High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity

Kaia Mattioli,[1,2] Pieter-Jan Volders,[1,3,4] Chiara Gerhardinger,[1,5] James C. Lee,[1,6] Philipp G. Maass,[1,7,8] Marta Melé,[1,5,9,12] and John L. Rinn[1,5,10,11,12]

[1]*Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA;* [2]*Department of Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts 02115, USA;* [3]*Department of Biomolecular Medicine, Ghent University, 9000 Ghent, Belgium;* [4]*VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium;* [5]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA;* [6]*Department of Medicine, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Cambridge CB2 0QQ, United Kingdom;* [7]*Genetics and Genome Biology Program, Sickkids Research Institute, Toronto, Ontario M5G 0A4, Canada;* [8]*Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A1, Canada;* [9]*Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Catalonia 08034, Spain;* [10]*Department of Pathology, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02115, USA;* [11]*Department of Biochemistry, University of Colorado, BioFrontiers Institute, Boulder, Colorado 80301, USA*

Transcription initiates at both coding and noncoding genomic elements, including mRNA and long noncoding RNA (lncRNA) core promoters and enhancer RNAs (eRNAs). However, each class has a different expression profile with lncRNAs and eRNAs being the most tissue specific. How these complex differences in expression profiles and tissue specificities are encoded in a single DNA sequence remains unresolved. Here, we address this question using computational approaches and massively parallel reporter assays (MPRA) surveying hundreds of promoters and enhancers. We find that both divergent lncRNA and mRNA core promoters have higher capacities to drive transcription than nondivergent lncRNA and mRNA core promoters, respectively. Conversely, intergenic lncRNAs (lincRNAs) and eRNAs have lower capacities to drive transcription and are more tissue specific than divergent genes. This higher tissue specificity is strongly associated with having less complex transcription factor (TF) motif profiles at the core promoter. We experimentally validated these findings by testing both engineered single-nucleotide deletions and human single-nucleotide polymorphisms (SNPs) in MPRA. In both cases, we observe that single nucleotides associated with many motifs are important drivers of promoter activity. Thus, we suggest that high TF motif density serves as a robust mechanism to increase promoter activity at the expense of tissue specificity. Moreover, we find that 22% of common SNPs in core promoter regions have significant regulatory effects. Collectively, our findings show that high TF motif density provides redundancy and increases promoter activity at the expense of tissue specificity, suggesting that specificity of expression may be regulated by simplicity of motif usage.

[Supplemental material is available for this article.]

Transcription factors (TFs) regulate gene expression by binding to DNA regulatory elements at both coding and noncoding genomic elements, including mRNA and long noncoding RNA (lncRNA) promoters and enhancers. Classically, promoters and enhancers have been defined as distinct categories of regulatory elements. However, recent findings suggest that promoters and enhancers share a common regulatory code, as transcription is initiated at both (Core et al. 2008; Engreitz et al. 2016). Indeed, at both promoters and enhancers, RNA polymerase II (Pol II) binds to a 50- to 100-bp stretch of DNA termed the "core promoter" and transcribes in both the sense and antisense directions—a phenomenon known as bidirectional transcription (Andersson 2015). Such transcription at promoters typically produces long, stable polyadenylated transcripts in the sense direction and short, unstable, nonpolyadenylated transcripts in the antisense direction (Andersson 2015). At enhancers, highly unstable RNAs, named enhancer RNAs (eRNAs), are produced in a bidirectional manner (The FANTOM Consortium and the RIKEN PMI and CLST [DGT] 2014).

Although almost all promoters exhibit bidirectional transcription, in some cases, this bidirectional transcription results in two stable transcripts that are arranged in a "head-to-head" orientation (one on the sense strand and one on the antisense strand). These so-called "divergent" transcripts are abundant in the human genome, are evolutionarily conserved, and often comprised of two highly expressed individual core promoter sequences (Trinklein et al. 2004). It remains unclear, however, whether their high expression levels are a byproduct of having two promoters in close proximity or whether it is an inherent property of their DNA sequence. Additionally, these divergent transcript pairs can also include lncRNAs, but whether divergent lncRNA promoters have distinct regulatory properties compared to divergent mRNA promoters is also unknown.

Like mRNAs, lncRNAs are transcribed by Pol II, canonically spliced, and polyadenylated. However, lncRNAs also show similarities to enhancers: They have similar chromatin environments

[12]These authors contributed equally to this work.
Corresponding author: marta.mele.messeguer@gmail.com
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.242222.118. Freely available online through the *Genome Research* Open Access option.

29:344–355 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/19; www.genome.org

(Marques et al. 2013), and they often act as enhancers themselves by activating the transcription of nearby genes (Rinn and Chang 2012; Ørom and Shiekhattar 2013). As a class, lncRNAs are known to be more lowly expressed and more tissue specific than protein-coding genes (Cabili et al. 2011; Derrien et al. 2012; Molyneaux et al. 2015). Although lncRNAs are less conserved than protein-coding genes, their promoters—and the TF binding sites within their promoters—are well-conserved (Ponjavic et al. 2007; Melé et al. 2017), suggesting that a conserved regulatory logic controls lncRNA transcription. However, the rules that govern lncRNA transcription and that determine their higher tissue specificity remain unclear. For example, it is unknown whether lncRNAs are more tissue specific than mRNAs due to differences in their TF binding profiles.

In this work, we address the fundamental question: Is there an underlying "code" in lncRNA and mRNA promoter and enhancer sequences that accounts for their established differences in tissue specificity and abundance? To address this, we used a massively parallel reporter assay (MPRA)—in which thousands of regulatory sequences of interest are assayed in a single experiment (Melnikov et al. 2012; Patwardhan et al. 2012)—to dissect core promoter sequence properties at high resolution and across multiple cell types. MPRAs have previously uncovered important characteristics of promoters and enhancers (Nguyen et al. 2016; Arnold et al. 2017), but to date a systematic analysis of whether intrinsic features of DNA sequence are responsible for differential activity at lncRNA promoters, protein-coding gene promoters, and enhancers has not been performed.

Results

Divergent lncRNA core promoters are strong and ubiquitously expressed

We first defined five biotypes: (1) eRNAs (RNAs emerging from bidirectionally transcribed enhancers that do not overlap protein-coding genes); (2) intergenic lncRNAs (lincRNAs); (3) divergent lncRNAs (lncRNAs that share promoters with either a protein-coding gene or another lncRNA in the antisense direction); (4) mRNAs; and (5) divergent mRNAs (mRNAs that share promoters with either another protein-coding gene or a lncRNA in the antisense direction) (Fig. 1A; Methods). Note that here the term "divergent" refers to the presence of a stable annotated transcript in the antisense direction, not the potential bidirectionality of the promoter itself. Because the TSSs of lncRNAs can be more poorly annotated than the TSSs of mRNAs, which could bias results when comparing them (Lagarde et al. 2017), we carefully selected a set of high-confidence TSSs defined by The FANTOM5 Consortium across all biotypes. Specifically, we used the stringent set of enhancer TSSs (for eRNAs) and promoter TSSs (for the remaining biotypes) defined as "robust" in the FANTOM5 project (Andersson et al. 2014; The FANTOM Consortium and the RIKEN PMI and CLST [DGT] 2014). For the promoter TSSs, we only considered TSSs that were within 50 bp of an annotated gene start site (Methods). In total, our genome-wide set of core promoter regions included 29,807 eRNAs, 4280 lincRNAs, 1713 divergent lncRNAs, 14,332 mRNAs, and 4235 divergent mRNAs. Analysis of cap analysis of gene expression followed by sequencing (CAGE-seq) data across 550 tissues and cell types (973 samples) for each TSS confirmed that mRNAs were more highly expressed and less tissue specific than lncRNAs and eRNAs (Supplemental Fig. S1). Additionally, for both lncRNAs and mRNAs, divergent transcripts were more highly expressed and less tissue specific than their nondivergent counterparts (Supplemental Fig. S1; Supplemental Methods, Genome-wide Analysis section).

To experimentally test the previously mentioned computational predictions and dissect the contribution of DNA sequence to the observed expression and tissue-specificity patterns, we designed an MPRA in which we could assess the activity of 2078 unique TSSs encompassing all five biotypes (564 eRNAs, 525 lincRNAs, 353 divergent lncRNAs, 599 mRNAs, and 137 divergent mRNAs) expressed across three diverse human cell lines: K562 (chronic myelogenous leukemia), HepG2 (liver carcinoma), and HeLa (cervical adenocarcinoma) (Fig. 1A,B; Supplemental Table S1; Methods). Since most TF motifs and ChIP-seq peaks were enriched near the TSS (Supplemental Fig. S2), we designed oligonucleotides that spanned the core promoter (from 80 bp upstream to 34 bp downstream from the TSS) (Methods). We linked each core promoter to a minimum of 15 unique 11-nt barcodes to ensure redundancy across sequencing measurements (Supplemental Table S2). We performed a minimum of four replicates and a maximum of 12 replicates per condition. We measured a sequence's ability to drive transcription—termed "MPRA activity"—by calculating the fold change between RNA barcode counts and input DNA library barcode counts after normalizing for sequencing depth (Methods). MPRA activity measurements across replicates within a given condition were highly correlated (Supplemental Fig. S3).

