Promoter sequence and architecture determine expression variability and confer robustness to genetic variants

Hjörleifur Einarsson[1], Marco Salvatore[1], Christian Vaagensø[1], Nicolas Alcaraz[1,2], Jette Bornholdt Lange[1,3], Sarah Rennie[1], Robin Andersson[1,*]

[1]Department of Biology, University of Copenhagen, 2200, Copenhagen, Denmark
[2]Current address: Novo Nordisk Foundation Center for Protein Research (CPR), University of Copenhagen, 2200, Copenhagen, Denmark
[3]Current address: Adcendo ApS, 2200, Copenhagen, Denmark
*To whom correspondence should be addressed: robin@binf.ku.dk

Abstract

Genetic and environmental exposures cause variability in gene expression. Although most genes are affected in a population, their effect sizes vary greatly, indicating the existence of regulatory mechanisms that could amplify or attenuate expression variability. Here, we investigate the relationship between the sequence and transcription start site architectures of promoters and their expression variability across human individuals. We find that expression variability is largely determined by a promoter's DNA sequence and its binding sites for specific transcription factors. We further demonstrate that flexible usage of transcription start sites within a promoter attenuates variability, providing transcriptional and mutational robustness.

Introduction

Transcriptional regulation is the main process controlling how genome-encoded information is translated into phenotypes. Hence, understanding how transcriptional regulation influences gene expression variability is of fundamental importance to understand how organisms are capable of generating proper phenotypes in the face of stochastic, environmental, and genetic variation. Through differentiation, cells acquire highly specialized functions, but need to still maintain their general abilities to accurately regulate both essential pathways as well as responses to changes in the environment. To achieve robustness, regulatory processes must be capable of attenuating expression variability of essential genes (Bartha et al. 2018), while still allowing, or possibly amplifying (Urban and Johnston 2018; Eldar and Elowitz 2010), variability in expression for genes that are required for differentiation or responses to environmental changes and external cues. How cells can achieve such precision and robustness remains elusive.

Genetic variation affects the expression level (Stranger et al. 2007; Pickrell et al. 2010; Montgomery et al. 2010) of the majority of human genes (GTEx Consortium 2017; Storey et al. 2007; Lappalainen et al. 2013a). However, genes are associated with highly different effect sizes, with ubiquitously expressed or essential genes frequently being less affected (GTEx Consortium 2017). This indicates that genes associated with different regulatory programs are connected with different mechanisms or effects of mutational robustness (Payne and Wagner 2015). Multiple transcription factor (TF) binding sites may act to buffer the effects of mutations in promoters (Spivakov et al. 2012), and promoters can have highly flexible transcription start site (TSS) architectures (Carninci et al. 2006; Akalin et al. 2009; Lehner 2008). This indicates that the sequence and architecture of a promoter may influence its variability in expression across individuals.

Previous studies aimed at identifying processes involved in the regulation of gene expression variability have indeed revealed regulatory features mostly associated with the promoters of genes, such as CpG islands and TATA-boxes (Ravarani et al. 2016;

Sigalova et al. 2020; Morgan and Marioni 2018), the chromatin state around gene TSSs (Faure et al. 2017), and the propensity of RNA polymerase II to pause downstream of the TSS (Boettiger and Levine 2009). As of yet these studies have relied on model organisms or single cell sequencing of cell lines, and regulatory features have not been thoroughly studied from the perspective of variability in promoter activity or across human individuals. Furthermore, it is unclear if regulation of variability mainly acts to attenuate variability to achieve stable expression for certain genes or if independent regulatory processes act in parallel to amplify variability for other genes.

Here, we provide a comprehensive characterization of the sequences, TSS architectures, and regulatory processes determining variability of promoter activity across human lymphoblastoid cell lines (LCLs). We find that variability in promoter activity is to a large degree encoded within the promoter sequence. Furthermore, the presence of binding sites for specific transcription factors, including those of the ETS family, are highly predictive of low promoter variability independently of their impact on promoter expression levels. In addition, we demonstrate that differences in the variability of promoters reflect their involvement in distinct biological processes, indicating a selective tradeoff between stability and plasticity of promoters. Finally, we show that flexibility in TSS usage can attenuate promoter variability and identify switches between proximal TSSs due to genetic variants as a novel mechanism that confers mutational robustness to gene promoters. Our study provides fundamental insights into transcriptional regulation, revealing shared mechanisms that can buffer stochastic, environmental, and genetic variation and how these affect the responsiveness and cell-type restricted activity of genes.

Results

TSS profiling reveals variability in promoter activity across individuals

To characterize human variability in promoter activities, we profiled TSSs using CAGE (Cap Analysis of Gene Expression (Takahashi et al. 2012); Fig. 1A) across 108 Epstein-Barr virus transformed LCLs (Auton et al. 2015) of African origin (89 from

Yoruba in Ibadan, Nigeria (YRI) and 19 from Luhya in Webuye, Kenya (LWK)). The samples had a balanced sex ratio (56 females and 52 males) and no observable population stratification in the expression data (Supplementary Fig. S1). With CAGE, TSSs can be mapped with single base pair resolution and the relative number of sequencing reads supporting each TSS gives simultaneously an accurate estimate of the abundance of its associated RNA (Kawaji et al. 2014). The LCL panel CAGE data therefore give us a unique opportunity to both estimate variability in promoter activity and characterize the regulatory features influencing such variability.

We identified 29,001 active promoters of 15,994 annotated genes (Frankish et al. 2019) through positional clustering of proximal CAGE-inferred TSSs on the same strand (Fig. 1A) (Carninci et al. 2006) detected in at least 10% of individuals (Supplementary Table 1). This individual-agnostic strategy ensured a focus on promoters that are consistently active across multiple individuals while also allowing for the measurement of variability in promoter activity across the panel. For example, the CAGE data revealed that the promoters of gene *RPL26L1*, encoding a putative component of the large 60S subunit of the ribosome, and transcription factor gene *SIX3* have highly different variance yet similar mean expression across individuals (Fig. 1B-D).

We used the squared coefficient of variation (CV^2) as a measure of promoter expression dispersion, revealing how the normalized expression across individuals deviates from the mean for each identified promoter. We observed that the promoter CV^2 decreases by increasing mean expression (Fig. 1E) (Eling et al. 2018; Kolodziejczyk et al. 2015; Sigalova et al. 2020). To account for this bias, we subtracted the expected dispersion for each promoter according to its expression level (Kolodziejczyk et al. 2015; Newman et al. 2006). Importantly, rank differences in promoter dispersion were maintained for each expression level after adjustment, as seen for promoters of genes *RPL26L1* and *SIX3* (Fig. 1E,F). This strategy thus allowed us to investigate how promoter architecture and sequence determine variability in promoter activity across the panel separately from its impact on expression level (Fig. 1F).

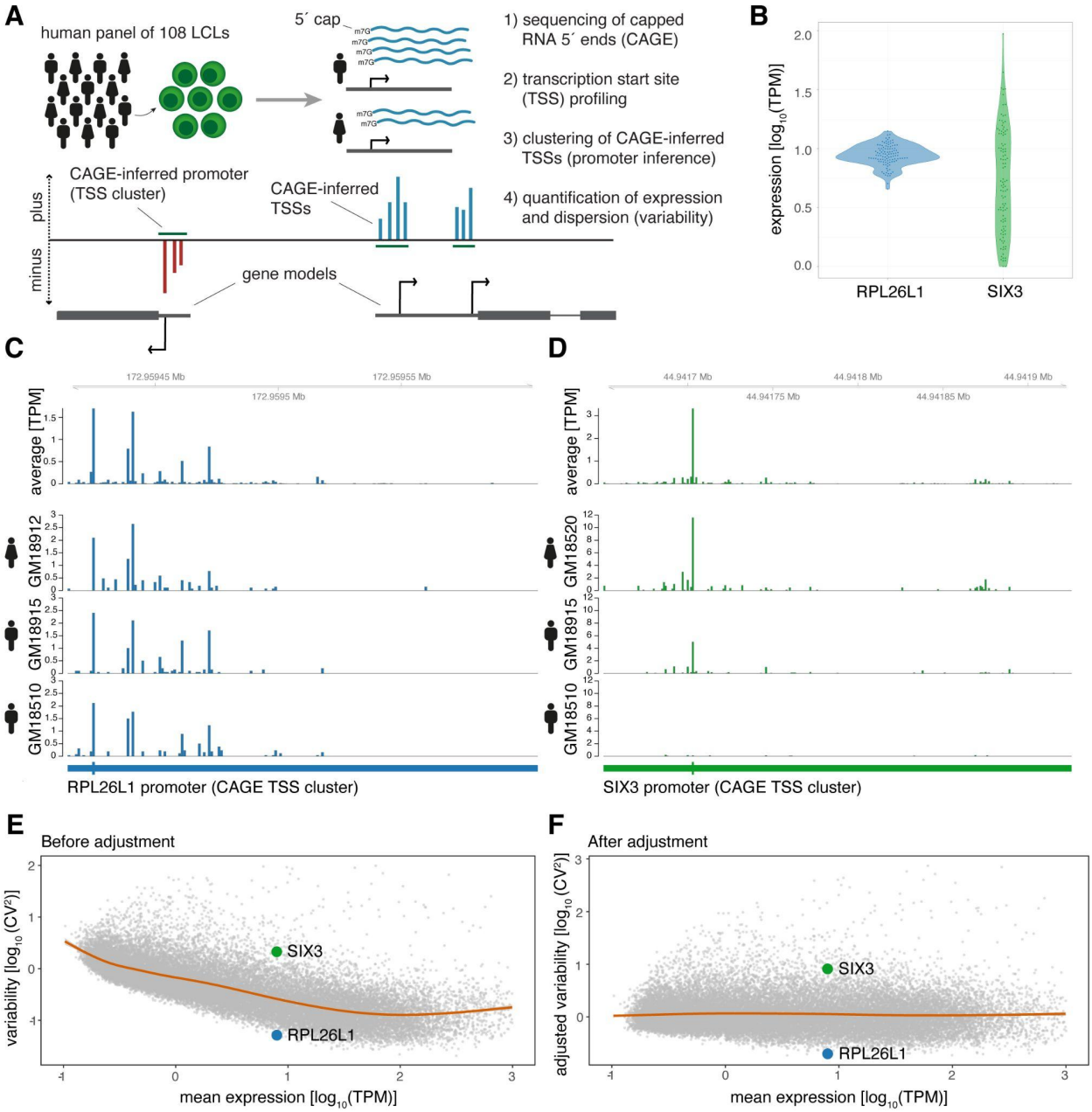

Figure 1: CAGE profiling of TSSs reveals diverse promoter variability across individuals. A: Illustration of the experimental design and approach for measuring promoter activity and variability. Capped 5' ends of RNAs from LCLs derived from 108 individuals were sequenced with CAGE, followed by individual-agnostic positional clustering of proximal CAGE-inferred TSSs (first 5' end bp of CAGE reads). The expression level of the resulting CAGE-inferred promoters proximal to annotated gene TSSs were quantified in each individual and used to measure promoter variability. **B:** Example of promoter activity (TPM normalized count of CAGE reads) across individuals for a low variable promoter (gene *RPL26L1*) and a highly variable promoter (gene *SIX3*) with similar average expression across the panel. **C-D:** Genome tracks for two promoters showing average TPM-normalized CAGE data (expression of CAGE-inferred TSSs) across individuals (top track) and TPM-normalised CAGE data for three individuals (bottom tracks) for a low variable promoter (panel C, gene *RPL26L1*) and a highly variable promoter (panel D, gene *SIX3*). **E-F:** The CV^2 (squared coefficient of variation) and mean expression relationship of

29,001 CAGE-inferred promoters across 108 individuals before (E) and after (F) adjustment of the mean expression-dispersion relationship. The CV^2 and mean expression are $\log_{10}$ transformed, orange lines show loess regression lines fitting the dispersion to the mean expression level, and example gene promoters from B-D are highlighted in colors.

Promoter expression variability is encoded in the promoter sequence

To investigate if local sequence features at promoters determine their variability in activity, we applied machine learning (convolutional neural network (CNN); Supplementary Fig. S2A; see Methods) to discern low variable promoters (N=5,054) from highly variable promoters (N=5,683) based on their local DNA sequence alone. Strikingly, the resulting model was capable of distinguishing between these promoter classes with high accuracy (area under receiving operating curve (AUC)=0.84 for the out of sample test set; Supplementary Fig S2B), equally well for highly and low variable promoters (per-class test set F1 scores of 0.76 and 0.77, respectively).

