The American Journal of Human Genetics, Volume *107*

Supplemental Data

A Survey of Rare Epigenetic Variation in 23,116 Human Genomes Identifies Disease-Relevant Epivariations and CGG Expansions

Paras Garg, Bharati Jadhav, Oscar L. Rodriguez, Nihir Patel, Alejandro Martin-Trujillo, Miten Jain, Sofie Metsu, Hugh Olsen, Benedict Paten, Beate Ritz, R. Frank Kooy, Jozef Gecz, and Andrew J. Sharp

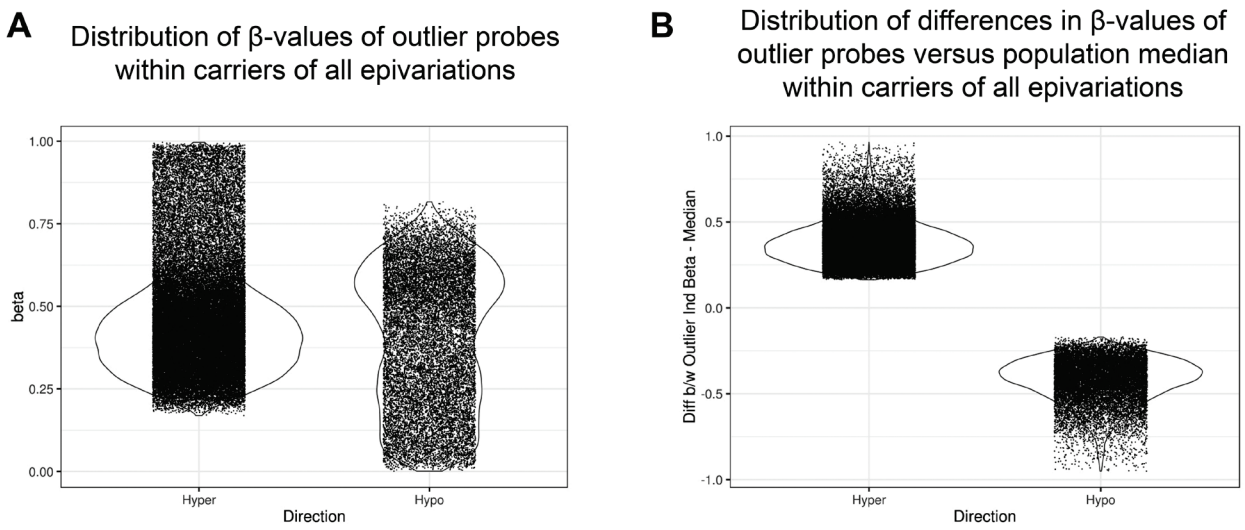

Figure S1. β-value distributions of outlier probes at all epivariations. Violin plots showing **(A)** Distributions of β-values of outlier probes within carriers of all epivariations, and **(B)** Distribution of differences in β-values of outlier probes versus population median within carriers of all epivariations. Full underlying data is shown in Table S5.

