



Figures and figure supplements

5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA

Marketa Tomkova et al

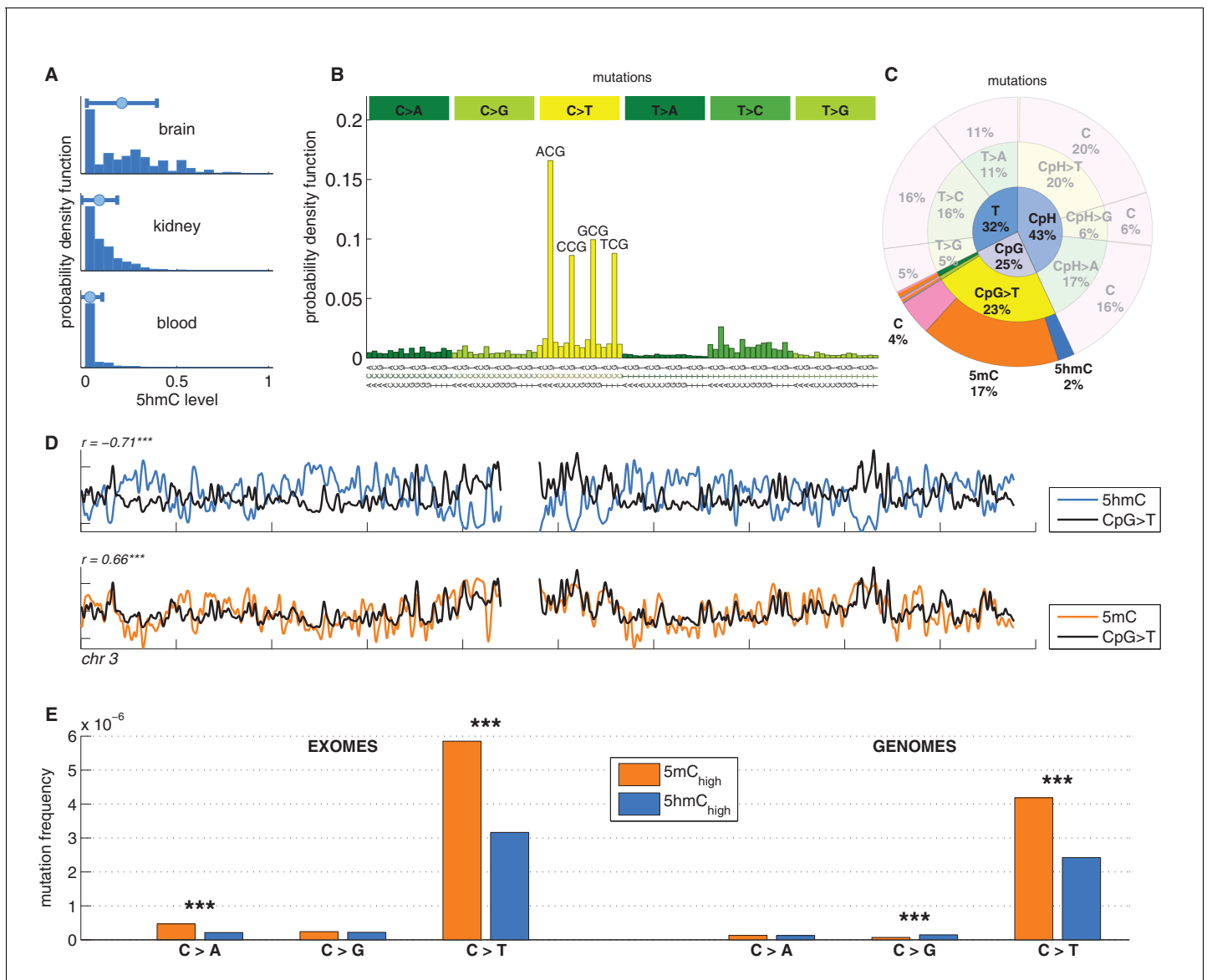


Figure 1. C>T mutations are common in the genome but depleted in 5hmC sites compared to 5mC sites. (A) Distribution of 5hmC in a CpG context in brain compared to kidney and blood. (B) Frequency of SNVs in brain cancer exomes, stratified by sequence context, normalised by frequency of trinucleotides. (C) Distribution of single-nucleotide variants (whole genomes) in brain cancer according to type, context and modification state. (D) CpG>T mutation frequency (black), 5hmC (blue) and 5mC (orange) density in 100 kbp windows of chromosome 3, smoothed with a Gaussian filter ($n = 50$, $\sigma = 2.5$). (E) Average fraction of mutated sites for 5mC_{high} vs. 5hmC_{high} over all patient samples (CpG sites only; *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$, see Materials and methods).

DOI: [10.7554/eLife.17082.003](https://doi.org/10.7554/eLife.17082.003)

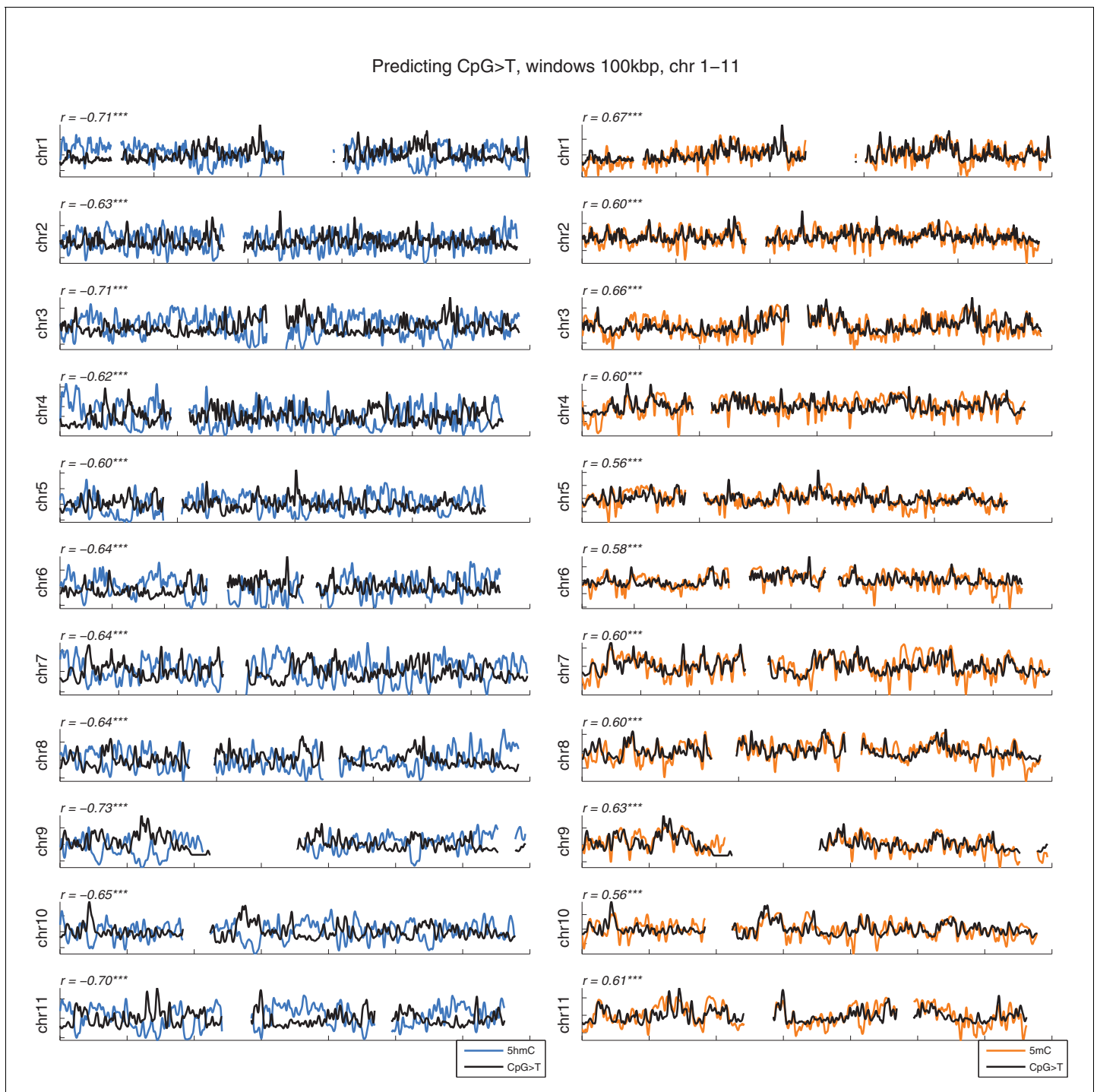


Figure 1—figure supplement 1. Distribution of CpG>T mutations vs modifications across all chromosomes. CpG>T mutation frequency (black), 5 hmC (blue) and 5 mC (orange) density in 100 kbp windows, smoothed with a Gaussian filter ($n = 50$, $\sigma = 2.5$).