To assess which sequence features the CNN had learned to distinguish the classes, we calculated importance scores (Shrikumar et al. 2019) for each nucleotide in the input sequences for predicting low and high promoter variability. This approach can be used to identify properties or short stretches of DNA indicative of amplifying or attenuating expression variability. We applied motif discovery on clustered stretches (metaclusters) of the input sequences with high importance scores (Shrikumar et al. 2020) and matched the identified motifs to known TF binding motifs (Fornes et al. 2020). This strategy revealed TFs indicative of either high or low promoter variability (Fig. 2A-C). Noteworthy, we observed motifs for the ETS superfamily of TFs, including ELK1, ETV6, and ELK3, among low variable promoters, and motifs for PTF1A, ASCL2, and FOS-JUN heterodimer (AP-1) among highly variable promoters. These results demonstrate that the promoter sequence and its putative TF binding sites are predictive of the expression variability of a promoter.

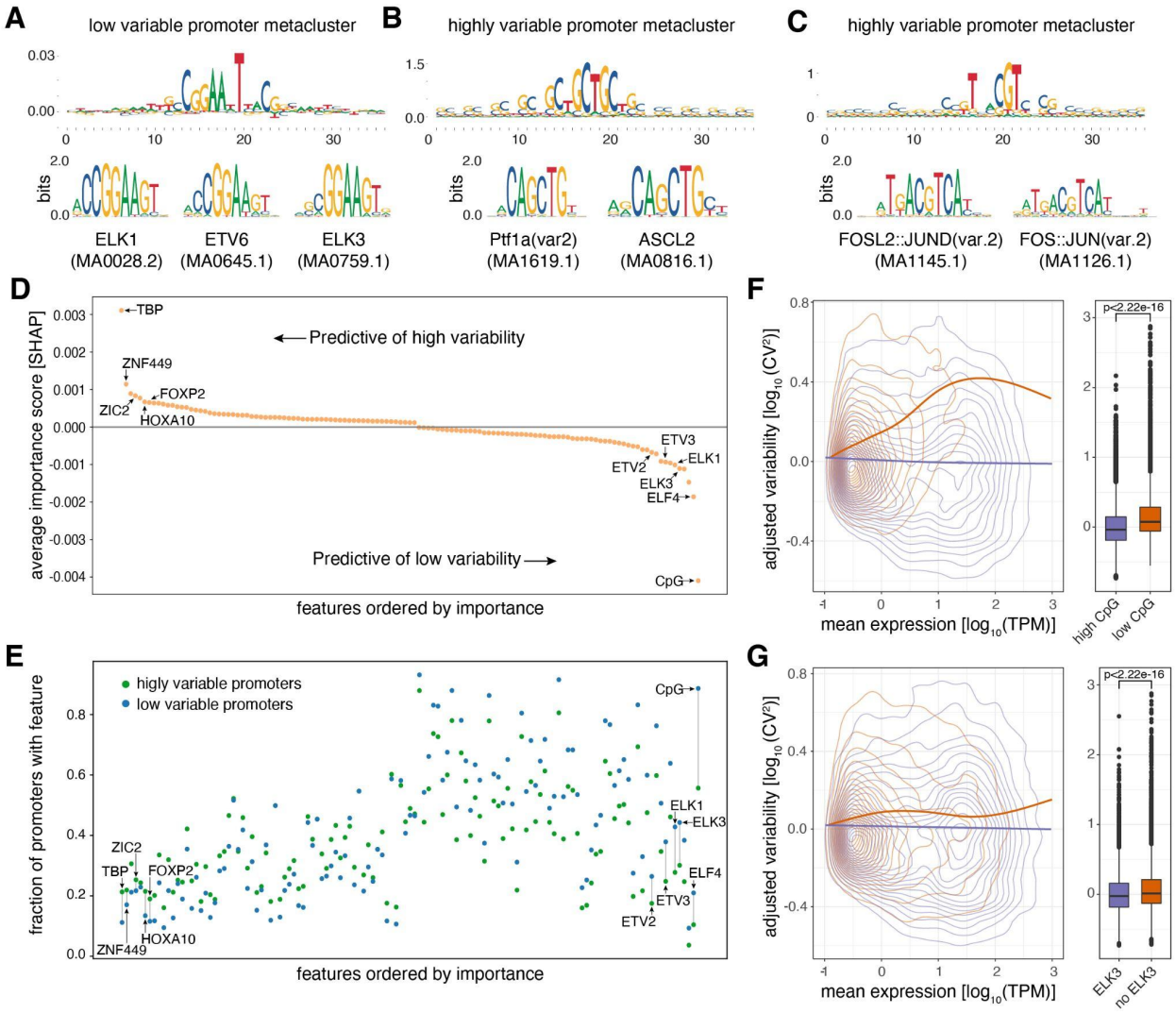

Figure 2: Promoter sequence features are highly predictive of promoter variability. A: Sequence logo of a metacluster (top) identified for low variable promoter sequences that matches known TF motifs (bottom) for ETS factors ELK1, ETV6, and ELK3. **B-C:** Sequence logos of two metaclusters (top) identified for highly variable promoter sequences that match known TF motifs (bottom) for PTF1A and ASCL2 (B) and FOSL2-JUND and FOS-JUN heterodimers (C). **D:** Average contribution (SHAP values) of each of the 125 TFs identified as important for predicting promoter variability. TFs are ordered by their average contribution to the prediction of highly variable promoters and selected TFs are highlighted. For a full version of the plot see Supplementary Figure S4A. **E:** The frequency of predicted TF binding sites (presence/absence) in highly variable (green) and low variable (blue) promoters. TFs follow the same order as in A and the same selected TFs are highlighted. For a full version of the plot see Supplementary Figure S4B,C. **F-G:** Promoters split into groups based on the presence/absence of high CpG content (F), and predicted binding sites of ELK3 (G). For both features displayed in panels F and G, the left subpanel displays the relationship between $\log_{10}$-transformed mean expression levels and $\log_{10}$-transformed adjusted CV^2 as a 2D kernel contour density plot with loess regression lines shown separately for each promoter group. The right subpanels display box-and-whisker plots of the differences in $\log_{10}$-transformed adjusted CV^2 between the two promoter groups. For box-and-whisker plots, central band: median;

boundaries: first and third quartiles; whiskers: +/- 1.5 IQR. P-values were determined using the Wilcoxon rank-sum test (***: p-value<0.05).

Sequence features of promoters are highly predictive of promoter variability

To systematically test how predictive TF binding sites at active promoters are of their variability, we made use of TF binding sites predicted from motif scanning for 746 TFs (Fornes et al. 2020). TF binding site profiles and low/high CpG content (Supplementary Fig. S3A) were collected for each identified promoter and the resulting feature data were used to train a machine learning (random forest) classifier TFs associated with either high or low variability (low variable N=5,054, highly variable N=5,683). Feature selection (Kursa and Rudnicki 2010) identified 125 of the 746 TFs to be important for classification, and a classifier based on these selected features demonstrated high predictive performance (AUC=0.78; per-class F1 score of 0.73 and 0.68 for low and highly variable promoters, respectively; Supplementary Fig. S3B), reinforcing the strong link observed between DNA sequence and promoter variability (Fig. 2).

Reverse engineering of the random forest classifier (SHAP, Shapley additive explanations) (Lundberg and Lee 2017) allowed us to calculate how much each of the 125 selected features contributed to the prediction of variability class for each promoter and whether the feature on average is indicative of amplifying or attenuating variability of expression when present in the promoter sequence (Fig. 2D; Supplementary Fig. S4A). We identified the presence of high observed/expected CpG ratio and TATA-binding protein (TBP) binding sites (TATA-boxes) to be the strongest predictive features of low and high promoter variability, respectively. Compared to TATA-box and high CpG content status, the remaining TFs contribute only marginally on their own to high or low promoter variability (Fig. 2D; Supplementary Fig. S4A). Interestingly, TFs associated with highly variable promoters are mostly associated with tissue specific or developmental regulation (e.g., FOXP2, HOXA10) while TFs predictive of low promoter variability are generally associated with ubiquitous activity across cell types and a diverse range of basic cellular processes (e.g., ELK1, ELF4, ETV3). In addition, TFs predictive of high variability (e.g., ZIC2, ZNF449, HOXA10) tend to have binding sites in

relatively few highly variable promoters while TFs predictive of low promoter variability (e.g., ELK1, ELK3) show a propensity for having binding sites present in a large number of promoters (Fig. 2E; Supplementary Fig. S4B,C). This suggests that variably expressed promoters have diverse TF binding profiles and that the regulatory grammar for promoter stability is less complex.

Although the adjusted dispersion of promoters was separated from their expression level (Fig. 1E), we observed that the presence of binding sites for some TFs that are predictive of promoter variability are also associated with promoter expression level (Supplementary Fig. S5). Importantly, despite this association, the effect of identified features on promoter variability is still evident across a range of promoter expression levels (Fig. 2F,G). This is particularly apparent for CpG islands, which we found to have an attenuating effect on promoter variability regardless of expression level (Fig. 2F).

Interestingly, many of the TFs identified as being predictive of low variability (e.g., ELK1, ELK3, ELF4, ETV2, ETV3) belong to the ETS family of TFs (Fig. 2D; Supplementary Fig. S4A), a large group of TFs that are conserved across Metazoa and characterized by their shared ETS domain that binds 5′-GGA(A/T)-3′ DNA sequences (Sharrocks 2001). ETS factors are important regulators of promoter activities in lymphoid cells (Hepkema et al. 2020), but are generally involved in a wide range of crucial cellular processes such as growth, proliferation, apoptosis, and cellular homeostasis (Kar and Gutierrez-Hartmann 2013; Oikawa and Yamada 2003; Suico et al. 2017). Furthermore, multiple ETS factors can bind in a redundant manner to the same promoters of housekeeping genes (Hollenhorst et al. 2007, 2011). We observed that the motifs of individual ETS family members are independently strong predictors of low promoter variability (Fig. 2D; Supplementary Fig. S4A) and matches to these were found in a relatively high number of promoters (Fig. 2E). However, the shared DNA binding domain of ETS factors makes it hard to discern individual factors based on their binding motifs alone (Fig. 2A). Although the ETS TFs are also associated with higher promoter activity, we observed an attenuating effect on variability across all expression levels (Fig. 2G). In addition, the degree of promoter variability decreases by an increasing number of

non-overlapping ETS binding sites (Supplementary Fig. S6A), regardless of promoter expression level (Supplementary Fig. S6B), suggesting that multiple ETS binding events can cooperate in an additive manner to stabilize promoter variability across individuals.

Taken together, our results demonstrate that the promoter sequence can influence both low and high promoter variability across human individuals independently from its impact on expression level. Several TFs were identified as contributing partially to the variability in promoter expression, while a lower complexity was identified for the regulatory grammar of stable promoters, being highly associated with higher CpG content and ETS binding sites.

Variability in promoter activity provides mechanisms of plasticity and robustness for distinct biological functions

The high performance of predicting promoter variability from local DNA sequence and the distinct TF binding profiles associated with low and highly variable promoters suggest distinct mechanisms for attenuating or amplifying variability to provide robustness or plasticity, respectively. This argues that selection of robustness over plasticity should be reflective of distinct biological processes where these mechanisms provide increased evolutionary fitness. Supporting this hypothesis, we observed that low variable promoters were highly enriched with basic cellular housekeeping processes, in particular metabolic processes (Fig. 3A). In contrast, highly variable promoters were enriched with more dynamic biological functions, including signalling, response to stimulus, and developmental processes (Fig. 3A).

