Published in final edited form as:

Nat Biotechnol. 2017 February ; 35(2): 145–153. doi:10.1038/nbt.3754.

Genome-wide mapping of autonomous promoter activity in human cells

Joris van Arensbergen^{1,*}, Vincent D. FitzPatrick², Marcel de Haas¹, Ludo Pagie¹, Jasper Sluimer¹, Harmen J. Bussemaker^{2,*}, and Bas van Steensel^{1,*}

¹Division of Gene Regulation, Netherlands Cancer Institute, 1016 HM Amsterdam, the Netherlands ²Department of Biological Sciences, Columbia University, New York, NY 10027, USA

Abstract

Previous methods to systematically characterize sequence-intrinsic activity of promoters have been limited by relatively low throughput and the length of sequences that could be tested. Here we present Survey of Regulatory Elements (SuRE), a method to assay more than 10⁸ DNA fragments, each 0.2–2kb in size, for their ability to drive transcription autonomously. In SuRE, a plasmid library is constructed of random genomic fragments upstream of a 20bp barcode and decoded by paired-end sequencing. This library is then transfected into cells and transcribed barcodes are quantified in the RNA by high throughput sequencing. When applied to the human genome, we achieved a 55-fold genome coverage, allowing us to map autonomous promoter activity genome-wide. By computational modeling we delineated subregions within promoters that are relevant for their activity. For instance, we show that antisense promoter transcription is generally dependent on the sense core promoter sequences, and that most enhancers and several families of repetitive elements act as autonomous transcription initiation sites.

INTRODUCTION

Promoters harbor the transcription start site (TSS) and various other sequences that control transcription initiation through the binding of trans-acting factors¹. Various genome-wide methods have been developed to map endogenous promoter activity^{2–5}. These methods have identified tens of thousands of human promoters, often at nucleotide resolution, and have provided estimates of their relative activity in many cell types. A limitation of these maps is that they provide information about where the promoters are located, but not how their

*Correspondence: j.v.arenbergen@nki.nl (JvA); hjb2004@columbia.edu (HJB); b.v.steensel@nki.nl (BvS).

Accession Codes

SuRE data sets are available at the Gene Expression Omnibus, accession GSE78709.

Author Contributions

J.v.A. Conceived and developed the SuRE assay, designed and performed experiments, analyzed data, wrote the manuscript. V.D.F. Developed algorithms, analyzed data and wrote the manuscript. L.P. Developed algorithms and analyzed data. M.d.H. Performed experiments. J.S. Performed experiments. H.J.B. Developed algorithms, analyzed data and wrote the manuscript. B.v.S. Designed experiments, analyzed data and wrote the manuscript.

Competing Financial Interests Statement

The authors declare no competing financial interest.

activity is controlled. Proximal sequences, distal enhancers, local chromatin context, and 3D conformation of the genome may all contribute to promoter activity. There is currently no estimate of the relative importance of these factors. Large-scale perturbative approaches are needed to tackle this problem systematically.

One important perturbation strategy is to take sequence elements out of their native context, to separate regulatory activities that are intrinsic to the underlying sequence from those that are extrinsic to it. Several highly multiplexed reporter assays have been developed for this purpose. One class of methods combines random barcodes located in the transcription unit with synthetic upstream promoter or enhancer sequences^{6–12}. This approach is particularly suited to systematic mutagenesis of selected regulatory elements; however, both the length of the tested elements (~150bp) and the level of multiplexing ($10^4 - 10^5$) are limited by DNA synthesis technology. A variant approach uses mutagenized or randomly assembled small enhancer fragments of up to several hundreds of basepairs^{13–15}, also with a multiplexing level between 10^4 and 10^5 . A complementary strategy that uses shotgun cloning into a reporter plasmid was used to screen several hundreds of kilobases of genomic DNA for enhancer activity in mouse cells¹⁶. Furthermore, a cell-sorting strategy was used to screen nearly 10^5 random DNA fragments from nucleosome-depleted regions (which are likely to contain enhancers and promoters) for regulatory activity in mouse cells¹⁷. At substantially higher throughput, near-saturating coverage of the entire *Drosophila* genome was achieved with STARR-seq^{18, 19}. However, this approach is only suitable to detect enhancer activity and not promoter activity. Moreover, like all other methods reported so far, it has not been applied on a scale sufficient to cover entire mammalian genomes.

Here, we present Survey of Regulatory Elements (SuRE), a method that overcomes some of these limitations. Instead of short synthetic promoter sequences, SuRE queries random genomic fragments in the size range of 0.2–2kb, which is long enough to include most elements that constitute fully functional promoters. Moreover, with SuRE it is possible to achieve a throughput of $>10^8$ fragments, which is sufficient to redundantly scan the entire human genome at an average base coverage of ~55-fold.

We demonstrate the feasibility of this approach in cultured human cells. SuRE data can be interpreted as maps of promoter “autonomy”, i.e., the degree to which sequences across the genome can act as promoters in the absence of other regulatory elements. Additionally, because each promoter is represented by many partially overlapping random fragments, it is possible to delineate the regions that contribute to its activity. We present a computational strategy for this purpose. The SuRE maps provide unique opportunities to gain new insights into the biology of human promoters and enhancers.

RESULTS

SuRE method and library preparation

The SuRE experimental strategy consists of three main steps (Fig. 1a, Supplementary Fig. 1). First, genomic DNA is randomly fragmented and subjected to size selection to obtain 0.2–2kb long fragments. These are ligated *en masse* into a plasmid, immediately upstream of a promoter-less transcription unit that contains a random 20 bp barcode near its 5′ end.

High-throughput paired-end sequencing of the resulting library associates each barcode with the genomic start and end positions and orientation of the corresponding fragment (Supplementary Fig. 1). Finally, the library is transiently transfected into cultured cells, where the vast majority of plasmids remains episomal and hence is not subject to chromosomal position effects.

Only fragments that contain a functional promoter will drive transcription into barcoded mRNA. These barcodes are counted after reverse transcription, PCR amplification, and high-throughput sequencing. Using the barcode-to-fragment table, a genome-wide map of promoter activity can then be constructed (Fig. 1a). We define activity detected in this way as *autonomous* promoter activity, because the reporter plasmid does not contain a promoter nor any other regulatory elements.

We generated a human SuRE plasmid library with an estimated complexity of ~270 million unique genomic fragments. Of these, we were able to map ~150 million to their barcode, resulting in a 55-fold coverage of the human genome on average (Supplementary Fig. 2a), with 96% of the mappable genome covered at least 15-fold (Supplementary Fig. 2a).

Genome-wide map of autonomous promoter activity in human cells

We transiently transfected the library into human K562 erythroleukemia cells. Two independent replicate experiments cumulatively yielded 111,851,687 SuRE reads across 26,501,576 distinct barcodes. More technical details about the data are provided in Supplementary Fig. 2.

As expected, the resulting SuRE activity map shows a pattern of peaks that overlap frequently with known transcription start sites and histone modifications marking active promoters, such as H3K4me3 and H3K27ac (Fig. 1b). A peak detection algorithm²⁰ identified 55,453 SuRE peaks at an estimated false discovery rate of 5% and with at least 2-fold enrichment of SuRE signal over background (Supplementary Dataset 1). SuRE activity is enriched in previously annotated active promoters, and to a lesser degree in enhancers and certain repetitive elements (see below), but depleted from repressed ('Repr') or quiescent ('Quies') parts of the genome²¹ (Fig. 1c, Supplementary Fig. 2c). Promoters and enhancers together explain 26% of the SuRE peaks (see below).

