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[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@bio.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 can be largely explained by a promoter's DNA sequence and its binding sites for specific transcription factors. We show that promoter expression variability reflects the biological process of a gene, demonstrating a selective trade-off between stability for metabolic genes and plasticity for responsive genes and those involved in signaling. Promoters with a rigid transcription start site architecture are more prone to have variable expression and to be associated with genetic variants with large effect sizes, while a flexible usage of transcription start sites within a promoter attenuates expression variability and limits genotypic effects. Our work provides insights into the variable nature of responsive genes and reveals a novel mechanism for supplying transcriptional and mutational robustness to essential genes through multiple independent transcription start site regions within a promoter.

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 (Eldar and Elowitz, 2010; Urban and Johnston, 2018), 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 (Montgomery et al., 2010; Pickrell et al., 2010; Stranger et al., 2007) of the majority of human genes (GTEx Consortium, 2017; Lappalainen et al., 2013; Storey et al., 2007). 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 (Akalin et al., 2009; Carninci et al., 2006; 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 (Morgan and Marioni, 2018; Ravarani et al., 2016; Sigalova et al., 2020), 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). These studies have relied on model organisms or focused on transcriptional noise across single cells. As of yet, 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, notwithstanding possible genotypic differences. 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 expression variability and identify switches between proximal core promoters 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) (Supplementary Table. 1). 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 2). This individual-agnostic strategy ensured a focus on promoters that are 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).

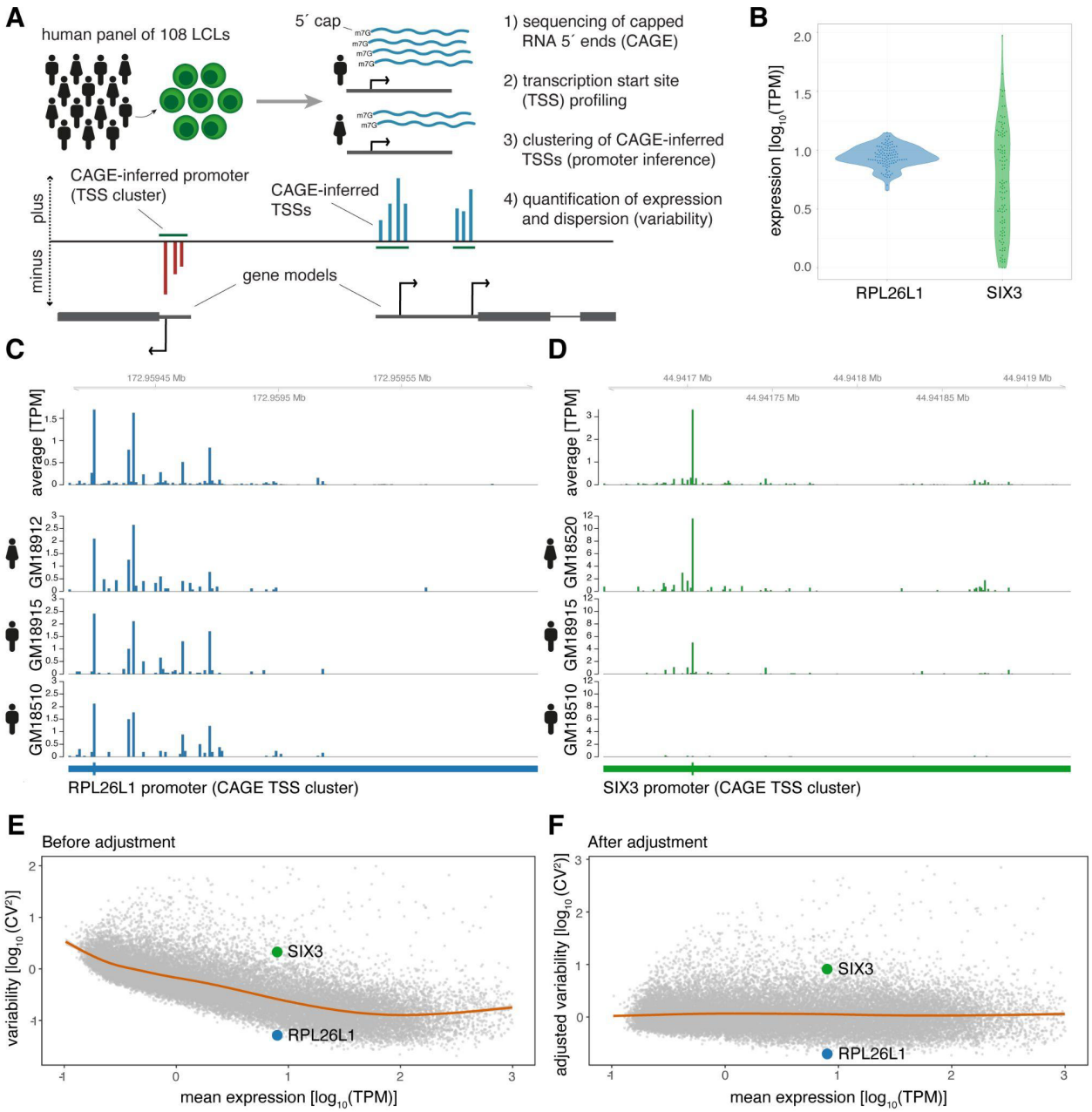

Figure 1: CAGE profiling of TSSs reveals diverse promoter variability across individuals. A: Illustration of the experimental design and approach for measuring promoter activity and variability. Capped 5' ends of RNAs from LCLs derived from 108 individuals were sequenced with CAGE, followed by individual-agnostic 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-normalized 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.

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).

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. 2A; see Methods) to discern low variable promoters (N=5,054) from highly variable promoters (N=5,683) based on their local DNA sequence alone. We considered the genomic reference sequence to model the intrinsic component of variability encoded within the promoter sequence independently of local genetic variants within the panel. 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 2B), 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 using DeepLift (Shrikumar et al., 2019) for each nucleotide in the input sequences for predicting low and high promoter variability. This

approach relies on backpropagation of the contributions of all neurons in the CNN to the input features, nucleotides, and can therefore be used to identify properties or short stretches of DNA indicative of amplifying or attenuating expression variability. We applied motif discovery on clustered stretches, so called metaclusters, of the input sequences with high importance scores (Shrikumar et al., 2020) and matched the identified metaclusters 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, associated with 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.

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. 3A) were collected for each identified promoter and the resulting feature data were used to train a machine learning (random forest) classifier features associated with either high or low variability (low variable N=5,054, highly variable N=5,683). Feature selection (Kursa and Rudnicki, 2010) identified 124 of the 746 TFs as well as CpG ratio to be important for classification, and a classifier based on these selected features demonstrated high predictive performance (AUC=0.79; per-class F1 score of 0.73 and 0.68 for low and highly variable promoters, respectively; Supplementary Fig. 3B), reinforcing the strong link observed between DNA sequence and promoter variability (Supplementary Fig 2B).

