Published in final edited form as:

Nat Methods. 2014 May ; 11(5): 559–565. doi:10.1038/nmeth.2885.

FIREWACh: High-throughput Functional Detection of Transcriptional Regulatory Modules in Mammalian Cells

Matthew Murtha[1], **Zeynep Tokcaer-Keskin**[1], **Zuojian Tang**[2], **Francesco Strino**[3], **Xi Chen**[4], **Yatong Wang**[1], **Xiangmei Xi**[1], **Claudio Basilico**[1], **Stuart Brown**[2], **Richard Bonneau**[4,5], **Yuval Kluger**[3], and **Lisa Dailey**[1]

[1]Department of Microbiology, New York University School of Medicine, New York, NY, 10016

[2]Center for Health Informatics and Bioinformatics, New York University School of Medicine, New York, NY 10016

[3]Department of Biology and Biological Sciences, Yale University, New Haven, CT 06511

[4]Department of Biology, New York University, New York, NY 10003

[5]Courant Institute of Mathematical Sciences, Department of Computer Science, NY, NY 10012

Abstract

Promoters and enhancers establish precise gene transcription patterns. The development of functional approaches for their identification in mammalian cells has been complicated by the size of these genomes. Here we report a new method called FIREWACh (***F***unctional ***I***dentification of ***R***egulatory ***E***lements ***W****ithin* ***A****ccessible* ***Ch****romatin*), a high-throughput functional assay for directly identifying active promoter and enhancer elements. FIREWACh simultaneously assessed over 80,000 DNA fragments derived from "nucleosome-free regions" within embryonic stem cell (ESC) chromatin to identify 6,364 new active regulatory elements. Many FIREWACh DNAs represent newly discovered ESC-specific enhancers and their analyses identified enriched binding site motifs for ESC transcription factors including SOX2, OCT4 (POU5f1), and KLF4. Thus FIREWACh identifies endogenous regulators of gene expression and can be used for the discovery of key cell-specific transcription factors. The application of FIREWACh to additional cultured cell types will facilitate functional annotation of the genome and expand our view of transcriptional network dynamics.

corresponding author: Lisa.Dailey@nyumc.org.

ACCESSION NUMBERS to all FIREWACh data sets will be provided subsequent to manuscript acceptance.

Competing Financial Interests
There are no competing financial interests to report.

AUTHOR CONTRIBUTIONS
L.D conceived and designed the study; Y.W. constructed the NFR-GFP-LV reporter libraries and viral stocks; M.M. developed the FIREWACh technology and generated all cell lines, FACs sorting and PCR rescue and analysis of FIREWACh DNAs; Z.T. performed all processing and genomic alignment of Illumina data; S.B generated the in silico DNA datasets; Z.T.K. and X.X performed luciferase assays; C.B. contributed critical discussions; X.C., F.S, Y.K., and R.B. provided bioinformatic support and analyses; M.M. and L.D. wrote the manuscript with additional contributions from X.C and R.B.

Introduction

Embryonic development relies on the establishment of precise temporally- and spatially regulated gene expression patterns. A fundamental determinant of this is the interaction of transcription factors (TFs) with their DNA binding sites within *cis*-regulatory modules (CRMs) of promoters and enhancers, leading to the activation or repression of the associated gene. Thus understanding the regulatory mechanisms underlying distinct gene expression networks requires the global identification of differentially active CRMs within mammalian genomes. Traditional approaches for identifying active CRMs rely on functional assays of individually transfected reporter plasmids harboring putative regulatory regions for a gene of interest[1].

However, the laborious nature of this approach and the cost of DNA synthesis at the required scale prohibit the use of such reporter assays as a tool for global identification of CRMs, and functional approaches for CRM discovery have largely been supplanted by surrogate genome-wide assays mapping protein-DNA interactions for specific TFs, coactivators, cohesion complex proteins, mediator, or histone modifications associated with active promoters and enhancers[2]. The combined data from many such studies have provided an expanded view of chromatin landscapes, and insights into the relationship between gene expression and the dynamics of chromatin modification and remodeling. However, the genomic loci defined by these marks typically span several kilobases and are generally too broad to define the specific DNA sequences mediating promoter or enhancer function. Furthermore, ChIP studies are predictive, but not proof, of the specific locations of CRMs, and validation using functional assays is required for a role in transcriptional activation to be considered definitive.