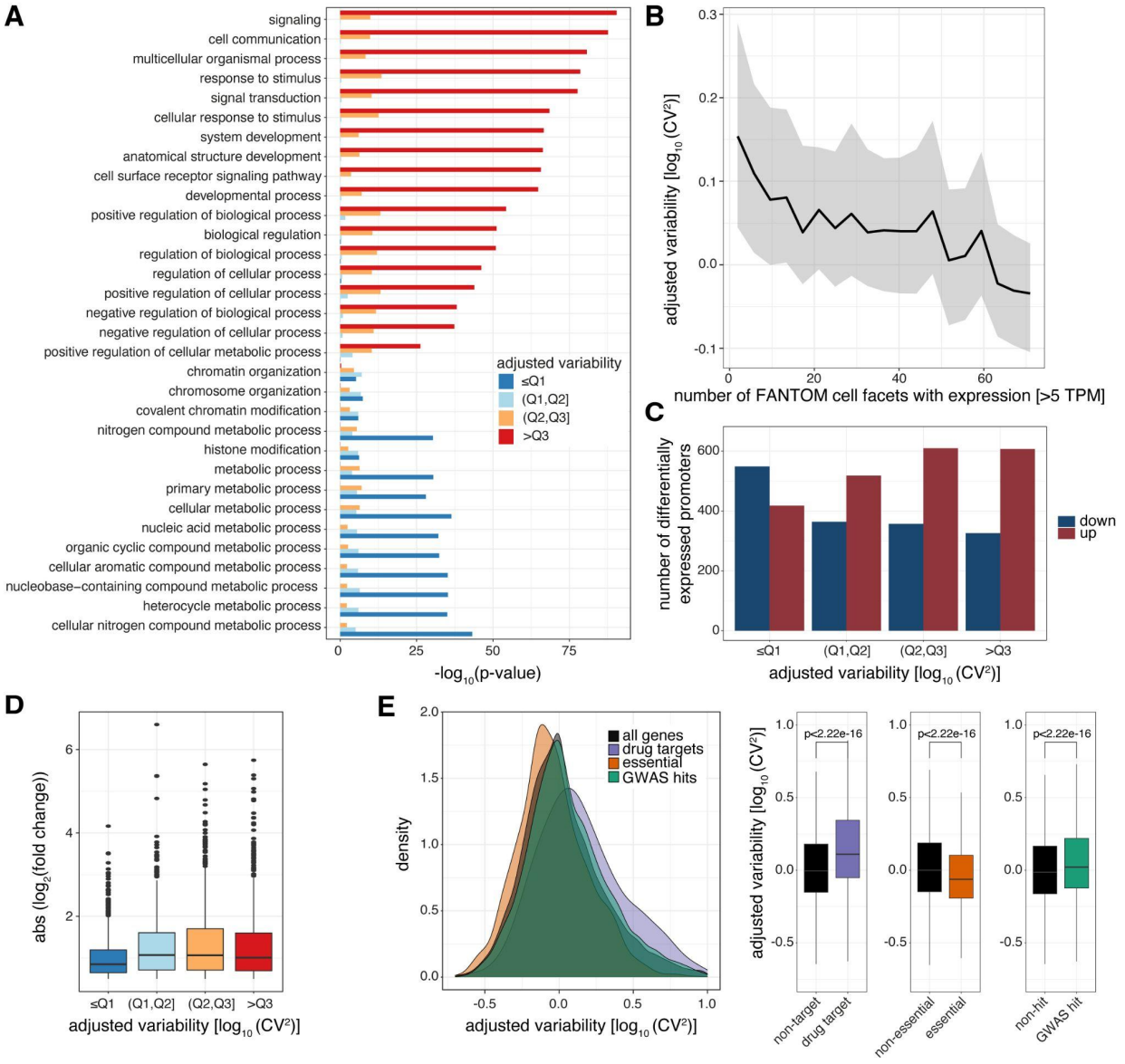

Figure 3: Levels of promoter variability are reflective of distinct biological processes and a selective trade off between robustness and plasticity. A: GO term enrichment, for biological processes, of genes split by associated promoter variability quartiles (Q1, Q2, Q3). Top 10 GO terms of all groups are displayed and ranked based on p-values of the >Q3 quartile group. **B:** Median promoter variability (line) and interquartile range (shading), as a function of the number of FANTOM cell facets (grouping of FANTOM CAGE libraries associated with the same Cell Ontology term) that the associated gene is expressed in (mean facet expression >5TPM). **C:** The number of differentially expressed promoters, split by variability quartiles, after 6h TNFα treatment. Promoters are separated into down-regulated (blue) and up-regulated (red). P-values were calculated using Fisher's exact test. **D**: Absolute $\log_2$ fold change of differentially expressed promoters, split by variability quartiles, after 6h of TNFα treatment. **E:** Distribution of promoter variability associated with drug-targets (purple), essential (orange), or GWAS hits (green) genes, compared to all promoters (black). Left: density plot of promoter variability per gene category. Right: Box-and-whisker plots of promoter variability split by each category of

genes. P-values were determined using the Wilcoxon rank-sum test. For all box-and-whisker plots, central band: median; boundaries: first and third quartiles; whiskers: +/- 1.5 IQR.

Interestingly, the same features found to be predictive of promoter variability across individuals, including CpG-islands and TATA-boxes (TBP binding sites), are also associated with low and high transcriptional noise across individual cells (Morgan and Marioni 2018; Faure et al. 2017), and the presence of a TATA-box is also associated with high gene expression variability in flies (Sigalova et al. 2020). This suggests that some of the same underlying regulatory mechanisms that dictate low or high transcriptional noise across single cells are maintained and conserved between humans and flies at an individual level and manifested to control low and high expression variability across a population, respectively, as well as housekeeping or restricted activity across cell types.

In agreement, genes of highly variable promoters tend to have higher transcriptional noise than those of low variable ones across GM12878 single cells (Cohen's d=0.411, $p<2.2\times10^{-16}$, two sample t-test; Supplementary Fig. S7A; Supplementary Table 2). Furthermore, we observed an inverse correlation between variability in promoter activity and the number of cell types (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014) and tissues (GTEx Consortium 2017) the corresponding gene is expressed in (Spearman's rank correlation $\rho = -0.21$ and -0.15 for cell types and tissues, respectively, $p< 2.2\times10^{-16}$; Fig. 3B; Supplementary Fig. S7B), demonstrating that highly variable promoters are more cell-type and tissue specific in their expression. To investigate if highly variable promoters are more prone to respond to external stimuli, we profiled GM12878 TSSs and promoter activities with CAGE before and after treatment with tumor necrosis factor (TNFα; Supplementary Table 3). This revealed an enrichment of up-regulated promoters after 6 hours of treatment among highly variable promoters ($\log_2$ odds ratio (OR)=1.529, $p=4.563\times10^{-8}$, Fisher's exact test) while low variable promoters were mostly unaffected or down-regulated ($\log_2$ OR=2.175, $p<2.2\times10^{-16}$, Fisher's exact test; Fig. 3C). In addition, low variable promoters had a weaker response (Fig. 3D).

In line with these findings, we observed drug-target genes and genes with GWAS hits to be regulated by highly variable promoters but essential genes to be regulated by low variable promoters (Fig. 3E). In contrast, when we compared promoter expression between these same groups of genes we observed no association with drug-targets or GWAS-associated genes. Although essential genes are associated with higher promoter expression, this association is comparably weaker than that with promoter variability (Supplementary Fig. S7C).

Taken together, our results demonstrate the importance of low promoter variability for cell viability and survival in different conditions and reveal the responsiveness of highly variable promoters. They further indicate that the variability observed in promoter activity across individuals is strongly associated with the regulation of its associated gene, the expression breadth across cell types, and to some extent also the transcriptional noise across single cells, which reflects a selective tradeoff between high stability and high responsiveness and specificity.

Promoters with low variability have flexible transcription initiation architectures

Promoters are associated with different levels of spread of their TSS locations, which has led to their classification into broad or narrow (sharp) promoters according to their positional width (Akalin et al. 2009; Lehner 2008; Carninci et al. 2006). Although the shape and distinct biological mechanisms of these promoter classes, e.g., housekeeping activities of broad promoters, are conserved across species (Carninci et al. 2006; Hoskins et al. 2011), the selective pressure for positional dispersion of TSSs and its association with promoter variability are poorly understood.

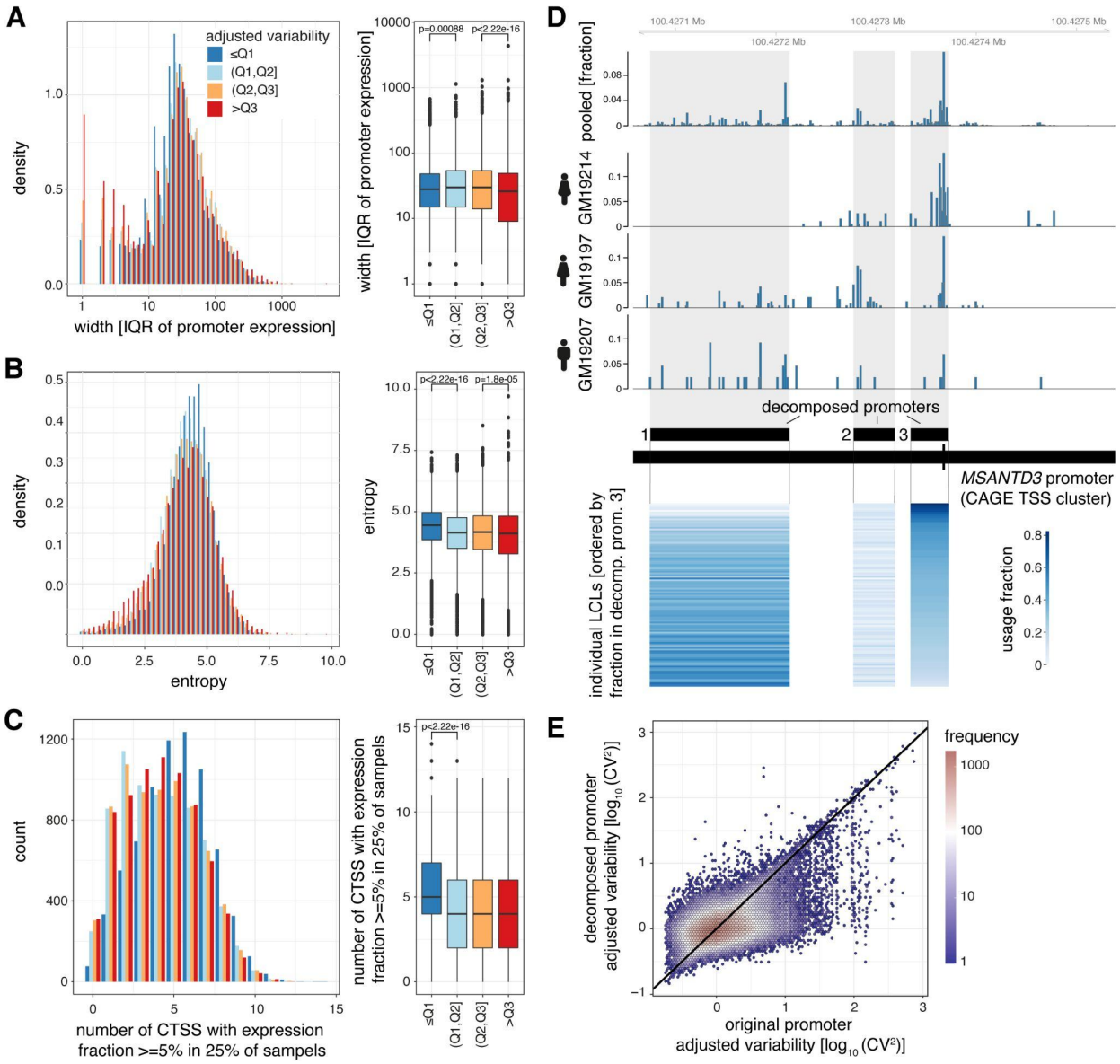

Figure 4: Low variable promoters exhibit flexibility in transcription initiation architecture. A-C: Promoter shape metrics for promoters split by variability quartiles. The left subpanels display the distribution of IQRs (widths containing the 25th to 75th percentiles of contained CAGE signal) (A), Shannon entropy (B) and the number of TSSs with expression fraction >=5% in 25% of samples (C). The right subpanels display box-and-whisker plots of the differences in these metrics across promoters split by variability quartiles (central band: median; boundaries: first and third quartiles; whiskers: +/- 1.5 IQR.). **D:** Upper panel displays genome tracks showing TSS usage contribution (fraction out of total expression) to the overall promoter expression of gene *MSANTD3* across the panel (top track) and for three individuals with variable contributions (bottom tracks). Lower panel displays heatmaps showing the contribution of decomposed promoters (TSS sub-clusters, shaded in genome tracks) to the overall expression across all 108 individuals, ordered based on contribution of the dominant decomposed promoter. **E:** The relationship between $\log_{10}$-transformed adjusted CV^2 of the original promoter and $\log_{10}$-transformed adjusted CV^2 of local-maxima decomposed promoters as a 2D density chart.

Surprisingly, analysis of promoter widths revealed only a weak relationship with promoter variability. We observed an enrichment of highly variable promoters within narrow promoters having an interquartile range (IQR) of their CAGE tags within a width of 1 to 5 bp ($P<2.2\times10^{-16}$, OR=2.04, Fisher's exact test). Low variable promoters, on the other hand, were enriched among those of size 10 to 25 bp ($P<2.2\times10^{-16}$, OR=1.44, Fisher's exact test), but beyond this width the association is lost (Fig 4A). To simultaneously capture the spread of TSSs and their relative frequencies within a promoter, we considered the Shannon entropy as a measure of TSS positional dispersion (Hoskins et al. 2011). We observed that low variable promoters are associated with a higher entropy than promoters with high variability (Fig. 4B). In addition, low variable promoters tend to have more TSSs substantially contributing to the overall expression of the promoters across individuals (Fig. 4C). We reasoned that a weaker association between low promoter variability and broad width than with high entropy may be due to low variable promoters being composed of multiple independent clusters of TSSs (multi-modal peaks). Indeed, decomposition of multi-modal peaks within the CAGE TSS profiles of promoters (Supplementary Table 4) demonstrated that higher entropy reflects an increased number of decomposed promoters, as indicated by their number of local maxima of CAGE signals (Supplementary Fig. S8A).

