The American Journal of Human Genetics, Volume *107*

Supplemental Data

Promoter CpG Density Predicts Downstream Gene Loss-of-Function Intolerance

Leandros Boukas, Hans T. Bjornsson, and Kasper D. Hansen

Contents

1. **Supplemental Table 1**: Promoter coordinates for cases where our promoter filtering procedure selected a non-canonical promoter. The table contains the promoter coordinates and transcript ENSEMBL ids of both the canonical, as well as the alternative transcript that was selected. All coordinates refer to hg19.

2. **Supplemental Table 2**: Promoter coordinates for 11,059 transcripts where our filtering procedure selected a reliable promoter.

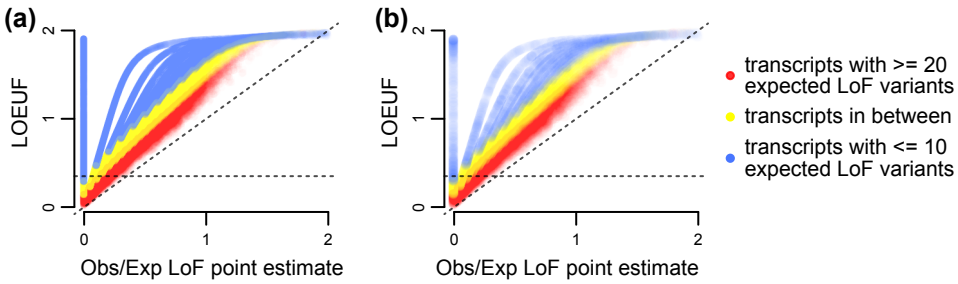
3. **Supplemental Table 3**: predLoF-CpG predictions for genes unascertained in gnomAD. Prediction probabilities are provided in the "prediction_probability_of_high_LoF_intolerance_by_predLoF-CpG" column. Probabilities > 0.75 correspond to genes predicted as highly LoF-intolerant, and probabilities < 0.25 to genes predicted as non-highly LoF-intolerant. ENSEMBL gene/transcript ids and coordinates of the promoters used for prediction are also provided; all coordinates refer to hg19.

4. **Supplemental Table 4**: Prediction probabilities for genes unascertained in gnomAD, which received a prediction probability between 0.25 and 0.75 by predLoF-CpG, and therefore remained unclassified. These prediction probabilities are provided in the "prediction_probability_of_high_LoF_intolerance_by_predLoF-CpG" column. We provide this table for completeness, but do not recommend using these probabilities for clinical decision making.

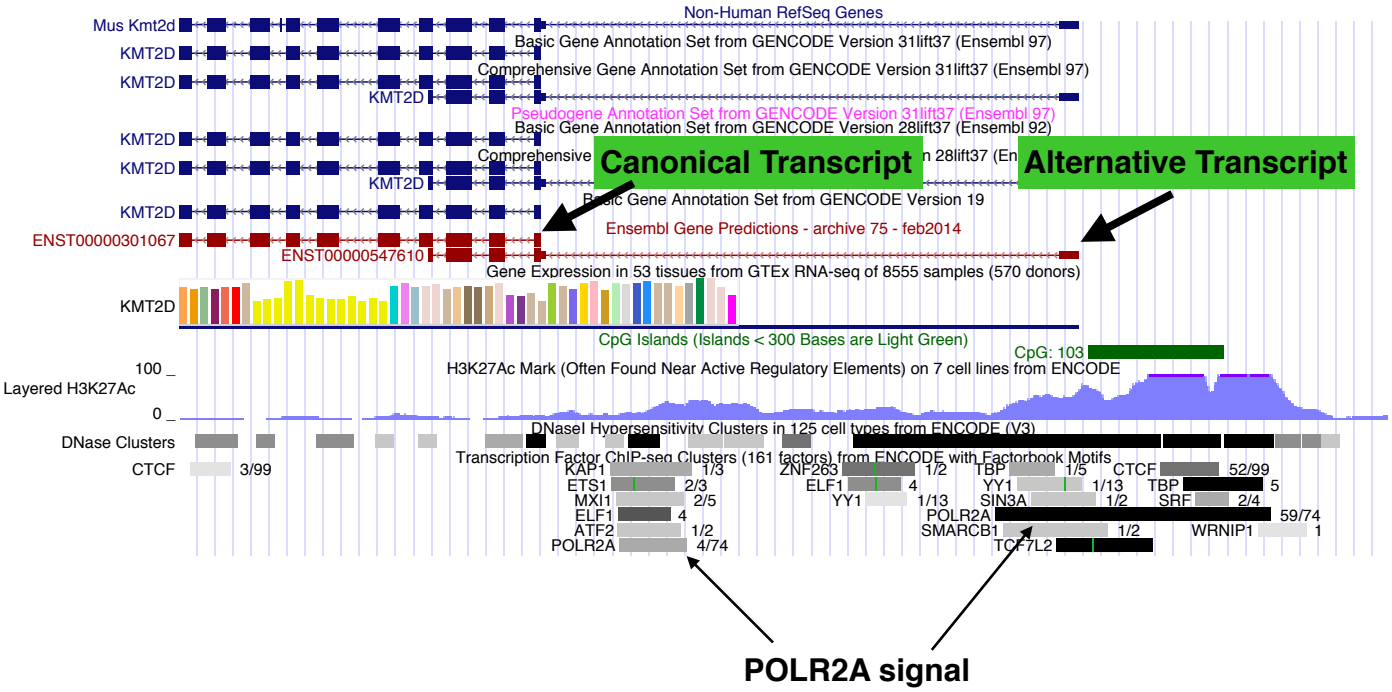
5. **Supplemental Table 5**: Similar to Supplemental Table 3, but for 101 genes with expected LoF variants between 10 and 20 that were classified as highly LoF-intolerant by predLoF-CpG but had LOEUF ≥ 0.35.