We first validated the MPRA by comparing core promoter activity measurements to negative controls; as expected, core promoters were significantly more active than random sequences in all three cell types (Fig. 1C; Supplemental Fig. S4). In general, MPRA activities correlated well with endogenous CAGE-seq expression (Supplemental Fig. S5). eRNAs had the lowest activity, followed by lincRNAs, which is consistent with the CAGE-seq results and indicates that lincRNA core promoters are stronger than eRNA core promoters (Fig. 1C; Supplemental Fig. S4). As we saw using CAGE-seq expression, we found that divergent mRNAs were more active than nondivergent mRNAs, and that divergent lncRNAs were more active than intergenic lncRNAs (Fig. 1C; Supplemental Fig. S4). This implies that, on average, an individual divergent promoter is stronger than an individual nondivergent promoter. Therefore, the higher CAGE-seq expression levels observed in divergent lncRNAs compared to lincRNAs cannot solely be explained by having two promoters in close proximity. When looking at expression-matched core promoters only, these results were substantially weakened (Supplemental Fig. S6), indicating that we are capturing innate expression differences between biotypes.

We further tested whether our MPRA could recapitulate endogenous cell-type–specificity patterns. Briefly, we recalculated tissue-specificity values using K562, HepG2, and HeLa CAGE-seq expression data only (termed "cell-type specificity") and found that 67% of sequences agreed in CAGE-seq and MPRA cell-type–specificity designations (i.e., were classified as either specific in both or nonspecific in both) (Fig. 1D). Consistently, eRNAs and lincRNAs were more tissue specific than mRNAs and divergent transcripts (Fig. 1E). Thus, the DNA sequences of core promoter regions alone drive part of the tissue-specificity pattern that is present endogenously.

We next sought to determine whether differences in expression patterns between biotypes are associated with known core promoter elements. Core promoters are often classified into two types: ubiquitously expressed promoters (associated with CpG islands and a depletion of TATA box motifs) and tissue-specific

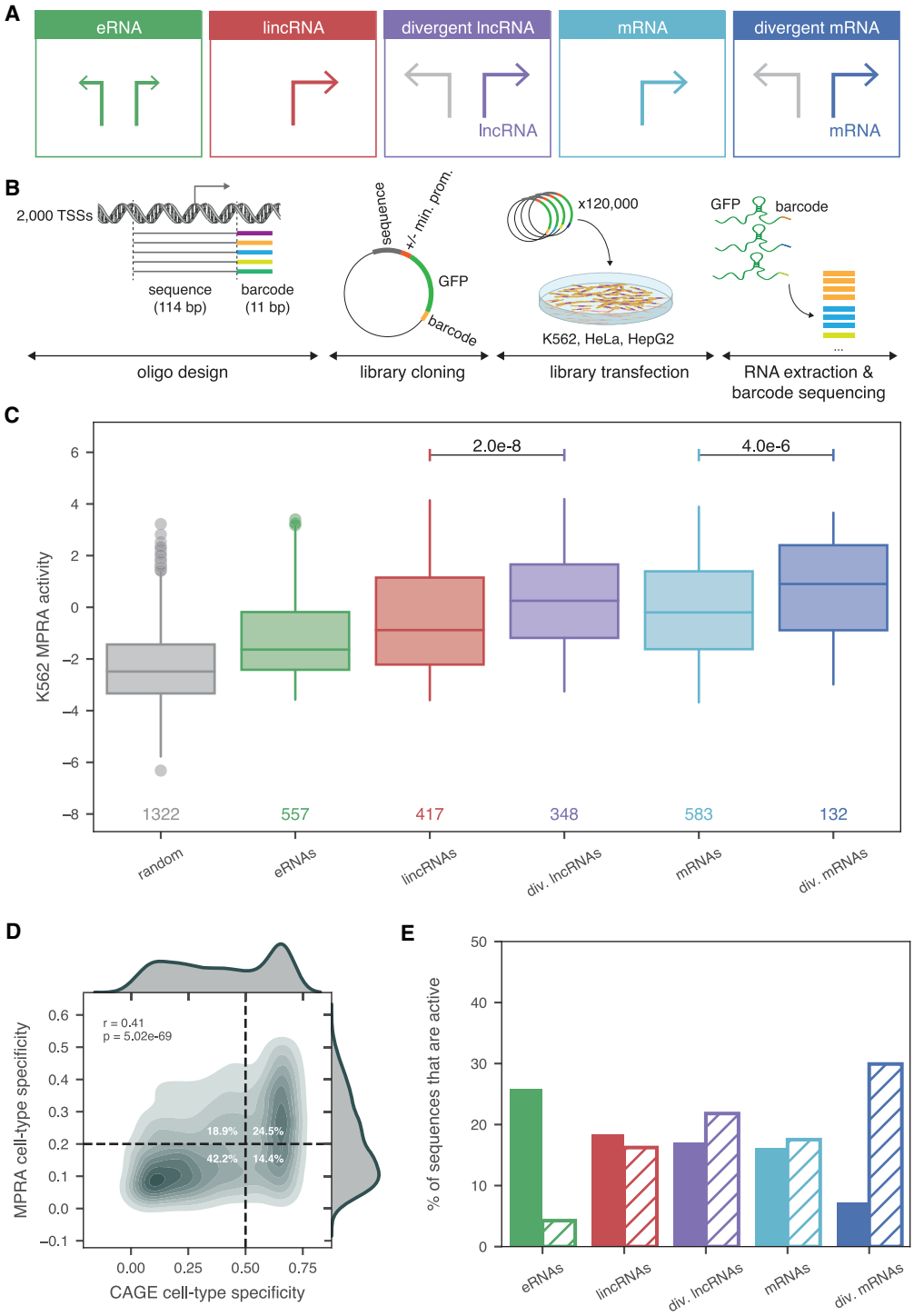

Figure 1. Core promoter sequences of different TSS classes vary in strength and cell-type specificity. (*A*) Overview of TSS classification based on element class (lncRNA, mRNA, or eRNA) and presence or absence of a divergent stable transcript arising from the same promoter region on the antisense strand. (*B*) Schematic of MPRA experimental design. (Min. prom.) Minimal promoter. (*C*) Comparison of MPRA activities (fold change between normalized RNA barcodes and input DNA barcodes) of the reference sequences of each TSS class to negative control sequences in K562 cells. Only TSSs that meet the quality criteria of three or more barcodes represented each with five or more DNA and RNA counts are plotted and *n* values are shown. *P*-values listed are from a two-sided Wilcoxon test. (*D*) Correlation between CAGE cell-type specificity calculated across HeLa, HepG2, and K562 (*x*-axis) and MPRA cell-type specificity across the same three cell lines (*y*-axis). The *upper right* and *lower left* quadrants correspond to sequences that agree with CAGE and MPRA and make up 67% of sequences. Dashed *horizontal* and *vertical* line thresholds for specificity were determined from the distribution of specificity values, shown as density plots on the *top* and to the *right* of the main plot. Spearman's ρ and *P*-value are shown. (*E*) Percent of sequences that are active in only one cell type (solid bars) or all three cell types (K562, HepG2, and HeLa; hatched bars) within each biotype.

promoters (enriched for TATA box and Initiator [Inr] motifs) (Medina-Rivera et al. 2018). As expected, we found that more ubiquitously expressed biotypes had higher CpG content (Supplemental Fig. S7A). All biotypes had similar numbers of sequences containing Inr motifs (Supplemental Fig. S7B). Very few sequences (~3%) had canonical TATA box motifs, which are traditionally associated with tissue-specific expression. Although eRNAs and lincRNAs had more TATA boxes than divergent lncRNAs and divergent mRNAs, mRNAs had a relatively high number of TATA boxes and equally high numbers of both TATA boxes and Inr motifs together (Supplemental Fig. S7C,D). Thus, it would appear that tissue specificity cannot be explained by core promoter elements alone, because mRNAs, which are more ubiquitously expressed than eRNAs and lincRNAs, are enriched for more canonical tissue-specific core promoter elements such as the TATA box.