Interestingly, the decomposed promoters of gene *MSANTD3* (Fig. 4D) clearly illustrate that the activity of sub-clusters of TSSs within promoters and their contributions to the overall activity of the encompassing promoter can vary to a great extent between individuals. To assess in general how individual decomposed promoters influence the overall promoter variability, we calculated the expression-adjusted dispersion (adjusted CV^2) of local-maxima decomposed promoters. Remarkably, many of the decomposed promoters show a vastly different variability across individuals compared to the promoters they originate from (Fig. 4E). This disagreement indicates that individual TSSs within the same promoter may operate and be regulated independently of each other, which may contribute to the overall buffering or plasticity of the promoter and, in turn, the gene. As multi-modal peaks are mainly found to be associated with low

variable promoters of ubiquitously expressed genes, we hypothesize that this flexibility in TSS usage may act as a compensatory mechanism to stabilize their expression.

Alternative TSSs of low variability promoters confer mutational robustness

While genetic variants associated with gene expression levels (expression quantitative trait loci, eQTLs) frequently occur within gene promoters, they are rarely found associated with housekeeping or ubiquitously expressed genes, and when they are, they have limited effect sizes (GTEx Consortium 2017). One explanation for this observation is that mutations that would significantly alter the expression of such genes would be detrimental to cell viability. Alternatively, the rare and limited effects of eQTLs on housekeeping genes might be due to mechanisms promoting mutational robustness.

To test if flexibility in TSS usage within a promoter may cause mutational robustness, we performed local eQTL analysis on promoters (within 25kb). We tested both the association between the genotypes of common genetic variants (MAF ≥ 10%) and the expression of promoters (promoter eQTL, prQTLs; Fig. 5A, top) as well as their association with the contribution of decomposed promoters to the overall expression of the encompassing (non-decomposed) promoter (fraction eQTL, frQTL; Fig. 5A, bottom).

2,457 promoters were associated with at least one prQTLs (5% FDR; Supplementary Table 5). While prQTLs were observed across all levels of promoter variability, they were more commonly associated with highly variable promoters (Fig. 5B). Fewer prQTL single nucleotide polymorphisms (SNPs) and, in general, common variants were found proximal to low variable promoters, indicating a negative selection for these. However, the effect size for the most significant prQTL variant (lead SNP) for each promoter was positively associated with the expression variability of the promoter (Spearman's rank correlation $\rho = 0.16$, $p < 2.2 \times 10^{-16}$, Supplementary Fig. S8B). This indicates that, in addition to having fewer proximal genetic variants, low variable promoters are less likely to have prQTLs with large regulatory effects.

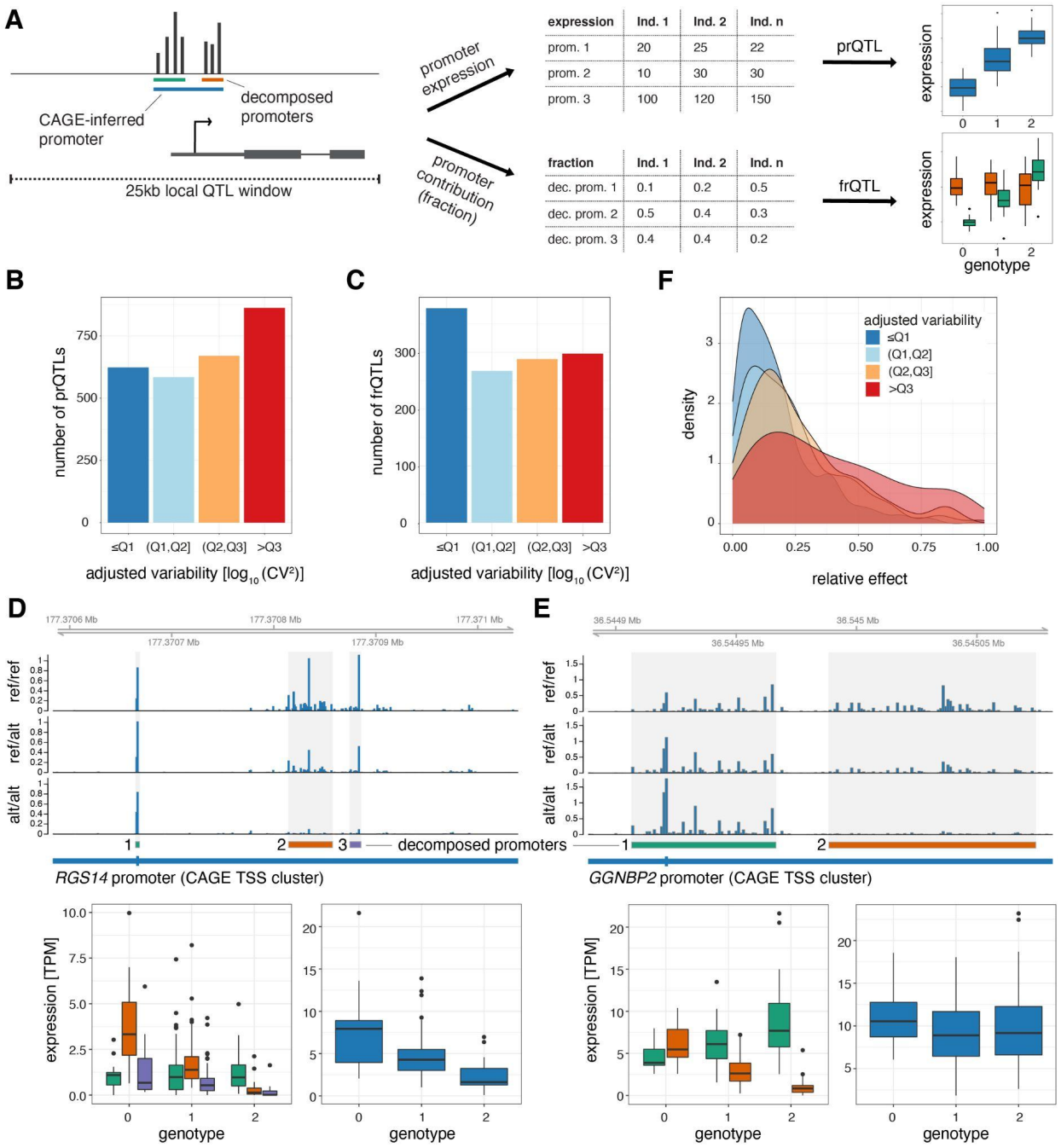

Figure 5: Plasticity in TSS usage promotes mutational robustness. A: Illustration of the strategy for testing the effects of genetic variants on promoter expression (prQTLs, TPM-normalized CAGE counts) and on decomposed promoter contribution to the encompassing promoter expression (frQTLs, ratios of TPM-normalized CAGE counts between decomposed and encompassing promoters). For both approaches only SNPs within 25kb of the promoter summit were tested. **B:** Number of prQTLs detected (FDR<0.05), split by promoter variability quartiles. **C:** Number of encompassing promoters with at least one frQTL detected for a contained decomposed promoter (FDR<0.05), split by encompassing promoter variability quartiles. **D-E:** Examples of two promoters associated with frQTLs for a highly variable promoter with limited buffering of promoter expression (panel D, gene *RGS14*) and for a low variable

promoter with strong buffering of promoter expression (panel E, gene *GGNBP2*). Upper panels display genome tracks showing average TPM-normalized CAGE data across homozygous individuals for the reference allele (top track), heterozygous individuals (middle track), and homozygous individuals for the variant (alternative) allele (bottom track). The bottom left subpanels display box-and-whisker plots of the differences in TPM-normalized CAGE data between genotypes for each decomposed promoter. The bottom right subpanels display box-and-whisker plots of the differences in TPM-normalized CAGE data between the three genotypes for the original encompassing promoter. For all box-and-whisker plots, central band: median; boundaries: first and third quartiles; whiskers: +/- 1.5 IQR. **F.** Density plot of the maximal relative change in expression between reference and variant alleles (relative effect size) for the most significant frQTL of each broad promoter with FDR≥5%, split by variability quartiles.

We identified 1,230 promoters to be associated with at least one frQTL (5% FDR; Supplementary Table 6). Unlike the prQTLs, the frQTLs were more commonly associated with sub-clusters of TSSs (decomposed promoters) from low variable promoters (Fig. 5C). Conceptually, the frQTLs can affect TSS usage and overall expression levels to different degrees, as exemplified by the promoters of genes *RGS14* and *GGNBP2* (Fig. 5D,E). Gene *RGS14* has three decomposed promoters localized within its promoter (Fig. 5D), for which SNP rs56235845 (chr5:177371039 T/G) was strongly associated with the contribution to the overall promoter activity for only decomposed promoters 1 and 2 (frQTL beta=0.210, -0.181, -0.062; FDR=2.42×10^{-5}, 2.54×10^{-8}, 2.64×10^{-2}, for decomposed promoters 1, 2, and 3, respectively). Despite the limited association of the variant with decomposed promoter 3, it still had a noticeable association with the overall promoter activity (prQTL beta=-2.47, FDR=3.57×10^{-5}; Fig. 5D, bottom right). In contrast, SNP rs9906189 (chr17:36549567 G/A) was strongly associated with the contribution to the overall promoter activity for both decomposed promoters of gene *GGNBP2* (frQTL beta=0.222, -0.222; FDR=2.05×10^{-26}, 2.05×10^{-25}, for decomposed promoters 1 and 2, respectively), but in opposite directions (Fig 5E). Interestingly, this switch in TSS usage translates into a limited effect on the overall promoter activity (prQTL beta=-0.063, FDR=0.989; Fig. 5E, bottom right).

Both examples, a partial shift and a switch in decomposed promoter usage, are indicative of plasticity in TSS usage, which can secure tolerable levels of steady-state mRNA. Although frQTLs were associated with promoters across the wide spectrum of promoter variabilities (Fig. 5A), they showed a striking difference in their relative effect

on the overall promoter activity (maximal relative change in expression between reference and variant alleles; Fig. 5E). frQTLs associated with highly variable promoters tend to have a larger relative effect on the overall promoter activity compared to frQTLs associated with low variable promoters (Fig. 5E). This association is further maintained at the gene level (adjusted RNA-seq (Lappalainen et al. 2013b) CV^2; Supplementary Fig. S8C; Supplementary Table 7), demonstrating that individual differences in TSS usage contribute to low promoter variability and, in turn, low gene variability. In total, we found 286 promoters (out of 1,230) of 284 genes to be associated with stabilizing frQTLs, for which the same SNP was associated with at least two decomposed promoters (5% FDR) with relative effects in opposite directions (Supplementary Table 8). TSS usage flexibility thus confers mutational robustness that stabilizes the variability of promoters and their associated genes.

Taken together, integrating prQTLs and frQTLs provides novel insights into how genetic variation can influence promoter regulation and its potential impact on gene expression. We demonstrate that low variable promoters, characterized by multiple decomposed promoters (multi-modal TSS usage) are less affected by the presence of genetic variants compared to highly variable promoters. In addition, we find that part of this tolerance can be explained by a, previously unreported, mechanism of mutational robustness through plasticity in TSS usage.

Discussion

In this study we provide an extensive characterization of promoter-associated features influencing variability in promoter activity across human individuals and demonstrate their importance for determining stability, responsiveness, and specificity. Overall, we show that the local DNA sequence, putative TF binding sites, and transcription initiation architecture of promoters are highly predictive of promoter variability.

Although the TF-based model was able to predict promoter variability well (AUC=0.78 on the test set), it did not perform as well as the convolutional neural network model

(AUC=0.84 on the test set), which was trained on DNA sequence alone. This indicates that even though the TF binding grammar of promoters influences their variability, additional information influencing variability may be encoded within the sequence of TSS-proximal regions. For instance, di- or tri-nucleotide sequence patterns and the AT-richness of promoters, that influence local nucleosome positioning (Segal et al. 2006), impose different requirements for chromatin remodelling activities (Lorch et al. 2014) at gene promoters of low and high variability, which in turn may affect their responsiveness.

Notably, many of the regulatory features we, and others (Sigalova et al. 2020), have identified to be predictive of promoter variability, including the presence or absence of CpG islands and TATA boxes, have previously been linked with different levels of transcriptional noise as inferred from single-cell experiments (Morgan and Marioni 2018; Faure et al. 2017). This suggests that variability in promoter activity across individuals partly reflects the stochasticity in gene expression across cells. Given that the underlying sources of variation are different, e.g., genetic and environmental versus stochastic, this indicates that mechanisms that contribute to the buffering of stochastic noise at a single cell level can also contribute to the attenuation of genetic and environmental variation at an individual level.