To verify these results, we repeated the SuRE experiments with a focused library derived from 9 selected regions of the human genome²² (Table S1), together spanning 1.3 Mb. This library had an average 212-fold coverage of the included base-pairs. Due to its lower complexity and higher coverage it yielded highly reproducible results (Pearson's $r = 0.99$; Figure S3a). Within the regions probed by this focused library, 45 out of 50 peaks (90%) previously identified in the genome-wide SuRE dataset also showed enriched signals in the focused SuRE dataset (Figure S3b). Similarly, out of 55 TSSs with a genome-wide SuRE enrichment of at least 2-fold, 53 (96%) showed enriched signals in the focused SuRE dataset (Figure S3c). This indicates that the false discovery rate of genome-wide SuRE peaks is low. Finally, for 23 promoters we compared SuRE peak heights to signals obtained by conventional reporter assays with individually cloned constructs. This showed an overall $r^2 = 0.73$ (Supplementary Fig. 3d).

Autonomous promoter activity explains a large fraction of gene expression

To determine to what extent the autonomous activity of known promoters correlates with their endogenous activity we compared the genome-wide SuRE map to levels of engaged RNA polymerases just downstream of TSSs, as determined by the GRO-cap method⁵. We focused on a curated set (see **Methods**) of 28,844 TSSs annotated by the GENCODE project²³. Notably, SuRE and GRO-cap signals are substantially correlated ($r^2 = 0.54$; Fig. 1d). Similar results were obtained when only comparing TSSs which showed expression in both SuRE and GRO-cap ($r^2 = 0.43$), and in a comparison with transcription activity detected by the CAGE method²¹ ($r^2 = 0.49$; Supplementary Fig. 2d). Thus, a substantial part of promoter activity is reproduced by sequence elements <2kb from the TSS, i.e., in the absence of distal enhancers, chromatin context and 3D organization.

Promoters of widely expressed (“housekeeping”) genes typically show more relative autonomy (i.e., SuRE signal divided by GRO-cap signal) than those of cell-type specific genes (Fig. 1e). Yet, we also identify many housekeeping promoters with a low level of promoter autonomy, for example promoters of genes that encode histones (Fig. S2e). Relative promoter autonomy is inversely correlated with the number of enhancers near the promoters in the native genomic context (Fig. 1f). This cannot be explained by differences in local gene density (Fig. S2f). These results support the notion that autonomous promoters as detected by SuRE are less dependent on distal enhancers than non-autonomous promoters.

Divergent transcription is generally autonomous

Endogenously, most human promoters drive divergent transcription, with stable transcripts produced in the sense orientation and unstable short transcripts originating upstream in the antisense orientation³. We expected that in SuRE this antisense transcription might be detected if a promoter is inserted in reverse orientation, as the transcript would be stabilized by the plasmid-encoded transcription unit. Indeed, SuRE detects extensive activity of promoters in the antisense direction (Fig. 2a–c). The antisense activity is on average 2–3 fold weaker but it correlates with the sense activity (Fig. 2b–d; $r^2 = 0.48$). We conclude that divergent transcription initiation is generally an autonomous feature of human promoters, and can be assayed by SuRE.

Delineation of promoter regions that drive autonomous transcription

In SuRE, each promoter is represented by a series of partially overlapping fragments with different sizes and different start and end positions. This offers the opportunity to identify critical sequence regions. For example, around the promoter of *NUP214*, multiple fragments that only include ~100 bp upstream of the annotated sense TSS show high SuRE signals (Fig. 3a), indicating that this region together with the TSS is sufficient to drive transcription autonomously. For a more quantitative analysis, we developed a generalized linear modeling (GLM) method based on Poisson statistics, which effectively deconvolves the SuRE data and identifies the promoter subregions that contribute most to the genome-wide autonomous transcription activity (see **Methods**). When applied to *NUP214*, this confirms that the proximal ~100 bp upstream of the TSS is primarily responsible for its autonomous activity (Fig. 3b).

To understand which parts of human promoters are generally required for optimal autonomous transcription, we aggregated SuRE data according to the start and end positions of each query fragment relative to the nearest TSS (Fig. 3c, **top triangle**). This shows that, as expected, most activity is contributed by the core promoter and sequences within a few hundred bp upstream; inclusion of longer upstream regions on average does not increase reporter activity. Increasing the length of the sequence included downstream of the TSS tends to reduce reporter activity, which may in part be due to the inclusion of splice sites (Supplementary Fig. 2j) or other elements that are not compatible with the reporter design. Application of GLM to all promoters combined yielded a similar conclusion: significant contributions to sense transcription are primarily provided by the core promoter region itself and sequences up to ~200 bp upstream (Fig. 3d, **blue curve**). These analyses illustrate how SuRE data can be used to identify critical sequence regions within promoters, both individually and genome-wide.

Requirements for autonomous antisense transcription

Sequence motif analysis of antisense TSS regions has suggested the presence of an independent antisense core promoter which may be responsible for antisense transcription^{5, 24, 25}. Indeed, two antisense core promoters were found to drive transcription autonomously *in vitro*²⁴. On the other hand, the sense and antisense core promoters have been proposed to function in a cooperative manner²⁵. To date, the functional interdependence of the sense and antisense core promoters has not been addressed through systematic deletion experiments. We therefore used the randomly overlapping fragment information as illustrated above to gain insight into the requirements for antisense transcription.

Virtually all *NUP214* fragments that show antisense SuRE activity extend at least ~200bp to include the annotated sense TSS, suggesting that the sense core promoter (here defined as -50 to +50 bp relative to the annotated TSS) is critical for antisense transcription (Fig. 3e). GLM confirmed this conclusion and found no evidence that the antisense TSS subregion is needed for antisense transcription (Fig. 3f). Indeed, genome-wide analysis shows that promoter fragments that include the forward core promoter generally exhibit the highest SuRE activity in antisense orientation (Fig. 3c, **bottom triangle**). GLM applied to all promoters combined also indicated that antisense transcription is dependent on essentially the same sequence region (including the sense core promoter) as sense transcription (Fig. 3d, **red curve**). Analysis of a well-defined set of sense-antisense TSS pairs⁵ (Fig. 3g) underscores this general conclusion.

Inspection of raw SuRE data and GLM profiles of individual promoters covered by our focused library revealed several interesting examples and exceptions to this general trend. For example, transcription from both the main sense and main antisense TSSs of *SLC50A1* requires the same subregion located between them; however, an alternative sense TSS upstream and an additional antisense TSS downstream appear to be non-autonomous, because no GLM signal is detectable at these sites (Fig. 3h). In the *WDR47* gene, antisense transcription does not require the antisense TSS subregion, but rather depends on a subregion that is also the primary driver of sense transcription, thus representing an example

of the general trend (Fig. 3i). Finally, the sense and antisense TSSs at the *HIST1H2BD* gene are each primarily driven by distinct local sequence elements (Fig. 3j). Thus, exceptions exist to the general rule that antisense transcription is driven by sequence subregions nearby the sense TSS.

