The American Journal of Human Genetics, Volume *108*

Supplemental information

Pervasive *cis* effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression

Paras Garg, Alejandro Martin-Trujillo, Oscar L. Rodriguez, Scott J. Gies, Elina Hadelia, Bharati Jadhav, Miten Jain, Benedict Paten, and Andrew J. Sharp

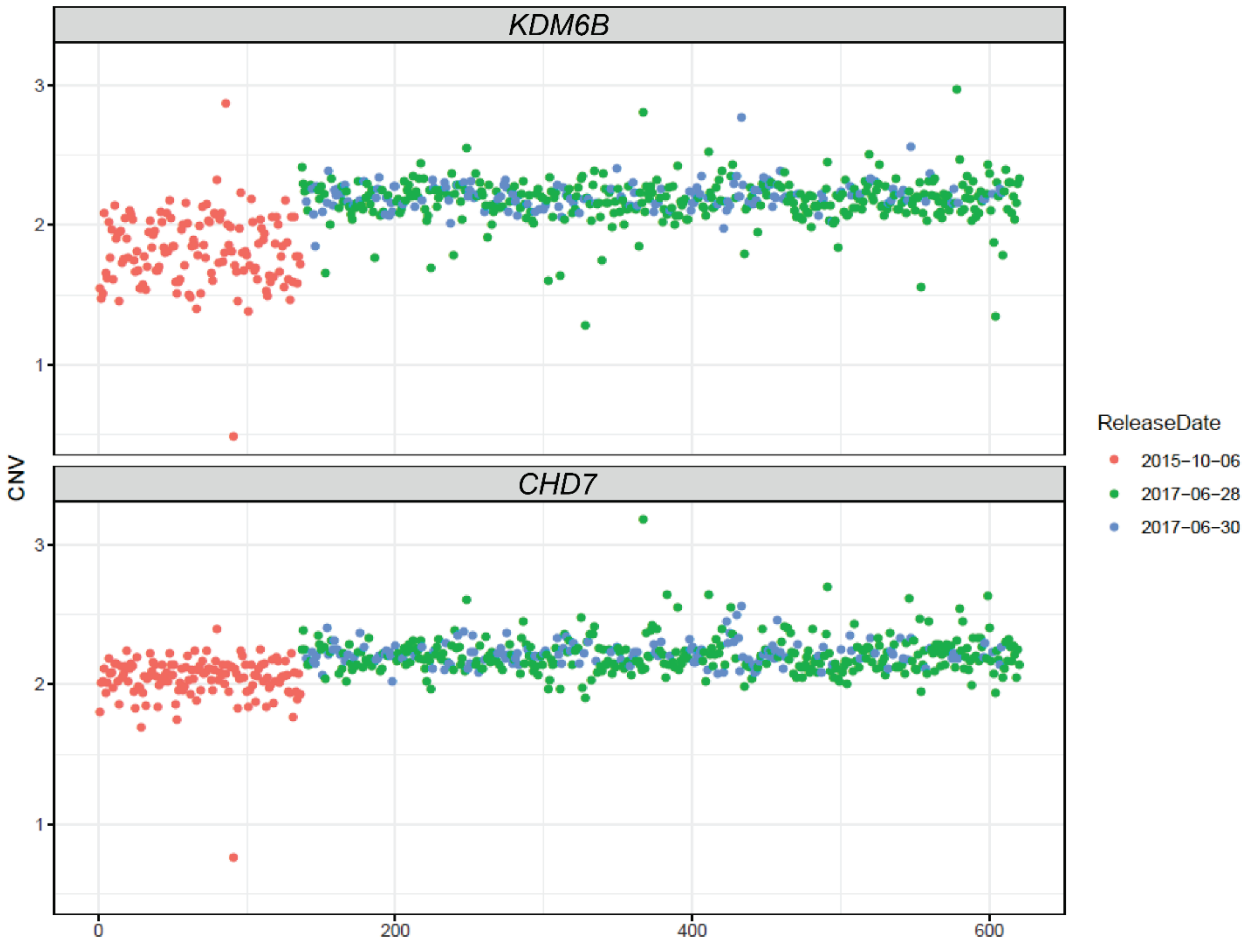

Figure S1. Batch effects in *CNVnator* copy number estimates from WGS in the GTEx cohort. Using *CNVnator* data for two highly constrained genes that should remain copy-number invariant in the normal population (*KDM6B* [MIM: 611577], chr17:7,839,904-7,854,796 and *CHD7* [MIM: 608892], chr8:60,678,740-60,868,028, hg38), we observed a strong batch effect in the GTEx cohort, whereby copy numbers derived from WGS data with release date October 6th 2015 (*red points*) were systematically shifted compared to later data releases (*blue and green points*). Based on these observations, we removed from further analysis all 135 samples with release October 6th 2015.