6. **Supplemental Figures S1-S15.**

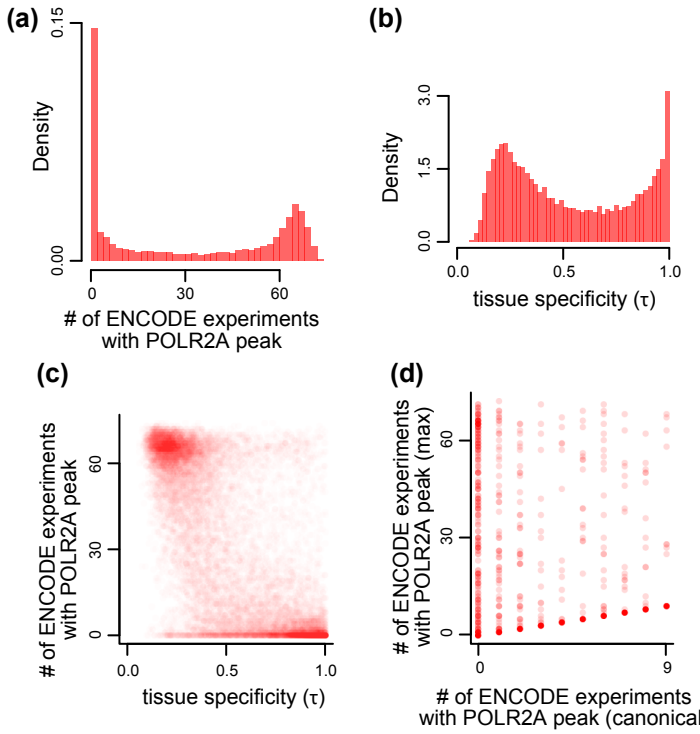
Supplemental Figures



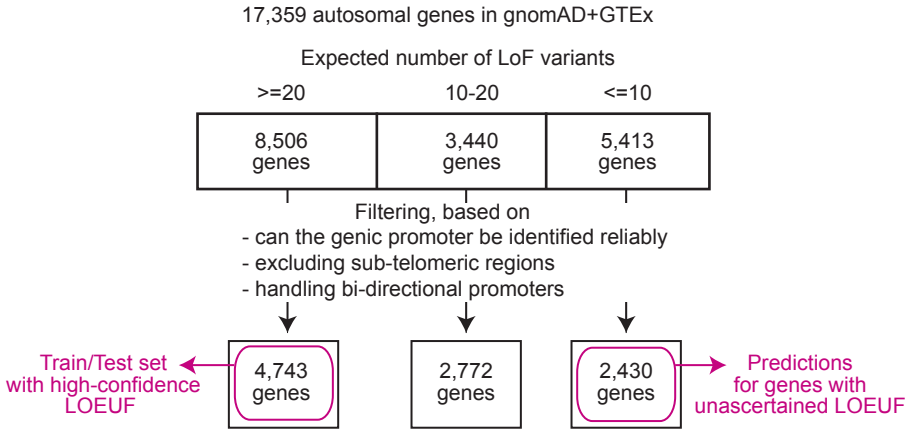
Supplemental Figure S1. Assessing the reliability of LOEUF estimates. Scatterplots of the point estimates of the observed/expected proportion of loss-of-function variants (x axis), against LOEUF (y axis; defined as the upper bound of the 90% confidence interval around the point estimate). Each point corresponds to a transcript. The horizontal line corresponds to the 0.35 cutoff for highly LoF-intolerant genes. Shown for: **(a)** all transcripts, and **(b)** canonical transcripts only (based on GENCODE annotation).



