Supplementary Materials

GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction

Valentina Iotchkova[1,2], Graham R.S. Ritchie[1,2], Matthias Geihs[1], Sandro Morganella[2], Josine L. Min[3], Klaudia Walter[1], Nicholas Timpson[3], UK10K Consortium, Ian Dunham[2], Ewan Birney[2] and Nicole Soranzo[1,4,5]

1. Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1HH, United Kingdom; 2. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom; 3. MRC Integrative Epidemiology Unit, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol, BS8 2BN, United Kingdom; 4. Department of Haematology, University of Cambridge, Cambridge CB2 0AH, United Kingdom; 5. The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge.

Contents

1 Introduction

In this Supplementary Text we give further information regarding method validation and specifications, and the real data enrichment analysis results.

2 Effective number of annotations

For each of the 27 GWAS studies (Online Methods), we downloaded summary statistics, which we used for greedy pruning of the sets of genetic variants per trait (Figure 1). Next, we annotated each variant based on DNaseI hypersensitive site overlap (or LD $r^2 > 0.8$ with such a variant). In order to calculate the effective number of independent annotations we adapted the approach proposed in Galwey, 2009[1]. Specifically, for each phenotype we took the binary (annotated/not annotated) matrix and calculated the eigenvalues λ_i of its correlation matrix. Then the effective number of features was defined as $M_{eff} = (\sum_{i=1}^{M} \sqrt{\lambda_i})^2/(\sum_{i=1}^{M} \lambda_i)$, where M = 424 was the total number of annotations (DNaseI hypersensitive sites) used. As a result we found between 186 and 194 effective numbers of annotations for the 27 traits (Figure S1) and choose to use the most conservative for a Bonferroni correction of all traits, which gives -log10 P-value threshold of 3.59 at the 95% significance level.

3 GWAS P-value dependence on the number of LD proxies

During exploratory data analysis, we investigated whether the GWAS association p-value was correlated to the number of LD proxies ($r^2 > 0.8$) a variant would have. Figure S2, shows the results for an example trait (HGB), which as expected, indicates a clear increase in significance of the p-values as the number of proxies increase. Similar patterns were observed for all other traits and are not shown here.

4 Enrichment of GWAS variants in DNaseI hypersensitive sites

For each of the 27 GWAS studies, we performed enrichment analysis for each 424 ENCODE and Roadmap Epigenomics cell type using GARFIELD such that we calculated the Fold Enrichment at GWAS thresholds ranging from $T < 10^{-1}$ to $T < 10^{-8}$ (in powers of 10) and then tested that enrichment at the four most significant thresholds, from $T < 10^{-5}$ to $T < 10^{-8}$. Figures S3, S4 and S5 present enrichment wheel plots for each trait. Results highlight sets of traits with cell type specific enrichment as well as ones with an overall enrichment.

5 GARFIELD memory usage and CPU time

Maximum memory usage and CPU time estimates for the pruning and annotation stage and the permutation stage of GARFIELD are reported in Figure S6, where the latter has been estimated by a general approach (100 000 permutations) as well as a fast approximation based on an adaptive number of permutations (min 100) so that no more permutations are performed if the enrichment p-value can no longer reach significance. Estimates are based on the 27 phenotypes and 424 DHS cell types.

6 Effect of MAF, distance to nearest TSS and number of LD proxies

To investigate the effects of the possible confounding features, we added a correction for each of the features, one at a time, and compared the resulting proportions of significant annotations (DHS data) for each trait to a model that does not account for any of the features. We performed this at the 10^{-8} GWAS threshold in all 27 GWAS datasets. We found that in nearly all settings we got fewer enrichments deemed as significant when using a feature correction (Figure S7), unsurprisingly except for MAF (all GWAS were performed on common variants). This means that a number of these enrichments can be explained by our chosen features and therefore not accounting for them can introduce confounding.

To further investigate the effect of each possible confounding feature relative to accounting for all of them together, we dropped each one at a time and re-run the enrichment analyses for the 10^{-8} GWAS threshold in all 27 GWASs (Figure S8). We found that accounting for each combination of features produces larger number of significant results as opposed to the model which accounts for all three features together, except for the one accounting for number of proxies and distance to nearest TSS (and no MAF) (again due to all variants in our analysis being common).

7 Overlap between HepG2 and Caco-2 DNaseI annotations

We used bedtools to calculate the intersection of HepG2 and Caco-2 peaks, where 1 base pair overlap was used to define a shared peak. As a result we found over 52% of Caco-2 peaks overlapping HepG2 peaks (Figure S9 topleft). Furthermore, we calculated the number of UK10K variants overlapping HepG2 and Caco-2 peaks (Figure S9 topright). Finally, we calculated the same for LDL GWAS variants after LD pruning and proxy annotating (Figure S9 bottomleft) and the subset of independent variants at the 10^{-8} threshold (Figure S9 bottomright).

8 Enrichment in 25 state genome segmentations in 127 cell types

Similarly to the DNaseI hypersensitive site data, GARFIELD enrichment analysis was run for each of the 27 phenotypes and each genome segmentation state and cell type. The number of annotations was $25 \times 127 = 3175$ and the significance p-value threshold after Bonferroni correction for the effective number of annotations was 3.7×10^{-5}, for which 10^6 permutations were run for each cell type.

Volcano plots of the results for the $T < 10^{-5}$ GWAS significance threshold are shown in Figures S9, S10 and S11, and state information and cell type information is given in Supplementary Tables S7 and S6, respectively.

References

[1] Nicholas W Galwey. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic epidemiology*, 33(7):559–68, November 2009.

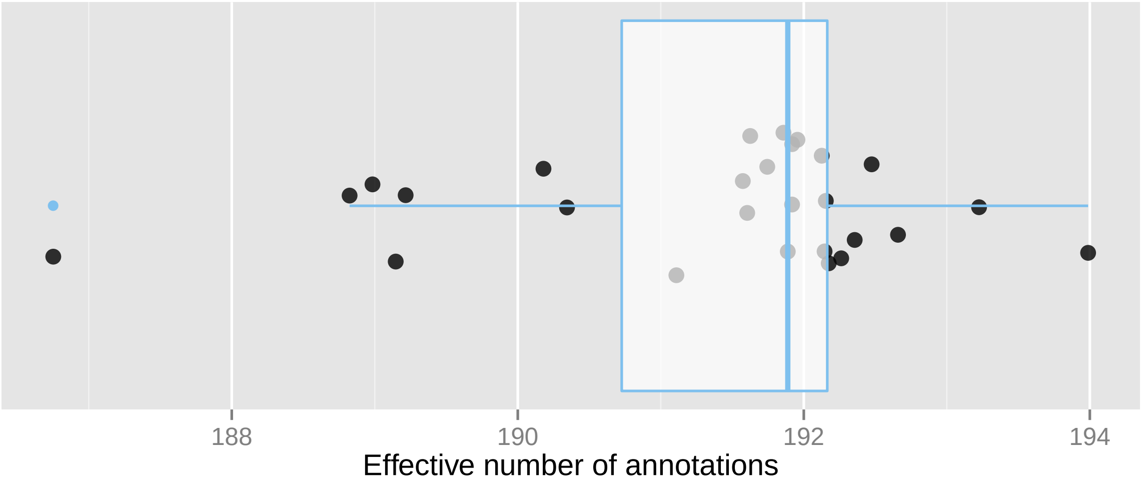


Figure S1: Estimates of the effective number of annotations for 27 phenotypes (dots), out of a total of 424 DNaseI hypersensitive site cell types used.