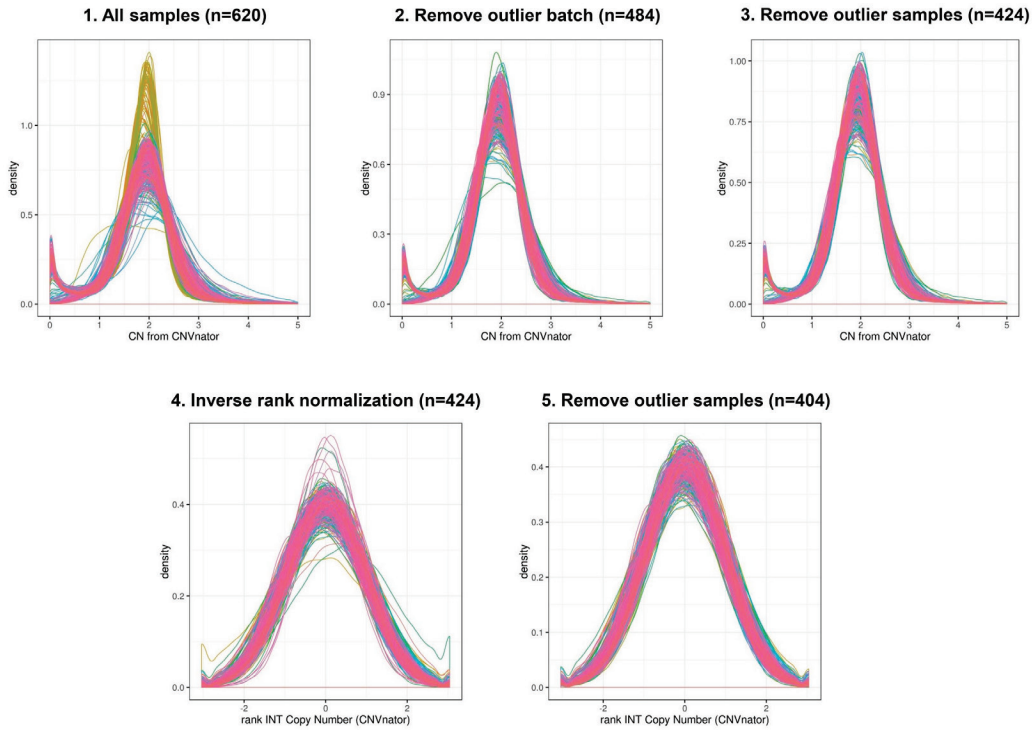

Figure S2. VNTR copy number distributions within the GTEx cohort after sample filtering and data normalization. Based on *CNVnator* copy number estimates of 89,893 autosomal -VNTRs, we generated density plots at each step of quality control and normalization. We initially analyzed WGS data from 620 individuals, but after removal of batch effects, samples that were consistent outliers at invariant constrained genomic loci, or outliers for VNTR copy number by principal component analysis and density plots, we used a final cohort of 404 samples in our analysis.

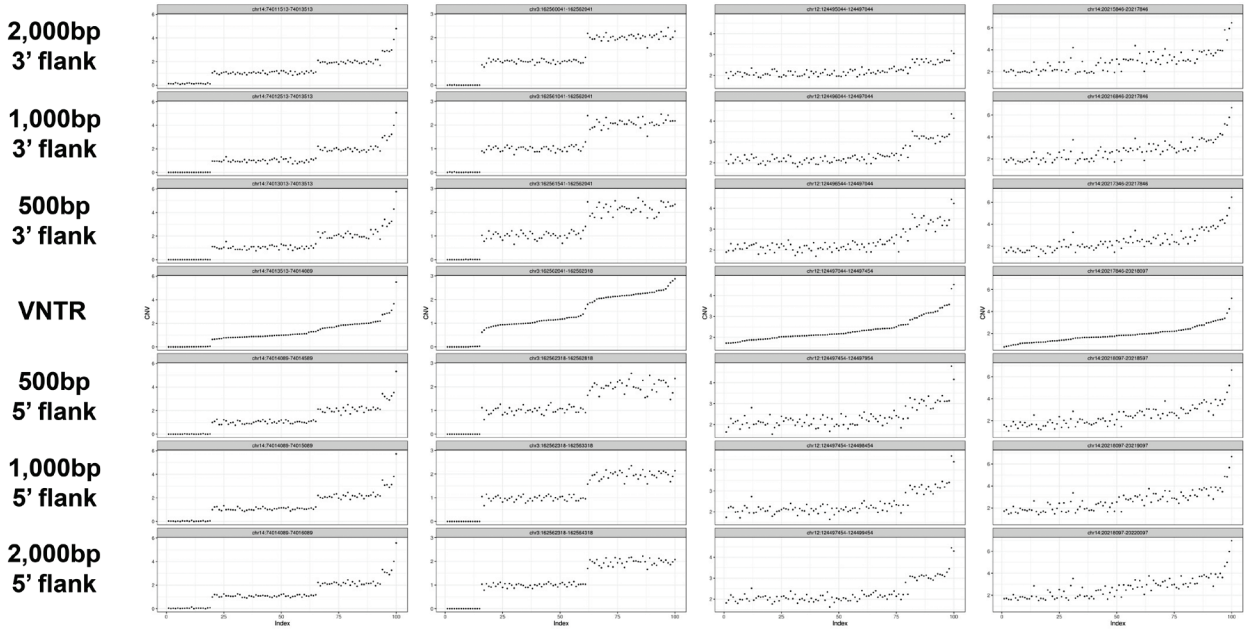

Figure S3. VNTR copy number estimates using *CNVnator* can be confounded by the presence of larger underlying CNVs. In situations where a VNTR is embedded within a larger copy number variable region, copy number estimates for the VNTR based on *CNVnator* read depth can be influenced by underlying variations of the wider region. To identify VNTRs that were subject to this confounder, we analyzed the 3' and 5' 500bp, 1kb and 2kb regions flanking each VNTR using *CNVnator*, and then correlated the values of the 1kb flanks with VNTR copy number. Shown are data from four representative loci that were removed from further analysis. Within each locus, samples are ordered based on the estimated VNTR copy number, revealing that the observed estimates of VNTR copy number are highly correlated with variation in the flanking regions, and likely simply reflect a larger underlying CNV, rather than changes in length of the VNTR array itself.