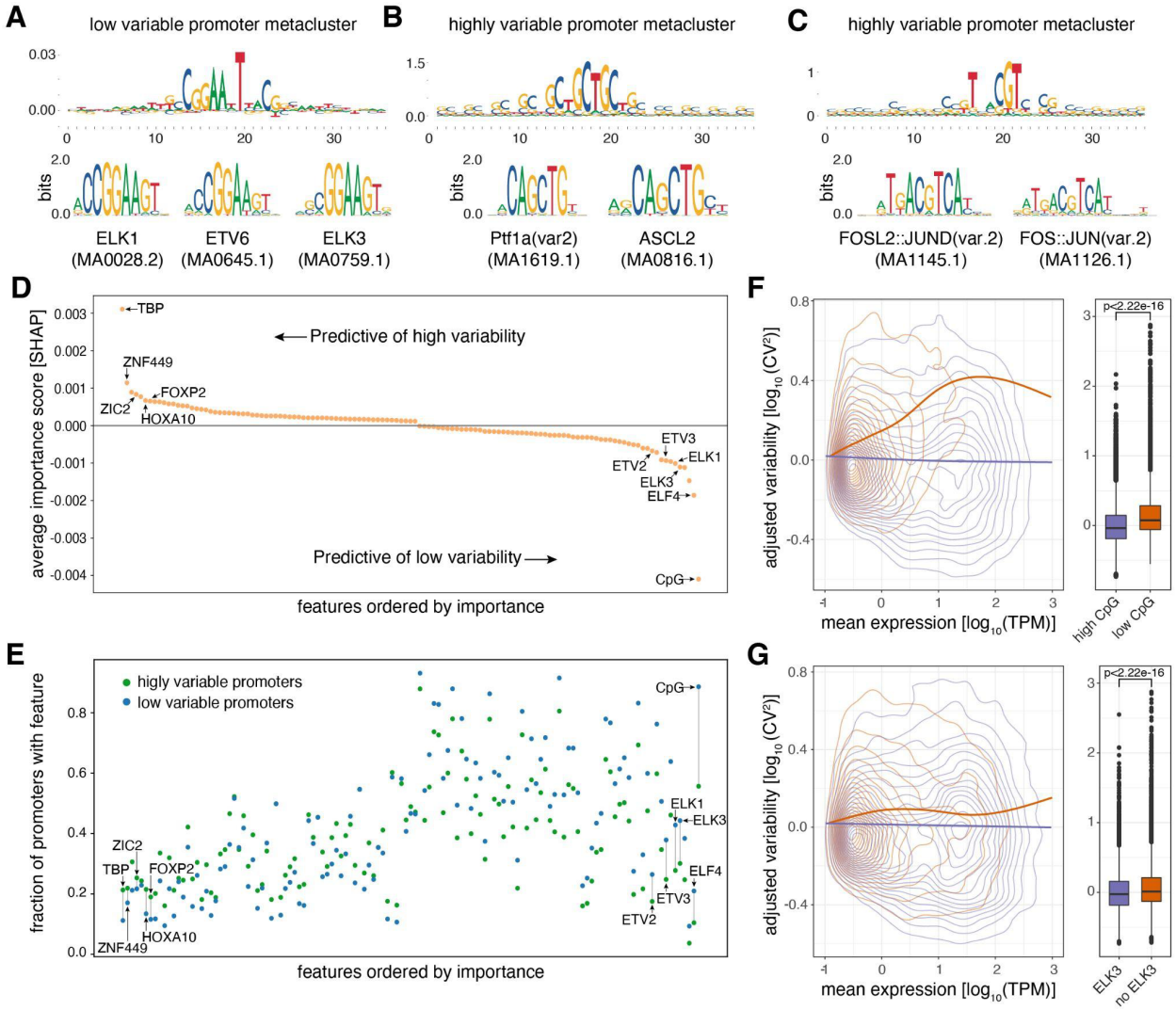

Figure 2: Promoter sequence features are highly predictive of promoter variability. A: Sequence logo of a metacluster (top) identified for low variable promoter sequences that matches known TF motifs (bottom) for ETS factors ELK1, ETV6, and ELK3. **B-C:** Sequence logos of two metaclusters (top) identified 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 CpG content and each of the 124 TFs identified as important for predicting promoter variability. Features 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 4A. **E:** The frequency of predicted TF binding sites (presence/absence) in highly variable (green) and low variable (blue) promoters. TFs are ordered as in D. For a full version of the plot see Supplementary Figure 4B,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 adjusted $\log_{10}$-transformed CV^2 with loess regression lines shown separately for each promoter group. The right subpanels display box-and-whisker plots of the differences in adjusted $\log_{10}$-transformed CV^2 between the two promoter groups (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).

Reverse engineering of the random forest classifier (SHAP, Shapley additive explanations) (Lundberg and Lee, 2017) allowed us to calculate the marginal contribution of each of the 125 selected features 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. 4A). 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. Although the remaining TFs contribute only marginally on their own compared to TATA-box and high CpG content status (Fig. 2D; Supplementary Fig. 4A), a baseline model (decision tree) based on CpG ratio and TBP binding site presence alone yielded worse performance than the full model (AUC=0.71 versus 0.79 for the baseline model and the full model, respectively; Supplementary Fig. 3B). This demonstrates that the TF binding grammar contributes to a promoter's expression variability. Interestingly, TFs associated with highly variable promoters are mostly related to 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. 4B,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. 5). 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. 4A), 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., 2011, 2007). We observed that the motifs of individual ETS family members are independently strong predictors of low promoter variability (Fig. 2D; Supplementary Fig. 4A) 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. 6A), regardless of promoter expression level (Supplementary Fig. 6B), suggesting that multiple ETS binding sites can either facilitate cooperativity between ETS factors or provide robustness 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 signaling, response to stimulus, and developmental processes.

Interestingly, the same features found to be predictive of low and high promoter variability across individuals, including CpG-content and TATA-boxes (TBP binding sites), are also associated with low and high transcriptional noise across individual cells (Faure et al., 2017; Morgan and Marioni, 2018). In addition, 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.

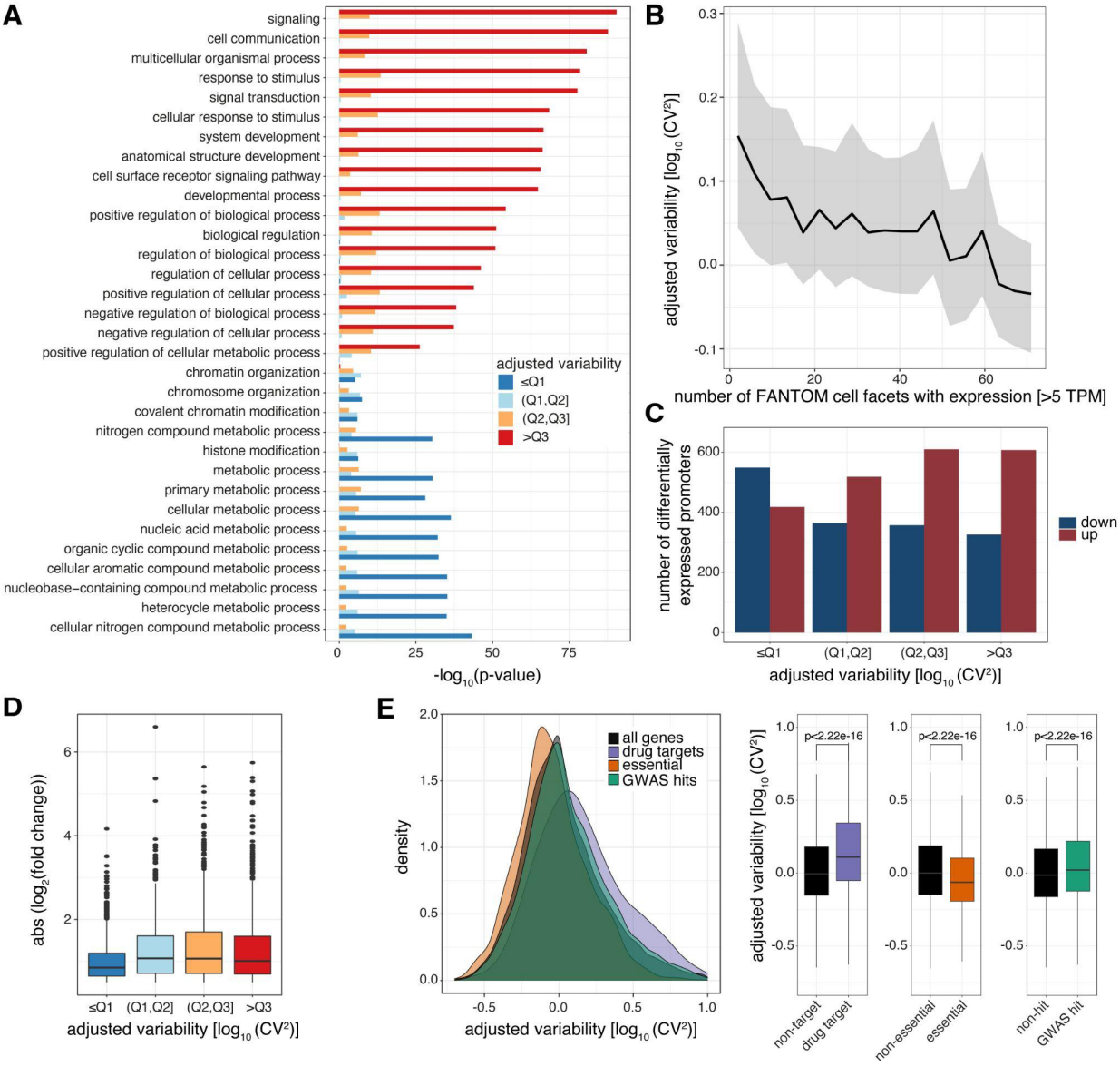

Figure 3: Levels of promoter variability are reflective of distinct biological processes and a selective trade off between robustness and plasticity. A: GO term enrichment, for biological processes, of genes split by associated promoter variability quartiles (Q1, Q2, Q3). Top 10 GO terms of all groups are displayed and ranked based on p-values of the >Q3 variability 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 >5 TPM). **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.