Recently, several groups have developed massively parallel reporter assays (MPRAs) that permit the simultaneous analysis of hundreds of thousands of reporter plasmids and, thereby, functional assessment of transcriptional activation properties of large numbers of genomic regions[3–6]. However, MPRAs have primarily been used for the detailed dissection of the functional components of previously identified transcriptional regulatory DNA elements rather than as a tool for the discovery of CRMs in mammalian cells. Even with these advances, the enormous size and complexity of mammalian genomes, and the concomitant number of required reporter plasmids, remain among the primary challenges to using functional approaches for the *de novo* discovery of CRMs.

We previously showed that the efficiency of identifying biologically relevant transcriptional regulatory elements can be dramatically increased by focusing the functional analysis on DNA isolated from nucleosome-free regions (NFRs) [7,8] i.e. genomic regions in which nucleosomes are relatively depleted and/or highly destabilized[9]. Importantly, NFRs are where active regulatory elements reside, and represent only 2% of chromatin. Thus focusing a functional analysis on NFR-derived DNA reduces the search space to the most relevant portion of the genome and eliminates the need for *a priori* selection based on criteria such as phylogenetic sequence conservation or chromatin marks.

We have adapted this basic strategy to develop a high-throughput functional assay for the identification of active CRMs, named FIREWACh (*F*unctional *I*dentification of *R*egulatory *E*lements *W*ithin *A*ccessible *Ch*romatin), and have applied this analysis to murine embryonic stem cells (ESCs). We report the identification of more than 6,000 new transcriptional regulatory modules for both promoters and enhancers, and provide evidence that genomic loci detected by FIREWACh correspond to elements that regulate endogenous gene expression in ESCs. Accordingly, analysis of ESC-specific enhancers discovered using FIREWACh can identify enriched binding site motifs for key ESC transcription factors, further demonstrating the utility of FIREWACh to identify essential components of transcriptional networks. The application of FIREWACh documented in this report has dramatically expanded the number of functionally-defined, validated CRMs active in ESCs; its more general application in a range of cell types will permit functional annotation of the genome and facilitate the interpretation of histone- and other chromatin marks for the generation of more accurate transcriptional network models.

RESULTS

Lentiviral reporter library preparation with ESC NFR DNAs

We have previously shown that DNA can be easily isolated from accessible chromatin regions by incubating permeabilized nuclei with restriction enzymes[7,8]. This results in the selective digestion and release of DNA from NFRs, and the diffusion of these molecules out of the nucleus into the surrounding buffer. Following centrifugation to pellet the nuclei and undigested chromatin, the released NFR DNAs can be recovered from the supernatant. The resulting DNA population is enriched for regulatory regions in the virtual absence of background DNA, making it feasible to use reporter-based functional assays to interrogate the DNA population for elements capable of activating transcription (Fig. 1).

We used murine embryonic stem cells (ESCs) as they have been the subject of a multitude of genome-wide ChIP[10–12]- and DNase studies and, accordingly, these annotated chromatin features provide a valuable platform for the evaluation of putative CRMs identified using FIREWACh[13]. ESC nuclei were exposed to either HaeIII or RsaI restriction enzymes (recognition sites 5′-GGCC and 5′-GTAC, respectively), and the two separate NFR-DNA populations were isolated. The HaeIII- or RsaI NFR DNAs were amplified using LMPCR and inserted within the lentiviral (LV) reporter plasmid FpG5, to create two distinct NFR-GFP-LV libraries. FpG5 is a derivative of the self-inactivating FUW lentivirus[14] and contains a cloning site for insertion of the NFR DNAs immediately upstream of a minimal promoter and GFP coding sequences, as well as a hygromycin resistance gene for selection of stably transduced cells (Fig.1). A positive control construct, FGF4enhLV, was created by insertion of *Fgf4* enhancer DNA sequences, which are specifically active in ESCs[15] upstream of the minimal promoter within FpG5. Illumina sequencing revealed a total of 84,240 elements in the two NFR-DNA libraries that were found to be, on average, 154 bp in length and to align with unique positions in the mouse reference genome. These loci strongly correlated with annotated DNaseI-accessible loci in ESCs (AUROC = 0.86, Fig. 2a **and** Supplementary Figure 1), and comprised approximately 4% of the total DNA within accessible chromatin of ESCs (Supplementary Note). In contrast, random DNA fragments with a similar size distribution generated by *in silico* digestion of the mouse genome displayed only weak correspondence with DNaseI-accessible regions, as expected (AUROC=0.52 Fig. 2a **and** Supplementary Figure 1). Together these results confirm that DNAs within the NFR-GFP-LV libraries derive from accessible chromatin regions in ESCs.