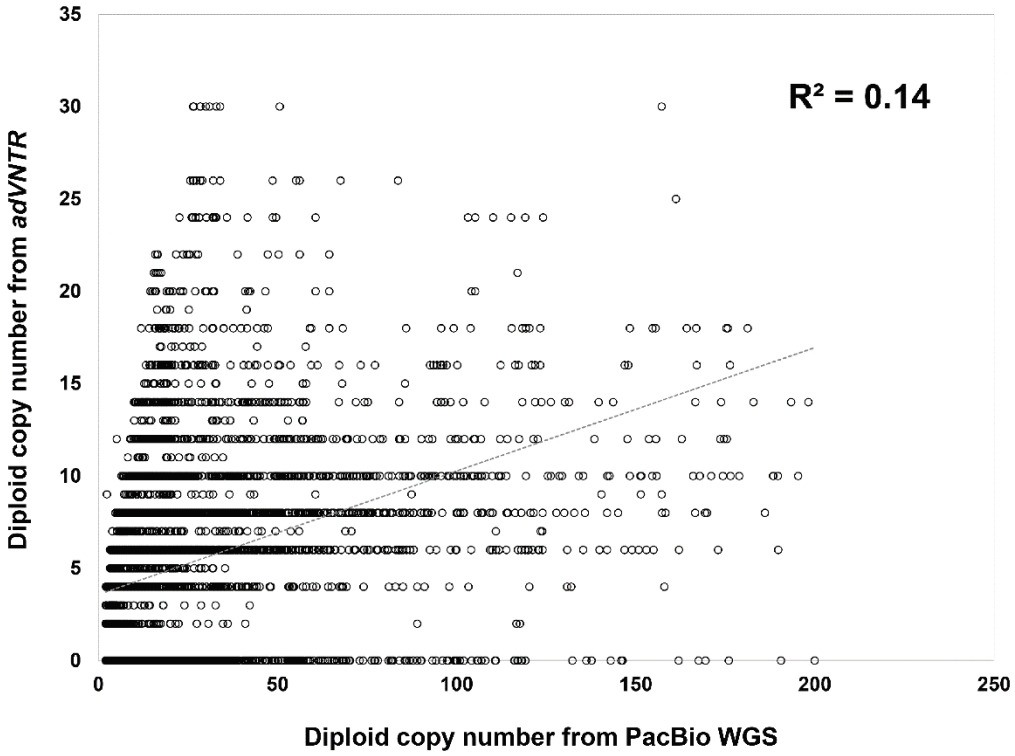

Figure S4. Poor performance of *adVNTR* for genotyping VNTRs. Using data for the same set of 1,891 VNTRs in 14 individuals as shown in Figure 1B (see Methods), we used *adVNTR* to generate VNTR genotypes. When compared with direct genotypes generated from PacBio long-read WGS in these same individuals, we observed an R^2=0.14, indicating generally poor accuracy of this tool when applied to this set of VNTRs (Table S3).

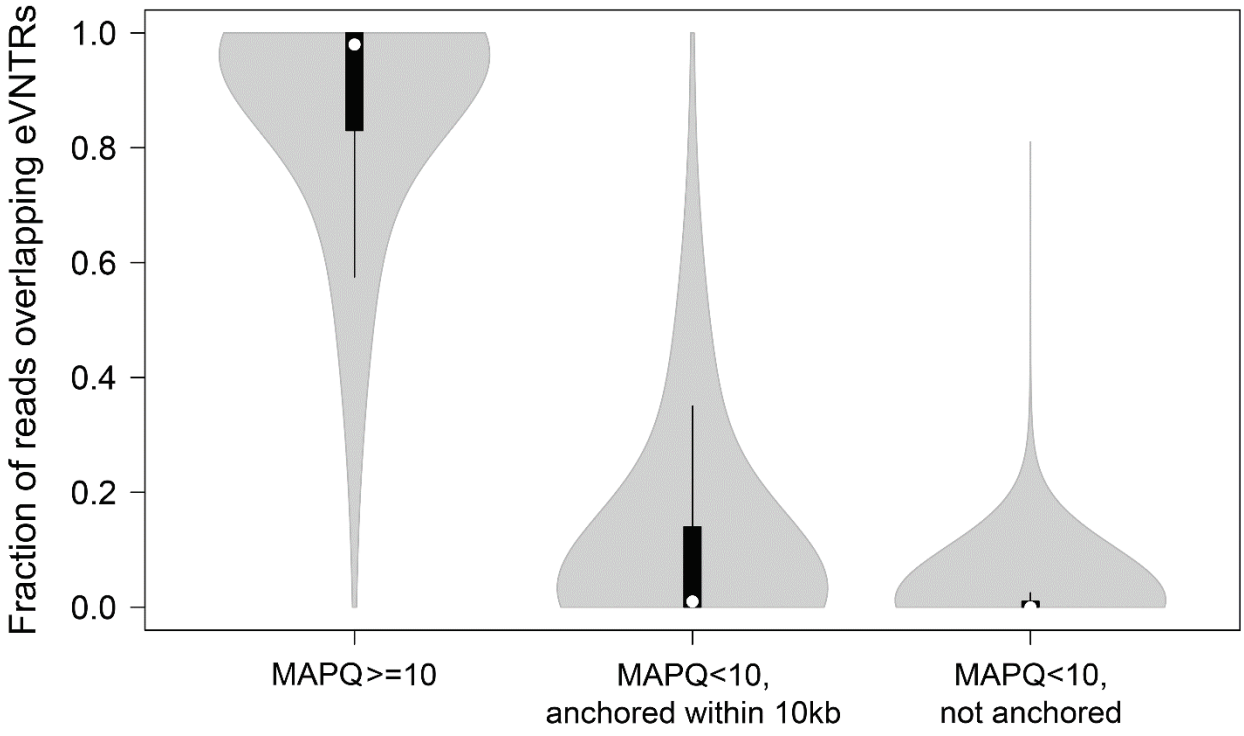

Figure S5. Assessment of the reliability of Illumina reads mapping to VNTR loci. We analyzed Illumina reads mapping to 2,980 autosomal GTEx eVNTRs in a Yoruban sample (NA18874), classifying them into three categories: (i) MAPQ≥10, (ii) MAPQ<10, but with a mate-pair that mapped reliably within ±10kb. (iii) MAPQ<10, without a mate pair that was anchored within ±10kb. Violin plots show the fraction of reads in each of these three categories at each eVNTR locus. Overall, copy number estimates for the vast majority of eVNTRs were based on reliably mapped reads, with only a single eVNTR containing >50% of unreliably mapped reads. Within each violin, the median is indicated by the white dot, box limits indicate the 25th and 75th percentiles, and whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles.

