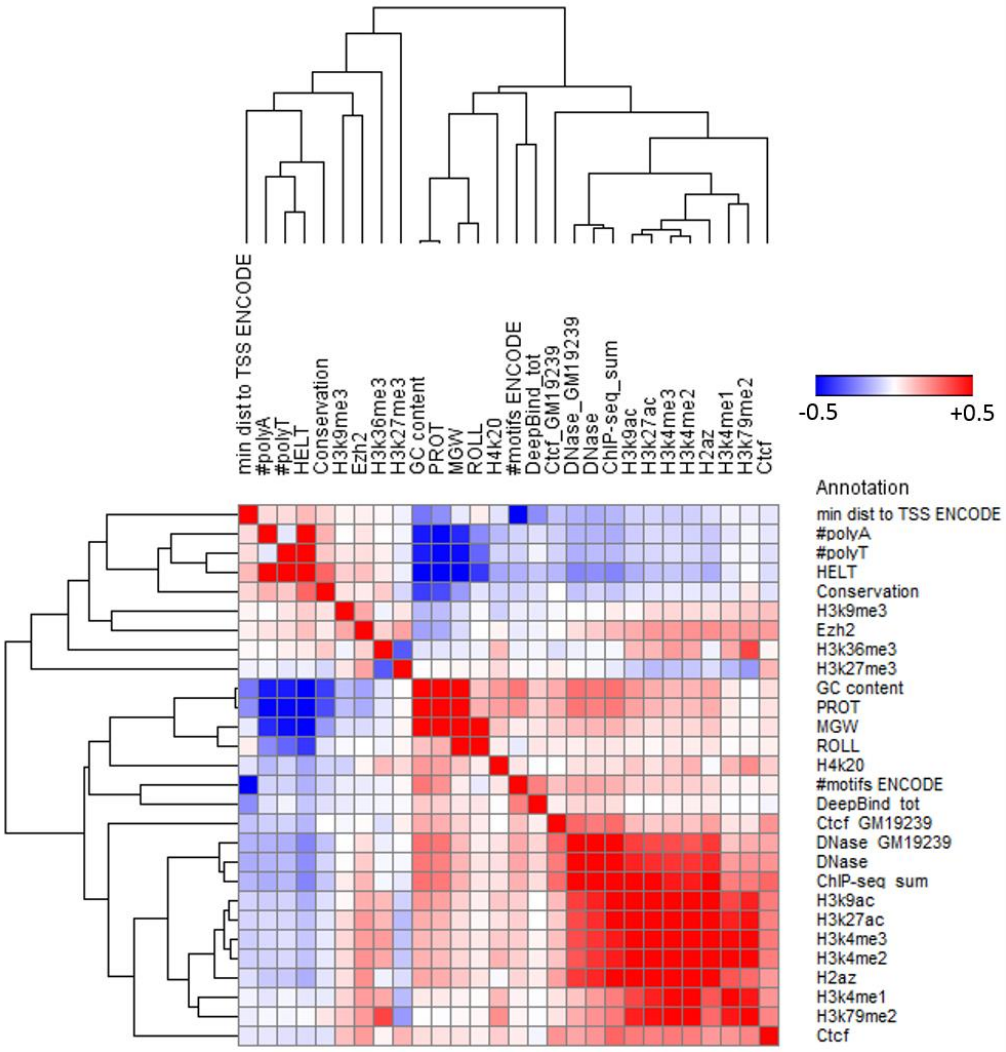
Supp. Table S1 (.xls file): CAGI regions annotations. **Training set:** Data provided for each sequence in the training set including: rsID, chromosome and position of the centered variant in hg19, sequences of the reference allele and alternate allele, normalized plasmid counts (ctrl.mean), RNA counts (exp.mean), log2 fold expression level (log2FC; RNA/plasmid), expression p-value (-log10P), BF corrected p-value (-log10_fdr) for the reference allele and the alternative allele from the combined LCL analysis (8 replicates) based on differential expression relative to the plasmid input by modeling a negative binomial with DESeq2 (Love, et al., 2014), log2 ratio (alternative/reference) of expression (LogSkew.Comb), allelic skew p-value (C.Skew.log10P), and allelic skew FDR (C.Skew.log10_fdr) , Regulatory_Hit (0 = no significant effect, 1 = regulatory effect) determined by applying a threshold of 0.01 for the Bonferroni corrected p-value, emVar_Hit (0 = no significant effect, 1 = allelic skew) based on a t-test on the log-transformed RNA-seq/plasmid ratios across replicates to test whether the reference and alternate allele had significantly different activity with a q-value threshold of 0.05, lead variant in the eQTL analysis, the eQTL associated gene, the eQTL beta, the eQTL t-statistic, and the eQTL p-value. For the two test sets the provided data includes: rsID, chromosome and position of the centered variant in hg19, sequences of the reference allele and alternate allele, lead variant in the eQTL analysis, the eQTL associated gene, the eQTL beta, the eQTL t-statistic, and the eQTL p-value. For part I (first test set): RefAllele.log2FC, AltAllele.log2FC and Regulatory_Hit are provided in the answer key. For part II (second test): LogSkew.Comb and emVar_hit are provided in the answer key. **Test set – part I and II:** For both training and test sets we calculated sequence-based predicted epigenetic properties of reference and alternative allele: number of ENCODE