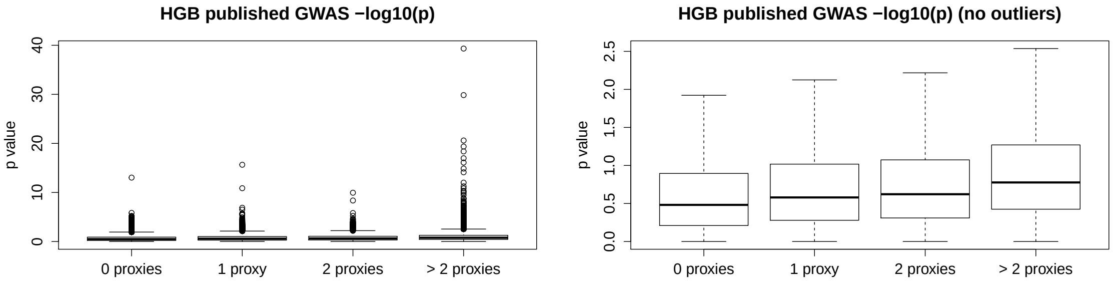
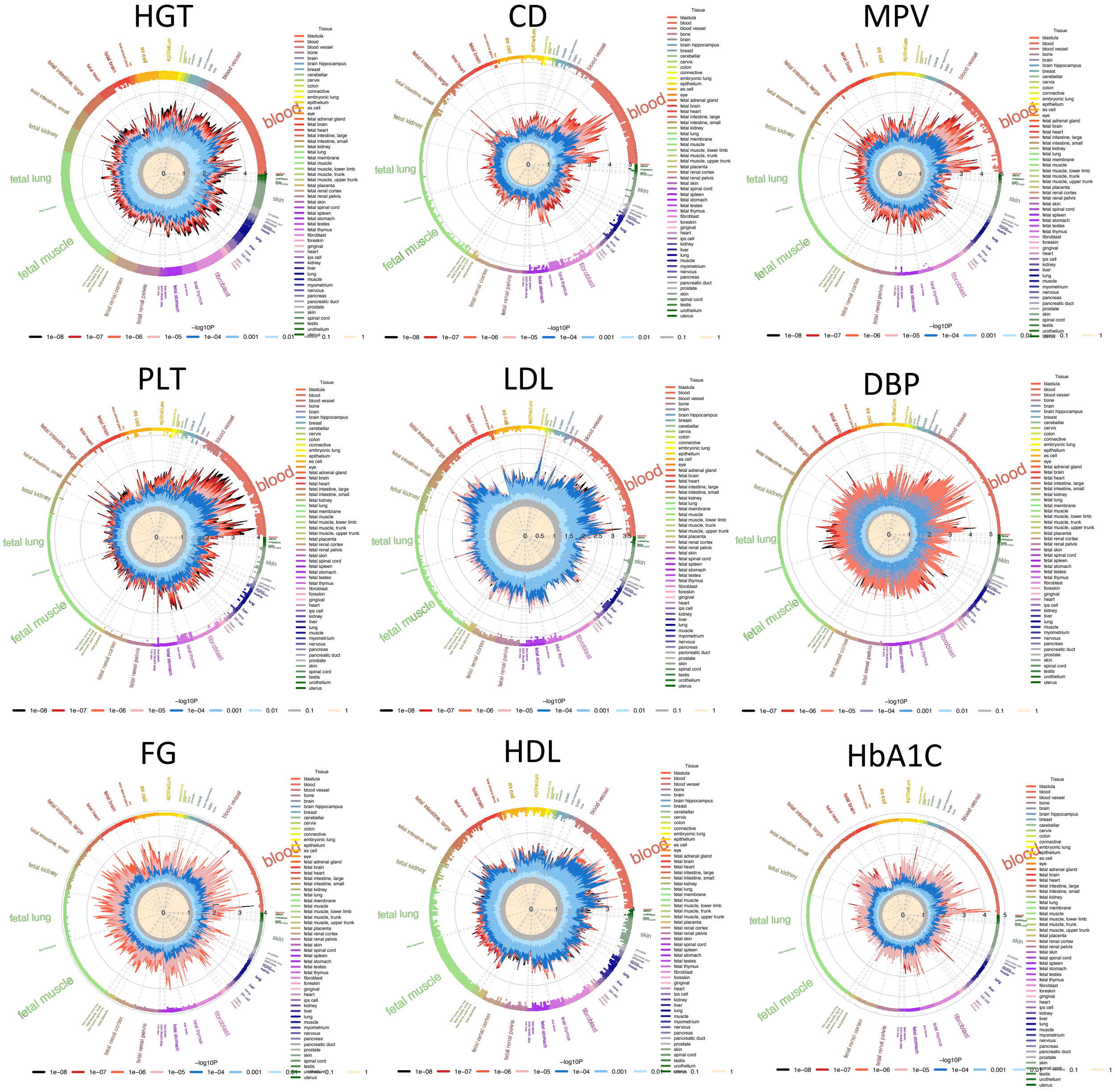


Figure S2: GWAS P-value dependence on the number of high LD proxies ($r^2 >= 0.8$) for the HGB phenotype.


HGT

CD

MPV

PLT

LDL

DBP

FG

HDL

HbA1C

Figure S3: Enrichment of genome-wide association analysis p-values in DNaseI hypersensitive sites (hotspots) for HGT, CD, MPV, PLT, LDL, DBP, FG, HDL and HbA1C. Radial lines show FE values at eight GWAS -log10P-value thresholds (T) for all ENCODE and Roadmap Epigenomics DHS cell lines, sorted by tissue on the outer circle. Dots in the outer side of the circle denote significant enrichment (if present) at $T < 10^{-5}$ (outermost) to $T < 10^{-8}$ (innermost).

