nature biotechnology

Genome-wide mapping of autonomous promoter activity in human cells

Joris van Arensbergen[1], Vincent D FitzPatrick[2,3], Marcel de Haas[1], Ludo Pagie[1], Jasper Sluimer[1], Harmen J Bussemaker[2,3] & Bas van Steensel[1]

Previous methods to systematically characterize sequence-intrinsic activity of promoters have been limited by relatively low throughput and the length of the sequences that could be tested. Here we present 'survey of regulatory elements' (SuRE), a method that assays more than 10^8 DNA fragments, each 0.2–2 kb in size, for their ability to drive transcription autonomously. In SuRE, a plasmid library of random genomic fragments upstream of a 20-bp barcode is constructed, and decoded by paired-end sequencing. This library is used to transfect cells, and barcodes in transcribed RNA are quantified by high-throughput sequencing. When applied to the human genome, we achieve 55-fold genome coverage, allowing us to map autonomous promoter activity genome-wide in K562 cells. By computational modeling we delineate subregions within promoters that are relevant for their activity. 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.

Promoters harbor the transcription start site (TSS) and various other sequences that control transcription initiation through the binding of *trans*-acting factors[1]. 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 activity is controlled. Proximal sequences, distal enhancers, local chromatin context, and three-dimensional (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 in order 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 (~150 bp) 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 hundred base pairs[13–15], also with a multiplexing level of 10^4–10^5. A complementary strategy that uses shotgun cloning into a reporter plasmid was used to screen several hundred kilobases of genomic DNA for enhancer activity in mouse cells[16]. 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 cells[17]. 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 SuRE, a method that overcomes some of these limitations. Instead of short synthetic promoter sequences, SuRE queries random genomic fragments ranging in size from 0.2 to 2 kb, 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', that is, 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.

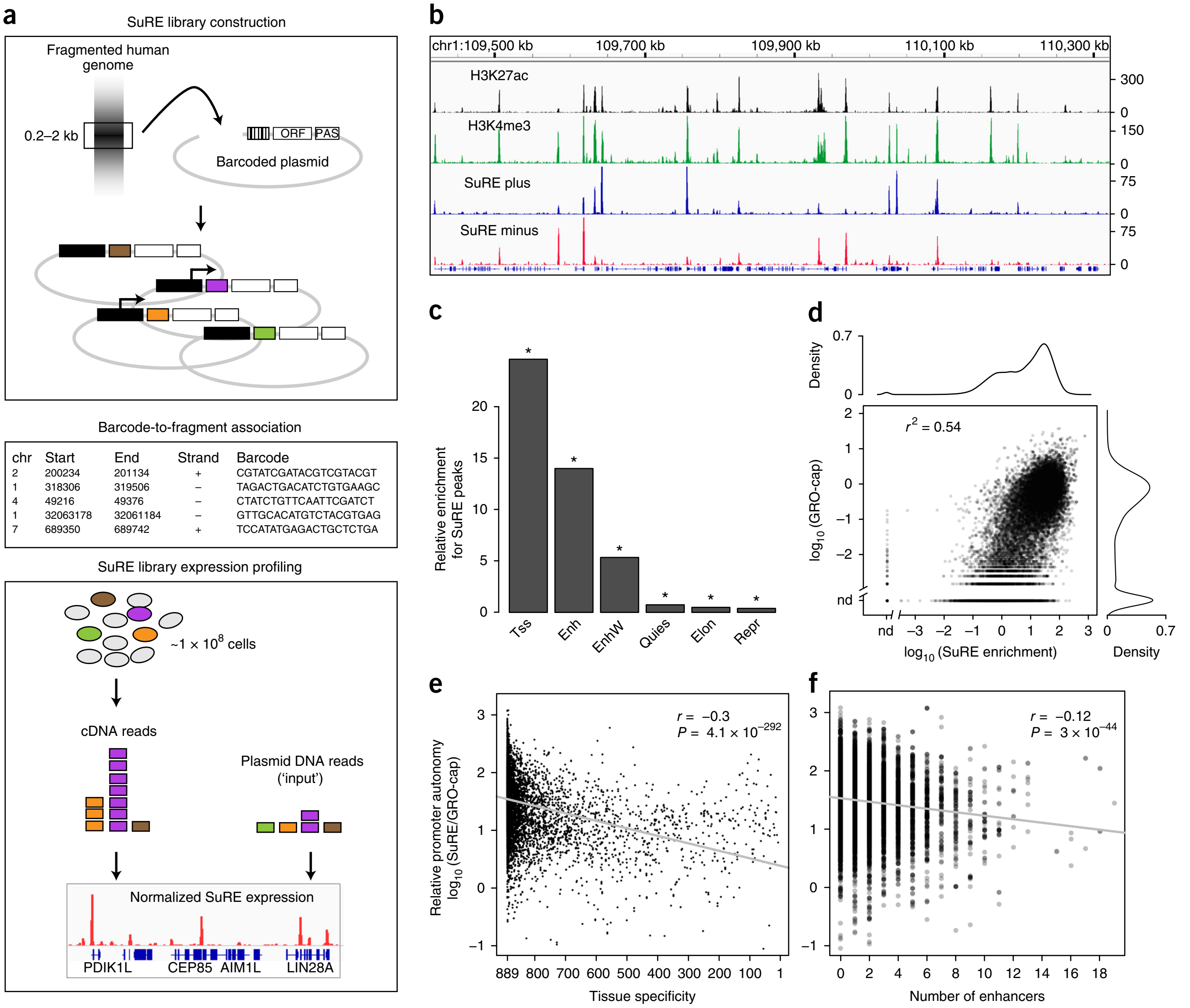
RESULTS

SuRE method and library preparation

The SuRE experimental strategy consists of three main steps (**Fig. 1a** and **Supplementary Fig. 1**). First, genomic DNA is randomly

[1]Division of Gene Regulation, Netherlands Cancer Institute, Amsterdam, the Netherlands. [2]Department of Biological Sciences, Columbia University, New York, New York, USA.[3]Department of Systems Biology, Columbia University Medical Center, New York, New York, USA. Correspondence should be addressed to J.v.A. (j.v.arensbergen@nki.nl), H.J.B. (hjb2004@columbia.edu), or B.v.S. (b.v.steensel@nki.nl).