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. 7A; Supplementary Table 3). 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. 7B), demonstrating that highly variable promoters are more cell-type and tissue specific in their expression.

The restricted expression (Supplementary Fig. 7B) and biological processes (Fig. 3A) of genes associated with highly variable promoters and their TF grammar (Fig. 2D,E) led us to hypothesize that these are more prone to respond to external stimuli. Tumor necrosis factor (TNFα) induces an acute and time-limited gene response to NFkB signaling (Nelson et al., 2004; Turner et al., 2010), with negligible impact on chromatin topology (Jin et al., 2013), and is therefore suitable to study gene responsiveness. We profiled GM12878 TSSs and promoter activities with CAGE before and after 6 hours treatment with TNFα (Supplementary Table 4). This revealed an enrichment of up-regulated promoters among highly variable promoters (odds ratio (OR)=1.529, $p=4.563\times10^{-8}$, Fisher's exact test) as posited, while low variable promoters were mostly unaffected or down-regulated (OR=0.459, $p<2.2\times10^{-16}$, Fisher's exact test; Fig. 3C). In addition, low variable promoters had a weaker response (Fig. 3D).

Interestingly, 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. 7C).

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; Carninci et al., 2006; Lehner, 2008). 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.

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 (Supplementary Fig. 8A). To simultaneously capture the spread of TSSs and their relative frequencies compared to total RNA expression within a promoter, we considered a width-normalized Shannon entropy as a measure of TSS positional dispersion (Hoskins et al., 2011). This measure will discern promoters whose relative TSS expression is concentrated to a small subset of their widths (low entropy) from those with a more even spread (high entropy). We observed that low variable promoters are associated with a higher entropy than promoters with high variability (Fig. 4A). Consistently, low variable promoters tend to have more TSSs substantially contributing to their overall expression across individuals (Supplementary Fig. 8B). 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 signals of promoters (Supplementary Table 5) demonstrated that higher entropy reflects an increased number of decomposed promoters, as indicated by their number of local maxima of CAGE signals (Fig. 4B).

Interestingly, the decomposed promoters of gene *UFSP2* (Fig. 4C,E) 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. In contrast, the decomposed promoters of gene RIT1 (Fig. 4D,F) contribute equally to the overall activity of the encompassing promoter across individuals. To assess in general how individual decomposed promoters influence the overall promoter variability, we calculated the expression-adjusted dispersion (adjusted $\log_{10}$-transformed CV^2) of local-maxima decomposed promoters. Remarkably, many of the decomposed promoters showed a vastly different variability across individuals compared to the promoters they originate from (Supplementary Fig. 8C). 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 highly multimodal 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.

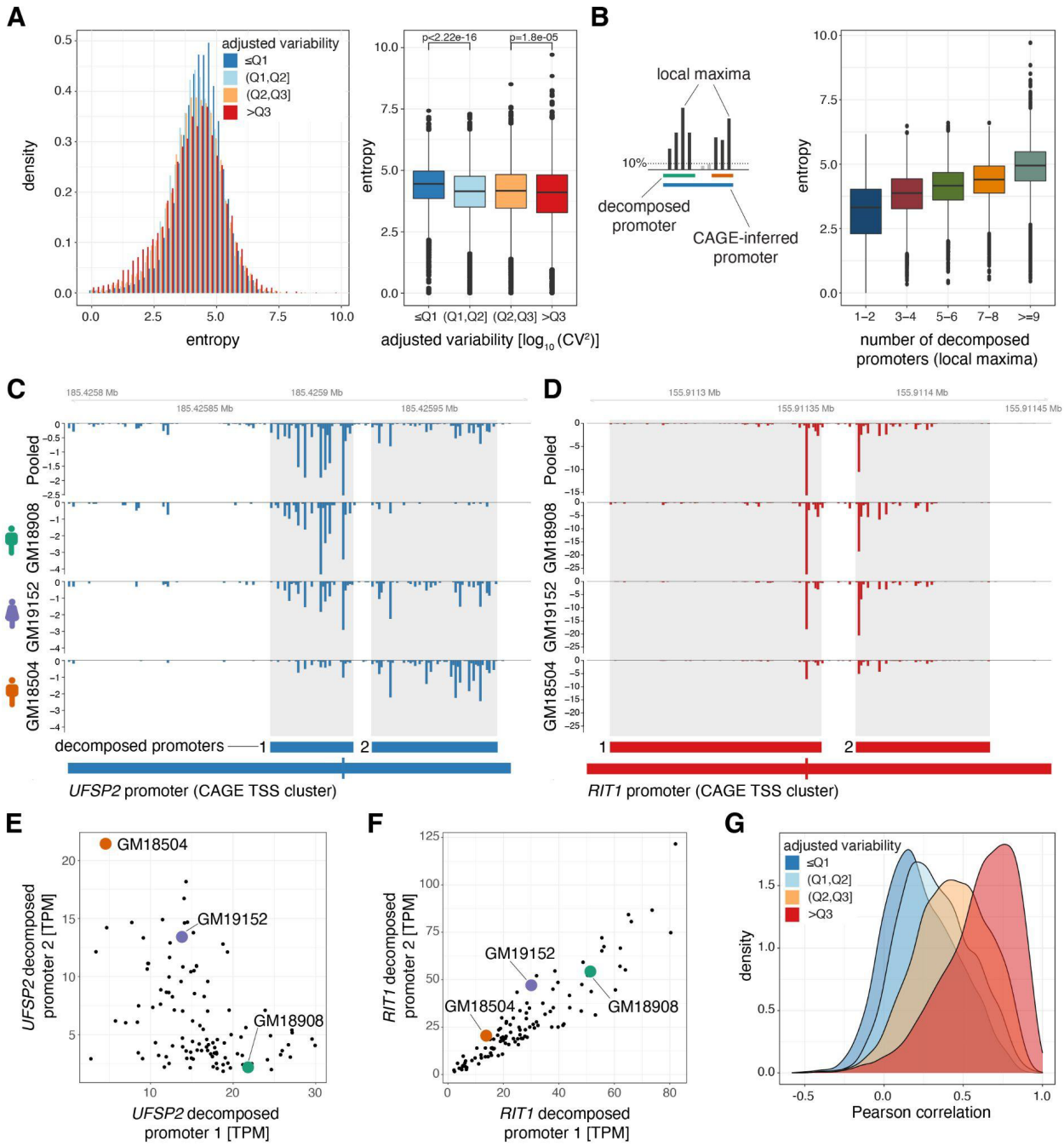

Figure 4: Low variable promoters exhibit flexibility in transcription initiation architecture. A: Promoter shape entropy for promoters split by variability quartiles, displayed as densities (left) and in a box-and-whisker plot (right). **B:** Illustration of the local maxima decomposition approach (left; see Materials and Methods) and box-and-whisker plot displaying the relationship between the Shannon entropy and the number of local maxima-inferred decomposed promoters. **C-D:** Examples of two promoters each containing two decomposed promoters exhibiting low correlation across individuals (panel C, gene UFSP2) and high correlation across individuals (panel D, gene RIT1). Both panels display genome tracks of average, TPM-normalized CAGE-inferred TSS expression levels across the panel (Pooled, top track) and for three individuals (GM18908, GM19152, GM18504, lower tracks). Below the genome tracks, the original promoter and resulting decomposed promoters (shaded in genome tracks) are shown. **E-F:** Relationship between TPM-normalized CAGE expression of

decomposed promoter 1 (x-axis) and 2 (y-axis) across all 108 LCLs for example genes *UFSP2* (E) and *RIT1* (F). The expression values for individuals included in panels B and C are highlighted. **G:** Densities of the lowest Pearson correlation between all pairs of decomposed promoters originating from the same promoter across all CAGE-inferred promoters with at least two decomposed promoters. For all box-and-whisker plots, central band: median; boundaries: first and third quartiles; whiskers: +/- 1.5 IQR.