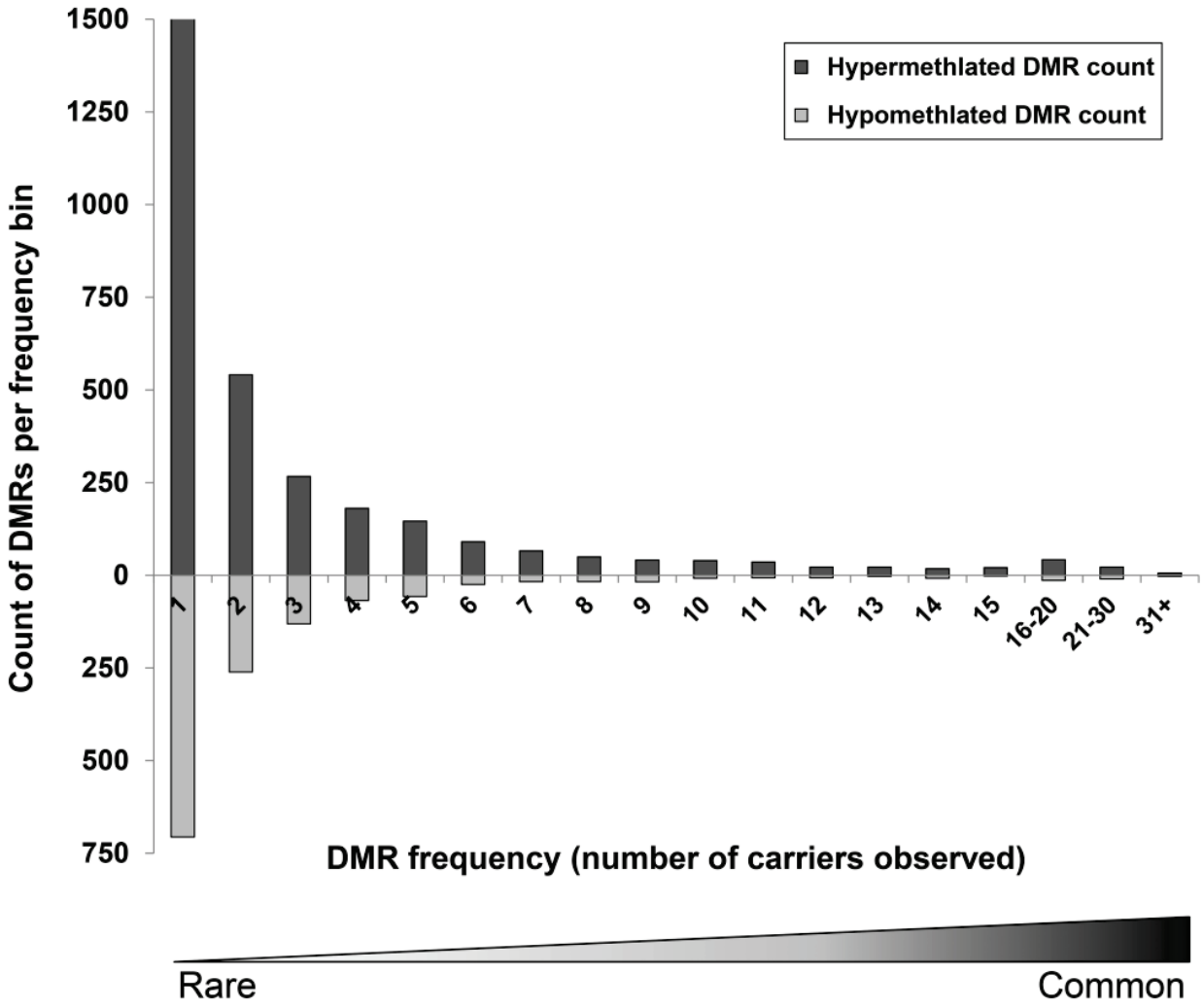

Figure S2. Frequency spectrum of autosomal hypermethylated and hypomethylated epivariations in the human genome. We identified 13,879 autosomal epivariations in 7,653 individuals, corresponding to 4,452 unique autosomal loci (Table S3). Many epivariations are rare events, with 2,193 of 4,452 (49.3%) epivariations being singletons, i.e. observed in only a single individual. Overall, there was an ~2.3-fold excess of hypermethylated compared to hypomethylated epivariations, with 3,095 loci showing gains of methylation, and 1,329 loci showing losses. In addition, 28 loci showed bidirectional changes, with either hyper- or hypomethylation observed in different samples at the same locus (not shown).

