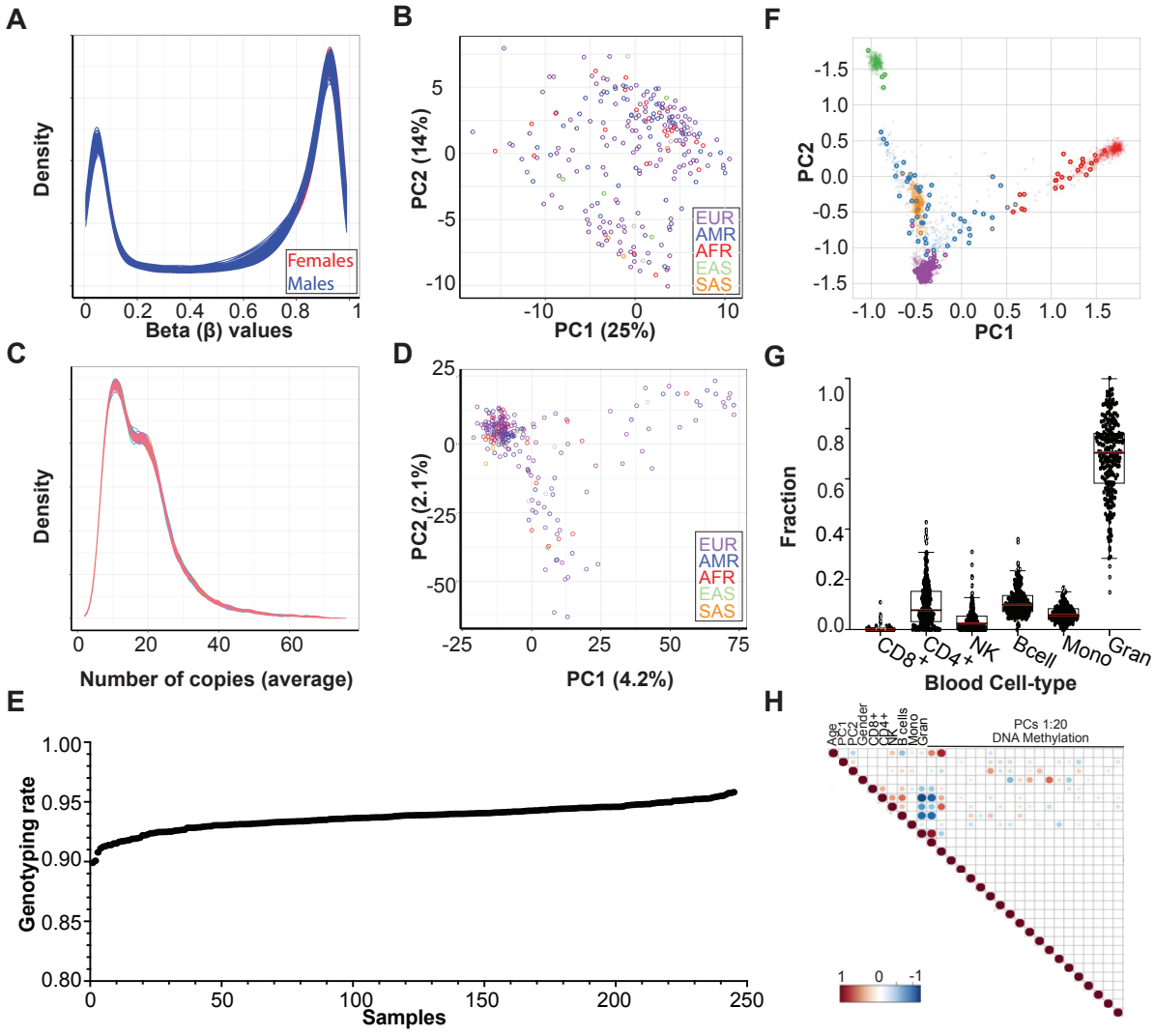
Supplemental Fig S11. Additional examples of functional STRs that regulate both DNA methylation and gene expression levels. (A) and **(B)** show associations of the length of fm-mSTRs (x axis, average of allelic size) with expression (right plot, y axis) and DNA methylation levels (left plot, y axis) of the target CpGs and gene, respectively. **(A)** Scatter plot showing an

