Predicting Autonomous Promoter Activity Based on Genome-wide Modeling of Massively Parallel Reporter Data

Vincent D. FitzPatrick

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

ABSTRACT

Predicting Autonomous Promoter Activity Based on Genome-wide Modeling of Massively Parallel Reporter Data

Vincent D. FitzPatrick

Existing methods to systematically characterize sequence-intrinsic activity of promoters are 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 a billion DNA fragments in parallel for their ability to drive transcription autonomously. In SuRE, a plasmid library is constructed of random genomic fragments upstream of a barcode and decoded by paired-end sequencing. This library is transfected into cells and transcribed barcodes are quantified in the RNA by high-throughput sequencing. By computationally analyzing the resulting data using generalized linear models, we succeed in delineating subregions within promoters that are relevant for their activity on a genomic scale, and making accurate predictions of expression levels that can be used to inform minimal promoter reporter construct design. We also show how our approach can be extended to analyze the differential impact of single-nucleotide polymorphisms (SNPs) on gene expression.

Table of Contents

List of Figures

Acknowledgements

My experience in the biological science has been filled with many wonderful people. I am so lucky to have the chance to learn from and work with all of them, but there are a few that I would like to acknowledge in particular.

First, I would like to thank Terry Horton and Larry Heaney, who saw a biologist hiding within an uncertain undergraduate. I am also grateful to Vinzenz Unger, who gave me my first shot in a molecular biology lab.

I will be forever grateful for the guidance and mentorship of my adviser, Harmen Bussemaker. Thank you for your enthusiasm, your encouragement, your knowledge and your patience. Without you, I may never have discovered my love for statistics and for teaching, gifts that I will carry for the rest of my life. I hope that, wherever I end up, I can be half the mentor you have been to me.

Thank you to my core committee members, Larry Chasin and Lars Dietrich, for your valuable advice. I would also like to thank Tuuli Lappalainen and Itsik Pe'er who, with Lars and Larry, provided excellent feedback that helped me revise my dissertation into its current form. Thank you to Ron Prywes and Sarah Kim Fein, and to the rest of the Columbia Department of Biological Sciences who provided me a home for the past seven years.

One of the greatest joys of my PhD has been the semi-regular calls with Joris van Arensbergen and Bas van Steensel. Your help and feedback have been invaluable, and

without your hard work and the work of so many other members of the van Steensel lab, none of this would have been possible.

My time in the Bussemaker lab has been one of many happy memories. I would like to particularly thank Chaitanya Rastogi, Judith Kribelbauer, and Tomas Rube for all your knowledge, advice, and great conversation. I would like to thank Xiaoting Li for a delightful collaboration. I would also like to acknowledge Ron Tepper, an inspiration who left us too soon.

Thank you to all my students, from whom I learned so much. My time teaching at Columbia and Cold Spring Harbor has been among my favorite moments these past few years. In particular, I would like to thank my CSHL co-instructor Sean Davis.

A very big thank you to Sarah, my wonderful wife. Thank you for being there through my ups and downs, my deadlines and revisions. You have been the best partner I could ever hope to have.

Finally, I would like to thank my family. You made sure I never lost my curiosity, and through your endless love and encouragement, you made me the man I am today.

1 Introduction

The human genome has on the order of 100,000 promoters, which are defined as regions of DNA capable of driving transcription [1]. At the most basic level, for protein-coding genes and many non-coding RNA (ncRNA) genes, promoters must include the position at which RNA polymerase II (RNAPII) initiates transcription. Modern RNA sequencing techniques allow us to identify the position of transcription start sites (TSSs) with single nucleotide precision [2]. While the locations of initiation events demonstrate the presence of a promoter in the immediate vicinity, identifying the specific parts within these promoter regions that are responsible for driving transcription initiation requires more information.

One common structural feature of promoters is the core promoter region, i.e., the region extending roughly 50bp up- and downstream of the TSS [3]. At the core promoter, general transcription factors (GTFs) bind to the DNA and assemble in a step-wise fashion before recruiting and positioning RNAPII for initiation. The complex of GTFs and RNAPII bound together at the core promoter constitute the pre-initiation complex (PIC). In addition to RNAPII, a minimal PIC includes the GTFs TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH [4]. Additional GTFs and associated proteins can be part of the PIC as well. PIC positioning is guided by the presence of core promoter elements, i.e., short DNA sequences that can be bound by components of the PIC. For example, the TATA-

box is typically located 30bp upstream of the TSS and is recognized by TATA-binding protein (TBP), a component of TFIID. Other core promoter elements found in vertebrates include the BRE motifs (bound by TFIIB and flanking the TATA-box), the Initator motif (Inr, overlapping the TSS), and the downstream core elements (DCEs) [4, 5].

While the concept of a core promoter region seems to suggest a trivial means of identifying the boundaries of promoter regions (i.e. locating core promoter elements or the binding sites of PIC proteins), the reality of transcription initiation is more complicated. It is rare for some of these core promoter elements to appear in the same promoter region, and different elements are typically associated with different promoter architectures [3, 6]. Additionally, while simpler organisms tend to have more fixed positions of initiation determined by core promoter elements, mammalian promoters show a greater diversity in the distribution of initiation sites. Mammalian promoters with TATA-boxes reflect the "focused" initiation sites of simpler organisms, but only include about 15% of promoters [4]. These tend to be associated with highly regulated genes. In contrast, housekeeping gene promoters rarely include clearly recognizable core promoter elements, and exhibit a "broad" pattern, where initiation can occur at any number of sites over a broader promoter region [3, 4]. These promoters tend to be associated with CpG islands, defined as regions of increased density of CpG dinucleotides. However, these CpG dinucleotides do not appear play a direct role in recruiting the transcriptional machinery, and can be associated with active or inactive promoters depending on their methylation state [7, 8].

Tightly regulated genes reveal another complication in using core promoter regions to define promoter boundaries. The underlying DNA sequences are present in cellular contexts where the core promoter is bound and active, as well as cellular contexts where the core promoter is not bound and inactive. Clearly, the presence of a core promoter alone is not sufficient to determine whether active transcription will occur. Identifying the local regions that influence the activity of the core promoter region can help delineate the broader promoter region.

One mediating factor in determining promoter activity is chromatin context. Chromatin is the macromolecular complex formed by DNA, histones and other proteins. An average of 147bp of DNA winds around each histone octamer to form a nucleosome [9]. Histones can be chemically modified to alter the degree to which nucleosome are packed together, which in turn can modify the accessibility of DNA. DNA accessibility is important for the formation of the PIC. Active promoters are generally associated with nucleosome-depleted regions (NDRs), flanked by the upstream –1 nucleosome and the downstream +1 nucleosome. In active NDRs, these flanking nucleosomes, and other nearby nucleosomes, tend to carry specific chromatin marks and histone protein variants [4, 9, 10]. Together, the position of nucleosomes and NDRs, their associated chromatin marks and histone variants provide some information about the structure of the promoter. Experiments like ChIP-seq and DNase-seq allow us to map the distribution of these chromosome features [4, 10]. However, these do not fully delineate which promoter sequences influence transcription activity. Histone modifications and positioning depend on the presence of other DNA-binding proteins, including specific transcription factors

other than the PIC-associated GTFs [10]. Only a subset of the sequence in and near the NDR may be necessary for modifying and positioning histones and recruiting the PIC. Additionally, the consistency of nucleosome positioning relative to the core promoter can differ across promoter classes [3, 4], suggesting that NDRs are an imprecise way of delineating precise promoter regions.

Due to the important role that non-GTF TFs play in recruiting the components of the PIC, the identification of transcription factor binding sites (TFBSs) in promoter regions is a strong indicator of which sequences may be driving transcription. TFBSs can be identified in several different ways. ChIP-seq experiments can identify the places where specific TFs bind *in vivo* [5]. Motif models can be constructed from prior binding experiments (including *in vivo* experiments like ChIP-seq and *in vitro* experiments like protein binding microarrays [11] and SELEX [12]), then used to predict binding based purely on DNA sequence. Finally, motif models can be combined with *in vitro* experimental data that identify accessible regions or sub-regions occupied by any DNA-binding proteins, such as DNase footprinting [13]. One important limitation to these approaches is that available datasets are limited to a subset of known TFs. While recent high-throughput methods have greatly expanded the number of TFs for which TFBS data are available [11, 14], these vary in quality and do not cover all of the thousands of known TFs. This makes it difficult to identify important regulatory sequences if they are bound by an uncharacterized TF.

Additionally, the identification of TFBSs does not directly implicate these regions in driving nearby transcription. Non-GTF TFs that regulate a nearby promoter tend to

bind upstream in what is called the proximal promoter region [5]. However, the presence of a TFBS in the proximal promoter region does not guarantee that the associated TF plays an active role in transcription initiation. TFs can bind in complexes called cis-regulatory modules (CRMs), and the composition of a CRM can alter the effect of individual TFs on downstream transcription. In some cases, the same TF can be associated with activation or repression depending on its binding partners [15]. Additionally, TFBSs can occur in regions outside of the proximal promoter, including in enhancers and the gene body itself [5]. Enhancer-binding TFs can influence the transcription of distal promoters via DNA looping. This is further complicated by the fact that enhancers are frequently transcribed themselves, either due to proximal TF binding or due to being brought into contact with the transcriptional machinery of promoters [16].

On the small scale, there is a simple experimental method that is capable of assessing the transcriptional activity of specific genomic sequences: promoter bashing. In promoter bashing, a candidate promoter sequence capable of driving transcription is inserted upstream of a reporter gene [17]. Subsequently, the promoter sequence is modified, either through mutation or deletion, usually on the 5' or 3' end [17]. The functional consequences of these changes are compared to the original construct either through transfection or genomic insertion, followed by expression and experimental detection of the reporter gene [17]. Deletions or mutations that result in decreases in transcription have probably affected functional sequences, while changes that do not cause a change in expression suggest that the altered DNA is non-functional. Some changes in the DNA sequences may even lead to increases in expression by removing

repressive elements or improving the spatial organization of the construct. If performed iteratively, promoter bashing can produce a minimal regulatory element, capable of performing the functional role of the original sequence without including unnecessary sequences [17]. A similar approach can probe the functional properties of enhancers by adding a target promoter to the plasmid construct. These minimal promoter and enhancer elements can be useful in future experiments.

While promoter and enhancer bashing are useful, they are labor-intensive and can only be used to probe a small number of regulatory regions at a time. In contrast, a number of massively-parallel reporter assays (MPRAs) have been developed to probe the regulatory activities of thousands to millions of specific genomic regions in parallel [18, 19]. Typically, a large number of regulatory sequences are inserted into a plasmid construct similar to those used in promoter bashing, except for the addition of a unique DNA sequence that ensures that each reporter transcript can be distinguished from all others [20]. This is often a short random barcode sequence that can be mapped back to the inserted regulatory region by sequencing the plasmid library. Upon transfection, this barcode is expressed as part of the reporter construct, and can then be selectively sequenced to measure the relative expression driven by the associated regulatory element. A notable exception is STARR-seq, which probes enhancer activity by inserting enhancer sequences downstream of the reporter gene itself [21]. Enhancers capable of driving expression of an upstream minimal promoter lead to their self-transcription. Unfortunately, this approach has limited approach in promoters, which typically transcribe in a downstream direction and must therefore be inserted upstream of the

reporter construct. Thus, promoter-based MPRAs typically include a barcode-based reporter construct architecture [20]. These plasmid libraries are transfected en masse into cell culture, followed by isolation and sequencing of expressed reporter constructs, allowing for the parallel measurement of the relative expression of the entire library.

Promoter-based MPRAs vary in their ability to accurately discriminate functional regions, both locally and on a genome-wide scale. Assays that test tens to hundreds of thousands of sequences have been applied to small sections of the genome, and typically are targeted at specific regions thought to contain functional elements *a priori* [22]. This can provide useful information about the specific boundaries of functional elements within these targeted regions, just as promoter bashing does in individual cases. However, it lacks the genome-wide scale that might be useful to the broader research community, whose targets of interest lie outside the selected regions.

To perform a promoter activity assay that can reveal the location of specific functional elements throughout the genome, the total number of elements tested must be several orders of magnitude larger. To this end, we have developed Survey or Regulatory Elements (SuRE), a genome-wide promoter-based MPRA that has been used to probe the activity of hundreds of millions of human genomic elements in parallel. SuRE is the subject of this dissertation. My primary focus has been the statistical analysis of SuRE data, which has aided in the development and analysis of SuRE experiments by my colleagues in the lab of Bas van Steensel at the Netherlands Cancer Institute.

In Chapter 2, I will introduce statistical background that is important to understanding my approach to modelling SuRE data. This includes basic information on

generalized linear models (GLMs), as well as a summary of some common regularization methods that can improve predictions of GLMs in certain contexts.

Chapter 3 is an overview of the SuRE protocol, as well as a summary of some initial results generated by myself and my colleagues and originally published in *Nature Biotech*. Our approach to SuRE experimental data produces a map of normalized autonomous promoter activity, which can reveal insights into the functional organization of promoters and enhancers throughout the genome. My contributions were particularly focused on the spatial patterns of transcription at promoters. I analyzed the global expression patterns of elements at various positions relative to annotated transcription starts sites, as well as and in conjunction with the autonomous activity of bidirectional promoter pairs. Also discussed are my initial attempts to model the high-resolution spatial promoter activity of specific promoter regions using penalized GLMs. Finally, I explored the specific relationship between genome-wide patterns in CpG density and SuRE expression.

In Chapter 4, I develop the penalized SuRE-GLM approaches used in Chapter 3 by scaling these models up to the entire genome. Leveraging the same SuRE experimental data used in Chapter 3, I generate a higher-resolution genome-wide promoter activity track that can be used to accurately predict expression of novel reporter constructs. This track is also used to provide finer-scale insights into spatial patterns of promoter organization. I also integrate results from other SuRE experiments into a novel multivariate SuRE-GLM approach, which allows for the analysis of differential expression across multiple cell culture conditions or cell types. Then, I demonstrate the

use of SuRE-GLM to identify regulatory SNP-variants by combining SuRE experiments from separate genomes.

Finally, in Chapter 5 I discuss future directions for the analysis of SuRE results.

2 Generalized Linear Models

2.1 Introduction

Generalized linear models (GLMs) are a general class of model that extend the basic regression framework of linear models to include response variables with error distributions that are not normal. GLMs are a powerful tool for identifying the relationship between biological datasets and a wide variety of explanatory variables. Of particular interest are GLMs that can be used to model count data, of which the normal distribution is often a poor approximation. GLMs based on discrete distributions, such as the Poisson, binomial, and multinomial distributions, may be appropriate in different contexts. In this chapter, we describe various conceptual and technical aspects of GLMs that are relevant to the analyses presented in Chapters 3 and 4.

2.1.1 Linear models

In a linear model, an observation (y_i) is modelled as being drawn from a normal distribution centered on a mean (μ_i):

$$y_i = \mu_i + \epsilon_i$$

Here, ϵ_i is the normally-distributed error term. The mean μ_i is not bounded, and ϵ_i is drawn from a continuous distribution, which means that the expectations and observations in a linear model can take on any real value, positive or negative [23].

The mean is specified as the linear combination of the products of one or more covariates (i.e. explanatory variables, x_{ik}) and their respective coefficients (β_k):

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_K x_{iK} = X_i^{\top} \boldsymbol{\beta}$$

Here β_0 represents an intercept shared across all observations. The interpretation of coefficients is straightforward: a unit change in x_{i1} produces a linear change of β_1 in the expected mean of y_i. A positive relationship between X_k and $\boldsymbol{\mu}$ is indicated by a positive coefficient, while a negative relationship results in a negative coefficient. However, the relative magnitude of a coefficient is not directly informative about the magnitude of the effect of the corresponding covariate on the responses, since different covariates can vary on different scales. For this reason, covariates are often standardized prior to the regression, so that coefficients uniformly represent the change expected in the response given a change of one standard deviation in the covariate [23].

For a set of observations, the likelihood function for the model is:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{e^{-(y_i-\mu_i)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

and the log-likelihood is:

$$\ell(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} -\frac{(y_i - \mu_i)^2}{2\sigma^2} - \log\left(\sigma\sqrt{2\pi}\right)$$

The left-hand side of this expression depends on the sum of squares $\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(y_i - \mu_i)^2$, where each term in the sum is the square of the difference between an observation and the corresponding mean, i.e. the error term ϵ_i.

In ordinary least squares (OLS) linear regression, the coefficients of a linear model are estimated by minimizing the sum of squares, thus maximizing the likelihood function. This minimum has a closed form, so the estimates can be achieved using simple

matrix operations [23]. This makes OLS regression particularly efficient relative to other modelling approaches, which do not have closed-form solutions [24].

2.1.2 Generalizations

Generalized linear models model observations in similar ways as a linear model, but with two key differences. The first is how the parameters of the GLM are specified by the covariates. In a GLM, we also calculate a linear sum of contributions of each covariate to each response:

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_K x_{iK} = X_i^{\mathsf{T}} \boldsymbol{\beta}$$

However, while in a linear model, $z_i = \mu_i$, the mean parameter of the normal error distribution, in a GLM the relationship between this sum z_i and the estimated parameters of the error distribution θ_i need not assume this functional form [24]. The link function $g(\theta)$ describes the transformation needed to apply to θ_i to return it to z_i. For example, in a model with a log-link function, z_i corresponds to the natural logarithm of the associated parameter: $g(\theta_i) = \log(\theta_i) = z_i$. To calculate the parameter θ_i given z_i, we use the inverse of the link function. In the case of the log-link, the inverse is the exponential function, so $\theta_i = g^{-1}(z_i) = e^{z_i}$. In a linear model, the link function is called the identity function, which is its own inverse.

The second major difference between a given GLM and a linear model is the error distribution, which defines the probability of observing y_i given the parameter θ_i [24]. A linear model uses a normal distribution as its error function, while a GLM can use a variety of different distributions. This includes continuous and discrete distributions, with infinite and non-infinite support. In the case of distributions with more than one

parameter, each can be modelled separately with its own link function and coefficients, or one or more can be modelled as fixed across all observations, as with the standard deviation parameter in a linear model [24]. By allowing for non-normal error distributions, GLMs can better reflect the structure of the observations, and can capture different relationships between the parameters and the properties of the response variables, such as relationships between the variance and the mean of a distribution.

In most cases, a "canonical" link function exists for a given distribution [24]. The canonical link functions are the most commonly used link functions for their respective distributions, although others can be used. The canonical link function usually reflects the properties of the distribution, and allows for straightforward interpretation of covariates. For example, the mean parameter of the Poisson distribution can assume any positive real value, and therefore the log-link function ensures that the mean will assume values within this support.

As in a linear model, the coefficients of a GLM are estimated by maximizing the log-likelihood, or equivalently minimizing the negative log-likelihood. However, in most cases this solution does not have a closed form, and iterative methods must be used to estimate the coefficients [24].

2.2 Families

2.2.1 Poisson distribution

The Poisson distribution is a discrete probability distribution used to model the probability of a observing a fixed number of independently occurring events given their

rate. For example, if the we assume that the transcription rate of a given gene is fixed and transcription is rare enough that subsequent transcription events are independent, then the number of transcripts produced after a fixed amount of time within a cell would be Poisson distributed. It is parameterized by single parameter, the mean λ. The probability mass function is given by:

$$f(k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

An important property of the Poisson distribution is that the mean is equal to the variance. This means that higher rates produce higher variances, a property observed in many sequencing datasets.

In Poisson regression, the canonical link function for the Poisson distribution is the log-link. Thus $\log(\lambda_i) = z_i = X_i^{\mathsf{T}}\boldsymbol{\beta}$, and $\lambda_i = e^{X_i^{\mathsf{T}}\boldsymbol{\beta}}$. A consequence is that an expected mean rate λ_i can only assume positive real values. Each covariate has a multiplicative effect on the expected mean, so that a unit increase in a covariate with coefficient β produces a multiplicative change of e^{β}. When the coefficient is positive, this produces a larger rate; when the coefficient is negative, the rate shrinks.

The likelihood function in Poisson regression is:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

and the log-likelihood function is:

$$\ell(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} -\lambda_i + y_i \log(\lambda_i) - \log(y_i!) \propto \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} -\lambda_i + y_i \log(\lambda_i)$$

The Poisson distribution is commonly used for modeling DNA sequence counts [25], especially when the observed count for any one gene or locus is dwarfed by the total sequencing depth.

2.2.2 Binomial distribution

The binomial distribution is a discrete distribution used to model the outcome of one or more Bernoulli trials, which result in either a "failure" (0) or a "success" (1). Given a fixed probability of success p and total number of trials n, the binomial distribution describes the probability of observing a given number of successes k, where $0 \leq k \leq n$. The probability mass function is:

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The expected number of successes is $E(k) = np$. The variance of the binomial distribution is $V(k) = np(1-p)$. Thus, for fixed p, the variance increases linearly with the sample size n (and therefore with the mean as well). For fixed n, the variance is maximized at $p = \frac{1}{2}$ and decreases non-linearly as p approaches 0 or 1.

The binomial and Poisson distributions are related in two important ways. First, for large n and small p, the Poisson distribution with $\lambda = np$ serves as a good approximation of the binomial distribution. As this is often the case for large sequencing datasets (e.g. the proportion of all RNA-seq reads corresponding to a single gene is small), the Poisson distribution can often be substituted even when the binomial

distribution is technically more appropriate. Second, given two Poisson-distributed variables with means λ_1 and λ_2 and total count $n = k_1 + k_2$, the count k_1 is a binomial-distributed variable with probability $p_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2}$.

In binomial regression, n_i is known and fixed for each observation k_i, while p_i is estimated using a link function that restricts it to values between 0 and 1. Most often, the canonical logit or log-odds link function is used, such that

$$\log\left(\frac{p_i}{1-p_i}\right) = z_i = X_i^{\mathsf{T}}\boldsymbol{\beta}$$

When $z_i = 0$, $p_i = \frac{1}{2}$. As z_i increases, p_i approaches 1. Similarly, p_i approaches 0 as z_i decreases. The inverse of the logit link function,

$$p_i = \frac{1}{1 + e^{-X_i^{\mathsf{T}}\boldsymbol{\beta}}}$$

is a logistic function. This is why binomial regression models that use the logit link are often referred to as logistic regression.

The likelihood function in binomial regression is:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \binom{n_i}{y_i} {p_i}^{y_i}(1-p_i)^{n_i - y_i}$$

and the log-likelihood function is given by:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log\left(\binom{n_i}{y_i}\right) + y_i \log(p_i) + (n_i - y_i)\log(1-p_i)$$

2.2.3 Multinomial distribution

The multinomial distribution extends the binomial distribution to cases with more than two outcomes. Each outcome is assigned its own probability p_j, such that $\sum_{j=1}^{J} p_j = 1$. The probability mass function is:

$$f(\boldsymbol{k}) = \frac{n!}{\prod_{j=1}^{J} k_j} \prod_{j=1}^{J} {p_j}^{k_j}$$

Just as the counts of two Poisson-distributed variables are binomial-distributed given their sum, the counts of J Poisson-distributed variables are multinomial-distributed given their sum. In such a case,

$$p_j = {\lambda_j}\Big/{\sum_{j\prime=1}^{J} \lambda_{j\prime}}$$

In multinomial regression, the parameters of the distribution are linked to the covariates through a functional form that reflects this relationship:

$$p_j = \frac{e^{z_{ij}}}{\sum_{j\prime=1}^{J} e^{z_{ij\prime}}},$$

where

$$z_{ij} = {X_i}^{\intercal} \boldsymbol{\beta}_j$$

Since the sum of p_j is restricted to 1, often the first outcome is used as the base case and z_1 is fixed to 0. For $j > 1$, each outcome has its own set of coefficients. In the case of two outcomes, we have

$$p_{i2} = \frac{e^{z_{i2}}}{e^{z_{i1}} + e^{z_{i2}}} = \frac{e^{z_{i2}}}{e^{0} + e^{z_{i2}}} = \frac{e^{z_{i2}}}{1 + e^{z_{i2}}} = \frac{1}{1 + e^{-z_{i2}}}$$

This last term is the same as in binomial logistic regression. As such, multinomial regression is often referred to as multinomial logistic regression.

2.3 Regularization in GLMs

When fitting GLMs, several problems can emerge due to the structure of the model and properties of the data. In some cases, these problems will cause common GLM regression software to fail to converge to an estimate of the model parameters within a reasonable time frame. In other cases, the models will converge but will produce poor predictions or erroneous interpretations of the results. Fortunately, there are extensions of the standard GLM approach that can help address many of these challenges. These extensions are called regularization.

Regularization introduces additional constraints to the GLM by penalizing some property of the coefficients. Different regularization methods are used to fit models with different properties, and to address different problems. While many different regularization schemes exist, this section will focus on three of the most popular: LASSO, ridge, and elastic net penalization.