Fewer overlapping TF motifs in lincRNAs and enhancers contribute to their lower expression levels and higher cell-type specificity

Our earlier results showed that MPRA can partially recapitulate endogenous patterns of gene expression, including abundance—for which MPRA activity is a proxy—and cell-type specificity. Therefore, we aimed to further understand what sequence features could be contributing to the lower abundance and higher tissue specificity of eRNA and lincRNA core promoters. To that end, we focused on two main features: TF motif architecture within a core promoter sequence and the cell-type specificity of the TFs themselves that are present within a core promoter sequence. To determine core promoter TF motif architecture, we first mapped motifs (corresponding to 519 TFs) within our core promoter sequences using FIMO (Grant et al. 2011). Since the presence of a computationally predicted motif does not always indicate physiological binding of the TF (Wasserman and Sandelin 2004), we then intersected these mapped motifs with ChIP-seq peaks corresponding to 771 TFs (218 of which we had motifs for) (Mei et al. 2017) and considered only the motifs that overlap a corresponding ChIP-seq peak (Methods). We divided TF motif architecture into two components: (1) the number of independent motif binding sites in linear sequence space; and (2) the number of overlapping motifs, which should be proportional to the number of different TFs that can bind to a specific sequence pattern. As a proxy for the number of independent motif binding sites, we used the number of base pairs covered by at least one motif in a given sequence. As a proxy for the number of overlapping motifs, we used the maximum coverage of motifs per sequence. As a proxy for the cell-type specificity of the TFs themselves, we calculated the mean cell-type specificity (across HepG2, HeLa, and K562) of all TF motifs within a given promoter (Fig. 2A).

To test the relative importance of these three components (number of base pairs covered by a motif, maximum coverage of motifs, and average TF cell-type specificity), we calculated the proportion of the variation in both MPRA activity and MPRA cell-type specificity that can be explained by each measurement using a general linear model (Methods). The number of overlapping motifs explains a slightly higher proportion of the variation than the number of base pairs covered by a motif when predicting either mean MPRA activity or MPRA cell-type specificity (Fig. 2B,C). Conversely, the cell-type specificities of the TFs themselves explain relatively little of the variation in MPRA activity and cell-type specificity (Fig. 2B,C). We also evaluated how much individual TF motifs contribute to sequence activity (Methods). No single TF motif was able to explain >1.5% of the variation (Supplemental Fig. S8B). Overall, our model suggests that having highly overlapping motifs is substantially predictive of higher transcriptional activity and decreased cell-type specificity.

Next, we looked at the motif architecture in biotypes that are known to be tissue specific (such as eRNAs and lincRNAs) compared to biotypes that are known to be ubiquitous (such as mRNAs and divergent genes). We observed that tissue-specific biotypes had both fewer base pairs covered by a motif and fewer overlapping motifs than ubiquitously expressed biotypes (Fig. 2D). We then classified individual lincRNA and mRNA TSSs as being either ubiquitously expressed (>0 CAGE tpm in >90% of samples), tissue specifically expressed (>0 CAGE tpm in <10% of samples), or dynamically expressed (a subset of tissue-specific genes, where in at least one sample the TSS is expressed at >50 CAGE tpm). Ubiquitously expressed TSSs within each biotype had both more base pairs covered by a motif and more overlapping motifs than tissue-specific and dynamic TSSs (Fig. 2E).

Some TF motifs are highly similar to each other, creating potential redundancies in motif databases (Mathelier et al. 2014). To control for this, we used two independent methods to cluster similar motifs. First, we performed unbiased clustering of the 519 motifs using MoSBAT (Lambert et al. 2016), resulting in 223 motif clusters (Supplemental Fig. S9A). Second, we used a list of 108 non-redundant 8-mer motifs generated using protein binding microarrays across 671 TFs (Mariani et al. 2017). We then recalculated the preceding metrics (number of base pairs covered by a motif and the maximum motif coverage) after removing each set of redundant motifs. We found that for both metrics, ubiquitously expressed biotypes had higher maximum coverage values than tissue-specific biotypes (Supplemental Figs. S9B,C, S10, S11). Moreover, DNA regions that harbor many overlapping TF motifs are more conserved than those harboring only one TF motif (Fig. 2F) and more DNase accessible across tissues (Supplemental Fig. S12). Thus, using computationally mapped TF motifs, endogenous TF binding events via ChIP-seq, and unique TF clusters, we observe that high and ubiquitous expression is correlated with many overlapping motifs.

Targeted deletions refine functional TF motifs in lncRNA promoters

Our results suggest that overlapping TF motifs that can be bound by many different TFs—potentially in different contexts—are associated with increased expression and decreased tissue specificity. We thus hypothesized that disruption of highly overlapping motifs should have larger effect sizes than disruption of more specific motifs. To test this, we performed a second MPRA across the core promoters of 21 disease-associated lncRNAs, five nearby mRNAs, and five nearby eRNAs (Supplemental Table S3) and tested the effect of single-nucleotide deletions across each core promoter in HepG2 and K562 cells (Fig. 3A; Supplemental Table S4). To ensure that we covered all motifs surrounding the TSS, we included two tiles for each TSS (from 183 bp upstream to 69 bp upstream and 89 bp upstream to 25 bp downstream from the TSS). Thus, this strategy allows us to assess the contribution of each individual nucleotide to core promoter activity independently in a single experiment (Patwardhan et al. 2009).

First, we confirmed that test core promoter sequences had significantly more activity than negative control sequences in both cell types (Supplemental Fig. S13). Next, we calculated the "effect size" of each deletion as a fold change in MPRA activity relative to the full reference sequence. In order to determine how deletion