Relationship between CpG content and autonomous promoter activity

Promoter regions in mammalian genomes often contain CpG islands, regions that have a relatively high ratio between the observed CpG dinucleotide density and the expected density, given the local C+G content²⁶. CpG content has previously been linked to promoter activity^{12, 27}. When binned by their observed and expected CpG density (Fig. 4a), SuRE fragments around TSSs form two distinct populations that can be separated by a ~50% observed/expected CpG ratio, consistent with a previous classification of promoters²⁷. However, the relationship between SuRE expression level and CpG content for individual fragments takes a different form (Fig. 4b): expression is highest when the observed and expected CpG density are equal, and decays gradually with decreasing CpG observed/expected ratio. Notably, this relationship is largely independent of the CpG density *per se* (i.e., the highest expression occurs along the diagonal in Fig. 4b).

This result most likely reflects the evolutionary history of promoters. A low observed/expected CpG ratio is thought to be the result of conversion of methylated cytosine (which primarily occur in CpG dinucleotides) to thymine by deamination²⁸. Our data suggest that autonomous promoters have been protected from this loss, presumably because they have remained consistently hypomethylated in the germline throughout evolution.

Enhancers act as autonomous promoters

In their native context, enhancers can also act as promoters, although the resulting transcripts (termed eRNAs) tend to be unstable^{29, 30}. For a subset of enhancers, stimulus-induced eRNA production precedes mRNA transcription from the target promoters^{31, 32}, suggesting that enhancers may be transcribed independently of their target promoter. On the other hand, significant correlations between physical promoter–enhancer interactions and the production of eRNAs have been reported^{30, 31, 33} and it has been shown that enhancer transcription can be dependent on the presence of the target promoter³⁴. We therefore used our SuRE data to investigate to what degree transcription initiation from enhancers is autonomous. The locus control region (LCR) of the β -globin gene cluster, a potent multi-enhancer region³⁵, showed several clear bi-directional SuRE signals (Fig. 5a). Analysis of 47,020 predicted active enhancers in K562 cells²¹ revealed SuRE signals for the majority (Fig. 5b, c), although the overall level of activity is approximately 10-fold lower than for promoters (cf. Fig. 3a and Fig. 5c; Fig. 5d). We conclude that eRNA production is generally autonomous, i.e. it generally does not require interactions of the enhancer with its target promoter *in cis*. We cannot rule out that the transfected plasmids interact with their target promoters *in trans*³⁶.

Notably, the ENCODE classification of enhancers as ‘weak’ or ‘strong’²¹ correlates with the strength of SuRE signals ($p < 2.2 \times 10^{-16}$, Wilcoxon test) (Fig. 5b–d). SuRE signal also correlates positively with the endogenous levels of H3K27ac (Fig. 5e), the histone modification most characteristic of active enhancers³⁷. Furthermore, the ability of ~130 bp

fragments derived from ENCODE-annotated enhancers to activate a minimal promoter in a previous reporter assay³⁸ shows a significant ($p = 4 \times 10^{-4}$) positive correlation with the SuRE signal for the same enhancers (Fig. 5f). These results indicate that the level of autonomous transcription initiation from enhancers is related to enhancer strength.

Dissection of regulatory element interplay in the alpha-globin LCR

To further illustrate the value of SuRE for dissecting regulatory mechanisms, we focused on the alpha-globin locus, which harbors a locus control region that can activate several globin genes over a distance of >50kb. The locus control region contains several separate enhancers known as R1-4. In mouse these enhancers work in an additive manner and no single element is critical for globin expression³⁹. Treatment of K562 cells with hemin is known to increase expression of several of the genes in the alpha-globin locus⁴⁰, which we confirmed by RT-qPCR (Figure 5g). Although R2 can be activated by hemin⁴¹, it is not known whether other elements in the region contribute to the response to hemin. Comparison of SuRE profiles obtained from hemin-treated and control cells (Figure 5h) revealed that R2 was exclusively activated by hemin. This indicates that activation of the three genes occurs selectively via elevated activity of enhancer R2, without contributions of any of the other enhancer or promoter sequences. This example illustrates how SuRE may be used to identify key elements in dynamic regulatory mechanisms.

Autonomous promoter activity in repetitive elements

ENCODE-annotated promoters and enhancers in K562 account for only 26% of the genome-wide SuRE peaks (Supplementary Fig. 2i). Several families of repetitive elements show significant ($p < 0.01$ after multiple testing correction) overlap with SuRE peaks, in particular the ERVL-MaLR and ERV1 retrotransposons (Fig. 6a), which account for another 19% of the peaks. Certain subfamilies within these families exhibit specific and high SuRE signals, for example the LTR12C subfamily of solitary long terminal repeats (Fig. 6b, c). For some repeat subfamilies (e.g., LTR12C) the average SuRE activity resembles that of promoters in terms of strength and directional bias, whereas for others (e.g., MER41B) the relatively weak signal and the balanced bidirectional activity are more reminiscent of enhancers (Fig. 6b, c).

Note that technologies like CAGE and GRO-cap have difficulty mapping transcription initiation activity uniquely to specific repeat instances in the genome⁴², whereas SuRE maps are based on paired-end sequencing reads that generally include unique sequences flanking the repeat instances, yielding a much more detailed map of promoter activity in repetitive regions. For example, autonomous promoter activity could be unambiguously assigned to a LTR12C insertion in the β -globin locus (Fig. 5a). In addition, GLM analysis of partially overlapping SuRE fragments, similar to what we applied to promoters (cf. Fig. 3d), pinpointed the precise sequence regions that generally contribute to autonomous promoter activity across hundreds of LTR12C variants (Fig. 6d). These data extend earlier analyses of single LTRs⁴³ and again indicate that essentially the same sequence elements contribute to sense and antisense transcription.

Sense-oriented run-on transcription⁵ is detectable downstream of LTR12C insertions with high SuRE activity (Fig. 6e). This is not found for insertions with low SuRE activity and not in the antisense direction (Supplementary Fig. 4). This indicates that the autonomously active LTR12 copies drive downstream intergenic transcription in their endogenous context and may produce long non-coding RNAs.

Non-annotated SuRE peaks may be cryptic promoters

Of the 55,453 SuRE peaks, only 45% are accounted for by ENCODE-annotated promoters and enhancers or ERVL-MaLR and ERV1 retrotransposons. Of the 30,548 remaining ‘unexplained’ peaks only 15% overlap with a TSS or enhancer annotated in one of 889 cell sources assayed by the FANTOM project. The unexplained peaks however do show enrichment for epigenetic marks of promoter activity, such as H3K4me3 or DNase I hypersensitivity (Supplementary Fig 5a, b). Their average SuRE signal is substantially above background, while they produce almost no GRO-cap signal (Supplementary Fig 5c, d). These peaks may thus represent cryptic promoters that fail to initiate transcription in the native chromatin setting. One function of chromatin may be to suppress such cryptic promoter activity.

DISCUSSION

These results establish SuRE as a high-throughput tool to functionally deconstruct large genomes and systematically identify elements that drive autonomous transcription activity. SuRE stands out from previous high-throughput promoter assays by its 100–1,000 fold larger scale, sufficient to survey the entire human genome at >50x coverage. Furthermore, the partial overlap of the query fragments makes it possible to use the SuRE data as a massive “promoter truncation” experiment and delineate the minimal regions required for autonomous activity, both for individual promoters and genome-wide.

Our GLM approach, which enhances the spatial resolution of SuRE by an order of magnitude, indicates that sequence elements that contribute to promoter autonomy are generally concentrated in regions <200 bp upstream of the TSS. The high density of regulatory information proximal to the TSS is in line with the findings in yeast and *Drosophila*^{10, 19}. Specific promoters may require additional elements further upstream; it is a matter of definition whether such elements should be considered as part of the promoter or as proximal enhancer elements.