DOI: [10.7554/eLife.17082.004](https://doi.org/10.7554/eLife.17082.004)

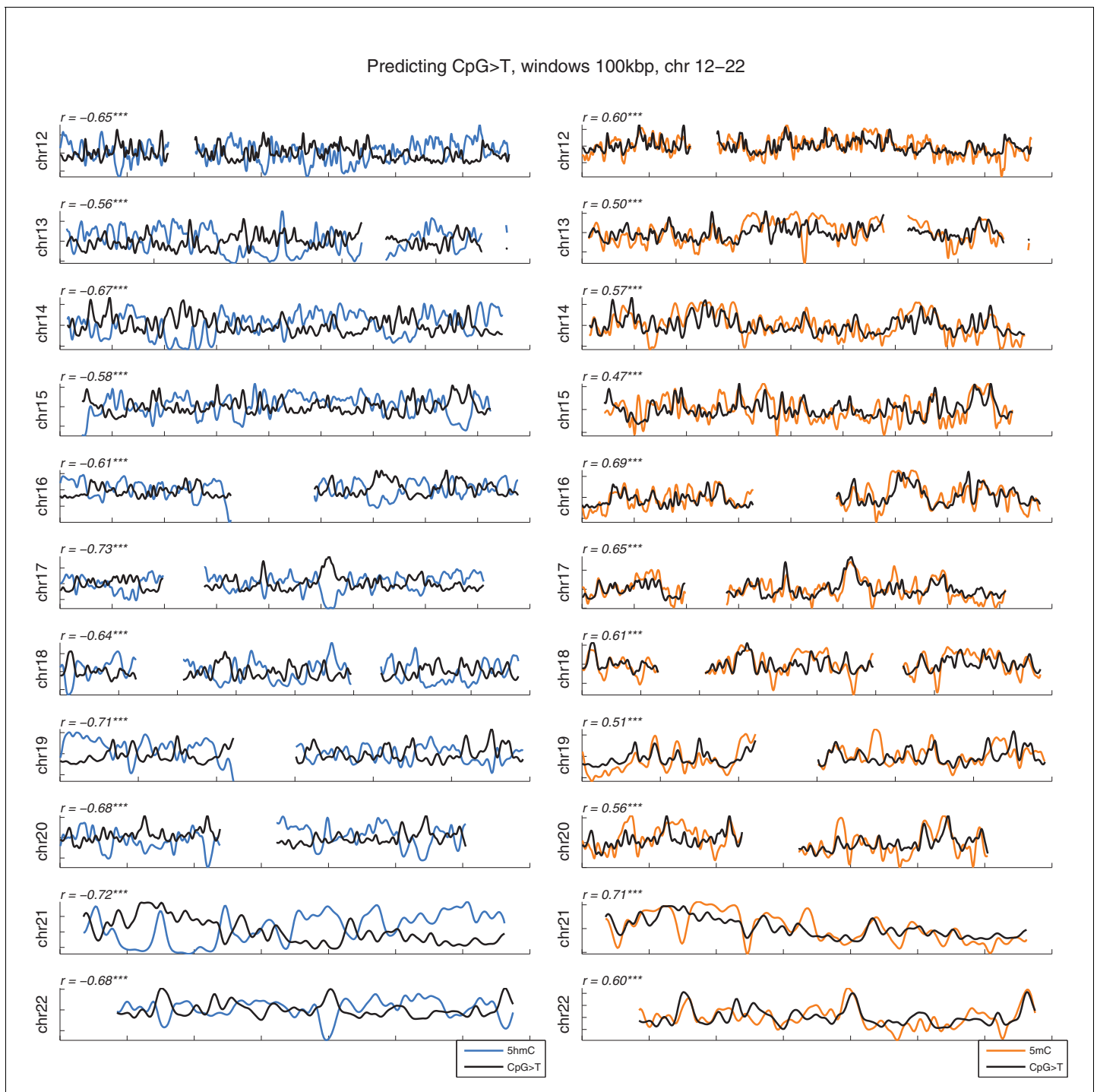


Figure 1—figure supplement 2. Distribution of CpG>T mutations vs modifications across all chromosomes. CpG>T mutation frequency (black), 5 hmC (blue) and 5 mC (orange) density in 100 kbp windows, smoothed with a Gaussian filter ($n = 50$, $\sigma = 2.5$).

DOI: [10.7554/eLife.17082.005](https://doi.org/10.7554/eLife.17082.005)

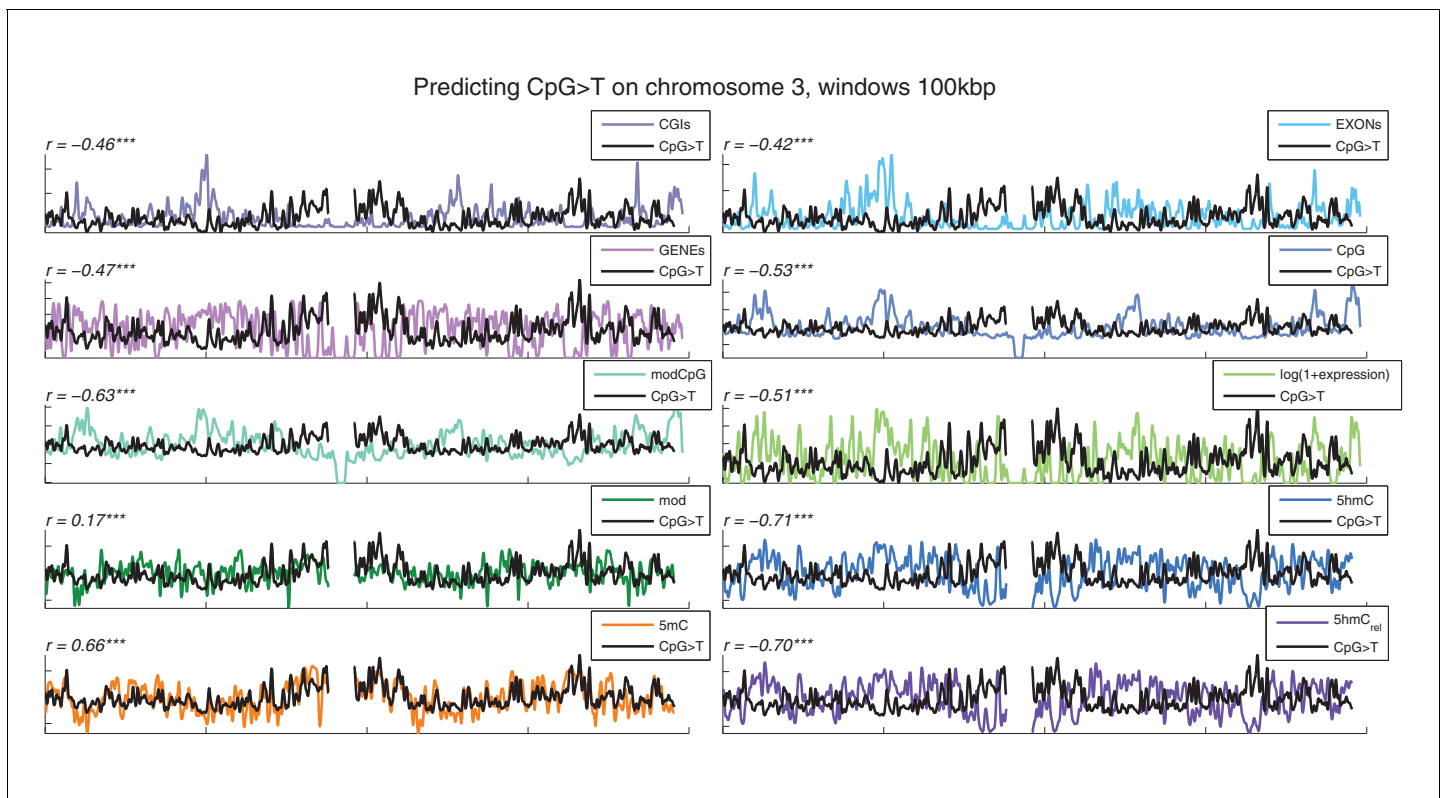


Figure 1—figure supplement 3. Distribution of CpG>T mutations vs other genomic features. CpG>T mutation frequency (black) and several genomic features in 100 kbp windows on chromosome 3, smoothed with a Gaussian filter ($n = 50$, $\sigma = 2.5$). CGIs: density of CpG islands, EXONs: density of exons, GENEs: density of genes, CpG: density of CpGs, modCpG: density of CpGs with *mod level* $\geq 10\%$; and average modification levels: mod, 5 hmC, 5 mC, and 5 hmC_{rel}.

DOI: [10.7554/eLife.17082.006](https://doi.org/10.7554/eLife.17082.006)