If the effects of significant changes in TSS usage across the panel are masked by compensatory changes in alternative TSS usage within the same promoter, this would be revealed by low or even negative expression correlation between decomposed promoters (e.g., decomposed promoters 1 and 2 of *UFSP2,* Fig. 4C,E). Remarkably, we observed a strong association between promoter variability and the minimal expression correlation between decomposed promoter pairs within a promoter (Fig. 4G). Strikingly, low variable promoters are associated with weakly or even negatively correlated pairs of decomposed promoters. In stark contrast, highly variable promoters are associated with moderately or highly correlated pairs of decomposed promoters. This demonstrates that decomposed promoters may operate independently of each other, and that their alternative usages confer stability to the overall promoter expression. The association between decomposed promoter correlation and overall promoter stability was maintained when all decomposed promoter pairs were considered (Supplementary Fig. 8D), and could not be explained by CpG island status (Supplementary Fig. 8E). However, decomposed promoter expression correlation was associated with promoter width (Supplementary Fig. 8F), demonstrating that complex promoters with multiple independent core promoters (decomposed promoters) require larger promoter width, while broad promoter width does not necessarily lead to lower promoter variability.

Taken together, our results demonstrate that flexible and independent core promoters within broad promoters provide stability to the overall expression of a large subset of gene promoters with low variability.

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. In addition, the rare and limited effects of eQTLs on housekeeping genes might be due to mechanisms promoting mutational robustness. Our results (Fig. 4) indicate that a flexible TSS architecture within a promoter may provide such a mechanism and thereby mask the effects genetic variants may have on individual decomposed promoters.

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 6). 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. 9A). This indicates that, in addition to having fewer proximal genetic variants, low variable promoters are less likely to have prQTLs with large regulatory effects.

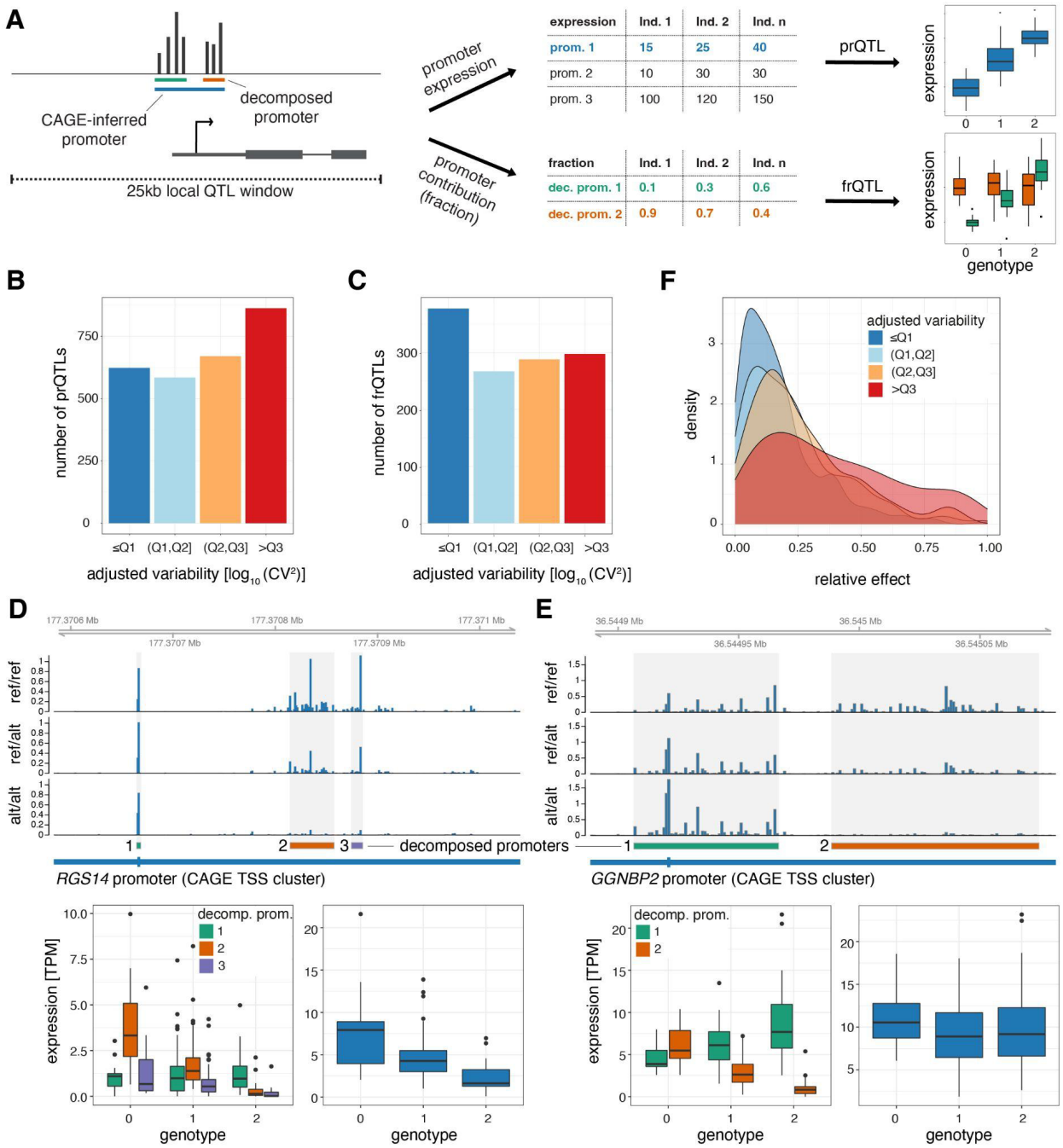

Figure 5: Plasticity in TSS usage promotes mutational robustness. A: Illustration of the strategy for testing the effects of genetic variants on promoter expression (prQTLs, TPM-normalized CAGE counts) and on decomposed promoter contribution to the encompassing promoter expression (frQTLs, ratios of TPM-normalized CAGE counts between decomposed and encompassing promoters). For both approaches only SNPs within 25kb of the promoter CAGE signal 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 7). Unlike the prQTLs, the frQTLs were more commonly associated with 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. 5C), 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. 5F). 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. This association is further maintained at the gene level (adjusted RNA-seq (Lappalainen et al., 2013) CV^2; Supplementary Fig. 9B; Supplementary Table 8), demonstrating that individual differences in decomposed promoter 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 9). 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 common genetic variants can influence TSS usage in humans 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. The prevalence of independent decomposed promoters within low variable promoters, as suggested by low pairwise correlation, indicates an extensive regulatory role of TSS plasticity in attenuating expression variability of essential genes.

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 model based on TF binding site sequence and CpG island status was able to predict promoter variability well (AUC=0.79 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 additional information influencing variability may be encoded within the sequence of TSS-proximal regions. For instance, di- or tri-nucleotide sequence patterns and stretches of high AT-richness, which influence local nucleosome positioning (Segal et al., 2006), impose different requirements for chromatin remodeling 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 (Faure et al., 2017; Morgan and Marioni, 2018). 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 is expected 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 clusters of TSSs (decomposed promoters) within a larger promoter is fundamentally different from switches between alternative promoters (Garieri et al., 2017; Valen et al., 2008; Zhang 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). Interestingly, ETS factors, here associated with low variable promoters, can interact with transcriptional co-activators and chromatin modifying complexes (Göös et al., 2022), and may therefore play a role in TSS selection in promoters with multi-modal architectures (Lam et al., 2019). 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, the presence of weak or negatively correlated expression patterns between decomposed promoters for a large number of promoters suggests that such buffering events will be revealed for more genes with an increased sample size.

A flexibility in TSS usage 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.

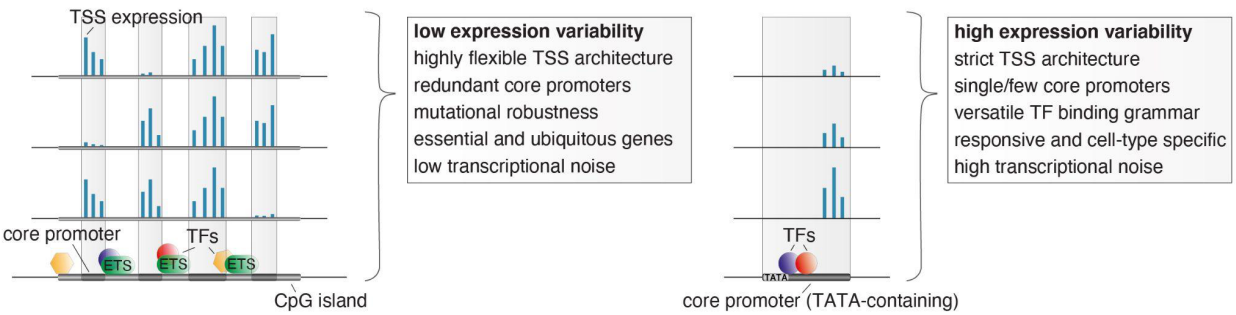

Figure 6: Unifying mechanisms influencing the variability in expression across individuals, the specificity in expression across cell types, and the stochasticity in expression across individual cells. Low variable promoters (left) frequently associate with high CpG content (CpG islands), multiple binding 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.