Despite a clear association with high promoter CpG content and housekeeping genes, low variable promoters were not strongly associated with a broader width, which one would expect from promoters in CpG islands and with housekeeping activity (Carninci et al. 2006). Rather, we found that low variability requires a certain minimum promoter width, which can encompass a transcription initiation architecture competent of attenuating variability through flexible TSS usage across individuals. Switching between proximal TSSs (core promoters) within a larger promoter is fundamentally different from switches between alternative promoters (Zhang et al. 2017; Valen et al. 2008; Garieri et al. 2017), which will more likely lead to differences in transcript and protein isoforms. Rather, a flexible initiation architecture enables several points of entries for RNA polymerase II to initiate in the same promoter, ensuring gene expression across

different cell types (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014; Kawaji et al. 2006) and developmental stages (Haberle et al. 2014). Here we show that such flexibility also attenuates variability across individuals for the same cell type. We further demonstrate that plasticity in TSS usage within a promoter confers a, previously unreported, layer of mutational robustness that can buffer the effects of genetic variants, leading to limited or no impact on the overall promoter expression. Of note, it is likely that such buffering events will be revealed for more genes with an increased sample size or a focus on other cell types.

A flexibility in TSS usage thus ensures transcriptional robustness of genes both in different environments and in the face of genetic variation. Since promoter shape is generally conserved across orthologous promoters (Carninci et al. 2006; Hoskins et al. 2011), it is plausible that robustness through flexible TSS usage is a conserved mechanism. In support, genetic variants affect promoter shape for ubiquitously expressed genes in flies with limited effect on promoter expression (Schor et al. 2017). Changes in promoter shape in flies thus likely recapitulates the plasticity in TSS usage across human LCLs, despite apparent differences in core promoter elements and regulatory features between flies and humans. Future investigations will reveal the characteristics of core promoters capable of switching within a promoter, and how these compare across species, which is unfeasible with current data due to the proximity between decomposed promoters.

Taken together, our results favor a model in which the regulation of transcriptional noise across single cells reflects specificity across cell types and dispersion across individuals with shared mechanisms conferring stochastic, genetic and environmental robustness (Fig. 6). There are several implications of this model. First, the link between low transcriptional noise and low individual variability of promoters and their associations with ubiquitous and essential genes indicate that regulatory mechanisms that ensure broad expression across cell types may enforce low variability across individuals and single cells. Second, our results indicate that encoding responsiveness or developmentally restricted expression patterns of gene promoters may require high

stochasticity in expression across single cells, which in turn may disallow ubiquitous expression across cell types. Thus, it is likely that increased variability is not just reflecting the absence of regulatory mechanisms that attenuate variability but the presence of those that amplify it. Finally, given that mutational robustness through flexible TSS usage is mostly associated with low variable genes, this implies that cell-type restricted, responsive and developmental genes may be more susceptible for trait-associated genetic variants, which finds support in the literature (Timshel et al. 2020; Finucane et al. 2015; Kundaje et al. 2015).

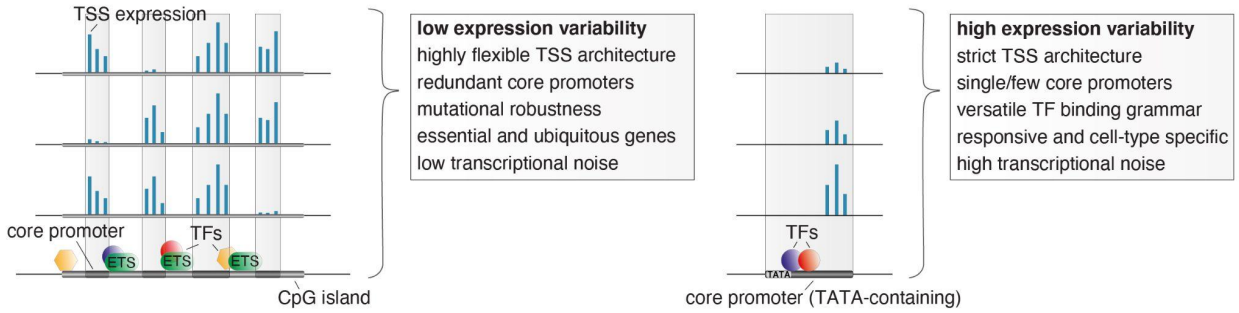

Figure 6: Unifying mechanisms influencing the variability in expression across individuals, the specificity in expression across cell types, and the stochasticity in expression across individual cells. Low variable promoters (left) frequently associate with high CpG content (CpG islands), multiple binding sites of ETS factors, and a highly flexible transcription initiation architecture arising from multiple independent and redundant core promoters. These stabilizing features along with a less complex TF binding grammar likely also act to buffer transcriptional noise across single cells and cause ubiquitous expression across cell types. The flexibility in redundant core promoter activities confers a novel layer of mutational robustness to genes. Highly variable promoters (right), on the other hand, are associated with a highly versatile TF regulatory grammar, TATA boxes, and low flexibility in TSS usage. These features cause, in addition to high expression variability between individuals, a responsiveness to external stimuli, cell-type restricted activity, high transcriptional noise across single cells, and less tolerance for genetic variants.

Methods

LCL cell culturing

Epstein-Barr virus immortalized LCLs were obtained from the NIGMS Human Genetic Cell Repository at Coriell Institute for Medical Research. Cells were incubated at 37°C at 5% carbon dioxide in the Roswell Park Memorial Institute (RPMI) Medium 1640 supplemented with 2mM L-glutamine and 20% of non-inactivated fetal bovine serum and antibiotics. Cell cultures were split every few days for maintenance. As these cell lines were freshly purchased, mycoplasma contamination screening was not undertaken.

CAGE library preparation, sequencing and mapping

CAGE libraries were prepared as described elsewhere (Andersson et al. 2014b; Takahashi et al. 2012) from total RNA from LCLs. Some libraries underwent a second round of size selection by e-gelling to remove primer dimers. Reads were trimmed to remove linker sequences, filtered for minimum sequence quality of Q30 in 50%, and rRNA reads matching U13369.1 were removed using rRNAdust (version 1.06 (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014)). Mapping to the human reference genome (hg38) was performed using BWA (version 0.7.15-r1140) allowing for max two mismatches per read. To limit mapping biases, reads were re-mapped using the WASP pipeline (van de Geijn et al. 2015) and BWA using the same parameters, taking into account biallelic SNVs (Lowy-Gallego et al. 2019). Reads with a mapping quality of 30 were retained for further analyses.

CAGE tag clustering, filtering and quantification

CAGE-defined transcription start sites (CTSSs) were identified from 5’ ends of CAGE read using CAGEfightR (version 1.10 (Thodberg et al. 2019)). The expression of CTSSs was quantified from the number of CAGE reads sharing 5’ ends. Positional clustering of CTSSs (tag clustering) within 60 bp of each other was performed on pooled CAGE data, including all 108 LCLs, using only CTSS with at least 1 read in at least 5 libraries. The expression of each tag cluster (CAGE-inferred promoter) in each individual cell line was

quantified by aggregating the expression of CTSSs falling within the defined tag cluster region. Expression levels were converted to tags per million (TPM), by normalizing the expression count of each tag cluster in each library to its library size scaled by 10^6. Tag clusters were annotated using GENCODE (hg38, version 29) and subsequently filtered to be proximal to GENCODE-annotated TSSs (within 1000bp upstream) and have at least 10 read counts in more than 10 libraries. The resulting 29,001 gene-associated CAGE-inferred promoters were later decomposed by a local maxima decomposition approach to split multi-modal tag clusters (https://github.com/anderssonlab/CAGEfightR_extensions).

RNA-seq data analysis

Gene expression data quantified in the recount2 project (Collado-Torres et al. 2017) using Geuvadis YRI RNA-seq data (Lappalainen et al. 2013b) was downloaded using the recount R package. Only genes with more than 1 transcript per million in at least 10% of YRI samples were considered for further analyses.

scRNA-seq data analysis

GM12878 10X Genomics scRNA-seq data (Osorio et al. 2019) was downloaded from Gene Expression Omnibus (GSE126321) and processed using Seurat (version 4.0.3, (Hao et al. 2021)). Cells with a proportion of mitochondrial reads lower than 10% and a library size smaller than 2.5 times the standard deviation from the average library size were considered. The expression of genes with read counts observed in at least 10 cells were normalized using scran (version 1.18.7, (Lun et al. 2016)) and retained for expression variability calculation.

Measuring expression variability across individuals

The raw dispersion of each CAGE tag cluster was calculated using the squared coefficient of variation (CV^2) of TPM-normalized tag cluster expression across the LCL panel and subsequently $\log_{10}$-transformed. Adjustment of the mean expression-dispersion relationship was performed by subtracting the expected dispersion for each promoter according to its expression level, using a running median

of raw dispersions by mean expression level across the panel (Kolodziejczyk et al. 2015; Newman et al. 2006). The same strategy was used to calculate the adjusted dispersion of gene expression from RNA-seq and scRNA-seq data.

Neural network model, training and hyperparameter tuning

The neural network model is described in Supplementary Figure 2A. The neural network model uses as input one-hot-encoded DNA sequence (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]) with fixed length of 600 bp as input to predict low and highly variable promoter activity as output. The model consists of one convolutional layer with 128 hidden units and a kernel size of 10, followed by global average pooling and two dense layers with 128 and 2 nodes, respectively. Batch normalization and dropout (0.1) were applied after each layer. The relu (Agarap 2019) activation function was used in all layers except the final layer, in which a sigmoid activation function was used to predict the variability class (low or high).

Regions from chromosomes 2 and 3 were only used as the test set and regions from the remaining chromosomes were used for training and hyperparameter tuning with a 5-fold cross-validation. Hyperparameters were manually adjusted to yield the best performance on the validation set. The neural network model was implemented and trained in Keras (version 2.3.0, https://github.com/fchollet/keras) with TensorFlow backend (version 1.14 (Abadi et al. 2016)) using the Adam optimizer (Kingma and Ba 2017) with a learning rate of 0.0001, batch size of 64, and early stopping with the patience of 15 epochs.

We initially used the first and third quartiles (Q1 and Q3) to distinguish low variable promoters (≤Q1) from highly variable promoters (>Q3), corresponding to an adjusted CV^2 of -0.1490 and 0.1922, respectively. To reduce false positives, we slightly adjusted the thresholds for low and highly variable promoters to -0.20 and 0.25, respectively. Thus, the final training and test sets for the neural network model together consisted of 5,054 low variable and 5,683 highly variable promoters. To ensure consistency, the same thresholds were used for training and testing with Random Forest (see below).

Motif discovery using DeepLIFT and TF–MoDISco

To interpret the neural network model we used DeepLIFT (Shrikumar et al. 2019), a feature attribution method, to compute importance scores in an input sequence. The importance scores were supplied to TF-MoDISco (Transcription Factor Motif Discovery from Importance Scores (Shrikumar et al. 2020)) to identify motifs of DNA stretches (seqlets) with high importance for the predictions. DeepLIFT and TF-MoDISco were used independently for the low variable and highly variable promoters. TF-MoDISco identified 18,035 seqlets for low variable promoters and 21,942 seqlets for highly variable promoters by using the importance scores from DeepLIFT over a width of 15 bp with a flank size of 5 bp and a FDR threshold of 0.05. The seqlets identified were merged in 41 and 47 metaclusters for low and highly variable promoters, respectively.

We used Tomtom (MEME package 5.1.1 (Gupta et al. 2007)) and the resulting metaclusters as input to match known TF motifs (in MEME format) from the JASPAR database (release 2020, hg38 (Fornes et al. 2020)). To match TF motifs with metaclusters, we compared each non-redundant JASPAR vertebrate frequency matrix with the metacluster using Tomtom based on the Sandelin and Wasserman distance (Sandelin and Wasserman 2004). Matches were considered those with a minimum overlap between query and target of 5 nucleotides and a p-value < 0.05.

Random forest, Boruta and SHAP analysis

CpG observed/expected ratio was calculated in windows covering +/- 500bp around the pooled summit TSS within each promoter. High CpG content was defined as promoters with CpG observed/expected ratio >0.5. Predicted transcription factor binding sites for 746 TFs (hg38) were obtained from JASPAR (release 2020, hg38 (Fornes et al. 2020)) and for each TF, presence/absence was obtained by overlapping predicted TF binding sites with promoters considered in the modelling. Together, the CpG content status and the presence/absence of predicted TF binding sites were used as features for predicting high and low variability using Random Forest (Pedregosa et al. 2011).

Similarly to the neural network model, promoters from chromosomes 2 and 3 were only used as the test set. The remaining promoters were used for training and hyperparameter tuning with 5 fold cross-validation. The Random Forest model was implemented and trained in Scikit-learn (version 2.3.0) with 500 trees, a maximum depth of trees of 10, 50 samples split per node, and 50 samples to be at a leaf node. The remaining hyperparameters were kept with default values.