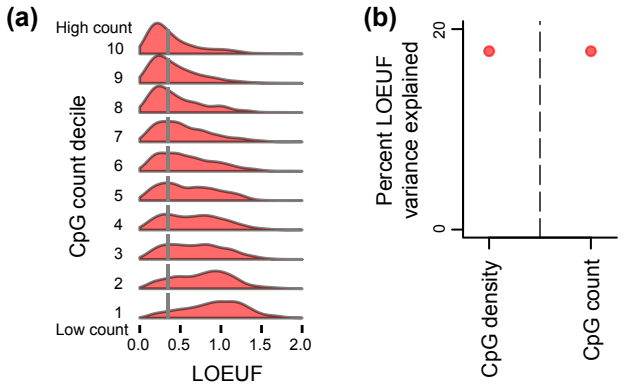
Supplemental Figure S2. UCSC genome browser screenshot of a 10kb region containing the transcriptional start sites of the canonical and one alternative *KMT2D* transcript. The precise coordinates are chr12:49,446,107-49,456,107. The sequence of the canonical transcript extends beyond the 10kb region shown.



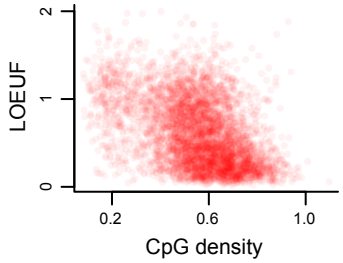
Supplemental Figure S3. Assessing the relationship between tissue specificity of gene expression and POLR2A binding at the canonical promoter. (a) The distribution of the number of ENCODE ChIP-seq experiments showing POLR2A peaks, for all canonical promoters (4 kb regions centered around the TSS) in Ensembl (hg19 assembly). **(b)** The distribution of τ computed using gene-level expression quantifications from GTEx. **(c)** Scatterplot of τ against the number of ENCODE ChIP-seq experiments showing POLR2A peaks at the canonical promoter. Each point corresponds to a gene-promoter pair. **(d)** Scatterplot of the number of ENCODE ChIP-seq experiments showing POLR2A peaks at the canonical (x axis) promoter versus the corresponding number at the promoter with the greatest number of detected peaks (out of all the alternative promoters of a gene; y axis). Each point corresponds to a promoter pair for a single gene; shown are only genes that are broadly expressed ($\tau < 0.6$) but whose canonical promoter shows POLR2A binding in less than 10 ENCODE experiments.



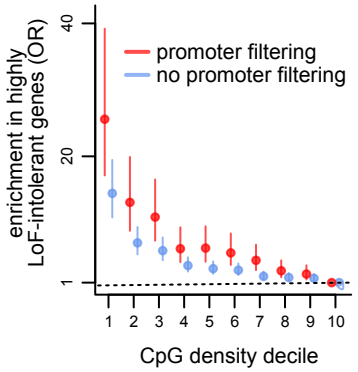
Supplemental Figure S4. Partitioning genes according to the reliability of their LOEUF estimates and promoter annotation. Schematic illustrating our approach (see Methods for details). We start with 17,359 genes that: a) are present in both GTEx and gnomAD, b) reside in autosomes, and c) their promoters are not subtelomeric. We then filter these according to whether they have reliable promoter annotations, and in cases of pairs of genes with overlapping promoters we only keep one pair. This gives us the set of high-confidence genes that we use to establish the relationship between CpG density and LOEUF and to train predLoF-CpG, and the set of unascertained genes to which we apply predLoF-CpG.