With a minor modification of the reporter design (Supplementary Fig. 6) SuRE should also be suited to survey the entire human genome specifically for functional enhancer activity (i.e., the ability of genomic fragments to activate a cis-linked minimal promoter) with a similar throughput and coverage as described here. In conjunction with complementary functional genomics strategies^{6–10, 12–18, 44} this will help dissect the sequence determinants of promoter and enhancer activity, and unravel the complex interplay of the possibly more than one million regulatory elements in the human genome²¹.

ONLINE METHODS

SuRE library preparation

The SuRE vector was constructed using standard molecular biology techniques. It is based on a pSMART backbone (Addgene plasmid # 49157; a gift from James Thomson) and contains a green fluorescent protein (GFP) open reading frame followed by a SV40 derived polyadenylation signal (PAS). To generate a barcoded SuRE vector library, 30 µg SuRE vector was digested with NheI (#R0131; NEB) and XcmI (#R0533; NEB) and a gel extraction was performed on the vector. Barcodes were generated by performing 10 PCR reactions of 100 µl each containing 5 µl 10 µM primer 256JvA, 5 µl 10 µM primer 264JvA and 1 µl 0.1 µM template 254JvA (see Supplementary Table 2 for oligonucleotide sequences). A total of 14 PCR cycles (1' at 96 °C, 14x(20" at 96 °C, 20" at 60 °C, 20" at 72 °C), hold at 10 °C) were performed using MyTaq™ Red Mix (#BIO-25043; Bioline), yielding ~30 µg barcodes. Barcodes were purified by phenol-chloroform extraction and isopropanol precipitation, digested overnight with 80 units AvrII (#ER1561; Thermo Fischer) and purified using magnetic beads (#AC60050; GC Biotech). Vector and barcodes were then ligated in 3 reactions of 100 µl with each containing 5 µg digested SuRE vector and 5 µg digested barcodes, 20 units NheI (#R0131S; NEB), 20 units AvrII, 10 µl of 10x CutSmart buffer, 10 µl of 10mM ATP, 10 units T4 DNA ligase (#10799009001 Roche). A cycle- ligation of 6 cycles was performed (10' at 22 °C and 10' at 37 °C), followed by 20' heat- inactivation at 80 °C. The ligation reaction was purified by magnetic beads and digested with 40 units of XcmI (#R0533S; NEB) for 3 hours, and size selected by gel-extraction, yielding 5–10 µg barcoded SuRE vector.

To insert genomic DNA into the barcoded vector, DNA was isolated from 40 million K562 cells and 250 µg was fragmented using NEBNext® dsDNA Fragmentase (#M0348; NEB), size selected (0.5–2kb) using gel-extraction (#11696505001; Roche), repaired using End-It™ DNA End-Repair Kit (#ER0720; Epicentre) and A-tailed using Klenow HC 3->5 exo⁻ (#M0212L; NEB). We also obtain many smaller elements in the final library (Supplementary Fig. 2b) presumably because size-selection is imperfect and smaller fragments preferentially contribute to the final plasmid library. Five µg of A-tailed genomic fragments were ligated with 5 µg barcoded SuRE vector in a 600 µl reaction using the Takara ligation kit v1.0 (#6021; Takara). The ligation product was purified by phenol-chloroform extraction and isopropanol precipitation and then digested in a 600 µl reaction with 60 units of Plasmid-Safe™ ATP-Dependent DNase (# E3101K; Epicentre) for 3 hours to digest away any non-ligated vector, again purified by phenol-chloroform extraction and isopropanol precipitation, taken up in 20 µl water, purified with magnetic beads and taken up in 20 µl water. This material was then electroporated into CloneCatcher DH5G electrocompetent *E. Coli* (#C810111; Genlantis) in 4 separate electroporations with each 5µl ligation product and 20µl bacteria, each transferred to 500 ml standard Luria Broth (LB) plus kanamycin (50µg/ml), grown overnight and together purified using a GIGA plasmid purification kit (#10091; Qiagen), yielding ~10 mg of SuRE library. The choice of plasmid backbone and bacteria used for expanding the plasmid pool were key to obtaining a highly complex library with low bias in A/T content. This allowed us to achieve a sufficiently homogenous representation of the genome. This protocol takes an experienced person about 5 days to

complete. Day 1: preparation of vector and barcodes; Day 2: ligation of barcodes onto vector, genomic DNA isolation and fragmentation; Day 3: genomic DNA size-selection, repair and A-tailing, ON ligation of barcoded vectors and A-tailed inserts; Day 4: Purification of ligation product and electroporation of library; Day 5: GIGA plasmid purification. The typical yield of ~ 10 µg can be used for 50 transfections on 100 million cells.

Focused SuRE library

In addition to the above genome-wide SuRE library, we also generated a library from 9 pooled Bacterial Artificial Chromosomes (BACs), collectively covering 1.3 Mb of the human genome (Supplementary Table. 1). This library was prepared essentially the same as the genome-wide library except that size-selection was performed for elements of 0.1kb–1kb and that only 100 ng barcoded vector was used with 100 ng of size-selected BAC inserts. The ligation product was phenol-chloroform purified, isopropanol precipitated and taken up in 16 µl water. Four µl was electroporated into 20 µl bacteria and transferred to 250 ml LB plus kanamycin (50 µg/ml). This yielded an approximate library complexity of ~3 million unique clones and we mapped ~25% of these elements to their barcode, as the library was somewhat under-sequenced.

SuRE library characterization by iPCR (barcode-to-fragment association; Supplementary Fig. 1)

To associate the barcodes with the linked genomic fragments, we digested 4 µg SuRE library with I-CeuI (#R0699S; NEB), followed by magnetic bead purification (1:1 ratio beads:DNA solution). Of this, 2 µg was self-ligated overnight at 16 °C in a total volume of 2 ml (#10799009001; Roche), and purified using phenol-chloroform extraction and isopropanol precipitation. To reduce the size of the genomic fragments this material was digested for 1 hour with 10 units of frequent cutter Nla III (#R0125S; NEB) or 10 units of HpyCH4V (#R0620L; NEB), bead purified and self-ligated again in a final volume of 1 ml. This material was purified by phenol-chloroform extraction and isopropanol precipitation, treated with 25 units of Plasmid-Safe™ ATP-Dependent DNase for 1 hour and purified again with phenol-chloroform and isopropanol precipitation. To facilitate PCR, the resulting mini-circles were linearized by digesting with I-SceI (#R0694S; NEB) in a volume of 25 µl. Finally, 10 cycles of PCR (1' at 98°C, 10x(15" at 98°, 15" at 60°, 20" at 72°)) with Phusion high-fidelity DNA Polymerase (#M0530L; NEB) were performed on 2.5 µl of the I-SceI digested material using primers 151AR (containing the S1 and p5 adapter) and (index variants of) 117JvA (containing the S2, index and p7 adapter). The PCR product was bead purified and subjected to high-throughput paired-end sequencing on an Illumina MiSeq, HiSeq2000 or HiSeq2500.

Cell culture and transfection

K562 (ATCC® CCL-243™) were cultured according to supplier's protocol. Every 3 months all cells in culture were screened for Mycoplasma using PCR (Takara; # 6601). Cells were transiently transfected using Amaxa Nucleofector II, program T-016 and nucleofection buffer as published previously. For K562, 2 biological replicates were done of each 100 million cells (5 million per cuvette with each 10 µg plasmid) and harvested after 24 hours

(see below). For the focused library experiments, 2 biological replicates of each 10 million cells were done per condition (standard, hemin, solvent control). In the hemin treatment experiment, treatment was started with 50 μ M hemin (Sigma; #51280-1G) or solvent control 1 hour after nucleofection and cells were harvested 24 hours later.