Received 4 May; accepted 1 December; published online 26 December 2016; doi:10.1038/nbt.3754



chr	Start	End	Strand	Barcode
2	200234	201134	+	CGTATCGATACGTCGTACGT
1	318306	319506	–	TAGACTGACATCTGTGAAGC
4	49216	49376	–	CTATCTGTTCAATTCGATCT
1	32063178	32061184	–	GTTGCACATGTCTACGTGAG
7	689350	689742	+	TCCATATGAGACTGCTCTGA

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 ~1-Mb genomic region showing histone modifications H3K27ac and H3K4me3 (ref. 21) 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 chromatin[21]. TSS, active promoter; Enh, enhancer; EnhW, weak enhancer; Repr, polycomb repressed; Elon, elongation; Quies, quiescent. Asterisks, significant enrichment or depletion ($P < 0.01$, Monte Carlo test, after multiple testing correction). (**d**) Correlation between endogenous promoter activity (measured by GRO-cap[5]) and SuRE enrichment at TSSs. The density plots show the data distribution over each axis. nd, not detected. (**e**) Correlation between relative promoter autonomy ($\log_{10}$(SuRE enrichment/GRO-cap)) and tissue specificity (number of cell types and tissues in which each TSS is active, out of 889 tested[45]). Gray 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**. The *P*-values in **e**,**f** refer to the *P*-value of the Pearson correlation.

fragmented and subjected to size selection to obtain 0.2- to 2-kb-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, cultured cells are transiently transfected with the library, 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 (the reporter plasmid itself does not contain a promoter or 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 with 96% of the mappable genome covered at least 15-fold (**Supplementary Fig. 2a**).

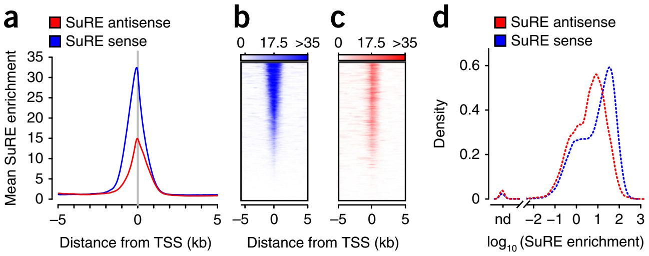


Figure 2 Autonomous divergent promoter activity. (**a**) Mean SuRE enrichment at all TSSs and their 5-kb flanking regions. (**b**,**c**) SuRE enrichment aligned to all TSSs in the sense (**b**) and antisense (**c**) orientation, sorted by sense signal intensity. (**d**) Distribution of SuRE enrichment levels at all TSSs; nd, not detected.

Genome-wide map of autonomous promoter activity in human cells

We transiently transfected human K562 erythroleukemia cells with the library. 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 Figure 2g**–**i**.

As expected, the resulting SuRE activity map shows a pattern of peaks that overlap frequently with known TSSs and histone modifications marking active promoters, such as H3K4me3 and H3K27ac (**Fig. 1b**). A peak detection algorithm[20] identified 55,453 SuRE peaks at an estimated false-discovery rate of 5% and with at least twofold enrichment of SuRE signal over background (**Supplementary Data Set 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 in repressed (‘Repr’) or quiescent (‘Quies’) parts of the genome[21] (**Fig. 1c** and **Supplementary Fig. 2b**). 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 nine selected regions of the human genome[22] (**Supplementary Table 1**), together spanning 1.3 Mb. This library had an average 212-fold coverage of the included base pairs. Owing to its lower complexity and higher coverage it yielded highly reproducible results (Pearson’s r = 0.99; **Supplementary Fig. 3a**). Within the regions probed by this focused library, 45 out of 50 peaks (90%) previously identified in the genome-wide SuRE data set also showed enriched signals in the focused SuRE data set (**Supplementary Fig. 3b**). Similarly, out of 55 TSSs with a genome-wide SuRE enrichment of at least twofold, 53 (96%) showed enriched signals in the focused SuRE data set (**Supplementary Fig. 3c**). 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 the variance in 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 method[5]. We focused on a curated set (Online Methods) of 28,844 TSSs annotated by the GENCODE project[23]. Notably, SuRE and GRO-cap signals were substantially correlated (r^2 = 0.54; **Fig. 1d**). Similar results were obtained when only comparing TSSs that showed expression in both SuRE and GRO-cap (r^2 = 0.43), and in a comparison with transcription activity detected by the CAGE method[21] (r^2 = 0.49; **Supplementary Fig. 2d**). Thus, a substantial part of promoter activity is reproduced by sequence elements <2 kb from the TSS, that is, 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 identified many housekeeping promoters with a low level of promoter autonomy, for example, promoters of genes that encode histones (**Supplementary Fig. 2e**). Relative promoter autonomy was 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 (**Supplementary Fig. 2f**). 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 orientation[3]. We expected that in SuRE this antisense transcription might be detected if a promoter was inserted in reverse orientation, as the transcript would be stabilized by the plasmid-encoded transcription unit. Indeed, SuRE detected extensive activity of promoters in the antisense direction (**Fig. 2a**–**c**). The antisense activity was, on average, two-to threefold weaker but it correlated 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 of different size and with 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 showed 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 (Online Methods). When applied to *NUP214*, this confirmed that the proximal ~100 bp upstream of the TSS was 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). Most activity was contributed by the core promoter and sequences within a few hundred bp upstream; inclusion of longer upstream regions, on average, did not increase reporter activity. Increasing the length of the sequence included downstream of the TSS tended to reduce reporter activity, which might, 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 were primarily provided by the core promoter