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 (Finucane et al., 2015; Kundaje et al., 2015; Timshel et al., 2020).

Materials and methods

LCL cell culturing

Epstein-Barr virus immortalized LCLs (Supplementary Table 1) 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. All 108 LCLs were grown unsynchronized for 5-7 passages and harvested when they reached 20+ million cells. As these cell lines were freshly purchased, mycoplasma contamination screening was not undertaken.

CAGE library preparation, sequencing, and mapping

CAGE libraries were prepared in 10 batches in total as described elsewhere (Andersson et al., 2014b; Takahashi et al., 2012) from 1,500 ng total RNA from each LCL. 23 libraries (Supplementary Table 1) underwent a second round of size selection (Invitrogen E-Gel) to remove excessive primer dimers. The libraries were quality checked using an Agilent 2100 Bioanalyzer system with an RNA pico chips kit (Agilent) and quantified using DNA 1000 chips kit (Agilent). Pooled libraries (Supplementary Table 1) were sequenced with spiked-in PhiX on an Illumina HiSeq 2500 machine single-end for 50 cycles using v4 sequencing chemistry (Illumina Inc.) and a custom sequencing primer (Takahashi et al., 2012). Libraries were split by barcode and reads were trimmed to remove linker sequences and filtered for a minimum sequence quality of Q30 in 50% of base pairs using the FASTX-Toolkit. rRNA reads matching subsequences of the human ribosomal DNA complete repeating unit (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 a maximum edit distance of 2. To reduce mapping bias, reads were re-mapped using the WASP pipeline (van de

Geijn et al., 2015) and BWA, taking into account biallelic SNVs (Lowy-Gallego et al., 2019). Reads with a mapping quality of 20 were retained for further analyses. Sample-related information, including CAGE run batch ID and E-gel information, are provided in Supplementary Table 1.

CAGE tag clustering, filtering, and quantification

CAGE-defined transcription start sites (CTSSs) were identified from 5' ends of mapped CAGE reads for each strand separately. The expression of CTSSs for each LCL was quantified from the number of CAGE reads sharing 5' ends using CAGEfightR (version 1.10) (Thodberg et al., 2019). To identify broad promoters that could potentially encompass multiple alternative core promoters, we performed lenient positional clustering of CTSSs (tag clustering) (Carninci et al., 2006), grouping CTSSs on the same strand within 60 bp of each other. To exclude rare promoters within the panel, tag clustering was performed on CTSSs with at least 1 CAGE read in at least 5 LCLs. The expression of each tag cluster (CAGE-inferred promoter) in each individual LCL was quantified by aggregating the expression of all CTSSs falling within the defined tag cluster region. To allow capture of flexible TSS usage within promoters across the panel, no support filtering was performed at CTSS level for expression quantification. Expression levels were converted to tags per million (TPM), by normalizing the expression count of each tag cluster in each library as a fraction of its number of mapped CAGE reads, scaled by 10^6. Tag clusters were filtered to be proximal to GENCODE-annotated TSSs (hg38, version 29, within 1000bp upstream) and to have at least 10 read counts in more than 10 LCLs. The resulting 29,001 gene-associated CAGE-inferred promoters were later decomposed by local maxima decomposition to split multi-modal tag clusters (https://github.com/anderssonlab/CAGEfightR_extensions, version 0.1.1). First, for each CAGE-inferred promoter, local maxima of within-promoter CTSSs with the highest pooled expression separated by at least 20 bp were identified. Second, decomposition was performed for each local maxima separately in decreasing order of pooled expression level. For each local maxima, the fraction between the pooled expression of each CTSS to that of the local maxima was calculated. All CTSSs associated with at least 10% of the local maxima signal that were not gapped by more than 10 bp with CTSS expression less than this value were retained in a new

decomposed promoter. For smoothing purposes, neighboring non-zero CTSSs within 1 bp distance of CTSSs fulfilling the fraction criterion were also included. Subsequently, decomposed promoters were merged if positioned within 1 bp from each other.

Geuvadis YRI 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., 2013) 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 expression variability calculation.

GM12878 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 sequencing depth deviating less than 2.5 times the standard deviation from the average sequencing depth across cells 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 promoter (or decomposed promoter) expression across the LCL panel and subsequently $\log_{10}$-transformed. Adjustment of the mean expression-dispersion relationship was performed by subtracting the expected $\log_{10}$-transformed dispersion for each promoter according to its expression level, using a running median (width 50, step size 25) of raw dispersions ($\log_{10} CV^2$) ordered by mean expression level (TPM) across the panel, as described elsewhere (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. Promoters were grouped by variability according to the quartiles of expression-adjusted dispersions (≤Q1, (Q1, Q2], (Q2, Q3], >Q3).

Neural network model, training and hyperparameter tuning

A simple neural network architecture was designed to learn to predict low and high variability from DNA sequence. The neural network model uses as input one-hot-encoded DNA sequences (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]) from the human reference genome (hg38) to predict low and highly variable promoter activity as output. Although CAGE-inferred promoters varied in width, we made use of fixed-length 600 bp sequences for each promoter centered on its pooled CAGE summit CTSS. 600 bp was used to make sure that sequences influencing promoter variability contained within regions that could cover a central open chromatin site (150-300bp) as well as within flanking nucleosomal DNA (150-200bp) were captured, where also most of the expression output of a promoter originate (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014).

The model (Supplementary Fig. 2A) 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 activation function (Agarap, 2019) 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 adjusted dispersion).

Promoter sequences from chromosomes 2 and 3 were used as the test set and those 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 $\log_{10}$-transformed 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. 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 and decision tree classifiers (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 for each nucleotide in the 600bp input sequences for low and highly variable promoters. DeepLIFT relies on backpropagation of the contributions of all neurons in the neural network to the input features, nucleotides, and was used to estimate the importance of each position and nucleotide in the input sequences to predict high and low variability. The resulting importance scores were supplied to TF-MoDISco (Transcription Factor Motif Discovery from Importance Scores) (Shrikumar et al., 2020) to identify DNA stretches (seqlets) with high importance for the predictions. DeepLIFT and TF-MoDISco were run independently on the input sequences for 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 an 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) to match the resulting metaclusters to known TF motifs (in MEME format) from the JASPAR database (release 2020, hg38) (Fornes et al., 2020). We compared each non-redundant JASPAR vertebrate frequency matrix with the metaclusters 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

To identify broad-scale trends of high CpG content, we calculated CpG observed/expected ratio in windows +/- 500bp around the pooled summit CTSS of

each promoter. Calculated CpG ratios revealed a bimodal distribution that informed on thresholding high CpG content promoters as those with CpG observed/expected ratio >0.5 (Supplementary Fig. 3A).

Predicted transcription factor binding sites for 746 TFs with scores of 500 or greater ($P < 10^{-5}$) (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 modeling. 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).

As with 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 TF features, only 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. SHAP calculates Shapley values, a game theoretic approach for optimal credit allocation during cooperation, which can be used to estimate the marginal contribution of each feature to a model's predictions.

Decision tree baseline model

To evaluate the contribution of TF binding site presence for predicting promoter variability, we trained a baseline model based on CpG content status and TATA-box

presence only. CpG content status (CpG observed/expected ratio > 0.5) and the presence/absence of predicted TBP binding sites were used as features for predicting high and low variability using a decision tree classifier. The decision tree model was implemented and trained in Scikit-learn (version 2.3.0) using default parameters. Training and test data were defined as for the CNN and Random Forest models.

Tissue-, cell-type specificity and gene annotations

RNA-seq 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.

CAGE 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. Unstimulated GM12878 cells and those stimulated with 25ng/ml TNF-α for 6 hours were harvested with four replicates for each condition. Cell culturing, CAGE library preparation and mapping were 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 standard library size adjustment

and a generalized linear model with DESeq2 (version 1.30.1) (Love et al., 2014). Promoters with FDR-adjusted p-value ≤ 0.05 were considered to be differentially expressed.

Correlation analysis of decomposed promoter expression

To test if decomposed promoters could act independently of each other, we calculated Pearson correlation coefficients of LCL expression between pairs of decomposed promoters originating from the same promoter. We focused on promoters with at least 2 decomposed promoters significantly contributing to the overall expression of the promoter. Specifically, we considered decomposed promoters whose expression accounted for at least 5% of the overall promoter expression in at least half of all LCLs, resulting in 37,663 decomposed promoters of 14,889 promoters. To avoid potential bias introduced from a variable number of decomposed promoters per promoter, we considered the lowest correlation across decomposed promoter pairs within a promoter.