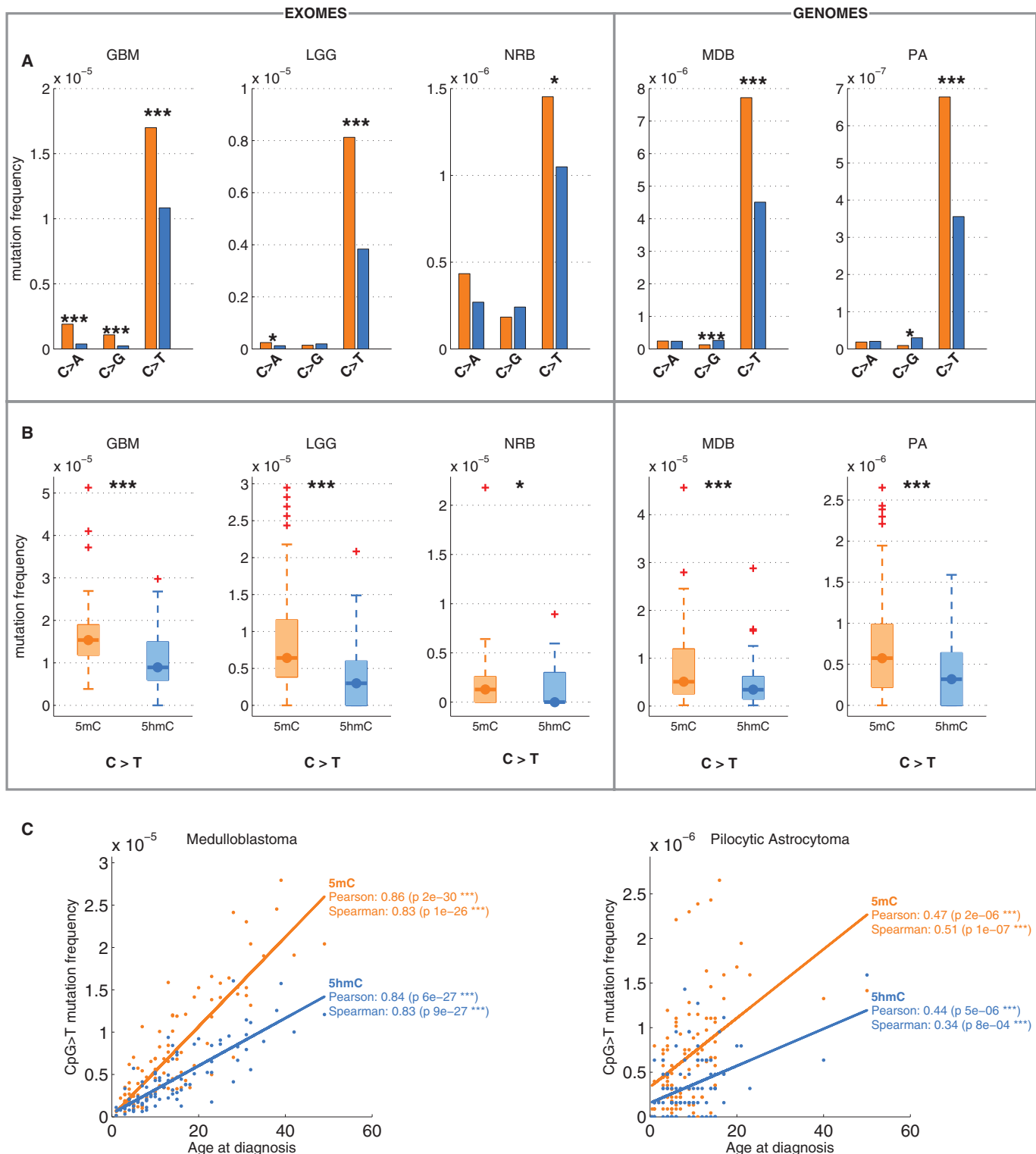


Figure 2. Differential mutation frequency between 5mC and 5hmC is present in all 5 brain cancer types and correlates with age at diagnosis. (A) Average fraction of mutated CpG sites for 5mC_{high} vs. 5hmC_{high} computed separately for each cancer type. (B) Box plot of C>T mutation frequency, as Figure 2 continued on next page

Figure 2 continued

shown in A. (C) Correlation of whole genome CpG>T mutation frequency with age at the time of diagnosis in patients with Medulloblastoma and Pilocytic Astrocytoma.

DOI: [10.7554/eLife.17082.007](https://doi.org/10.7554/eLife.17082.007)

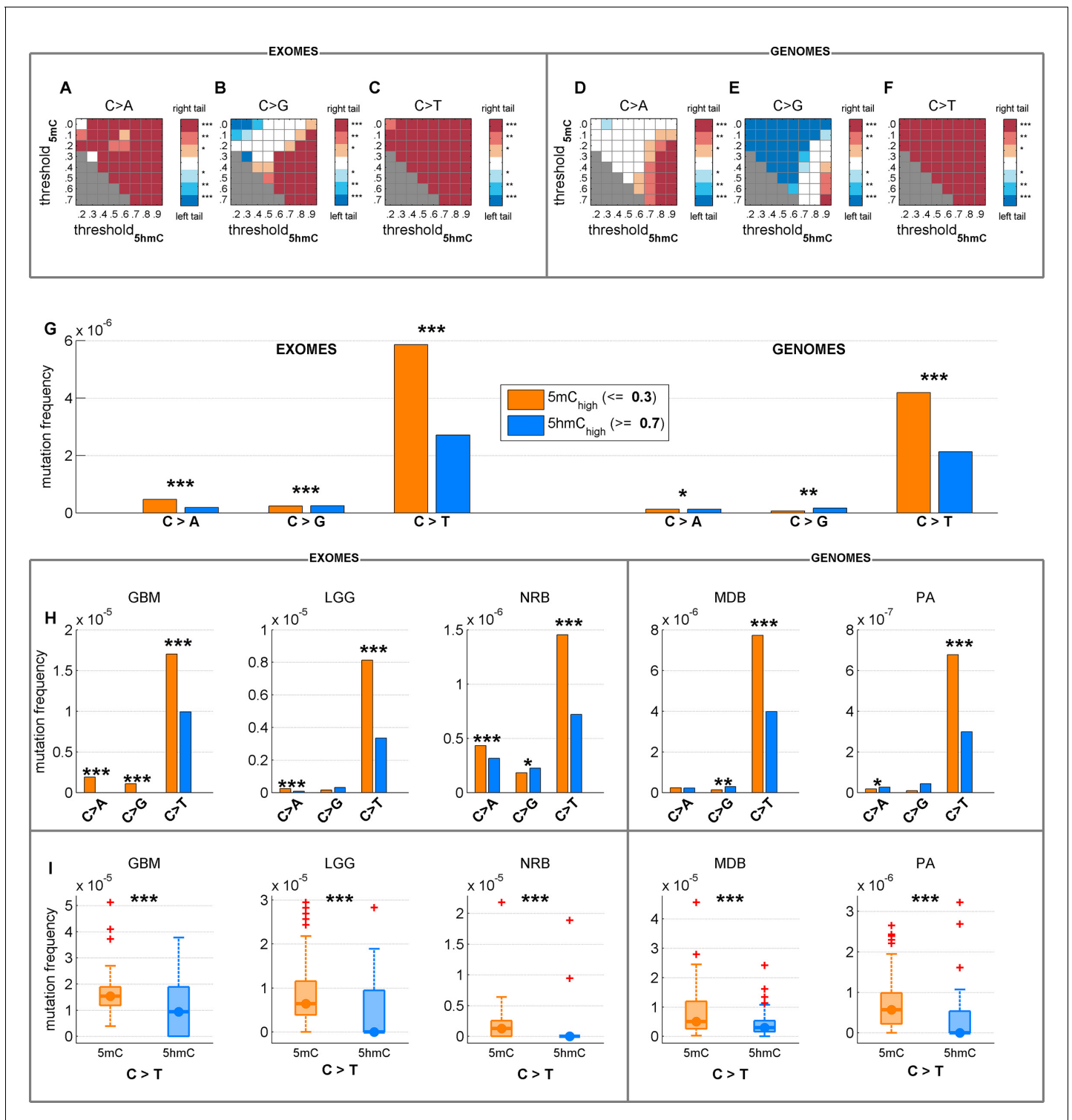


Figure 2—figure supplement 1. Depletion of C>T mutations in 5hmC_{high} is relatively insensitive to varying definitions of 5mC_{high} and 5hmC_{high}. (A–F) Significance of a difference in mutation frequency in 5mC_{high} and 5hmC_{high}, for a range of values of threshold_{5mC} and threshold_{5hmC} (5mC_{high} is defined as sufficiently modified sites with 5hmC_{rel} \leq threshold_{5mC}; 5hmC_{high} is defined as sufficiently modified sites with 5hmC_{rel} \geq threshold_{5hmC}). One-sided paired Wilcoxon sign-rank test was used. Red colour represents a significant increase of mutation frequency in 5mC_{high} (right tail test) whereas blue colour represents elevated mutations in 5hmC_{high} (left tail test). (G–I) C>T mutation frequency for 5mC_{high} vs. 5hmC_{high} with threshold_{5mC} = 0.3 and threshold_{5hmC} = 0.7.

DOI: [10.7554/eLife.17082.008](https://doi.org/10.7554/eLife.17082.008)