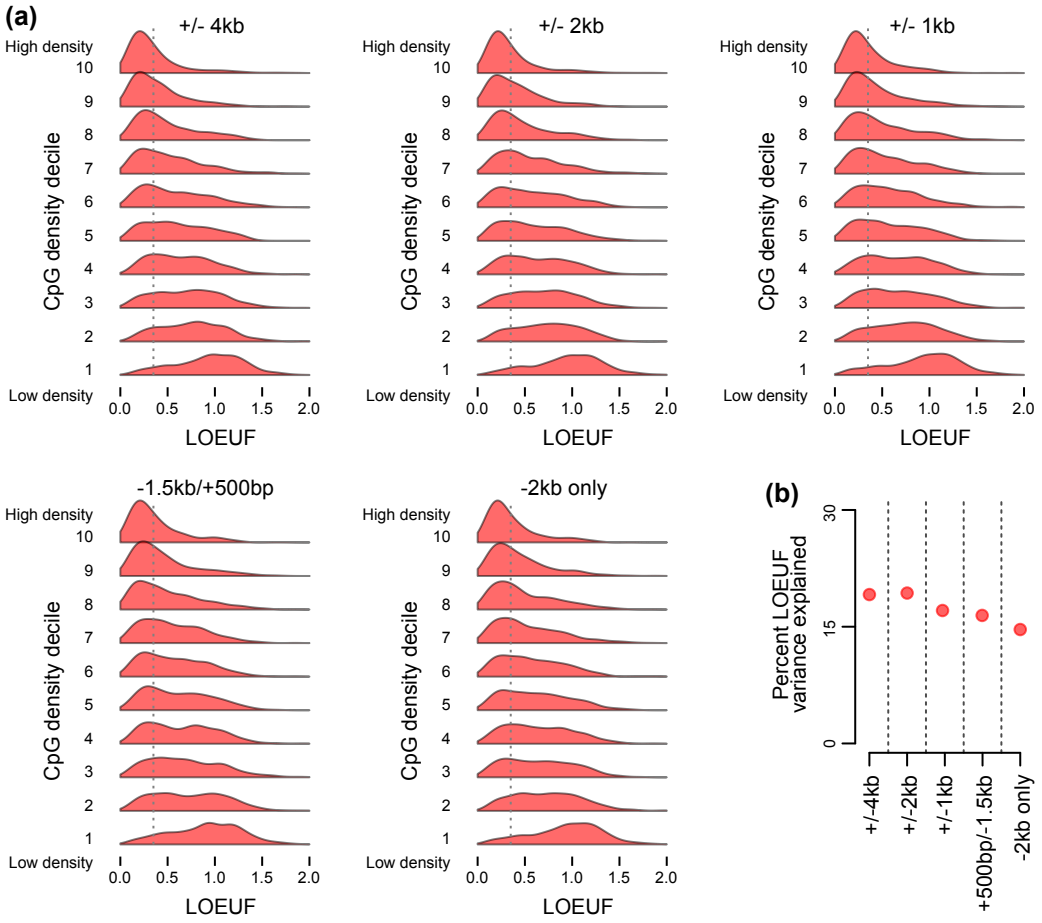
Supplemental Figure S5. The relationship between CpG count and downstream gene LOEUF. (a). Like Figure 1A, but with the CpG count of a promoter instead of CpG density. **(b)**. The percentage of LOEUF variance (adjusted r^2) that is explained by either promoter CpG density or CpG count.



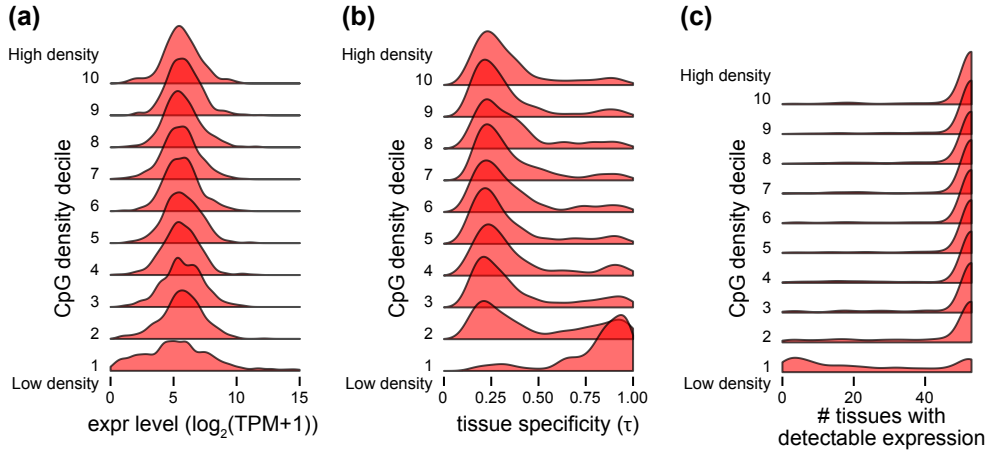
Supplemental Figure S6. Scatterplot of promoter CpG density (o/e CpG ratio) against downstream gene LOEUF. Each point corresponds to a promoter-gene pair.