Mapping QTLs

We tested both the association between the genotypes of common genetic variants (MAF ≥ 10%) and the expression of promoters (promoter eQTL, prQTLs) as well as their association with the contribution of decomposed promoters to the overall expression of the encompassing (non-decomposed) promoter (fraction eQTL, frQTL). 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 (Supplementary Table 1) 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 expression 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 considered only decomposed promoters whose expression accounted for at least 5% of the overall promoter expression in at

least half of all LCLs and promoters with at least 2 such decomposed promoters, resulting in 37,663 decomposed promoters of 14,889 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 the 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. 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%.

Code availability

Code for data analysis performed in this study is publicly available on GitHub: https://github.com/anderssonlab/Einarsson_et_al_2022/.

Data availability

GM12878 scRNA-seq data were retrieved from GEO (accession number GSE126321). Processed Geuvadis RNA-seq gene expression data were retrieved from recount2 (Collado-Torres et al., 2017) (accession number ERP001942). Processed GTEx RNA-seq gene expression data were retrieved from the GTEx portal (https://www.gtexportal.org/home/datasets, version-8). Predicted transcription factor binding sites for 746 TFs were obtained from JASPAR 2020 (http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2020/hg38/).

Acknowledgements

We thank members of the Andersson lab for rewarding discussions. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala, 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.

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 [grants NNF18OC0052570, NNF20OC0059796].

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, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *ArXiv160304467 Cs*.

Agarap AF. 2019. Deep Learning using Rectified Linear Units (ReLU). *ArXiv180308375 Cs Stat*.

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. doi:10.1186/gb-2009-10-4-r38

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Müller F, Consortium TF, Forrest ARR, Carninci P, Rehli M, Sandelin A. 2014a. An atlas of active enhancers across human cell types and tissues. *Nature* **507**:455–461. doi:10.1038/nature12787

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. doi:10.1038/ncomms6336

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Yingrui, Liu S, Liu Xiao, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter

I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Grocock R, Humphray S, James T, Kingsbury Z, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo ML, Fulton L, Ananiev V, Belaia Z, Beloslyudtsev D, Bouk N, Chen C, Church D, Cohen R, Cook C, Garner J, Hefferon T, Kimelman M, Liu C, Lopez J, Meric P, O'Sullivan C, Ostapchuk Y, Phan L, Ponomarov S, Schneider V, Shekhtman E, Sirotkin K, Slotta D, Zhang H, Balasubramaniam S, Burton J, Danecek P, Keane TM, Kolb-Kokocinski A, McCarthy S, Stalker J, Quail M, Davies CJ, Gollub J, Webster T, Wong B, Zhan Y, Campbell CL, Kong Y, Marcketta A, Yu F, Antunes L, Bainbridge M, Sabo A, Huang Z, Coin LJM, Fang L, Li Q, Li Z, Lin H, Liu B, Luo R, Shao H, Xie Y, Ye C, Yu C, Zhang F, Zheng H, Zhu H, Alkan C, Dal E, Kahveci F, Garrison EP, Kural D, Lee WP, Leong WF, Stromberg M, Ward AN, Wu J, Zhang M, Daly MJ, DePristo MA, Handsaker RE, Banks E, Bhatia G, Del Angel G, Genovese G, Li H, Kashin S, McCarroll SA, Nemesh JC, Poplin RE, Yoon SC, Lihm J, Makarov V, Gottipati S, Keinan A, Rodriguez-Flores JL, Rausch T, Fritz MH, Stütz AM, Beal K, Datta A, Herrero J, Ritchie GRS, Zerbino D, Sabeti PC, Shlyakhter I, Schaffner SF, Vitti J, Cooper DN, Ball EV, Stenson PD, Barnes B, Bauer M, Cheetham RK, Cox A, Eberle M, Kahn S, Murray L, Peden J, Shaw R, Kenny EE, Batzer MA, Konkel MK, Walker JA, MacArthur DG, Lek M, Herwig R, Ding L, Koboldt DC, Larson D, Ye Kai, Gravel S, Swaroop A, Chew E, Lappalainen T, Erlich Y, Gymrek M, Willems TF, Simpson JT, Shriver MD, Rosenfeld JA, Bustamante CD, Montgomery SB, De La Vega FM, Byrnes JK, Carroll AW, DeGorter MK, Lacroute P, Maples BK, Martin AR, Moreno-Estrada A, Shringarpure SS, Zakharia F, Halperin E, Baran Y, Cerveira E, Hwang J, Malhotra A, Plewczynski D, Radew K, Romanovitch M, Zhang C, Hyland FCL, Craig DW, Christoforides A, Homer N, Izatt T, Kurdoglu AA, Sinari SA, Squire K, Xiao C, Sebat J, Antaki D, Gujral M, Noor A, Ye Kenny, Burchard EG, Hernandez RD, Gignoux CR, Haussler D, Katzman SJ, Kent WJ, Howie B, Ruiz-Linares A, Dermitzakis ET, Devine SE, Kang HM, Kidd JM, Blackwell T, Caron S, Chen W, Emery S, Fritsche L, Fuchsberger C, Jun G, Li B, Lyons R, Scheller C, Sidore C, Song S, Sliwerska E, Taliun D, Tan A, Welch R, Wing MK, Zhan X, Awadalla P, Hodgkinson A, Li Yun, Shi X, Quitadamo A, Lunter G, Marchini JL, Myers S, Churchhouse C, Delaneau O, Gupta-Hinch A, Kretzschmar W, Iqbal Z, Mathieson I, Menelaou A, Rimmer A, Xifara DK, Oleksyk TK, Fu Yunxin, Liu Xiaoming, Xiong M, Jorde L, Witherspoon D, Xing J, Browning BL, Browning SR, Hormozdiari F, Sudmant PH, Khurana E, Tyler-Smith C, Albers CA, Ayub Q, Chen Y, Colonna V, Jostins L, Walter K, Xue Y, Gerstein MB, Abyzov A, Balasubramanian S, Chen J, Clarke D, Fu Yao, Harmanci AO, Jin M, Lee D, Liu J, Mu XJ, Zhang J, Zhang Yan, Hartl C, Shakir K, Degenhardt J, Meiers S, Raeder B, Casale FP, Stegle O, Lameijer EW, Hall I, Bafna V, Michaelson J, Gardner EJ, Mills RE, Dayama G, Chen K, Fan X, Chong Z, Chen T, Chaisson MJ, Huddleston J, Malig M, Nelson BJ, Parrish NF, Blackburne B, Lindsay SJ, Ning Z, Zhang Yujun, Lam H, Sisu C, Challis D, Evani US, Lu J, Nagaswamy U, Yu J, Li W, Habegger L, Yu H, Cunningham F, Dunham I, Lage K, Jespersen JB, Horn H, Kim D, Desalle R, Narechania A, Sayres MAW, Mendez FL, Poznik GD, Underhill PA, Mittelman D, Banerjee R, Cerezo M, Fitzgerald TW, Louzada S, Massaia A, Yang F, Kalra D, Hale W, Dan X, Barnes KC, Beiswanger C, Cai H, Cao H, Henn B, Jones D, Kaye JS, Kent A, Kerasidou A, Mathias R, Ossorio PN, Parker M, Rotimi CN, Royal CD, Sandoval K, Su Y, Tian Z, Tishkoff S, Via M, Wang Y, Yang H, Yang L, Zhu J, Bodmer W, Bedoya G, Cai Z, Gao Y, Chu J, Peltonen L, Garcia-Montero A, Orfao A, Dutil J, Martinez-Cruzado JC, Mathias RA, Hennis A, Watson H, McKenzie C, Qadri F, LaRocque R, Deng X, Asogun D, Folarin O, Happi C, Omoniwa O, Stremlau M, Tariyal R, Jallow M, Joof FS, Corrah T, Rockett K, Kwiatkowski D, Kooner J, Hien TT, Dunstan SJ, ThuyHang N, Fonnie R, Garry R, Kanneh L, Moses L, Schieffelin J, Grant DS, Gallo C, Poletti G, Saleheen D, Rasheed A, Brooks LD, Felsenfeld AL, McEwen JE, Vaydylevich Y, Duncanson A, Dunn M, Schloss JA. 2015. A global reference for human genetic variation. *Nature*