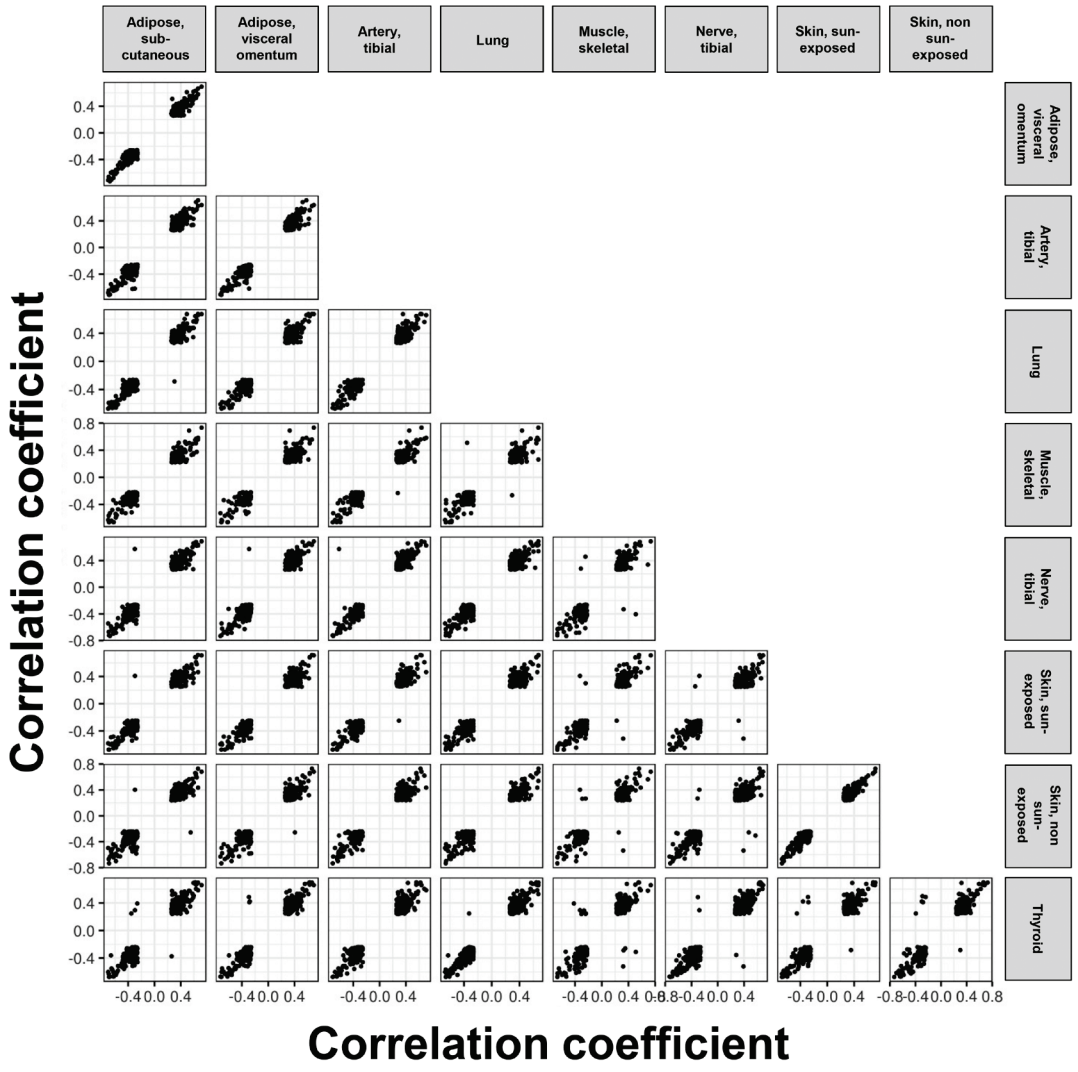

Figure S6. Pairwise correlation patterns of significant eVNTR:gene associations across eight tissues. Each point shows the R values of an individual gene:eVNTR pair that was significant in both tissues. In nearly all cases, the directionality of the observed associations are concordant among different tissues, with only 0.6% of eVNTR:gene pairs showing opposite direction of effect in different tissues, consistent with our results representing genuine associations.