Separate analysis of the HaeIII and RsaI NFR DNAs showed that both NFR populations displayed comparable alignment with DNaseI-accessible sites but the genomic regions targeted by each enzyme were largely distinct and non-overlapping (Fig. 2b). Indeed, HaeIII was more likely to target promoter-proximal regions than RsaI (Fig. 2c), likely due to differences in recognition sequence GC content. Thus, the combined use of two enzymes with distinct recognition sequences increases genomic coverage and better captures the diversity of regulatory elements within ESC chromatin.

Functional detection of transcriptional regulatory modules

The lentiviral reporter system since permits the individual activity of thousands of cloned NFR DNAs to be assessed *en masse* following a single transduction. ESCs were transduced with the FpG5 or FGF4enhLV control lentiviruses, or each NFR-GFP-LV library using a multiplicity of infection previously determined to maximize the number of transduced cells while favoring single copy integration events per cell. This consideration is critical for interrogating the activity of individual NFRs as the presence of multiple reporter constructs per cell would increase the false positive rate. The number of ESCs transduced was at least ten fold the estimated complexity of the libraries to increase the likelihood that all NFR-GFP-LVs would be represented in the transduced cell population. While FpG5-transduced cells did not exhibit detectable GFP expression even after Hygromycin selection, GFP$^+$ cells were easily detected for Fgf4enhLV and HaeIII- and RsaI NFR library-transduced cells following Hygromycin selection (Fig 3a and b, Supplementary Figure 2). Independent transductions were performed to create two Biological Replicate (BioRep) samples for each NFR-GFP-LV library. Quantitative flow cytometry analysis showed that 4.9% and 4.5% of cells within RsaI_BioReps 1 and 2, respectively, and 9.5% and 11% of HaeIII_BioReps 1 and 2, respectively, displayed activated GFP expression (Fig. 3b **and** Supplementary Figure 2).

GFP$^+$ cells were isolated using FACS to a purity of >90% (Fig 3a). To ascertain that GFP$^+$ cells harbored LV transgenes with cloned NFR-DNAs capable of activating transcription, genomic DNA was prepared from the GFP$^+$ transduced cells and used as template to recover the NFR-DNAs from integrated LV using PCR. The rescued DNAs were recloned into the FpG5 LV reporter to create secondary NFR-GFP-LV libraries. 63% of cells transduced with the secondary libraries displayed activated GFP expression following transduction of ESCs and selection in hygromycin, demonstrating a dramatic enrichment for transcriptionally active elements compared to the primary NFR-GFP-LV Libraries (Supplementary Figure 3).

As a further test, NFR DNAs recovered from GFP$^+$ cells transduced by the primary NFR-GFP-LV libraries were shuttled into a luciferase reporter plasmid and individually assessed for their ability to activate luciferase expression in transfected ESCs. 78% (42/54) activated luciferase expression more than two-fold above the basal level (Fig. 3c **and** Supplementary Figures 4 and 5). In contrast, only 19.5% (8/41) of similarly tested DNAs recovered from the input library NFR DNAs, and 3% (1/30) of random genomic DNA fragments activated luciferase expression (Fig. 3c **and** Supplementary Figures 4 and 5). In addition to exhibiting a greater percentage of active CRMs, the FIREWACh elements demonstrate a wide range of activities from two fold to >100 fold induction and a ten-fold greater median for luciferase activity than input library NFR DNAs (Supplementary Figure 4).