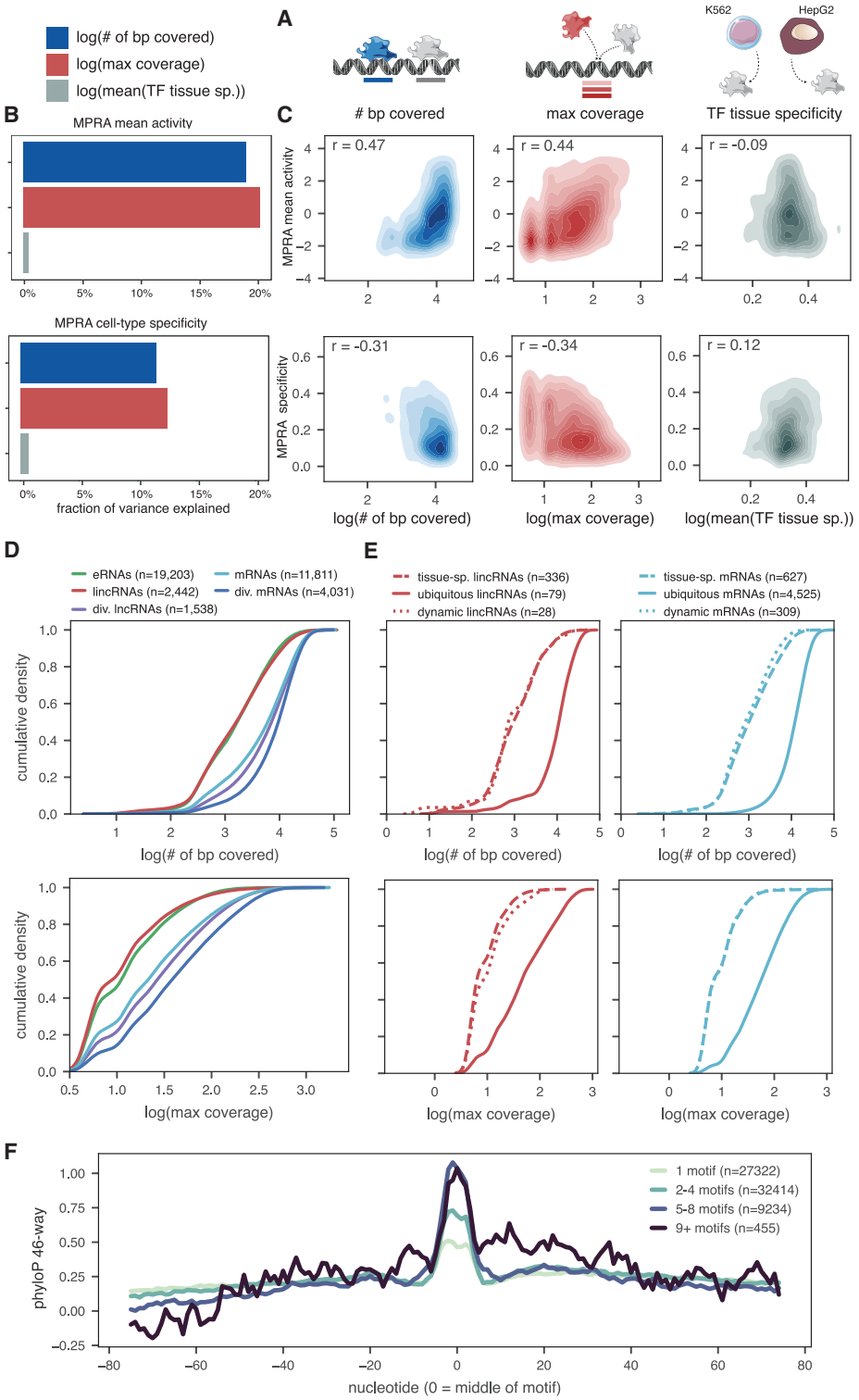

Figure 2. Coverage of TFs within a binding site explains expression levels and cell-type–specificity variability. (*A*) Schematic of the three metrics used to model the capacity to drive transcription and cell-type specificity in the MPRA. For each metric, only TF motifs that have been validated by ChIP (i.e., overlap a ChIP peak for the cognate TF) are considered. (*B*) Fraction of variance explained by each of the three metrics for either mean MPRA activity (*top*) or MPRA cell-type specificity (*bottom*). (*C*) Correlation between the three metrics (*x*-axis) and either the mean MPRA activity (*top*) or MPRA cell-type specificity (*bottom*) across HeLa, HepG2, and K562. Spearman's ρ is shown. (*D*) Cumulative density plot of the number of base pairs covered by a motif across all biotypes (*top*) and the maximum motif coverage across all biotypes (*bottom*). (*E*) Cumulative density plot for number of base pairs covered (*top*) and maximum motif coverage (*bottom*) either within lincRNAs (*left*) or within mRNAs (*right*), looking only at TSSs that are defined as tissue-specific (tissue-sp.), ubiquitous, or dynamically expressed (see text). (*F*) Metaplot of the average phyloP 46-way placental mammal conservation score centered on motif regions, broken up by how many individual TF motifs map to the region. In all plots, only sequences with at least one validated motif were considered.

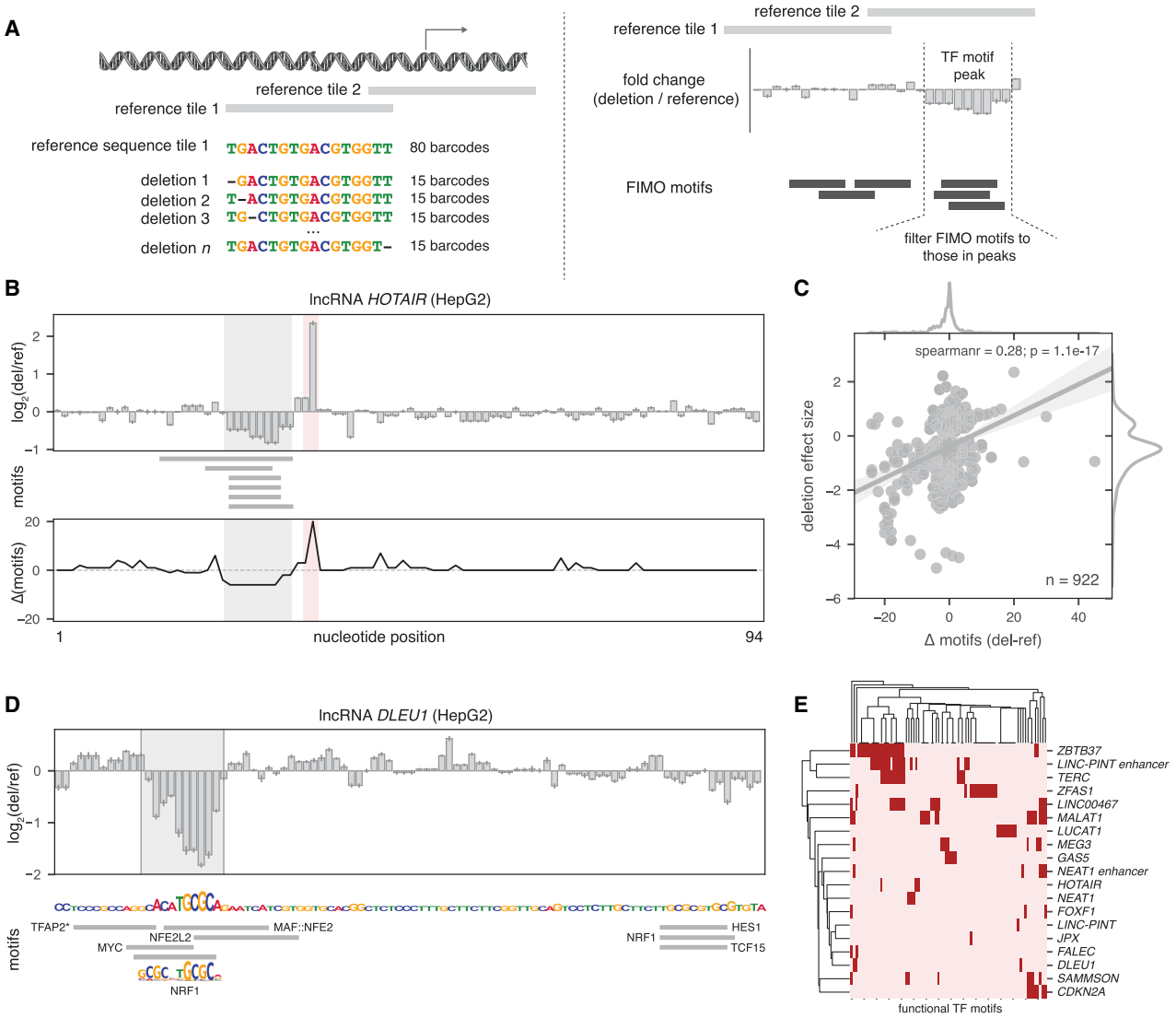

Figure 3. Targeted deletions refine TF motifs within lncRNA promoters. (*A*) Schematic of the single-nucleotide deletion MPRA design (*left*) and the output interpretation (*right*). (*B*) MPRA deletion profile for the lncRNA *HOTAIR* promoter (*top*), the positions of computationally mapped motifs in the reference sequence (*middle*), and the number of motifs predicted to be gained or lost due to the single-nucleotide deletions (*bottom*). Shaded areas represent the strongest gain (red) or loss (gray) of activity. (*C*) Correlation between the number of motifs predicted to be disrupted (*x*-axis) and the effect size of deletions (*y*-axis) for all significant deletions in HepG2. (*D*, *top*) MPRA deletion profile for the lncRNA *DLEU1* promoter. Shaded area is a called peak. (*Bottom*) *DLEU1* sequence (plotted with letter heights proportional to loss of activity in the MPRA) and computationally mapped motifs (gray boxes). The sequence logo for NRF1 is shown. (*) TFAP2A, TFAP2B, and TFAP2C all map to the noted gray box. (*E*) Heatmap showing all computationally mapped motifs that overlap deletion peaks in HepG2.

effect sizes correlate with TF motif profiles, we calculated the number of computationally mapped motifs that are lost (or gained) in each deletion sequence relative to the full reference sequence (Fig. 3B). Individual nucleotides that overlap a predicted motif are important in maintaining transcription, as deletion of each nucleotide independently shows a strong loss of activity (Fig. 3B, gray shaded area). Additionally, we also saw deletions with gain-of-function effects, for example, deleting a single nucleotide in the lncRNA *HOTAIR* core promoter is predicted to create 20 new TF motifs and causes a strong increase in activity (Fig. 3B, red shaded area). These observations extended to the remaining core promoters: Deletion effect sizes were generally correlated with the number of motifs computationally predicted to be affected by each deletion (Fig. 3C; Supplemental Fig. S14A).

Moreover, single-nucleotide deletions can be used to better identify functional DNA regulatory motifs than computational motif mapping, because the strategy directly tests whether specific nucleotides are required for transcription in a particular cellular context (Supplemental Fig. S14B). We therefore took advantage of the fact that functional DNA regulatory regions appear as "peaks" in the deletion effect size map and intersected these peaks with computationally mapped motifs (Fig. 3A). Of all of the computationally mapped motifs in these sequences, 41% and 49% were found to be functional in the tested cell line, i.e., overlapped deletion peak regions, in HepG2 and K562, respectively. For example, the lncRNA *DLEU1*, which is frequently lost in lymphocytic leukemia (Liu et al. 1997), contains eight predicted TF motifs, but only one motif (NRF1) significantly overlapped the peak found

via single-nucleotide deletions (Fig. 3D). Therefore, we hypothesize that NRF1, which has a known role in the immune system (Suliman et al. 2010), is the primary and direct regulator of *DLEU1*. Consistent with this, NRF1 also has a corresponding ChIP peak in the *DLEU1* promoter. In total, we were able to determine a wide range of functional TF motifs in 15 lncRNAs, three eRNAs, and three mRNAs (Fig. 3E; Supplemental Fig. S15). These results show the utility of MPRA in combination with single-nucleotide deletions to refine functional TFs.