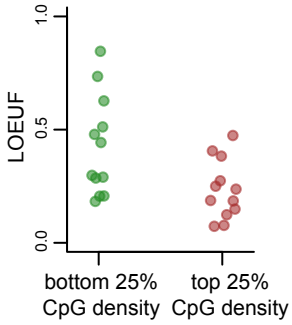
Supplemental Figure S7. The effect of filtering for high-confidence promoter annotations on the relationship between CpG density (o/e CpG ratio) and LOEUF. Like Figure 1b, but shown both for the 4,859 genes with high-confidence promoter annotations (red), and for 6,656 genes with canonical (based on GENCODE) promoter annotations and at least 20 expected LoF variants, without further promoter filtering (blue).



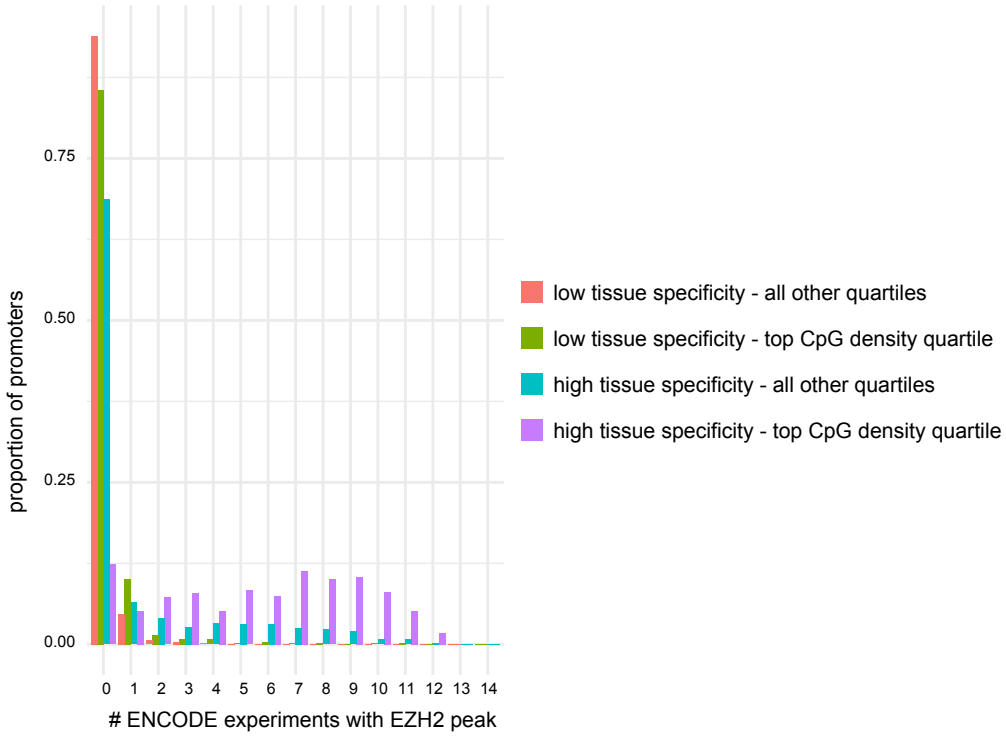
Supplemental Figure S8. The impact of promoter definition on the relationship between CpG density (o/e CpG ratio) and LOEUF. (a) Like Figure 1A, but with different choices of the interval around the transcription start site that is defined as the promoter. The "-" sign refers to upstream of the TSS in the 5′ direction (that is, taking gene strandedness into account). **(b)** The percentage of LOEUF variance (adjusted r^2) that is explained by promoter CpG density, for each of the promoter definitions in (a).