2.3.1 LASSO

Least absolute shrinkage and selection operator (LASSO) [26] is a regularization approach that places a penalty on the $\mathcal{L}_1$-norm, or sum of the absolute value, of the coefficients in a model. In applying LASSO to GLMs, optimization of the objective function takes the following form:

$$\min_{\boldsymbol{\beta}} \left[-\frac{1}{N} \ell(\boldsymbol{\beta}) + \lambda_1 \sum_{k=1}^{p} |\beta_k| \right]$$

Here $\ell(\boldsymbol{\beta})$ is the log-likelihood function from the unpenalized GLMs discussed above, N is the number of observations, p is the number of coefficients (usually excluding β_0, the intercept), and λ_1 is the $\mathcal{L}_1$ penalty parameter that must be selected *a priori*.

In general, this penalization method tends to shrink coefficients towards zero [26]. This "shrinkage" effect helps address a common problem in regression analysis: overfitting. Overfitting occurs when a model is fit too exactly to the sample data such that the model does not generalize [24]. While only a fraction of the variance in a sample can be explained by a given set of covariates, an overfit model will erroneously assign some of the residual variation to these covariates as well. This will tend to produce errors when the model is used to make predictions based on new data. With LASSO regularization, the $\mathcal{L}_1$ penalty partially counteracts the effects of overfitting [26]. Even when increasing the magnitude of a coefficient would improve the log-likelihood for a small number of observations by overfitting, these improvements must be fairly large in order to overcome the resulting increase in the coefficient penalty. As a result, coefficients in a LASSO model reflect conservative estimates of the true coefficients. This is particularly useful in cases where $p > N$, which leads inevitably to overfitting in unpenalized models.

In many cases, a LASSO model will result in some fraction of all coefficients being set to exactly zero [27]. This sparsity can be a useful property of LASSO models. In some situations, *a priori* information suggests that only a subset of all covariates is expected to contribute to the outcome. For example, if we tried to predict gene expression

based on the presence of many different transcription factor binding motifs in gene promoters, we might reasonably assume that only a subset of all transcription factors is active in a given cellular context, even if we are unaware of which transcription factors are active *a priori*. Unlike an unpenalized model, a LASSO model will assign non-zero coefficients to only a fraction of the covariates, separating the covariates into active and inactive sets. This is a form of feature selection, allowing for the identification of a subset of covariates of interest.

LASSO requires a single λ_1 be chosen before fitting a model. Obviously, it is difficult to know the appropriate choice beforehand. In the extreme cases, a very large $\lambda_1$will produce a null or intercept model, where all coefficients are set to 0, while a very small $\lambda_1$will produce a model that approaches the unpenalized model. Selection of $\lambda_1$is usually based on cross-validation [26]. In most cases, the cross-validated error of models using a series of $\lambda_1$values will be minimized at some point between these two extreme cases, suggesting that the predictive power of our model is most improved at a particular non-zero value of λ_1.

2.3.2 Ridge regression

Ridge regression, also known as Tikhonov regularization, is a regularization method that places a penalty on the $\mathcal{L}_2$-norm, or sum of squares, of the model coefficients:

$$\min_{\boldsymbol{\beta}} \left[-\frac{1}{N} \ell(\boldsymbol{\beta}) + \lambda_2 \sum_{k=1}^{p} {\beta_k}^2 / 2 \right]$$

As in the case of LASSO, ridge regression can help address over-fitting by shrinking the coefficients towards zero [28]. Due to the fact that the square of a small coefficient will produce a penalty very close to zero, ridge regression generally does not set any coefficients to zero. This makes ridge regression a poor choice when feature selection is desired. However, one advantage of ridge regression over LASSO is that it helps to address (multi-)collinearity.

Collinearity occurs when some covariates are highly correlated. In the extreme case, consider two identical covariates. Assume that an unpenalized model that included only one of these covariates would assign the coefficient β^*. When both covariates are included in an unpenalized model, any values (β_1, β_2) such that $\beta_1 + \beta_2 = \beta^*$ will produce a model with identical results. Without a unique optimal solution to the objective function, an unpenalized model will fail to converge. Even in cases where only partial collinearity exists, unpenalized methods can result in convergence issues.

With LASSO regression, identical covariates can still give rise to convergence issues. For example, a model where $\beta_1 = \beta^*$ and $\beta_2 = 0$ will produce the same result as a model where $\beta_1 = 0$ and $\beta_2 = \beta^*$. In cases where covariates are collinear but not identical, LASSO may converge, but will often set some of these covariates to zero in an arbitrary manner that is highly sensitive to noise. In contrast, ridge regression tends to have a "grouping effect", such that correlated covariates receive similar coefficients. This better reflects the relationship between each covariate and the response variable than arbitrarily assigning this effect to a single covariate.

As in LASSO, the λ_2 penalty parameter used in ridge regression must be selected *a priori*. This is typically accomplished in a similar manner using cross-validation.

2.3.3 Elastic net

Elastic net regularization combines both the $\mathcal{L}_1$ penalty of LASSO and the $\mathcal{L}_2$ penalty of ridge regression in a single model [27]. Rather than using λ_1 and λ_2, these penalties are reparameterized using $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}$ such that $\lambda > 0$, $0 \leq \alpha \leq 1$. The objective function is therefore:

$$\min_{\boldsymbol{\beta}} \left[-\frac{1}{N} \ell(\boldsymbol{\beta}) + \alpha\lambda \sum_{k=1}^{p} |\beta_k| + (1-\alpha)\lambda \sum_{k=1}^{p} {\beta_k}^2 / 2 \right]$$

By combining the $\mathcal{L}_1$ and $\mathcal{L}_2$ penalties, elastic net regression produces models that have the advantages of both ridge and LASSO. Some feature selection occurs due to the $\mathcal{L}_1$ penalty, so a subset of coefficients will be set to zero. Both penalties produce shrinkage that helps avoid overfitting. The $\mathcal{L}_2$ penalty produces a grouping effect that encourages correlated covariates to have similar coefficients.

The relative strength of these effects can be tuned using the α parameter. In the extreme case of $\alpha = 0$ or 1, we have pure ridge or LASSO regression, respectively. Intermediate values of α produce a compromise between the two. For a given α value, Friedman et al. [29] have developed a cyclical coordinate descent algorithm that can efficiently fit models over a "λ path" of decreasing λ values. Using this algorithm, cross-validation can be used to select an optimal λ value given α, and this procedure can be repeated at a series of α values to validate the two penalty parameters.

2.4 Alternative approaches

In addition to the methods mentioned above, the GLM framework has been extended in many different ways. In designing the models used in this thesis, I considered but ultimately rejected a number of these methods for practical reasons. Nevertheless, these methods have some compelling properties, and may occur to readers with a statistical modelling background. For these reasons, I will discuss a few of these modelling approaches below.

2.4.1 Overdispersed count distributions

A commonly observed feature of sequencing experiments is overdispersion, i.e. when the variability of a dataset exceeds what is expected given a statistical model [25]. This may be the result of underlying biological heterogeneity orthogonal to the variables of interest, or due to experimental processes such as PCR duplication. Regardless of the source, GLM approaches that relax the variance assumptions of the default count distributions can be useful in modelling sequencing datasets.

As alternatives to the Poisson distribution, there exist several common GLM extensions that are used to model overdispersed count data. The quasi-Poisson approach assumes a linear relationship between the variance and the mean, such that $V(k|\lambda) = \theta\lambda$ where λ is the expected value for k and $\theta \geq 1$. This approach has been used previously to model RNA-seq data [30]. However, it is important to note that the quasi-likelihood approach does not correspond to the likelihood function of any known probability distribution. This makes it more difficult to interpret the results of a quasi-Poisson model.

A more popular Poisson alternative is the negative binomial (NB) distribution. NB has been employed extensively for modelling overdispersed RNA-seq data [25]. In the NB distribution, $V(k|\lambda) = \lambda(1 + \theta\lambda)$. Unlike the quasi-Poisson approach, the NB corresponds to a well-characterized probability distribution. Another interesting property of the NB distribution is that it can be considered as a Poisson mixture distribution, where the underlying mixture of Poisson means is distributed according to the two-parameter gamma distribution [30].

For the binomial distribution, a popular overdispersion model is the beta-binomial distribution [31, 32]. As the name suggests, the beta-binomial distribution can be understood as a mixture distribution, where the success probability p of the binomial-distributed variable is itself distributed according to a two-parameter Beta distribution.

For the multinomial distribution, a common way to deal with overdispersion is to use the Dirichlet-multinomial distribution. Just as the multinomial is the multivariate extension of the binomial distribution, the Dirichlet distribution is the multivariate extension of the Beta distribution. The Dirichlet-Multinomial is commonly used for machine learning applications such as topic modelling [33], and has seen some applications in the modelling of biological sequencing data [34].

All the models discussed above have been implemented in R, the programming platform used for my analyses. The quasi-Poisson is implemented as part of the base R function glm() from the stats package. Negative binomial regression can be implemented using the glm.nb() function in the MASS package [35], or using specialized sequencing packages like DESeq2 [36]. Beta-binomial regression has been implemented in VGAM

[37] and Dirichlet-multinomial regression has been implemented in the MGLM package [38]. While our data showed some signs of overdispersion, initial testing showed that the approaches described here could not be scaled up to the necessary scale. Additionally, all the overdispersed models described here lacked an implementation with flexible regularization options. Given the collinearity of the covariates in our models, I deemed it necessary to use a model that included some form of penalization.

One overdispersion model that I have neglected to mention is the linear (or normal) model, and the related multivariate normal model. Linear models can be used for modelling count data [39], and both the univariate and multivariate normal distribution have been implemented as part of glmnet [29], making these models just as easy to implement as their penalized Poisson and multinomial counterparts. However, there are a couple of problems with applying a normal model to count data. For one, a normal model assumes that there is a uniform variance regardless of predicted mean. In our data, there is a clear relationship between variance and mean. Second, normal models make predictions that do not make sense in the context of count data, such as predicting a negative mean. These properties make a count distribution preferable to a normal distribution in our case.

2.4.2 Zero-inflated distributions

Our data showed signs of zero-inflation, where values of zero occurred more often than we would expect given our count distribution models. This may be due to the fact that, as will be explained later, these experiments require transfection of a plasmid into a cell before any expression can occur. Observed counts of zero therefore represent a

mixture of plasmids that failed to transfect and plasmids that transfected but did not express. Ideally, we could implement a model that accounts for this zero-inflation.

Zero-inflated Poisson models have been applied to RNA-seq datasets [40], and implementations in R exist [41]. A similar approach is used for the zero-inflated negative binomial model [41]. These approaches model the probability of a structural zero (e.g. probability of no transfection in our case) using a logit link-function, as in binomial regression. Unfortunately, these R implementations suffer from the same drawbacks as the overdispersed models discussed above: they are difficult to scale, and lack penalization options.

One alternative method I considered to address our zero-inflation was a compound Poisson-Poisson distribution. This distribution is sometimes called the Neyman Type A distribution [42]. With this model, I assume that the number of transfection events N_i per plasmid is Poisson distributed based on some global mean θ, and then the observed response variable is the sum of N_i Poisson-distributed counts with a mean λ_i dependent upon the specific covariates associated with plasmid i. This model has the advantage of capturing both zero-inflation (when $N_i = 0$) and overdispersion (variance has a similar relationship with the mean as in the negative binomial distribution). While this may represent my preferred model for this data, no R implementations exist for this distribution.

2.4.3 Alternative regularization approaches

In the model discussed below, we used elastic net penalization to regularize the coefficients corresponding to different spatial bins along the genome. This penalization

approach helps address the inherent collinearity present in adjacent bins, but the model does not explicitly include the spatial relationship between bins. Given that we might expect regulatory sequences to co-locate (e.g. clusters of TFBSs), I considered alternative models that took spatial relationships into account as part of the penalization.

Smoothing splines are used to find a smooth polynomial (most often cubic) function that models the relationship between a covariate and a response variable subject to some smoothing penalty that mediates between a perfect interpolation and a linear relationship [43]. There are base R functions that allow splines to be used in GLMs via the glm() function. By using genomic position as a covariate, such a model could fit a continuous function relating each position in the genome to an activity level, such that adjacent genomic positions would have similar activity levels. However, genomic fragments do not overlap a single genomic position, but a range of positions. The base R implementation of smoothing splines does not allow for an integration over a range of covariate values, making this approach intractable.

An alternative penalization scheme that allows for the inclusion of structural information is the fused LASSO. Rather than penalizing the absolute value of covariates, fused LASSO penalizes the absolute value of the difference of adjacent covariates [44]. R implementations of fused LASSO exist, as in the genlasso package. However, initial tests showed some issues with scaling, and GLM extensions were unavailable. For these reasons, I chose to implement an elastic net model.

3 Survey of Regulatory Elements

This chapter was reproduced from the following publication:

> J. van Arensbergen, **V.D. FitzPatrick**, M. de Haas, L. Pagie, J, Sluimer, H.J. Bussemaker†, and B. van Steensel†
> Genome-wide mapping of autonomous promoter activity in human cells.
> ***Nat. Biotechnol.*** 35(2):145-153 (2017)

The author of this thesis contributed to many aspects of the computational analyses in this paper, and in particular was solely responsible for Figures 3.3, 3.4, and 3.6d in this chapter.

3.1 Introduction

Promoters harbor the transcription start site (TSS) and various other sequences that control transcription initiation through the binding of trans-acting factors [45]. Various genome-wide methods have been developed to map endogenous promoter activity [2, 46-48]. 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 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 [49-55]. 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 [18, 19, 56], 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 cells [57]. 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 [22]. At substantially higher throughput, near-saturating coverage of the entire *Drosophila* genome was achieved with STARR-seq [21, 58]. 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.

3.2 Methods

3.2.1 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 (Figure 3.8b) 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.

3.2.2 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 16μl water. Four μl was electroporated into 20ul bacteria and transferred to 250 ml LB plus kanamycin (50μl/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.

3.2.3 SuRE library characterization by iPCR

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.

3.2.4 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 Mycoplasm 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 µM hemin (Sigma; #51280-1G) or solvent control 1 hour after nucleofection and cells were harvested 24 hours later.

3.2.5 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 minutes (#04716728001; Roche) and DNase I was inactivated by addition of 1µl 25mM EDTA and incubation at 70ºC for 10 minutes. Next, cDNA was produced by first adding 1 µl of 10 µM gene specific primer targeting the GFP ORF (247JvA) and 1 µl dNTP (10mM each) and incubating for 5 minutes 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 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 doen in parallel per biological replicate. Each 20 µl reaction was then PCR amplified (1' 96 °C, 20x(15'' 96 °C, 15'' 60 °C, 15'' 72 °C)) in a 100 µl 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 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.

3.2.6 Mapping of iPCR sequencing data

Paired-end reads are trimmed, using cutadapt (version 1.2.1), 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) [59], 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 where converted to bedpe files using custom scripts.

3.2.7 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 2 replicate PCRs each on 2µl 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.embnet.org/index.php/embnetjournal/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 BioConductor [60].

3.2.8 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 ($r^2 = 0.98$; Figure 3.9e). 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.

3.2.9 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 µl 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 65C° 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.

3.2.10 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 [1]. 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 +/– 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 3.2a) were also generated using CoverageView, using 50bp bins, except for the PRO-seq data in Figure 3.6e 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 Figure 3.1d, 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 3.3c), 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 3.5b-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 3.2b,c and Figure 3.5b 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 [61]. 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 ±500bp from their center.

For the single-locus analysis in Figure 3.3a and e, only genomic fragments are shown that were detected in the cDNA. In Figure 3.6e 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 3.6a we focused on the mappable part

of the genome which we obtained by concatenating all adjacent 36-mer mappable regions from ENCODE (wgEncodeCrgMapabilityAlign36mer.bw).

3.2.11 Penalized Generalized Linear Modeling

To create Figures 3.3b,d,f,h-j and 3.6d, 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 3.3g, we used stable/unstable peak pairs identified in K562 GRO-cap [48], 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 [62].

3.2.12 Data sources

- As a reference for transcription start sites (TSSs) we used GENCODE [63] version 19 TSSs (gencode.v19.annotation.gff3.gz) downloaded from http://www.gencodegenes.org/releases/19.html.
- FANTOM phase 1 data [1] was downloaded from http://fantom.gsc.riken.jp/5/tet/#!/search/hg19.cage_peak_counts_ann_decoded.osc.txt.gz.
- ENCODE chromHMM annotations [64] in K562 were downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgSegmentation/wgEncodeAwgSegmentationChromhmmK562.bed.gz.
- CAGE data [64] (wgEncodeRikenCageK562CellPapAlnRep1.bam) was downloaded from http://moma.ki.au.dk/genome-mirror/cgi-bin/hgFileUi?db=hg19&g=wgEncodeRikenCage.
- GRO-cap data [48] (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).

3.2.13 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) [65] 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").

3.2.14 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 [60].

3.2.15 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.

3.2.16 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.

3.2.17 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 3.1f and 3.6a 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 [66]). The overlap distribution in 10,000 random circular permutations was used to compute a p-value for enrichment.

The p-values in Figure 3.1e,f and 3.5e,f refer to the p-value of the Pearson correlation.

3.3 Results

3.3.1 SuRE method and library preparation

The SuRE experimental strategy consists of three main steps (Figure 3.1a, Figure 3.7). 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 (Figure 3.7). 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 (Figure 3.1a). 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 (Figure 3.8a), with 96% of the mappable genome covered at least 15-fold (Figure 3.8a).

3.3.2 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 Figure 3.7.

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 (Figure 3.1b). A peak detection algorithm [65] 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 genome [64] (Figure 3.1c, Figure 7c). 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 genome [67] (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 3.8a). 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 3.9b). 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 3.9c). 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 $r^2 = 0.73$ (Figure 3.9d).

3.3.3 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 method [48]. We focused on a curated set (see Methods) of 28,844 TSSs annotated by the GENCODE project [63]. Notably, SuRE and GRO-cap signals are substantially correlated ($r^2 = 0.54$; Figure 3.1d). 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 method [64] ($r^2 = 0.49$; Figure 3.8d). 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 (Figure 3.1e). Yet, we also identify many housekeeping promoters with a low level of promoter autonomy, for example promoters of genes that encode histones (Figure 3.8e). Relative promoter autonomy is inversely correlated with the number of enhancers near the promoters in the native genomic context (Figure 3.1f). This cannot be

explained by differences in local gene density (Figure 3.8f). These results support the notion that autonomous promoters as detected by SuRE are less dependent on distal enhancers than non-autonomous promoters.

3.3.4 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 [2]. 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 (Figure 3.2a-c). The antisense activity is on average 2-3 fold weaker but it correlates with the sense activity (Figure 3.2b-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.

3.3.5 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 (Figure 3.3a), 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 (Figure 3.3b).

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 (Figure 3.3c, 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 (Figure 3.8j) 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 (Figure 3.3d, blue curve). These analyses illustrate how SuRE data can be used to identify critical sequence regions within promoters, both individually and genome-wide.

3.3.6 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 [48, 68, 69]. Indeed, two antisense core promoters were found to drive transcription autonomously *in vitro* [68]. On the other hand, the sense and antisense core promoters have been proposed to function in a cooperative manner [69]. 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 (Figure 3.3e). GLM confirmed this conclusion and found no evidence that the antisense TSS subregion is needed for antisense transcription (Figure 3.3f). Indeed, genome-wide analysis shows that promoter fragments that include the forward core promoter generally exhibit the highest SuRE activity in antisense orientation (Figure 3.3c, 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 (Figure 3.3d, red curve). Analysis of a well-defined set of sense-antisense TSS pairs [48] (Figure 3.3g) 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 (Figure 3.3h). 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 (Figure 3.3i). Finally, the sense and antisense TSSs at the *HIST1H2BD* gene are each primarily driven by distinct local sequence elements (Figure 3.3j). Thus, exceptions exist to the general rule that antisense transcription is driven by sequence subregions nearby the sense TSS.

3.3.7 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 [70]. CpG content has previously been linked to promoter activity [55, 71]. When binned by their observed and expected CpG density (Figure 3.4a), 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 promoters [71]. However, the relationship between SuRE expression level and CpG content for individual fragments takes a different form (Figure 3.4b):

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 Figure 3.4b).

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 [7]. Our data suggest that autonomous promoters have been protected from this loss, presumably because they have remained consistently hypomethylated in the germline throughout evolution.

3.3.8 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 [16, 72]. For a subset of enhancers, stimulus-induced eRNA production precedes mRNA transcription from the target promoters [73, 74], 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 [72, 73, 75] and it has been shown that enhancer transcription can be dependent on the presence of the target promoter [76]. 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 [77], showed several clear bi-

directional SuRE signals (Figure 3.5a). Analysis of 47,020 predicted active enhancers in K562 cells [64] revealed SuRE signals for the majority (Figure 3.5b,c), although the overall level of activity is approximately 10-fold lower than for promoters (cf. Figure 3.3a and Figure 3.5c; Figure 3.5d). 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* [78].

Notably, the ENCODE classification of enhancers as 'weak' or 'strong' [64] correlates with the strength of SuRE signals ($p < 2.2 \times 10^{-16}$, Wilcoxon test) (Figure 3.5b-d). SuRE signal also correlates positively with the endogenous levels of H3K27ac (Figure 3.5e), the histone modification most characteristic of active enhancers [79]. Furthermore, the ability of ~130 bp fragments derived from ENCODE-annotated enhancers to activate a minimal promoter in a previous reporter assay [61] shows a significant ($p = 4 \times 10^{-4}$) positive correlation with the SuRE signal for the same enhancers (Figure 3.5f). These results indicate that the level of autonomous transcription initiation from enhancers is related to enhancer strength.

3.3.9 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 expression [80]. Treatment of K562 cells with hemin is known to increase expression of several of the genes in the alpha-globin locus [81], which we confirmed by RT-qPCR (Figure 3.5g). Although R2 can be activated by hemin [82], 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 3.5h) 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.

3.3.10 Autonomous promoter activity in repetitive elements

ENCODE-annotated promoters and enhancers in K562 account for only 26% of the genome-wide SuRE peaks (Figure 3.8i). 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 (Figure 3.6a), 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 (Figure 3.6b, 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 (Figure 3.6b, c).

Note that technologies like CAGE and GRO-cap have difficulty mapping transcription initiation activity uniquely to specific repeat instances in the genome [83], 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 (Figure 3.5a). In addition, GLM analysis of partially overlapping SuRE fragments, similar to what we applied to promoters (cf. Figure 3.3d), pinpointed the precise sequence regions that generally contribute to autonomous promoter activity across hundreds of LTR12C variants (Figure 3.6d). These data extend earlier analyses of single LTRs [84] and again indicate that essentially the same sequence elements contribute to sense and antisense transcription.

Sense-oriented run-on transcription [48] is detectable downstream of LTR12C insertions with high SuRE activity (Figure 3.6e). This is not found for insertions with low SuRE activity and not in the antisense direction (Figure 3.10). This indicates that the autonomously active LTR12 copies drive downstream intergenic transcription in their endogenous context and may produce long non-coding RNAs.

3.3.11 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 (Figure 3.11a, b). Their average SuRE signal is substantially above background, while they produce almost no GRO-cap signal (Figure 3.11c, 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.

3.4 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* [53, 58]. 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 (Figure 3.12) 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 [18, 19, 21, 22, 49-53, 55-57, 85] 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 [64].

3.4.1 Accession Codes

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

3.4.2 Acknowledgements

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).

3.4.3 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.