526:68–74. doi:10.1038/nature15393
Bartha I, di Iulio J, Venter JC, Telenti A. 2018. Human gene essentiality. *Nat Rev Genet* **19**:51–62. doi:10.1038/nrg.2017.75
Boettiger AN, Levine M. 2009. Synchronous and stochastic patterns of gene activation in the Drosophila embryo. *Science* **325**:471–473. doi:10.1126/science.1173976
Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**:626–635. doi:10.1038/ng1789
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. doi:10.1038/nbt.3838
Eldar A, Elowitz MB. 2010. Functional roles for noise in genetic circuits. *Nature* **467**:167–173. doi:10.1038/nature09326
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. doi:10.1016/j.cels.2018.06.011
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, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple CA, Ishizu Y, Young RS, Francescatto M, Alam I, Albanese D, Altschuler GM, Arakawa T, Archer JAC, Arner P, Babina M, Rennie S, Balwierz PJ, Beckhouse AG, Pradhan-Bhatt S, Blake JA, Blumenthal A, Bodega B, Bonetti A, Briggs J, Brombacher F, Burroughs AM, Califano A, Cannistraci CV, Carbajo D, Chen Y, Chierici M, Ciani Y, Clevers HC, Dalla E, Davis CA, Detmar M, Diehl AD, Dohi T, Drablos F, Edge ASB, Edinger M, Ekwall K, Endoh M, Enomoto H, Fagiolini M, Fairbairn L, Fang H, Farach-Carson MC, Faulkner GJ, Favorov AV, Fisher ME, Frith MC, Fujita R, Fukuda S, Furlanello C, Furino M, Furusawa J, Geijtenbeek TB, Gibson AP, Gingeras T, Goldowitz D, Gough J, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M, Hara M, Harbers M, Harshbarger J, Hasegawa A, Hasegawa Y, Hashimoto T, Herlyn M, Hitchens KJ, Ho Sui SJ, Hofmann OM, Hoof I, Hori F, Huminiecki L, Iida K, Ikawa T, Jankovic BR, Jia H, Joshi A, Jurman G, Kaczkowski B, Kai C, Kaida K, Kaiho A, Kajiyama K, Kanamori-Katayama M, Kasianov AS, Kasukawa T, Katayama S, Kato S, Kawaguchi S, Kawamoto H, Kawamura YI, Kawashima T, Kempfle JS, Kenna TJ, Kere J, Khachigian LM, Kitamura T, Klinken SP, Knox AJ, Kojima M, Kojima S, Kondo N, Koseki H, Koyasu S, Krampitz S, Kubosaki A, Kwon AT, Laros JFJ, Lee W, Lennartsson A, Li K, Lilje B, Lipovich L, Mackay-Sim A, Manabe R, Mar JC, Marchand B, Mathelier A, Mejhert N, Meynert A, Mizuno Y, de Lima Morais DA, Morikawa H, Morimoto M, Moro K, Motakis E, Motohashi H, Mummery CL, Murata M, Nagao-Sato S, Nakachi Y, Nakahara F, Nakamura T, Nakamura Y, Nakazato K, van Nimwegen E, Ninomiya N, Nishiyori H, Noma S, Noazaki T, Ogishima S, Ohkura N, Ohimiya H, Ohno H, Ohshima M, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov DA, Pain A, Passier R, Patrikakis M, Persson H, Piazza S, Prendergast JGD, Rackham OJL, Ramilowski JA, Rashid M, Ravasi T, Rizzu P, Roncador M, Roy S, Rye MB, Saijyo E, Sajantila A, Saka A, Sakaguchi S, Sakai M, Sato H, Savvi S, Saxena A, Schneider C, Schultes EA, Schulze-Tanzil GG, Schwegmann A, Sengstag T, Sheng G, Shimoji H, Shimoni Y, Shin JW, Simon C, Sugiyama D, Sugiyama T, Suzuki M, Suzuki N, Swoboda RK, ’t Hoen PAC, Tagami M, Takahashi N, Takai J, Tanaka H, Tatsukawa H, Tatum Z, Thompson M, Toyodo H, Toyoda T, Valen E, van de Wetering M, van den Berg LM,

Verado R, Vijayan D, Vorontsov IE, Wasserman WW, Watanabe S, Wells CA, Winteringham LN, Wolvetang E, Wood EJ, Yamaguchi Y, Yamamoto M, Yoneda M, Yonekura Y, Yoshida S, Zabierowski SE, Zhang PG, Zhao X, Zucchelli S, Summers KM, Suzuki H, Daub CO, Kawai J, Heutink P, Hide W, Freeman TC, Lenhard B, Bajic VB, Taylor MS, Makeev VJ, Sandelin A, Hume DA, Carninci P, Hayashizaki Y. 2014. A promoter-level mammalian expression atlas. *Nature* **507**:462–470. doi:10.1038/nature13182

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. doi:10.1016/j.cels.2017.10.003

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K, Ripke S, Day FR, Consortium R, Consortium SWG of the PG, Consortium TR, Purcell S, Stahl E, Lindstrom S, Perry JRB, Okada Y, Raychaudhuri S, Daly MJ, Patterson N, Neale BM, Price AL. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**:1228–1235. doi:10.1038/ng.3404

Fornes O, Castro-Mondragon JA, Khan A, Van Der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A. 2020. JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Res* **48**:D87–D92. doi:10.1093/nar/gkz1001

Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**:D766–D773. doi:10.1093/nar/gky955

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. doi:10.1038/s41467-017-01467-7

Göös H, Kinnunen M, Salokas K, Tan Z, Liu X, Yadav L, Zhang Q, Wei G-H, Varjosalo M. 2022. Human transcription factor protein interaction networks. *Nat Commun* **13**:766. doi:10.1038/s41467-022-28341-5

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**:204–213. doi:10.1038/nature24277

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8**:R24. doi:10.1186/gb-2007-8-2-r24

Haberle V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, Gehrig J, Dong X, Akalin A, Suzuki AM, van IJcken WFJ, Armant O, Ferg M, Strähle U, Carninci P, Müller F, Lenhard B. 2014. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507**:381–385. doi:10.1038/nature12974

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184**:3573-3587.e29. doi:10.1016/j.cell.2021.04.048

Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, Chandrashekhar M, Hustedt N, Seth S, Noonan A, Habsid A, Sizova O, Nedyalkova L, Climie R, Tworzyanski L, Lawson K, Sartori MA, Alibeh S, Tieu D, Masud S, Mero P, Weiss A, Brown KR, Usaj M, Billmann M, Rahman M, Constanzo M, Myers CL, Andrews BJ, Boone C, Durocher D, Moffat J. 2017. Evaluation and Design of Genome-Wide

CRISPR/SpCas9 Knockout Screens. *G3 Bethesda Md* **7**:2719–2727. doi:10.1534/g3.117.041277
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. doi:10.1101/2020.11.26.400218
Hollenhorst PC, McIntosh LP, Graves BJ. 2011. Genomic and Biochemical Insights into the Specificity of ETS Transcription Factors. *Annu Rev Biochem* **80**:437–471. doi:10.1146/annurev.biochem.79.081507.103945
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. doi:10.1101/gad.1561707
Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, Yang L, Boley N, Andrews J, Kaufman TC, Graveley BR, Bickel PJ, Carninci P, Carlson JW, Celniker SE. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* **21**:182–192. doi:10.1101/gr.112466.110
Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*. doi:10.1038/nature12644
Kar A, Gutierrez-Hartmann A. 2013. Molecular mechanisms of ETS transcription factor-mediated tumorigenesis. *Crit Rev Biochem Mol Biol* **48**:522–543. doi:10.3109/10409238.2013.838202
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. doi:10.1186/gb-2006-7-12-r118
Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, Shin JW, Kojima-Ishiyama M, Kawano M, Murata M, Ninomiya-Fukuda N, Ishikawa-Kato S, Nagao-Sato S, Noma S, Hayashizaki Y, Forrest ARR, Carninci P. 2014. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res* **24**:708–717. doi:10.1101/gr.156232.113
Keany E. 2020. BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. Zenodo. doi:10.5281/ZENODO.4247618
Kingma DP, Ba J. 2017. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs*.
Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, Marioni JC, Teichmann SA. 2015. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell* **17**:471–485. doi:10.1016/j.stem.2015.09.011
Kundaje A, Ernst J, Yen A, Heravi-Moussavi A, Zhang Z, Amin V, Schultz MD, Sarkar A, Wu Y-C, Pfenning A, Wang X, ClaussnitzerYaping Liu M, Alan Harris R, Shoresh N, Leung D, David Hawkins R, Hong C, Tam A, Scott Hansen R, Kaul R, Bansal MS, Feizi S, Karlic R, Kim A-R, Mercer TR, Neph SJ, Polak P, Ray P, Siebenthall KT, Sinnott-Armstrong NA, Thurman RE, Wu J, Zhang B, Zhou X, Abdennur N, Adli M, Akerman M, Barrera L, Antosiewicz-Bourget J, Ballinger T, Barnes MJ, Bates D, Bell RJA, Bennett DA, Bianco K, Bock C, Boyle P, Brinchmann J, Caballero-Campo P, Camahort R, Carrasco-Alfonso MJ, Charnecki T, Chen H, Chen Z, Cheng JB, Cho S, Chu A, Chung W-Y, Cowan C, Athena Deng Q, Deshpande V, Diegel M, Ding B, Durham T, Echipare L, Edsall L, Flowers D, Genbacev-Krtolica O, Gifford C, Gillespie S, Giste E, Glass IA, Gnirke A, Gormley M, Gu H, Gu J, Hafler DA, Hangauer MJ, Hariharan M, Hatan M, Haugen E, He Y, Heimfeld S, Herlofsen S, Hou Z, Humbert R, Issner R, Jackson AR, Jia H, Jiang P, Johnson AK, Kadlecek T, Kamoh B, Kapidzic M, Kent J, Kim A, Kleinewietfeld M, Klugman S, Krishnan J, Kuan S, Kutyavin T, Lee AY, Lee K, Li J, Li N, Li Y, Ligon KL, Lin S, Lin Y, Liu J, Liu Yuxuan, Luckey CJ, Ma YP, Maire C, Marson A, Mattick JS, Mayo M, McMaster M,