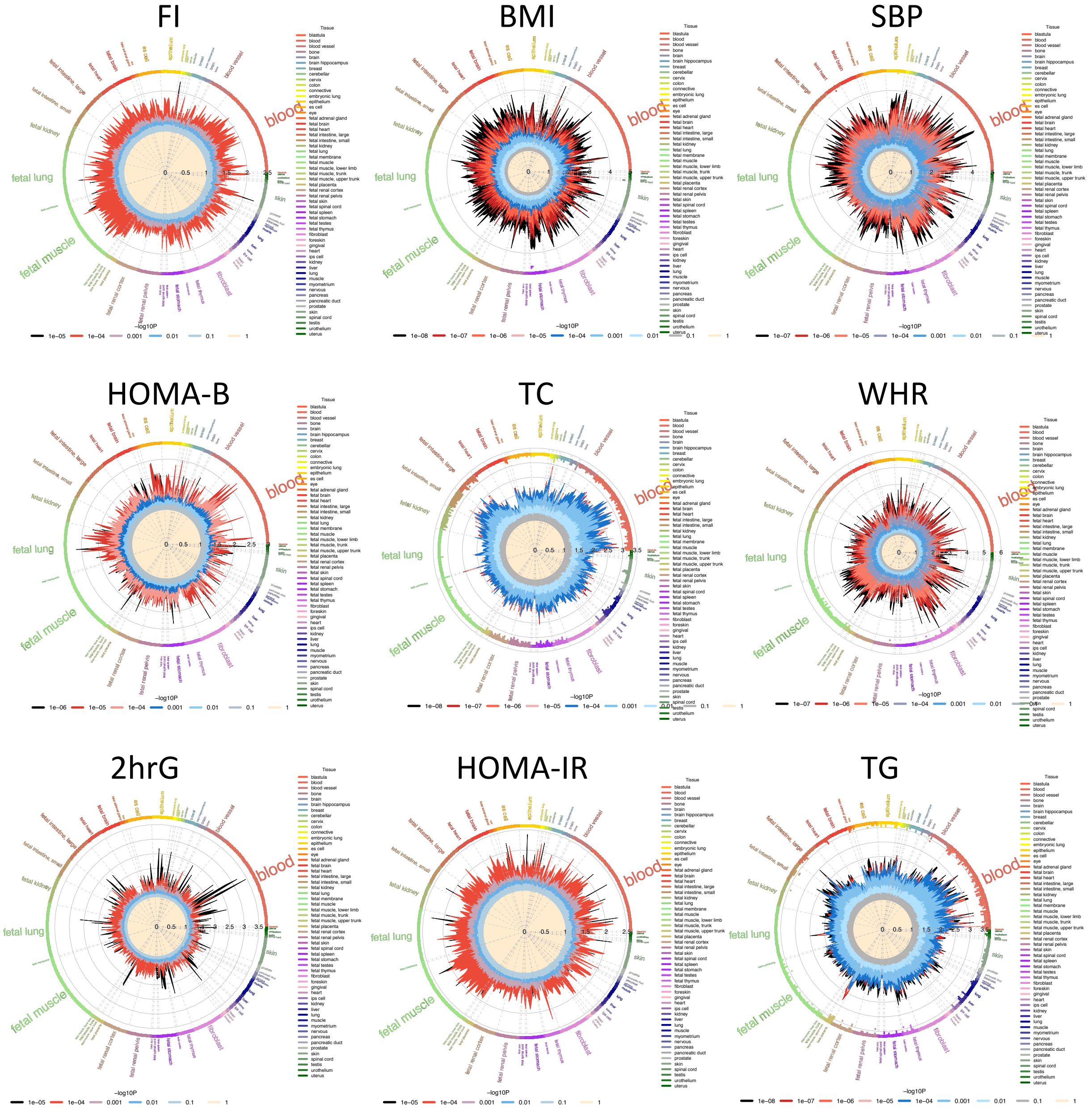
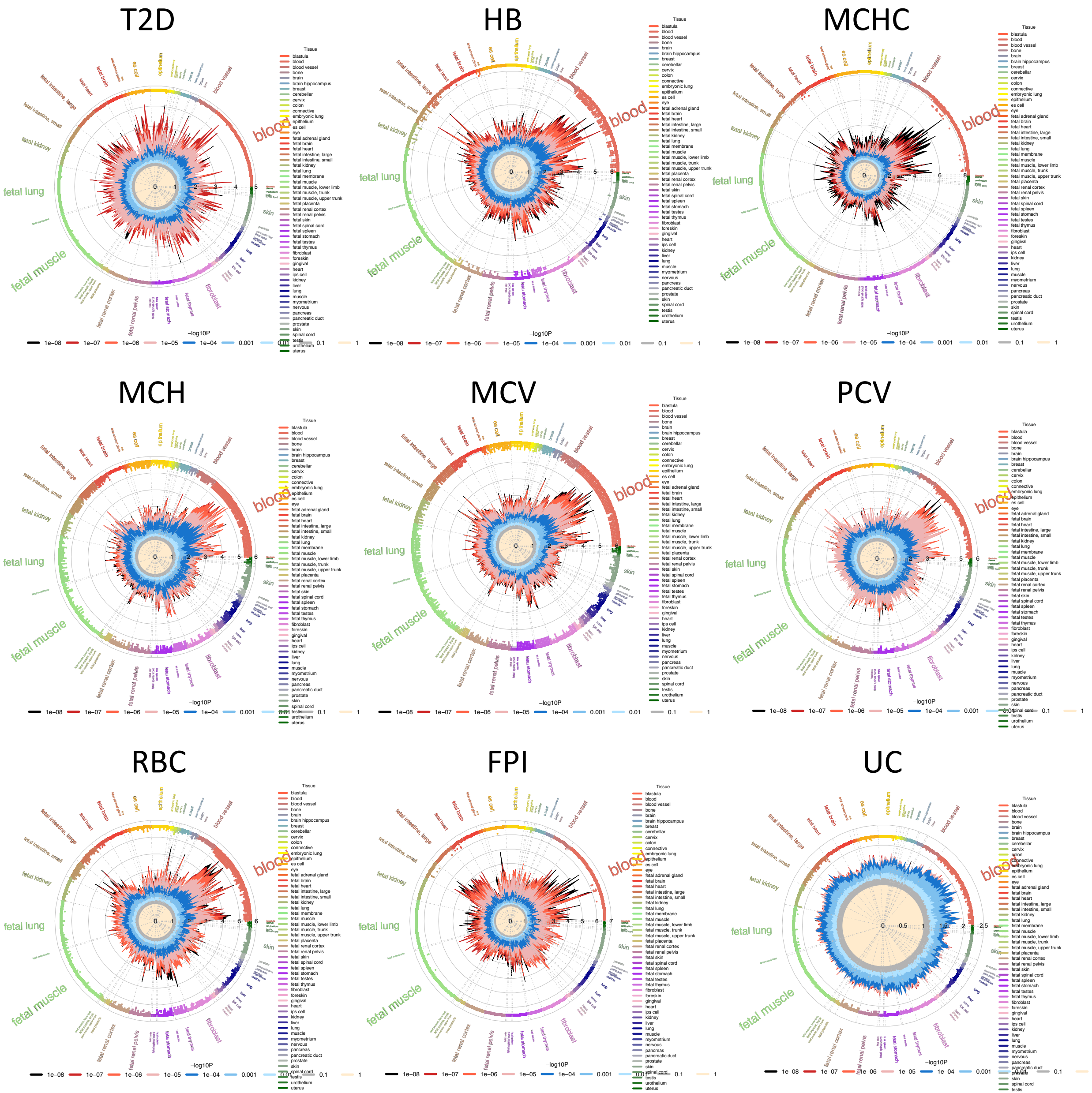