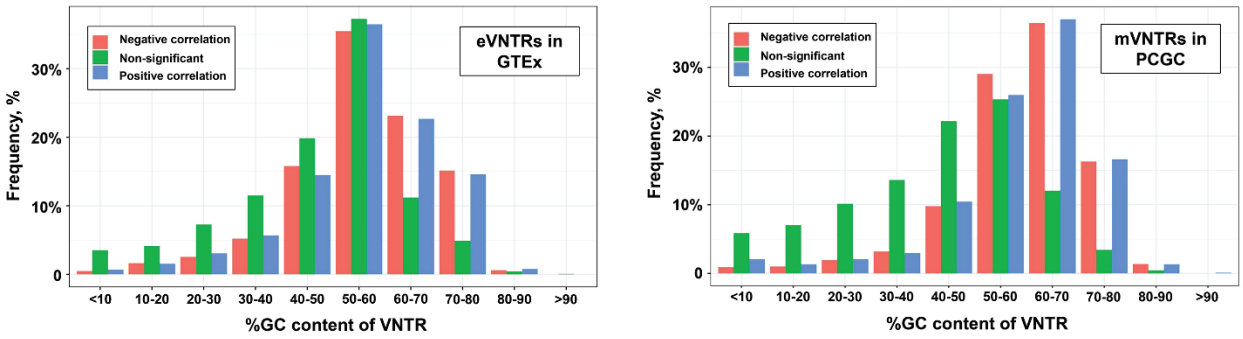

Figure S7. Significant eVNTRs and mVNTRs are biased towards higher GC content. We observed that both putatively functional eVNTRs and mVNTRs showed a clear trend to be composed of motifs with higher GC-content than the background of all VNTRs tested. This trend was stronger for mVNTRs, which is consistent with the Illumina 850k array preferentially sampling CpG in GC-rich regions of the genome, and the observation that most mVNTRs are located physically close (<5kb) from the CpGs that they associated with.

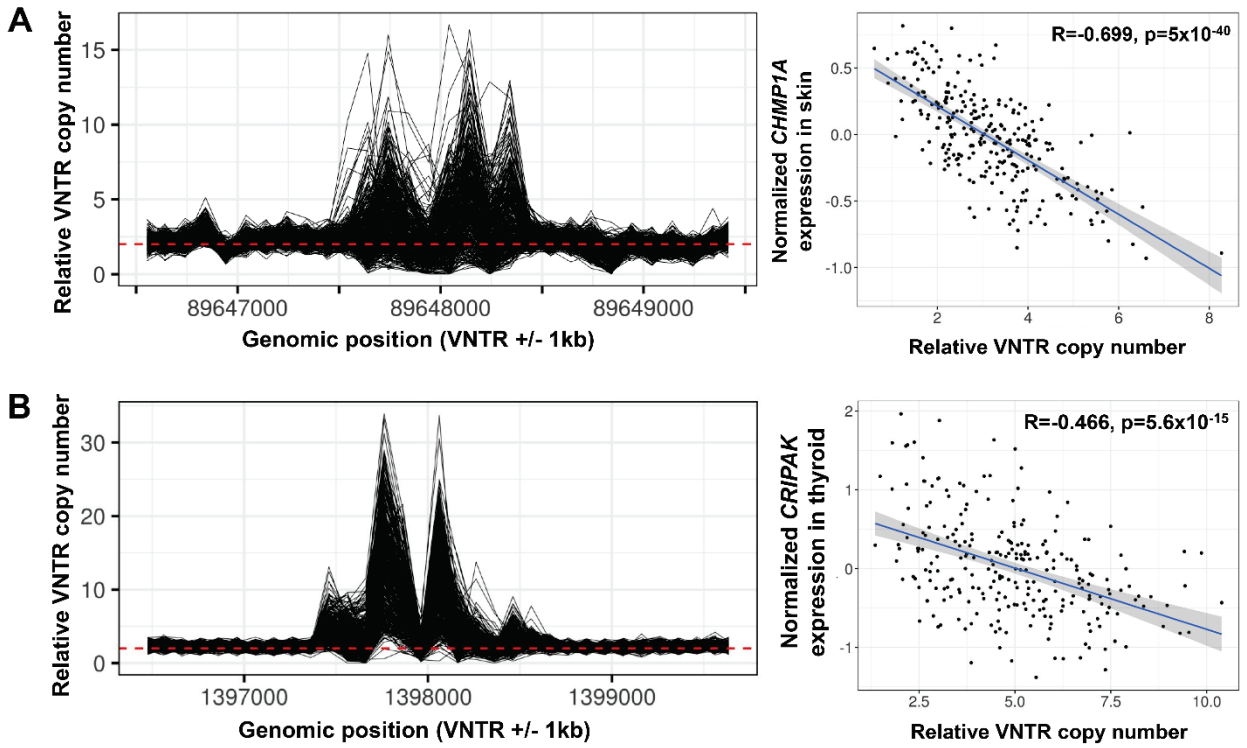

Figure S8. Two example eVNTR loci detected in the GTEx cohort. (A) A 42mer repeat (chr16:89,647,518-89,648,445, hg38) located intronic within *CHMP1A* [MIM: 164010] associates negatively with *CHMP1A* expression in multiple tissues (shown is data from skin, sun exposed lower leg). **(B)** A 45mer repeat (chr4:1,397,437-1,398,660, hg38) located 1.4 kb downstream of *CRIPAK* [MIM: 610203] associates negatively with *CRIPAK* expression in multiple tissues (shown is expression data from thyroid). *CNVnator* locus plots show estimated copy number per 100bp bin over the VNTR region, extending 1kb each side, with the red dashed line indicating diploid copy number equal to that of the reference genome.