RNA extraction and reverse transcription

RNA was isolated using Trisure (#BIO-38032; Bioline) and polyA RNA was purified using Oligotex from Qiagen (#70022; Qiagen). PolyA RNA was divided into 10 μ l reactions containing 500 ng RNA and treated with 10 units DNase I for 30 minutes (#04716728001; Roche) and DNase I was inactivated by addition of 1 μ l 25mM EDTA and incubation at 70°C for 10 minutes. Next, cDNA was produced by first adding 1 μ l of 10 μ M gene specific primer targeting the GFP ORF (247JvA) and 1 μ l dNTP (10mM each) and incubating for 5 minutes at 65°C. Then 4 μ l of RT buffer, 20 units RNase inhibitor (#EO0381; ThermoFisher Scientific), 200 units of Maxima reverse transcriptase (#EP0743; ThermoFisher Scientific) and 2.5 μ l water was added and the reaction mix was incubated for 30 minutes at 50°C followed by heat-inactivation at 85° for 5 minutes. Per biological replicate of the genome-wide library, 20–30 reactions were done in parallel. For the focused library, 4 reactions were done in parallel per biological replicate. Each 20 μ l reaction was then PCR amplified (1' 96 °C, 20x(15" 96 °C, 15" 60 °C, 15" 72 °C)) in a 100 μ l reaction with MyTaq™ Red Mix and primers 151AR (containing the S1 and p5 adapter) and (index variants of) 211JvA (containing the S2, index and p7 adapter) for 21 cycles. Reactions were then pooled and 500 μ l was purified using a PCR purification kit (#BIO-52060; Bioline) and then size-selected using e-gel (#G6400EU; Invitrogen) and subjected to single read 50 bp high throughput sequencing on an Illumina HiSeq2000 or HiSeq2500.

Mapping of iPCR sequencing data

Paired-end reads are trimmed, using cutadapt (version 1.2.1) {cutadapt: <http://journal.embnet.org/index.php/embnetjournal/article/view/200>}, to remove the adapter sequences from the forward (CCTAGCTAACTATAACGGTCCTAAGGTAGCGAACCAGTGAT) and the reverse reads (CCAGTCGT). The remaining read sequences are then trimmed from the first occurring NlaIII/HpyCH4V restriction site (CATG/TGCA) onward. Trimmed reads with length < 6 bp were removed from further processing. Next, reads were aligned to the human genome reference sequence (hg19, including only chr1-22, chrX, chrY, chrM) using Bowtie2 (version 2.1.0)⁴⁸, with a maximum insert length set to 4kb. All read pairs not aligned as 'proper pair' were excluded from further processing. The resulting bam files were converted to bedpe files using custom scripts.

SuRE normalization

Data were processed using custom R scripts (<https://www.R-project.org>). To normalize SuRE expression data, we first characterized the barcode frequencies in the plasmid library. More specifically, we digested 1 μ g library with I-SceI (#R0694S; NEB) in 25 μ l to linearize the plasmids, then performed 2 replicate PCRs each on 2 μ l I-SceI digested material, using the same protocol as for the cDNA but for 8 cycles. Because of the high complexity of the library (~270 million) the aim was not to get a quantitative readout for each barcode, but

rather to identify potentially over-represented barcodes and/or regions of the genome and normalize for that (see below for validation). The PCR product was e-gel size-selected and subjected to single-read 50 bp high throughput sequencing on an Illumina HiSeq2500 or HiSeq2000.

In total we obtained ~40 million reads per PCR replicate. From these reads barcode counts were determined using cutadapt version 1.2.1 (<http://journal.embnnet.org/index.php/embnnetjournal/article/view/200>) to remove the adapter (GCTAGCTAACTATAACGGTCCTAAGGTAGCGAA) from the sequence. To determine genome-wide input coverage ('input') we took all fragments mapped in the iPCR step, initializing the read count to a pseudo count of 1 for each. The barcode counts determined for the input plasmid libraries were then added to these initial counts.

Raw SuRE expression data was determined by counting barcodes in cDNA, discarding those not identified in the iPCR mapping. Barcodes with identical genomic positions accounted for 5% of the library and mostly corresponded to iPCR barcode read errors; input and cDNA counts for these fragments were aggregated. To obtain SuRE enrichment profiles, cDNA read numbers were normalized (to reads per billion) and genome-wide coverage was calculated and divided by a similarly generated genome-wide input coverage (i.e. 'input' normalized to reads per billion). Throughout the manuscript the combined data from the biological replicates is used unless indicated otherwise. We created BigWig files for the profiles thus obtained using the GenomicRanges package in BioConductor⁴⁹.

Validation using the focused SuRE library

To assess if the sequencing depth for the input is sufficient we sequenced our focused BAC library input deeply and then by down-sampling established that normalization by a deeply sequenced input (average = 10 reads per barcode) gave essentially the same result ($r^2 = 0.98$ for TSSs) as normalization by lowly sequenced input (average = 0.1 read per barcode). This thus strongly suggests there is no large systematic differences in plasmid representation that affect our final results, presumably in large part because of the redundant representation of each part of the genome.

Furthermore to assess systematic transfection biases, we used the focused library and compared pre- and post-transfection plasmid abundances. We find that coverage is highly similar between pre- and post-transfection libraries ($r^2 = 0.98$; Supplementary Fig. 3e). In addition, in neither library there is any correlation between insert length and representation (data not shown), presumably because the typical insert-size (~1000bp) only represents 25% of the total plasmid-size. We conclude that the use of the pre-transfection library as input does not compromise the results.

Post-transfection plasmid extraction

Per replicate, 10 million cells were transfected with our focused SuRE library. After 24 hours, cells were spun down, washed with PBS, spun down again and taken up in 500 μ l nuclear extraction buffer (10mM NaCl, 2mM MgCl, 10mM Tris-HCl (pH 7.8), 5mM DTT, 0.5% NP40). Cells were incubated on ice for 5 minutes, and nuclei were spun down at 7000g and washed twice more with nuclear extraction buffer. The resulting pellet was taken

up in 500 µl miniprep buffer 1 (#BIO-52057) and purified as 2 minipreps according to the manufacturer's protocol. Per replicate 5 µg was digested in 50ul with 2.5ul Sce-1 for 2 hours, heat inactivated at 65°C for 20 minutes and two PCRs with 2.5ul of this material were amplified as described above to characterize the barcode frequencies in the pre-transfection plasmid library. For comparison, 1µg of pre-transfection library was subjected to the same protocol from the Sce-1 digest onwards.

Annotations and data analysis

As a reference for transcription start sites (TSSs) we used GENCODE version 19 TSSs (downloaded from <http://www.genecodegenes.org>). We focused on TSSs located on chr1-22 or chrX. To filter out TSSs based on computational analysis for which no empirical evidence is available, we required them to be identified as being expressed in at least one of the samples assayed in the FANTOM5 phase 1 project⁴⁵. The FANTOM5 phase 1 project profiled RNA expression using CAGE in 889 cell-types, cell-lines and tissues and used these data to identify 184,827 TSSs (intervals representing clusters of mapped 5' ends of mRNAs). This intersection yielded a curated set of 28,844 GENCODE TSSs which we refer to throughout the manuscript.