Figure S4: Enrichment of genome-wide association analysis p-values in DNaseI hypersensitive sites (hotspots) for FI, BMI, SBP, HOMA-B, TC, WHR, 2hrG, HOMA-IR and TG. Radial lines show FE values at eight GWAS -log10P-value thresholds (T) for all ENCODE and Roadmap Epigenomics DHS cell lines, sorted by tissue on the outer circle. Dots in the outer side of the circle denote significant enrichment (if present) at $T < 10^{-5}$ (outermost) to $T < 10^{-8}$ (innermost).


T2D HB MCHC

MCH MCV PCV

RBC FPI UC

Figure S5: Enrichment of genome-wide association analysis p-values in DNaseI hypersensitive sites (hotspots) for T2D, HB, MCHC, MCH, MCV, PVC, RBC ,FPI and UC. Radial lines show FE values at eight GWAS -log10P-value thresholds (T) for all ENCODE and Roadmap Epigenomics DHS cell lines, sorted by tissue on the outer circle. Dots in the outer side of the circle denote significant enrichment (if present) at $T < 10^{-5}$ (outermost) to $T < 10^{-8}$ (innermost).

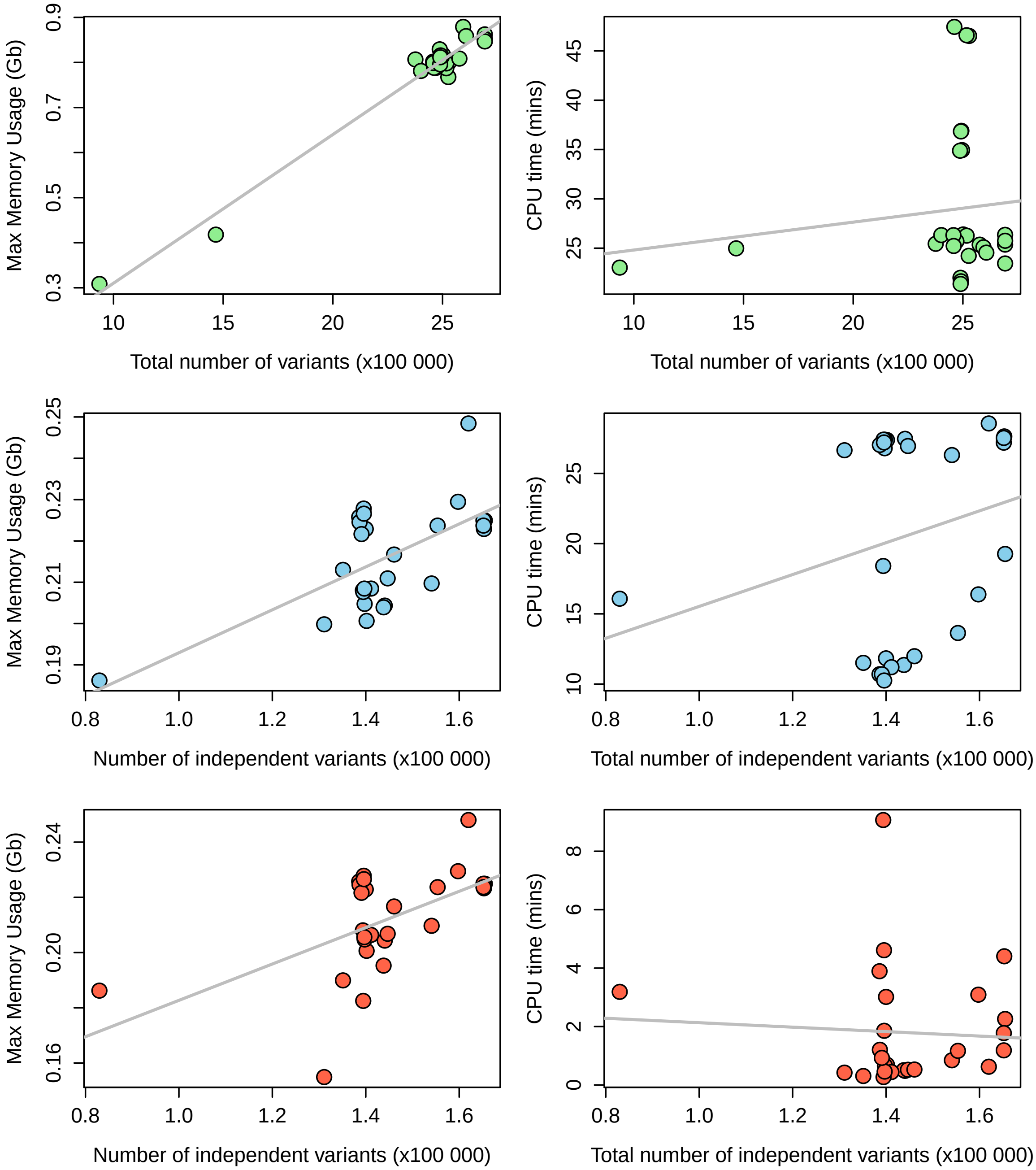


Figure S6: GARFIELD CPU time and maximum memory usage for the pruning step (Top), general FE and P-value calculation algorithm (Middle) and fast FE and P-value calculation algorithm (Bottom). Estimates are based on all traits analysed for 424 annotations at the 10^{-8} genome-wide significance thresholds.