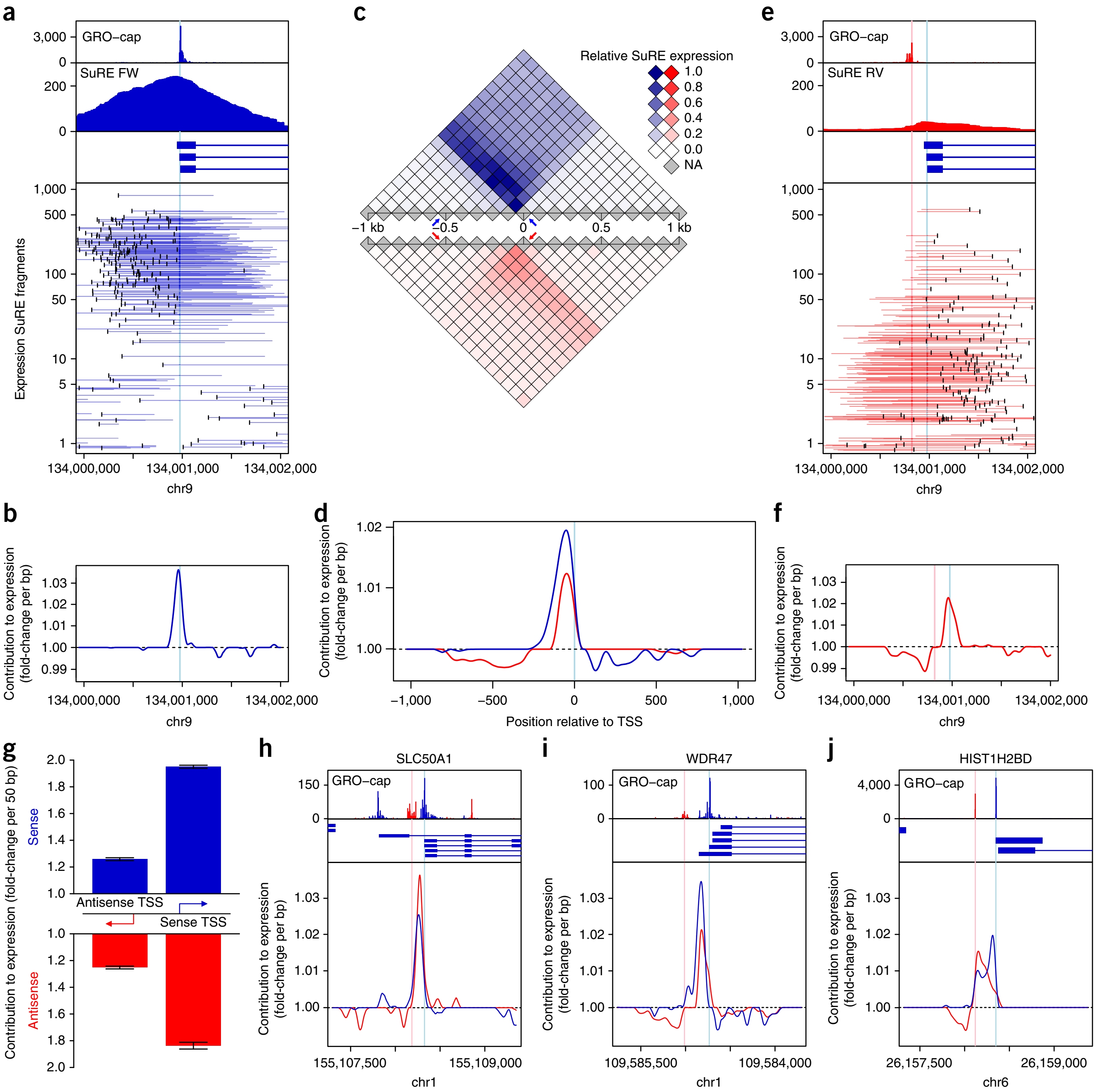


Figure 3 Partially overlapping query fragments allow for delineation of regions that drive promoter activity. (**a**) Top tracks: GRO-cap expression, SuRE enrichment, and alternative transcripts; bottom panel: SuRE expression of individual genomic fragments around the *NUP214* TSS in the sense orientation. 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 50-bp genomic sequence bins to predict expression in a multiplicative manner. The model fit was repeated using shifted versions of the same bins to avoid artifacts 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 left-most colored arrows mark all fragments starting at –500 ± 50 bp and the right-most 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 pairs[5], using 50-bp 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. Note for panels **b**, **d**, **f**–**j**: when interpreting these plots, it should be kept in mind that a fold-change per bp of 1.01 corresponds to a fold-change per 50 bp of $(1.01)^{50} = 1.6$.