To assign an expression level to GENCODE TSSs, the BioConductor package CoverageView (version 1.4.0) was used to retrieve the mean SuRE or GRO-cap expression from the respective BigWig files for the interval $\pm$ 500bp around the TSS, using either total expression or expression in the sense orientation as indicated. Thus, where an expression level is assigned to a TSS (i.e. in all density plots and scatter plots) the expression level represents the mean over a 1kb region. Metaprofiles (e.g., Figure 2a) were also generated using CoverageView, using 50bp bins, except for the PRO-seq data in Figure 6e which was generated using 1kb bins because of the sparser nature of the data.

In log-transformed data representations on data-sets that also contain zero's, such as the comparison of GRO-cap and SuRE at GENCODE TSSs in Fig 1d, a pseudo-count of half the minimal non-zero measurement was used to calculate correlations and visualize all values.

We used the FANTOM data to determine the tissue specificity of each TSS. We considered any (center of a) FANTOM phase 1 TSS that fell within 500 bp of the GENCODE TSS, retrieved the number of samples in which each was detected and used the highest (i.e. least tissue specific) number. In the comparison of tissue specificity or proximal enhancers with promoter autonomy only endogenously active promoters (mean GRO-cap > 0.25) which were also detected in SuRE (SuRE >0) were used (n=13,815). In the analysis of the relation between relative promoter autonomy and enhancers, any enhancer (ENCODE state 'Enh') was considered that was within 5–50 kb on either side of the considered TSS and at least 5kb away from any GENCODE TSS.

To assess the spatial profile of contribution to autonomous expression of successive intervals relative to the TSS (Figure 3c), we created a 2D histogram by binning both the start and end position of each SuRE fragment in 100 bp increments. In this analysis, we only included GENCODE TSSs that were expressed in at least one tissue in FANTOM phase 1.

In the analyses of Figure 5b-d only those ENCODE chromatin states were used for which their center was at least 5 kb away from GENCODE TSSs in either direction ('Enh'; n=18257), 'EnhW'; n=28763, 'Quies'; n=36627). Heatmaps in Figure 2b, c and Figure 5b were ordered based on the signal in the full 10 kb interval.

In the comparison of enhancer expression in SuRE with enhancer strength, we used all enhancer elements tested by the authors for which significant activity was found (~20%) using a comparison to a scrambled control³⁸. In addition we required enhancers to be at least 3 kb (rather than 5 kb, in order to have a large enough sample) from a TSS (n=189). For these, we compared enhancer activity (normalized CRE-seq signal using the *Hsp68* minimal promoter) with SuRE activity over a window ± 500 bp from their center. For the single-locus analysis in Figure 3a and e, only genomic fragments are shown that were detected in the cDNA. In Figure S2e histone genes were indicated that contained 'HIST' in their name and to avoid redundancy, alternative TSSs were only plotted if they were at least 500bp from the previous. For Figure S2a we focused on the mappable part of the genome which we obtained by concatenating all adjacent 36-mer mappable regions from ENCODE (wgEncodeCrgMapabilityAlign36mer.bw).

Penalized Generalized Linear Modeling

To create Figures 3b, d, f, h-j and 6d, we used the R package glmnet (<http://CRAN.R-project.org/package=glmnet>) to fit an elastic net Poisson log-linear regression model to SuRE counts, based on a design matrix indicating, for each consecutive 50 bp genomic window, the fraction of bases in that window included in the SuRE fragment. Elastic net combines LASSO regression (penalty on absolute value of the coefficients) and ridge regression (penalty on the square of the coefficients). Together they reduce overfitting of the bin coefficients that can result from the high multicollinearity of adjacent bins. To avoid bias due to the specific choice of bin positions, we performed this fit for all 50 possible ways of positioning the windows relative to the TSS, and then assigned to each base pair in the genome the average of the regression coefficients for all 50 windows containing it, one from each fit, resulting in a smooth curve. Equal ridge and LASSO penalties were used for all regressions ($\alpha = 0.5$). A $\log(\lambda)$ value of 0 was used for NUP214, -1.5 for the BAC, LTR12C and whole-genome regressions. For Figure 3g, we used stable/unstable peak pairs identified in K562 GRO-cap⁵, assigning stable peaks to the sense strand and unstable peaks to the antisense strand. TSS positions correspond to the center of each peak. LTR12C positions were determined via global pairwise alignment of RepeatMasker-annotated genomic LTR12C sequences to the Dfam consensus sequence⁵⁰.

Code availability

Software source codes for SuRE sequencing data processing and Penalized Generalized Linear Modeling are provided as supplementary information.

Data sources

- As a reference for transcription start sites (TSSs) we used GENCODE²³ version 19 TSSs (gencode.v19.annotation.gff3.gz) downloaded from <http://www.gencodegenes.org/releases/19.html>.

- FANTOM phase 1 data⁴⁵ was downloaded from http://fantom.gsc.riken.jp/5/tet/#!/search/hg19.cage_peak_counts_ann_decoded.osc.txt.gz.
- ENCODE chromHMM annotations²¹ in K562 were downloaded from <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgSegmentation/wgEncodeAwgSegmentationChromhmmK562.bed.gz>.
- CAGE data²¹ (wgEncodeRikenCageK562CellPapAlnRep1.bam) was downloaded from <http://moma.ki.au.dk/genome-mirror/cgi-bin/hgFileUi?db=hg19&g=wgEncodeRikenCage>.
- GRO-cap data⁵ (GSM1480321_K562_GROcap_wTAP_plus.bigWig and GSM1480321_K562_GROcap_wTAP_minus.bigWig) was downloaded from <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1480321>.
- Annotation of repetitive elements were taken from the UCSC table browser, track 'Repeats', track 'RepeatMasker' (downloaded 26-01-2015).

Peak calling on SuRE signal

To detect peaks of enrichment in the genome-wide SuRE-seq signal, we applied the MACS2 peak calling tool (version 2.1.0)²⁰ using (non-default) options “-g hs --bw 2000 --nomodel --keep-dup all --nolambda --slocal 1500”) to the 2 biological replicates of the cDNA data (‘treatment data’) and the genome-wide input coverage (‘control data’, see above; “Annotations, normalization and integrated data analysis”).

Overlap of SuRE peak summits, TSS, enhancers, and repetitive elements

The SuRE peaks were annotated by determining overlap of the peak summits with ‘Tss’ and combined ‘Enh’ and ‘EnhW’ regions taken from the ENCODE annotation, and with repetitive regions taken from the repeatmasker annotation (see above: “Data Sources”). Overlap was determined using the GenomicRanges package of BioConductor⁴⁹.

qPCR of globin genes

Treatment of K562 cells with hemin or solvent control was performed in triplicate as described above. RNA extraction and DNase digestion for ~ 1µg RNA were performed as described above, but no polyA purification was done. Next, cDNA was produced by adding 0.5 µl of 10 µM oligo dT, 0.5 µl 50ng/ul random hexamers and 1 µl dNTP (10mM each) and incubating for 5 minutes at 65°C. Then 4 µl of first strand buffer, 20 units RNase inhibitor (#EO0381; ThermoFisher Scientific), 1µl of Tetro reverse transcriptase (#BIO-65050; Bioline) and 2 µl water was added and the reaction mix was incubated for 10 minutes at 25°C followed by 45 minutes at 45°C and heat-inactivation at 85° for 5 minutes. qPCR was performed on the Roche LightCycler480 II using the Sensifast SYBR No-ROX mix (#BIO-98020). All expression levels were normalized to the internal control TBP and then expressed as relative to the 24 hour solvent treated control. Primer sequences can be found in Supplementary Table 2.