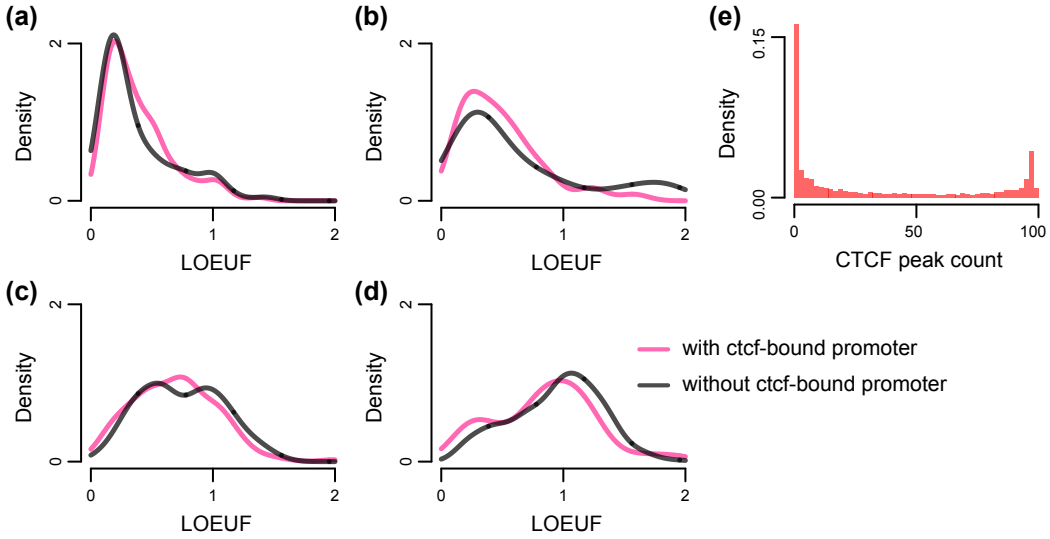
Supplemental Figure S9. Distributions of downstream gene expression level and tissue specificity across promoter CpG density (o/e CpG ratio) deciles. (a) The distribution of expression level within CpG density deciles. **(b)** The distribution of tissue specificity (τ) within CpG density deciles. **(c)** The distribution of the number of tissues with detectable expression (defined as median TPM > 0.3) within CpG density deciles. Both expression level and τ were computed from the GTEx dataset (see Methods). In all three figures, CpG density deciles are labeled 1-10, with 1 the most CpG-poor decile and 10 the most CpG-rich.



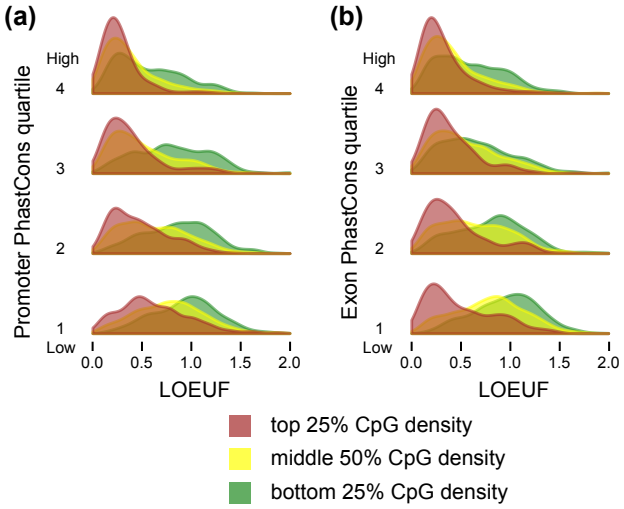
Supplemental Figure S10. The relationship between promoter CpG density (o/e CpG ratio) and loss-of-function intolerance of key human developmental regulators. Each point corresponds to a gene. 46 key human developmental regulators were obtained from the supplemental material of Akalin et al. [1] (see Methods). The 25th and 75th CpG density percentiles were computed from the empirical CpG density distribution of these genes and were equal to 0.58 and 0.8, respectively.