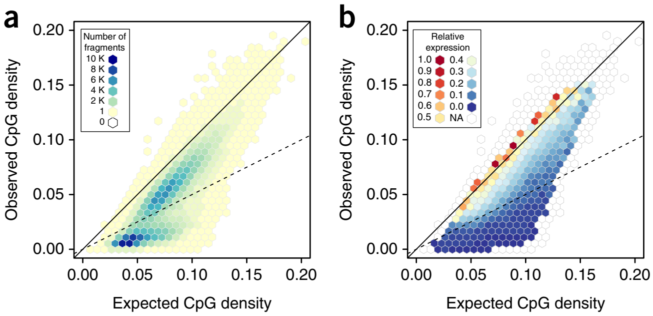


Figure 4 Relationship between CpG islands and gene expression. (**a**) Distribution of all mappable SuRE fragments, regardless of their expression level, in terms of their CpG characteristics. Only fragments that overlap an annotated TSS were included. The color scale indicates the number of fragments 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**.

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 that 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, it has been proposed that the sense and antisense core promoters function in a cooperative manner[25]. 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 showed antisense SuRE activity extended at least ~200 bp 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 was needed for antisense transcription (**Fig. 3f**). Indeed, genome-wide analysis showed that promoter fragments that included the forward core promoter generally exhibited the highest SuRE activity in antisense orientation (**Fig. 3c**, bottom triangle). GLM applied to all promoters combined also indicated that antisense transcription was 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 pairs[5] (**Fig. 3g**) underscored this general conclusion.

Inspection of raw SuRE data and GLM profiles of individual promoters covered by our focused library revealed several interesting examples of and exceptions to this general trend. For example, transcription from both the main sense and main antisense TSS of *SLC50A1* required the same subregion located between them; however, an alternative sense TSS upstream and an additional antisense TSS downstream appeared to be non-autonomous, because no GLM signal was detectable at these sites (**Fig. 3h**). In the *WDR47* gene, antisense transcription did not require the antisense TSS subregion, but rather depended on a subregion that was 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 were 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 content[26]. 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 formed two distinct populations that could be separated by a ~50% observed/expected CpG ratio, consistent with a previous classification of promoters[27]. However, the relationship between SuRE expression level and CpG content for individual fragments took a different form (**Fig. 4b**). Expression was highest when the observed and expected CpG density were equal, and decayed gradually with decreasing CpG observed/expected ratio. Notably, this relationship was 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 deamination[28]. 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 promoter[34]. 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 region[35], showed several clear bi-directional SuRE signals (**Fig. 5a**). Analysis of 47,020 predicted active enhancers in K562 cells[21] revealed SuRE signals for the majority (**Fig. 5b**,**c**), although the overall level of activity was approximately tenfold lower than for promoters (cf., **Figs. 2a**,**d** and **5c**,**d**). We conclude that eRNA production is generally autonomous, that is, 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] correlated with the strength of SuRE signals ($P < 2.2 \times 10^{-16}$, Wilcoxon test) (**Fig. 5b**,**d**). SuRE signals also correlated positively with the endogenous levels of H3K27ac (**Fig. 5e**), the histone modification most characteristic of active enhancers[37]. Furthermore, the ability of ~130-bp fragments derived from ENCODE-annotated enhancers to activate a minimal promoter in a previous reporter assay[38] showed a significant ($P = 4 \times 10^{-4}$) positive correlation with the SuRE signal