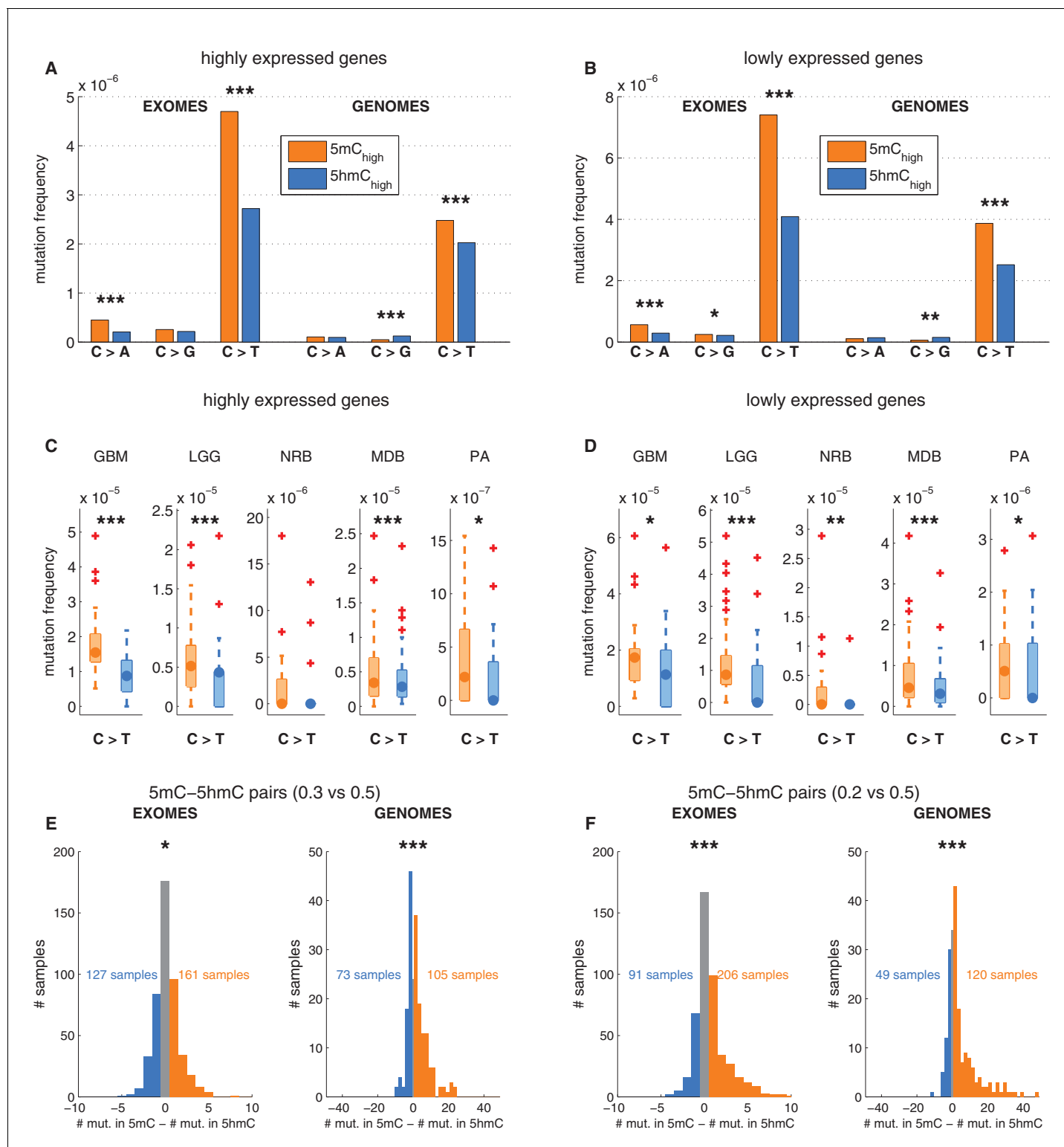


Figure 3. Depletion of C>T mutations in 5hmC sites is not explained by gene expression or regional mutation rate variation. (A–B) Frequency of mutations in 5mC_{high} vs 5hmC_{high} sites within highly expressed (A) or lowly expressed (B) genes (see Materials and methods). (C–D) Boxplot visualisation of C>T mutation frequency for each cancer type. (E) For each patient sample, the overall difference in mutations in paired sites was calculated and compared using a Wilcoxon signed-rank test. Shown here is a histogram of samples by the difference in mutations for paired 5mC and 5hmC sites (negative values shown blue, positive in orange; see Materials and methods for details). Mutations in 5mC sites exceed paired 5hmC sites, causing a shift to the right. (F) Same as E but using a more stringent definition of 5mC (only sites with threshold_{5mC} ≤ 0.2).

Figure 3 continued on next page

Figure 3 continued

DOI: [10.7554/eLife.17082.009](https://doi.org/10.7554/eLife.17082.009)

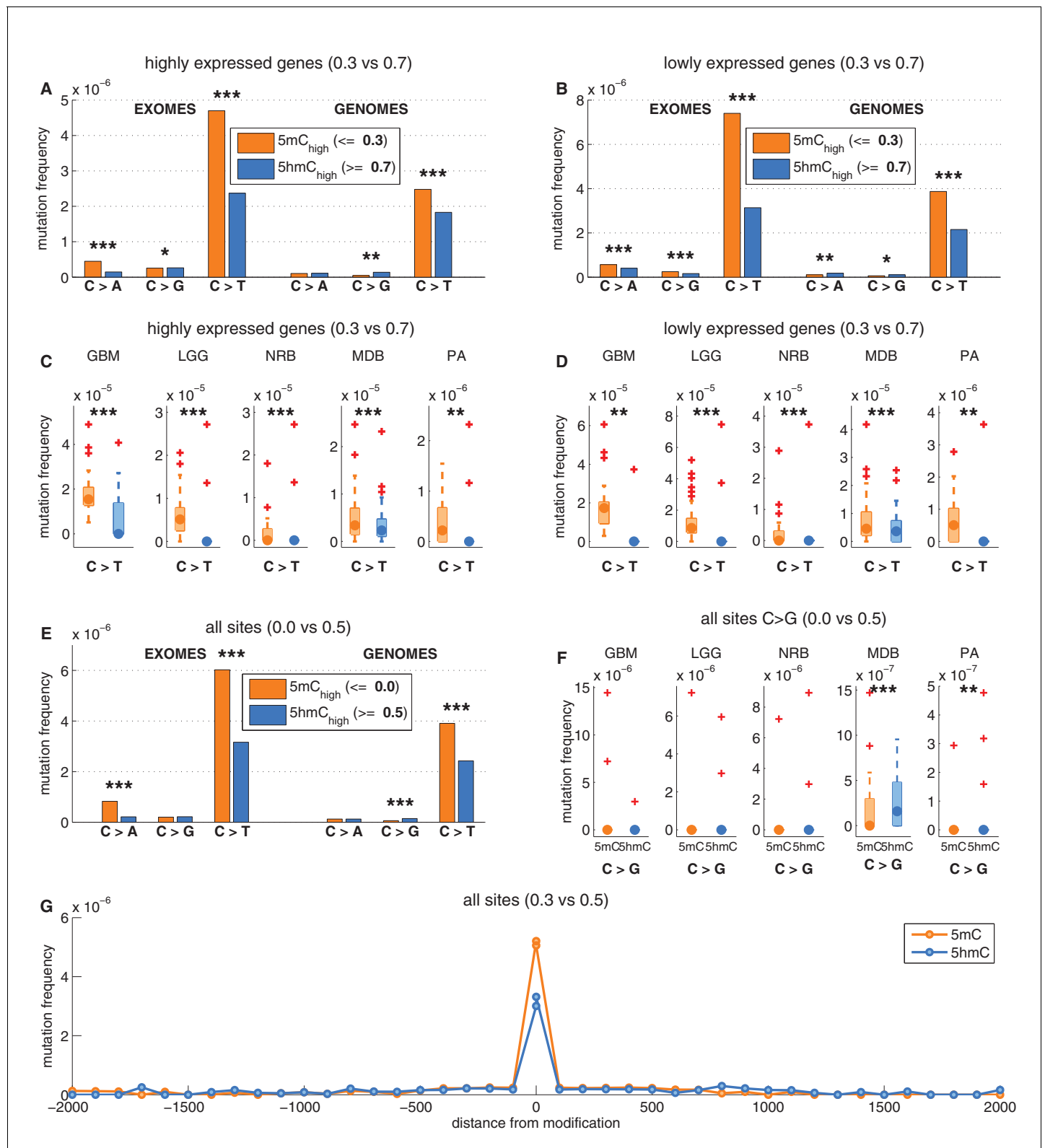


Figure 3—figure supplement 1. Depletion of C>T mutations in 5hmC_{high} is relatively insensitive to varying definitions of 5mC_{high} and 5hmC_{high}. (A–D) C>T mutation frequency for 5mC_{high} vs. 5hmC_{high} in highly vs. lowly expressed genes with threshold_{5mC} = 0.3 and threshold_{5hmC} = 0.7. (E–F) C>G mutation frequency with threshold_{5mC} = 0.0 and threshold_{5hmC} = 0.5. (G) Mutation frequency around aligned 5mC and 5hmC sites.

DOI: [10.7554/eLife.17082.010](https://doi.org/10.7554/eLife.17082.010)

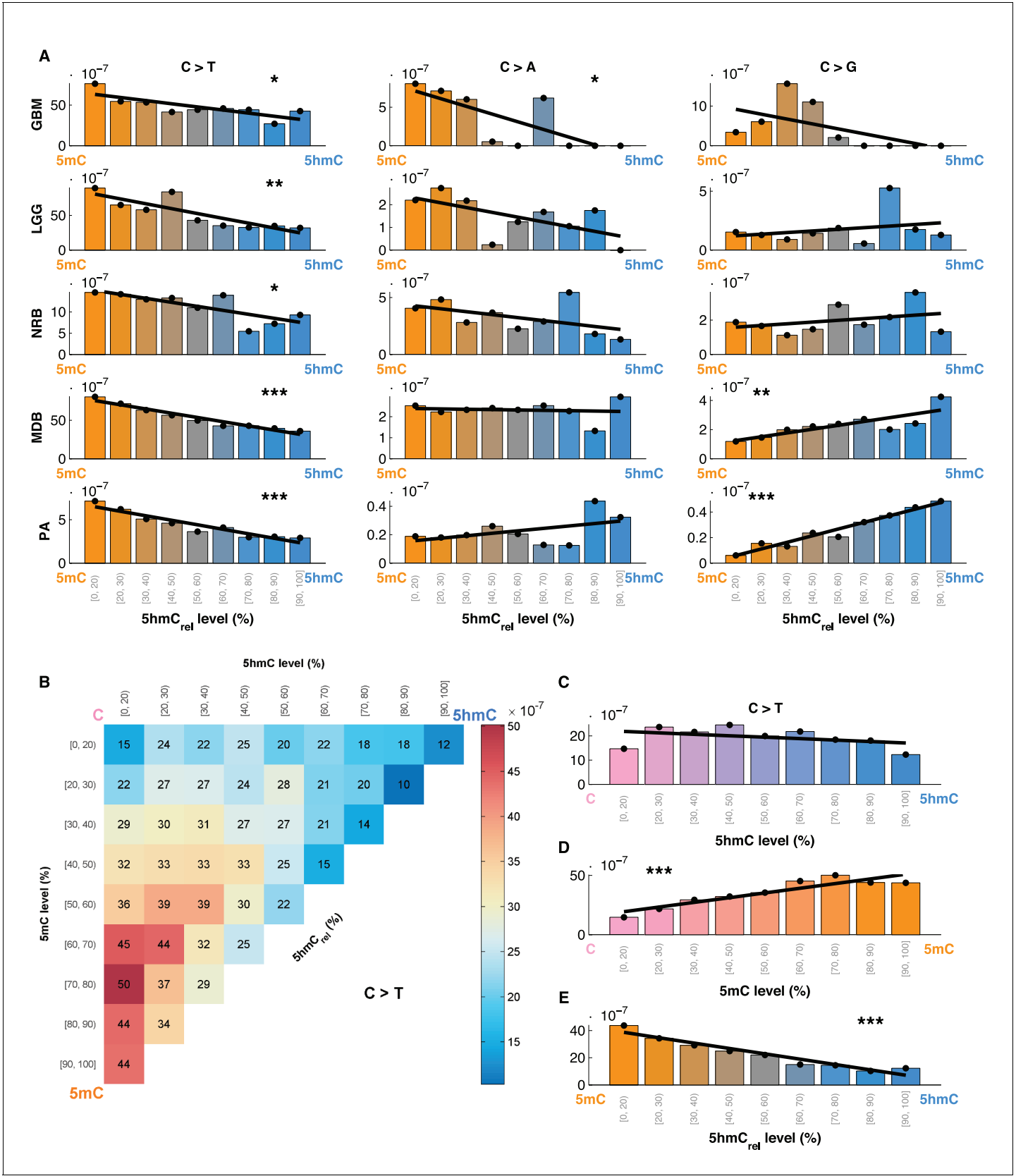


Figure 4. Mutation frequency negatively correlates with 5hmC_{rel} level per base. (A) Fraction of mutated CpG sites as a function of 5hmC_{rel} levels by mutation and cancer type. Bins to the left represent sites predominantly methylated, while bins to the right contain increasingly hydroxymethylated