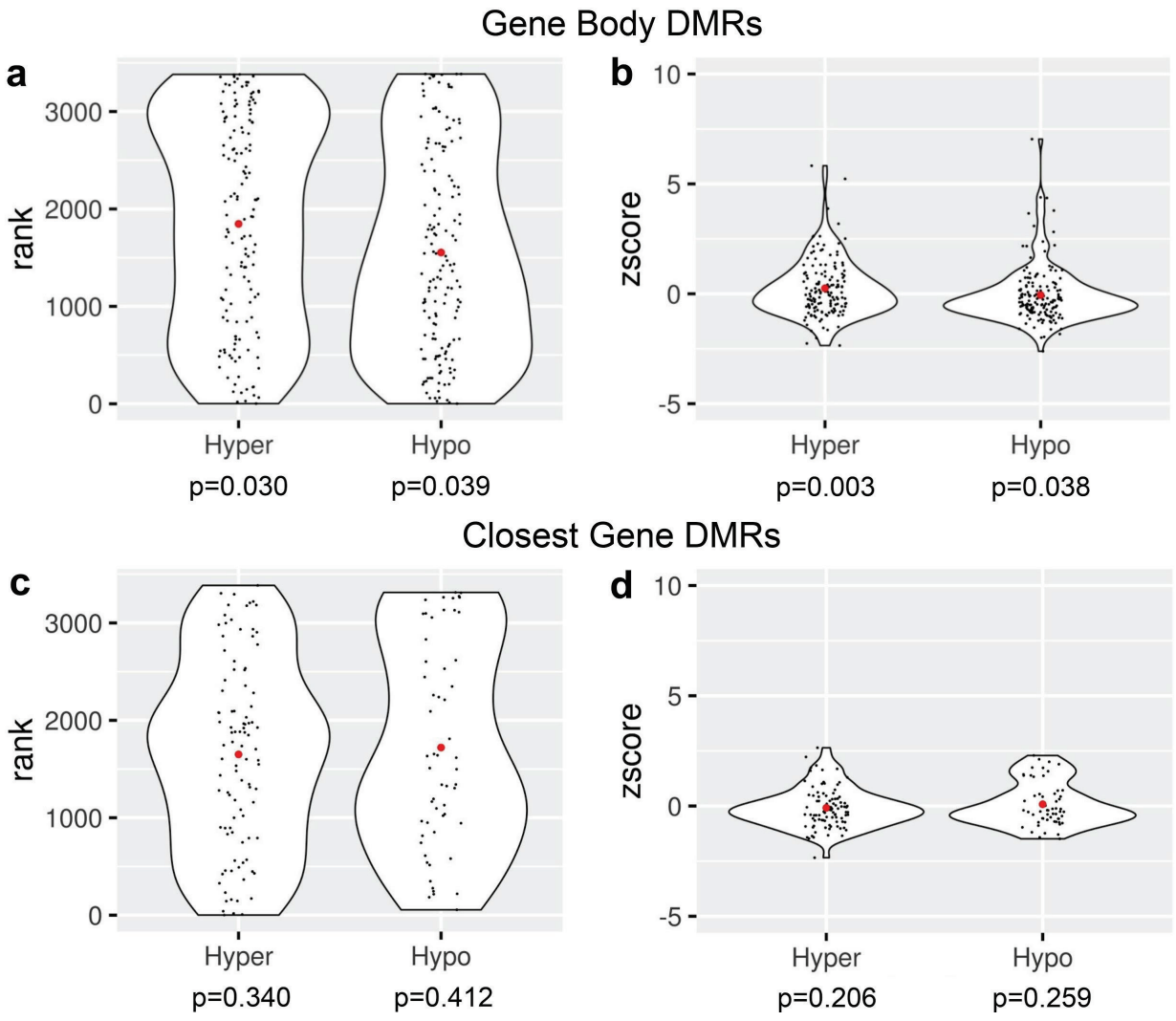

Figure S3. Effects on expression of epivariations that overlap gene bodies (excluding promoter regions), and the effect of intergenic epivariations on the closest gene. Using RNAseq data from 3,560 individuals in the BIOS cohort, we observed weak but nominally significant effects of epivariations located within gene bodies, but not for intragenic epivariations. **(A,B)** Gene body methylation changes showed a weak positive correlation with expression, while **(C,D)** epivariations at intergenic loci showed no significant correlation with expression of the closest gene. However, we do note evidence of bimodality in the observed distributions for intergenic epivariations, which might indicate that while some have positive effects on nearby gene expression, others are suppressive. P-values were generated using 10,000 permutations.