FIGURES

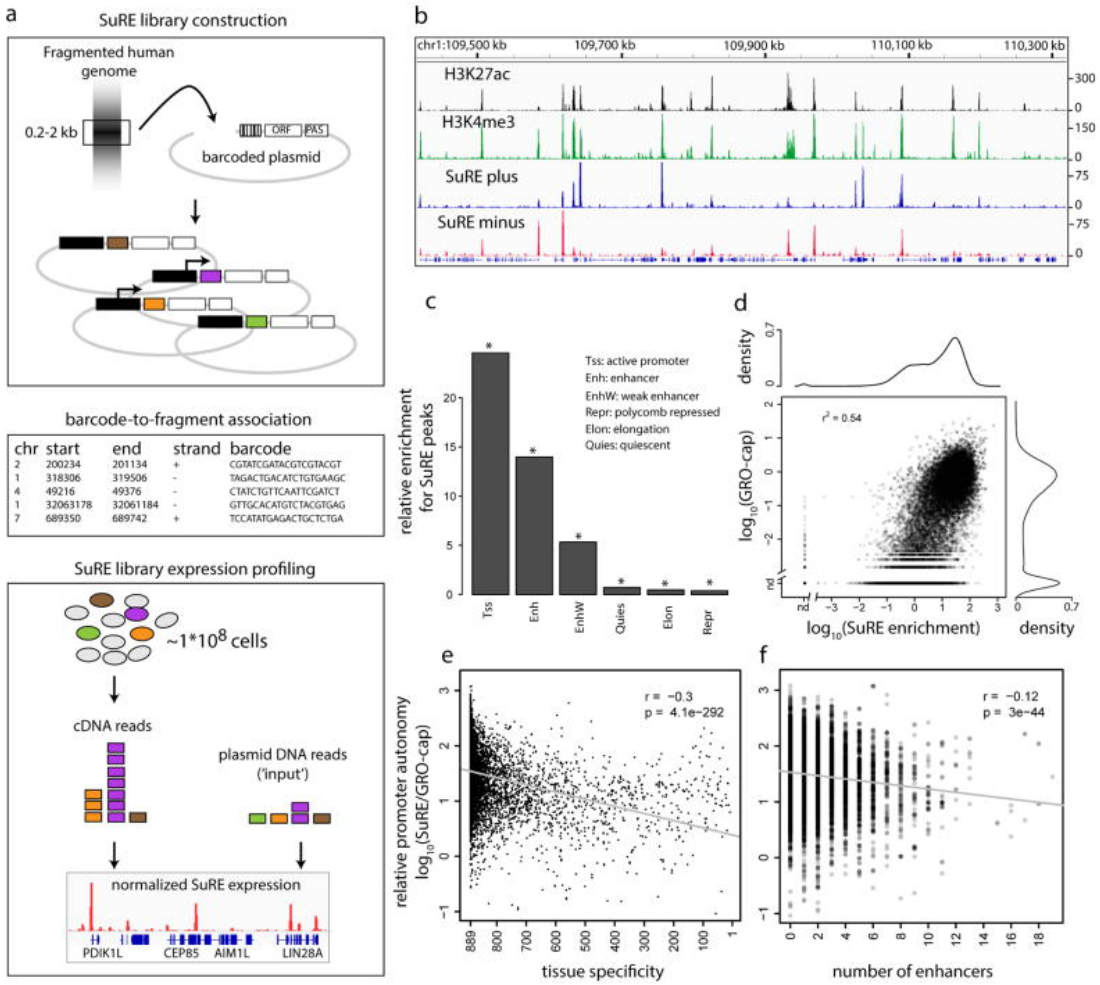

Figure 3.1 SuRE provides a genome-wide map of autonomous promoter activity.

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 H3K4me3[64] 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 [64]. **d.** Correlation between endogenous promoter activity (measured by GRO-cap [48]) 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 [1]). 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**.

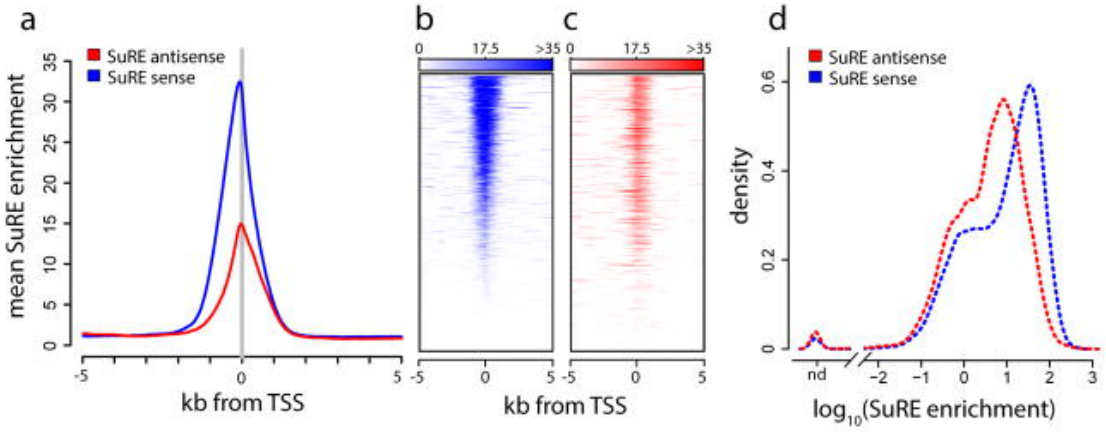

Figure 3.2 Autonomous divergent promoter activity.

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.

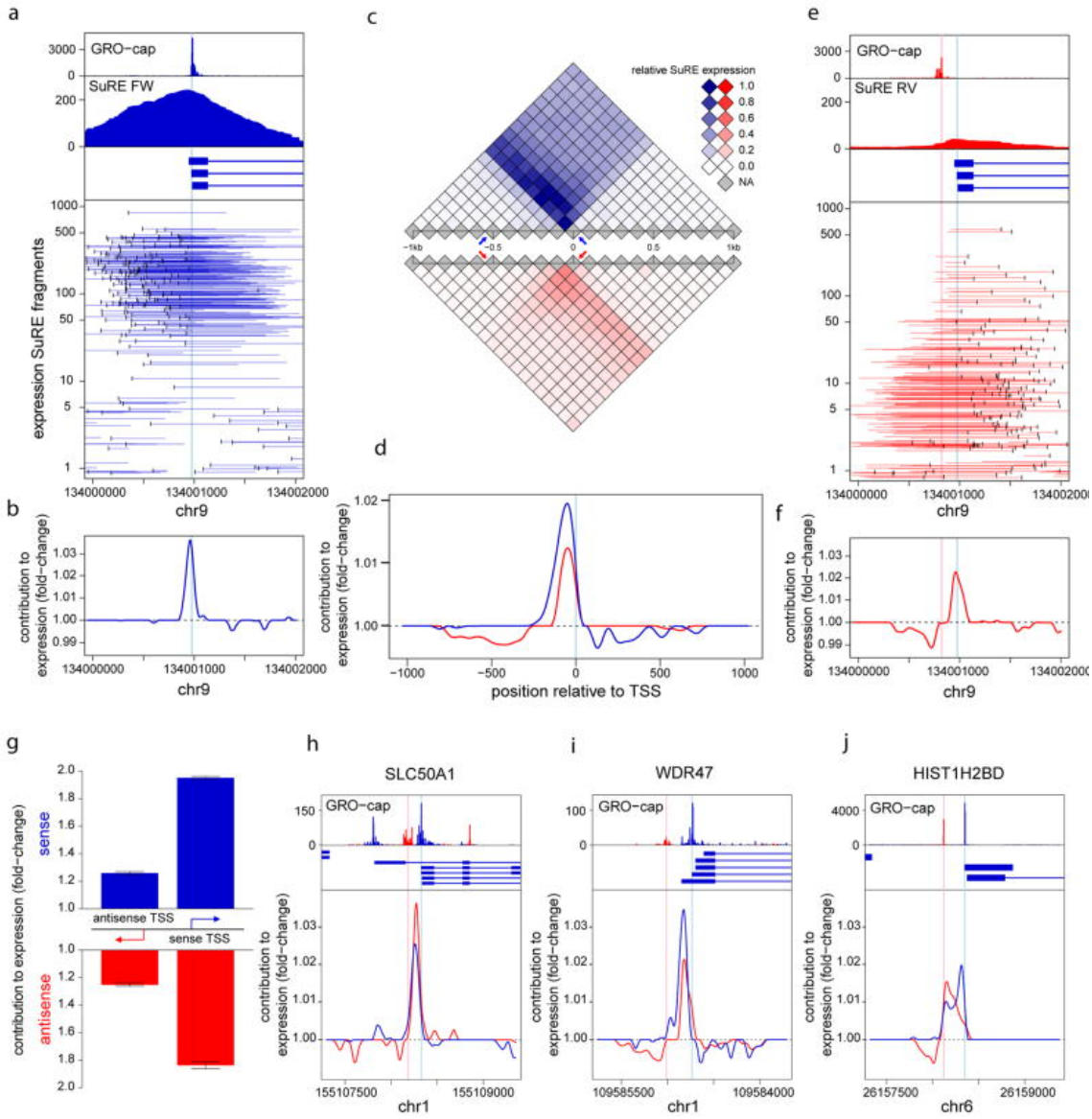

Figure 3.3 Partially overlapping query fragments allow for delineation of regions that drive promoter activity.

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 pairs [48], 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.

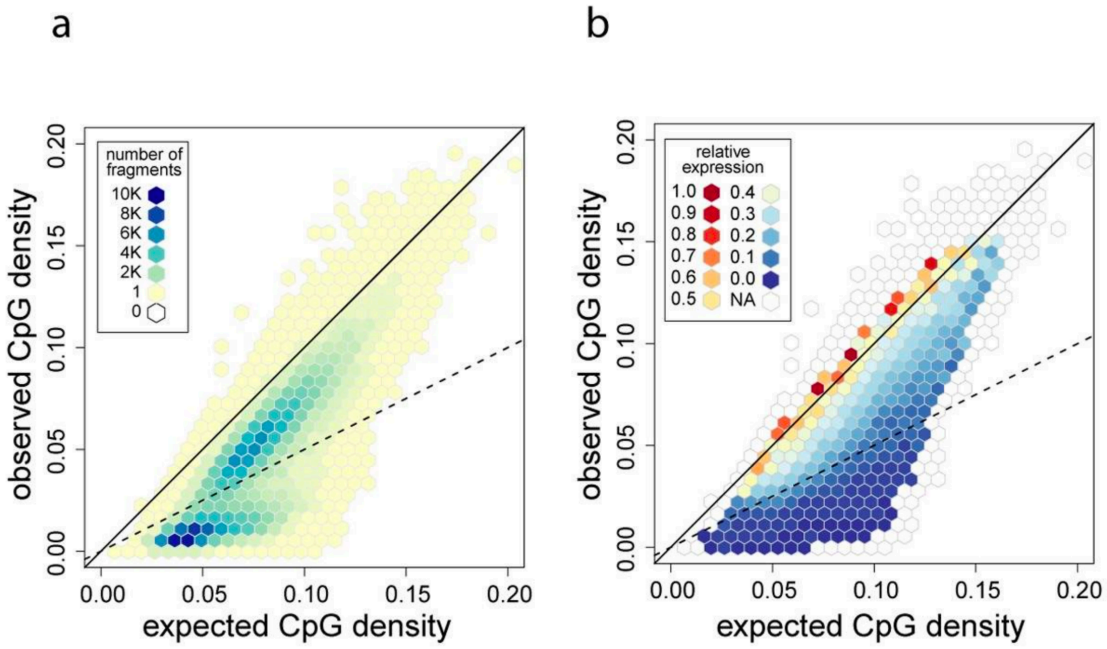

Figure 3.4 Relationship between CpG islands and gene expression.

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.**

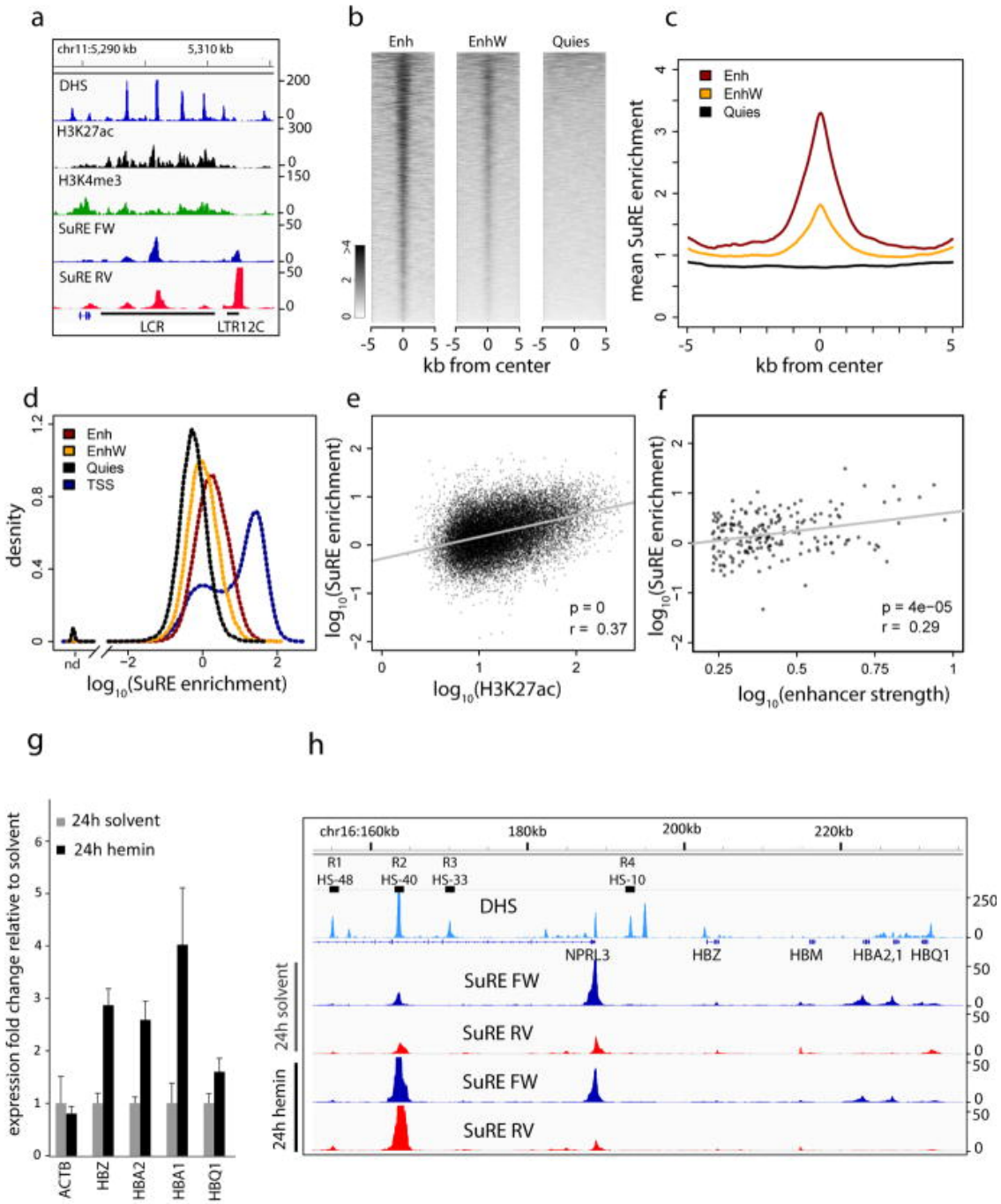

Figure 3.5 Autonomous transcription from enhancers.

a. SuRE data indicate that three of the five DNase hypersensitive sites (DHS) [64] 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') [64], 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 enhancers [61] 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 signal [64]. The bottom 4 tracks show SuRE enrichment before and after hemin induction for the plus strand (blue) and minus strand (red).

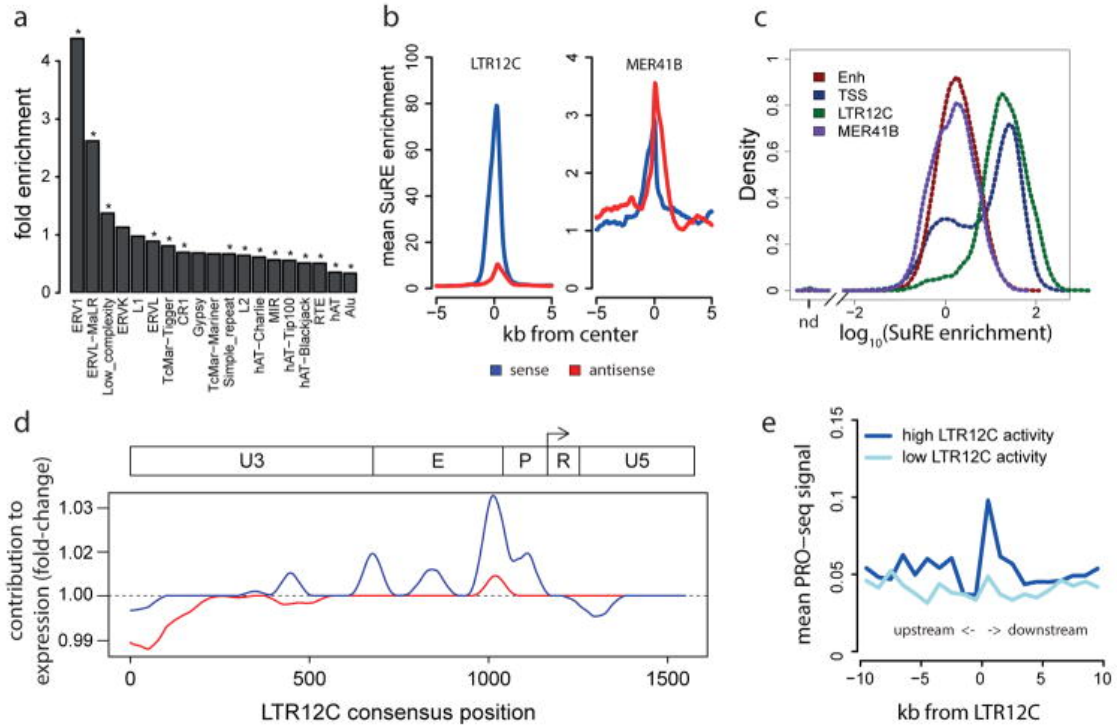

Figure 3.6 Autonomous transcription from specific repeat elements.

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 [86, 87] U3, promoter (P), enhancer (E), transcribed (R) and U5 elements. **e.** Average endogenous run-on transcription [48] 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.

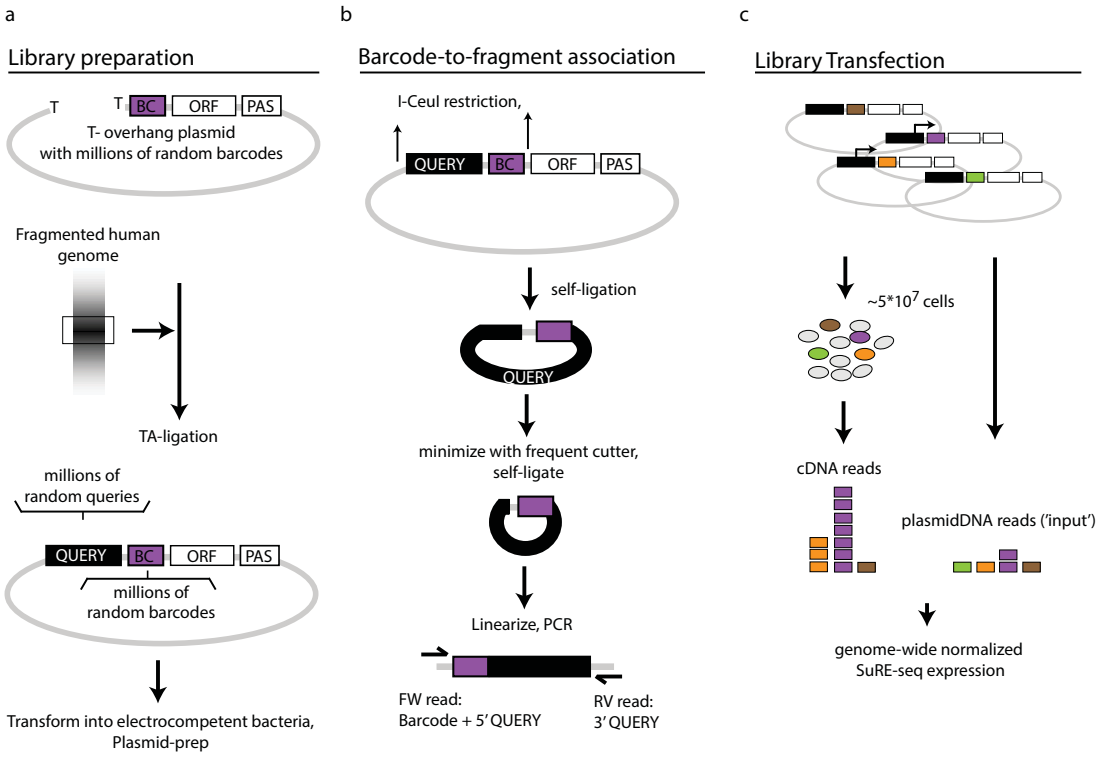

Figure 3.7 Detailed schematic representation of SuRE methodology.

See Methods for detailed description. **a.** Size-selected and A-tailed random fragments ('queries') of the human genome are inserted in bulk into barcoded T-overhang plasmids by ligation. BC, barcode; ORF, open reading frame; PAS, polyadenylation signal. **b.** The library is digested by endonuclease I-CeuI so that the barcode with the query sequence is released. This is then self-ligated and again digested with a frequent cutter restriction enzyme to reduce the insert size. After another self-ligation the circle is linearized, PCR amplified and subjected to high-throughput sequencing. **c.** Per biological replicate ~50 million cells are transfected. Those plasmids that contain promoter activity in the direction of the barcode will transcribe the barcode into RNA. Cells are harvested after 24 hours, RNA is extracted, polyA purified, reverse transcribed, PCR amplified and subjected to high-throughput sequencing. By normalization to estimated barcode frequencies in the SuRE plasmid library a genome-wide SuRE expression profile is generated.

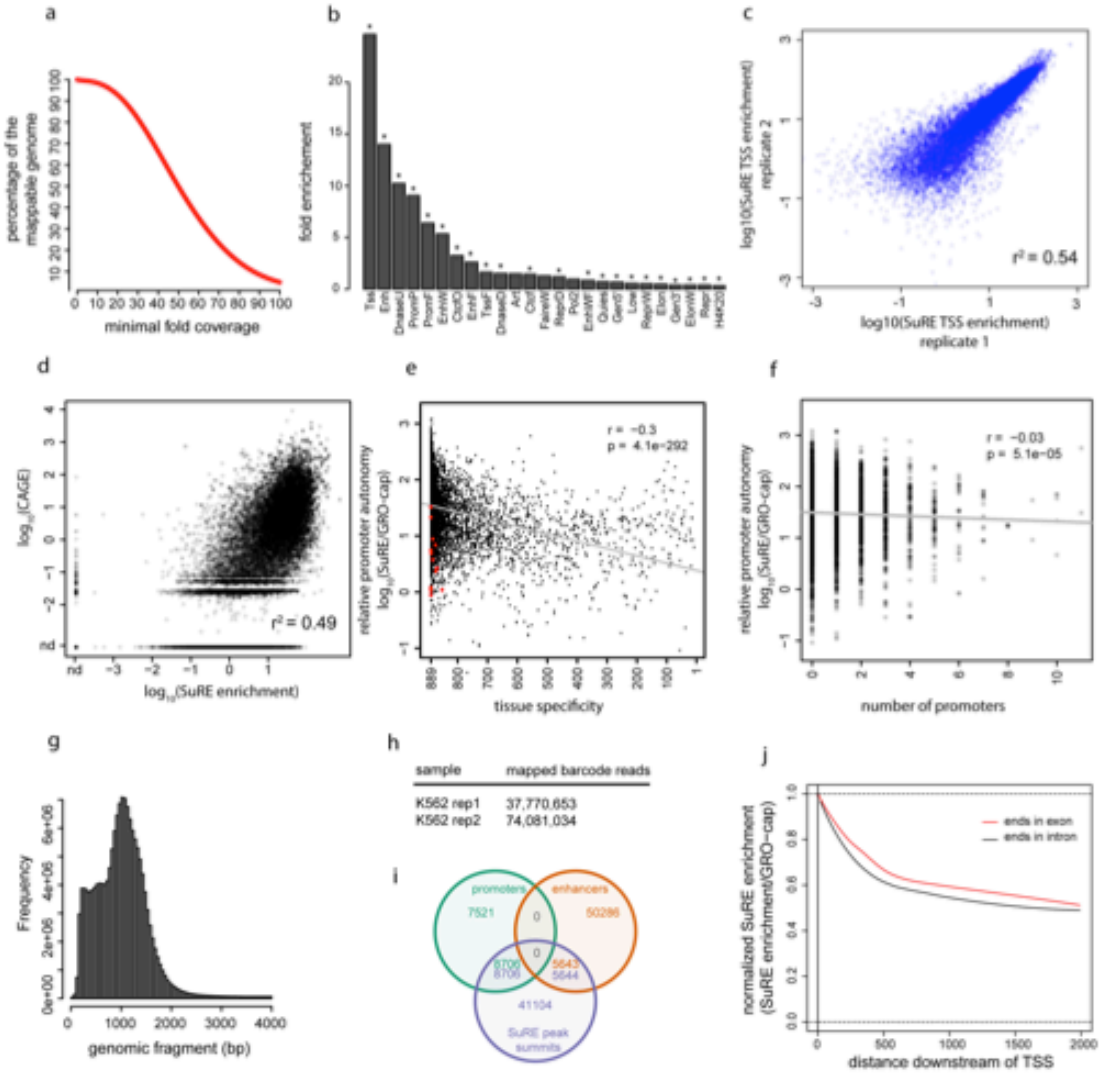

Figure 3.8 SuRE genome coverage, reproducibility and peaks.

a. Coverage of the human genome by unique elements in the SuRE library. **b**. Distribution (fold enrichment) of SuRE peaks among the 25 types of chromatin. **c**. Correlation of SuRE enrichment between biological replicates at TSSs. **d**. Correlation between CAGE1 and SuRE at the TSSs. **e**. Same as Fig. 1e but with Histone genes indicated in red. Correlation between relative promoter autonomy (log10 (SuRE/GRO-cap)) and tissue specificity (number of cell types and tissues in which each TSS is active, out of 889 tested). Grey line shows linear fit. **f**. Correlation between relative promoter autonomy and the total number of promoters (ENCODE chromatin type 'Tss') that are found in a fixed window of 5-50 kb from the TSS. **g**. Size distribution of genomic fragments in the SuRE library. **h**. Number of reads (per individual replicate) of barcodes in cDNA. Only barcodes linked to a unique genomic fragment were counted. **i**. Venn diagram representing the overlap between the summits of SuRE peaks as called by the MACS algorithm3 and ENCODE-annotated promoters ('Tss') and enhancers

('Enh' and 'EnhW' combined) . Because >1 peak summit can overlap a ENCODE annotation, overlaps are given for each direction of the comparison in the color of the annotation. **j**. Relative SuRE expression (SuRE/GRO-cap) of SuRE fragments for which the 3%apos; ends either in an intron (black) or an exon (red). Expression is normalized to GRO-cap to avoid systematic biases resulting from possible correlations between gene structure and expression level. A LOESS curve was separately fit to the logratios for all exon- and intron-terminal fragments using the distance each fragment ended downstream of the corresponding TSS, then predicted ratios were normalized to a maximum of 1. ENCODE annotation, overlaps are given for each direction of the comparison in the color of the annotation.