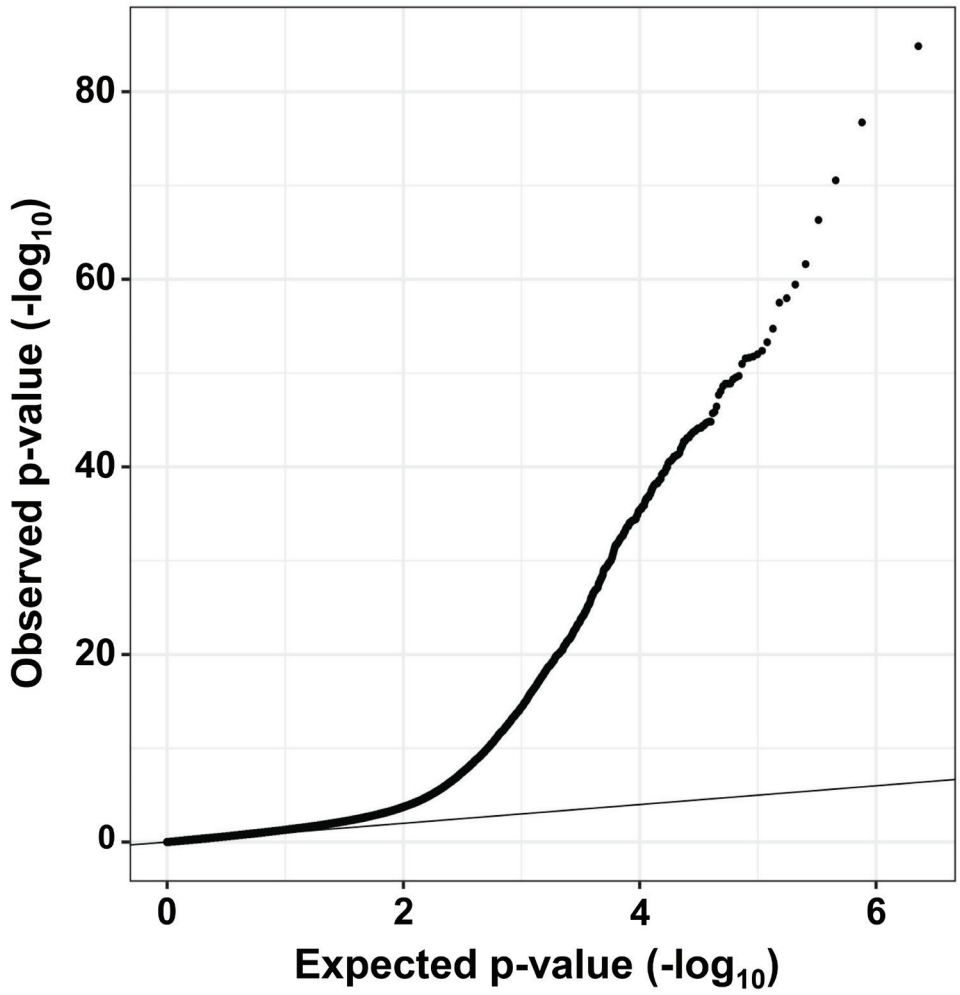

Figure S9. QQ plot showing the distribution of observed versus expected p-values for mVNTRs in whole blood from the PCGC. In this cohort we observed some evidence for genomic inflation (λ=1.297), although with a clear enrichment for significant associations compared to the null. We therefore chose to apply a more stringent multiple testing correction to ensure robust associations in the PCGC cohort.

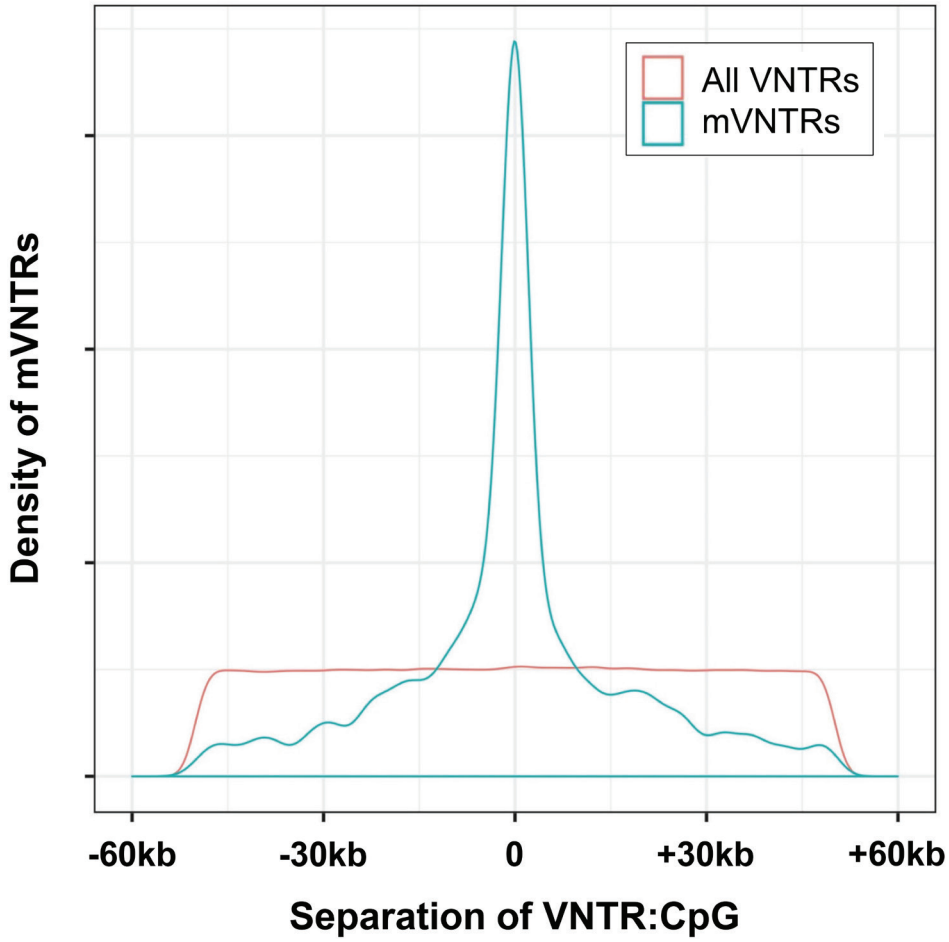

Figure S10. Significant mVNTRs show strong enrichments to be located within close proximity to the CpG whose methylation level they are associated with. These results mirror similar observations made for SNV mQTLs,[65] and for eVNTRs in the GTEx cohort (Figure 1E). However, we note an approximate order of magnitude difference in the distances over which significant eVNTRs and mVNTRs were typically observed to function.