Instead of selecting features directly from the Random Forest model, the BorutaShap package (Keany 2020) was used for feature selection. The main advantage of using the Boruta approach is that the features compete with their randomized version (or shadow feature) and not with themselves. Therefore a feature is considered relevant only if its score is higher than the best randomized feature. In this way, from the 746 original features, 125 features were kept. The features were selected using only promoters from the training set. Finally, the SHAP library (Lundberg and Lee 2017) was used to explain the importance of the 125 selected features for the two promoter classes.

Tissue-, cell-type specificity and gene annotations

Gene expression values across tissues were obtained from the GTEx consortium (GTEx Consortium 2017). Promoters were considered expressed in tissues in which their corresponding gene had ≥5 RPKM average expression across donors.

Gene expression values across cell types were obtained from the FANTOM5 project (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014). The average normalized (tags per million, TPM) expression per gene was calculated across samples associated with the same cell type facet (grouping of CAGE libraries according to Cell Ontology annotation of samples, according to (Andersson et al. 2014a)), and a gene was considered expressed in a cell type facet if the average expression was ≥5 TPM.

Gene lists for FDA approved drug-targets (Wishart et al. 2018), essential genes (Hart et al. 2017) and GWAS hits (MacArthur et al. 2017) were downloaded from the MacArthur Lab Repository (https://github.com/macarthur-lab/gene_lists).

GM12878 cell culturing, TNF-α stimulation and differential expression analysis

GM12878 cells were obtained from the NHGRI Sample Repository for Human Genetic Research at Coriell Institute for Medical Research. GM12878 cells were stimulated with 25ng/ml TNF-α for 0 and 6 hours prior to harvesting with four replicates for each condition. Cell culturing, CAGE library preparation and mapping was done as described above for the LCL panel. CAGE reads supporting each of the final filtered promoters identified in the LCL panel were counted for each replicate using CAGEfightR (version 1.10, (Thodberg et al. 2019)). Differential expression analysis of the aggregated CTSS counts was performed using DESeq2 (version 1.30.1 (Love et al. 2014)). Promoters with FDR-adjusted p-value ≤0.05 were considered to be differentially expressed.

Mapping QTLs

prQTLs and frQTLs were mapped using the MatrixEQTL R package (version 2.3, (Shabalin 2012)). We controlled for genetic population stratification and library preparation batches by including these as covariates. In addition, we included the first 5 principal components derived from normalized promoter expression values (TPM) as covariates for prQTLs.

For prQTL detection, all 29,001 promoters were tested using TPM-normalized values. For frQTLs, we calculated the fractional contribution of each decomposed promoter to the expression of its original promoter. To focus the frQTL analysis on relevant shifts in TSS usage, we selected only decomposed promoters overlapping the same non-decomposed promoter, which had at least 2 decomposed promoters with ≥0.05 fractional contribution in at least half of the samples, resulting in 37,663 decomposed promoters.

For each promoter, we tested common (minor allele frequency ≥10%) biallelic SNVs (Lowy-Gallego et al. 2019) at a maximum distance of 25kb from CTSS with maximum pooled CAGE signal within each promoter for association with changes in promoter expression levels or decomposed promoter contribution. Resulting p-values were FDR-adjusted according to the total number of promoters or decomposed promoters

tested genome-wide within the MatrixEQTL R package and prQTLs and frQTLs with FDR≤5% were retained. A promoter was associated with an frQTL if at least one of its decomposed promoters was associated with a frQTL at FDR<5%.

Acknowledgements

We thank members of the Andersson lab for rewarding discussions. This work was supported by funding from the Danish Council for Independent Research [grant 6108-00038], the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [grant 638173], and the Novo Nordisk Foundation [grant NNF18OC0052570]. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

Author contributions

H.E. and R.A. conceived the project; H.E. led the data analysis with support from N.A., S.R., and R.A.; M.S. performed machine learning; C.V. performed CAGE experiments with contributions from J.B.L.; R.A. supervised the project; H.E. and R.A. wrote the manuscript; all authors reviewed the final manuscript.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, et al. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *ArXiv160304467 Cs*. http://arxiv.org/abs/1603.04467 (Accessed October 27, 2021).

Agarap AF. 2019. Deep Learning using Rectified Linear Units (ReLU). *ArXiv180308375 Cs Stat*. http://arxiv.org/abs/1803.08375 (Accessed October 27, 2021).

Akalin A, Fredman D, Arner E, Dong X, Bryne JC, Suzuki H, Daub CO, Hayashizaki Y, Lenhard B. 2009. Transcriptional features of genomic regulatory blocks. *Genome Biol* **10**: 1–13.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014a. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Andersson R, Refsing Andersen P, Valen E, Core LJ, Bornholdt J, Boyd M, Heick Jensen T, Sandelin A. 2014b. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat Commun* **5**: 5336.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Bartha I, di Iulio J, Venter JC, Telenti A. 2018. Human gene essentiality. *Nat Rev Genet* **19**: 51–62.

Boettiger AN, Levine M. 2009. Synchronous and stochastic patterns of gene activation in the Drosophila embryo. *Science* **325**: 471–473.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**: 626–635.

Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. 2017. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol* **35**: 319–321.

Eldar A, Elowitz MB. 2010. Functional roles for noise in genetic circuits. *Nature* **467**: 167–173.

Eling N, Richard AC, Richardson S, Marioni JC, Vallejos CA. 2018. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Syst* **7**: 284-294.e12.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, et al. 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462–470.

Faure AJ, Schmiedel JM, Lehner B. 2017. Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Syst* **5**: 471-484.e4.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K, et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**: 1228–1235.

Fornes O, Castro-Mondragon JA, Khan A, Van Der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, et al. 2020. JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Res* **48**: D87–D92.

Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766–D773.

Garieri M, Delaneau O, Santoni F, Fish RJ, Mull D, Carninci P, Dermitzakis ET, Antonarakis SE, Fort A. 2017. The effect of genetic variation on promoter usage and enhancer activity.

Nat Commun **8**: 1190.
GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204–213.
Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8**: R24.
Haberle V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, Gehrig J, Dong X, Akalin A, Suzuki AM, et al. 2014. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507**: 381–385.
Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184**: 3573-3587.e29.
Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, Chandrashekhar M, Hustedt N, Seth S, Noonan A, et al. 2017. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 Bethesda Md* **7**: 2719–2727.
Hepkema J, Lee NK, Stewart BJ, Ruangroengkulrith S, Charoensawan V, Clatworthy MR, Hemberg M. 2020. Predicting the impact of sequence motifs on gene regulation using single-cell data. *bioRxiv* 2020.11.26.400218.
Hollenhorst PC, McIntosh LP, Graves BJ. 2011. Genomic and Biochemical Insights into the Specificity of ETS Transcription Factors. *Annu Rev Biochem* **80**: 437–471.
Hollenhorst PC, Shah AA, Hopkins C, Graves BJ. 2007. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. *Genes Dev* **21**: 1882–1894.
Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, et al. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* **21**: 182–192.
Kar A, Gutierrez-Hartmann A. 2013. Molecular mechanisms of ETS transcription factor-mediated tumorigenesis. *Crit Rev Biochem Mol Biol* **48**: 522–543.
Kawaji H, Frith MC, Katayama S, Sandelin A, Kai C, Kawai J, Carninci P, Hayashizaki Y. 2006. Dynamic usage of transcription start sites within core promoters. *Genome Biol* **7**: R118.
Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, Shin JW, Kojima-Ishiyama M, Kawano M, Murata M, et al. 2014. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res* **24**: 708–717.
Keany E. 2020. *BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values.* Zenodo https://zenodo.org/record/4247618 (Accessed October 25, 2021).
Kingma DP, Ba J. 2017. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs*. http://arxiv.org/abs/1412.6980 (Accessed October 27, 2021).
Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, et al. 2015. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell* **17**: 471–485.
Kundaje A, Ernst J, Yen A, Heravi-Moussavi A, Zhang Z, Amin V, Schultz MD, Sarkar A, Wu Y-C, Pfenning A, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.
Kursa MB, Rudnicki WR. 2010. Feature Selection with the **Boruta** Package. *J Stat Softw* **36**. http://www.jstatsoft.org/v36/i11/ (Accessed September 17, 2021).
Lappalainen T, Sammeth M, Friedländer MR, 'T Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013a. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–511.
Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013b. Transcriptome and

genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–511.
Lehner B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol* **4**: 170.
Lorch Y, Maier-Davis B, Kornberg RD. 2014. Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. *Genes Dev* **28**: 2492–2497.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.
Lowy-Gallego E, Fairley S, Zheng-Bradley X, Ruffier M, Clarke L, Flicek P, 1000 Genomes Project Consortium. 2019. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res* **4**: 50.
Lun ATL, McCarthy DJ, Marioni JC. 2016. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**: 2122.
Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* **2017-Decem**: 4766–4775.
MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**: D896–D901.
Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**: 773–777.
Morgan MD, Marioni JC. 2018. CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness. *Genome Biol* **19**: 13–81.
Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS. 2006. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**: 840–846.
Oikawa T, Yamada T. 2003. Molecular biology of the Ets family of transcription factors. *Gene* **303**: 11–34.
Osorio D, Yu X, Yu P, Serpedin E, Cai JJ. 2019. Single-cell RNA sequencing of a European and an African lymphoblastoid cell line. *Sci Data* **6**: 112.
Payne JL, Wagner A. 2015. Mechanisms of mutational robustness in transcriptional regulation. *Front Genet* **6**. http://journal.frontiersin.org/Article/10.3389/fgene.2015.00322/abstract (Accessed September 13, 2021).
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-Learn: Machine Learning in Python. *J Mach Learn Res* **12**: 2825–2830.
Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**: 768–772.
Ravarani CNJ, Chalancon G, Breker M, de Groot NS, Babu MM. 2016. Affinity and competition for TBP are molecular determinants of gene expression noise. *Nat Commun* **7**: 10417.
Sandelin A, Wasserman WW. 2004. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol* **338**: 207–215.
Schor IE, Degner JF, Harnett D, Cannavò E, Casale FP, Shim H, Garfield DA, Birney E, Stephens M, Stegle O, et al. 2017. Promoter shape varies across populations and affects promoter evolution and expression noise. *Nat Genet* **49**: 550–558.
Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**: 772–778.
Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinforma Oxf Engl* **28**: 1353–1358.
Sharrocks AD. 2001. The ETS-domain transcription factor family. *Nat Rev Mol Cell Biol* **2**: 827–837.

Shrikumar A, Greenside P, Kundaje A. 2019. Learning Important Features Through Propagating Activation Differences. *ArXiv170402685 Cs*. http://arxiv.org/abs/1704.02685 (Accessed September 17, 2021).

Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, Nair S, Kundaje A. 2020. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. *ArXiv181100416 Cs Q-Bio Stat*. http://arxiv.org/abs/1811.00416 (Accessed September 17, 2021).

Sigalova OM, Shaeiri A, Forneris M, Furlong EE, Zaugg JB. 2020. Predictive features of gene expression variation reveal mechanistic link with differential expression. *Mol Syst Biol* **16**: e9539.

Spivakov M, Akhtar J, Kheradpour P, Beal K, Girardot C, Koscielny G, Herrero J, Kellis M, Furlong EE, Birney E. 2012. Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biol* **13**: R49.

Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. 2007. Gene-Expression Variation Within and Among Human Populations. *Am J Hum Genet* **80**: 502–509.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. *Science* **315**: 848–853.

Suico MA, Shuto T, Kai H. 2017. Roles and regulations of the ETS transcription factor ELF4/MEF. *J Mol Cell Biol* **9**: 168–177.

Takahashi H, Lassmann T, Murata M, Carninci P. 2012. 5′ End-Centered Expression Profiling Using Cap-Analysis Gene Expression and Next-Generation Sequencing. *Nat Protoc* **7**: 542–561.

Thodberg M, Thieffry A, Vitting-Seerup K, Andersson R, Sandelin A. 2019. CAGEfightR: analysis of 5'-end data using R/Bioconductor. *BMC Bioinformatics* **20**: 487.

Timshel PN, Thompson JJ, Pers TH. 2020. Genetic mapping of etiologic brain cell types for obesity eds. R. Loos and N. Barkai. *eLife* **9**: e55851.

Urban EA, Johnston RJ. 2018. Buffering and Amplifying Transcriptional Noise During Cell Fate Specification. *Front Genet* **9**: 591.

Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, et al. 2008. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res* **19**: 255–265.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**: 1061–1063.

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**: D1074–D1082.

Zhang P, Dimont E, Ha T, Swanson DJ, FANTOM Consortium, Hide W, Goldowitz D. 2017. Relatively frequent switching of transcription start sites during cerebellar development. *BMC Genomics* **18**: 461.