inverse correlation between the length of an intronic AT repeat (Chr13:43,882,543:43,882,586) and methylation levels of CpGs present in the promoter region of *CCDCC122*, which results in increased expression of this gene. **(B)** Scatter plot showing an intronic ATTT repeat (Chr19:41,564,325-41,564,372) present within an enhancer element that modulates positively both the DNA methylation levels and expression of the overlapping gene *CEACAM21*. In scatter plots, individual samples are represented by single dots. Diagonal line indicates the best-fit across all data points. Genomic location for STR (black rectangle), gene (blue), CpG islands (green) and enhancer (grey bar) and promoter (red bar) regions according to GeneHancer annotation are shown above the plots.



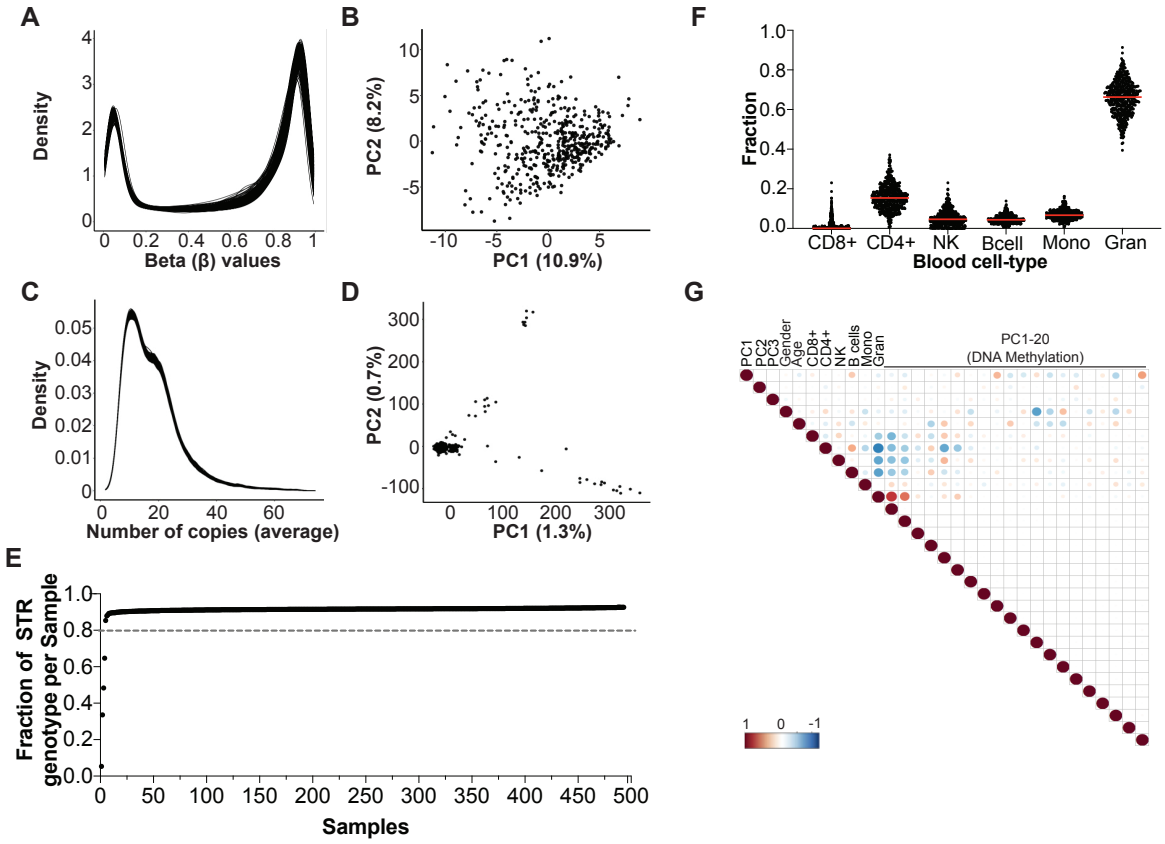
Supplemental Fig S12. LD structure of fm-mSTRs. Box plots showing the distribution of correlations (r^2, y axis) between fm-mSTRs (n=585) and their corresponding strongest associated SNVs (tagging SNV). STRs have been grouped according to the number of alleles observed in the discovery cohort (x axis). Pairwise Pearson correlation coefficients were calculated between the average STR allele size and dosage of the alternate SNV allele. Overall, fm-mSTRs with low number of alleles show higher correlation coefficient than more highly polymorphic STRs. Dashed line indicates threshold to be considered a strong correlation between STR and SNV genotypes. Boxes indicate Q1, Q3 and median, whiskers show 1.5x the interquartile range. (**B)** Density plot showing distance between best tagging SNV and STR midpoint (red line) compared to the whole set of tested SNV (grey shaded area). **(C)** Regional Manhattan plot showing associations between SNVs and fm-mSTR genotypes (Chr15:74,204,941-74,204,964). Each dot represents a single SNV. The best tagging SNV for the STR (Chr15:74,205,575) is represented by a red dot. Fm-