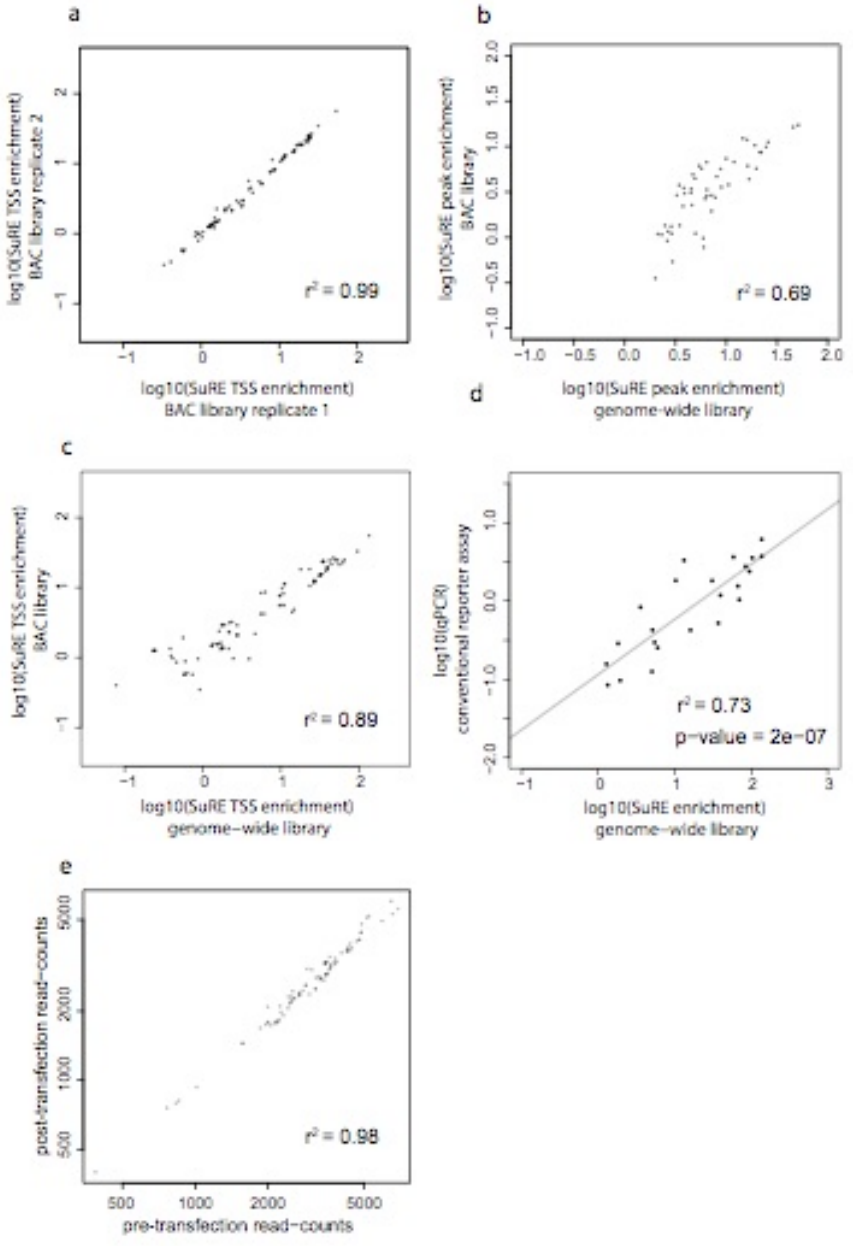

Figure 3.9 Focused BAC library.

a. Correlation between biological replicates for the focused SuRE library. Data is shown for all TSSs within in the BAC library. **b**. Correlation between SuRE enrichment obtained with the genome-wide library (x-axis) and the focused library (y-axis) for all peaks overlapping the BAC library. **c** . Same as (**b**) but for all TSSs in the BAC library. **d**. Correlation between SuRE enrichment obtained with the genome-wide library (x-axis) and a conventional reporter assay (y-axis) for 23 promoters. Grey line shows linear fit. **e**. Correlation between pre-transfection read-counts and post-transfection read-counts for all TSSs in the BAC library.

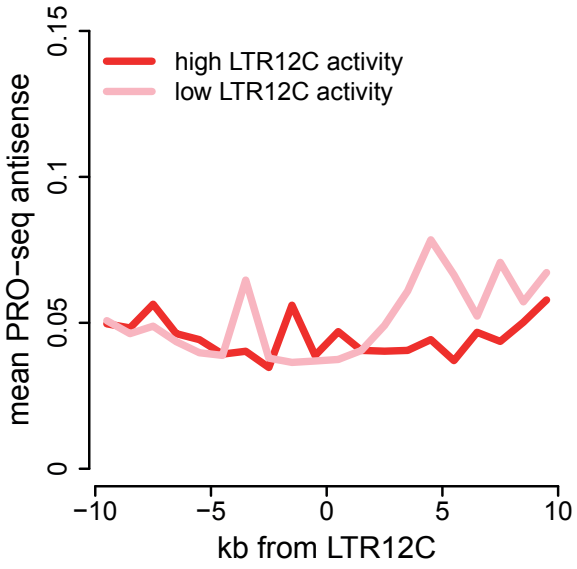

Figure 3.10 Run-on transcription around LTR12C elements, antisense.

Average PRO-seq run-on transcription activity4 around LTR12C elements as in Fig. 5e, but in antisense orientation.

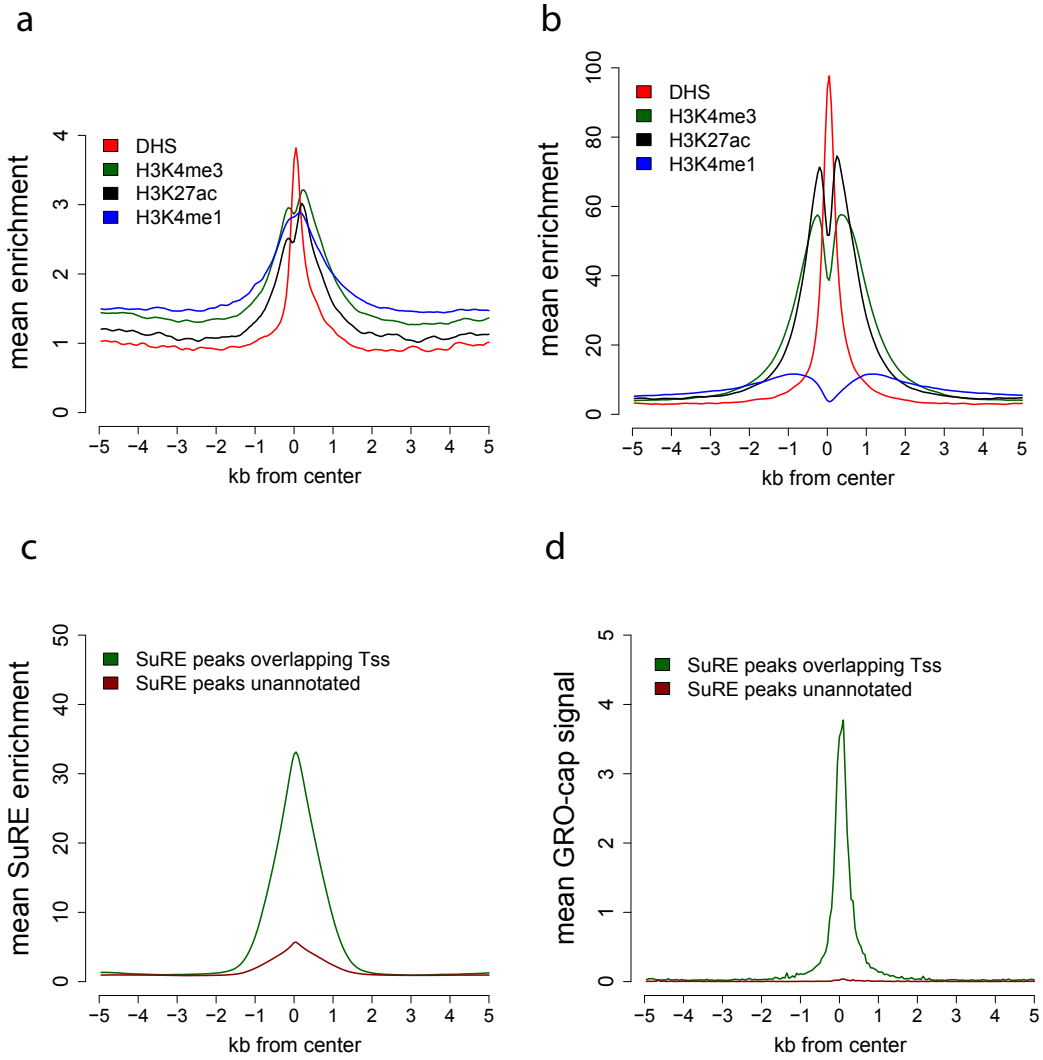

Figure 3.11 Chromatin marks associated to unannotated SuRE peaks.

a. Mean enrichment for 4 chromatin marks centered on the summit of unannotated SuRE peaks, i.e. peaks that did not overlap ENCODE annotated promoters or enhancers ('Tss' or 'Enh' chromatin state) or repetitive elements of the ERV1 or ERVL-MaLR family. **b**. Same as (**a**) but for SuRE peaks that overlapped encode annotated promoters. **c**. Mean SuRE enrichment for all peaks overlapping ENCODE annotated promoters (green) and unannotated SuRE peaks. **d**. Same as (**c**) but for mean GRO-cap signal.

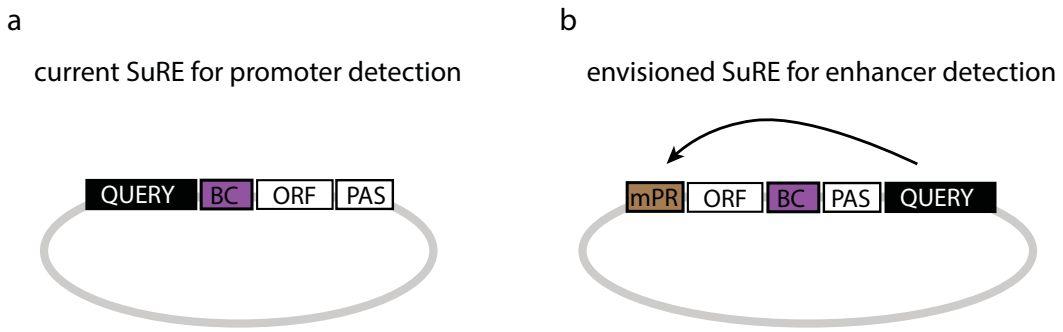

Figure 3.12 Envisioned SuRE methodology for enhancer detection.

a. Current SuRE reporter construct for promoter detection. **b.** Envisioned reporter construct for enhancer detection. Query: genomic fragment, BC: barcode, ORF: open reading frame, PAS: polyadenylation signal, mPR: minimal promoter.

4 Genome-wide Analysis with SuRE-GLM

4.1 Introduction

When a normalized activity track (cf. Figure 3.1a,b) is constructed, each SuRE element contributes to the normalized activity over its entire extent. This ensures that an active promoter region will be assigned some weight from all overlapping elements regardless of the relative position of the promoter within each element. However, this averaging procedure also assigns higher expression levels to low-activity regions proximal to active promoters, due to the fact that many elements overlapping the promoter will tend to extend into these inactive flanks. In cases of high coverage, this "piggy-backing" effect produces a broad triangular peak centered on the high-activity promoter and decreasing on each side as the proportion of elements that extend over the promoter diminishes (Figure 4.1b). In cases of lower coverage, random differences in the extent of elements overlapping up- and downstream of the active promoter can result in "false peaks" in the inactive flanks.

In all cases, the "piggy-backing" results in the assignment of higher normalized activity to inactive regions outside of the flanking active promoter. As a result, SuRE normalized expression profiles provide only a low-resolution map of promoter activity throughout the genome. This can be seen by comparing the SuRE normalized expression profile for the NUP214 promoter region (Figure 4.1b) to the expression of individual SuRE elements that contributed to this profile (Figure 4.1a). Many elements containing the region immediately upstream of the NUP214 TSS have high expression, regardless of

how far they extend up- and downstream. The smallest of these highly expressed elements are less than 200bp in length. Meanwhile, nearby elements that do not overlap the TSS-proximal region show very little expression. This suggests that a fairly small promoter region may be responsible for the expression of elements that extend over a larger area.

In contrast, the normalized activity profile is limited in the information it can provide about the boundaries of the active promoter region driving expression at a given locus. Based on this profile, we might expect an element extending from 500bp upstream to 200bp upstream of the NUP214 TSS to have fairly high expression, despite the fact that it does not extend into the region shared across all active SuRE elements. While the normalized expression profile peak does match the location of the center of this active region, the borders of the active region are ambiguous. Individual elements suggest that a minimal promoter capable of high expression could be less than 200bp long, yet the normalized profile transitions from high to low expression gradually over a much wider region. Even if boundaries for a "minimal promoter" were set to those parts of the region with a profile score greater than 50% of the peak height, the resulting promoter would be around 500bp long, much longer than necessary.

In some cases, this limitation can be overcome by looking for the smallest specific elements that show full expression, as we have with NUP214. However, this method has its own limitations. SuRE libraries differ in the number and length distribution of their elements, and within a library there can be considerable differences in coverage across different regions. For example, in the SuRE23 library (see Chapter 3) an average of 25

elements overlap a given position on each strand, considerably less than what we see in Figure 4.1a. Additionally, the mean length of elements in the SuRE23 library is close to 1kb, with only a small fraction of elements reaching lengths below 400bp. As a result, there are many instances where all the elements overlapping a small active promoter will greatly exceed the length of the active region itself.

Furthermore, the observed expression of a single SuRE element is a poor estimate of its true expression rate. In an ideal experiment, each SuRE expression count would follow a Poisson distribution. In reality, sequencing counts are typically overdispersed relative to the Poisson model [25]. Also, note that a small number of elements overlapping the active NUP214 promoter region show no expression. These likely represent elements that failed to transfect into any cells in each of the 3 SuRE biological replicates. These structural zeroes do not reflect the expression rate of the elements, as transfection must occur before expression can occur. Elements may transfect in all replicates, a subset of the replicates, or none at all. As a consequence of all these sources of experimental noise, we cannot draw accurate conclusions about the expression rate of individual elements.

Only by leveraging information from multiple overlapping elements can we identify active promoter regions at a higher resolution. To do so, we developed SuRE-GLM, a penalized generalized linear model that predicts SuRE element expression based on overlap with short, disjoint spatial bins. The model output is a coefficient track which represents the estimated contribution of each base-pair on the expression of an overlapping genomic fragment. These tracks can be used to predict the relative promoter

activity of any genomic fragment overlapping the regions covered by the SuRE library used to generate the model. They can also be used to provide insight into the structure and mechanisms of promoter activity in specific cellular contexts.

4.2 Methods

4.2.1 Datasets

In this chapter, we will focus on data from several SuRE libraries:

- SuRE23: This library was discussed in Chapter 3. Experiments using this library were performed in three cell types: K562, HT1080, and HEPG2. It was used for qPCR validation and in the multinomial model.
- SuRE34: This library was created from BAC regions covering several subsets of the genome, totaling 1.6Mbp in length. Experiments were performed using this library in K562 cells under two conditions, in the presence and absence of hemin. These results were used to validate the binomial GLM model.
- SuRE42-45: These libraries were constructed from four divergent genomes from the 1000 Genomes Project, twice independently for each genome. They were each used to perform experiments in K562 and HEPG2 cells. Together, they contain many more fragments (2.4 billion) than the SuRE23 library (150 million), at a smaller mean fragment length (~300bp). As a result, this library was used for the majority of the Poisson GLM analyses.
- SuRE49: This library was created from BAC regions covering several subsets of the genome that total 673kbp in length. Experiments were performed using this

library in K562 cells. The high genomic coverage associated with this library allowed us to use it for validation of the model built from libraries SuRE42-45.

4.2.2 SuRE-GLM Poisson model implementation

To predict the expression of SuRE elements in a single cell type, SuRE-GLM uses a log-link Poisson model with elastic net penalization. For the SuRE23 model, the covariates in this model correspond to short, non-overlapping strand-specific spatial bins of equal length that tile the length of all the regions covered by the SuRE library. Each SuRE element overlaps a set of consecutive bins. If an element overlaps an entire bin, it receives a value of 1 for that bin's covariate. If the element only partially overlaps a bin, as is usually the case for the start and end of an element, the covariate value for that bin is equal to the fraction of the bin overlapped by the element. For all bins not overlapped by an element, the corresponding covariate value is 0. This overlap covariate matrix is then fit to the SuRE element expression counts, summed over all replicates, using a Poisson GLM with elastic net penalization.

For the SuRE42-45 model, a slightly different approach was used to take advantage of higher coverage. Rather than equal-length bins, the endpoints of the covariate bins were generated by using the union of endpoints of all elements in all eight libraries. This allowed bin length to vary depending on the local density of elements. Additionally, covariate values were set to the square root of the bin length, which biased the model towards assigning larger weight to longer bins than to shorter ones.

The resulting coefficients estimate the effect of each bin on the log-expression rate of an element containing the entire bin. This means that the presence of each genomic bin in a reporter construct is predicted to have multiplicative effect on the expression of the construct, allowing SuRE-GLM to capture both activating and repressive effects without producing non-sensical predictions, such as the negative expression rates that could appear in a linear-link model. By dividing each bin coefficient by the length of the bin, or by the square root of the length in the case of SuRE42-45, we estimate the contribution of each base-pair within the bin. Each strand is modelled separately to allow for strand-specific differences in promoter activity. As a result, each model produces two strand-specific coefficient tracks. The model also produces an intercept, which accounts for the cumulative sequencing depths of the experiments, and does not factor into the track. Additional unpenalized parameters were included in the SuRE42-45 model, which will be discussed below.

4.2.3 Penalization

The elastic net GLM fits were implemented using the glmnet() function in the glmnet R package (http://CRAN.R-project.org/package=glmnet). Elastic net penalization is a natural choice for SuRE-GLM for several reasons. The L1 penalty promotes sparsity [26]. Previous research has suggested that only a small subset of the genome is transcriptionally active [2], and so we can reasonably expect that many parts of the genome will have little to no effect on promoter activity. The L2 penalty helps address collinearity [28]. Because the elements overlapping one genomic bin are very likely to overlap the adjacent genomic bin, the covariates for neighboring bins are highly collinear.

In an unpenalized fit, this can prevent a model from reaching convergence. With the L2 penalty, the coefficients for collinear bin covariates are encouraged to have similar values, or identical values in the case of perfectly collinear bin covariates, which is an intuitive behavior. Finally, both L1 and L2 help prevent overfitting, which is a particularly important feature, as overfitting is exacerbated by overdispersion and collinearity, both of which are present [28].

Prior to performing the full fit, the glmnet() function is used to validate the penalty parameters. Given an alpha penalty parameter value, this function efficiently fits a series of models, each with a different lambda penalty parameter value. To validate the lambda parameter for a given alpha, an optimal lambda value was chosen based on the model that maximized the log-likelihood of a test dataset (a random sample of 10% of the full dataset). The initial, largest lambda in the "lambda path" was selected to ensure that all coefficients in this model are penalized to zero, with the rest of the lambda path being a decreasing series evenly spaced in log-space. If the optimal lambda selected via this method was also the smallest lambda tested, a series of even smaller lambdas was tested. In all cases, the smallest lambda eventually produced an inferior model to some larger lambda, and the optimal lambda was selected. This lambda selection procedure was repeated at various values of alpha, and the optimal alpha value was based on the alpha-lambda pair that maximized the test dataset deviance ratio (Figure 4.2). Then the model was re-run using these optimized penalty values and the full dataset in order to produce a single model. In the case of the genome-wide libraries (SuRE23 and SuRE42-45),

validation was performed on a large region of the genome rather than on the entire dataset, which would be intractable.

The deviance ratio is a measure of the predictive ability of a given model. A deviance ratio equal to zero describes a model that produces predictions that are no better than those produced by the null (intercept) model, while a deviance ratio equal to one describes a perfect model that predicts mean expression values exactly equal to the observed cDNA count of each test fragment. For a given model, the deviance ratio gives the fraction of the log-likelihood difference between these two model extremes that is captured in the fit.

4.2.4 Standardization

SuRE-GLM fits differ from a default glmnet fit in some important ways. In a typical elastic net regression model, covariates are standardized prior to the fit to ensure that all covariates are at a similar scale [27], as the scale of the covariates can modify the effect of the penalization on the coefficient estimates. However, in SuRE-GLM all covariates are either in the same units (coverage of identical-length bins) or modified to reflect bin size (SuRE42-45) so standardization is unnecessary and was suppressed.

4.2.5 Strategy for performing genome-wide GLM fits

Additionally, in the case of genome-wide SuRE-GLM models, it is computationally intractable to fit the entire genome simultaneously using the R glmnet framework. In these cases, the genome was broken up into smaller subsets. In the SuRE23 model, one representative subset of the genome was used to validate the penalty

parameters and to estimate an intercept term. Then all genomic subsets were fit in parallel using the validated penalty parameters, with the model-specific intercept set to zero and the estimated intercept used as an offset in all the models to ensure a universal intercept across all subsets.

In the SuRE42-45 Poisson GLM, validation was performed using two 10Mb subsets of the genome. After the penalization parameters were validated, several additional unpenalized parameters were estimated using the average of coefficients from models fit to ten other 10Mb subsets. These parameters included:

- Library-specific intercepts: These reflect differences in sequencing depth for individual library experiments.
- iPCR coefficients: iPCR counts are defined as the number of times each element was observed in the barcode mapping procedure. They reflect relative library concentration indirectly, but our analysis suggested a library-specific and non-linear relationship. To allow the model to incorporate the information provided by iPCR, two covariates were included for each library separately: log(iPCR) and $[\log(\text{iPCR})]^2$.
- Length: Our analysis showed a log-linear relationship between mean expression and length, with longer fragments showing lower mean expression. This may be due to differences in transfection efficiency.

Once these unpenalized parameters were estimated, element-specific offsets were calculated and used in the genome-wide models.

4.2.6 Averaging over coefficient bin offsets

In models that use a fixed bin width, a single SuRE-GLM fit produces tracks with a noticeable "tiered" pattern resulting from the model structure, in which all positions within a single bin are assigned a single coefficient. The start position of the initial bin (and consequently all subsequent bins) is chosen arbitrarily, but the tiered structure can suggest the false impression that positions within a bin have some biological relationship. In theory, this could be avoided by fitting the model with a bin length of one base pair. In practice, these models are computationally intractable. To select the appropriate bin length, for the SuRE23 model, we tested a series of decreasing bin lengths. While smaller bin lengths generally produced models with better predictive power, the improvements were minor after a length of 50bp. To ameliorate the misleading "tiered" pattern, ten fits were run on the same dataset, but with the bin start positions shifting by a tenth of the bin length in each fit. The resulting tracks were then averaged, producing a "smoothed" average model track which was subsequently used for all downstream applications.