Using the luciferase assay- validation of individual elements we estimate the false-positive rate (FPR) of FIREWACh to be 0.22 (Supplementary Figure 5). We also note that 26% of positive control, Fgf4enhLV –transduced cells does not display activated GFP expression, suggesting that some FIREWACh lentiviral vectors, which are expected to integrate randomly throughout the host cell genome, will integrate within a context that dampens expression from the transgene (Supplementary Figure 2). Thus the 0.26 false-negative rate (FNR) observed for Fgf4enhLV-transduced cells provides an estimate for the overall false-negative rate of FIREWACh. Notably, however, over 95% of transgenes within the GFP^{-} fraction are ‘true negatives’ as none of 20 elements recovered from these cells tested positive in luciferase assays (Supplementary Figures 5 and 6). Factors affecting the FPR/FNR ratio are presented in Supplementary Figure. 7.

We conclude that NFR-derived DNAs within the integrated LV genomes of GFP+ cells are transcriptional activating modules and that FIREWACh is a highly selective tool to enrich these elements from a general NFR-DNA population.

FIREWACh identifies active endogenous CRMs within ESCs

Transgenic NFR-elements from each population of GFP+ cells were recovered and amplified using PCR, sequenced, and aligned to the mouse genome, resulting in the identification of 6,364 putative new transcriptional regulatory modules (Supplementary Table 1). These modules, which we refer to as “FIREWACh elements”, represent the subset of DNAs within the NFR-GFP-LV libraries able to activate transcription. We observed a good correlation between technical replicates, (on average Pearson *r* = 0.95, Supplemental Figure 8) which serve to measure the efficiency with which elements can be recovered and sequenced from the genomic DNA of sorted cells. Biological replicates showed lower correlation (Pearson *r* = 0.61, Supplementary Figure 8) likely due to incomplete transduction of library constructs and issues related to PCR bias (Supplementary Note M and Supplementary Figure 8). Examination of the overall genomic distribution of these elements showed that FIREWACh elements were more likely to localize within 5 kb of a known TSS than library NFR DNAs, and are therefore more likely to localize within a putative promoter (Fig. 4a).

A key consideration for experimental approaches that utilize reporter plasmid systems is to evaluate whether the CRMs identified by these methods are relevant for regulating endogenous gene expression of a given cell type *in situ*. Analysis using the Genomic Regions Enrichment of Annotations Tool (GREAT)[16] showed that genes associated with FIREWACh elements tended to be those typically expressed in the early mouse embryo (Theiler stages 3–5) and with roles in early embryonic development (Supplementary Table 6). In contrast, genes associated with input library DNAs were typically expressed at later developmental stages or in differentiated somatic cell lineages. A possible explanation for this latter observation is that accessible chromatin in ESCs is a feature of both active promoters for genes expressed in ESCs, and relatively inactive bivalent promoters thought to be ‘poised’ for activation after ESC differentiation[11,17]. Thus NFR DNAs in the input library derive from both transcriptionally active and poised regulatory regions whereas FIREWACh enriches for a subset of DNAs that are associated with genes whose expression patterns are more typical of cells of early embryos and cultured ESCs.

To examine this further, we used ESC RNA-seq data[18] to compare the read density for genes associated with promoter-proximal elements within FIREWACh-, random-, or input NFR-GFP-LV library DNAs This analysis showed notable association of expressed genes with proximal elements of both the input library and FIREWACh DNAs compared to random proximal DNA fragments, with proximal FIREWACh- associated genes displaying the highest expression levels of the three datasets. (Fig. 4b). Together these observations are consistent with the notion that the proximal FIREWACh DNAs represent modules within active promoters of ESCs.