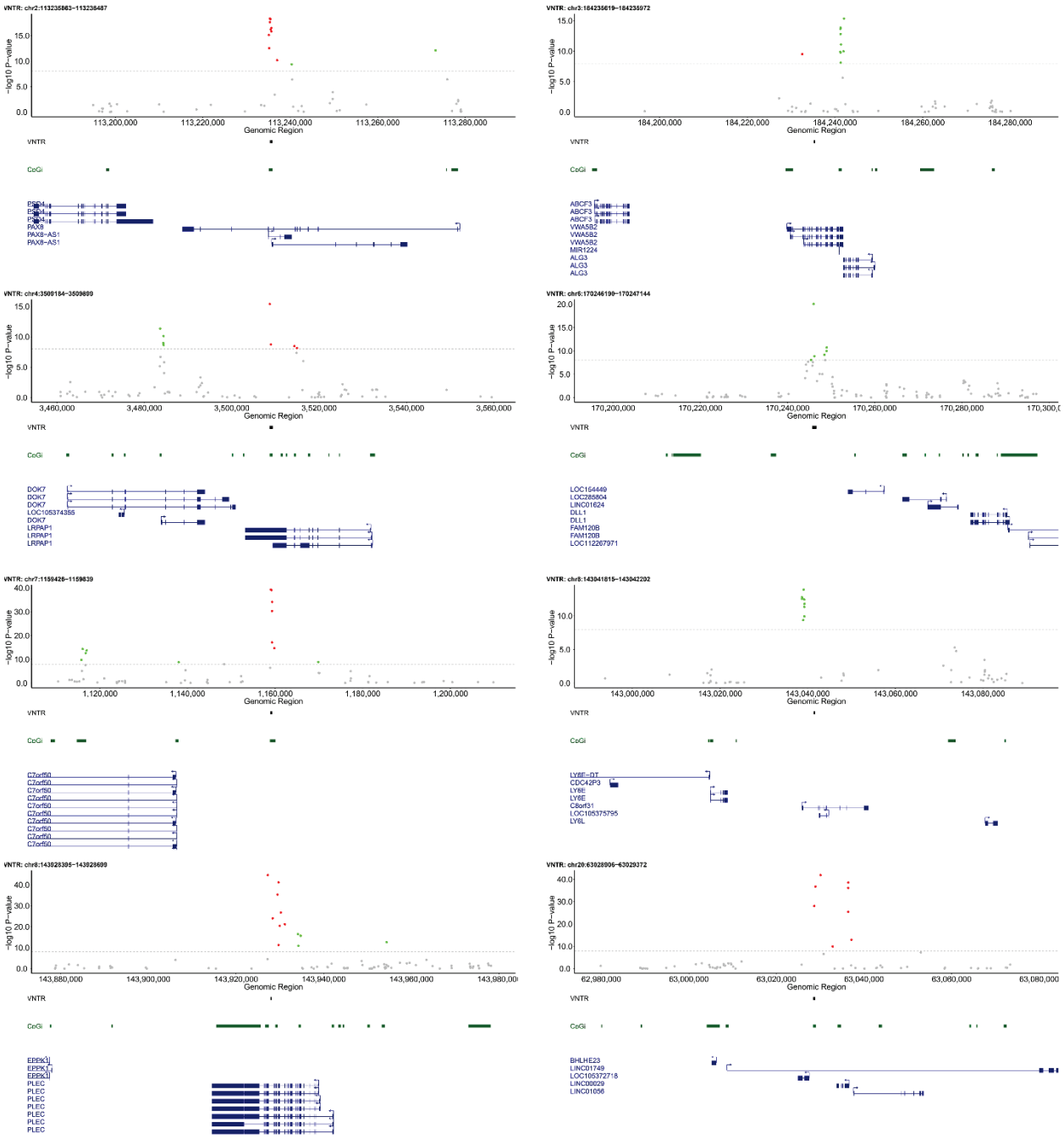

Figure S11. Eight example mVNTR loci detected in the PCGC cohort. Each plot shows associations between copy number of a VNTR and CpG methylation within ±50kb (hg38 coordinates). The horizontal dashed line indicates the significance threshold (p<0.01 after Bonferroni correction for the number of pairwise tests performed genome-wide), with significant

CpGs shown in color, with red representing positive correlations with VNTR copy number, and green indicating negative correlations. The location of the VNTR is indicated by the horizontal black bar in the center below each plot. Underneath each plot are shown the location of CpG islands (green bars) and Refseq genes (blue).