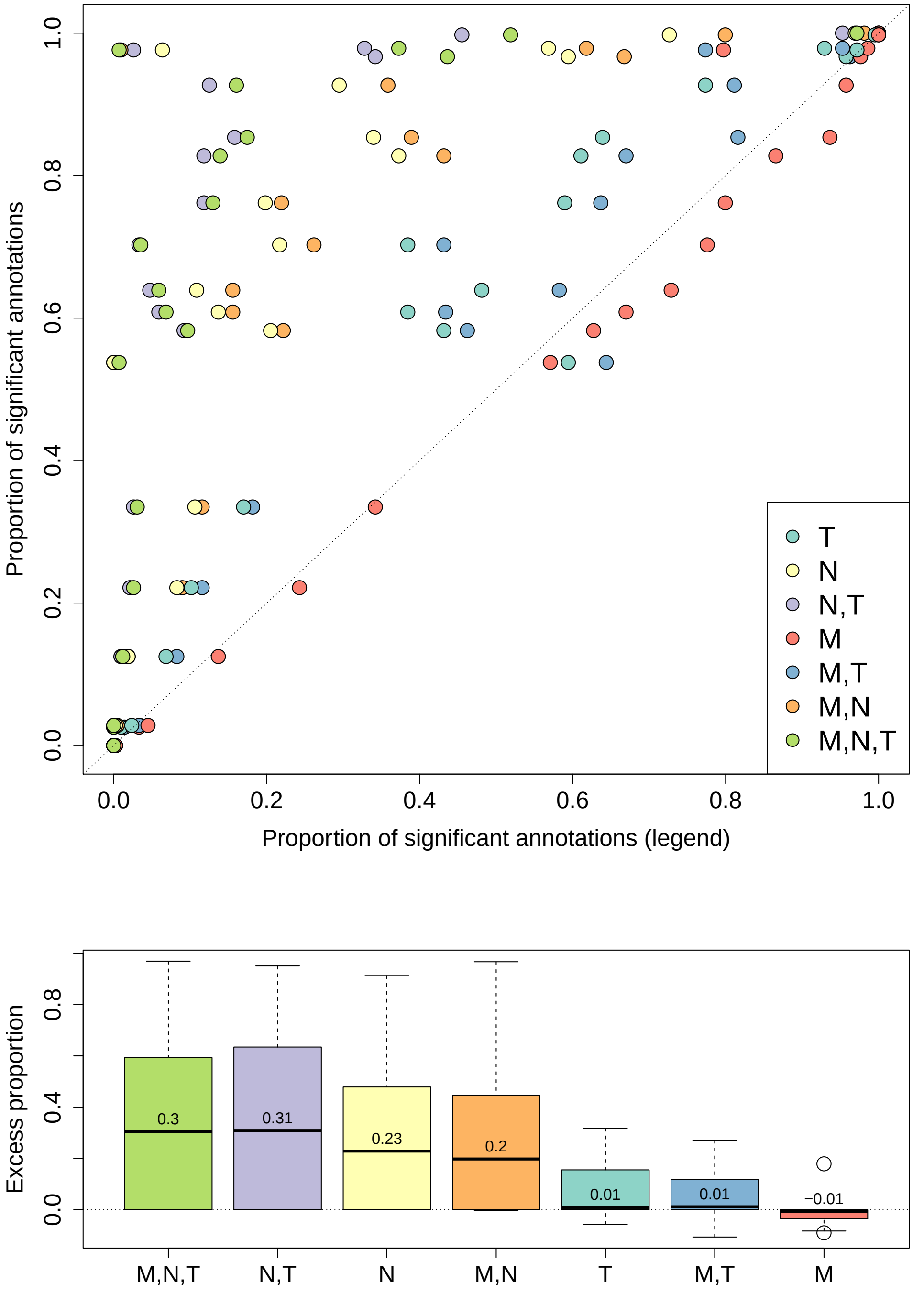
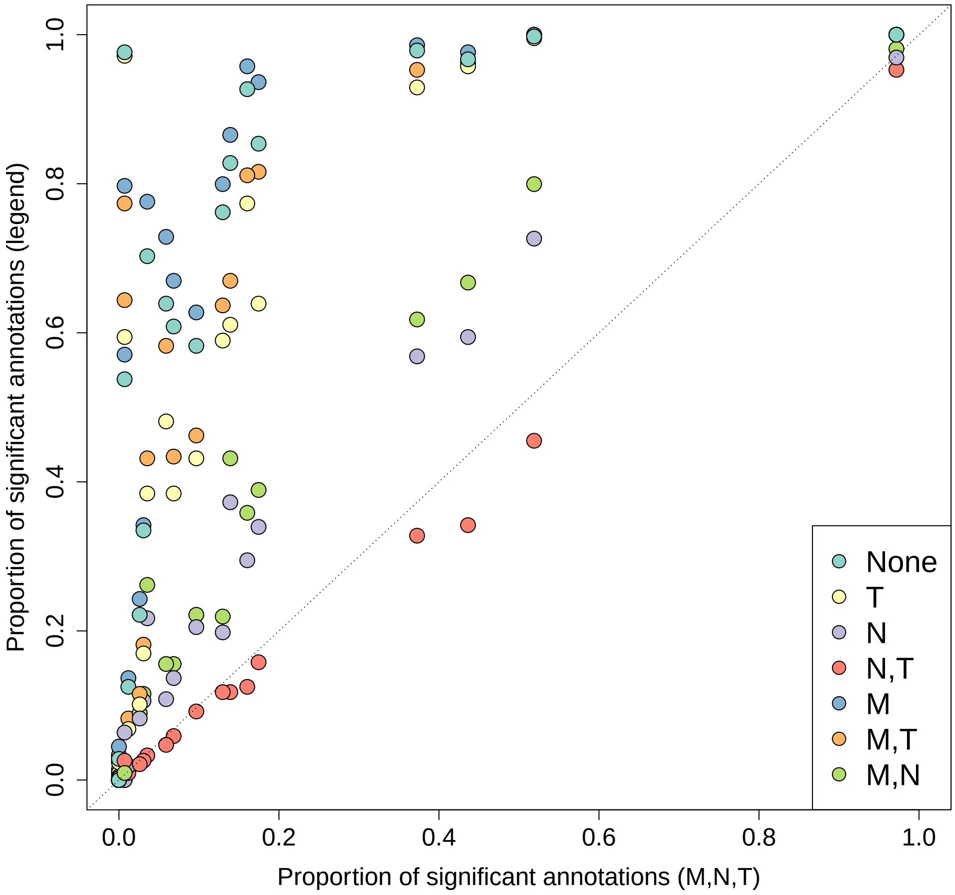


Figure S7: Effect of feature correction for each of the 27 GWAS studies at the 10^{-8} significance threshold. (Top) Proportion of significant annotations with respect to feature correction, where N denotes the number of LD proxies, M - MAF, and T - distance to the nearest TSS. Y-axis shows the corresponding values when no feature correction is employed. (Bottom) Excess proportion of significant annotations for no correction when compared to correcting for various features (x-axis). Numbers denote median excess values.