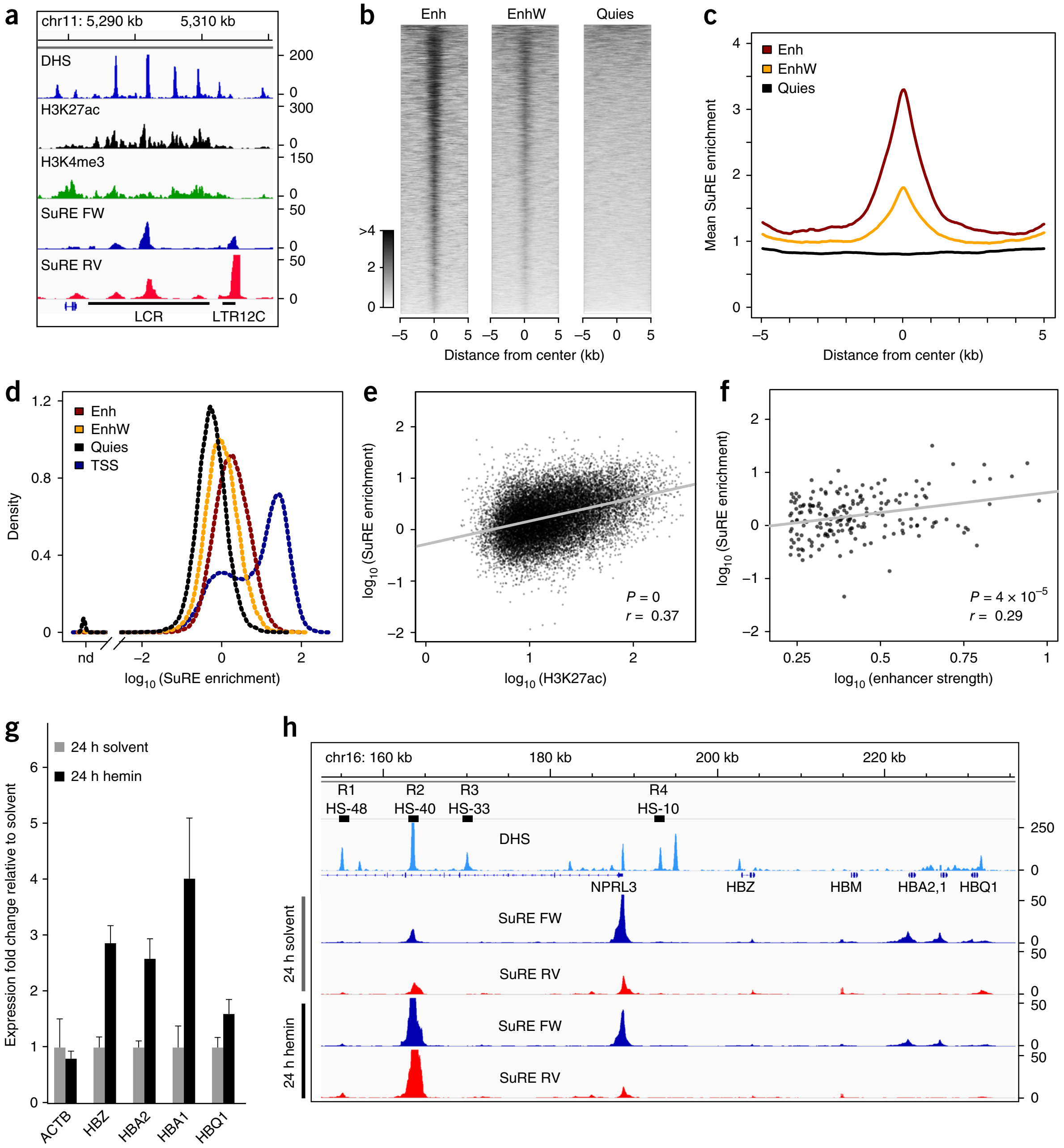


Figure 5 Autonomous transcription from enhancers. (**a**) SuRE data indicate that three of the five DNase hypersensitive sites (DHS)[21] in the β-globin locus control region show autonomous transcription activity. (**b**) SuRE signals (plus and minus strand combined) aligned to enhancers (Enh), weak 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. Gray line shows linear fit. (**f**) Correlation between enhancer strength of ~130-bp fragments from selected enhancers[38] and the mean SuRE expression in a 1-kb window around the center of these (n = 189). Gray line shows linear fit. The P-values in **e**,**f** refer to the P-value of the Pearson correlation. (**g**) Expression levels of four genes of the α-globin region and a negative control gene (*ACTB*) after 24 h 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 s.e.m. of three biological replicates. (**h**) Genomic region of the α-globin locus. The top track indicates conserved enhancers. The track below shows the DHS-seq signal[21]. The bottom four tracks show SuRE enrichment before and after hemin induction for the plus strand (blue) and minus strand (red).

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 α-globin LCR

To further illustrate the value of SuRE for dissecting regulatory mechanisms, we used our focused SuRE library to analyze the α-globin locus, which harbors a locus control region that can activate several globin genes over a distance of >50 kb. 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 expression[39]. Treatment of K562 cells with hemin is known to increase expression of several of the genes in the α-globin locus[40], which we confirmed by RT-qPCR (**Fig. 5g**). Although R2