Figure 4 continued on next page

Figure 4 continued

sites. Black line denotes linear regression fit (F-test for coefficient deviation from 0, see Materials and methods). **(B)** Distribution of CpG>T mutation frequency by modification type. The top left bin contains cytosines that are mostly unmodified, the bottom left bin contains exclusively methylated cytosines and the top right bin contains cytosines that are mostly hydroxymethylated. **(C)** Top row of B, i.e. distribution of mutations in unmethylated sites. **(D)** First column of B, i.e. distribution of mutations in sites without 5hmC. **(E)** Diagonal of B, i.e. distribution of mutations in highly modified sites. DOI: [10.7554/eLife.17082.011](https://doi.org/10.7554/eLife.17082.011)

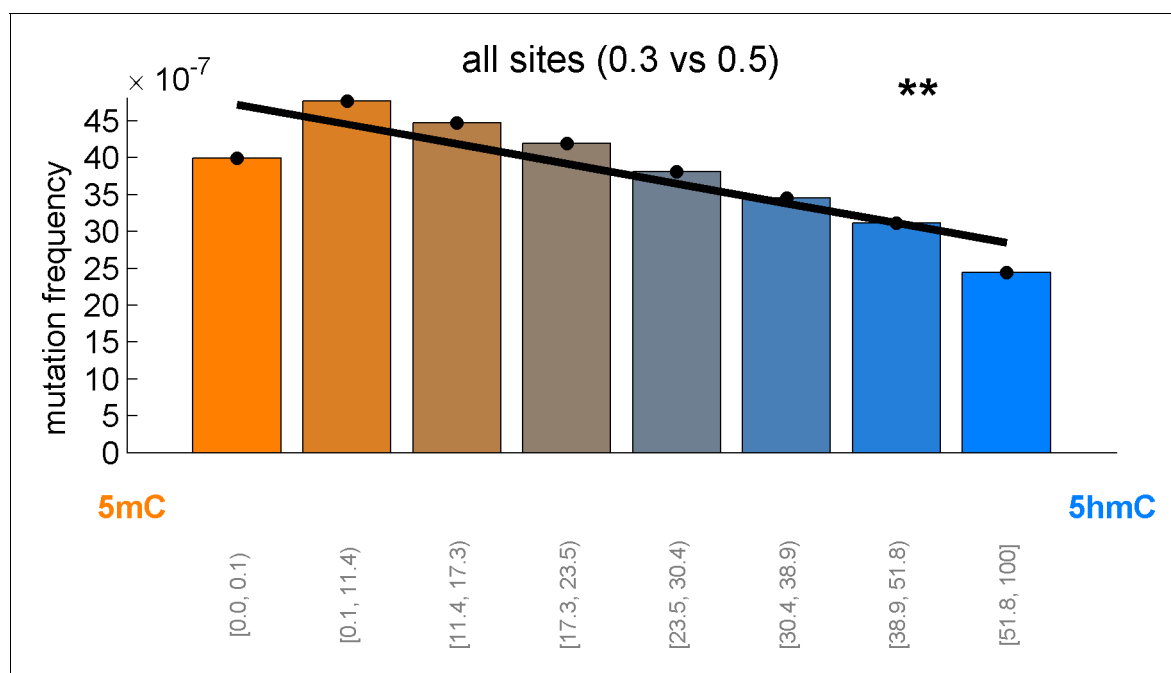


Figure 4—figure supplement 1. CpG>T mutation frequency as a function of 5hmC_{rel} levels with equal binning (each bin contains approximately the same number of sites).

DOI: [10.7554/eLife.17082.012](https://doi.org/10.7554/eLife.17082.012)

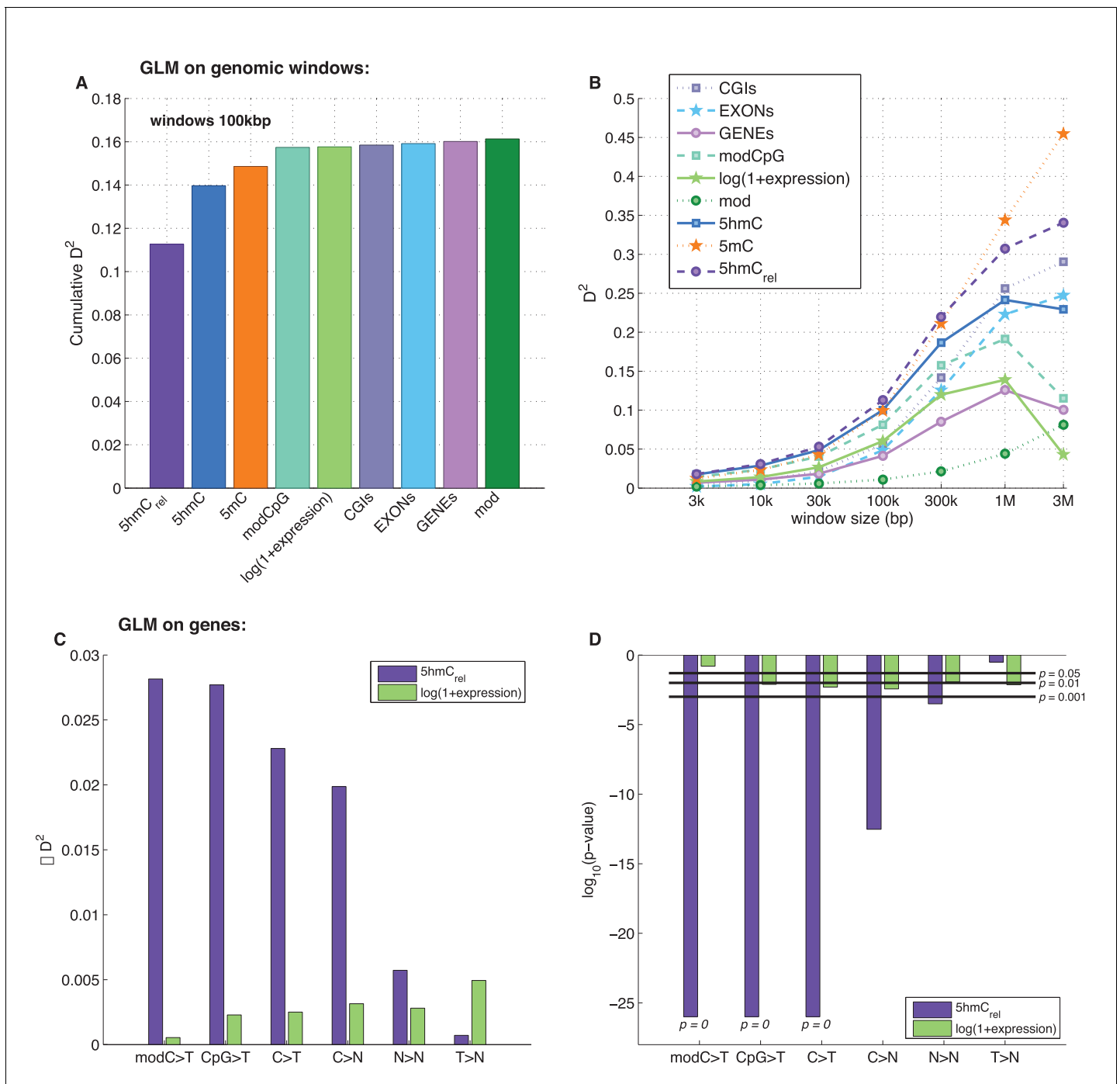


Figure 5. Predictors of mutations: 5hmC_{rel} compared to other genomic features. (A) Prediction of CpG>T mutation frequency (using whole genome sequencing only) in 100 kbp genomic windows. Predictors are sorted according to the D^2 in a univariate model. The height of the k^{th} bar denotes the D^2 of a model with the first k predictors. (B) Comparison of the nine predictors of CpG>T mutation features by D^2 in a univariate models, in a range of window sizes. (C) Prediction of different types of mutation frequency in genes. Increase in D^2 of a generalised linear model including 5hmC_{rel} over gene expression (purple) or gene expression over 5hmC_{rel} (green) (see Materials and methods). (D) Significance of observations in C (see Materials and methods).