4.2.7 Modelling multiple conditions

A binomial GLM was applied to the SuRE34 hemin dataset to capture differential expression due to hemin exposure. The model used a logit link function and elastic net penalization, with penalty parameter validation performed on the entire dataset. Fixed-length bins of 50bp were used with shifted smoothing similar to the SuRE23 Poisson model. The output of the binomial model is a single coefficient profile, where positive and negative coefficients indicate higher expression in the hemin and control conditions, respectively.

An elastic-net multinomial GLM was fit to the SuRE23 experiments in three cell types: HT1080, HEPG2, and K562. Penalty parameter validation, bin size, and smoothing were similar to that for the Poisson model, except that the resulting model yields three distinct coefficient tracks, with positive and negative coefficient values indicating relative over- and under-expression in the corresponding cell type. For the multinomial triangle plots shown in Figure 4.13, multinomial probabilities were calculated by first summing the coefficients on the plus strand for each gene within 2kb of the associated TSS, and then exponentiating this sum, and dividing by the sum over all three cell lines. This is equivalent to the link function used in the GLM, except it excludes the cell type-specific intercepts which capture sequencing depths. For each cell type, a related tissue was selected from those available in the GTEX RNA-seq database (skin, liver, and blood). For each tissue, the relative expression for all genes was calculated by dividing the normalized expression in that tissue by the mean across all GTEX tissues. To determine the statistical significance of the association between multinomial probabilities for each cell type and expression levels in the corresponding GTEX tissue sample, we used a one-sided Wilcox-Mann-Whitney rank sum test.

4.2.8 SNP SuRE-GLM modeling

SNPs were evaluated for allele-specific SuRE42-45 differential expression using two methods: (i) a simple Wilcox-Mann-Whitney (WMW) rank sum test, and (ii) a GLM-based approach.

For the WMW approach, element cDNA counts were normalized by their iPCR counts to remove biases due to element-specific library concentrations. These ratios were

then normalized by the mean ratio to control for differences in sequencing depth across libraries. Finally, a WMW test was applied to the normalized counts.

The GLM-based approach used the same model as the SNP-agnostic version described above, except that SNP covariates were included in the design matrix. For elements containing a given SNP, reference allele-containing elements received a value of –0.5, while alternative allele-containing elements received a value of +0.5. This ensured that the resulting coefficient describes the change in log-expression predicted to occur due to a change from the reference to the alternative allele. By using –0.5 and +0.5 instead of 0 and 1, the model avoids estimation biases due to the structure of the penalization scheme for SNPs that are covered in an unbalanced way by reference or alternative-containing elements.

Once the coefficients were extracted from the GLM, coefficient p-values were calculated. To do this, the model was repeated with shuffled allele assignments within each SNP, yielding an empirical null distribution. SNPs were then assigned to a specific bin based on the number of elements overlapping the reference and alternative alleles. This procedure was motivated by the observation that coefficient magnitude was strongly associated with sample size. There were 225 bins, corresponding to 15 equal-sized breakpoints each for the reference and alternative sample size. Within each bin, shuffled coefficients were used to estimate the null distribution of coefficients. The distributions of the negative and positive coefficients were fit separately. The empirical distribution was used for the 95% of coefficients closest to zero, due to the density of coefficients within this region. In the remaining 5% tails of the distribution, observations were

sparser, with the largest observed coefficient often exceeding the largest coefficient in the null distribution, making it impractical to use the empirical distribution to determine a p-value for these extreme coefficients. To remedy this, we sought a distribution that could capture the general shape of the tails, and settled on the shifted stretched exponential distribution. The shifted stretched exponential distribution has the form:

$$p(x) = \frac{\lambda}{\Gamma(1 + \beta^{-1})} e^{-(|x - x_0|\lambda)^{\beta}}$$

Here, x_0 is the 95% quantile discussed above, so $x \leq x_0 < 0$ for the left tail of the coefficient distributions and $x \geq x_0 > 0$ for the right tails.

These distributions were fit using the optim() function in R. Then these distributions were used to calculate the p-values for all SNPs.

4.3 Results

4.3.1 High-resolution genome-wide promoter activity map

SuRE-GLM reveals fine-scale spatial patterns in promoter activity. At the NUP214 TSS, normalized K562 SuRE42-45 activity shows a broad triangular peak (Figure 4.1b). Based on this track alone, it is unclear whether the breadth of this peak reflects a similarly broad promoter region driving the activity of overlapping fragments, or if a smaller promoter region is producing a broader peak due to the aforementioned "piggy-backing effect".

The genome-wide SuRE-GLM track (Figure 4.1c) suggests that the latter is true. According to the model, most positions within the peak have only a small effect on the

expression of overlapping genomic fragments. The exception is a small region immediately upstream of the TSS. This region corresponds to the region shared by the active individual SuRE constructs observed in Figure 4.1a. The information provided by the activity of many active and inactive fragments can be reduced to a single track using SuRE-GLM.

4.3.2 Spatial patterns in promoter activity

The preinitiation complex (PIC), which is responsible for Pol II transcription initiation, typically binds to core promoter sequences ±50bp from the TSS [48], while other transcription factors bind further upstream in the proximal promoter. Recent genome-wide TSS-mapping assays have shown widespread divergent transcription throughout the genome, with antisense transcripts typically initiating 90-120bp upstream of their sense-strand pairs [48], which suggests that transcription initiation occurs at separate, directional core promoters. This has led to some discussion about whether promoter activity is being driven primarily by these individual core promoters, or if a central proximal promoter is responsible for divergent activity at the two nearby core promoters [16, 48, 68, 69].

SuRE-GLM allows us to examine the spatial distribution of promoter activity more closely. On the sense strand, cross-correlation between GRO-cap and SuRE-GLM peaks at 50bp (Figure 4.3), suggesting that the proximal drivers of promoter activity tend to slightly precede endogenous transcription initiation sites upstream by this distance. This places the primary proximal drivers of promoter activity partially outside the canonical core promoter region. Meanwhile, the same proximal promoter regions appear

to be responsible for both sense and antisense transcription among the most highly expressed genes (Figure 4.4) and in the mean profile for all TSSs (Figure 4.5). The median sense profile lies slightly upstream of the annotated TSS position, with most sense profiles peaking within 100bp upstream and 50bp downstream. A subset of profiles peaks further up- or downstream of the annotated TSS, and may reflect alternative TSS promoter activity. Antisense profiles are generally highly correlated with sense profiles (Figure 4.6), with overlapping peaks of lower intensity (Figure 4.4). This suggests that the sequences responsible for driving divergent transcription are shared between sense and antisense TSS pairs, and that these regions tend to be just upstream of the core promoter. An example can be seen at the WDR55 TSS (Figure 4.7).

Sense and antisense profiles show a bias towards more negative coefficients moving away from the peak in the downstream and upstream directions, respectively (Figure 4.4). In both cases this represents reduction in expression in the direction of transcription for transcripts initiating near the peak. This may reflect the effects of downstream sequence features that decrease the transcription rate (pausing sites, e.g.), cause early termination before the reporter barcode is transcribed, or promote degradation of transcripts before reverse transcription can occur [48].

4.3.3 Accurate prediction of reporter construct relative expression

In addition to producing per-bp coefficient tracks, SuRE-GLM models allow us to predict the relative expression of novel reporter constructs with arbitrary start- and end-points. To make a prediction, we simply sum up all coefficients between the start and end of a hypothetical construct, then exponentiate the sum to get the predicted relative

expression rate. To test the quality of these predictions, we tested the expression level of 23 reporter constructs in K562 cells using RT-qPCR. We compared the results to predictions based on the mean normalized SuRE23 activity within the endpoints of the constructs (Figure 4.8a) as well as the SuRE-GLM predictions (Figure 4.8b). These constructs all contain an annotated TSS, and have a median length of 1050. SuRE-GLM improved the R^2 for the predictions from 0.73 to 0.78.

4.3.4 Validation on BAC libraries

To further test the predictive power of our SuRE-GLM models, we explored the ability of SuRE-GLM to predict the activity of thousands of smaller elements in the K562 SuRE49 experiment. These elements have a median length of ~300bp, and lie within 2kb of an annotated TSS within the regions covered by at least 100 elements in the SuRE49 BAC library. These elements allow us to test SuRE-GLM predictions on smaller constructs than were used in the qPCR-based validation experiments.

In Figure 4.9, genome-wide SuRE-GLM predictions based on our SuRE42-45 model accurately predict the relative expression of SuRE49 elements. This produces an overall R^2 of 0.65. This correlation is based on multiple different loci, suggesting that SuRE-GLM can be used to make comparisons of constructs both within and across different genomic loci.

4.3.5 Minimal promoter design with SuRE-GLM

Given that SuRE-GLM can accurately predict the expression of constructs of various length within a TSS region, SuRE-GLM predictions can aid in the identification

and design of minimal promoter regions for specific genes. In Figure 4.10a, we visualize the relative expression of all hypothetical elements within 2kb of the DGCR14 TSS. The endpoints of each element can be found by extending two lines parallel to the sides of the larger triangle, from the corresponding point to the base of the triangle. Grid patterns form in the predictions due to oscillations in the coefficient profile between positive and negative values. While many constructs show some predicted expression, there are three hypothetical constructs that maximize relative expression within this region. These red regions share an endpoint ~100bp downstream of the TSS, with the smallest option extending ~400bp upstream of the TSS. By leveraging SuRE predictions over a range of lengths and positions, researchers can use SuRE-GLM to isolate minimal promoters without having to perform multiple promoter-bashing experiments.

4.3.6 Binomial and multinomial models predict differential expression

In addition to predicting expression in a single cell type with a penalized Poisson model, SuRE-GLM can be extended to predict differential expression across multiple cell types or conditions using a penalized binomial and multinomial model. This approach requires data from multiple SuRE experiments using the same SuRE library in different cell cultures or conditions. Based on the spatial bins that a SuRE element overlaps, the SuRE-GLM model fits a model that predicts that element's distribution of counts over each experiment. The result is a high-resolution track of the regions responsible for differential expression across the cell types or conditions.

To test this differential SuRE-GLM approach, a SuRE library was constructed using a BAC library that included the R2 α-Locus Control Region (LCR), a region known

to be upregulated in the presence of hemin [88]. This library was transfected into two K562 cultures, one exposed to hemin and another exposed only to solvent. On both strands, the normalized SuRE activity track in the R2 α-LCR is substantially higher in the hemin condition than in the control (Figure 4.11b). While normalized expression peaks within the LCR, the difference in normalized activity extends beyond the bounds of this region. We then applied differential SuRE-GLM to these two conditions. Note that a binomial model was used, as this is the special case of the multinomial appropriate for two conditions. This model, wherein counts in the hemin condition were considered a "success" (see Section 2.2.2), reveals a narrower coefficient peak within the R2 α-LCR region on both strands (Figure 4.11a). This reflects the ability of multinomial SuRE-GLM to identify specific, biologically relevant regions responsible for differences in expression that only appear at lower resolution in normalized SuRE activity profiles.

We also applied the multinomial SuRE-GLM model to data from SuRE experiments using the genome-wide SuRE23 library discussed above. These experiments include the K562 results discussed previously, as well as results in HEPG2 and HT1080 cells. In Figure 4.12, we compare results from Poisson SuRE-GLM models fit separately to each cell type to a multinomial SuRE-GLM model fit to all three together. In both cases, a broad region immediately upstream of the annotated CA1 TSS is identified as contributing to expression in K562 cells. As this region is inactive in HT1080 and HEPG2 cells, the Poisson coefficient profile remains flat for these cell types. However, the multinomial coefficient profile for these two cell types is negative in the same region where it is positive in K562 cells. This reflects the fact that multinomial model captures

relative differences in expression across cell types. The result is that the coefficients of one cell type can be affected by the relative expression of another cell type, even if their independent Poisson coefficient profiles are flat.

Given that the different cell lines originate from distinct tissue types, we might expect that genes predicted to be over-expressed based on our SuRE-GLM model would also show increased expression in related tissues. To test this hypothesis, we compared the normalized SuRE-GLM predicted cell type-specific proportions to the relative expression of genes in related tissues according to GTEX [89]. The SuRE-GLM predictions indicate that genes overexpressed in skin cells are biased towards higher expression in the HT1080 cell line relative to the other two cell lines (Figure 4.13a, p-value $< 2.2 \times 10^{-16}$). HT1080 is a fibroblast cell line [90]. We see a similar relationship between genes overexpressed in the liver and the relative predicted expression in HEPG2 (Figure 4.13b, p-value $< 2.2 \times 10^{-16}$), which is derived from a hepatocellular carcinoma [91]. We did not observe a significant association between the relative expression of erythrocyte cell line K562 [92] predicted expression levels and overexpression in whole blood (Figure 4.13c), which may be the result of cell type heterogeneity in blood or divergent patterns of gene expression in K562 cells. Nevertheless, these patterns show that multinomial SuRE-GLM can predict biologically meaningful differences in expression patterns across different cell lines.

4.3.7 Identifying regulatory SNP variants with SuRE

A typical genome contains hundreds of thousands of SNPs that overlap regulatory regions [93]. In some cases, SNP variants in regulatory regions can affect the binding of transcription factors or other regulatory factors, which can lead to downstream changes in gene expression [94]. Methods that have been used to identify SNP variants that drive differential expression include genome-wide association studies (GWAS) [95, 96] and expression quantitative trait loci (eQTL) mapping [97, 98]. Unfortunately, these approaches are limited in their resolution due to linkage disequilibrium (LD), which makes it difficult to separate the effects of adjacent SNPs on expression. Identifying a causal SNP among a block of SNPs in LD remains a challenging problem in the search for non-coding regulatory variants.

One approach used to identify causal regulatory SNPs among a block of SNPs in LD is to use reporter assays [99, 100] in a similar fashion to promoter-bashing experiments, except that the variation in sequence across constructs is limited to the allele of a single SNP at a time. Given that many candidate SNPs can share an LD block, this method requires the creation of many constructs to find a single causal SNP. Creating synthetic constructs for all SNPs across many LD blocks is intractable.

As we have previously demonstrated, the SuRE experimental protocol makes it possible to deconvolve the effects of adjacent genomic regions with a fairly high resolution. Similarly, by comparing the results of SuRE experiments across different genomes, we can observe genome-specific differences in expression that may be attributed to sequence variants. Over one billion of the SuRE42-45 library elements

contain at least one SNP for which we observed both allele variants. In total, we observed both alleles for 5,919,293 SNPs, accounting for 57% of the known common SNPs world-wide (minor allele frequency >5%).

When we look at normalized SuRE activity alone, the libraries used in SuRE42-45 reveal pronounced differences in expression at certain loci, which correspond to differences in genotype (Figure 4.14). Most SuRE elements in these experiments contain less than 500bp of genomic DNA, and are therefore unlikely to contain more than a few SNPs that vary across the four genomes. This largely eliminates the issue faced in GWAS and eQTL analyses, as the effects of a SNP can be decoupled from the effects of most of the SNPs in the same LD block. As a consequence, differential SuRE expression can frequently be attributed to a single SNP contained within the differentially active region.

A simple comparison of normalized SuRE expression across different genomes can reveal some patterns in differential expression. In Figure 4.14, two individuals homozygous for the C allele show high expression on both strands, while the individual homozygous for the A allele shows almost no expression. The heterozygous individual shows intermediate expression. However, as described earlier in this chapter, normalized SuRE activity tracks capture only a subset of the information available in the full SuRE experimental data.

The allelic identity of SuRE42-45 SNPs was identified during the mapping procedure for the majority of fragments covering any given SNP. This allows us to separately measure the activity of fragments overlapping a particular allele of a given SNP (Figure 4.15), even in the cases where the corresponding genome is heterozygous. In

regions with some active expression, we can estimate the relative contribution of each allele to expression by integrating information across all elements from SuRE42-45 for which the relevant SNP position has been sequenced.

The simplest way to assess whether a SNP influences expression in SuRE42-45 is to directly compare the distribution of all elements containing the reference allele with the distribution of all elements containing the alternative allele. For each SNP, we calculated a p-value with a Wilcox-Mann-Whitney rank sum test. We repeated this process for all SNPs using shuffled allele assignments to generate a null distribution for p-values. At an FDR < 10%, we identified 22,986 SNPs that are differentially expressed across different allelic variants.

In addition to the WMW method, we wanted to test whether the GLM method could aid in the identification of differential SNPs. Elements that overlap a given SNP differ in terms of start- and endpoints, and therefore some elements may contain active regions that are missing in others. If active regions are asymmetrically distributed across elements containing one allele when compared to the other allele, this may produce a misleading signal when using the WMW test. By including both spatial bins and SNP covariates in the a single SuRE-GLM model, we can account for spatial differences across the elements corresponding to each allele. Additionally, by running a model that estimates the effects for all SNPs simultaneously, we may be able to disentangle the effects of adjacent SNPs that appear in some, but not all, of the same elements.

To implement the SNP-SuRE-GLM model, we used the same model as for the SNP-agnostic SuRE42-45 model, except that we included additional terms to capture the

effects of SNP allele variants. Elastic-net penalized models do not report p-values by default. To nevertheless assess the statistical significance of the model coefficient for each SNP, we constructed a coefficient null distribution by repeating the GLM model fit with shuffled allele assignments. We then constructed a series of null models based on these coefficients, as well as the sample size of the reference and alternative alleles (see Methods, Figure 4.16). The result was a p-value distribution that is much flatter than the WMW p-value distribution (Figure 4.17), suggesting that the GLM-based method better reflects the more likely scenario in which only a small fraction of SNPs is predicted to have a significant effect on local transcription activity. Indeed, the number of SNPs found to be significant at an FDR cutoff of 10% is 1,203 in K562 cells, a much smaller number of significant SNPs than was found using the WMW method.

To evaluate the ability of both methods to detect functional SNPs, we compared our results from both methods to the SNP2TFBS database, which lists TF motifs predicted to be disrupted by SNP variants. Based on the assumption that functional SNPs would be enriched in this list compared to non-functional SNPs, we checked to see whether SNPs with smaller p-values according to each method were enriched in the SNP2TFBS database (Figure 4.18). In K562 cells, both methods showed enrichment for SNP2TFBS SNPs at low p-values. For the one thousand smallest-ranked p-values, the WMW showed higher enrichment. However, beyond this initial group the GLM-based method shows higher enrichment. This suggests that the GLM-based method can capture the true differential activity of more SNPs by removing the false positives present in the WMW-based ranking.

FIGURES

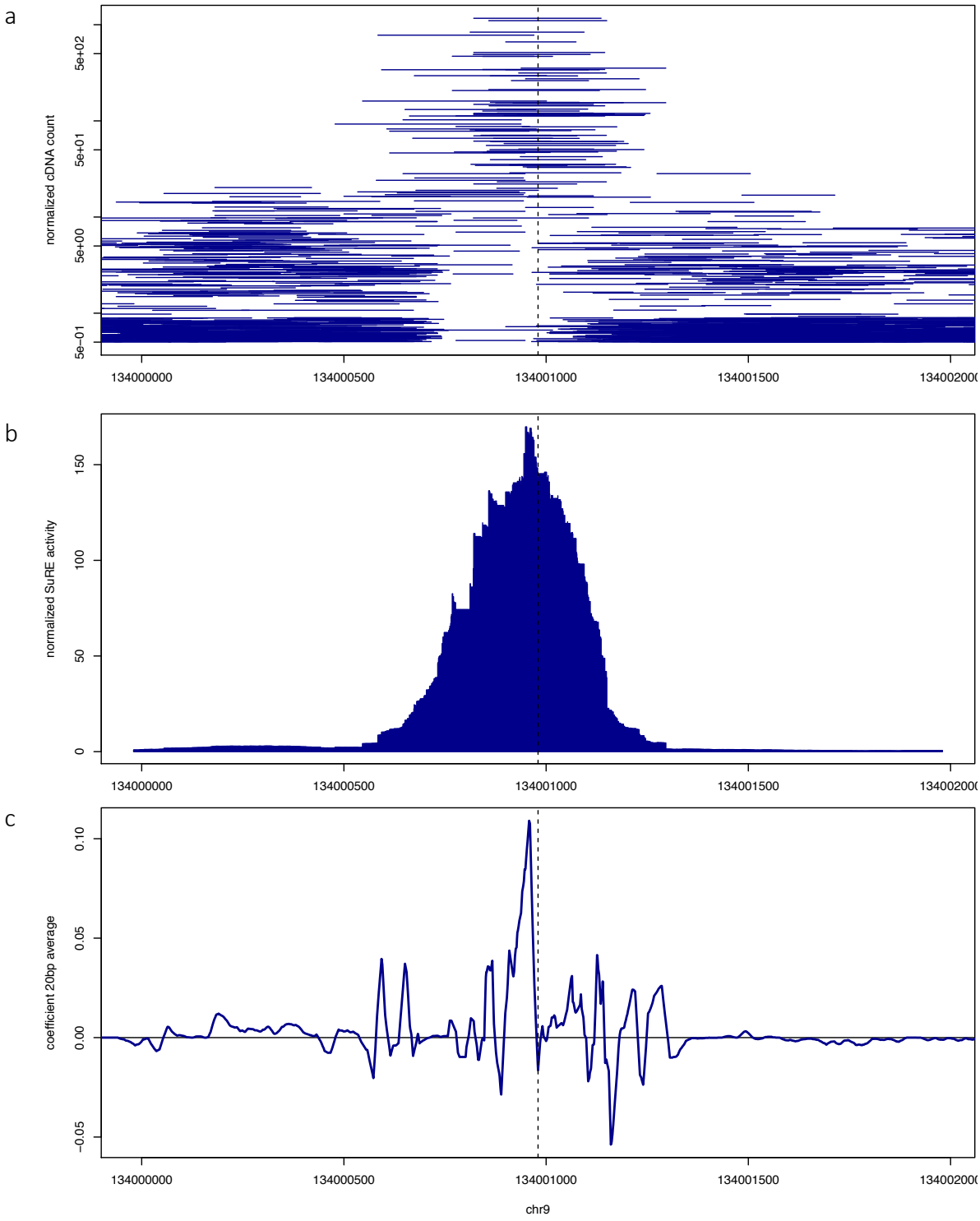
a
normalized cDNA count
5e+02
5e+01
5e+00
5e−01
134000000
134000500
134001000
134001500
b
normalized SuRE activity
150
100
50
0
c
coefficient 20bp average
0.10
0.05
0.00
−0.05
chr9

Figure 4.1 Comparison of cDNA counts, normalized activity, and coefficient profile

To visualize K562 cDNA counts for all SuRE42-45 elements near the NUP214 TSS, each element appears as a line extending between its upstream and downstream cut sites (**a**). A small random pseudocount has been added to each count to allow for visualization of overlapping and zero-count elements. Element counts have also been normalized to reflect library-specific scaling factors. These same offsets are used to generate a K562 normalized activity track (**b**). GLM estimates of expression fold-change per base pair were smoothed by a running average of 20bp (**c**) for a clearer visualization of the coefficient track.

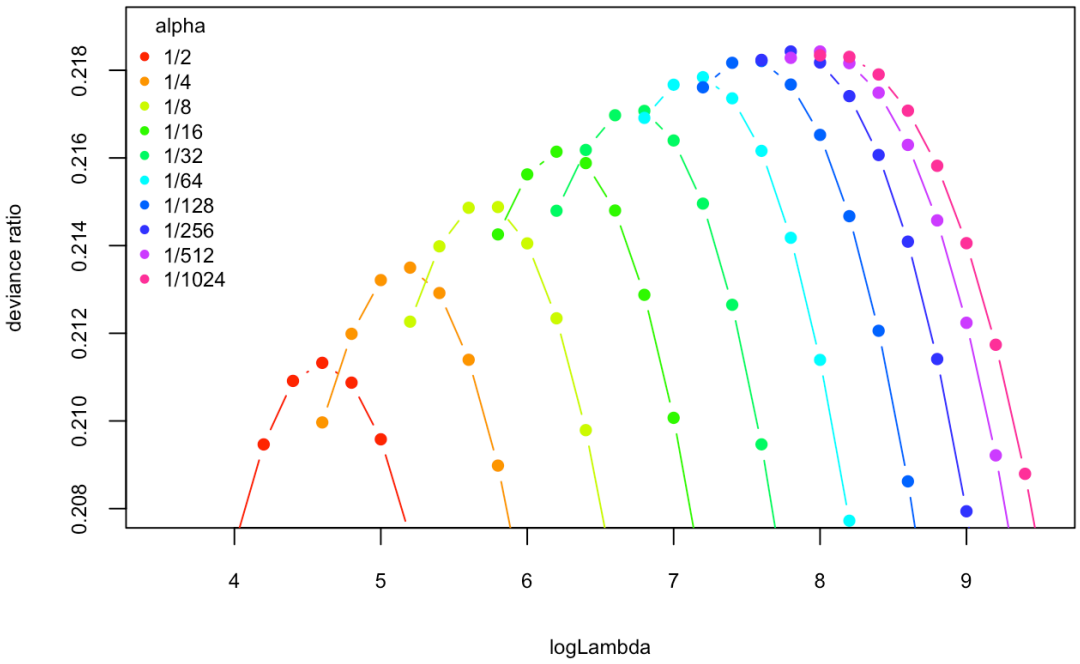

Figure 4.2 Cross-validated deviance ratios for SuRE-GLM fits

Each point represents a unique pair of alpha and lambda values tested in a 10Mb subset of chromosome 22 from the K562 experiments performed with the SuRE42-45 libraries. Each hyperparameter pair was used to model 90% of this subset, and predictions for the remaining 10% test set were used to assess the performance of each model. A higher deviance ratio indicates a more predictive model.