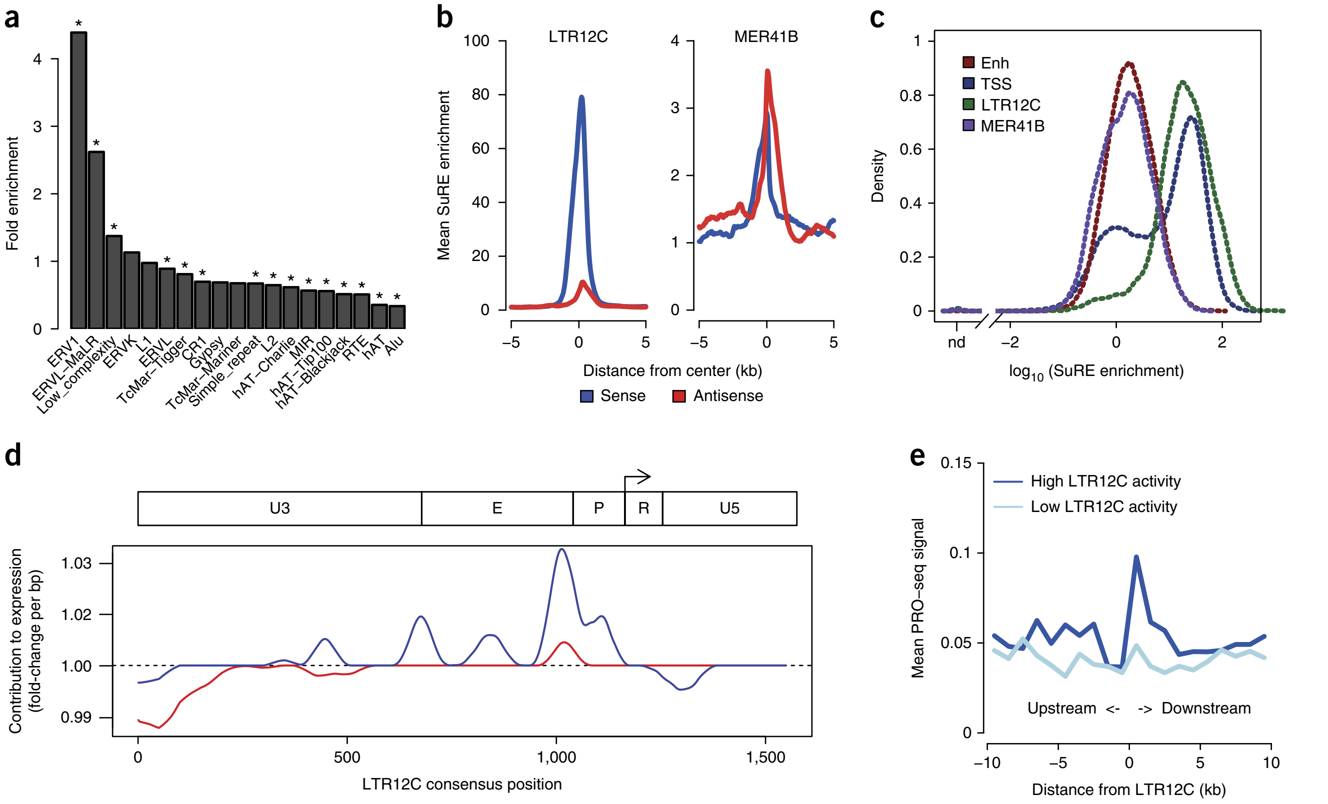


Figure 6 Autonomous transcription from specific repeat elements. (**a**) Enrichment of SuRE peaks among the major repeat families. Asterisks, significant enrichment or depletion ($P < 0.01$, Monte Carlo test, after multiple testing correction). (**b**) Mean SuRE enrichment of subfamilies LTR12C (left 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 **Figure 3b**, relative to previously annotated[46,47] U3, promoter (P), enhancer (E), transcribed (R), and U5 elements. (**e**) Average endogenous run-on transcription[5] 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.

can be activated by hemin[41], 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 (**Fig. 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 accounted for only 26% of the genome-wide SuRE peaks (**Supplementary Fig. 2i**). Several families of repetitive elements showed significant ($P < 0.01$ after multiple testing correction) overlap with SuRE peaks, in particular, the ERVL-MaLR and ERV1 retrotransposons (**Fig. 6a**), which accounted for another 19% of the peaks. Certain subfamilies within these families exhibited specific and high SuRE signals, for example, the LTR12C subfamily of solitary long terminal repeats (LTR; **Fig. 6b**,**c**). For some repeat subfamilies (e.g., LTR12C), the average SuRE activity resembled 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 were more reminiscent of those 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 genome[42], 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 an 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 LTRs[43] and again indicate that essentially the same sequence elements contribute to sense and antisense transcription.

Sense-oriented run-on transcription[5] was detectable downstream of LTR12C insertions with high SuRE activity (**Fig. 6e**). This was 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% were accounted for by ENCODE-annotated promoters and enhancers or ERVL-MaLR and ERV1 retrotransposons. Of the 30,548 remaining 'unexplained' peaks, only 15% overlapped with a TSS or enhancer annotated in one of 889 cell sources assayed by the FANTOM project. The unexplained peaks, however, did show enrichment for epigenetic marks of promoter activity, such as H3K4me3 or DNase I hypersensitivity (**Supplementary Fig. 5a**,**b**). Their average SuRE signal was substantially above background, while they produced 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

Our results show that SuRE can work as a high-throughput tool to functionally deconstruct large genomes and systematically identify elements that drive autonomous transcription activity. SuRE operates on a 100- to 1,000-fold larger scale than previous high-throughput promoter assays, sufficient to survey the entire human genome at >50× 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 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 genome[21].

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

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, and T. Rube in particular for suggesting the 2D analysis of CpG content. Supported by ERC Advanced Grant 293662 and NWO-ALW VICI (B.v.S.); NIH grants R01HG003008 and S10OD021764 (H.J.B.); and NIH grants R01HG003008 and S10OD021764 and T32GM008281 (V.D.F.). Addgene plasmid # 49157 is a gift from James Thomson, University of Wisconsin, Madison.

AUTHOR CONTRIBUTIONS

J.v.A. conceived and developed the SuRE assay, designed and performed experiments, analyzed data and 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

The authors declare no competing financial interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Kadonaga, J.T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 40–51 (2012).
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. USA* **100**, 15776–15781 (2003).
3. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
4. Kwak, H., Fuda, N.J., Core, L.J. & Lis, J.T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
5. Core, L.J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
6. Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
7. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
8. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
9. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).
10. Lubliner, S. *et al.* Core promoter sequence in yeast is a major determinant of expression level. *Genome Res.* **25**, 1008–1017 (2015).
11. Farley, E.K. *et al.* Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015).
12. Nguyen, T.A. *et al.* High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26**, 1023–1033 (2016).
13. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat. Biotechnol.* **30**, 265–270 (2012).
14. Smith, R.P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45**, 1021–1028 (2013).
15. Mogno, I., Kwasnieski, J.C. & Cohen, B.A. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* **23**, 1908–1915 (2013).
16. Dickel, D.E. *et al.* Function-based identification of mammalian enhancers using site-specific integration. *Nat. Methods* **11**, 566–571 (2014).
17. Murtha, M. *et al.* FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat. Methods* **11**, 559–565 (2014).
18. Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
19. Zabidi, M.A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
20. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
21. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
22. Osoegawa, K. *et al.* A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11**, 483–496 (2001).
23. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
24. Duttke, S.H. *et al.* Human promoters are intrinsically directional. *Mol. Cell* **57**, 674–684 (2015).
25. Scruggs, B.S. *et al.* Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* **58**, 1101–1112 (2015).
26. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
27. Landolin, J.M. *et al.* Sequence features that drive human promoter function and tissue specificity. *Genome Res.* **20**, 890–898 (2010).
28. Bird, A.P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
29. Andersson, R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays* **37**, 314–323 (2015).
30. Kim, T.K. & Shiekhattar, R. Architectural and functional commonalities between enhancers and promoters. *Cell* **162**, 948–959 (2015).
31. Hah, N., Murakami, S., Nagari, A., Danko, C.G. & Kraus, W.L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.* **23**, 1210–1223 (2013).
32. Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010–1014 (2015).
33. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
34. Kim, T.K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
35. Blom van Assendelft, G., Hanscombe, O., Grosveld, F. & Greaves, D.R. The beta-globin dominant control region activates homologous and heterologous promoters in a tissue-specific manner. *Cell* **56**, 969–977 (1989).