motifs hits (Kheradpour and Kellis, 2014) calculated using fimo (Grant, et al., 2011), minimal distance to transcription start site (TSS) calculated by taking the distance from the end of the motif to the end of the 150bp sequence, maximal length of polyA/T sub-sequence representing nucleosome disfavoring sequences, GC content, DNA shape features including: minor groove width (MGW), roll, propeller twist, helix twist (Zhou, et al., 2013). We calculate experimentally measured and locus specific properties for each region, using the hg19 coordinates and ENCODE data for GM12878 and GN19239 cell types (Consortium, 2012): we mark 1 or 0 if a peak from the following assays intersects with the region: DNase-seq for GM12878 and GM19239, GM19239 Ctcf and the following marks for GM12878: H3k9me3, H3k36me3, H3k4me3, Ezh2, H3k27me3_StdPkV2, H3k9ac, H3k27ac, H3k4me1, H2az, H3k27me3_StdPk, H3k79me2, H3k4me2, H3k04me1, Ctcf, H4k20, H3k04me3. For ChIP-seq data we sum over all transcription factors files for GM12878. Evolutionary conservation score is calculated using PhastCons (Siepel, et al., 2005). **ENCODE all:** For each TF in GM12878 we report the overlap with all regions provided for the CAGI eQTL challenge. **ENCODE reference:** reference file for each TF used.