Finally, we reexamined the idea that sequences that can be bound by many TFs are more broadly expressed. Indeed, we found that sequences that were active in both cell types had more of our detected functional TF motifs than sequences that were active in only one of the tested cell types ($P = 0.061$, one-sided Wilcoxon test) (Supplemental Fig. S16). This again suggests that the more TFs a sequence can bind, the broader its expression pattern.

More than 20% of genetic variants within core promoters have regulatory effects

We extended our single-nucleotide MPRA studies to examine how human variation affects promoter activity in contrast to engineered deletions. Briefly, we used MPRA to identify regulatory SNPs that could affect a sequence's ability to drive transcription in our set of 21 disease-associated lncRNA core promoters. The effect sizes of the tested SNPs were highly correlated with the deletion effect sizes (Supplemental Fig. S17A). More importantly, significant SNPs tended to occur in peaks corresponding to TF motifs (Supplemental Fig. S17B). In fact, 78% and 90% of significantly regulatory SNPs that decrease expression overlapped deletion-predicted TF peaks in HepG2 and K562, respectively, compared to only 9% and 5% of nonregulatory SNPs. The tumor suppressor lncRNA *MEG3*, for example, harbors one regulatory SNP shown to be mutated in breast cancer tumors by two separate studies (Forbes et al. 2017). This SNP lies in a functional TF peak predicted to harbor binding sites for the CREB family of TFs (Supplemental Fig. S17B). Together, these results show that our MPRA strategy can identify regulatory SNPs that disrupt functional TF motifs.

To gain a wider understanding of how genetic variation affects DNA regulatory elements, we next used MPRA to test all common SNPs annotated in our set of approximately 2000 core promoters in HepG2 and K562 (Fig. 4A). We correctly identified 100% and 71% of positive control variants as significantly regulatory in HepG2 and K562, respectively (Supplemental Fig. S18). As with the deletion effect sizes, SNP effect sizes also correlated with the number of predicted TF motifs disrupted by the SNP (Fig. 4B; Supplemental Fig. S19), again suggesting that disruption of multiple overlapping motifs is associated with larger expression changes.

Overall, we found that as many as 22% of SNPs in the tested TSS regions have an effect on promoter strength (Supplemental Table S5; Supplemental Fig. S20). We predict that this proportion would increase with a higher number of barcodes (Supplemental Fig. S21A) and replicates (Supplemental Fig. S21B). When we looked within each biotype, we found no differences in the number of regulatory SNPs or in the SNP effect sizes (Supplemental Fig. S22). We found that 55% of regulatory SNPs have an effect in only one of the two cell types (Fig. 4C).

Due to linkage disequilibrium (LD) in the human genome, multiple individual SNPs tend to be inherited together in haplotypes. However, how individual SNPs interact within a haplotype remains unclear. We therefore sought to determine whether individual SNPs in TSSs tend to interact additively (i.e., the effect of all SNPs together is equal to the sum of their individual effects) or epistatically (i.e., the effect of one SNP masks the effects of the other SNPs). We found that a minority of SNPs acted epistatically because only 16% and 22% of SNPs had a nonadditive effect in HepG2 and K562, respectively (Fig. 4D).

Finally, we sought to identify regulatory SNPs that are in LD with GWAS hits. We identified 96 and 36 such SNPs in HepG2 and K562, respectively (Supplemental Table S6). To analyze the putative relationship between the regulatory potential of an MPRA-tested SNP and the GWAS-associated phenotype, we selected SNPs that (1) are regulatory SNPs in both HepG2 and K562 cells (32 total); (2) disrupt known TF motifs; and (3) have nearby coding genes that are associated with the GWAS-associated phenotype. We identified three SNPs with significant regulatory effects in both HepG2 and K562 cells that are associated with levels of HDL cholesterol (rs3785098) (Willer et al. 2013), lung cancer (Wang et al. 2008) or schizophrenia (rs3101018) (Goes et al. 2015), and inflammatory bowel disease (IBD) (rs4456788) (Liu et al. 2015), respectively (Fig. 4E; Supplemental Fig. S23). For example, the IBD-associated SNP rs4456788 disrupts six TF motifs and shows significantly lower MPRA activity compared to the reference allele (Fig. 4E). As well as being associated with IBD, this SNP is known to be an eQTL for the protein-coding gene *ICOSLG* (The GTEx Consortium 2015); thus, this MPRA result could provide an important clue—and a testable hypothesis—as to the biological pathway that is responsible for this genetic association.

Discussion

Here, we have characterized the differences between lncRNA, mRNA, and eRNA core promoter sequences by combining computational predictions and experimental testing using high-throughput assays. Because many lncRNAs are thought to arise from enhancers (Marques et al. 2013) or bidirectional transcription stemming from protein-coding promoters (Sigova et al. 2013), we sought to determine whether lncRNA promoters are intrinsically different from enhancers and protein-coding promoters. Our findings suggest that the regulation of divergent lncRNAs and intergenic lncRNAs are quite different. Divergent lncRNAs have more TF motifs and consequently have stronger promoters than intergenic lncRNAs. Notably, higher expression levels of divergent lncRNAs compared to lincRNAs cannot solely be explained by having a nearby protein-coding promoter. Rather, we show that both divergent lncRNA and mRNA core promoters are intrinsically stronger than nondivergent lncRNA and mRNA promoters (Fig. 1C). Conversely, intergenic lncRNA TSSs are similar to enhancer TSSs, both in terms of their TF motif architecture and expression patterns, with both biotypes being highly tissue specific (Fig. 2D).

Our results suggest that core promoter sequences play important roles in determining transcript tissue specificity, because our MPRA results can partially recapitulate endogenous expression patterns (Fig. 1D,E). Importantly, using MPRA allows us to thoroughly interrogate the regulatory potential of DNA sequence alone while controlling for other factors such as chromatin differences and effects of post-transcriptional regulation. However, we recognize that this study has some limitations that are intrinsic to using episomal plasmids in MPRA. These limitations are reflected within our own data, in which sequence alone only accounts for ~50% of observed expression profiles when modeled. Nonetheless, our approach has characterized the baseline to which higher-order structural and epigenetic information can be added in order to gain a more complete picture of transcriptional regulation.