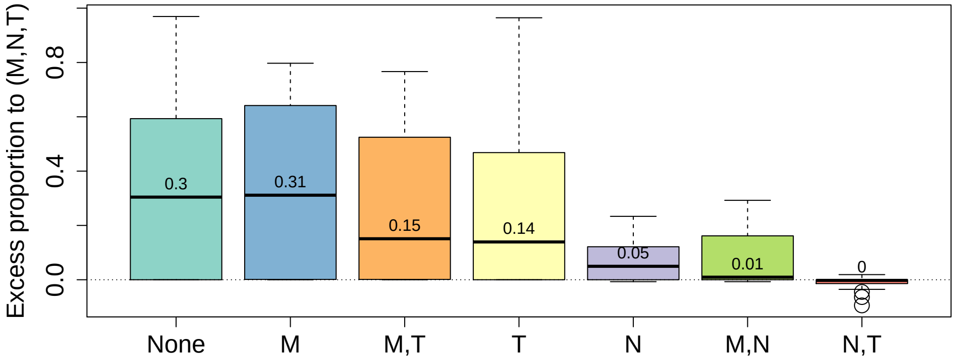


Figure S8: Effect of feature correction for each of the 27 GWAS studies at the 10^{-8} significance threshold. (Top) Proportin of significant annotations with respect to feature correction, where N denotes the number of LD proxies, M - MAF, and T - distance to the nearest TSS. Y-axis shows the corresponding values when correction for all three features. (Bottom) Excess proportion of significant annotations for correction on the x-axis when compared to correcting for all three features. Numbers denote median excess values.

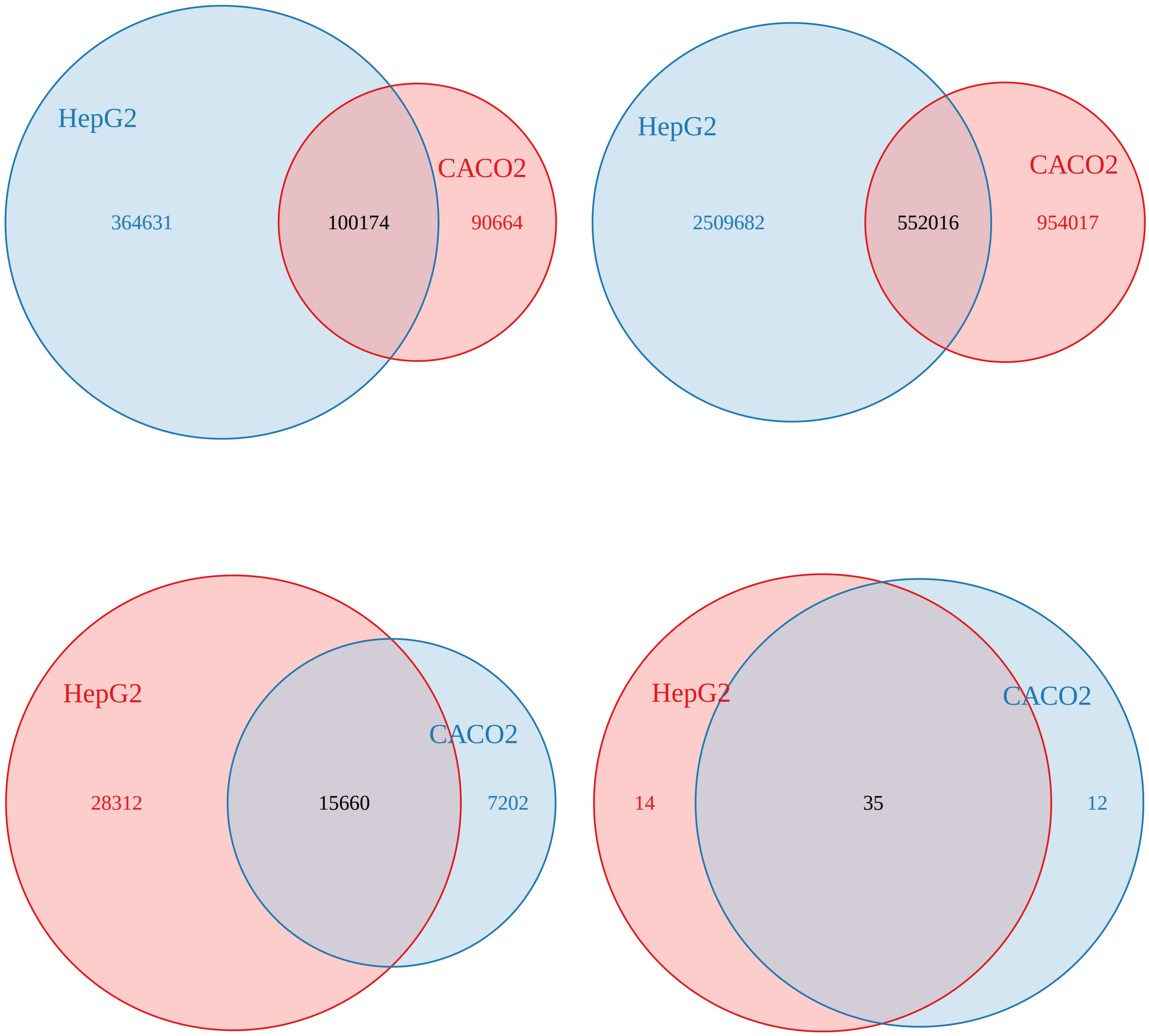


Figure S9: HepG2 and Caco-2 overlap. (Topleft) DNaseI peak overlap. (Topright) UK10K variant overlap. (Bottomleft) Pruned and proxy annotated LDL associated variants. (Bottomright) Pruned and proxy annotated LDL associated variants at the $T < 10^{-8}$ threshold.