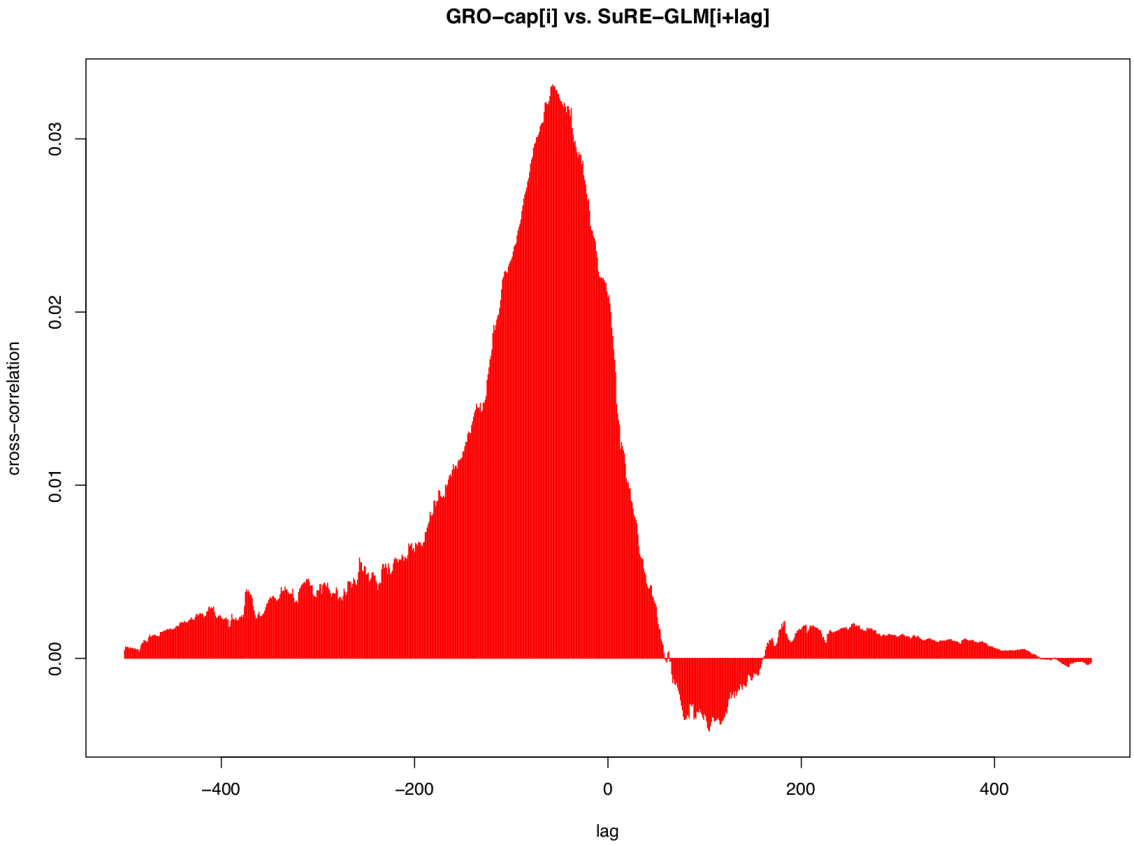

Figure 4.3 Cross-correlation of SuRE-GLM and GRO-cap in TSS regions

Cross-correlation of K562 SuRE42-45 GLM coefficient profile and K562 GRO-cap in the 2kb regions surrounding annotated TSSs. The cross-correlation peaks around 50bp, suggesting that the proximal drivers of promoter activity tend to precede endogenous transcription initiation sites upstream by this distance. This places the primary proximal drivers of promoter activity just outside the canonical core promoter region, which surrounds the transcription initiation site.

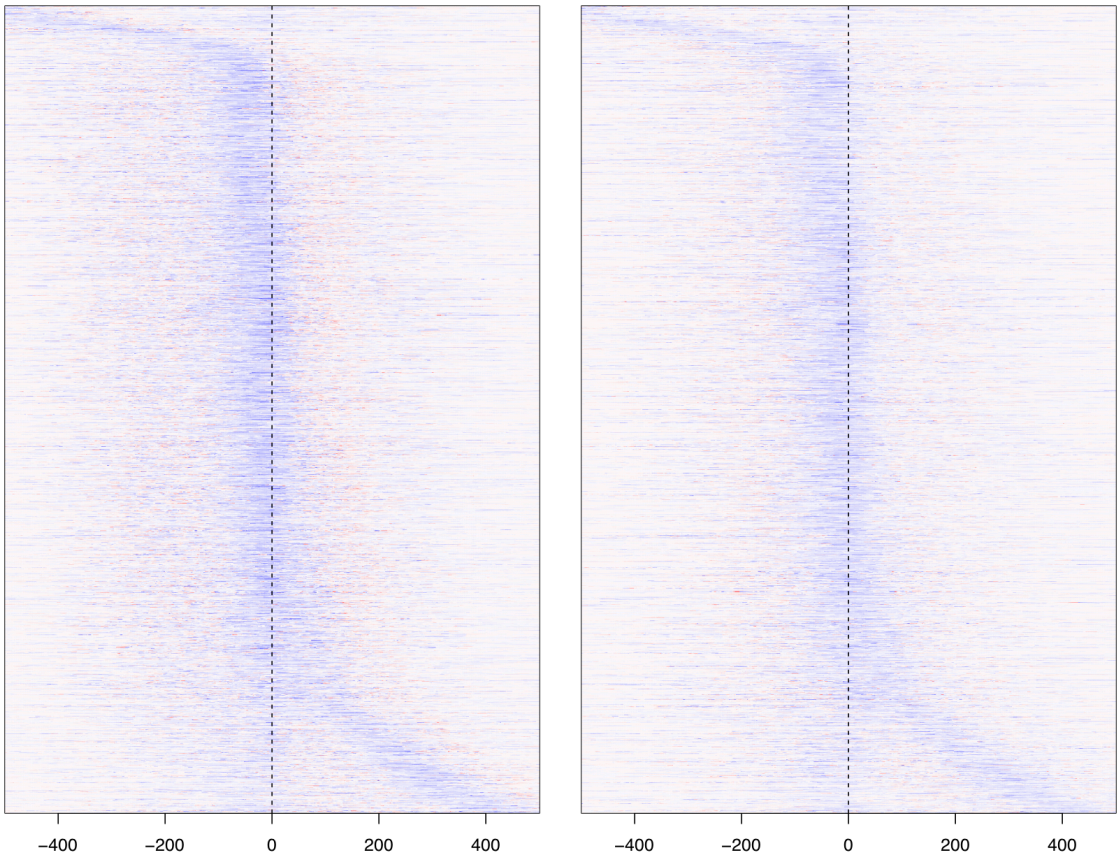

Figure 4.4 Sense and antisense strand profiles for 2000 most active TSS loci

Line plots of K562 SuRE-GLM coefficient profiles for the 2000 most active TSS loci for the sense and antisense strands, based on total K562 GRO-cap activity within a 1Kb window surrounding each TSS. Profiles are sorted vertically by the relative position of the sense GRO-cap profile peak. Sense and antisense profiles for each TSS are shown on the same line and are shown relative to the sense strand. Red and blue points reflect negative and positive GLM coefficients, respectively.

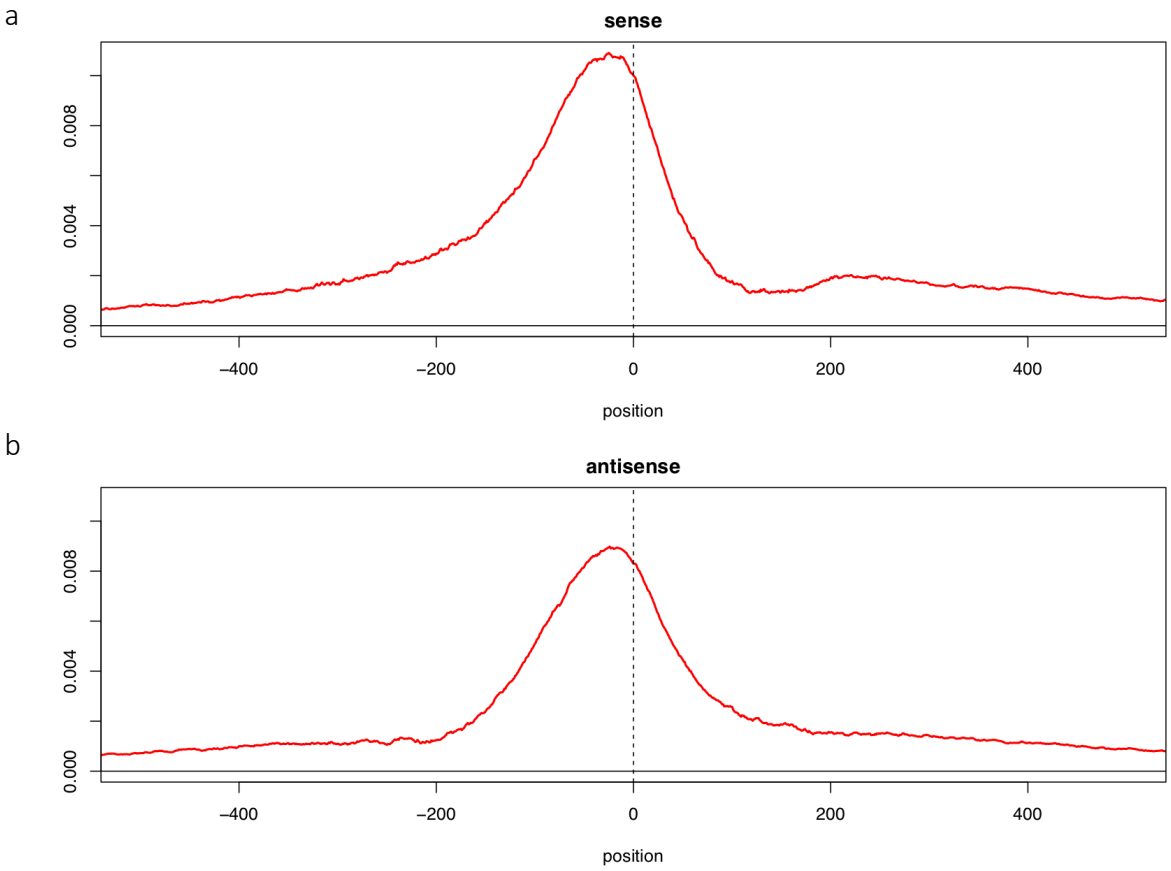

Figure 4.5 Mean GLM coefficients profile surrounding annotated TSSs

Mean profiles are based on the SuRE42-45 K562 GLM model. The sense and antisense plots are shown in relation to the sense strand. Both the sense and antisense profiles peak within 50bp upstream of the annotated TSS, while antisense profiles tend to peak within 50bp downstream of the annotated TSS, suggesting that unlike GRO-cap peaks, which show separate peaks for transcription initiation on the two strands, sense and antisense promoter activity profiles tend to overlap.

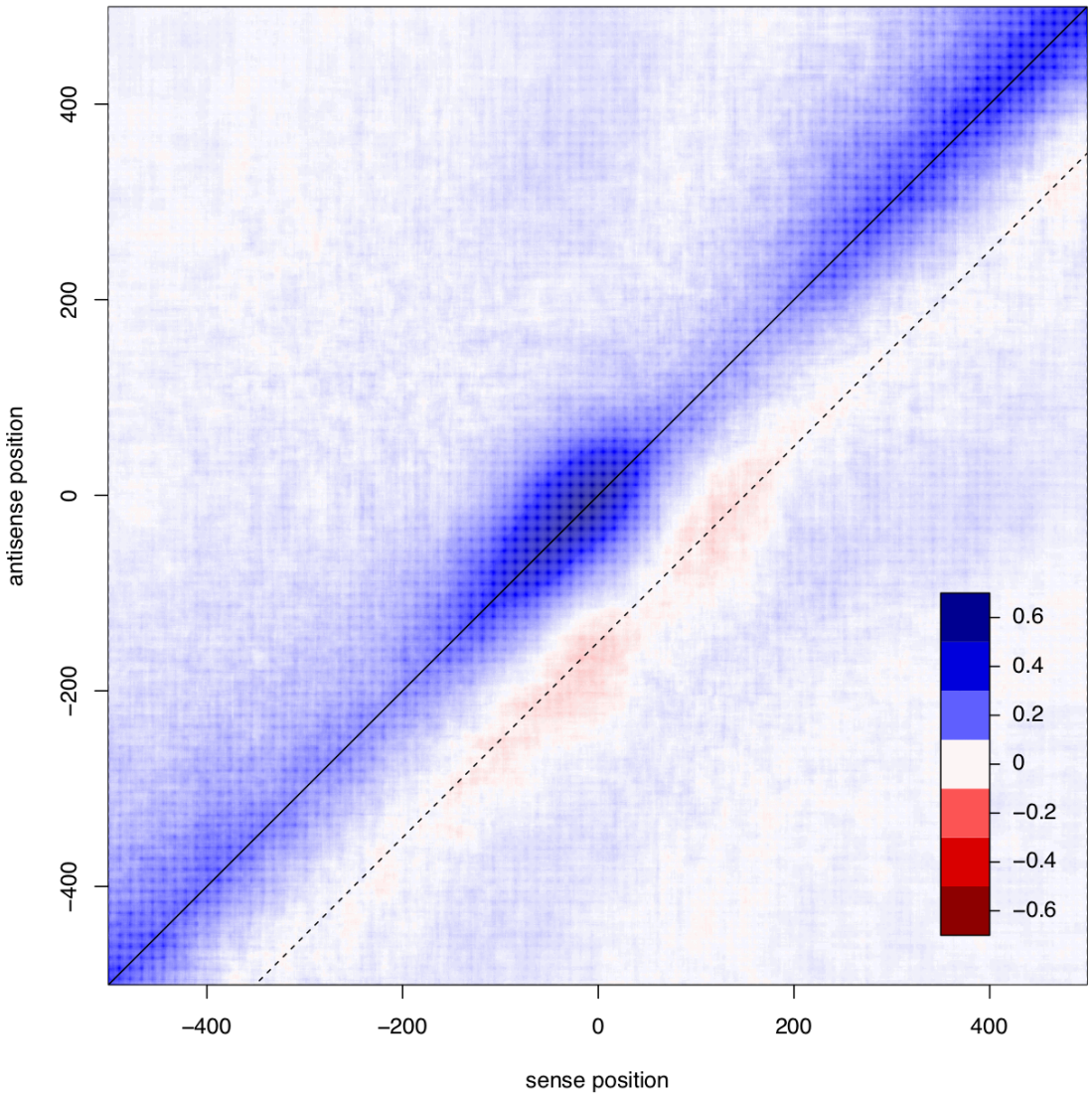

Figure 4.6 Correlation structure between sense and antisense SuRE profiles

Correlation structure between sense and antisense genome-wide SuRE42-45 K562 GLM profiles in the region surrounding annotated TSS. Positions are relative to the sense strand. Sense and antisense profiles are the most highly correlated at matching positions (solid line), and are somewhat anticorrelated about 150bp downstream (dotted line).

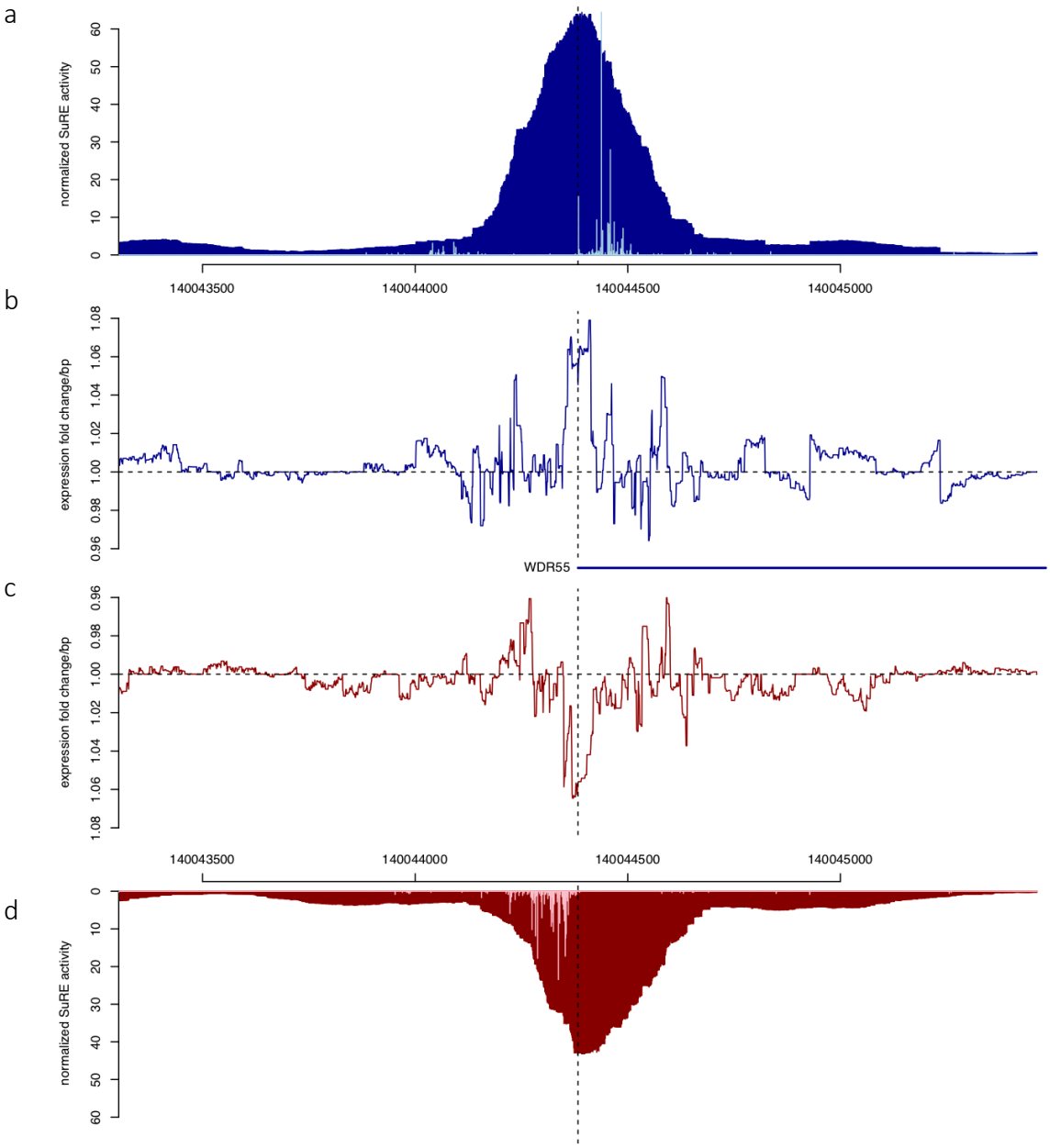

Figure 4.7 Normalized SuRE, GLM, and GRO-cap tracks in the WDR55 TSS region

Normalized SuRE, GLM, and GRO-cap tracks in the region surrounding the WDR55 TSS in K562 cells. Plots are shown for both the sense (blue) and antisense (red) strand. Both the normalized SuRE tracks (**a**, dark blue; **d**, dark red) and the GLM coefficient tracks (**b**, **c**) are based on K562 experiments using the SuRE42-45 libraries. The GRO-cap tracks (**a**, light blue; **d**, pink) capture the position of transcription initiation sites in this region. At the TSS, the tracks reflect a common pattern, with overlapping sense and antisense SuRE-GLM peaks separating downstream sense and upstream antisense GRO-cap peaks.

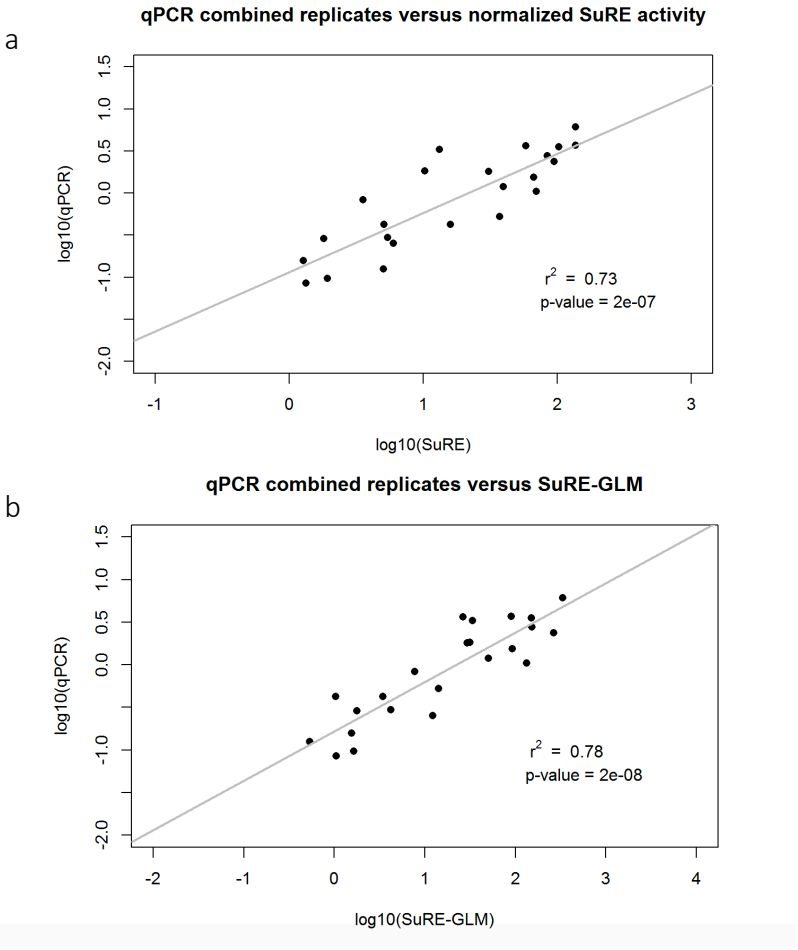

Figure 4.8 qPCR-based validation of predictions

Predictions based on normalized activity (**a**) or GLM-based estimates (**b**) from genome-wide K562 SuRE23. Normalized activity predictions are made based on the mean normalized activity level over the length of the tested element. GLM-based estimates are calculated by exponentiating the sum of all coefficients within the bounds of the element. SuRE-GLM marginally improves the correlation coefficient when compared to normalized SuRE predictions.