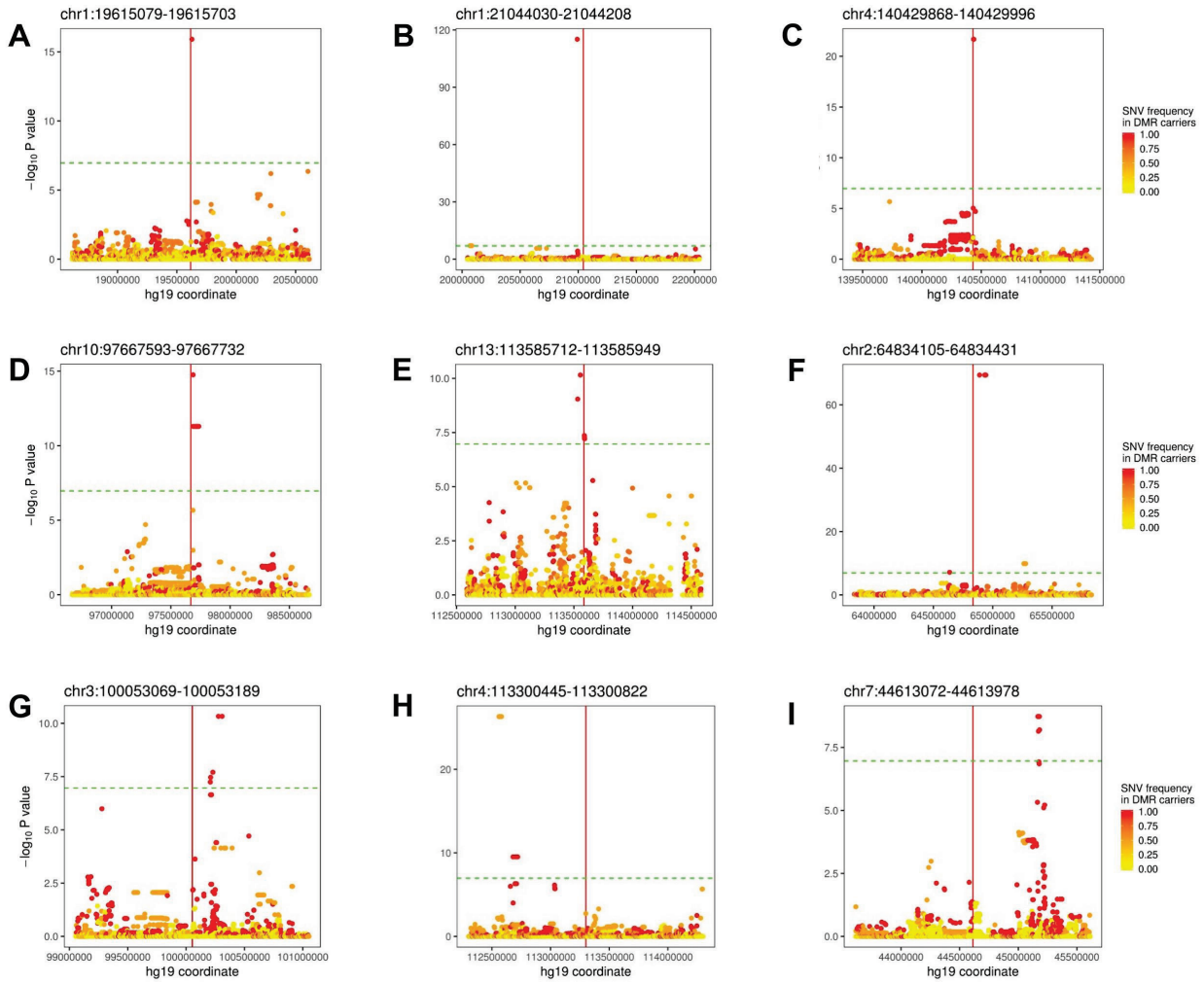

Figure S4. Example associations between epivariations and local SNVs. Using data from 933 individuals from the Women's Health Initiative in whom both methylation and SNV data were available, we performed association analysis with SNVs located within ±1Mb of each of 97 epivariations that were observed in multiple individuals. **(A-F)** At most loci, the strongest associations were with SNVs located within close proximity or overlapping the epivariation locus. **(G-I)** However, in some instances, significant associations occurred with SNVs located >500kb away. These data are consistent with many epivariations being secondary events that occur downstream of local genetic variants. In each plot, the relative position of the epivariation is indicated by the vertical red line, while the horizontal dashed green line indicates a significance threshold of 1% FDR.

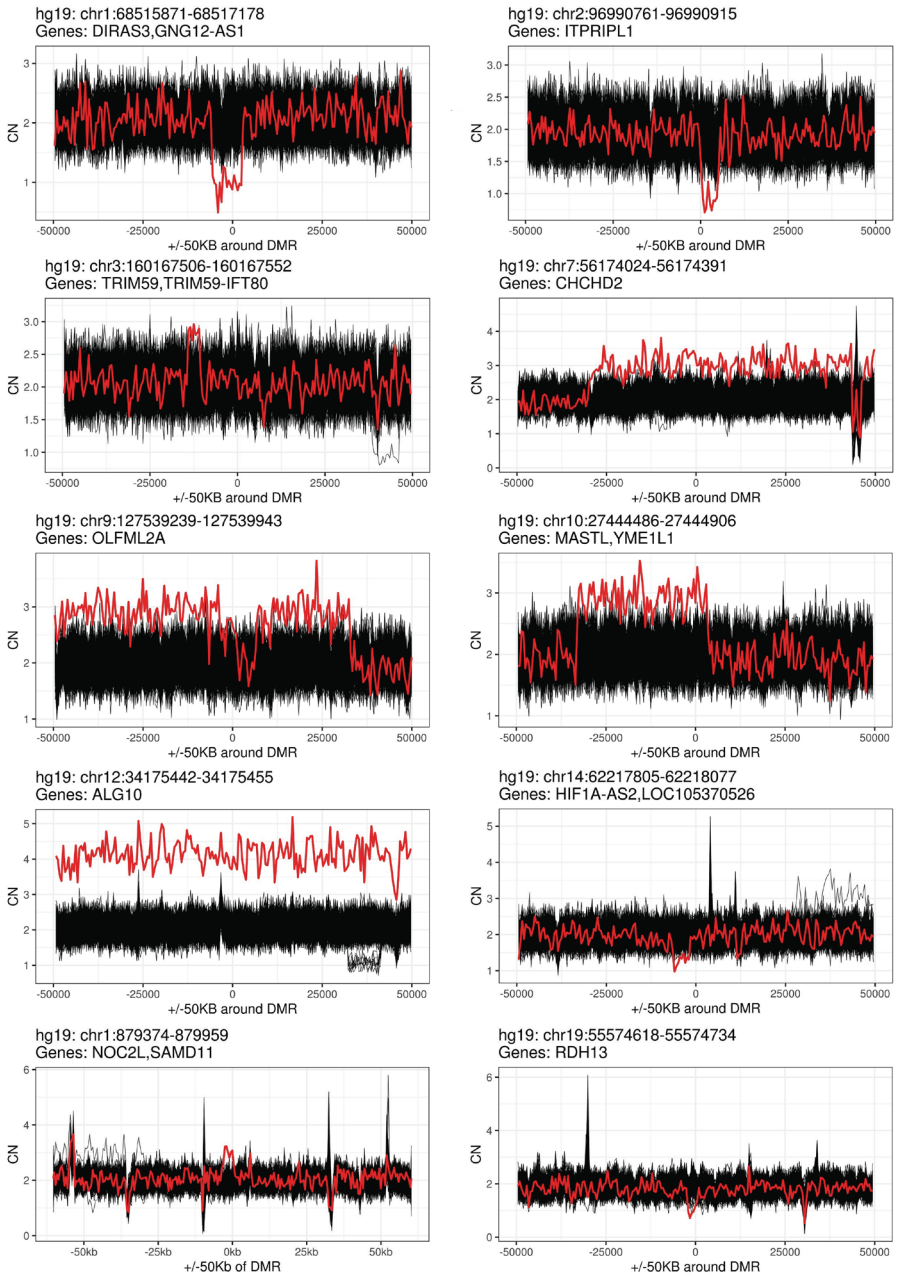

Figure S5. CNVs detected in epivariation carriers that lie in close proximity to the region of altered methylation. Using 477 individuals from the PPMI cohort who had both methylation and WGS data available, we identified ten epivariation carriers that had rare deletions or duplications located within +/-50kb of the epivariation. Compared to the background frequency of rare CNVs at these same loci in other samples from the PPMI cohort who did not carry an epivariation, this represents a 37.4-fold enrichment for rare CNVs to co-occur with an epivariation ($p=2x10^{-71}$). Each plot shows a ±50kb region around an epivariation. Lines show estimated diploid copy number per 500bp interval in 477 individuals from the PPMI cohort, with the epivariation carrier shown in red, and all other individuals who do not carry the epivariation in black. A full list of all rare CNVs detected within 50kb of epivariations in the PPMI cohort is listed in Table S13.