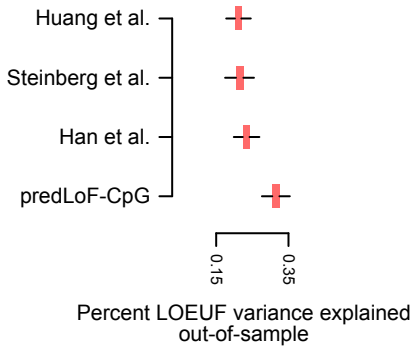
Supplemental Figure S11. The proportion of promoters with EZH2 peaks in 1-14 ENCODE experiments, stratified based on their CpG density (o/e CpG ratio) and downstream gene tissue specificity. Tissue specificity was quantified from the GTEx dataset using τ (Methods). Low tissue specificity corresponds to $\tau < 0.6$ and high tissue specificity corresponds to $\tau > 0.6$.



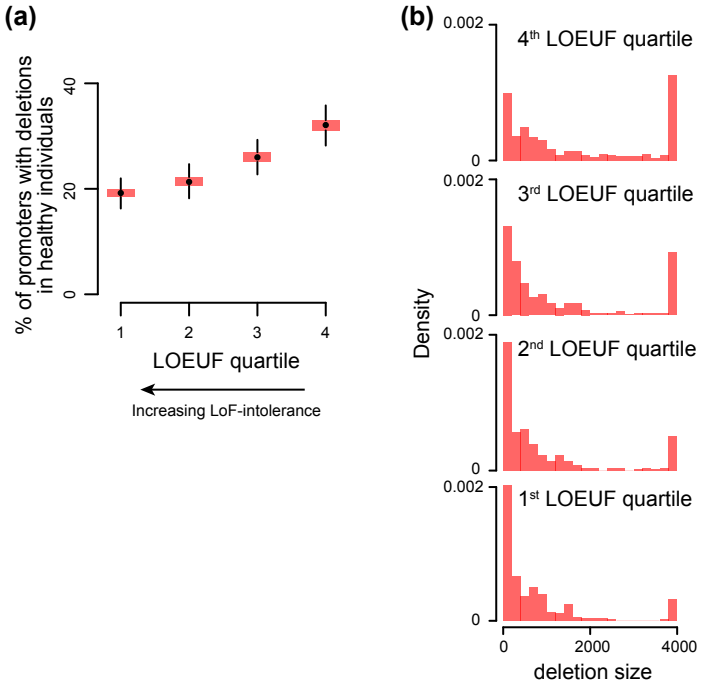
Supplemental Figure S12. Loss-of-function intolerance of CTCF-bound versus CTCF-unbound genes. The LOEUF distributions of well-ascertained genes, stratified according to whether their promoters show CTCF peaks in at least 70 ENCODE experiments (CTCF-bound), or in no experiments (CTCF-unbound; see also panel **(e)**). **(a)** Broadly expressed ($\tau < 0.6$) genes with high-CpG-density (top 25%) promoters. **(b)** Tissue specific ($\tau > 0.6$) genes with high-CpG-density (top 25%) promoters. **(c)** Broadly expressed ($\tau < 0.6$) genes with low-CpG-density (bottom 25%) promoters. **(d)** Tissue specific ($\tau > 0.6$) genes with low-CpG-density (bottom 25%) promoters. **(e)** The distribution of the number of ENCODE ChIP-seq experiments showing CTCF peaks, for well-ascertained gene promoters.



Supplemental Figure S13. The relationship between promoter CpG density (o/e CpG ratio) and loss-of-function intolerance conditional on promoter and exonic across-species conservation. (a) The distribution of LOEUF, stratified by promoter CpG density, in each quartile of promoter PhastCons score (Methods). **(b)** The distribution of LOEUF, stratified by promoter CpG density, in each quartile of exonic PhastCons (Methods). For both **(a)** and **(b)** quartiles are labeled from 1-4, with 1 being the least and 4 the most conserved, respectively.



Supplemental Figure S14. The percentage of out-of-sample LOEUF variance explained by the different predictors of LoF-intolerance. Each boxplots corresponds to a LoF-intolerance predictor as shown on the x-axis, and shows the sampling distribution of the adjusted r^2 after regressing the LOEUF of genes in the test set on the corresponding predictor. We performed 1,000 random train/test splits. For predLoF-CpG, the regression was performed on the prediction probably of high LoF-intolerance.



Supplemental Figure S15. The relationship between promoter deletions seen in healthy individuals and downstream gene loss-of-function intolerance. (a) The proportion of promoters harboring deletions across different strata of downstream gene loss-of-function intolerance. For each stratum, the distribution is obtained via the bootstrap. **(b)** The distribution of the size of deletions harbored by promoters across different strata of downstream gene loss-of-function intolerance.