Enrichment for defined combinations of histone modifications and other features can be used to predict the loci and activity status of promoters and enhancers[2,18–22]. We used several previously reported ESC ChIP-seq data sets (Supplementary Table 5) to assess the genomic regions corresponding to proximal and distal FIREWACH DNAs. Overall, both input library and selected FIREWACh elements were more enriched for chromatin features associated with active transcription than were random genomic DNAs (Supplementary Figure 9). Loci corresponding to Proximal FIREWACh elements were enriched for marks typically associated with active promoter regions, including H3K4me3, H3K27Ac, H3K9Ac, and cohesion protein Nipbl while displaying a paucity of H3K27me3 modified nucleosomes that are generally associated with repressed or poised promoters (Fig. 4c). Similar analysis of distal FIREWACh loci (i.e. those > 2 kb from a TSS) revealed a general enrichment of H3K4me1-, but not H4K4me3, consistent with the possibility that they correspond to enhancers (Fig. 4c). Importantly, distal FIREWACh loci were also found to be enriched for H3K27Ac-modified histones and the cohesion factor Nipbl, but not H3K27me3-modified nucleosomes, suggesting that these putative enhancers are active in ESCs (Fig. 4c).

Together these observations suggest that FIREWACh not only selects for DNAs capable of activating transcription in the reporter assay, but that these NFR-derived elements represent modules that actively participate in endogenous gene expression in ESCs *in situ*.

Distal FIREWACh Elements can act as ESC-specific enhancers

Comparison of Distal FIREWACh elements with enhancers that have been previously characterized or predicted in ESCs showed that 75 distal FIREWACh elements overlap with ‘super enhancers’, a class of regulatory regions recently described in ESCs[23], 573 with ‘regular enhancers’, 399 with computationally predicted enhancers[24] and 14 with *in-vivo* validated enhancers from the VISTA database[25]. While these previously predicted enhancers range in size from 1–30 kilobases, FIREWACh elements are typically an order of magnitude shorter, with an average length of approximately 150 bp. A comparison of genomic regions identified using FIREWACh and some of these alternative approaches is shown for the Sgk1 locus (Fig. 5a). Notably, as depicted in this example, many of the FIREWACh- predicted enhancers precisely overlap binding regions for the core ESC factors POU5f1, SOX2, and NANOG (Fig. 5a).

Enhancers are defined by their ability to activate transcription independently of their location, distance, or orientation with respect to gene promoters. To test whether distal FIREWACh elements can function as enhancers, 20 of these DNAs were inserted 2 kb downstream of the TK promoter within the pGL3-luciferase reporter plasmid and individually transfected into ESCs or mouse 3T3 fibroblasts. 90% (18/20) activated transcription more than two-fold in ESCs, in contrast to a control promoter-proximal FIREWACh element (PROX) that did not (Fig. 5b). Interestingly, only 25% (5/20) of these elements were also able to activate luciferase expression in transfected murine 3T3 fibroblasts (Fig. 5b), suggesting that distal FIREWACh DNAs identify active, ESC-specific enhancers.

Distal FIREWACh DNA analysis reveals key TF binding sites

The transcriptional circuitry of ESCs has been well characterized and many TFs required for the establishment or maintenance of the ESC state have been identified[10]. To determine whether the short, distal FIREWACh elements could be used for the identification of ESC-associated TF binding motifs, we used the AME module in the MEME suite[26] and analyzed a database of 651 known murine TF binding motifs (http://cisbp2.ccbr.utoronto.ca/) for those that are over-represented among the 3,789 distal FIREWACh DNAs using random genomic mouse DNA as a background or compared with distal input library DNA. These analyses returned motifs corresponding to SOX2, KLF4, and POU5f1 among those most highly enriched, as well as other ESC-associated TFs (Fig. 5c, Supplementary Fig. 8, and Supplementary Table 7). These observations a provide proof-of principle for FIREWACh as a tool to predict key transcription factors regulating gene expression in cell types with less well characterized transcriptional networks.

DISCUSSION

The development of a high throughput method for the functional identification of cis-regulatory modules in mammalian genomes has long been elusive, primarily due to the large size of these genomes. FIREWACh circumvents this impediment by limiting the analyses to the 2% of cellular chromatin within accessible regions, thereby dramatically reducing the search space while focusing on regions most relevant for endogenous transcriptional regulation in the examined cell type.