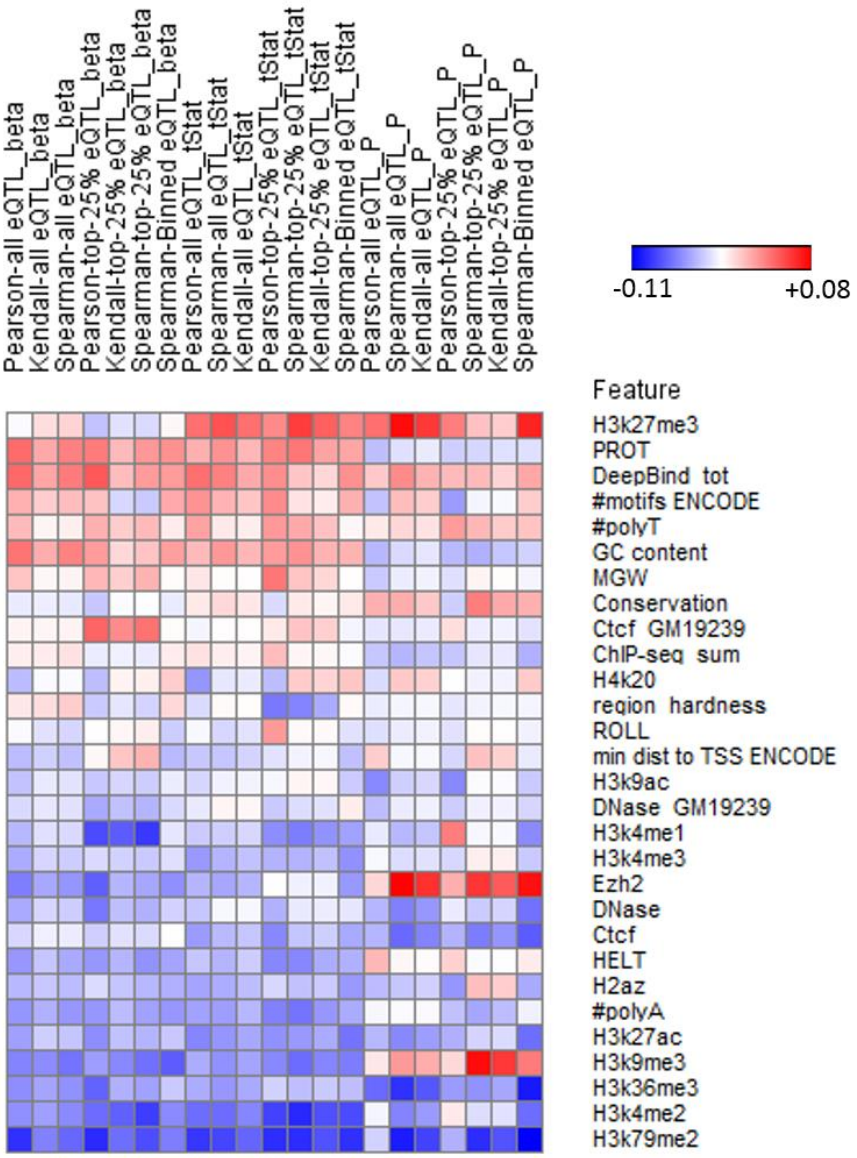
Supp. Table S2 (.xls file): Enrichment analysis of CAGI regions with epigenetic marks from ENCODE (Consortium, 2012). Enrichment results represented by overlapping bp/peaks with epigenetic annotations for each one of the sets: training, test part I and test part II and the corresponding p-value for the hypergeometric test.

Supp. Table S3 (.xls file): Bootstrap analysis of feature accuracy. This table corresponds to the rows and columns in the matrix represented in Figure 2a (Part I) and figure 2D (Part II). For each feature and statistical test we report the original performance, the mean performance, noise (mean/STD) and the standard deviation of resampling 100 times.