mSTR location is indicated by a vertical gold bar. **(D)** Distribution of STR length for the Chr15:74,204,941-74,204,964 STR according to its best tagging SNV genotype (Chr15:74,205,575). Black line indicates the median of STR length for each group.



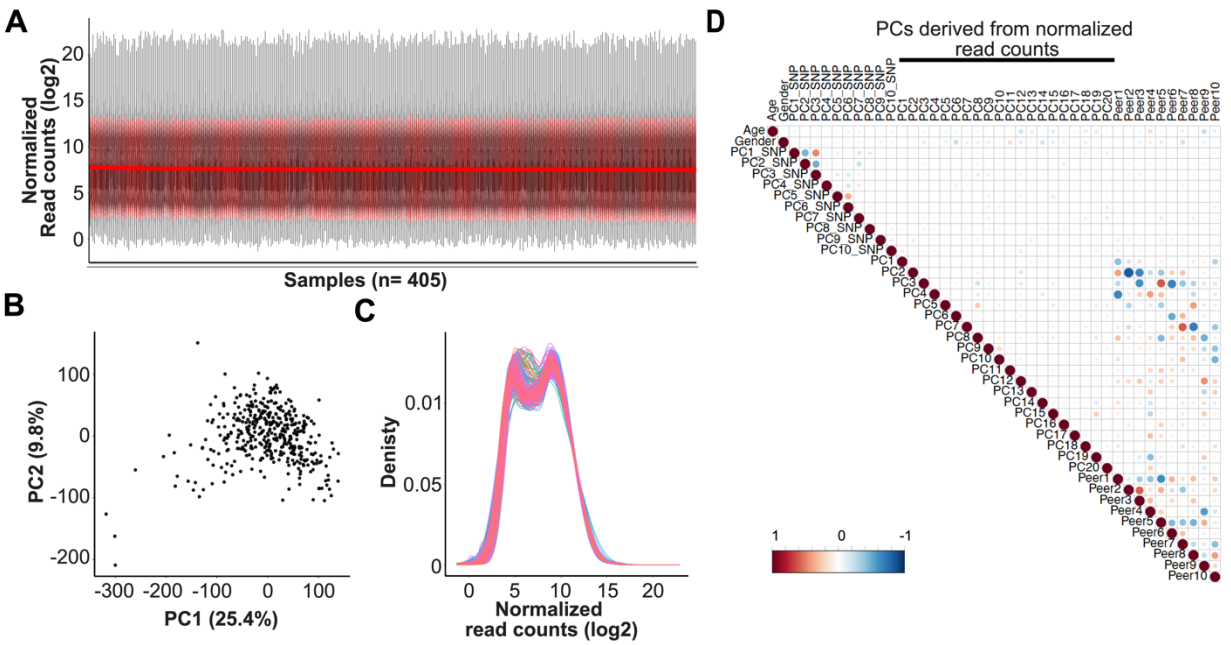
Supplemental Fig S13. Quality control of DNA methylation and HipSTR-derived genotypes in 245 samples collected by the PCGC consortium. (A) Density plot based on β values of CpGs present on Chromosome 1 that shows the expected bimodal distribution (0, unmethylated; 1, methylated) of β values in the human genome. Individual samples are represented by single lines, which are color-coded according to gender. **(B)** Scatter plot showing PC1 against PC2 calculated from β values of CpGs present on Chromosome 1. Individual samples are depicted by single circles, which are color-coded according to reported ancestry. **(C)** Density plot showing the distribution of the average of the allele size obtained from HipSTR for STRs present on