36. Ashe, H.L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N.J. Intergenic transcription and transinduction of the human beta-globin locus. *Genes Dev.* **11**, 2494–2509 (1997).
37. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
38. Kwasnieski, J.C., Fiore, C., Chaudhari, H.G. & Cohen, B.A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).
39. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer *in vivo. Nat. Genet.* **48**, 895–903 (2016).
40. Dean, A., Ley, T.J., Humphries, R.K., Fordis, M. & Schechter, A.N. Inducible transcription of five globin genes in K562 human leukemia cells. *Proc. Natl. Acad. Sci. USA* **80**, 5515–5519 (1983).
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.* **324**, 77–85 (2004).
42. Faulkner, G.J. *et al.* A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* **91**, 281–288 (2008).
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.* **76**, 2410–2423 (2002).
44. Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C. & Cohen, B.A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. USA* **109**, 19498–19503 (2012).
45. Forrest, A.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
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.* **280**, 35184–35194 (2005).
47. Temin, H.M. Structure, variation and synthesis of retrovirus long terminal repeat. *Cell* **27**, 1–3 (1981).

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 an SV40-derived polyadenylation signal (PAS). To generate a barcoded SuRE vector library, we digested 30 μg SuRE vector with NheI (#R0131; New England BioLabs (NEB)) and XcmI (#R0533; NEB) and performed a gel extraction 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 min at 96 °C, 14×(20 s at 96 °C, 20 s at 60 °C, 20 s 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 three 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 10× CutSmart buffer, 10 μl of 10 mM ATP, 10 units T4 DNA ligase (#10799009001 Roche). A cycle-ligation of six cycles was performed (10 min at 22 °C and 10 min at 37 °C), followed by 20 min heat-inactivation at 80 °C. The ligation reaction was purified by magnetic beads and digested with 40 units of XcmI (#R0533S; NEB) for 3 h, and size-selected by gel-extraction, yielding 5–10 μg of 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 (0.5–2 kb) 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 obtained many smaller elements in the final library (**Supplementary Fig. 2b**) presumably because size-selection is imperfect and smaller fragments preferentially contributed 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 h 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 *Escherichia coli* (#C810111; Genlantis) in four separate electroporations with each 5 μl of ligation product and 20 μl bacteria, each transferred to 500 ml standard Luria Broth (LB) plus kanamycin (50 μg/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, once established in the lab, takes about 5 d 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, overnight 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 mg 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 nine pooled bacterial artificial chromosomes (BACs), collectively covering 1.3 Mb of the human genome (**Supplementary Table 1**). This library was prepared essentially the same way as the genome-wide library except that size selection was performed for elements of 0.1–1 kb and that only 100 ng of 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 μl/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 undersequenced.

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 h 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 h 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, ten cycles of PCR (1 min at 98 °C, 10×(15 s at 98°,15 s at 60°, 20 s 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 adaptor) and (index variants of) 117JvA (containing the S2, index and p7 adaptor). 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, two biological replicates were done of each 100 million cells (5 million per cuvette with each 10 μg plasmid) and harvested after 24 h (see below). For the focused library experiments, two 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 μM hemin (Sigma; #51280-1G) or solvent control 1 h after nucleofection, and cells were harvested 24 h 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-μl reactions containing 500 ng RNA and treated with 10 units DNase I for 30 min (#04716728001; Roche) and DNase I was inactivated by addition of 1 μl 25 mM EDTA and incubation at 70 °C for 10 min. Next, cDNA was produced by first adding 1 μl of 10 μM gene-specific primer targeting the GFP ORF (247JvA) and 1 μl dNTP (10 mM each) and incubating for 5 min at 65 °C. Then 4 μl of RT buffer, 20 units RNase inhibitor (#EO0381; ThermoFisher Scientific), 200 units of Maxima reverse transcriptase (#EP0743; ThermoFisher Scientific) and 2.5 μl water was added, and the reaction mix was incubated for 30 min at 50 °C followed by heat inactivation at 85° for 5 min. Per biological replicate of the genome-wide library, 20–30 reactions were done in parallel. For the focused library, four reactions were done in parallel per biological replicate. Each 20 μl reaction was then PCR amplified (1 min 96 °C, 20×(15 s 96 °C, 15 s 60 °C, 15 s 72 °C)) in a 100-μl reaction with MyTaq Red Mix and primers 151AR (containing the S1 and p5 adaptor) and (index variants of) 211JvA (containing the S2, index and p7 adaptor) for 21 cycles. Reactions were then pooled and 500 μl 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 adaptor 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

doi:10.1038/nbt.3754

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 4 kb. All read pairs not aligned as a '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 µg library with I-SceI (#R0694S; NEB) in 25 µl to linearize the plasmids, then performed two replicate PCRs each on 2 µl I-SceI digested material, using the same protocol as for the cDNA but for eight 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.embnet.org/index.php/embnetjournal/article/view/200) to remove the adaptor (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 were analyzed 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 were used unless indicated otherwise. We created BigWig files for the profiles thus obtained using the GenomicRanges package in BioConductor[49].