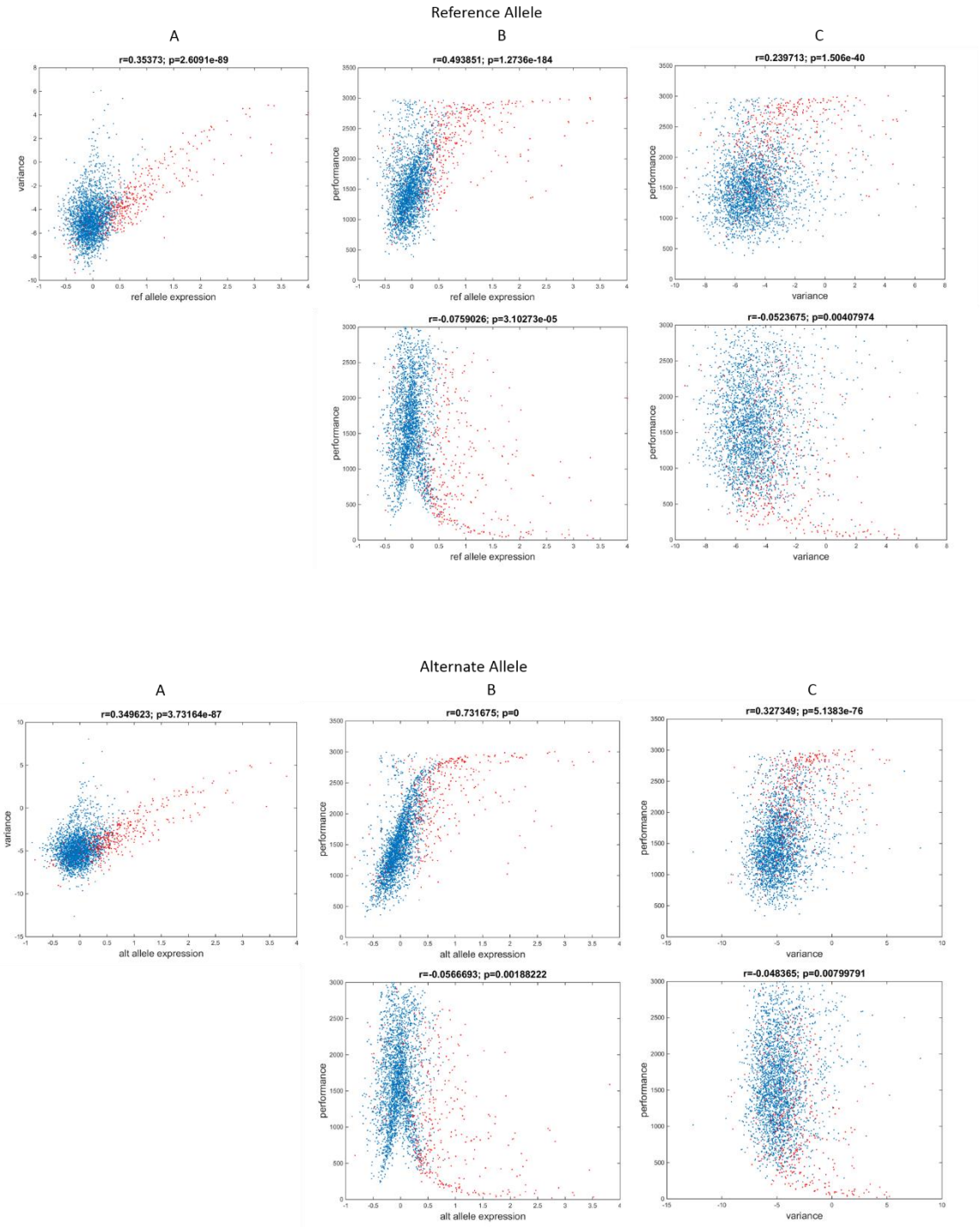
Supp. Note S1: Discussion of the properties that each statistical test highlights. For regression, Pearson correlation captures linear magnitude relationship between variables while Spearman correlation captures their ranking, Kendall correlation is similar to Spearman but is not effected by the distance between the ranks. Examining correlation for the top 25% expressing sequences or binning - attempts to reduce the noise in such correlations. For classification, AUROC represents the probability of a classifier to rank a randomly chosen positive instance higher than a negative one and is not dependent on the baseline probability that an instance is positive while the AUPRC is sensitive to this baseline probability. The Hypergeometric test represents the enrichment of the overlap between the set of true positives and the predicted positives.



Supp. Figure S1: correlation between features. The heat-map represents the Spearman correlation coefficient between every two features. The matrix is hierarchically clustered.



Supp. Figure S2: correlation between features and eQTL statistics (i.e. eQTL beta, eQTL t-statistics and e-QTL p-value). The heat-map represents the correlation coefficient across 7 statistical tests. The features are ranked from the most correlated to the least correlated.



Supp. Figure S3: Correlation (Spearman) between region's performance, variance and expression for reference and alternate allele. Red dots mark regulatory hit regions, blue dots mark all other regions. (A) Regions that are highly expressed are more variable

across replicates. When we define the region performance by the absolute difference between the predicted and observed values, we observe upper panel: a strong correlation between (B) expression and performance, (C) variance and performance. Using a more robust definition of performance decreases this bias - lower panel.

References:

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27(7):1017-8.

Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res 42(5):2976-87.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S and others. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15(8):1034-50.

Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, Rohs R. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res 41(Web Server issue):W56-62.