Chromosome 1 where single samples are depicted by individual lines. **(D)** Scatter plots of PC obtained from the average of diploid STR genotypes obtained from tested STRs present on Chromosome 1. Individual samples are color-coded according to ancestry. **(E)** Fraction of STRs with available genotypes across the selected 245 samples. Overall, all samples showed a genotyping rate >90%. Each dot represents a single sample. **(F)** Scatter plot showing PCs obtained from SNV genotypes using Peddy projected over the PCs of the thousand genomes samples. **(G)** Scatter plot showing distribution of the methylation-inferred fraction of the six main cell type present in blood across PCGC cohort. Overlay box plot show the distribution of values for each cell fraction, where box shows Q1 and Q3 and medium (red line). Whiskers show 1.5* Interquartile range. **(H)** Correlation matrix showing associations across potential covariates for the PCGC samples. Covariates include age, ancestry accounted by top two PCs obtained from SNVs, fraction of the six main types of blood cell types inferred from DNA methylation measurements and PCs obtained from β values. Direction and strength of the association are represented by the dot size and color (red, positive; blue, negative). For example, we observed a strong correlation between age and PC2.

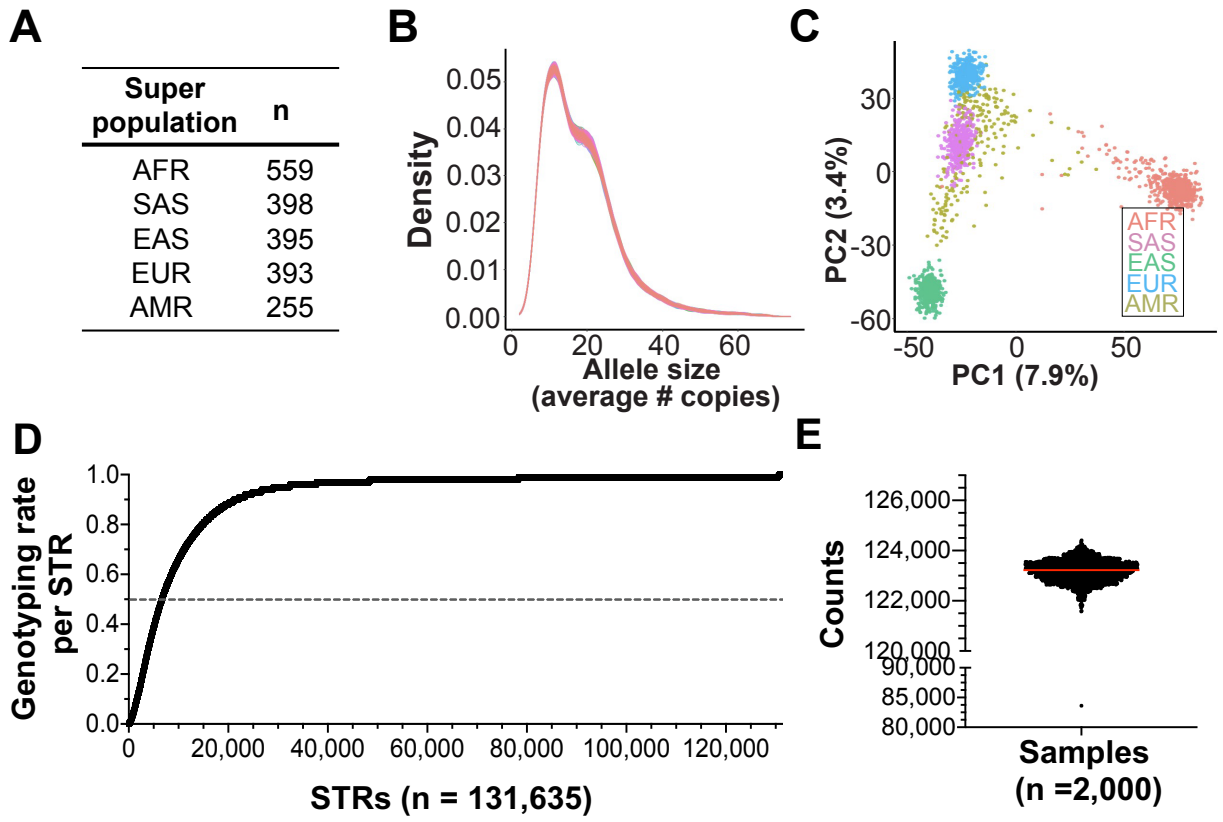


Supplemental Fig S14. Quality control of DNA methylation and HipSTR-derived STR genotypes for the quality-filtered 484 samples collected from the PPMI cohort. Density **(A)** and scatter **(B)** plots showing PC1 against PC2 obtained from β values from CpGs present on Chromosome 1. **(C)** Density plot showing the distribution of the average of the allele size obtained from HipSTR for STRs present on Chromosome 1. **(D)** Scatter plot showing PC1 against PC2 obtained from average of diploid STR genotypes. Samples are clustered according to their ancestry, however the PCs explain a very small fraction of the variance and were not included as covariates in association analysis. Variance explained by each PC is shown within parentheses. (**E)** Distribution of genotyping rate of tested STRs (n=131,635), *i.e.* fraction of STRs with available genotype, per sample, sorted low to high. Every dot represents a single sample. (**F)** Scatter plots shaped according to the frequency distribution of the inferred fraction (y axis) of the six main cell types present in blood (x axis) in the PPMI cohort. Red line depicts median per group. **(G)**

Correlation matrix showing the relationship across the potential covariates identified in the 484 samples collected from the PPMI cohort. These covariates include ancestry (represented by the three first principal components based on SNVs), gender, age, fraction of the six main types of blood cell types inferred from DNA methylation measurements and PCs obtained from β values. Direction and strength of the association are represented by the dot size and color (red, positive; blue, negative). For example, PC1 (DNA methylation) is strongly correlated with granulocyte composition.