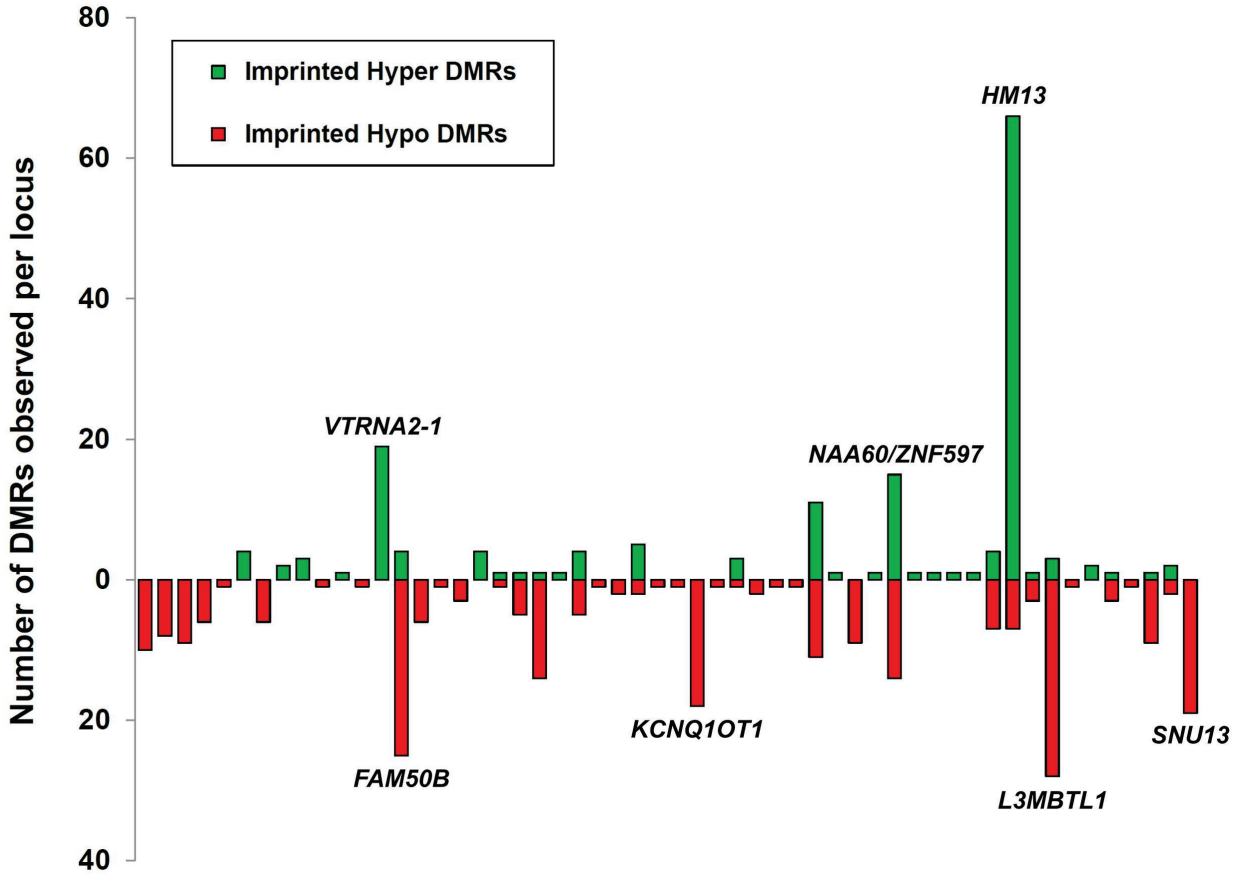

Figure S6. Many imprinted loci show frequent gains and losses of methylation. Nearly one third of imprinted loci with epivariations showed bidirectional changes, exhibiting both gains and losses of methylation in different individuals, a rate 85-fold higher than the background rate in the genome ($p=1.2x10^{-24}$, χ-square test). Several imprinted loci show frequent epivariation (labeled), with *HM13* being the most common epivariation observed in the entire study, and *L3MBTL1*, *NAA60/ZNF597* and *FAM50B* all showing estimated population frequencies >1 per 1,000 individuals. Imprinted loci are ordered according to genomic position, with full details listed in Table S3.