Supplementary Figures

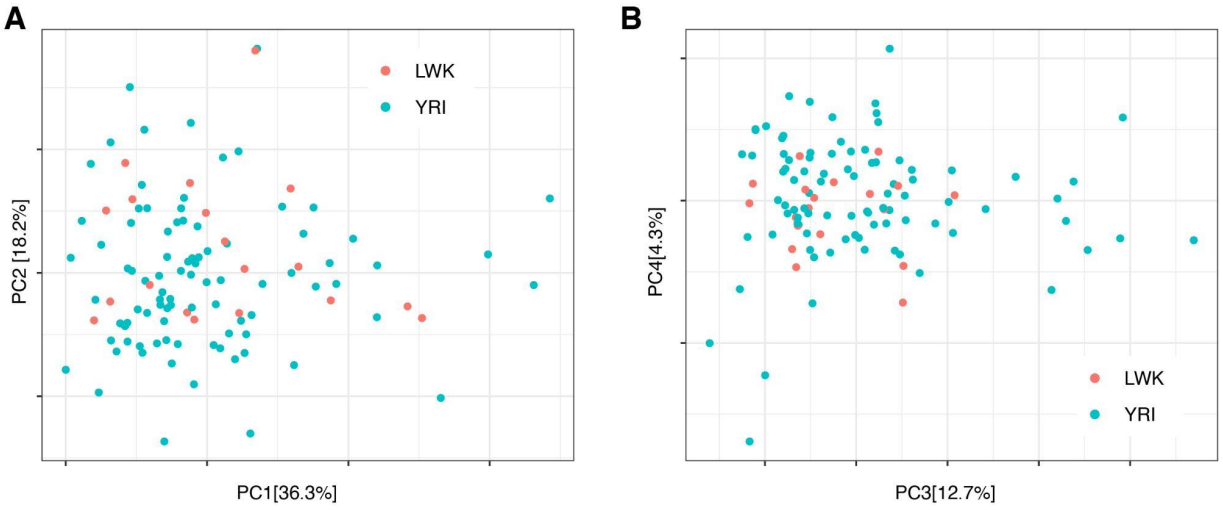

Figure S1. PCA plot of promoter expression (CAGE) across the LCL panel. 1st and 2nd (A), and 3rd and 4th (B) principal components (PCs), colored according to population (YRI and LWK). PCA was performed using TPM-normalized expression for all 29,001 promoters. Percentage of variation accounted for by each principal component is shown in brackets with the axis label.

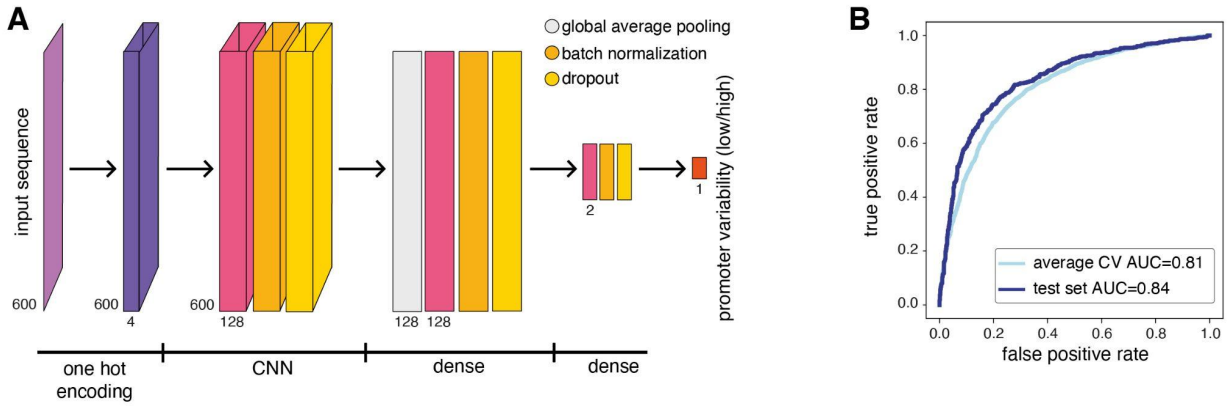

Figure S2. Neural network model and performance. A: Neural network architecture used for learning promoter variability from promoter sequence. The architecture is composed of one convolutional layer with 128 hidden units, followed by global average pooling and two dense layers with 128 and 2 nodes, respectively. **B:** Receiver-operating curves (ROC) for average cross validation (light blue, AUC=0.81) and the test set (dark blue, AUC=0.84).

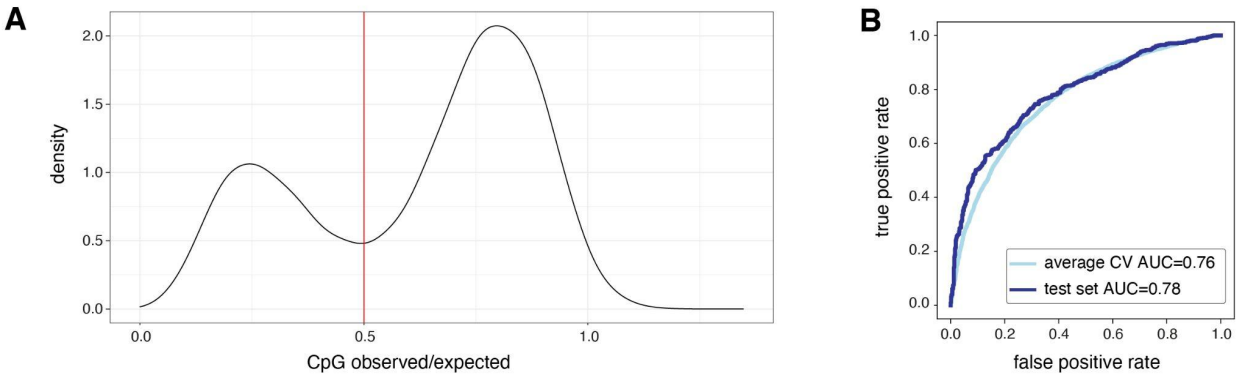

Figure S3. Random forest features and performance. A: Observed / expected CpG ratio calculated in windows covering +/- 500 bp around the CAGE summit position of considered promoters. Red vertical line marks the threshold (0.5) between low and high CpG content. **B:** Random forest model receiver-operating curves (ROC) for average cross validation (light blue, AUC=0.76) and the test set (dark blue, AUC=0.78).

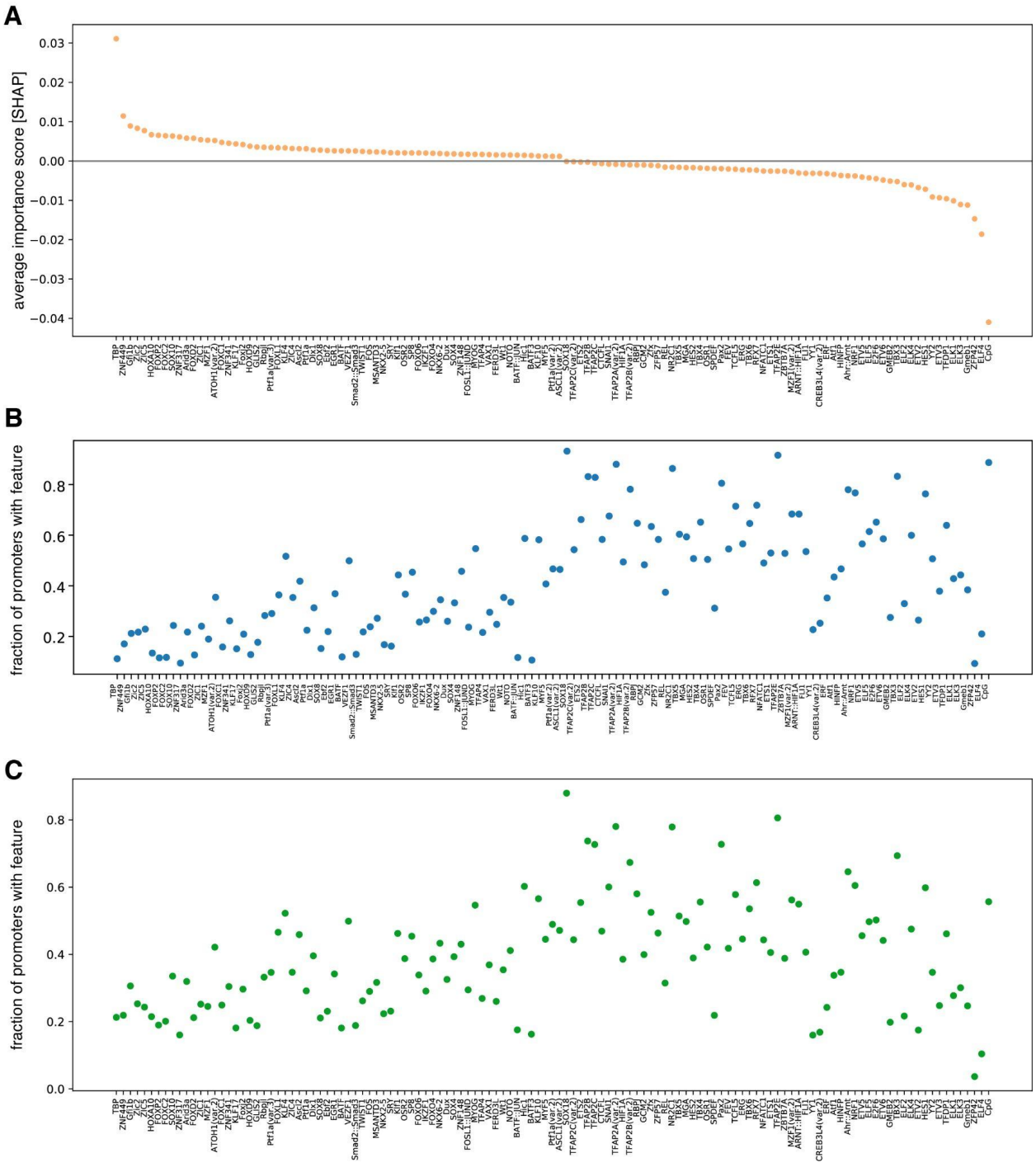

Figure S4. Full panel of promoter features found to be predictive of promoter variability. A: Average contribution (SHAP values) of each of the 125 TFs identified as important for predicting promoter variability. TFs are ordered by their average contribution to the prediction of highly variable promoters. **B-C:** The frequency of predicted TF binding sites (presence/absence) in low variable (B) and highly variable (C) promoters. TFs follow the same order as in panel A.

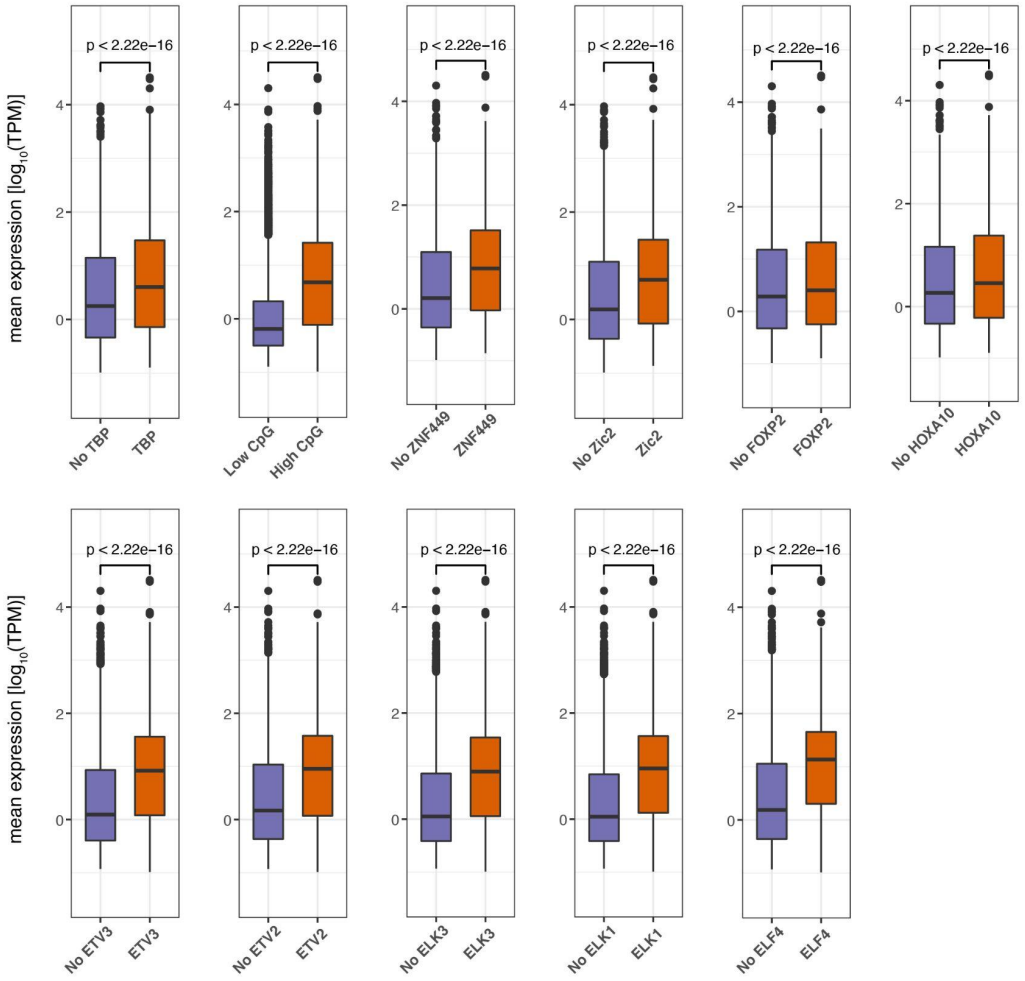

Figure S5. Association between TF binding sites and promoter expression level. Box-and-whisker plots displaying the difference in TPM normalized promoter expression between in the absence (blue) or presence (orange) of TF binding sites. For all box-and-whisker plots, central band: median; boundaries: first and third quartiles; whiskers: +/- 1.5 IQR. P-values were determined using the Wilcoxon rank-sum test (***: p-value<0.05).