DOI: [10.7554/eLife.17082.013](https://doi.org/10.7554/eLife.17082.013)

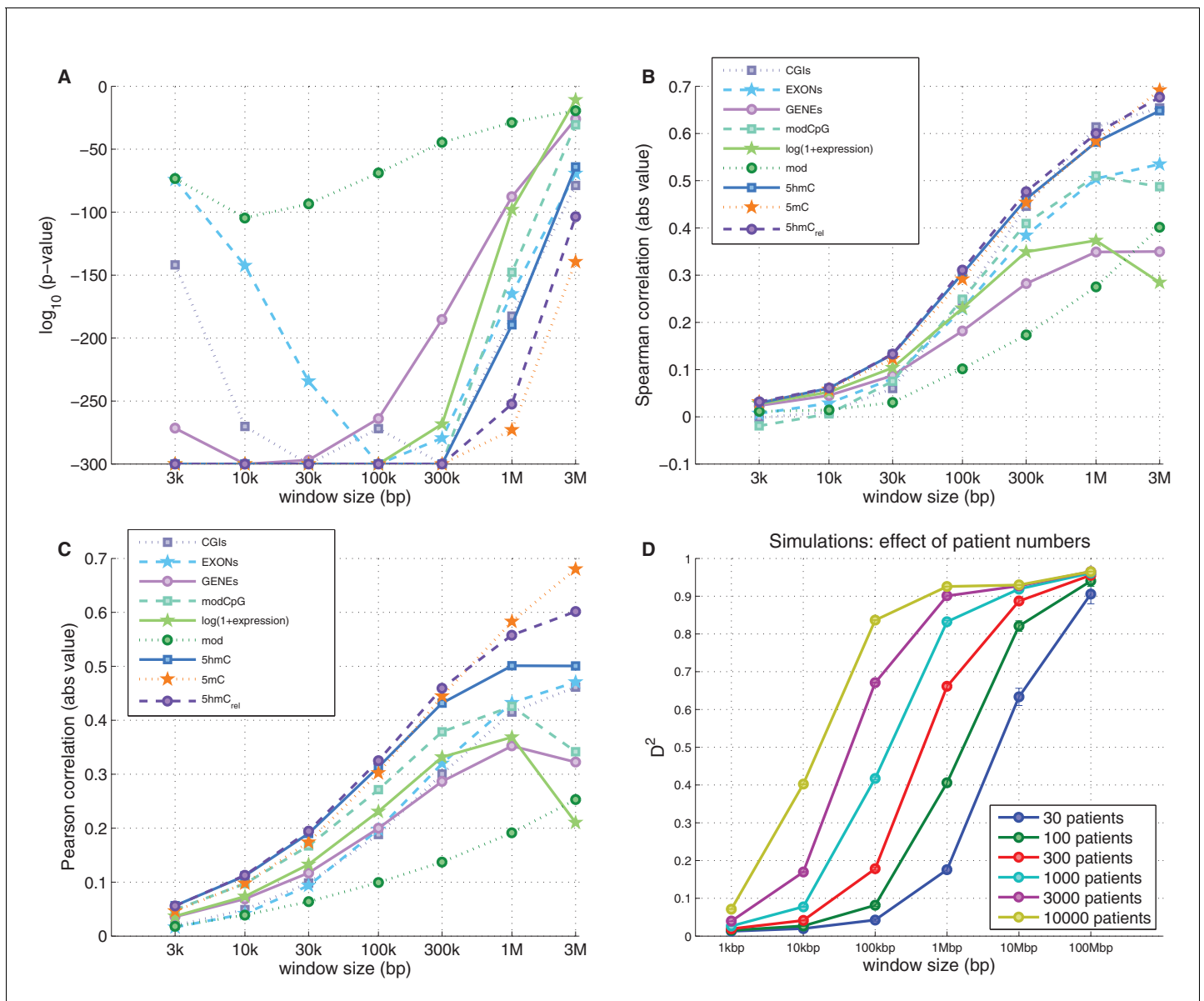


Figure 5—figure supplement 1. Genome-wide prediction of CpG>T mutation frequency: 5hmC_{rel} compared to other genomic features. (A–C) Comparison of nine predictors of CpG>T mutation frequency in a range of window sizes by p-value of univariate generalised linear models (A), Spearman correlation (B), and Pearson correlation (C). (D) Effects of window size and patient numbers on D^2 of GLM with one response variable (simulated mutation frequency) generated proportionally from a single ideal predictor (see Materials and methods for details).

DOI: [10.7554/eLife.17082.014](https://doi.org/10.7554/eLife.17082.014)

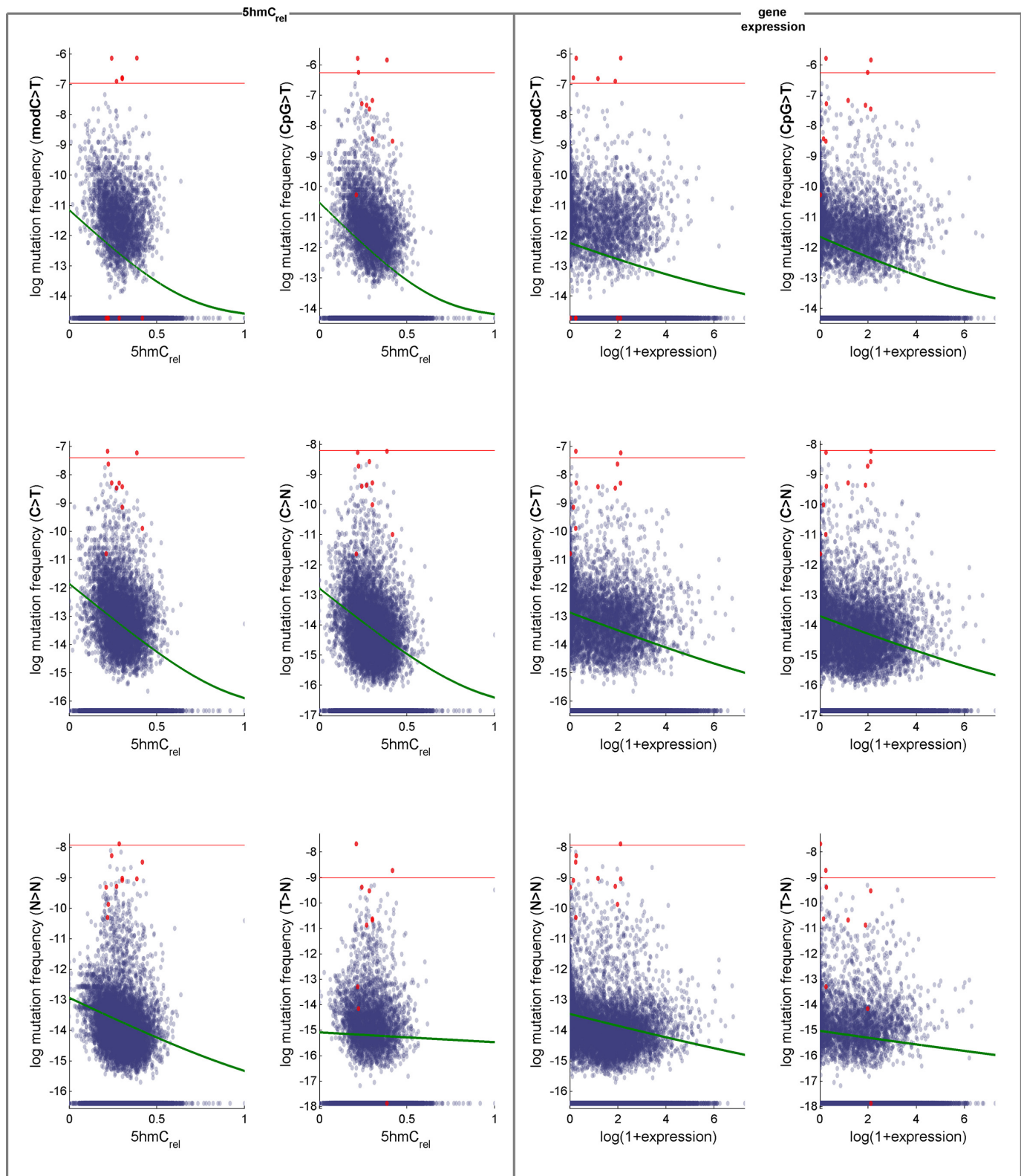


Figure 5—figure supplement 2. Effects of 5hmC_{rel} levels on gene mutability. Data for GLM with Poisson distribution (the fitted curve is in green). Genes defined as outliers in at least one definition of mutation frequency (above the red line) are plotted in red. For convenience, the mutation frequency is plotted on log-scale.

DOI: [10.7554/eLife.17082.015](https://doi.org/10.7554/eLife.17082.015)