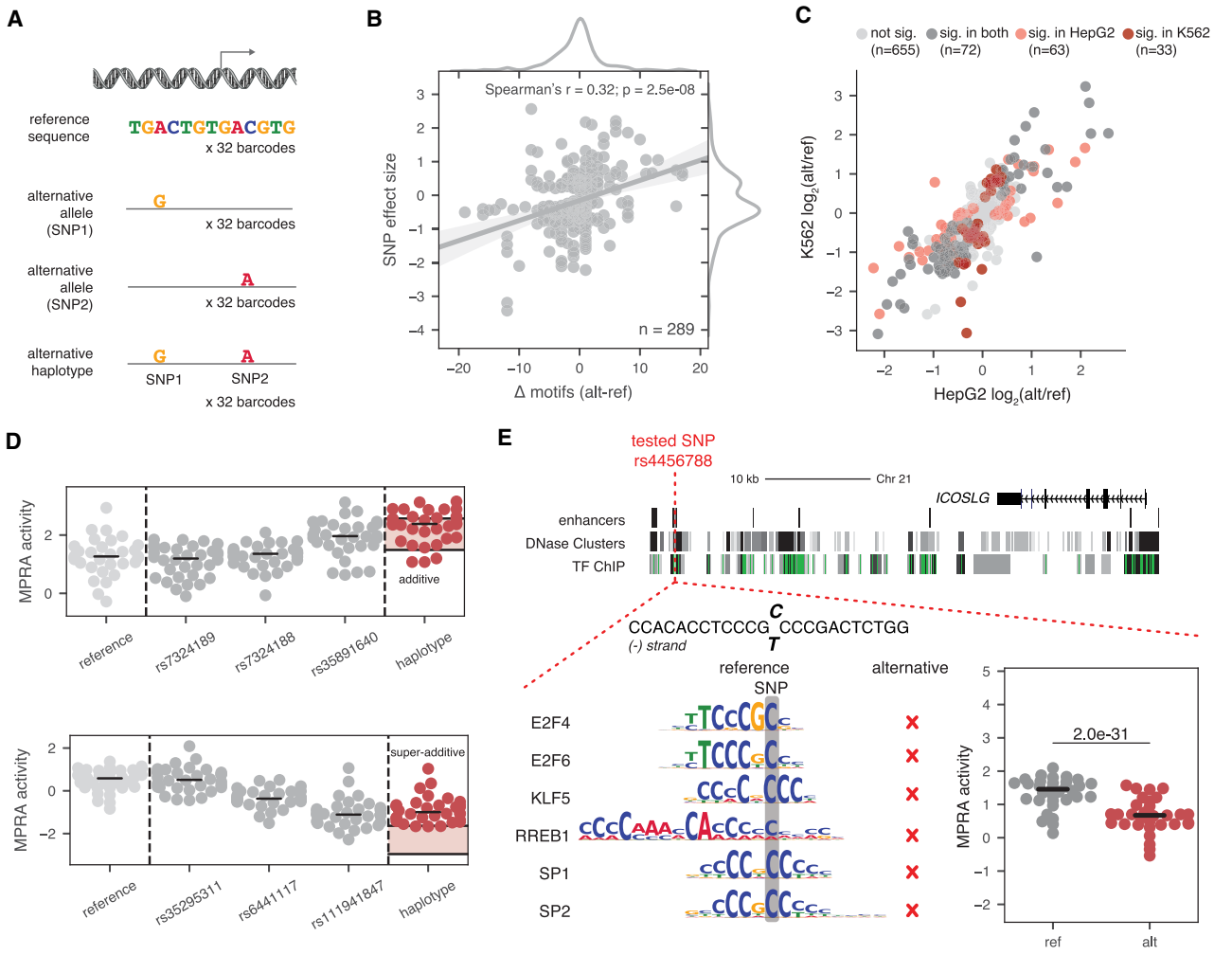

Figure 4. Twenty-two percent of SNPs in promoter and enhancer TSSs have regulatory effects. (*A*) Schematic of SNP and haplotype testing in MPRA. (*B*) Correlation between the number of TF motifs disrupted (*x*-axis) and the SNP effect size (*y*-axis) for all significant SNPs in HepG2. SNP effect size is the mean $\log_2$ fold change in MPRA activity between the alternative and reference alleles. (*C*) Correlation between SNP effect sizes in HepG2 (*x*-axis) and K562 (*y*-axis). (*D*) Examples of two haplotype effects, one additive (*top*) and one super additive (*bottom*). Dots represent barcode activity means across replicates for reference tile (light gray), individual SNP tiles (dark gray), or haplotype tiles (red). Shaded red area in the haplotype column refers to the 90% confidence interval surrounding the expected median additive effect. (*E*) Example of a SNP near *ICOSLG* that disrupts six TF motifs present on the reference allele. The difference in MPRA activity between the reference and alternative alleles in HepG2 is shown. *P*-value listed is from a two-sided Wilcoxon test.

Our data are consistent with a model in which highly abundant genes have complex TF binding profiles, with stretches of promiscuous DNA that can be recognized by many TFs (Fig. 5). Several lines of evidence point toward overlapping binding sites playing a role in determining abundance and tissue specificity. First, we see that a model trained on MPRA data finds the number of overlapping motifs to be highly predictive of abundance and anticorrelated to cell-type specificity (Fig. 2B,C). Second, we find that tissue-specific biotypes have fewer overlapping motifs than ubiquitously expressed biotypes (Fig. 2D). We also find that within one biotype, tissue-specific genes have fewer overlapping motifs than ubiquitously expressed genes (Fig. 2E). Third, we show that sequences that are expressed in both HepG2 and K562 have more functional motifs than sequences that are only expressed in one cell type (Supplemental Fig. S16). Finally, we see that both single-nucleotide deletions and SNPs that are predicted to disrupt more motifs have higher effect sizes (Figs. 3B, 4B). For example, a single-nucleotide deletion in the *HOTAIR* promoter generates 20 new TF motifs and subsequently increases promoter activity by fourfold (Fig. 3B).

By definition, overlapping binding sites are at the same distance from the TSS; interestingly, where TFs bind in relation to the TSS is important (Tabach et al. 2007). We speculate that this system would allow genes to maintain high and ubiquitous expression levels across different tissues and conditions despite the likely fluctuating expression levels of the TFs. Further, this redundancy could explain why knockdown of certain TFs often does not result in the misregulation of the expected target genes (Cusanovich et al. 2014), as other TFs would be able to bind to the same position. Thus, we propose that promoter specificity may be a function of simplicity in motif usage. Further work remains to be done to experimentally test the extent of the role that these overlapping motif profiles play in regulating expression and specificity.

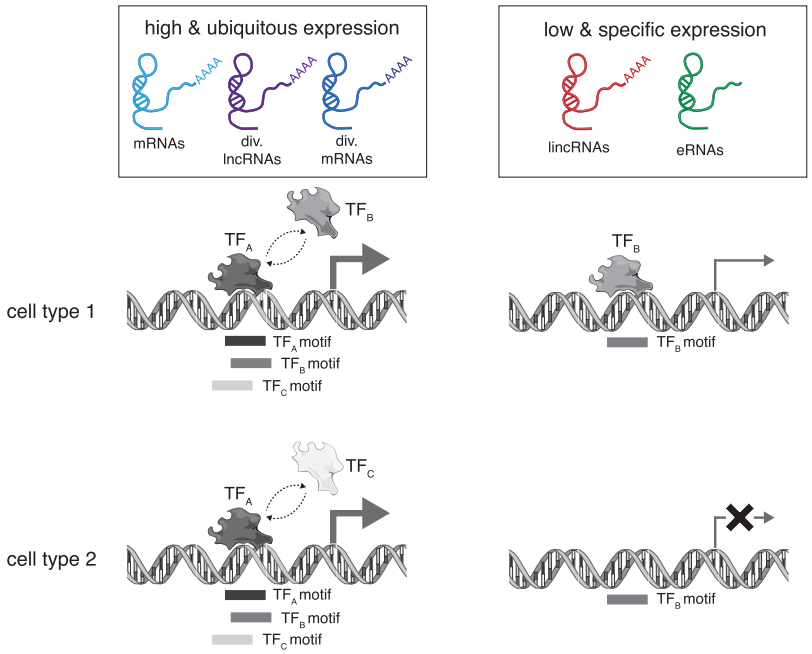

Figure 5. In this model, overlapping TF motifs are associated with high expression and low specificity. Schematic showing biotypes that are highly and ubiquitously expressed (*left*; thick arrow) have more overlapping TF binding sites (gray shaded boxes) and thus more TFs can bind both within a specific cell type or across cell types. Biotypes that are lowly and specifically expressed (*right*; thin arrow or crossed arrow) have fewer overlapping motifs and thus only a few TFs (one in the example) can bind. TF_A is present in cell types 1 and 2, whereas TF_B and TF_C are only present in cell type 1 and cell type 2, respectively.

Our findings also have evolutionary implications. Much attention has been given to the fact that enhancers and lincRNAs have rapid sequence turnover (Hon et al. 2017). Our findings are consistent with this notion. We find that tissue-specific TSSs, such as those of lincRNAs and eRNAs, have less complex motif profiles. Thus, they may be more likely to appear and disappear throughout evolutionary time. In fact, DNA regions with few overlapping motifs are more poorly conserved than DNA regions with many overlapping motifs (Fig. 2F). Conversely, highly transcribed genes have developed more complex TF binding patterns, which may have evolved to produce stable antisense transcripts because they provide a fitness advantage. Indeed, if we compare human and mouse orthologous genes that have gained a stable antisense transcript in either one of the lineages, they show an overall increase in expression (Supplemental Fig. S24). Thus, bidirectional transcription may not only allow for de novo gene origination but could also be an evolutionary mechanism to increase expression of the gene in the sense direction. This may also help explain why divergent mRNA–lncRNA pairs occur so frequently in the human genome (Sigova et al. 2013). Overall, this study sheds light on the important roles that core promoters play in complicated aspects of gene regulation, including divergent transcription and tissue specificity, across both coding and noncoding genes.

Methods

TSS biotype classification

mRNA and lncRNA TSSs were classified based on GENCODE v19 (Harrow et al. 2012) gene annotations. All TSSs from genes annotated as lncRNAs were classified as intergenic lncRNA (lincRNA) TSSs if they did not overlap any annotated protein-coding genes or as divergent lncRNA TSSs if the annotated TSS had an antisense FANTOM5 TSS within 1000 bp. Similarly, mRNA TSSs were classified as divergent mRNA TSSs if the annotated TSS had an antisense FANTOM5 TSS within 1000 bp. eRNA TSSs were also defined by FANTOM5 (Andersson et al. 2014) and had two TSSs each—a sense TSS and an antisense TSS—due to their inherent definition of being bidirectionally transcribed.