Supplemental Fig S15. Quality control of RNA sequencing data for quality-filtered 405 samples collected from the PPMI cohort. (A) Box plots showing distribution of expression data per sample (x axis). Median expression per sample is indicated by a single red dot. While red bars represent median ±2 standard deviation, grey bars represent 1.5 IQR of the gene expression levels per sample **(B)** Scatter plot showing PC1 against PC2 obtained from normalized read counts for 18,861 autosomal genes. Individual samples are represented by single dots. **(C)** Density plot showing the distribution of normalized read counts (log2 scale) across the tested samples. Individual samples are represented by a single colored line. **(D)** Correlation matrix showing the relationship across the potential covariates identified in the tested samples. These covariates include age, gender, ancestry (represented by the ten first principal components based on SNVs), PCs and PEER factors obtained from normalized read counts. Direction and strength of the association are represented by the dot size and color (red, positive; blue, negative). For example, PC2 shows a strong negative correlation with PEER factor 2.



Supplemental Fig S16. Quality control of HipSTR-derived estimates generated from PCR-free Illumina sequencing data from 2,000 samples collected by the 1KG project (A) Table showing the number of samples collected from the 1KG project corresponding to each of the five tested super-populations. **(B)** Density and **(C)** scatter plots of PCs obtained from the average of diploid genotypes for STRs present on Chromosome 1. Individual samples are represented by single line or color-coded (ancestry) circles in density and scatter plots, respectively. **(D)** Scatter plot showing genotyping rate per STR. Each dot represents an STR. **(E)** Violin plot showing number of available STR genotypes per sample after QC.

Supplemental Tables

Supplemental Table S2. Genomic location and feature of mSTRs that influence DNA methylation levels of multiple CpGs.

Genomic Region	n mSTRs
Promoter region (TSS±2 kb)	343
Enhancer (GeneHancer)	290
Coding regions	10
5', 3'UTR	110
Intronic	1,778
Intergenic	1,556

Supplemental Table S5. Genomic location and feature of fine-mapped mSTRs.

Genomic Region	n fm-mSTRs	Odds ratio	Fisher exact P-value
Promoter region (TSS±2 kb)	118	7.58	1.05×10^{-11}
Enhancer (GeneHancer)	55	3.03	1.05×10^{-11}
Coding regions	3	5.00	0.024
5'UTR, 3' UTR	41	5.55	2.20×10^{-16}
Intronic	273	0.84	0.3595
Intergenic	268	0.73	2.20×10^{-16}

The remaining Supplemental Tables (S1, S3, S4, S6, S7 and S8) are available as separate online Excel files.