Conventional reporter assay

Promoters were chosen to cover the entire SuRE enrichment range. For each promoter a region representing ~550bp upstream to ~50bp downstream of the TSS was PCR amplified using MyTaq™ Red Mix (#BIO-25043; Bioline), repaired using the End-It™ DNA End-Repair Kit (#ER0720; Epicentre) and cloned into the SuRE reporter vector lacking barcodes. PCR primers are listed in Supplementary Table 2. The SuRE reporter vector was generated as described above but after the first gel-extraction (after the Xcm1/Nhe1 digest), the vector was repaired using the End-It™ DNA End-Repair Kit and dephosphorylated using rSAP (M037PS; NEB). All constructs were purified by miniprep (#BIO-52057) and their sequence was confirmed by Sanger sequencing. One µg was nucleofected along with 0.2µg of a control plasmid (YFP expressed under the CMV promoter) into 2 million K562 cells. Expression was analyzed by RT-qPCR after 20 hours as described above for the globin genes. GFP expression was quantified and normalized to the internal control YFP. Results were then compared to the mean SuRE enrichment obtained for the interval covered by the cloned promoter region.

Statistics

All SuRE peaks were called with FDR ≤ 0.05; for each region the SuRE enrichment and the peak summit were recorded. We subsequently only considered peaks that showed at least a 2 fold enrichment in SuRE.

Enrichment of overlap between features in Figure 1f and 6a was defined as the ratio of the overlap on the generated data and on the overlap between the features where one feature set was circularly randomized within each chromosome (using R-package regioneR⁵¹). The overlap distribution in 10,000 random circular permutations was used to compute a p-value for enrichment.

The p-values in Figure 1e, f and 5e, f refer to the p-value of the Pearson correlation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the NKI Genomics Core Facility for technical support, J. Omar Yáñez Cuna for scripts and advice, and members of our laboratories for helpful discussions. Supported by ERC Advanced Grant 293662 and NWO-ALW VICI (BvS); NIH grant R01HG003008 (HJB); and NIH training grants T32GM008798 and T32GM008281 (VDF).

References

1. Kadonaga JT. Perspectives on the RNA polymerase II core promoter. Wiley Interdiscip Rev Dev Biol. 2012; 1:40–51. [PubMed: 23801666]
2. Shiraki T, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A. 2003; 100:15776–15781. [PubMed: 14663149]
3. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008; 322:1845–1848. [PubMed: 19056941]

4. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*. 2013; 339:950–953. [PubMed: 23430654]
5. Core LJ, et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet*. 2014; 46:1311–1320. [PubMed: 25383968]
6. Patwardhan RP, et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol*. 2009; 27:1173–1175. [PubMed: 19915551]
7. Sharon E, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*. 2012; 30:521–530. [PubMed: 22609971]
8. Melnikov A, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol*. 2012; 30:271–277. [PubMed: 22371084]
9. Kheradpour P, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res*. 2013; 23:800–811. [PubMed: 23512712]
10. Lubliner S, et al. Core promoter sequence in yeast is a major determinant of expression level. *Genome Res*. 2015; 25:1008–1017. [PubMed: 25969468]
11. Farley EK, et al. Suboptimization of developmental enhancers. *Science*. 2015; 350:325–328. [PubMed: 26472909]
12. Nguyen TA, et al. High-throughput functional comparison of promoter and enhancer activities. *Genome Res*. 2016; 26:1023–1033. [PubMed: 27311442]
13. Patwardhan RP, et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol*. 2012; 30:265–270. [PubMed: 22371081]
14. Smith RP, et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet*. 2013; 45:1021–1028. [PubMed: 23892608]
15. Mogno I, Kwasniewski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res*. 2013; 23:1908–1915. [PubMed: 23921661]
16. Dickel DE, et al. Function-based identification of mammalian enhancers using site-specific integration. *Nat Methods*. 2014; 11:566–571. [PubMed: 24658141]
17. Murtha M, et al. FIREWACH: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat Methods*. 2014; 11:559–565. [PubMed: 24658142]
18. Arnold CD, et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*. 2013; 339:1074–1077. [PubMed: 23328393]
19. Zabidi MA, et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*. 2015; 518:556–559. [PubMed: 25517091]
20. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008; 9:R137. [PubMed: 18798982]
21. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
22. Osoegawa K, et al. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res*. 2001; 11:483–496. [PubMed: 11230172]
23. Harrow J, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012; 22:1760–1774. [PubMed: 22955987]
24. Duttke SH, et al. Human promoters are intrinsically directional. *Mol Cell*. 2015; 57:674–684. [PubMed: 25639469]
25. Scruggs BS, et al. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol Cell*. 2015; 58:1101–1112. [PubMed: 26028540]
26. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*. 1987; 196:261–282. [PubMed: 3656447]
27. Landolin JM, et al. Sequence features that drive human promoter function and tissue specificity. *Genome Res*. 2010; 20:890–898. [PubMed: 20501695]
28. Bird AP. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*. 1980; 8:1499–1504. [PubMed: 6253938]
29. Andersson R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *Bioessays*. 2015; 37:314–323. [PubMed: 25450156]

30. Kim TK, Shiekhhattar R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell*. 2015; 162:948–959. [PubMed: 26317464]
31. Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res*. 2013; 23:1210–1223. [PubMed: 23636943]
32. Arner E, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*. 2015; 347:1010–1014. [PubMed: 25678556]
33. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature*. 2012; 489:109–113. [PubMed: 22955621]
34. Kim TK, et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature*. 2010; 465:182–187. [PubMed: 20393465]
35. Blom van Assendelft G, Hanscombe O, Grosveld F, Greaves DR. The beta-globin dominant control region activates homologous and heterologous promoters in a tissue-specific manner. *Cell*. 1989; 56:969–977. [PubMed: 2924354]
36. Ashe HL, Monks J, Wijgerde M, Fraser P, Proudfoot NJ. Intergenic transcription and transinduction of the human beta-globin locus. *Genes Dev*. 1997; 11:2494–2509. [PubMed: 9334315]
37. Rada-Iglesias A, et al. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*. 2011; 470:279–283. [PubMed: 21160473]
38. Kwasniewski JC, Fiore C, Chaudhari HG, Cohen BA. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res*. 2014; 24:1595–1602. [PubMed: 25035418]
39. Hay D, et al. Genetic dissection of the alpha-globin super-enhancer in vivo. *Nat Genet*. 2016
40. Dean A, Ley TJ, Humphries RK, Fordis M, Schechter AN. Inducible transcription of five globin genes in K562 human leukemia cells. *Proc Natl Acad Sci U S A*. 1983; 80:5515–5519. [PubMed: 6310580]
41. Tahara T, Sun J, Igarashi K, Taketani S. Heme-dependent up-regulation of the alpha-globin gene expression by transcriptional repressor Bach1 in erythroid cells. *Biochem Biophys Res Commun*. 2004; 324:77–85. [PubMed: 15464985]
42. Faulkner GJ, et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*. 2008; 91:281–288. [PubMed: 18178374]
43. Ling J, et al. The solitary long terminal repeats of ERV-9 endogenous retrovirus are conserved during primate evolution and possess enhancer activities in embryonic and hematopoietic cells. *J Virol*. 2002; 76:2410–2423. [PubMed: 11836419]
44. Kwasniewski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A*. 2012; 109:19498–19503. [PubMed: 23129659]
45. Consortium F, et al. A promoter-level mammalian expression atlas. *Nature*. 2014; 507:462–470. [PubMed: 24670764]
46. Yu X, et al. The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2. *J Biol Chem*. 2005; 280:35184–35194. [PubMed: 16105833]
47. Temin HM. Structure, variation and synthesis of retrovirus long terminal repeat. *Cell*. 1981; 27:1–3. [PubMed: 7326747]
48. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012; 9:357–359. [PubMed: 22388286]
49. Huber W, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015; 12:115–121. [PubMed: 25633503]
50. Hubley R, et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res*. 2016; 44:D81–89. [PubMed: 26612867]
51. Gel B, et al. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*. 2016; 32:289–291. [PubMed: 26424858]