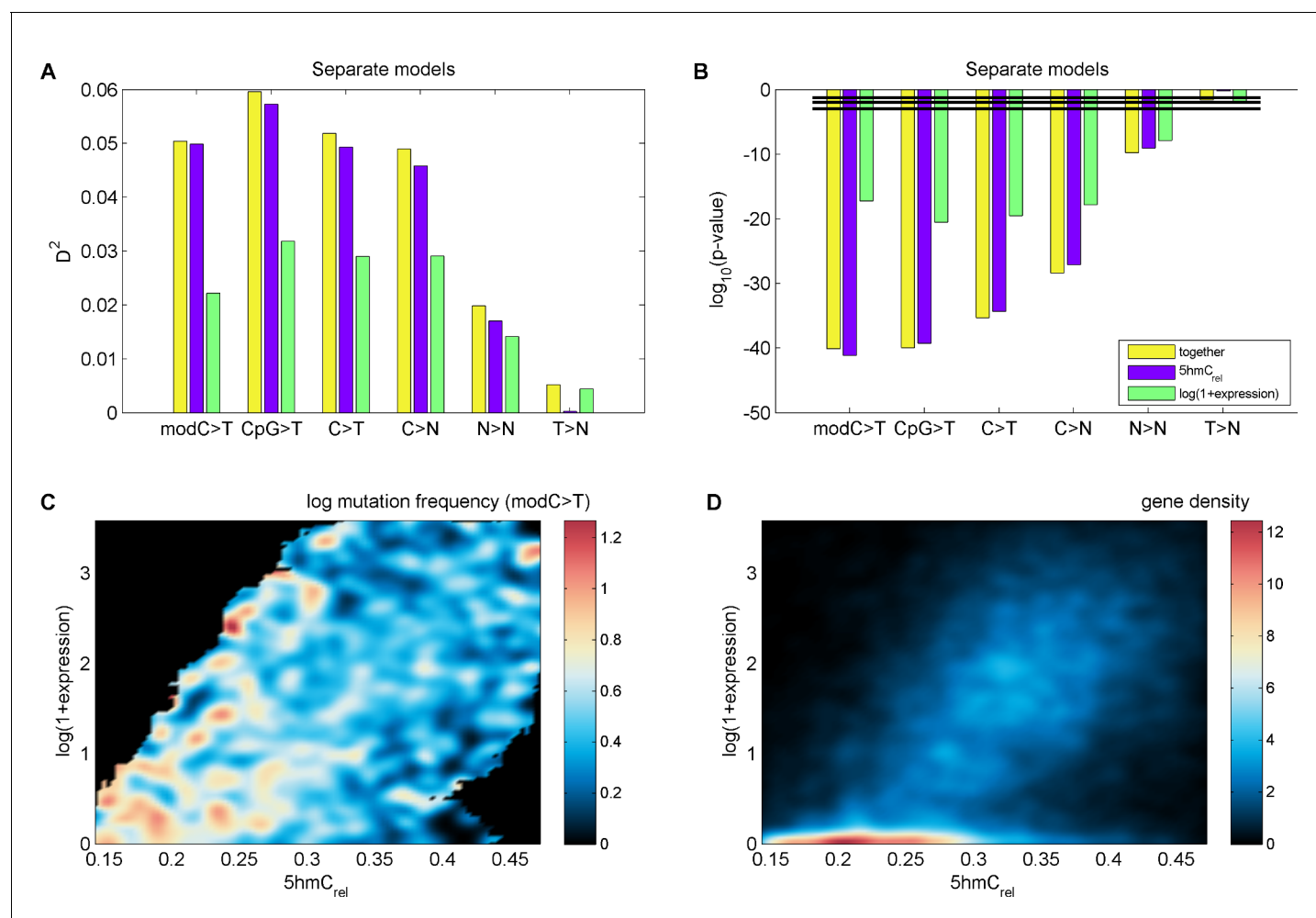


Figure 5—figure supplement 3. Effects of 5hmC_{rel} levels on gene mutability. (A–B) GLM results fitted separately for 5hmC_{rel} (purple) and gene expression (green) and both of them together (yellow). (C–D) Frequency of modC>T mutations of all genes (C) and gene density (D) in the space of 5hmC_{rel} and gene expression. Briefly, for figures C and D the space was limited to [quantile(x, 0.05), quantile(x, 0.95)] on both axes and then binned into 100x100 bins. In each bin, the average mutation frequency (in the form of $\log(\text{mutFreq} + \min(\text{mutFreq}(\text{mutFreq} > 0)))$) was computed. The resulting matrix was smoothed by applying a Gaussian filter (radius 5 bins, sigma 2) weighted by the number of genes in each bin (bins with $\geq 2/3$ missing values in their neighbourhood were set to NaN) and plotted with pcolor (NaN bins are shown in black).

DOI: [10.7554/eLife.17082.016](https://doi.org/10.7554/eLife.17082.016)

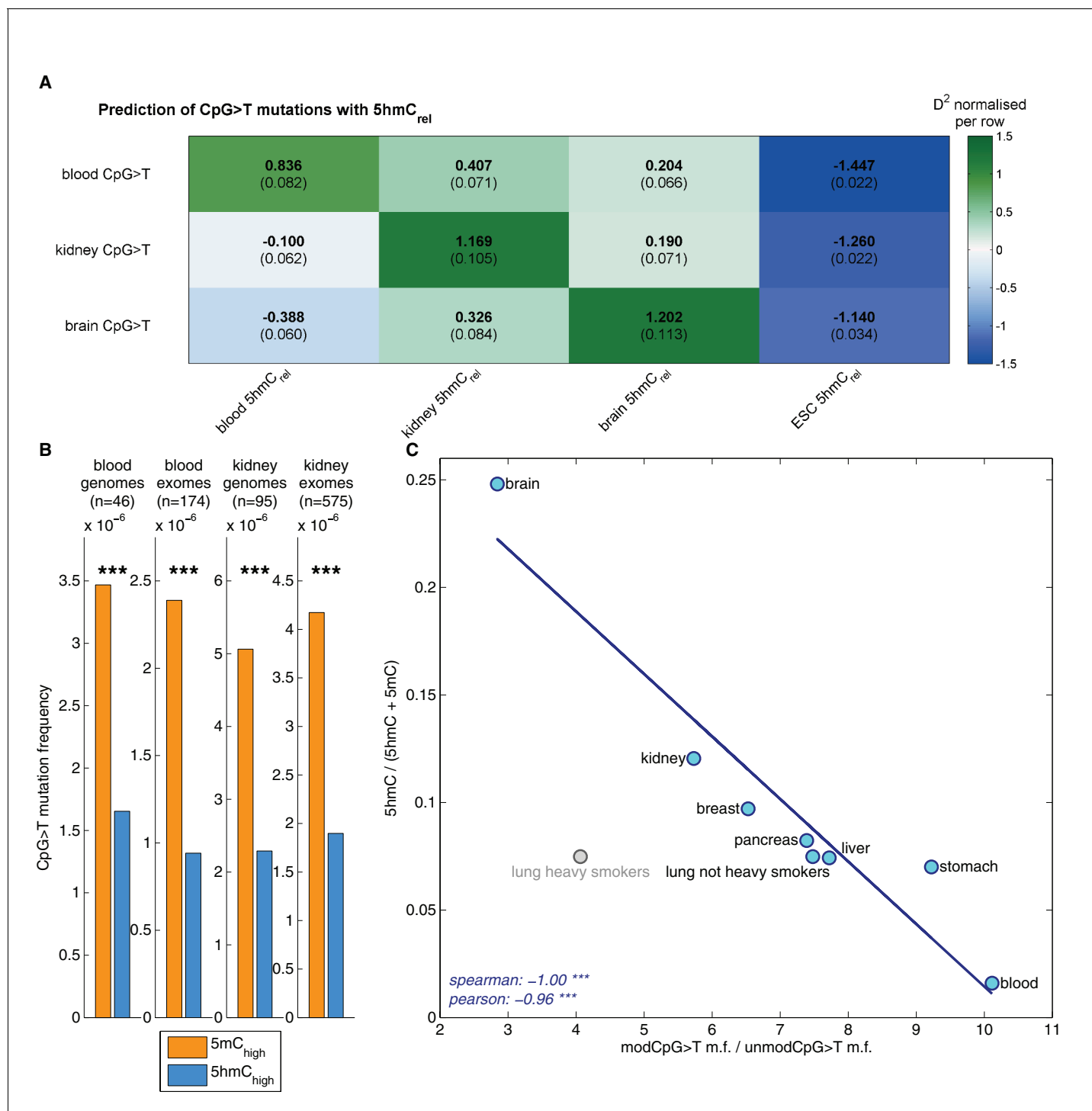


Figure 6. Decreased CpG>T mutation frequency in 5hmC is not limited to brain tissue. (A) Predictions of CpG>T mutation frequency in whole genome cancers in blood (AML), kidney and brain using 5hmC_{rel} maps from blood, kidney, brain and embryonic stem cells (ESC) in 100 kbp genomic windows. The values are z-score normalised per rows in order to normalise for different number of patients and mutations in each cancer type (the original D² values are in parentheses); the higher values of D² (green), the better predictions. (B) CpG>T mutation frequency in 5mC vs. 5hmC in kidney and blood. (C) Correlation of total 5hmC_{rel} levels (measured with HPLC) with frequency of CpG>T mutations in modified cytosines normalised by the frequency in unmodified cytosines in different tissues (see Materials and methods).

DOI: 10.7554/eLife.17082.017

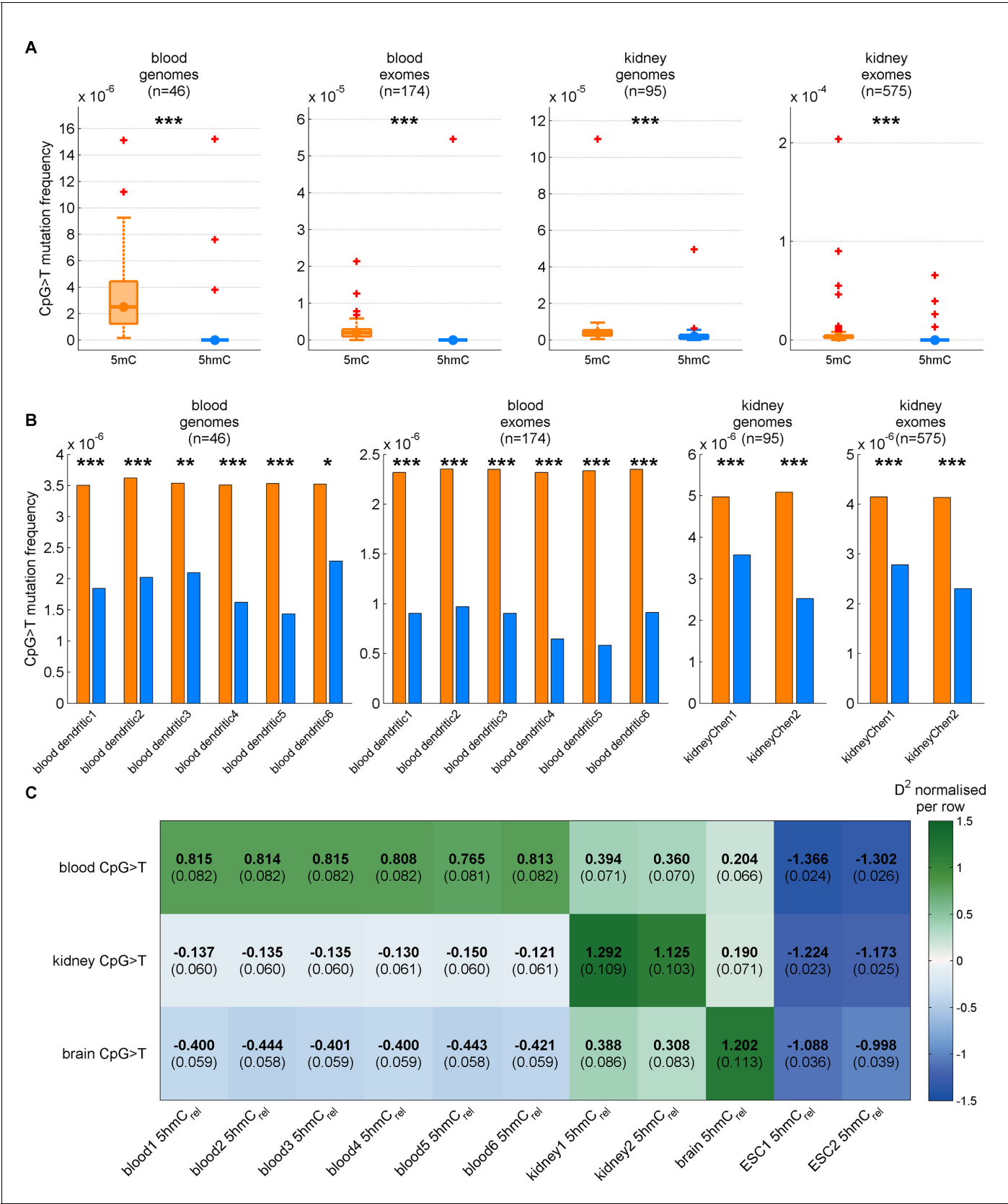


Figure 6—figure supplement 1. Decreased CpG>T mutation frequency in 5hmC is present in three tissues consistently for different replicates of modification maps. (A) CpG>T mutation frequency in 5mC compared to 5hmC in blood and kidney using modification maps from different replicates