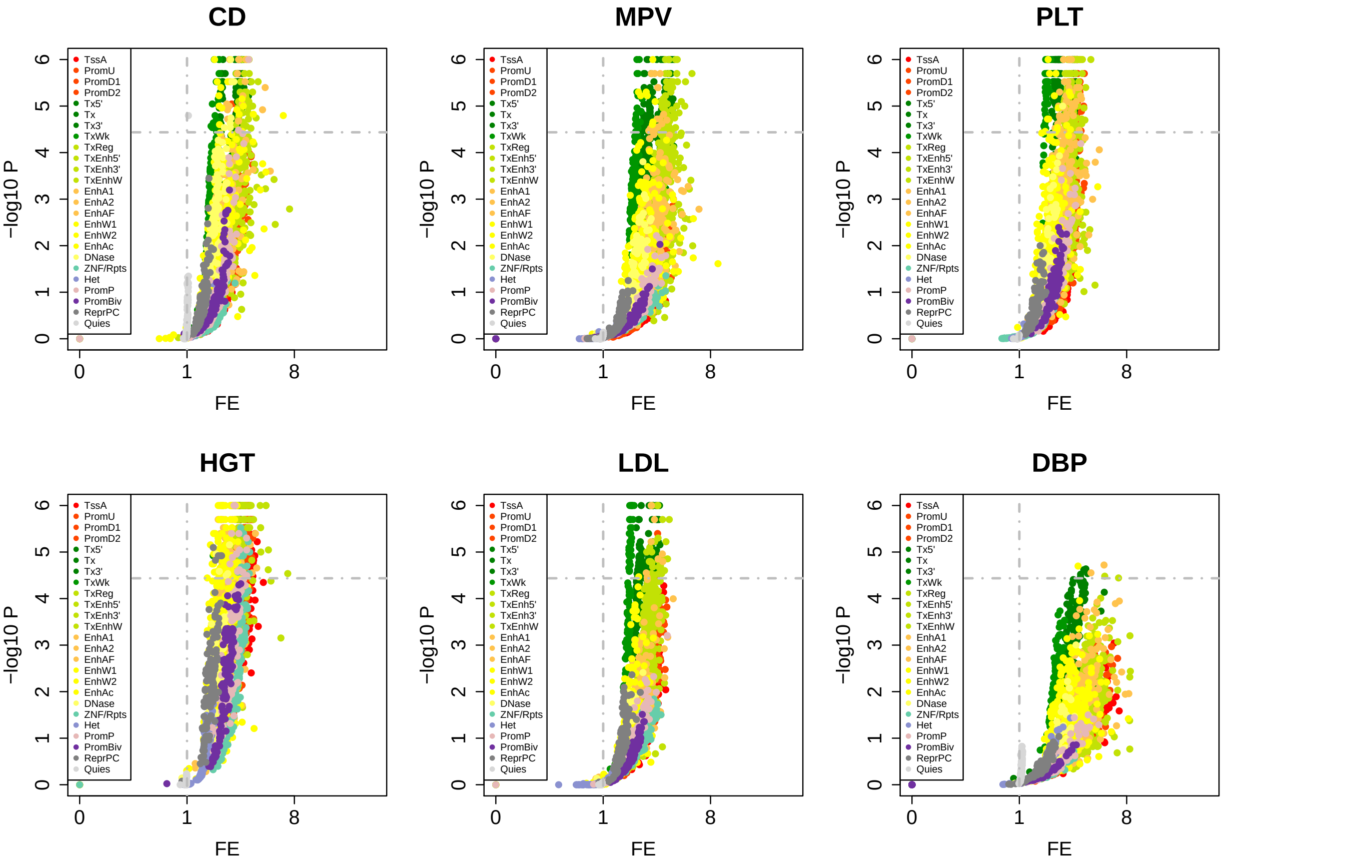


Figure S10: Fold enrichment and significance of GWAS associations in 25 segmentation states for 127 cell lines at the 10^{-5} GWAS significance threshold for CD, MPV, PLT, HGT, LDL and DBP.

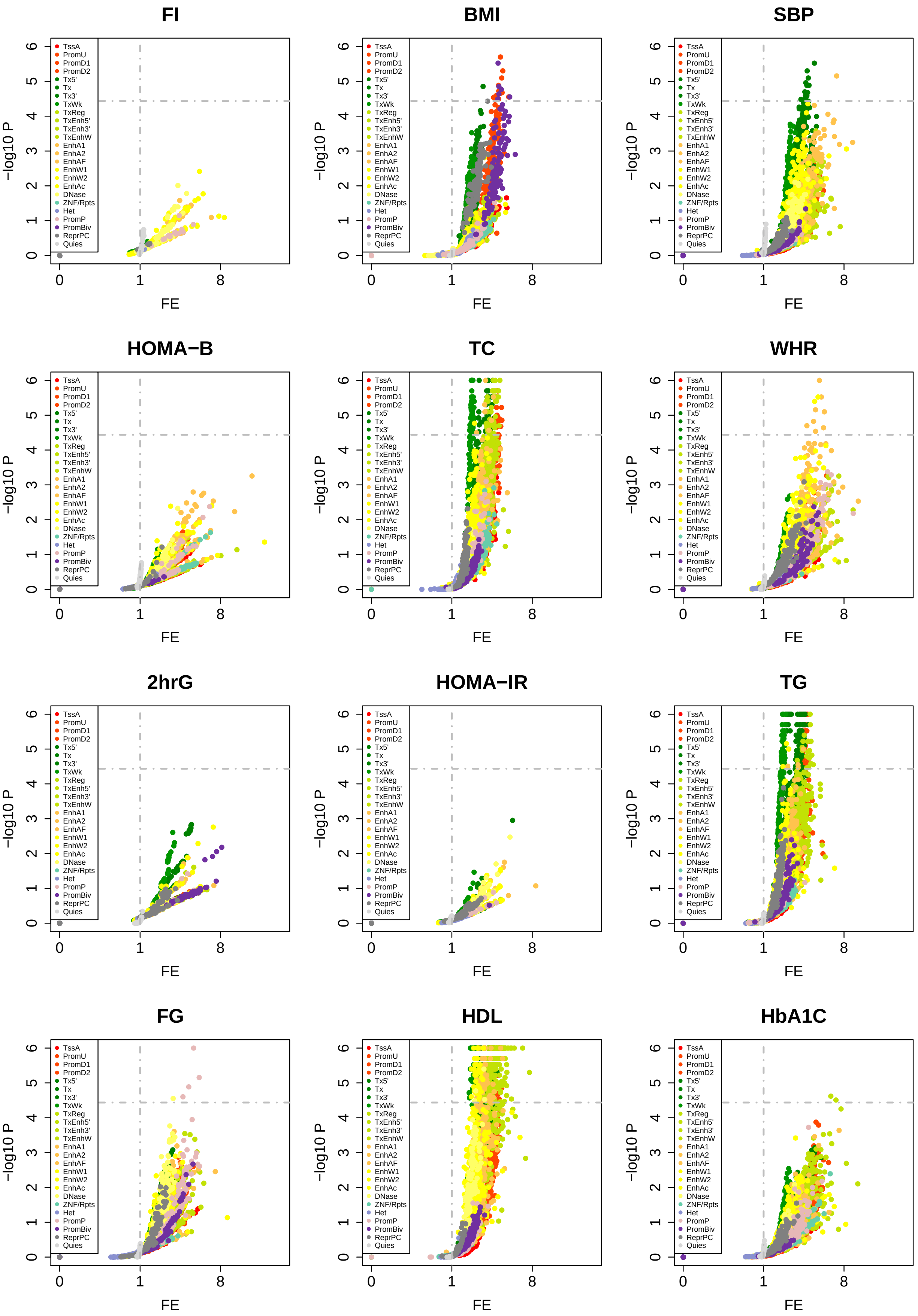


Figure S11: Fold enrichment and significance of GWAS associations in 25 segmentation states for 127 cell lines at the 10^{-5} GWAS significance threshold for FI, BMI, SBP, HOMA-B, TC, WHR, 2hrG, HOMA-IR, TG, FG, HDL and HbA1C.