Validation using the focused SuRE library. To assess if the sequencing depth for the input was sufficient we sequenced our focused BAC library input deeply and then by downsampling 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 shallowly sequenced input (average = 0.1 read per barcode). This thus strongly suggests there are 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 found that coverage was highly similar between pre- and post-transfection libraries (r^2 = 0.98; **Supplementary Fig. 3e**). In addition, in neither library was there any correlation between insert length and representation (data not shown), presumably because the typical insert size (~1,000 bp) represents only 25% of the total plasmid size. We conclude that the use of the pre-transfection library as input did not compromise the results.

Post-transfection plasmid extraction. Per replicate, 10 million cells were transfected with our focused SuRE library. After 24 h, cells were spun down, washed with PBS, spun down again, and taken up in 500 µl nuclear extraction buffer (10 mM NaCl, 2 mM MgCl, 10 mM Tris-HCl (pH 7.8), 5 mM DTT, 0.5% NP40). Cells were incubated on ice for 5 min, and nuclei were spun down at 7,000*g* 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 two minipreps according to the manufacturer's protocol. Per replicate 5 µg was digested in 50 µl with 2.5 µl Sce-1 for 2 h, heat inactivated at 65 °C for 20 min and two PCRs with 2.5 µl of this material were amplified as described above to characterize the barcode frequencies in the pre-transfection plasmid library. For comparison, 1 µg 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.gencodegenes.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 project[45]. 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 ± 500 bp 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 1-kb region. Metaprofiles (e.g., **Fig. 2a**) were also generated using CoverageView, using 50-bp bins, except for the PRO-seq data in **Figure 6e** and **Supplementary Figure 4**, which were generated using 1-kb bins because of the sparser nature of the data.

In log-transformed data representations on data sets that also contain zeros, such as the comparison of GRO-cap and SuRE at GENCODE TSSs in **Figure 1d**, a pseudo-count of half the minimal nonzero 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 5 kb away from any GENCODE TSS.

To assess the spatial profile of contribution to autonomous expression of successive intervals relative to the TSS (**Fig. 3c**), we created a two-dimensional (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 = 18,257), 'EnhW'; n = 28,763, 'Quies'; n = 36,627). Heatmaps in **Figures 2b**,**c** and **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 control[38]. 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**,**e**, only genomic fragments are shown that were detected in the cDNA. In **Supplementary Figure 2e** histone genes were indicated that contained 'HIST' in their name and to avoid redundancy, alternative TSSs were only plotted if they were at least 500 bp from the previous. For **Supplementary Figure 2a** 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

doi:10.1038/nbt.3754

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-cap[5], 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 sequence[50].

Data sources. • As a reference for transcription start sites (TSSs) we used GENCODE[23] version 19 TSSs (gencode.v19.annotation.gff3.gz) downloaded from http://www.gencodegenes.org/releases/19.html.

- FANTOM phase 1 data[45] was downloaded from http://fantom.gsc.riken.jp/5/tet/#!/search/hg19.cage_peak_counts_ann_decoded.osc.txt.gz.
- ENCODE chromHMM annotations[21] in K562 were downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgSegmentation/wgEncodeAwgSegmentationChromhmmK562.bed.gz.
- CAGE data[21] (wgEncodeRikenCageK562CellPapAlnRep1.bam) was downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRikenCage/.
- GRO-cap data[5] (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 two 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 BioConductor[49].

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 ~1 µg 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 50 ng/µl random hexamers and 1 µl dNTP (10 mM each) and incubating for 5 min at 65 °C. Then 4 µl of first strand buffer, 20 units RNase inhibitor (#EO0381; ThermoFisher Scientific), 1 µl of Tetro reverse transcriptase (#BIO-65050; Bioline) and 2 µl water was added and the reaction mix was incubated for 10 min at 25 °C followed by 45 min at 45 °C and heat-inactivation at 85° for 5 min. 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 gene *TBP* and then expressed as relative to the 24 h 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 ~550 bp upstream to ~50 bp 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. 1 µg was nucleofected along with 0.2 µg of a control plasmid (YFP expressed under the CMV promoter) into 2 million K562 cells. Expression was analyzed by RT-qPCR after 20 h 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 twofold enrichment in SuRE.

Enrichment of overlap between features in **Figures 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 regioneR[51]). The overlap distribution in 10,000 random circular permutations was used to compute a *P*-value for enrichment.

The *P*-values in **Figures 1e**,**f** and **5e**,**f** refer to the *P*-value of the Pearson correlation.

The analysis in **Figure 4** was inspired by ref. 52.

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

For software source codes of SuRE sequencing data processing and Penalized Generalized Linear Modeling, see **Supplementary Source Codes 1** and **2**.

48. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
49. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
50. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
51. Gel, B. *et al.* regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**, 289–291 (2016).
52. Rube, H.T. *et al.* Sequence features accurately predict genome-wide MeCP2 binding *in vivo. Nat. Commun.* **7**, 11025 (2016).