MPRA TSS selection

To select promoters to include in the MPRA, we used the FANTOM5 robust TSS set (The FANTOM Consortium and the RIKEN PMI and CLST [DGT] 2014). These TSSs are expressed robustly in CAGE-seq data (more than 10 CAGE reads in one sample and >10 tpm CAGE expression in at least one sample). Additionally, we only considered FANTOM5 TSSs that were within 50 bp of their cognate annotated GENCODE v19 transcript TSSs. eRNA TSSs were selected from the enhancer TSS set defined in the same FANTOM5 release (Andersson et al. 2014). Next, we selected TSSs based on their CAGE expression profiles. Specifically, we required the TSSs to either (1) be expressed >0.5 tpm across all replicates of at least one of the tested cell lines (HeLa, HepG2, or K562); or (2) have an average expression >0.5 tpm across all FANTOM5 samples (suggesting they had high baseline expression). Finally, we excluded any lncRNA TSSs arising from transcripts with high coding potential (phyloCSF >0) (Lin et al. 2011) or that overlapped a protein-coding gene in the sense direction.

Because the MPRA was lncRNA-focused, all lncRNA TSSs (lincRNAs and divergent lncRNA TSS) that filled the aforementioned criteria were included for testing in the MPRA. To control for the fact that there were more mRNA and eRNA TSSs than lncRNAs, we selected both expression-matched mRNA and eRNA TSSs (average expression across all FANTOM5 samples matching that of lncRNA TSSs) as well as randomly selected mRNA and eRNA TSSs for further analysis. We also included all protein-coding TSSs that were in close proximity to the selected lncRNA TSSs (<160 bp) in antisense and some of the most highly expressed eRNAs. Additional TSSs were included if they contained at least one SNP in LD with a GWAS hit in their core promoters (for additional details, see Supplemental Table S6). Overall, we ended up with 2078 TSSs for testing in MPRA (Fig. 1A; Supplemental Table S1). More details are available in Supplemental Methods (MPRA TSS Selection section).

MPRA pool design

Two 120,000 oligonucleotide (oligo) pools of 170 bp with 11-bp barcodes were designed. The first pool included core promoter sequences across biotypes and common SNPs falling in these regions (Supplemental Methods; Supplemental Tables S1, S2). The second pool included single-nucleotide deletions across the core promoters of 21 lncRNAs, five enhancers, and five mRNAs with two consecutive reference tiles each (Supplemental Methods;

Supplemental Tables S3, S4). Random and scrambled sequences were included in both pools as negative controls. More details are available in Supplemental Methods (MPRA Oligo Pool Design section).

MPRA cloning and transfection

Oligo pools were synthesized by Twist Biosciences and then cloned into plasmids to generate a library of constructs in which the regulatory sequence is upstream of a reporter gene (here, GFP) that is upstream of a unique barcode (Supplemental Methods). Constructs were transfected into live cells, and barcode expression was assayed by high-throughput RNA sequencing. A minimum of four and a maximum of 12 replicates were performed per condition (cell type and presence/absence of a minimal promoter) adding up to 32 total experiments (Supplemental Fig. S3). Results are based on the minimal promoter setup given the high similarity between replicates with and without the minimal promoter and the fact that more replicates were performed for this setup. More details are available in Supplemental Methods (MPRA Cloning, Transfection, and Sequencing section).

MPRA data analysis

All Python scripts and notebooks used to perform the MPRA analyses are available at https://github.com/kmattioli/2018__lncRNA_promoter_MPRA and provided as Supplemental Code.

Exact matches to known barcodes and six upstream constant nucleotides were mapped after quality-filtering the sequencing reads. Barcodes were filtered to those with five or more counts (in both DNA and RNA). Barcode activities were calculated as the log-transformed proportion of RNA barcodes to the proportion of DNA barcodes (after normalizing for sequencing depth) and were quantile-normalized across replicates. Element activities were calculated as the median activity value across all cognate barcodes, requiring three or more barcodes. Significantly active tiles were defined as those with barcode activities that were significantly higher than random negative control sequences according to a two-sided Wilcoxon test in ≥75% of replicates (Supplemental Methods). Because we had many more replicates in HepG2 than in other cell types and to ensure we had similar power when comparing across cell types (i.e., Fig. 1F), HepG2 replicates were downsampled 100 times and sequences were considered significant if they were significant by the rules above in ≥75% of samples. More details are available in Supplemental Methods (MPRA Analysis section).

Core promoter element analysis

The core promoter was defined as 80 bp upstream to 34 bp downstream from the TSS. CpG content was calculated by counting the number of "CG" dinucleotides in this region. Inr motifs were defined to be matches to the motif BBCABW (B = C/G/T, W = A/T) (Kugel and Goodrich 2017) within 5 bp of the TSS. TATA motifs were defined to be matches to the motifs TATAAA or TATATA within 55–15 bp upstream of the TSS. Position weight matrices for TF binding motifs were obtained from the JASPAR database (core, vertebrates, 2016 release) (Mathelier et al. 2014).

MPRA activity and tissue-specificity predictions

An ANOVA analysis was used to evaluate what properties contribute to MPRA activity and specificity. Specificity across the MPRA activity values for HepG2, K562, and HeLa was calculated using the τ metric as follows (Kryuchkova-Mostacci and Robinson-Rechavi 2017):

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x}_i)}{n-1};\ \hat{x}_i = \frac{x_i}{\max\limits_{1\le i\le n}(x_i)},$$

where x_i is the median activity of a TSS in cell type i; and n is the number of cell types. Briefly, τ calculates the average difference between the activity of a TSS in a given cell type and the TSS' maximal expression across all cell types. Thus, "ubiquitous" TSSs will have τ values close to zero while "tissue-specific" TSSs will have τ values close to one.

To perform the ANOVA analysis, the variance in activity/specificity that is explained by the general sequence features (listed in Supplemental Fig. S6A) was calculated. The variance explained by each parameter was calculated on its own and the optimal subset of parameters was computed. Because the parameters were highly correlated, the optimal subset consisted of only seven of 14 parameters yet explained 41% of the total variance (Supplemental Fig. S6A).

$$\begin{aligned}\text{MPRA activity} \sim\ & \text{CpG content} + \max(\text{coverage}) + \#\ \text{bp covered} \\ & + \text{CG content} + \text{CG content}^2 + \text{total \# motifs} \\ & + \text{total \#\ motifs}^2\end{aligned}$$

Motifs were then added into the model one by one.

$$\begin{aligned}\text{MPRA activity} \sim\ & \text{CpG content} + \max(\text{coverage}) + \#\ \text{bp covered} \\ & + \text{CG content} + \text{CG content}^2 + \text{total \# motifs} \\ & + \text{total \# motifs}^2 + \text{motif is present}\end{aligned}$$

Of the 382 motifs tested, 17 were found to explain a significant fraction of the variance (listed in Supplemental Fig. S6B). Combining the seven sequence features and the significant motifs in a model explained a total of 49% of the variance in MPRA activity.

This analysis was performed in R (version 3) using the leaps and tidyverse packages (R Core Team 2018).

ChIP-seq analysis and TF motif mapping

ChIP-seq files were downloaded from the Cistrome Data Browser (www.cistrome.org) (Mei et al. 2017) for 771 human TFs (Supplemental Table S7)—218 of which overlapped with the set of 519 JASPAR motifs. BEDTools (Quinlan and Hall 2010) was used to merge peaks for a given TF and then intersect the merged ChIP peaks with our set of promoters. Since Cistrome peaks were in hg38 and our promoters were in hg19, we first used liftOver (Hinrichs et al. 2006) to convert our promoters to hg38 coordinates. Motifs were mapped in sequences using FIMO (version 4.11.2) (Grant et al. 2011) with a P-value threshold of 1×10^{-5}. Motifs were assigned to ChIP-seq peaks if there was a FIMO motif mapped within 250 bp of the ChIP-seq peak.

Ubiquitous and tissue-specific TSS categorization

In order to categorize TSSs based on their expression profiles, we used the FANTOM5 CAGE-seq expression data for both TSSs and enhancers. We removed any FANTOM5 samples corresponding to experimental time courses or fractionated cells and then grouped the remaining samples by tissue or cell type (Supplemental Table S8). We then defined the following expression profiles: ubiquitously expressed (>0 CAGE tpm in >90% of these grouped samples), tissue specifically expressed (>0 CAGE tpm in <10% of these grouped samples), or dynamically expressed (a subset of

tissue-specific genes, where in at least one of the grouped samples the TSS is expressed at >50 CAGE tpm).