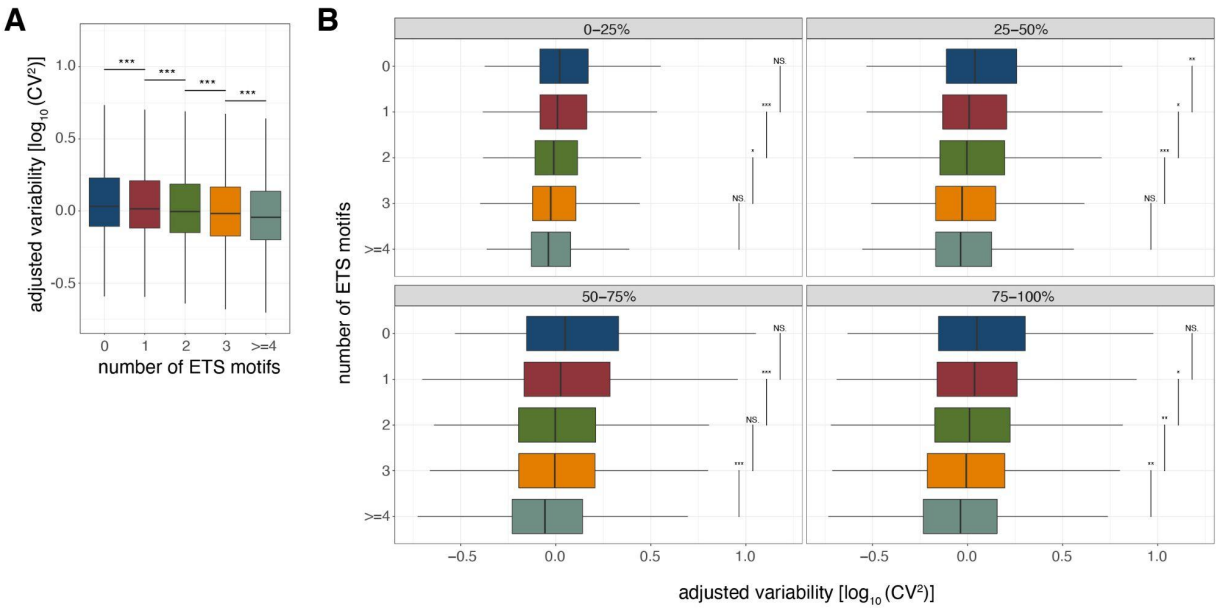

Figure S6. A: Variability of promoters grouped by their number of predicted non-overlapping ETS binding sites. **B.** Variability of promoters grouped by their number of predicted non-overlapping ETS binding sites split based on promoter expression level quartiles. For all box-and-whisker plots, central band: median; boundaries: first and third quartiles; whiskers: +/- 1.5 IQR. In both panels, outliers were not plotted. P-values were determined using the Wilcoxon rank-sum test (***: p-value<0.05).

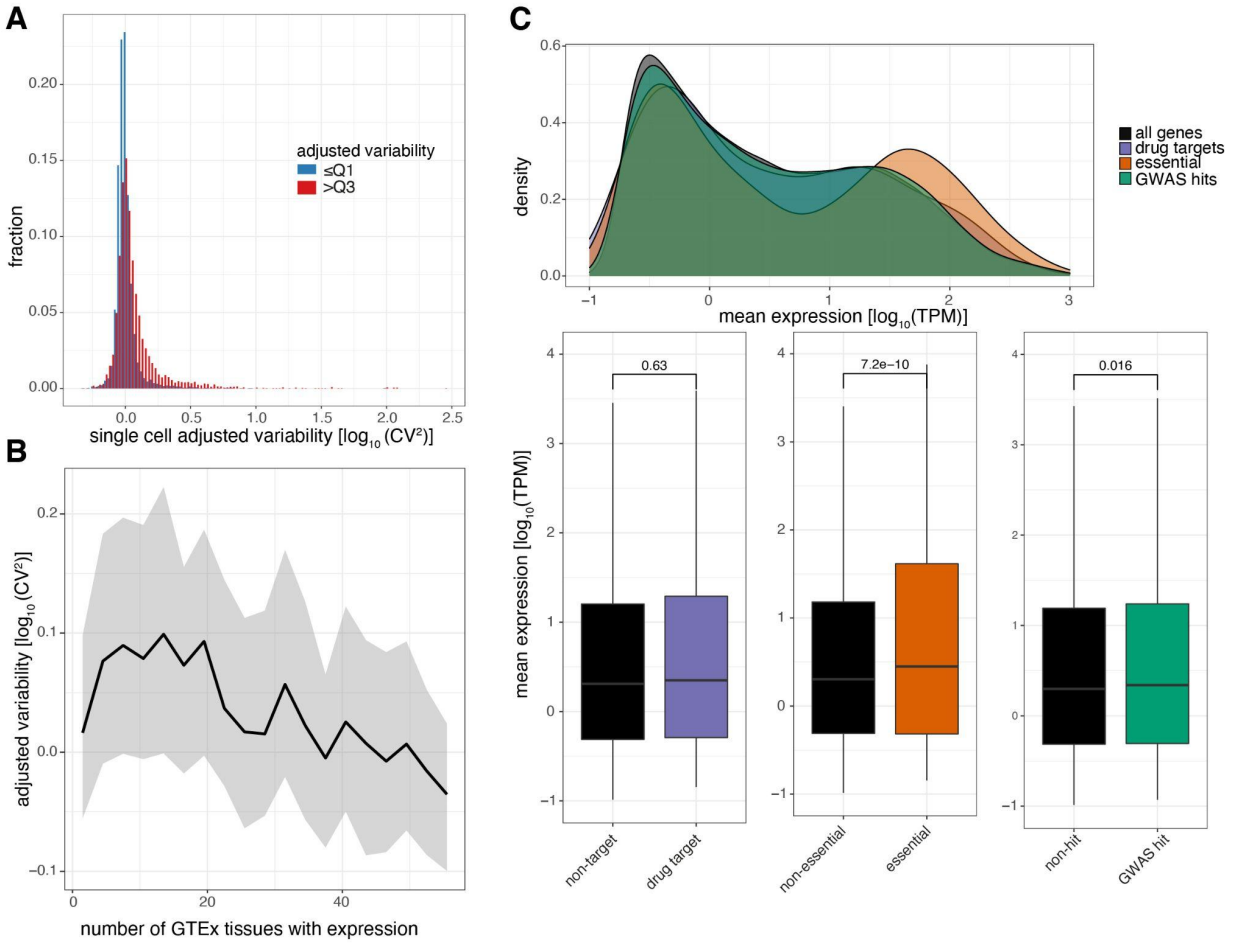

Figure S7. Levels of promoter variability are reflective of distinct biological processes. A: Distribution of single cell adjusted variability [$\log_{10}(CV^2)$] of genes for low variable promoters (blue) and highly variable promoters (red). **B:** Median promoter variability (line) and interquartile range (shading), as a function of the number of GTEx tissues the associated gene is expressed in (median tissue expression >5 RPKM). **C:** Distribution of promoter expression associated with drug-targets (purple), essential (orange), or GWAS hits (green) genes, compared to all promoters (black). Top: density plot of promoter expression per gene category. Bottom: Box-and-whisker plots of promoter expression split by each category of genes. P-values were determined using the Wilcoxon rank-sum test. For all box-and-whisker plots, central band: median; boundaries: first and third quartiles; whiskers: +/- 1.5 IQR.

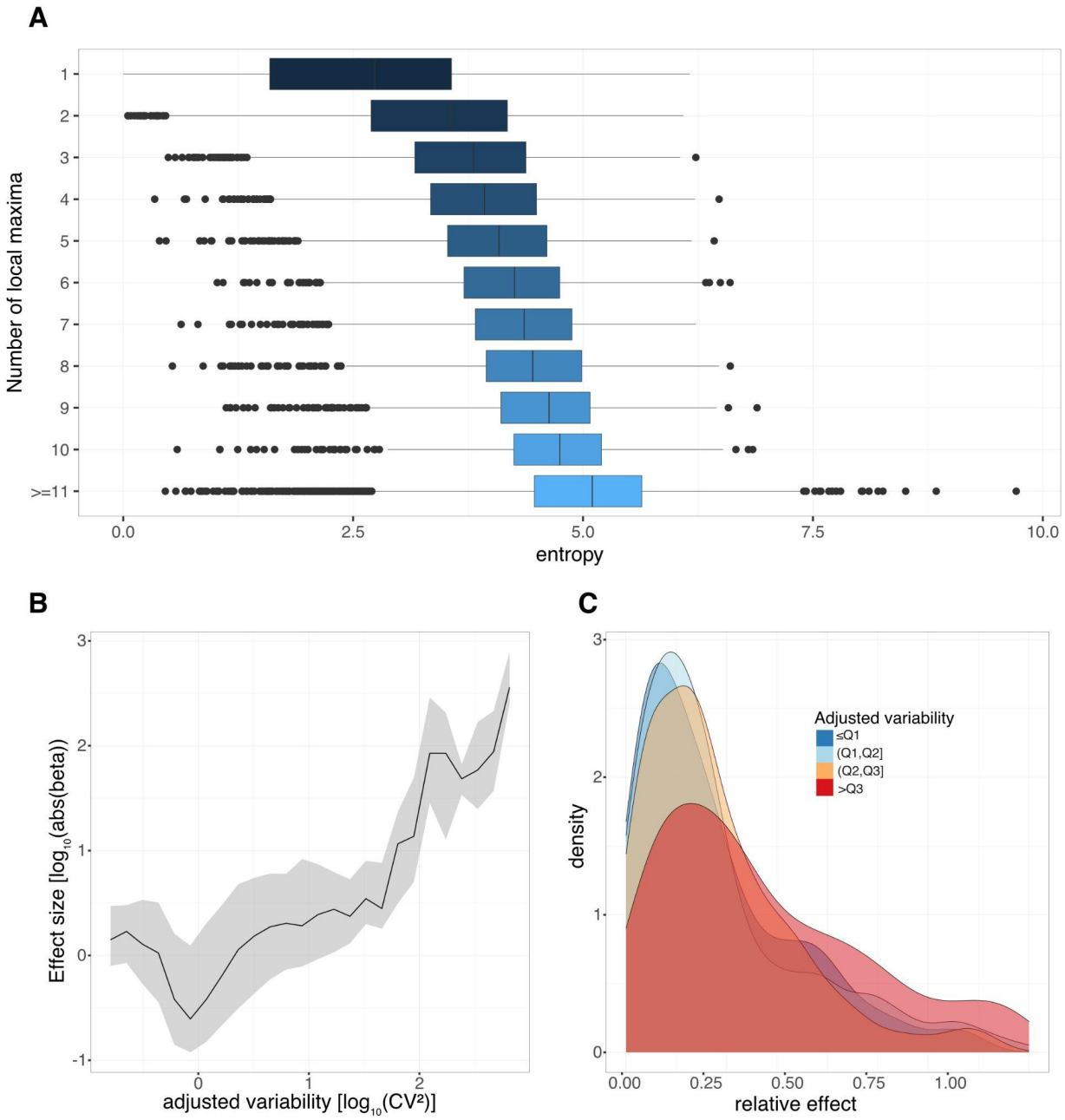

Figure S8. Low variable promoters are less affected by proximal genetic variation. A: Box-and-whisker plot displaying the relationship between the Shannon entropy and the number of local maxima CAGE signals of promoters. Central band: median; boundaries: first and third quartiles; whiskers: +/- 1.5 IQR. **B:** Median effect size [log10(abs(beta))] for the most significant prQTL for each promoter (line) and interquartile range (shading), as a function of adjusted promoter variability. **C:** Density plot of the relative effect sizes for the most significant prQTL of each promoter with FDR<=5% split by RNA-seq-derived variability quartiles.