Metsky H, Mikkelsen T, Miller D, Miri M, Mukame E, Nagarajan RP, Neri F, Nery J, Nguyen T, O'Geen H, Paithankar S, Papayannopoulou T, Pelizzola M, Plettner P, Propson NE, Raghuraman S, Raney BJ, Raubitschek A, Reynolds AP, Richards H, Riehle K, Rinaudo P, Robinson JF, Rockweiler NB, Rosen E, Rynes E, Schein J, Sears R, Sejnowski T, Shafer A, Shen L, Shoemaker R, Sigaroudinia M, Slukvin I, Stehling-Sun S, Stewart R, Subramanian SL, Suknuntha K, Swanson S, Tian S, Tilden H, Tsai L, Urich M, Vaughn I, Vierstra J, Vong S, Wagner U, Wang H, Wang Tao, Wang Y, Weiss A, Whitton H, Wildberg A, Witt H, Won K-J, Xie M, Xing X, Xu I, Xuan Z, Ye Z, Yen C-A, Yu P, Zhang Xian, Zhang Xiaolan, Zhao J, Zhou Y, Zhu J, Zhu Y, Ziegler S, Boyer LA, Fisher SJ, Jones SJM, Sunyaev S, Tsai L-H, Wang W, Zhang MQ, Bernstein BE, Costello JF, Ecker JR, Meissner A, Milosavljevic A, Stamatoyannopoulos JA, Meuleman W, Bilenky M, Kheradpour P, Wang J, Ziller MJ, Whitaker JW, Ward LD, Quon G, Sandstrom RS, Eaton ML, Pfenning AR, Claussnitzer M, Liu Yaping, Coarfa C, Harris RA, Epstein CB, Gjoneska E, Xie W, Hawkins RD, Lister R, Gascard P, Mungall AJ, Moore R, Chuah E, Canfield TK, Hansen RS, Sabo PJ, Carles A, Dixon JR, Farh K-H, Kulkarni A, Li D, Lowdon R, Elliott G, Onuchic V, Rajagopal N, Sallari RC, Stevens M, Beaudet AE, De Jager PL, Farnham PJ, Haussler D, Li W, Marra MA, McManus MT, Thomson JA, Tlsty TD, Waterland RA, Chadwick LH, Hirst M, Ren B, Wang Ting, Kellis M. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**:317–330. doi:10.1038/nature14248

Kursa MB, Rudnicki WR. 2010. Feature Selection with the **Boruta** Package. *J Stat Softw* **36**. doi:10.18637/jss.v036.i11

Lam KC, Chung H-R, Semplicio G, Iyer SS, Gaub A, Bhardwaj V, Holz H, Georgiev P, Akhtar A. 2019. The NSL complex-mediated nucleosome landscape is required to maintain transcription fidelity and suppression of transcription noise. *Genes Dev* **33**:452–465. doi:10.1101/gad.321489.118

Lappalainen T, Sammeth M, Friedländer MR, 'T Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, Van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, Macarthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Lehrach H, Schreiber S, Sudbrak R, Carracedo Á, Antonarakis SE, Häsler R, Syvänen AC, Van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**:506–511. doi:10.1038/nature12531

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. doi:10.1101/gad.250704.114

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**:550. doi:10.1186/s13059-014-0550-8

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. doi:10.12688/wellcomeopenres.15126.2

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. doi:10.12688/f1000research.9501.2

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, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**:D896–D901. doi:10.1093/nar/gkw1133

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. doi:10.1038/nature08903

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.

Nelson DE, Ihekwaba AEC, Elliott M, Johnson JR, Gibney CA, Foreman BE, Nelson G, See V, Horton CA, Spiller DG, Edwards SW, McDowell HP, Unitt JF, Sullivan E, Grimley R, Benson N, Broomhead D, Kell DB, White MRH. 2004. Oscillations in NF-κB Signaling Control the Dynamics of Gene Expression. *Science* **306**:704–708. doi:10.1126/science.1099962

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. doi:10.1038/nature04785

Oikawa T, Yamada T. 2003. Molecular biology of the Ets family of transcription factors. *Gene* **303**:11–34. doi:10.1016/S0378-1119(02)01156-3

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. doi:10.1038/s41597-019-0116-4

Payne JL, Wagner A. 2015. Mechanisms of mutational robustness in transcriptional regulation. *Front Genet* **6**. doi:10.3389/fgene.2015.00322

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 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. doi:10.1038/nature08872

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. doi:10.1038/ncomms10417

Sandelin A, Wasserman WW. 2004. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol* **338**:207–215. doi:10.1016/j.jmb.2004.02.048

Schor IE, Degner JF, Harnett D, Cannavò E, Casale FP, Shim H, Garfield DA, Birney E, Stephens M, Stegle O, Furlong EEM. 2017. Promoter shape varies across populations and affects promoter evolution and expression noise. *Nat Genet* **49**:550–558. doi:10.1038/ng.3791

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. doi:10.1038/nature04979

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinforma Oxf Engl* **28**:1353–1358. doi:10.1093/bioinformatics/bts163

Sharrocks AD. 2001. The ETS-domain transcription factor family. *Nat Rev Mol Cell Biol* **2**:827–837. doi:10.1038/35099076

Shrikumar A, Greenside P, Kundaje A. 2019. Learning Important Features Through Propagating Activation Differences. *ArXiv170402685 Cs*.

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*.

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. doi:10.15252/msb.20209539
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. doi:10.1186/gb-2012-13-9-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. doi:10.1086/512017
Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET. 2007. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. *Science* **315**:848–853. doi:10.1126/science.1136678
Suico MA, Shuto T, Kai H. 2017. Roles and regulations of the ETS transcription factor ELF4/MEF. *J Mol Cell Biol* **9**:168–177. doi:10.1093/jmcb/mjw051
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. doi:10.1038/nprot.2012.005
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. doi:10.1186/s12859-019-3029-5
Timshel PN, Thompson JJ, Pers TH. 2020. Genetic mapping of etiologic brain cell types for obesity. *eLife* **9**:e55851. doi:10.7554/eLife.55851
Turner DA, Paszek P, Woodcock DJ, Nelson DE, Horton CA, Wang Y, Spiller DG, Rand DA, White MRH, Harper CV. 2010. Physiological levels of TNFα stimulation induce stochastic dynamics of NF-κB responses in single living cells. *J Cell Sci* **123**:2834–2843. doi:10.1242/jcs.069641
Urban EA, Johnston RJ. 2018. Buffering and Amplifying Transcriptional Noise During Cell Fate Specification. *Front Genet* **9**:591. doi:10.3389/fgene.2018.00591
Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, Marstrand TT, Tang MHE, Zhao X, Krogh A, Winther O, Arakawa T, Kawai J, Wells C, Daub C, Harbers M, Hayashizaki Y, Gustincich S, Sandelin A, Carninci P. 2008. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res* **19**:255–265. doi:10.1101/gr.084541.108
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. doi:10.1038/nmeth.3582
Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**:D1074–D1082. doi:10.1093/nar/gkx1037
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. doi:10.1186/s12864-017-3834-z