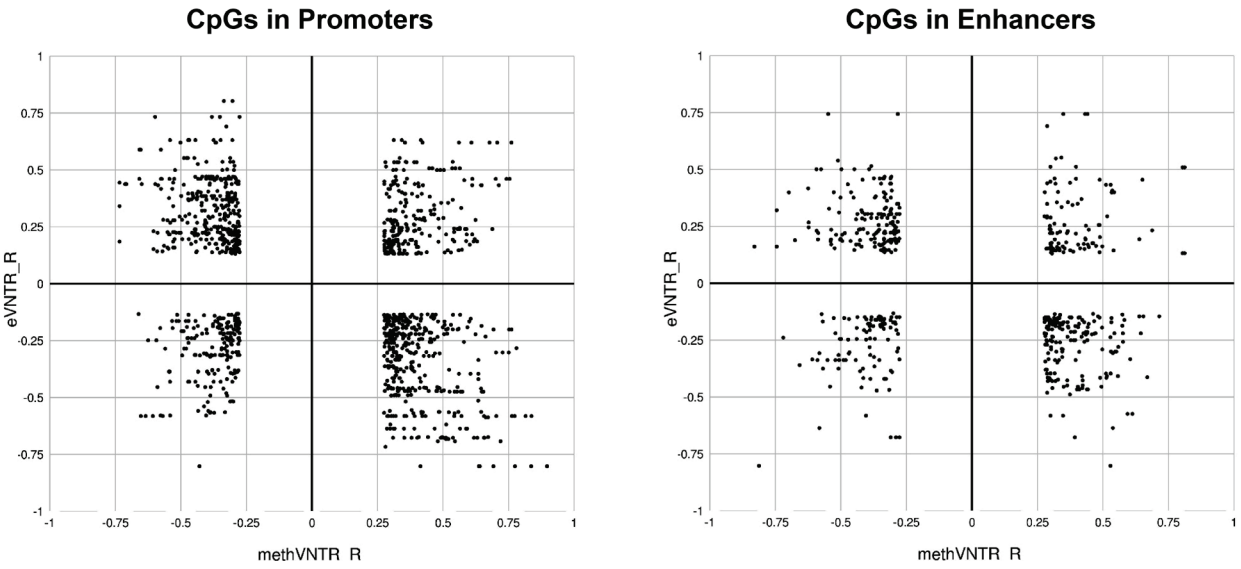

Figure S12. An inverse relationship between gene expression and methylation of regulatory elements associated with VNTRs. Considering VNTRs that were associated with both methylation of CpGs in annotated regulatory regions and expression of the genes they regulate, we compared the correlation coefficients between VNTR copy number and both methylation and expression. We observed that for both promoters ($p=5.8x10^{-10}$) and enhancers ($p=3.9x10^{-16}$) there was a significant inverse relationship of CpG methylation with gene expression, *i.e.* functional VNTRs preferentially showed opposite directionality of effects on methylation of regulatory elements and expression of the associated genes.

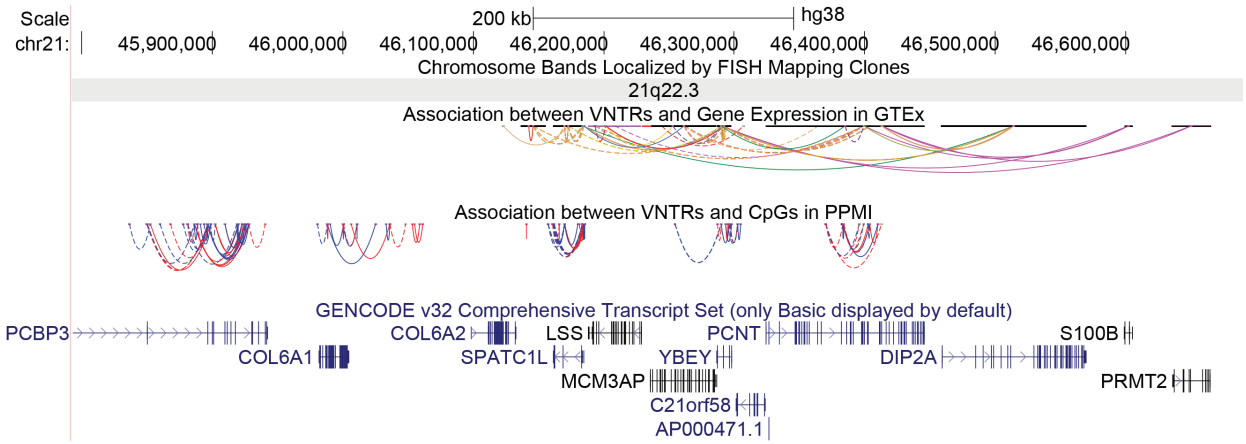

Figure S13. Screenshot showing UCSC Genome Browser tracks created to display eVNTRs and mVNTRs identified in our analysis. Shown is an example region of ~900 kb located at 21q22.3. Each line joins the location of an eVNTR:gene transcription start site, or mVNTR:CpG pair. For GTEx eVNTRs, line color indicates tissue type, while for mVNTRs blue and red lines represent positive and negative associations, respectively. Tracks are titled "Exp/Meth VNTR hub" accessible via the UCSC Genome Browser Track Hubs for both the hg19 and hg38 genome assemblies. A link is included in the Data and Code Availability section of the manuscript.

Acknowledgements

This work was supported by NIH grant NS105781 to AJS, NIH grant R01HG010485 to BP, NIH predoctoral fellowship NS108797 to OLR, and American Heart Foundation Postdoctoral Fellowship 18POST34080396 to AMT. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD018522. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc.(HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania

(MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v7.p2.

The Gabriella Miller Kids First Pediatric Research Program (Kids First) was supported by the Common Fund of the Office of the Director of the National Institutes of Health (www.commonfund.nih.gov/KidsFirst). Baylor College of Medicine's Human Genome Sequencing Center was awarded an administrative supplement (3U54HG003273-12S1) to sequence congenital cohort samples submitted by Christine Seidman through the Kids First program (1X01HL132370). The data analyzed and reported in this manuscript were accessed from dbGaP (www.ncbi.nlm.nih.gov/gap; accession number phs001138). Additional funds from the NHLBI grants U01HL098123, U01HL098147, U01HL098153, U01HL098162, U01HL098163, and U01HL098188 supported the assembly of the Pediatric Cardiac Genomics Consortium (https://benchtobassinet.com/About/AboutPCGC) cohort, and collection of the phenotypic data and samples (PMID: 23410879).

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmiinfo.org/data). For up-to-date information on the study, visit www.ppmiinfo.org. PPMI, a public-private partnership, is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, a full list of which can be found at www.ppmiinfo.org/fundingpartners.

WGS data for samples from the 1000 Genomes Project sample collection were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.