Figure 1. SuRE provides a genome-wide map of autonomous promoter activity

a. Schematic representation of the SuRE experimental strategy. ORF, open reading frame; PAS, polyadenylation signal. Colors indicate different barcodes. **b.** Representative ~1Mb genomic region showing histone modifications H3K27ac and H3K4me3²¹ that mostly mark active TSSs, and SuRE signals divided into plus and minus orientation. SuRE signal represents fold enrichment over input. **c.** Relative enrichment (compared to random) of SuRE peaks among the major types of chromatin²¹. **d.** Correlation between endogenous promoter activity (measured by GRO-cap⁵) and SuRE enrichment at TSSs. The density plots show the data distribution over each axis. nd, not detected. **e.** Correlation between relative promoter autonomy ($\log_{10}(\text{SuRE enrichment}/\text{GRO-cap})$) and tissue specificity (number of cell types and tissues in which each TSS is active, out of 889 tested⁴⁵). Grey line shows linear fit. **f.** Correlation between relative promoter autonomy and the total number of enhancers that are found in a fixed window of 5–50 kb from the TSS (regardless of the position of neighboring genes). The y-axis scale is the same as in **e**.

Figure 2. Autonomous divergent promoter activity

a. Mean SuRE enrichment at all TSSs and their 5kb flanking regions. **b, c.** SuRE enrichment aligned to all TSSs in the sense (**b**) and antisense (**c**) orientation, sorted by sense signal intensity. **d.** Distribution of SuRE enrichment levels at all TSSs; nd, not detected.

Figure 3. Partially overlapping query fragments allow for delineation of regions that drive promoter activity

a. Top tracks: GRO-cap expression, SuRE enrichment and alternative transcripts; bottom panel: SuRE expression of individual genomic fragments around the *NUP214* TSS in the sense orientation. The y-axis indicates the $\log_{10}$ -transformed number of reads for each genomic fragment; a random value between -0.2 and $+0.2$ was added to avoid overlap of fragments. The 5' end of each element is indicated by a black vertical bar. **b.** Contribution to autonomous promoter activity across the region surrounding the *NUP214* TSS, estimated using an elastic net Poisson regression model that uses fragment overlap with 50bp genomic sequence bins to predict expression in a multiplicative manner. The model fit was repeated using shifted versions of the same bins to avoid artefacts due to breakpoint choice. Shown are the exponentiated per base mean coefficients for all possible shifts. **c.** Mean SuRE

expression of genomic fragments with a similar start and end position (binned in 100 bp windows) relative to the nearest TSS. For example, the leftmost colored arrows mark all fragments starting at -500 ± 50 bp and the rightmost colored arrows mark all fragments ending at the TSS ± 50 bp; the square at the intersection shows the mean SuRE expression of all fragments that match both criteria. NA: fewer than 50 fragments in bin. **d.** Same as **(b)** but for all TSSs. **e.** Same as **(a)** but for antisense orientation. Here the 3' end of each element is indicated by black vertical bar. **f.** Same as **(b)** but for antisense orientation. **g.** Same model used in **(d)** was applied to a subset of sense-antisense TSS pairs⁵, using 50bp regions centered on the sense TSS (right) in one model and the antisense TSS (left) in a second. Expected fold-changes in sense (above) and antisense expression (below) are shown for the 50 bp region centered on the corresponding TSS. Error bars indicate standard error of Poisson regression coefficients. **h–j.** GRO-cap expression and alternative transcripts (top panels) and contribution to autonomous promoter activity as in **(b)** (bottom panels) for the genes *SLC50A1* (**h**), *WDR47* (**i**) and *HIST1H2BD* (**j**). In all panels, sense orientation is depicted in blue and antisense orientation in red.

Figure 4. Relationship between CpG islands and gene expression

a. Distribution of all mappable SuRE fragments, regardless of their expression level, in terms of their CpG characteristics. Only fragments that overlap an annotated TSS were included. The color scale indicates the number of fragments belonging to each hexagon bin. The lines denote when the observed CpG density per base pair equals 100% (solid) or 50% (dashed) of the value expected based on C+G content. **b.** Relationship between expression level and CpG characteristics. The color scale indicates the average cDNA read count per fragment in each hexagon bin. Lines are the same as in **a**.

Figure 5. Autonomous transcription from enhancers

a. SuRE data indicate that three of the five DNase hypersensitive sites (DHS)²¹ in the β -globin locus control region show autonomous transcription activity. **b.** SuRE signals (plus and minus strand combined) aligned to enhancers ('Enh'), weak enhancers ('EnhW') and quiescent parts of the genome ('Quies')²¹, each sorted by SuRE signal intensity. **c.** Average profiles of data in **b.** **d.** Distribution of SuRE enrichments as shown in **b** compared to TSSs. nd, not detected. **e.** Correlation between SuRE expression and H3K27ac signal for enhancers. Grey line shows linear fit. **f.** Correlation between enhancer strength of ~130 bp fragments from selected enhancers³⁸ and the mean SuRE expression in a 1 kb window around the center of these (n=189). Grey line shows linear fit. **g.** Expression levels of 4 genes of the alpha-globin region and a negative control gene (*ACTB*) after 24 hours of induction with hemin or the solvent control. Expression levels were normalized to TBP and visualized as fold-change relative to solvent control. Error bars indicate the SEM of 3 biological replicates. **h.** Genomic region of the alpha-globin locus. The top track indicates conserved enhancers. The track below shows the DHS-seq signal²¹. The bottom 4 tracks

show SuRE enrichment before and after hemin induction for the plus strand (blue) and minus strand (red).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 6. Autonomous transcription from specific repeat elements

a. Enrichment of SuRE peaks among the major repeat families. Asterisks: significant enrichment or depletion ($p < 0.01$ after multiple testing correction). **b.** Mean SuRE enrichment of subfamilies LTR12C (left panel; $n = 2,600$) and MER41B (right panel; $n = 2,764$) in the sense (blue) and antisense (red) direction. **c.** Distribution of SuRE enrichment levels (plus and minus strand combined) of LTR12C and MER41B repeats compared to enhancers and TSSs. nd, not detected. **d.** Contribution of LTR12C sequences to autonomous promoter activity, as in Fig. 3b, relative to previously annotated^{46, 47} U3, promoter (P), enhancer (E), transcribed (R) and U5 elements. **e.** Average endogenous run-on transcription⁵ levels in the sense orientation at indicated distances upstream or downstream of LTR12C repeats. High and low activity refers to top 50% and bottom 50% in SuRE enrichment.