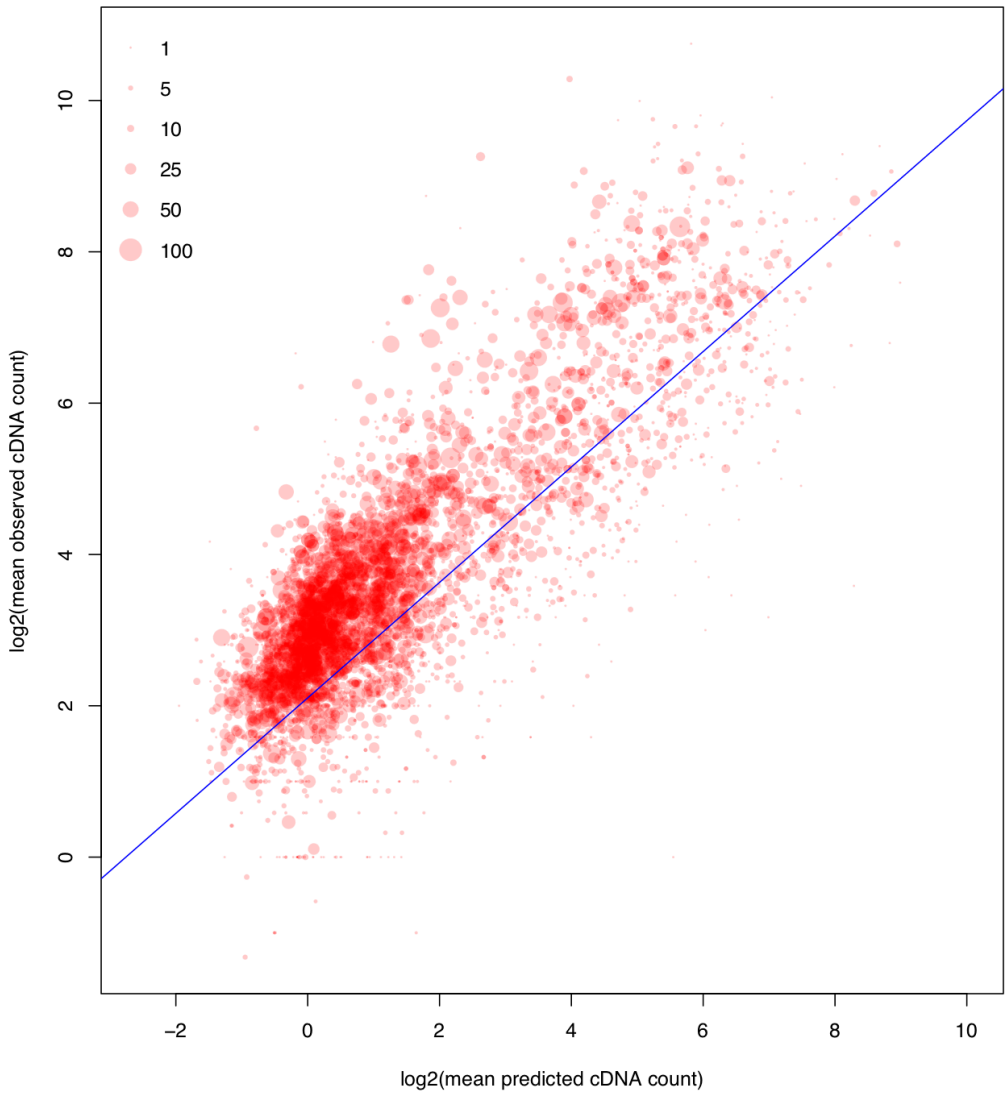

Figure 4.9 Predicted and observed expression of SuRE elements near annotated TSSs

SuRE49 elements were grouped based on whether they started within the same 100bp bin and ended within the same 100bp bin. Variations in coverage and fragment length resulted in bins of different size, represented by dot size. Predictions were based on the K562 genome-wide SuRE42-45 GLM model, while the observations resulted from the separate K562 SuRE49 experiment. The blue line indicates the line of best fit from a separate Poisson fit.

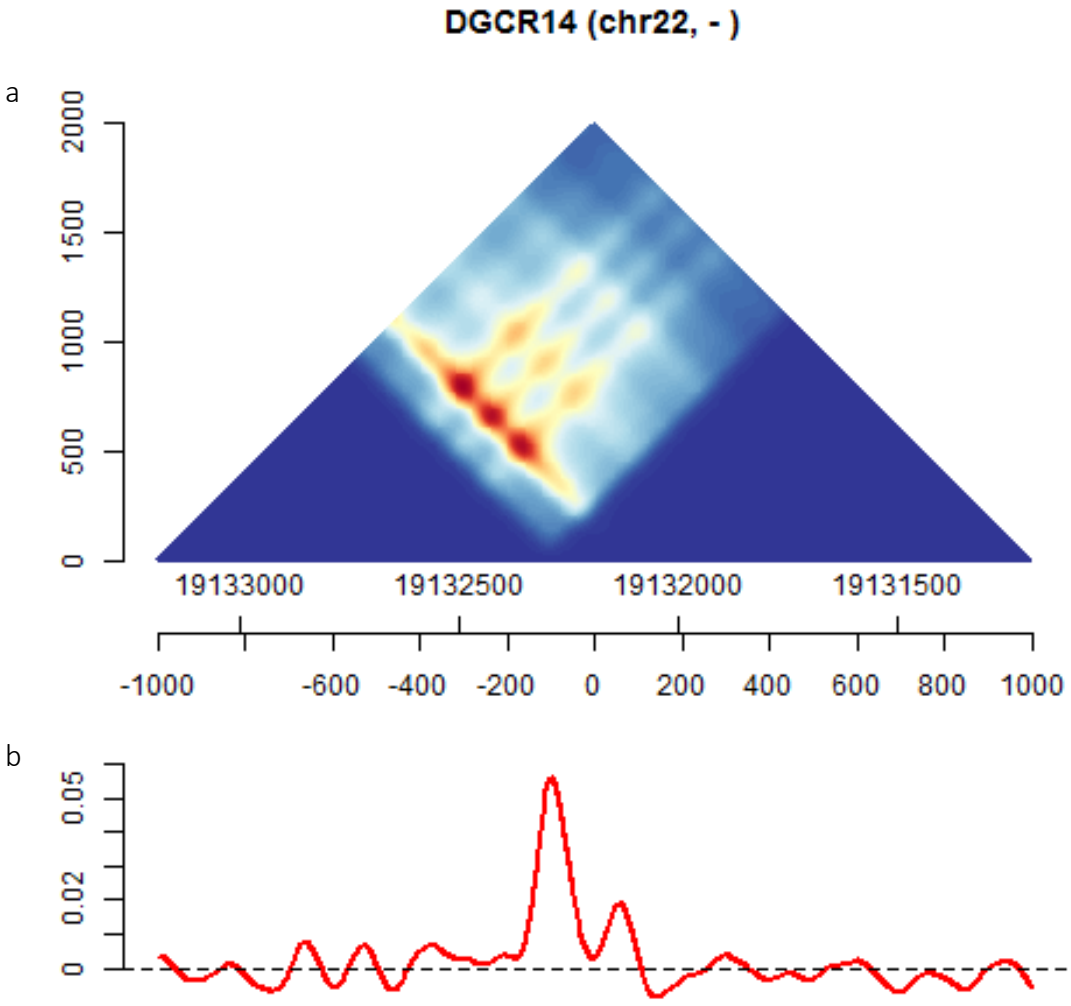

Figure 4.10 Prediction of expression for elements in the DGCR14 TSS region

K562 SuRE23 genome-wide GLM 2-D prediction profile (**a**) and 1-D coefficient profile (**b**) for the region surround the DGCR14 TSS. The position of each point in the 2-D profile reflects the center (x-axis) and length (y-axis) of a hypothetical element, while the color reflects the predicted relative expression of this element, ranging from no expression low (dark blue) to high (dark red).

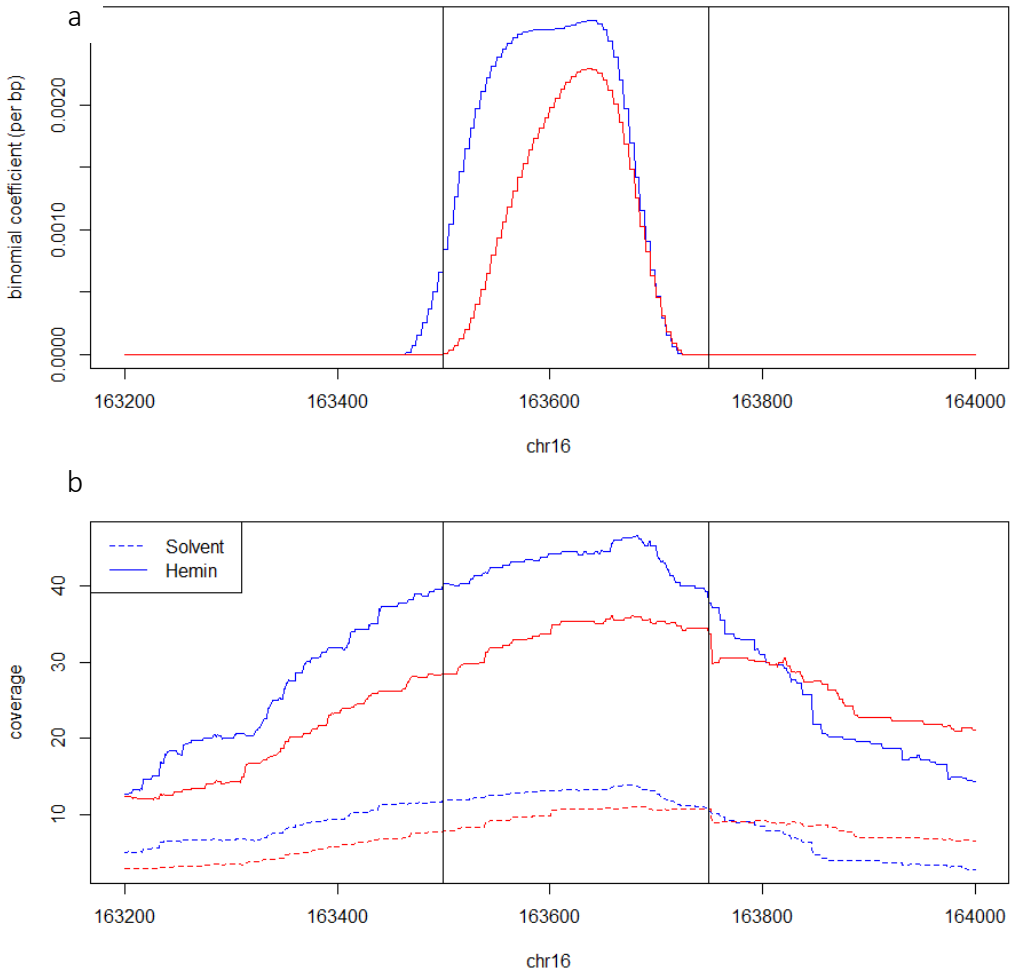

Figure 4.11 Multinomial SuRE-GLM identifies known regulatory region in hemin response experiment

(a) Binomial coefficient profiles on both strands (+ blue, - red) at the R2 α-Locus Control Region (shown by black lines). (**b)** Normalized SuRE activity for same region on both strands in two conditions: hemin (solid line) and solvent control (dotted line).

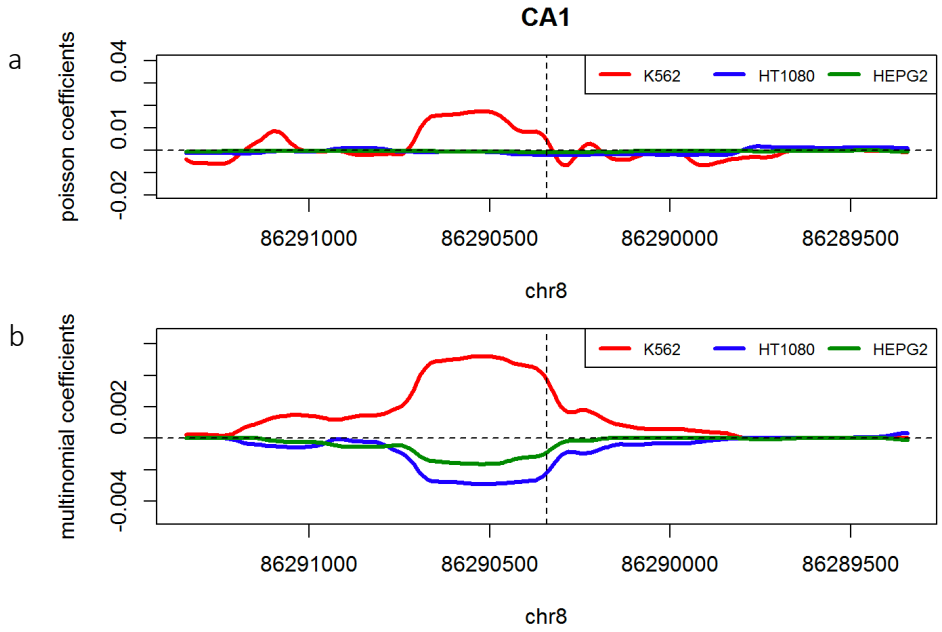

Figure 4.12 Poisson and multinomial coefficient profiles for three cell types at the CA1 TSS region

The Poisson profiles (**a**) are based on the genome-wide Poisson SuRE23 GLMs fit to each cell type separately, while the multinomial profiles (**b**) reflect a single SuRE23 genome-wide multinomial logistic GLM fit to the three cells together.

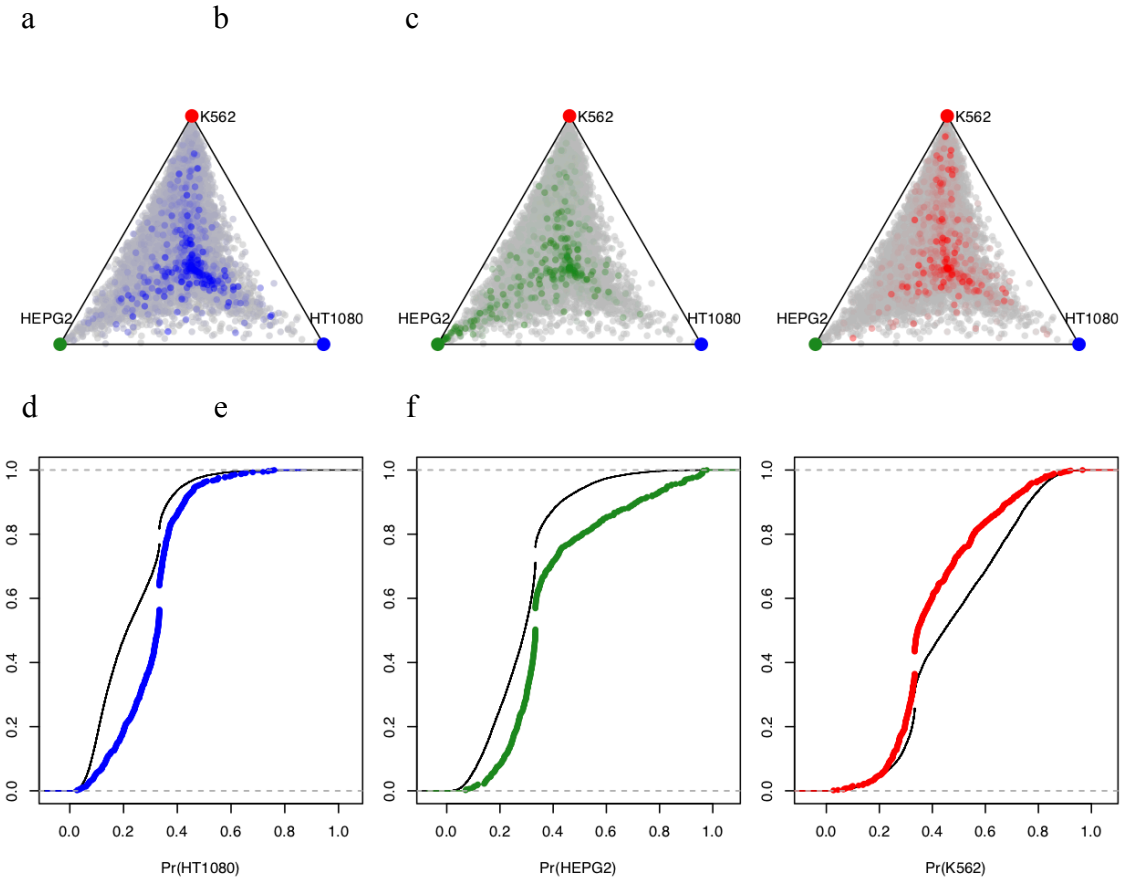

Figure 4.13 Normalized SuRE multinomial probabilities and observed relative GTEX expression in related tissues

Multinomial SuRE23 probabilities are projected on the triangle simplex (**a**-**c**), so that the proximity to each corner reflects the bias towards each cell type. Points are colored based on over-expression of the corresponding tissue in the GTEX RNA-seq database: skin (**a**), liver (**b**), or blood (**c**). Cumulative distributions (**d**-**f**) show the distribution of cell type-specific GLM probabilities for genes that are greater than 4x enriched in the corresponding tissue: blue (**d**), green (**e**), or red (**f**), as well as the distribution for genes that are not enriched (black).

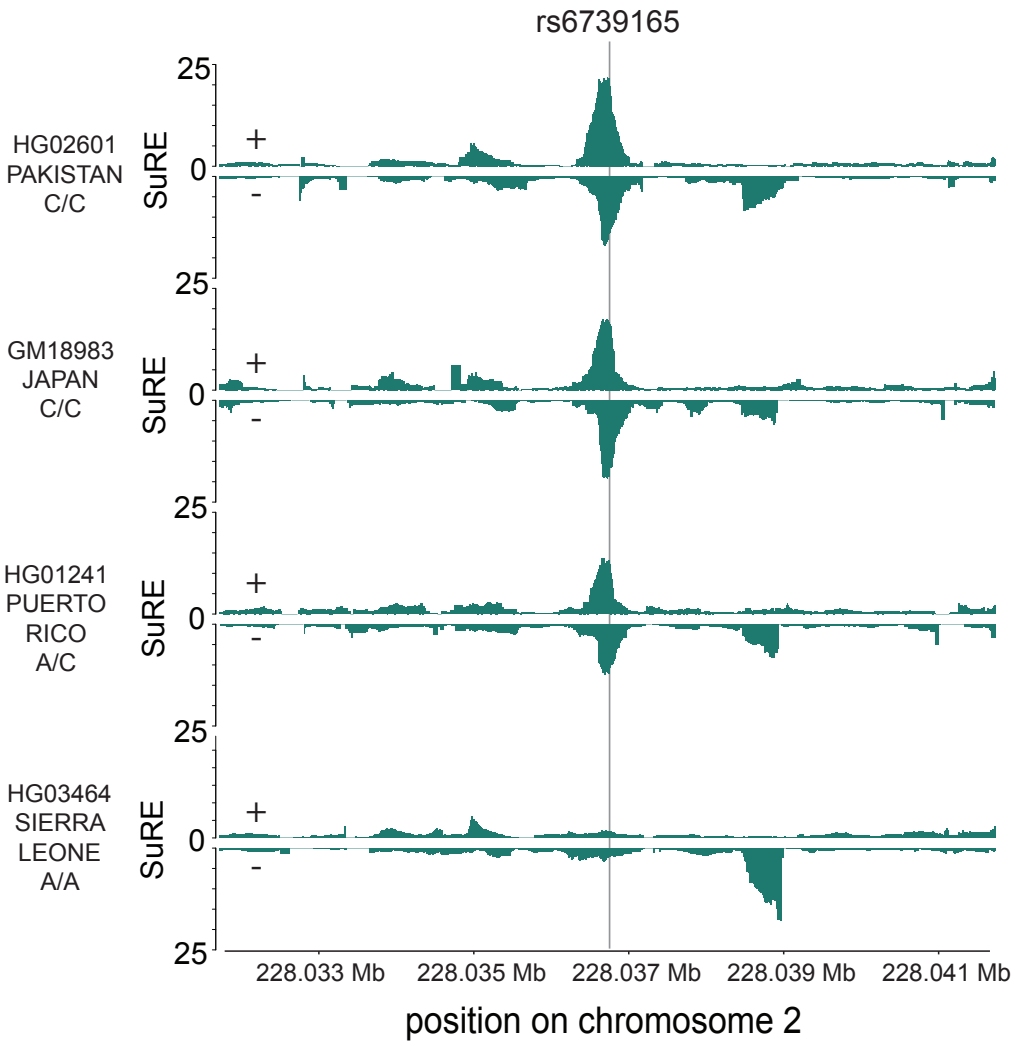

Figure 4.14 Normalized SuRE activity at SNP rs6739165

SuRE signals from the four genomes in an example locus, showing differential SuRE activity depending on the allele (A or C) present.

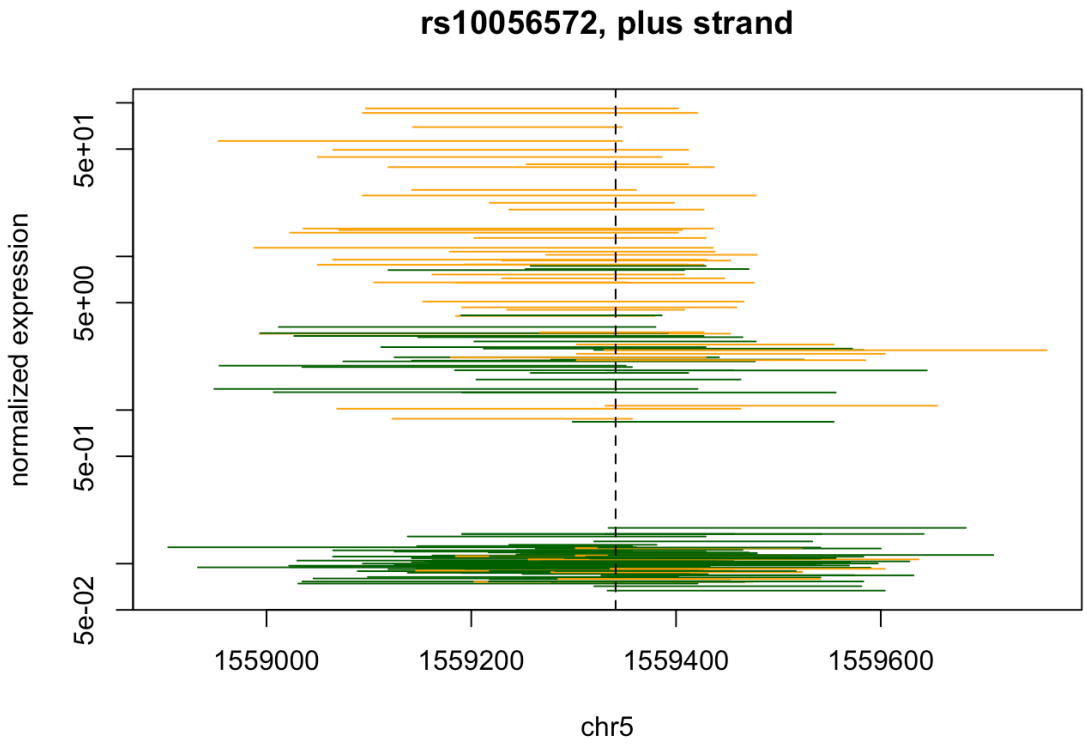

Figure 4.15 SuRE expression of individual fragments overlapping an example differentially expressed SNP locus

SuRE42-25 cDNA counts were normalized for library- and fragment-specific effects using GLM offsets. Fragments are color coded based on whether they contain the reference (T, green) or alternative (C, yellow) allele.

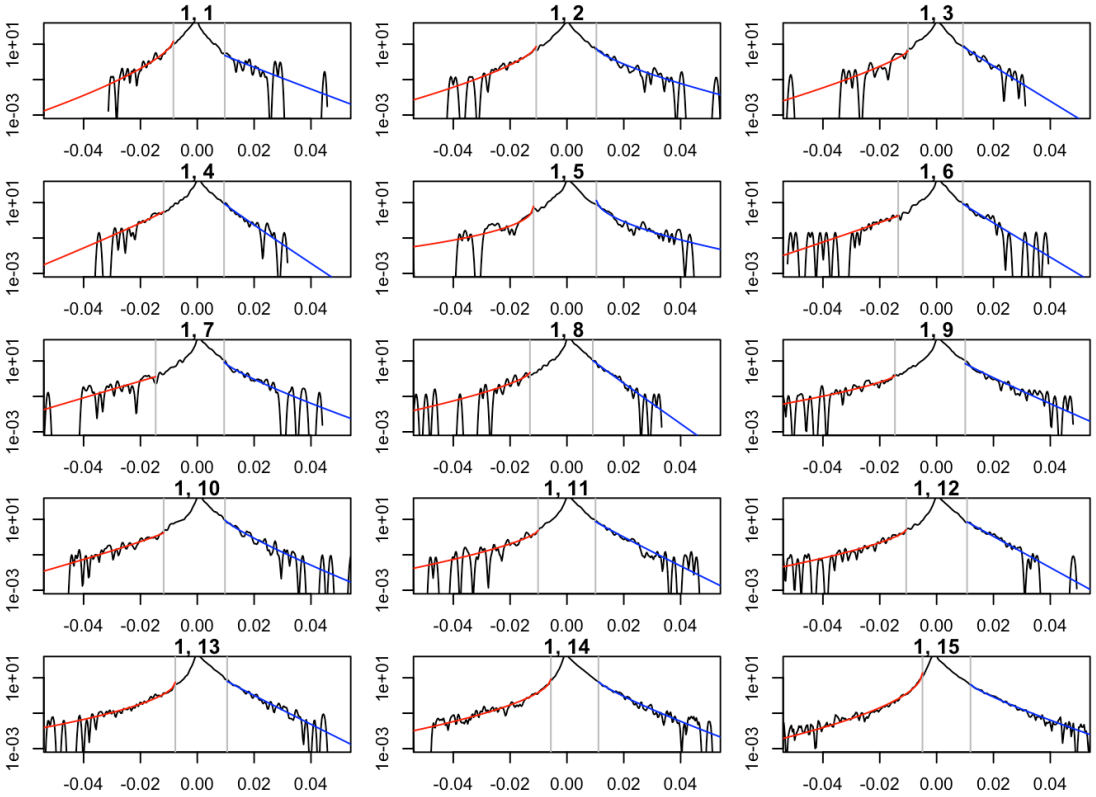

Figure 4.16 Estimated density function for a subset of SNP sample size bins

For each bin, the black line represents the kernel density estimates of the shuffled SNP coefficients, shown on a log-scale. The red and blue lines describe the shifted stretched exponential distributions fit to the 5% of coefficients in the negative and positive tails, respectively.