Figure 6—figure supplement 1 continued on next page

Figure 6—figure supplement 1 continued

merged together (A) and used separately (B). (C) Predictions of CpG>T mutation frequency in whole genome cancers in blood (AML), kidney and brain using different replicates of 5hmC_{rel} maps from blood, kidney, brain and embryonic stem cells (ESC) in 100 kbp genomic windows. The values are z-score normalised per rows in order to normalise for different number of patients and mutations in each cancer type (the original D² values are in parentheses); the higher values of D² (green colour), the better predictions.

DOI: [10.7554/eLife.17082.018](https://doi.org/10.7554/eLife.17082.018)

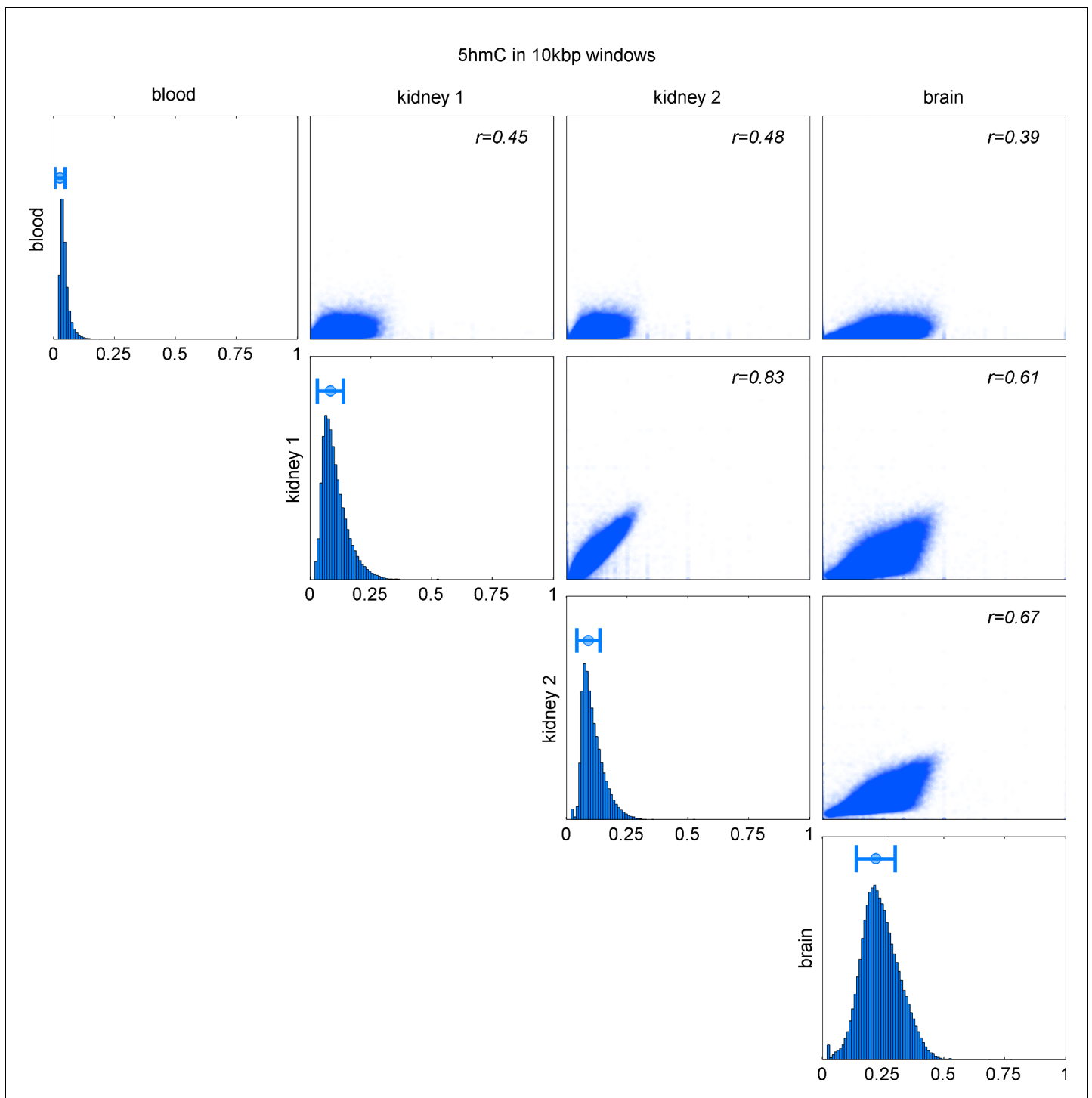


Figure 6—figure supplement 2. Comparison of 5hmC in 10 kbp windows in blood, kidney (2 replicates), and brain. Distribution of 5hmC values in each map and Pearson correlation of pairs of maps.

DOI: [10.7554/eLife.17082.019](https://doi.org/10.7554/eLife.17082.019)

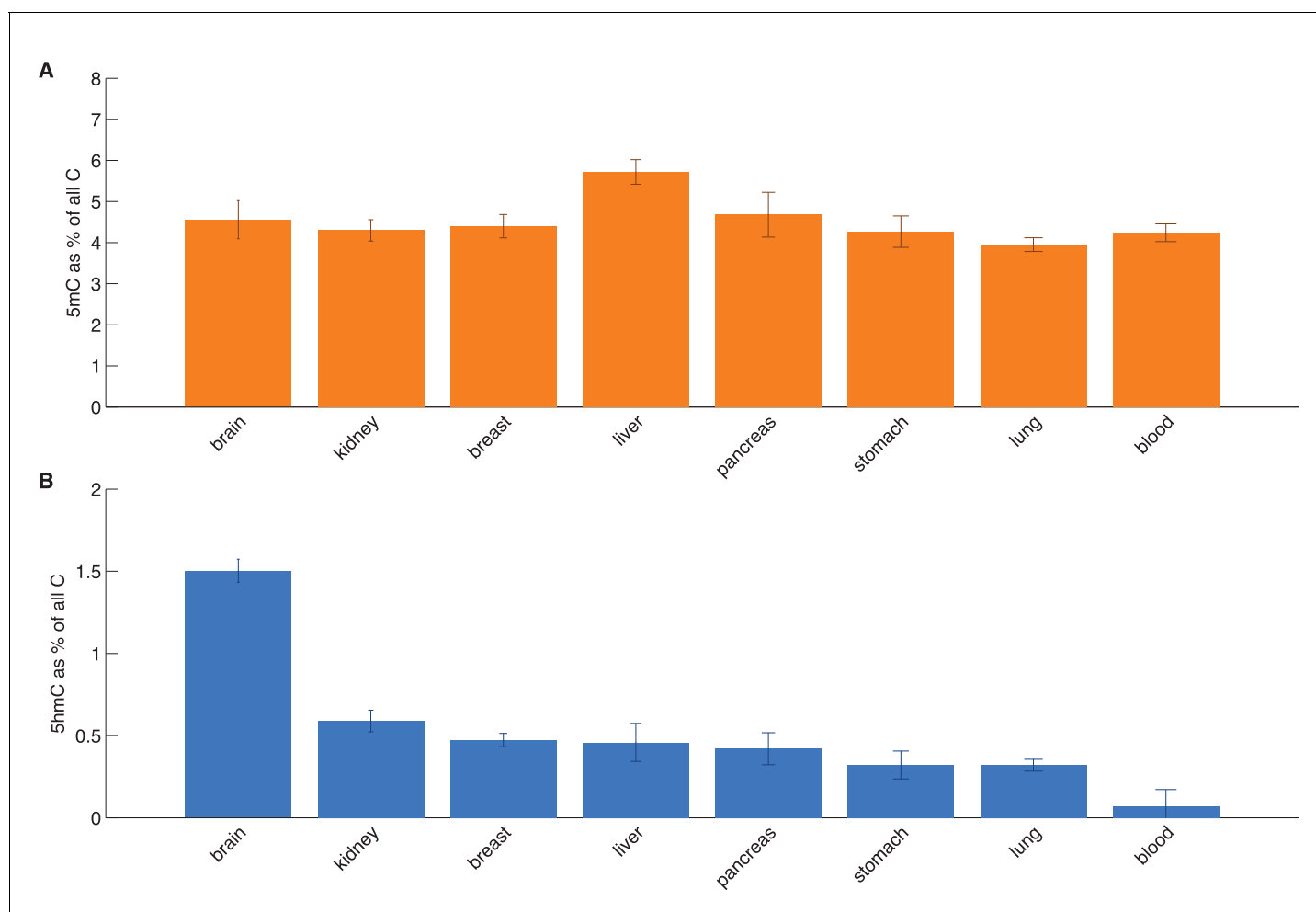


Figure 6—figure supplement 3. HPLC measurements of total 5mC and 5hmC in eight tissues. Average values with standard deviation of 5mC and 5hmC (as a percentage of total cytosine).

DOI: [10.7554/eLife.17082.020](https://doi.org/10.7554/eLife.17082.020)