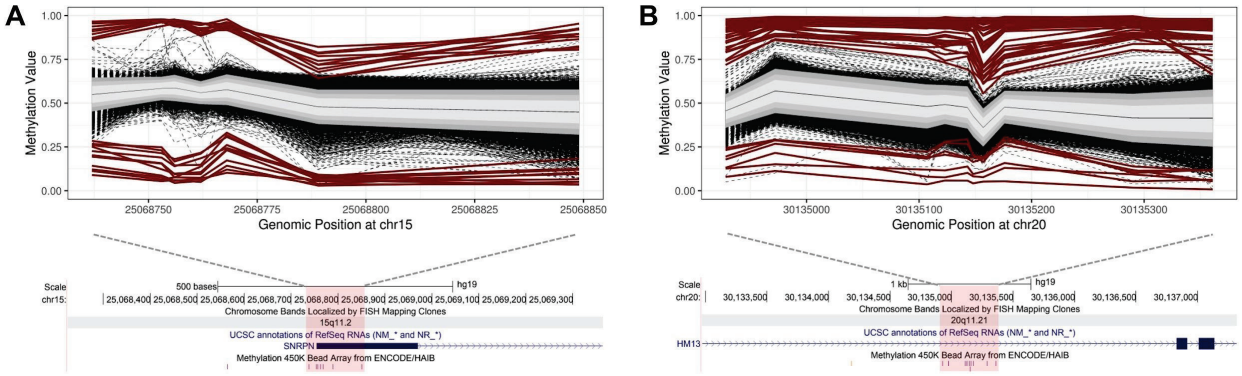

Figure S7. Recurrent bi-directional epivariations observed at imprinted loci. We identified **(A)** 22 individuals that carried either hyper- or hypomethylated epivariations at the promoter of the long isoform of *SNRPN* (chr15:25068737-25068851, hg19), and **(B)** 73 individuals with epivariations at the *HM13* locus (chr20:30134928-30135363, hg19). Both loci are normally methylated exclusively on the maternal allele. In each methylation plot, individuals with epivariations are shown as bold red lines, while the other ~23,000 samples tested are shown as dashed black lines. Grey shaded regions indicate 1, 1.5 and 2 Standard Deviations from the mean of the distribution, while the solid black line shows the mean β-value of the entire population. Below each methylation plot are screenshots from the UCSC Genome Browser showing the wider genomic region, which are zoomed out 10-fold from the regions shown in the methylation plots (indicated by the red shaded area), including genes and the location of CpGs assayed by probes on the Illumina 450k array.

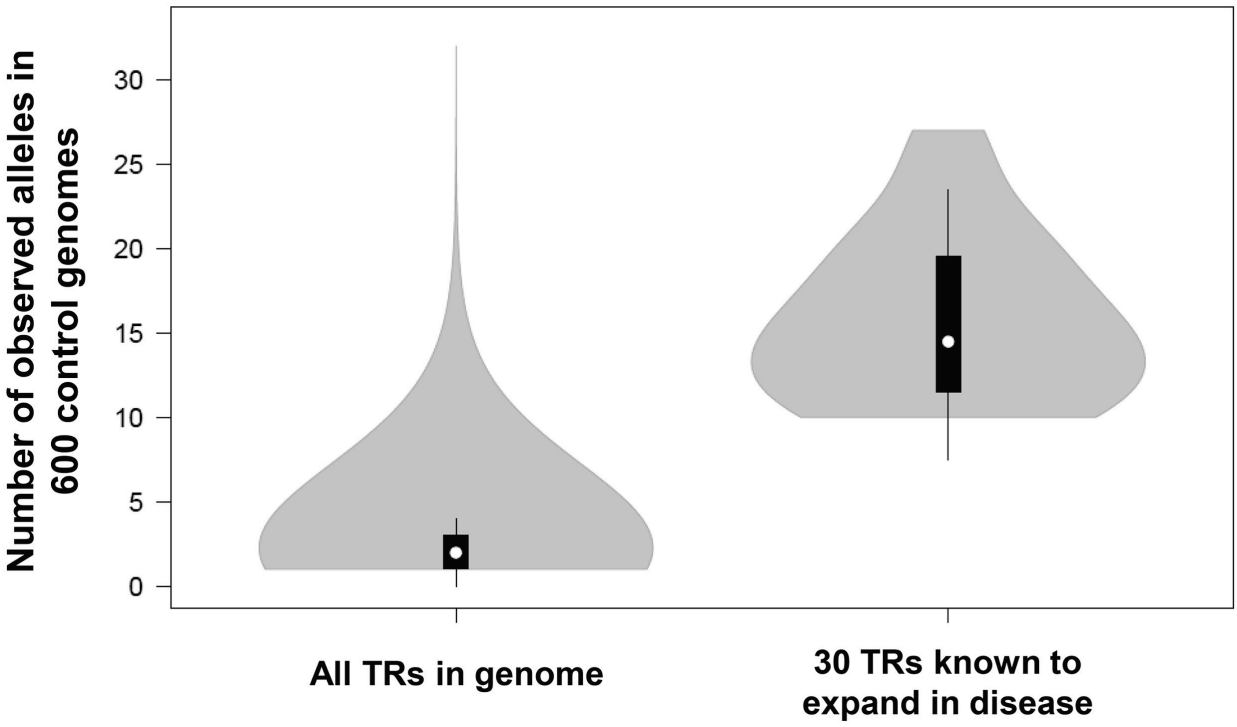

Figure S8. Pathogenic tandem repeats that show rare expansions in human disease show high levels of polymorphism in the general population. Using TR genotypes generated by *hipSTR* in 600 unrelated individuals who had undergone Illumina WGS, we observed that TRs that are known to undergo occasional expansion in human disease show high levels of polymorphism in the general population. Based on this observation, we hypothesized that population variability is a strong predictor of the potential of a TR to undergo rare expansion.