MPRA deletion analysis and functional TF motif mapping

Deletion effect sizes were defined as the $\log_2$ fold change between the mean activity of the deletion sequence across replicates and the mean activity of the reference sequence across replicates, resulting in a value per nucleotide. Peaks were defined as any stretch of ≥5 nt with effect sizes of ≥ −1.5 × the average standard deviation of the deletion effect sizes in that tile. Mapped motifs were said to be "functional" if ≥1 nt in the motif intersected a peak.

SNP and haplotype analysis

Regulatory SNPs were defined as those whose barcode activities were significantly different and consistent in direction between reference and alternative tiles using a two-sided Wilcoxon test in ≥75% of replicates (Supplemental Methods). Again, when comparing among cell types (i.e., Fig. 4C), HepG2 replicates were downsampled 100 times as previously mentioned in the "MPRA data analysis" section.

To determine additive haplotypes, first the expected additive haplotype effect size was found by summing the median $\log_2$ fold changes (alternative/reference activities) for each individual SNP in a haplotype. This effect was bootstrapped ($n = 1000$) to determine a 90% confidence interval, and a haplotype was considered additive if the actual median $\log_2$ fold change of the haplotype fell within this 90% confidence interval.

The GWAS catalog was downloaded from https://www.ebi.ac.uk/gwas/downloads. Raggr was used to calculate whether any of the MPRA-tested SNPs were in linkage disequilibrium ($r^2 < 0.6$) with any of the GWAS tag SNPs at http://raggr.usc.edu/.

Data access

The MPRA sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE117594. All scripts required to reproduce this work are available as Supplemental Code as well as on GitHub at https://github.com/kmattioli/2018__lncRNA_promoter_MPRA.

Acknowledgments

We thank Catherine Weiner, Abigail Groff, and Julia Rogers for thoughtful comments on the manuscript. We thank Lucas Janson and Kian Hong Kock for helpful discussions throughout the project. M.M. is a Gilead Fellow of the Life Sciences Research Foundation. K.M. is a National Science Foundation Graduate Research Fellow under grant no. DGE1144152. J.C.L. holds a Wellcome Trust Intermediate Clinical Fellowship (105920/Z/14/Z). J.L.R. is an HHMI faculty scholar. This work was supported by U.S. National Institutes of Health grant P01 GM099117.

Author contributions: K.M., M.M., and J.L.R. designed the project and wrote the manuscript. K.M. and M.M. designed the oligonucleotide libraries. K.M. and M.M. performed all MPRA analyses. P.-J.V. performed genome-wide ChIP-seq analysis and the MPRA model. C.G. and P.G.M. performed the MPRA experiments. J.C.L. contributed to the project design. All authors have read and approved the manuscript for publication.

References

Andersson R. 2015. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *Bioessays* **37:** 314–323. doi:10.1002/bies.201400162

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507:** 455–461. doi:10.1038/nature12787

Arnold CD, Zabidi MA, Pagani M, Rath M, Schernhuber K, Kazmar T, Stark A. 2017. Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat Biotechnol* **35:** 136–144. doi:10.1038/nbt.3739

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25:** 1915–1927. doi:10.1101/gad.17446611

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848. doi:10.1126/science.1162228

Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. 2014. The functional consequences of variation in transcription factor binding. *PLoS Genet* **10:** e1004226. doi:10.1371/journal.pgen.1004226

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22:** 1775–1789. doi:10.1101/gr.132159.111

Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES. 2016. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539:** 452–455. doi:10.1038/nature20149

The FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* **507:** 462–470. doi:10.1038/nature13182

Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, et al. 2017. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45:** D777–D783. doi:10.1093/nar/gkw1121

Goes FS, McGrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I, Nestadt G, Kenny EE, Vacic V, Peters I, et al. 2015. Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am J Med Genet B Neuropsychiatr Genet* **168:** 649–659. doi:10.1002/ajmg.b.32349

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018. doi:10.1093/bioinformatics/btr064

The GTEx Consortium. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348:** 648–660. doi:10.1126/science.1262110

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774. doi:10.1101/gr.135350.111

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34:** D590–D598. doi:10.1093/nar/gkj144

Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. 2017. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543:** 199–204. doi:10.1038/nature21374

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2017. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* **18:** 205–214. doi:10.1093/bib/bbw008

Kugel JF, Goodrich JA. 2017. Finding the start site: redefining the human initiator element. *Genes Dev* **31:** 1–2. doi:10.1101/gad.295980.117

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* **49:** 1731–1740. doi:10.1038/ng.3988

Lambert SA, Albu M, Hughes TR, Najafabadi HS. 2016. Motif comparison based on similarity of binding affinity profiles. *Bioinformatics* **32:** btw489. doi:10.1093/bioinformatics/btw489

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27:** i275–i282. doi:10.1093/bioinformatics/btr209

Liu Y, Corcoran M, Rasool O, Ivanova G, Ibbotson R, Grandér D, Iyengar A, Baranova A, Kashuba V, Merup M, et al. 1997. Cloning of two candidate tumor suppressor genes within a 10 kb region on chromosome 13q14,

frequently deleted in chronic lymphocytic leukemia. *Oncogene* **15:** 2463–2473. doi:10.1038/sj.onc.1201643

Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, et al. 2015. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47:** 979–986. doi:10.1038/ng.3359

Mariani L, Weinand K, Vedenko A, Barrera LA, Bulyk ML. 2017. Identification of human lineage-specific transcriptional coregulators enabled by a glossary of binding modules and tunable genomic backgrounds. *Cell Syst* **5:** 187–201.e7. doi:10.1016/j.cels.2017.06.015

Marques AC, Hughes J, Graham B, Kowalczyk MS, Higgs DR, Ponting CP. 2013. Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biol* **14:** R131. doi:10.1186/gb-2013-14-11-r131

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, et al. 2014. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42:** D142–D147. doi:10.1093/nar/gkt997

Medina-Rivera A, Santiago-Algarra D, Puthier D, Spicuglia S. 2018. Widespread enhancer activity from core promoters. *Trends Biochem Sci* **43:** 452–468. doi:10.1016/j.tibs.2018.03.004

Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, Zhu M, Wu J, Shi X, Taing L, et al. 2017. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res* **45:** D658–D662. doi:10.1093/nar/gkw983

Melé M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. 2017. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res* **27:** 27–37. doi:10.1101/gr.214205.116

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30:** 271–277. doi:10.1038/nbt.2137

Molyneaux BJ, Goff LA, Brettler AC, Chen HH, Brown JR, Hrvatin S, Rinn JL, Arlotta P. 2015. DeCoN: genome-wide analysis of in vivo transcriptional dynamics during pyramidal neuron fate selection in neocortex. *Neuron* **85:** 275–288. doi:10.1016/j.neuron.2014.12.024

Nguyen TA, Jones RD, Snavely AR, Pfenning AR, Kirchner R, Hemberg M, Gray JM. 2016. High-throughput functional comparison of promoter and enhancer activities. *Genome Res* **26:** 1023–1033. doi:10.1101/gr.204834.116

Ørom UA, Shiekhattar R. 2013. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* **154:** 1190–1193. doi:10.1016/j.cell.2013.08.028

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27:** 1173–1175. doi:10.1038/nbt.1589

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, et al. 2012. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat Biotechnol* **30:** 265–270. doi:10.1038/nbt.2136

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17:** 556–565. doi:10.1101/gr.6036807

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

R Core Team. 2018. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Ann Rev Biochem* **81:** 145–166. doi:10.1146/annurev-biochem-051410-092902

Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, et al. 2013. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci* **110:** 2876–2881. doi:10.1073/pnas.1221904110

Suliman HB, Sweeney TE, Withers CM, Piantadosi CA. 2010. Co-regulation of nuclear respiratory factor-1 by NFκB and CREB links LPS-induced inflammation to mitochondrial biogenesis. *J Cell Sci* **123:** 2565–2575. doi:10.1242/jcs.064089

Tabach Y, Brosh R, Buganim Y, Reiner A, Zuk O, Yitzhaky A, Koudritsky M, Rotter V, Domany E. 2007. Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PLoS One* **2:** e807. doi:10.1371/journal.pone.0000807

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. 2004. An abundance of bidirectional promoters in the human genome. *Genome Res* **14:** 62–66. doi:10.1101/gr.1982804

Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, Qureshi M, Dong Q, Gu X, Chen WV, et al. 2008. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* **40:** 1407–1409. doi:10.1038/ng.273

Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5:** 276–287. doi:10.1038/nrg1315

Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, et al. 2013. Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45:** 1274–1283. doi:10.1038/ng.2797

Received July 29, 2018; accepted in revised form January 17, 2019.