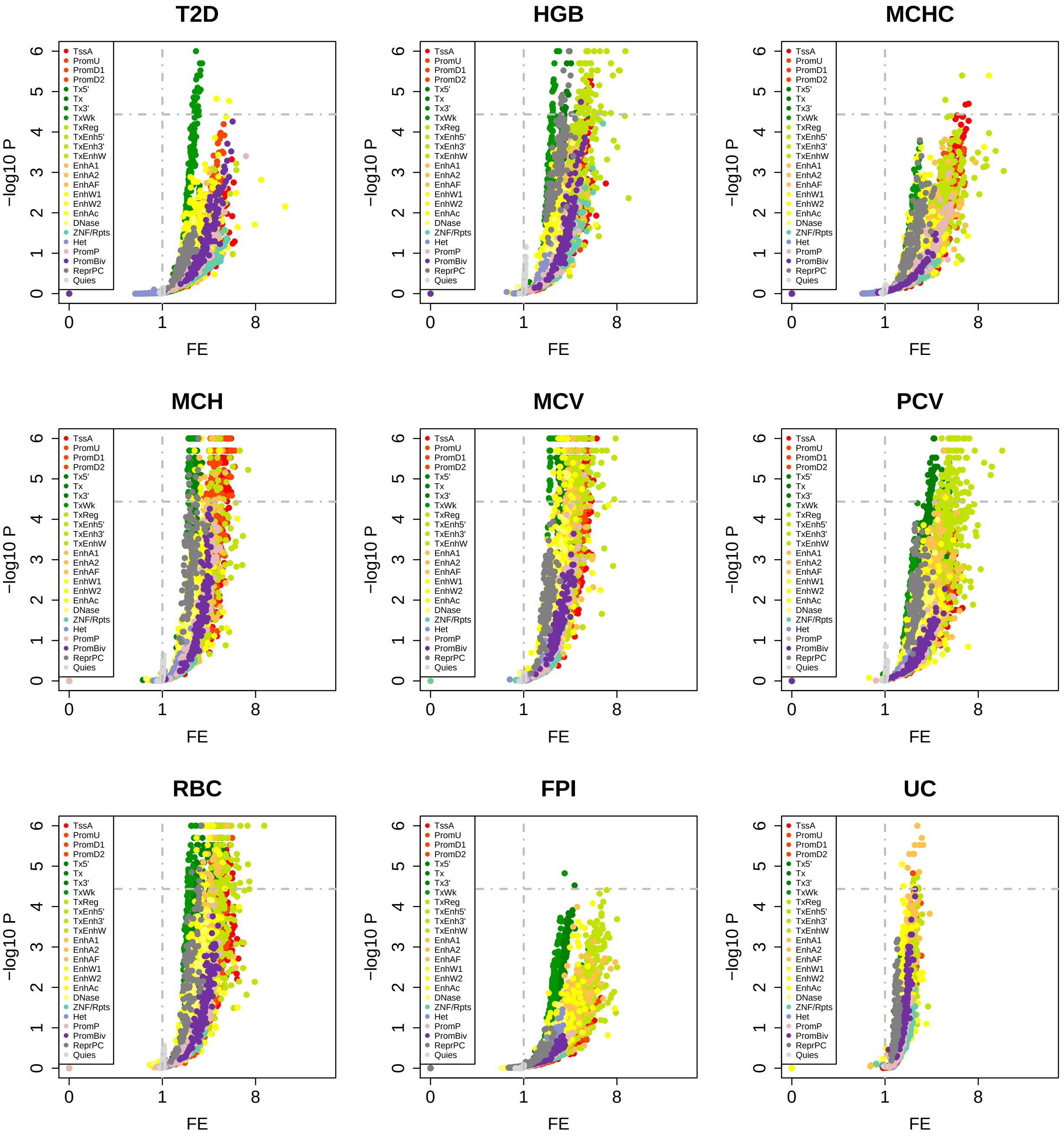


Figure S12: Fold enrichment and significance of GWAS associations in 25 segmentation states for 127 cell lines at the 10^{-5} GWAS significance threshold for T2D, HGB, MCHC, MCH, MCV, PCV, RBC, FPI and UC.

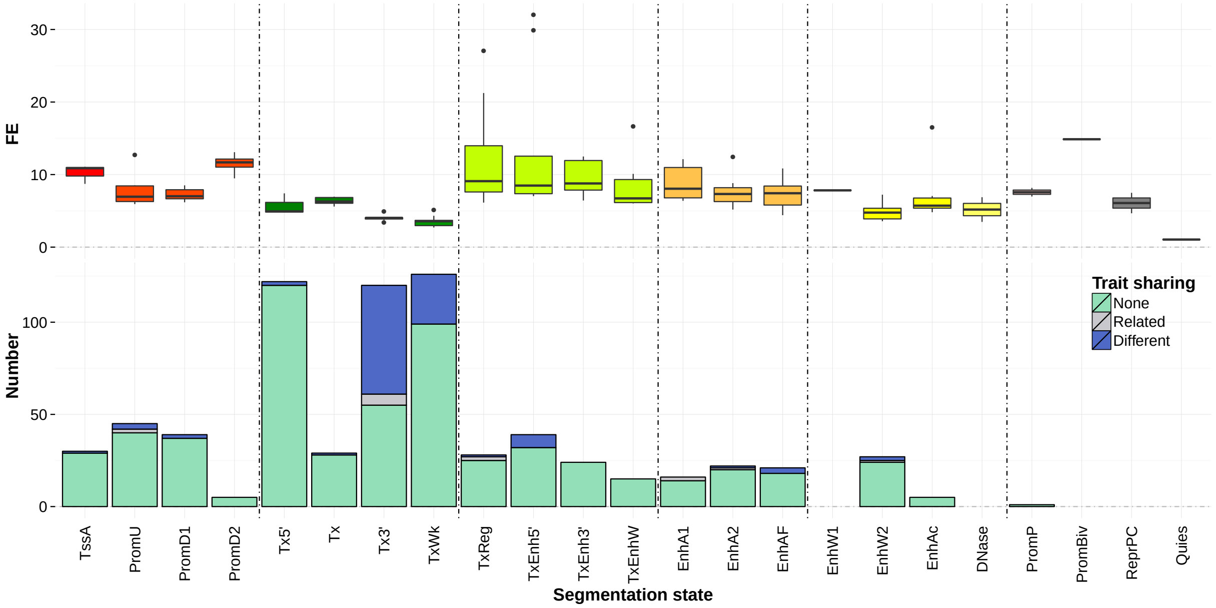


Figure S13: Fold enrichment levels and extent of sharing between traits for 25-state chromatin segmentations of the NIH Roadmap and ENCODE projects at the T$< 10^{-8}$ GWAS significance threshold. (Top) Distribution of significant FE values across the 27 traits considered, split by segmentation state and coloured to highlight predicted functional elements (e.g. bright red for active TSS, dark green for transcription, see Supplementary Table S7). (Bottom) Sharing of significantly enriched annotations with FE>2 across different phenotypes. The barplot displays the number of cell types where an annotation is uniquely enriched in a trait (light green), shared between closely related phenotypic traits (e.g. LDL cholesterol and TC, see Supplementary Table S3) (grey) or shared among non-correlated traits (e.g. TC and height) (blue).