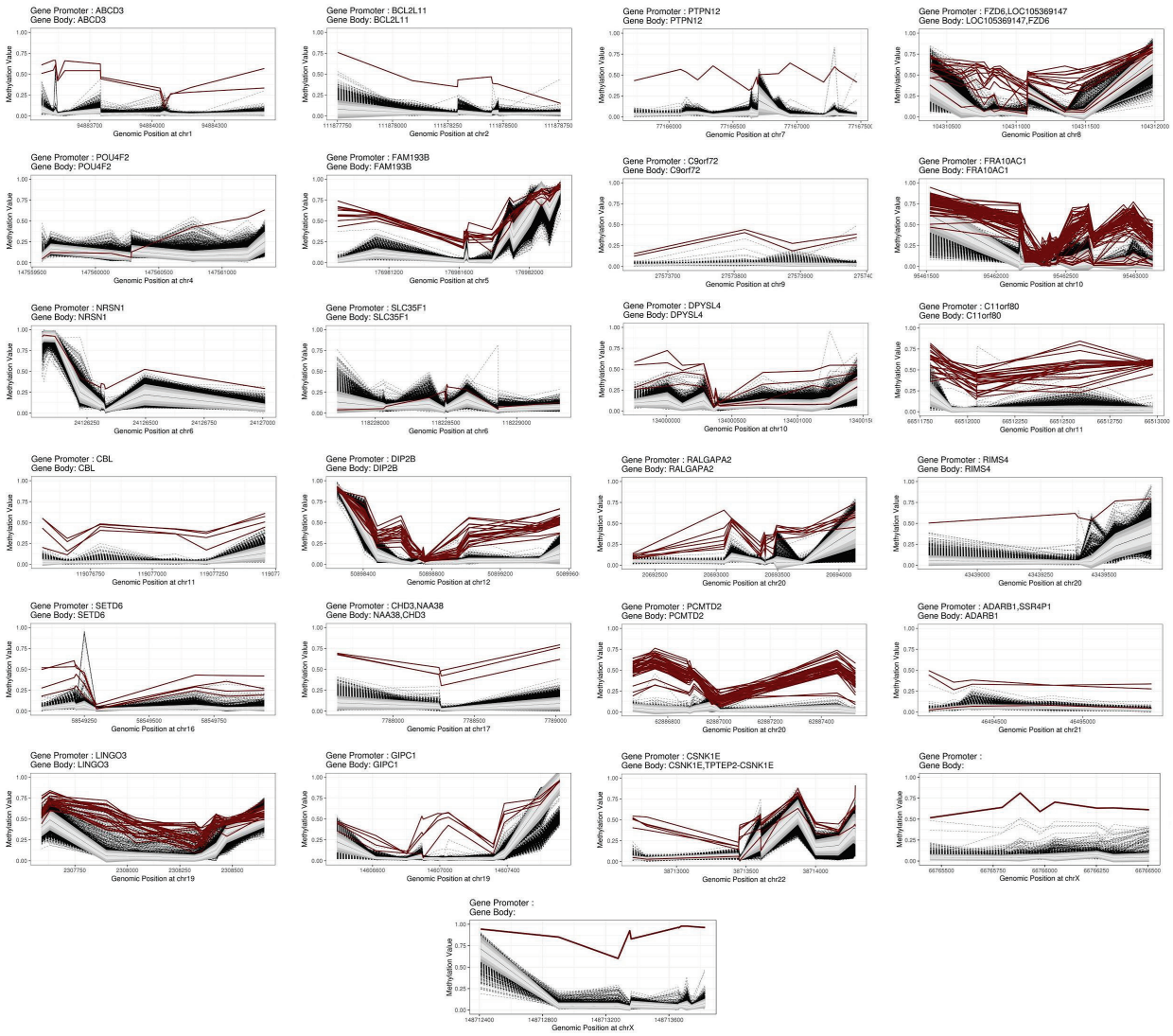

Figure S9. Methylation profiles observed at all unstable CGG repeats with epivariations. We identified 25 loci where highly polymorphic CG-rich TRs overlapped epivariations, suggesting that the observed hypermethylation events might be caused by expansion of these putatively unstable TRs. This included known CG-rich TRs that undergo occasional expansion and hypermethylation, such as *FRA10AC1*, *C11orf80*, *CBL2*, *C9orf72*, *DIP2B*, and *TMEM185A*. In each methylation plot, individuals with epivariations are shown as bold red lines, while the other ~23,000 samples tested are shown as dashed black lines. Grey shaded regions indicate 1, 1.5 and 2 Standard Deviations from the mean of the distribution, while the solid black line shows the mean β-value of the entire population.