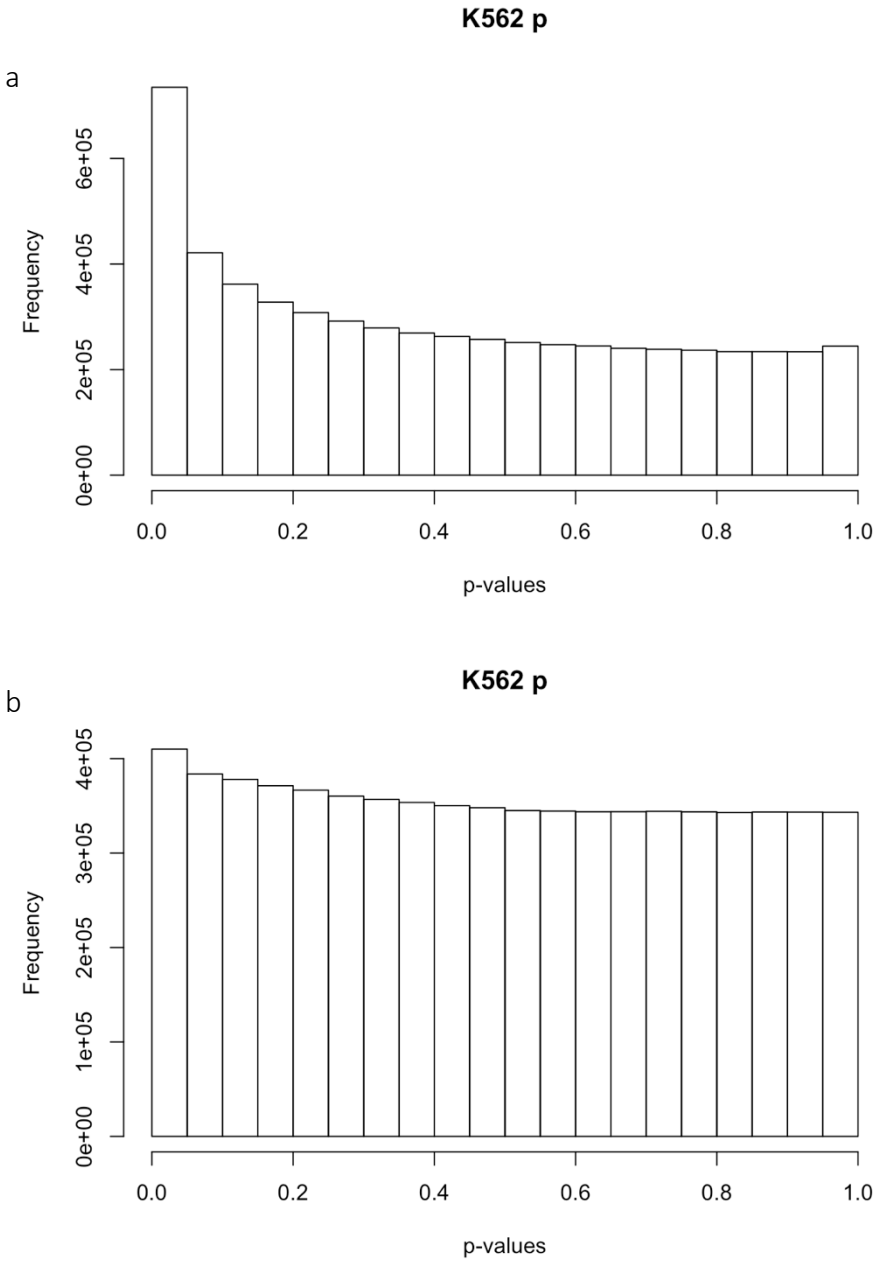

Figure 4.17 P-value distributions for differential SNP analysis methods

P-value distributions for (a) Wilcoxon-Mann-Whitney test and (b) GLM-based methods.

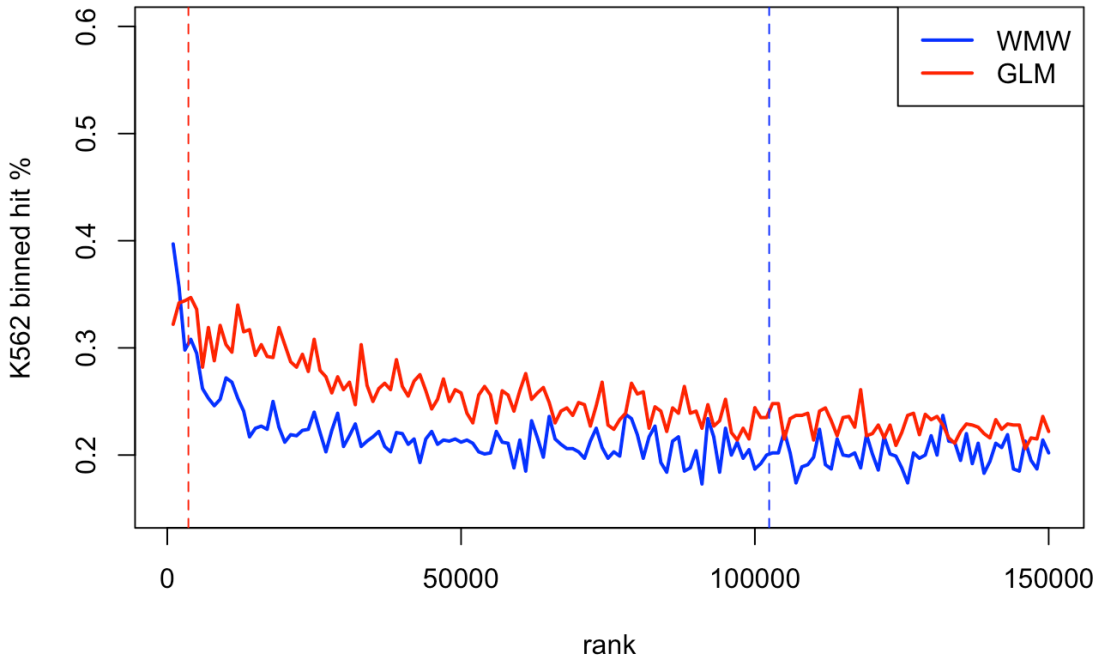

Figure 4.18 Validation of SNP differential analysis methods using predicted TFBS dataset

SNP2TFBS validation of SuRE results using Wilcox-Mann-Whitney p-values (blue) and GLM p-values (red). For each method, SNPs were placed in bins of 1000 based on their p-value rank, then the fraction of SNPs in each bin that also appears in the SNP2TFBS dataset was calculated. Dotted lines represent the 0.1 FDR cutoff for the corresponding method.

5 Future Directions

5.1 Webtool for promoter design and analysis

In Chapter 4, we introduced a method for the visualization of predicted reporter construct activity (Figure 4.12). This method is equivalent to an *in silico* promoter bashing experiment, and can be used to design optimal minimal promoters that can be used in a variety of experimental applications in similar cellular backgrounds. In the future, we aim to develop an interactive web-tool that can enable researchers to visualize and select promoter constructs in just this manner. Users will be able to select and visualize expression within a given cell type and in any genomic region, select a candidate construct with high expression within the region, and identify the start and end positions of this construct in just a few clicks. Similarly, users will be able to enter their own start and end coordinates, and the corresponding position within the triangle will be highlighted for easy assessment. Other information (such as predicted and experimentally verified transcription factor binding sites, SuRE library coverage, and genomic annotations) can be added below the visualization plot to aid in promoter design. This application will leverage SuRE-GLM to produce a valuable tool for experimental biologists, expediting the task of minimal promoter selection.

5.2 Database for minimal promoter elements

While the webtool above allows researchers to take their own criteria into account when designing minimal promoter constructs, a simplified optimization procedure can be automated to produce a database of minimal promoter elements. Within a set of annotated regions (such as gene promoters, enhancers, or repetitive elements), all possible constructs can be evaluated based on the SuRE-GLM model for a given cell type, and a single minimal promoter can be selected based on an algorithmic tradeoff of expression level and length. This can be understood as a more refined version of the peak-calling used in Chapter 3, but based on a SuRE-GLM track and additional length criteria. This will provide a resource to researchers interested in selecting a single promoter based on these simple criteria. Additionally, it will act as a more stringent and cell type-specific annotation source of transcriptionally active promoters than is currently available.

5.3 Motif analysis

As demonstrated by the CpG analysis in Chapter 3, SuRE can be a useful tool in analyzing sequence patterns that are associated with transcription activity. Ideally, SuRE reporter activities could be used to identify sequence motifs that are responsible for driving expression. Unfortunately, my initial motif discovery attempts were unsuccessful, in part due to the "piggy-backing" effect discussed in Chapter 4. This effect causes large inactive regions to become associated with high activity if there is an active promoter within a few hundred base pairs. Fitting a motif model to such a noisy signal fails because these "false positives" drown out the signal from "true positives". Only patterns

broadly associated with active regions, such as CpG dinucleotides, are readily recognized.

SuRE-GLM provides a remedy to this motif discovery problem. By removing the "piggy-backing" effect, positive SuRE-GLM track signals are largely isolated within smaller, functionally important regions where relevant transcription factors are bound. SuRE-GLM coefficients can be aggregated over small regions of identical lengths, and motif discovery programs can be used to predict these signals based on the underlying sequences of these regions. To test this method, I aggregated K562 SuRE23 GLM coefficients into 100bp bins in the 2kb region surrounding all annotated TSSs, then applied MatrixREDUCE [101] to this signal. The results are promising (Figure 5.1). MatrixREDUCE identifies a large signal from CpG dinucleotides as previously observed, but also discovers motifs that resemble whole and partial consensus sequences for transcription factors known to be active K562 cells. These include GABP (consensus CCGGAAG), CREB (CGTCA) [102], and C/EBP (CCAAT) [103]. This suggests that future motif discovery efforts, applied to a broader set of genomic regions, can produce biologically meaningful results.

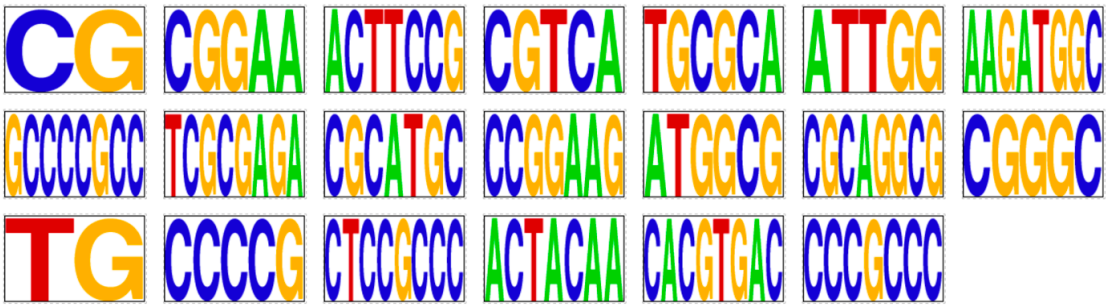

Figure 5.1 Motifs discovered via MatrixREDUCE in K562 TSS regions.

References

1. Consortium, F., et al., *A promoter-level mammalian expression atlas.* Nature, 2014. **507**(7493): p. 462-70.
2. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.* Science, 2008. **322**(5909): p. 1845-8.
3. Haberle, V. and A. Stark, *Eukaryotic core promoters and the functional basis of transcription initiation.* Nature reviews. Molecular cell biology, 2018. **19**(10): p. 621-637.
4. Haberle, V. and B. Lenhard, *Promoter architectures and developmental gene regulation.* Seminars in Cell & Developmental Biology, 2016. **57**: p. 11-23.
5. Lenhard, B., A. Sandelin, and P. Carninci, *Metazoan promoters: emerging characteristics and insights into transcriptional regulation.* Nat Rev Genet, 2012. **13**(4): p. 233-45.
6. Gagniuc, P. and C. Ionescu-Tirgoviste, *Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters.* BMC genomics, 2012. **13**: p. 512-512.
7. Bird, A.P., *DNA methylation and the frequency of CpG in animal DNA.* Nucleic Acids Res, 1980. **8**(7): p. 1499-504.
8. Cohen, N.M., E. Kenigsberg, and A. Tanay, *Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection.* Cell, 2011. **145**(5): p. 773-86.
9. Klemm, S.L., Z. Shipony, and W.J. Greenleaf, *Chromatin accessibility and the regulatory epigenome.* Nature Reviews Genetics, 2019. **20**(4): p. 207-220.
10. Venkatesh, S. and J.L. Workman, *Histone exchange, chromatin structure and the regulation of transcription.* Nature Reviews. Molecular Cell Biology, 2015. **16**(3): p. 178-189.
11. Newburger, D.E. and M.L. Bulyk, *UniPROBE: an online database of protein binding microarray data on protein–DNA interactions.* Nucleic Acids Research, 2008. **37**(suppl_1): p. D77-D82.
12. Riley, T.R., et al., *SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes.* Methods in molecular biology (Clifton, N.J.), 2014. **1196**: p. 255-278.
13. Sung, M.-H., et al., *DNase footprint signatures are dictated by factor dynamics and DNA sequence.* Molecular cell, 2014. **56**(2): p. 275-285.
14. Jolma, A., et al., *DNA-Binding Specificities of Human Transcription Factors.* Cell, 2013. **152**(1): p. 327-339.
15. Amoutzias, G.D., et al., *Choose your partners: dimerization in eukaryotic transcription factors.* Trends in Biochemical Sciences, 2008. **33**(5): p. 220-229.
16. Andersson, R., *Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model.* Bioessays, 2015. **37**(3): p. 314-23.

17. Wortzman, J., *Promoter-bashing as a tool for systems neuroscience : a study of olfactory processing in larval zebrafish*, in *Biomedical Sciences*. 2012, Tufts University.
18. Patwardhan, R.P., et al., *Massively parallel functional dissection of mammalian enhancers in vivo.* Nat Biotechnol, 2012. **30**(3): p. 265-70.
19. Smith, R.P., et al., *Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model.* Nat Genet, 2013. **45**(9): p. 1021-8.
20. Melnikov, A., et al., *Massively parallel reporter assays in cultured mammalian cells.* J Vis Exp, 2014(90).
21. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.
22. Murtha, M., et al., *FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells.* Nat Methods, 2014. **11**(5): p. 559-65.
23. Rencher, A.C. and G.B. Schaalje, *Linear Models in Statistics*. 2008: Wiley.
24. Nelder, J.A.a.B., R. J. , *Generalized Linear Models*, in *Encyclopedia of Statistical Sciences*. 2006.
25. Zhou, Y.H., K. Xia, and F.A. Wright, *A powerful and flexible approach to the analysis of RNA sequence count data.* Bioinformatics, 2011. **27**(19): p. 2672-8.
26. Tibshirani, R., *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society. Series B (Methodological), 1996. **58**(1): p. 267-288.
27. Zou, H., and Trevor Hastie, *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2005. **67**: p. 301–320.
28. Ng, A.Y., *Feature selection, L1 vs. L2 regularization, and rotational invariance*, in *Proceedings of the twenty-first international conference on Machine learning*. 2004, ACM: Banff, Alberta, Canada. p. 78.
29. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.
30. Auer Paul, L. and W. Doerge Rebecca, *A Two-Stage Poisson Model for Testing RNA-Seq Data*, in *Statistical Applications in Genetics and Molecular Biology*. 2011.
31. Smith, G.R. and M.R. Birtwistle, *A Mechanistic Beta-Binomial Probability Model for mRNA Sequencing Data.* PLOS ONE, 2016. **11**(6): p. e0157828.
32. Lindsey, J.K. and P.M.E. Altham, *Analysis of the Human Sex Ratio by Using Overdispersion Models.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1998. **47**(1): p. 149-157.
33. Elkan, C., *Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution*, in *Proceedings of the 23rd international conference on Machine learning*. 2006, ACM: Pittsburgh, Pennsylvania, USA. p. 289-296.

34. Nowicka, M. and M.D. Robinson, *DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics.* F1000Research, 2016. **5**: p. 1356-1356.
35. Venables, W.N.R., B. D. , *Modern Applied Statistics with S*. 2010: Springer Publishing Company, Incorporated. 495.
36. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.
37. Yee, T., *Vector Generalized Linear and Additive Models: With an Implementation in R*. 2015. 1-589.
38. Zhang, Y., et al., *Regression Models for Multivariate Count Data.* Journal of Computational and Graphical Statistics, 2017. **26**(1): p. 1-13.
39. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic acids research, 2015. **43**(7): p. e47-e47.
40. Zhou, Y., et al., *Classifying next-generation sequencing data using a zero-inflated Poisson model.* Bioinformatics, 2017. **34**(8): p. 1329-1335.
41. Zeileis, A., C. Kleiber, and S. Jackman, *Regression Models for Count Data in R.* 2008, 2008. **27**(8): p. 25.
42. Martin, D.C. and S.K. Katti, *Approximations to the Neyman Type A Distribution for Practical Problems.* Biometrics, 1962. **18**(3): p. 354-364.
43. Green, P.J. and B.W. Silverman, *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. 1993: Taylor & Francis.
44. Tibshirani, R., et al., *Sparsity and smoothness via the fused lasso.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. **67**(1): p. 91-108.
45. Kadonaga, J.T., *Perspectives on the RNA polymerase II core promoter.* Wiley Interdiscip Rev Dev Biol, 2012. **1**(1): p. 40-51.
46. 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**(26): p. 15776-81.
47. Kwak, H., et al., *Precise maps of RNA polymerase reveal how promoters direct initiation and pausing.* Science, 2013. **339**(6122): p. 950-3.
48. Core, L.J., et al., *Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers.* Nat Genet, 2014. **46**(12): p. 1311-20.
49. Patwardhan, R.P., et al., *High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis.* Nat Biotechnol, 2009. **27**(12): p. 1173-5.
50. Sharon, E., et al., *Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters.* Nat Biotechnol, 2012. **30**(6): p. 521-30.
51. Melnikov, A., et al., *Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.* Nat Biotechnol, 2012. **30**(3): p. 271-7.

52. Kheradpour, P., et al., *Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.* Genome Res, 2013. **23**(5): p. 800-11.
53. Lubliner, S., et al., *Core promoter sequence in yeast is a major determinant of expression level.* Genome Res, 2015. **25**(7): p. 1008-17.
54. Farley, E.K., et al., *Suboptimization of developmental enhancers.* Science, 2015. **350**(6258): p. 325-8.
55. Nguyen, T.A., et al., *High-throughput functional comparison of promoter and enhancer activities.* Genome Res, 2016. **26**(8): p. 1023-33.
56. Mogno, I., J.C. Kwasnieski, and B.A. Cohen, *Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants.* Genome Res, 2013. **23**(11): p. 1908-15.
57. Dickel, D.E., et al., *Function-based identification of mammalian enhancers using site-specific integration.* Nat Methods, 2014. **11**(5): p. 566-71.
58. Zabidi, M.A., et al., *Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation.* Nature, 2015. **518**(7540): p. 556-9.
59. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
60. Huber, W., et al., *Orchestrating high-throughput genomic analysis with Bioconductor.* Nat Methods, 2015. **12**(2): p. 115-21.
61. Kwasnieski, J.C., et al., *High-throughput functional testing of ENCODE segmentation predictions.* Genome Res, 2014. **24**(10): p. 1595-602.
62. Hubley, R., et al., *The Dfam database of repetitive DNA families.* Nucleic Acids Res, 2016. **44**(D1): p. D81-9.
63. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
64. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
65. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.
66. Gel, B., et al., *regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests.* Bioinformatics, 2016. **32**(2): p. 289-91.
67. Osoegawa, K., et al., *A bacterial artificial chromosome library for sequencing the complete human genome.* Genome Res, 2001. **11**(3): p. 483-96.
68. Duttke, S.H., et al., *Human promoters are intrinsically directional.* Mol Cell, 2015. **57**(4): p. 674-84.
69. Scruggs, B.S., et al., *Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin.* Mol Cell, 2015. **58**(6): p. 1101-12.
70. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(2): p. 261-82.
71. Landolin, J.M., et al., *Sequence features that drive human promoter function and tissue specificity.* Genome Res, 2010. **20**(7): p. 890-8.

72. Kim, T.K. and R. Shiekhattar, *Architectural and Functional Commonalities between Enhancers and Promoters.* Cell, 2015. **162**(5): p. 948-59.
73. Hah, N., et al., *Enhancer transcripts mark active estrogen receptor binding sites.* Genome Res, 2013. **23**(8): p. 1210-23.
74. Arner, E., et al., *Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells.* Science, 2015. **347**(6225): p. 1010-4.
75. Sanyal, A., et al., *The long-range interaction landscape of gene promoters.* Nature, 2012. **489**(7414): p. 109-13.
76. Kim, T.K., et al., *Widespread transcription at neuronal activity-regulated enhancers.* Nature, 2010. **465**(7295): p. 182-7.
77. Blom van Assendelft, G., et al., *The beta-globin dominant control region activates homologous and heterologous promoters in a tissue-specific manner.* Cell, 1989. **56**(6): p. 969-77.
78. Ashe, H.L., et al., *Intergenic transcription and transinduction of the human beta-globin locus.* Genes Dev, 1997. **11**(19): p. 2494-509.
79. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans.* Nature, 2011. **470**(7333): p. 279-83.
80. Hay, D., et al., *Genetic dissection of the alpha-globin super-enhancer in vivo.* Nat Genet, 2016.
81. Dean, A., et al., *Inducible transcription of five globin genes in K562 human leukemia cells.* Proc Natl Acad Sci U S A, 1983. **80**(18): p. 5515-9.
82. Tahara, T., et al., *Heme-dependent up-regulation of the alpha-globin gene expression by transcriptional repressor Bach1 in erythroid cells.* Biochem Biophys Res Commun, 2004. **324**(1): p. 77-85.
83. Faulkner, G.J., et al., *A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE.* Genomics, 2008. **91**(3): p. 281-8.
84. 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**(5): p. 2410-23.
85. Kwasnieski, J.C., et al., *Complex effects of nucleotide variants in a mammalian cis-regulatory element.* Proc Natl Acad Sci U S A, 2012. **109**(47): p. 19498-503.
86. 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**(42): p. 35184-94.
87. Temin, H.M., *Structure, variation and synthesis of retrovirus long terminal repeat.* Cell, 1981. **27**(1 Pt 2): p. 1-3.
88. Vernimmen, D., *Uncovering Enhancer Functions Using the α-Globin Locus.* PLOS Genetics, 2014. **10**(10): p. e1004668.
89. Consortium, T.G., *The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans.* 2015. **348**(6235): p. 648-660.
90. Rasheed, S., et al., *Characterization of a newly derived human sarcoma cell line (HT-1080).* Cancer, 1974. **33**(4): p. 1027-33.

91. Mersch-Sundermann, V., et al., *Use of a human-derived liver cell line for the detection of cytoprotective, antigenotoxic and cogenotoxic agents.* Toxicology, 2004. **198**(1-3): p. 329-40.
92. Lozzio, C.B. and B.B. Lozzio, *Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome.* Blood, 1975. **45**(3): p. 321-34.
93. The Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**: p. 68.
94. Gusev, A., et al., *Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases.* American journal of human genetics, 2014. **95**(5): p. 535-552.
95. Gallagher, M.D. and A.S. Chen-Plotkin, *The Post-GWAS Era: From Association to Function.* American journal of human genetics, 2018. **102**(5): p. 717-730.
96. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic acids research, 2017. **45**(D1): p. D896-D901.
97. Consortium, G.T., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**: p. 204.
98. Albert, F.W. and L. Kruglyak, *The role of regulatory variation in complex traits and disease.* Nature Reviews Genetics, 2015. **16**: p. 197.
99. Tewhey, R., et al., *Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay.* Cell, 2016. **165**(6): p. 1519-1529.
100. Ulirsch, J.C., et al., *Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits.* Cell, 2016. **165**(6): p. 1530-1545.
101. Foat, B.C., H.J. Bussemaker, and A.V. Morozov, *Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.* Bioinformatics, 2006. **22**(14): p. e141-e149.
102. Xia, H., et al., *Gene Expression Profile Regulated by CREB in K562 Cell Line.* Transplantation Proceedings, 2016. **48**(6): p. 2221-2234.
103. Ferrari-Amorotti, G., et al., *The biological effects of C/EBPalpha in K562 cells depend on the potency of the N-terminal regulatory region, not on specificity of the DNA binding domain.* The Journal of biological chemistry, 2010. **285**(40): p. 30837-30850.