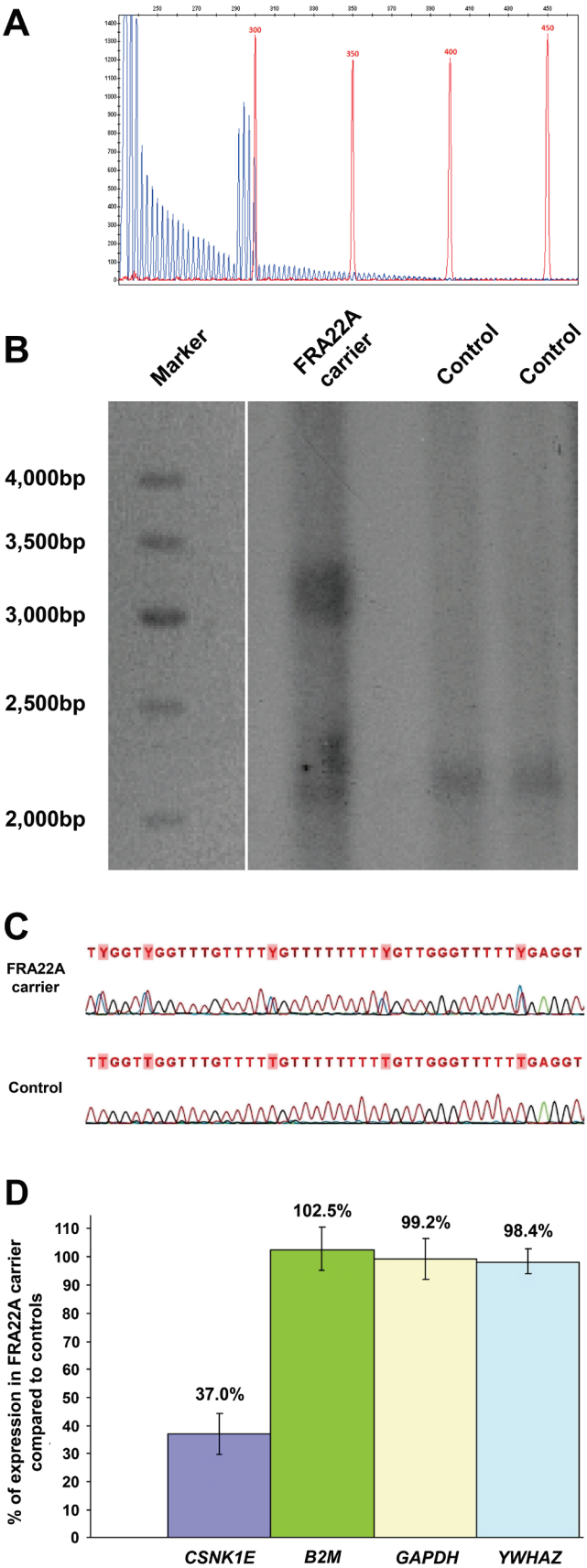
A
300
350
400
450
B
Marker
FRA22A carrier
Control
Control
4,000bp
3,500bp
3,000bp
2,500bp
2,000bp
C
FRA22A carrier
TYGGTYGGTTTGTTTTYGTTTTTTTTYGTTGGGTTTTTYGAGGT
Control
TTGGTTGGTTTGTTTTTGTTTTTTTTTGTTGGGTTTTTTGAGGT
D
% of expression in FRA22A carrier compared to controls
110
100
90
80
70
60
50
40
30
20
10
0
37.0%
102.5%
99.2%
98.4%
CSNK1E
B2M
GAPDH
YWHAZ

Figure S10. Validation that a methylated repeat expansion within the 5'UTR of *CSNK1E* underlies the FRA22A fragile site. In an individual expressing the FRA22A fragile site, we identified a novel insertion allele associated with hypermethylation and reduced expression of *CSNK1E*. **(A)** Repeat primed-PCR of the CGG repeat (chr22:38,713,353-38,713,380) showed a characteristic saw-tooth pattern on the fluorescence trace, with periodicity of 3bp, indicative of a triplet repeat expansion (data not shown). **(B)** Southern blot identified a novel smeared fragment of approximately 3.2-kb in the patient in addition to the expected fragment of 2.2-kb seen in controls, which, together with the PCR result, indicate the presence of an expanded poly(CGG) tract of approximately 340 repeats. **(C)** Analysis of CpG sites in the promoter of *CSNK1E* using both Sanger bisulfite sequencing and Pyrosequencing showed methylation levels of 40-50% in the FRA22A carrier, while control samples were essentially unmethylated. **(D)** Finally, using real-time RT-PCR in lymphoblastoid cells, we observed that in the FRA22A-carrier, expression of *CSNK1E* was reduced to ~37% of the level observed in controls ($p=0.003$, Wilcoxon rank-sum test). Overall, these results indicate that an expansion of a CGG repeat in the 5'UTR of *CSNK1E* results in allelic methylation and silencing of the gene, and represents the molecular defect underlying the FRA22A FSFS.