Alternate approaches for limiting search space have restricted the functional analysis to individually cloned non-coding genomic regions exhibiting a high degree of DNA sequence conservation[27,28]. This approach has permitted the functional identification of 1659 murine enhancers active at E11, and 81 new CRMs in developing sea urchin embryos[29]. More recently, high throughput functional approaches STARR-seq and eFS have been developed for the identification of enhancers in Drosophila[30,31]. In STARR-seq, sheared DNA fragments are inserted within a non-coding portion of the reporter plasmid transcription unit and enhancer activity is detected in transfected cells by the ability to self-transcribe, permitting identification of 5,499 elements[30]. eFS[31] instead is a highly parallel functional screen for identifying cell-specific enhancers within developing Drosphila embryos. Several hundred genomic DNA segments with predicted roles in mesodermal gene expression are cloned within GFP reporter plasmids. These are used to create thousands of transgenic flies harboring a single reporter plasmid and a second construct for selection of mesodermal cells. Recovery of the transgenic CRMs from GFP$^+$ selected cells identifies developmentally relevant mesodermal enhancers whose analysis can permit motif-based predictions for TF activators.

While both STARR-seq and eFS are elegant systems for CRM discovery in Drosophila, their successful application to the analysis of large mammalian genomes is uncertain. The mouse genome is roughly 23 times that of *D. melanogaster*, suggesting that initial STARR-seq libraries would require over 200 million unique plasmids, and a preliminary study using STARR-seq analysis of human DNAs identified only 6 enhancers from a plasmid library consisting of 1.3 million unique genomic regions derived from 1MB of human DNA. In addition, the relevance of STARR-seq- identified enhancers to endogenous gene expression is less clear since many correspond to closed chromatin regions[30]. While eFS focuses on potentially relevant portions of the genome, an eFS-like approach using transgenic mouse embryos for screening genomic segments would be prohibitively expensive for most researchers.

While the performance of FIREWACh is generally robust there are some limitations. FIREWACh does not provide a quantitative measure of CRM ‘strength’, primarily due to the context-dependent effects of integration site noted above. In addition, issues of non-linear amplification and PCR bias make the uniform recovery of these elements challenging. These considerations, as well as incomplete transduction, likely underlie the lower correlations noted between Biological Replicates (Supplementary Figure 8). Future development of FIREWACh will endeavor to improve these features, perhaps by adopting strategies such as those employed by MPRAs[3–6]. As was mentioned, MPRAs have been used to study position effects or DNA sequence variants of several synthetic or previously identified mammalian or yeast transcriptional regulatory elements but have not yet been applied as tools for the discovery of new elements. While a prohibitively large number of plasmid constructs would be required for analyzing sonicated or enzymatically digested mammalian genomic DNA, combining MPRAs with the focused analysis of NFRs could result in a powerful new variation of both approaches for CRM identification.

Genome coverage by FIREWACh is not comprehensive, nor is it intended to be. Rather than aspiring to the low-resolution prediction of ‘all’ potentially active genomic regions or TF binding sites, FIREWACh identifies functionally validated active modules at high resolution. Enzymes with distinct DNA sequence recognition properties are used to increase the genomic coverage for the genomic features interrogated. Because enzymes are used to isolate these DNAs, each fragment has a discrete 5′ and 3′ end, simplifying genomic alignment and obviating the need for complicated peak-finding algorithms. Most importantly, the short DNAs generated by FIREWACh permit the identification of enriched DNA sequence motifs and relevant TFs in the selected DNAs, as evidenced by the ability of FIREWACh to identify DNA binding motifs for TFs known to play key roles in ESCs.

Application of FIREWACh to less characterized cell types may reveal key TFs with previously unknown roles in their transcriptional regulation.

It is noteworthy that motifs for the insulator binding protein CTCF[32] were among those found to be enriched for elements in the input NFR DNA library (Supplementary Figure 10 and Supplementary Table 7), suggesting that additional types of regulatory elements within accessible chromatin regions are likely to be present in our NFR DNA libraries. These elements, which would include insulators, repressive elements, or matrix attachment regions, could be functionally identified using alternate vectors and screening strategies.

In summary, the use of FIREWACh to identify active CRMs in mouse embryonic stem cells has dramatically increased the number of functionally validated CRMS in this important cell type. The more general application of FIREWACh to a wider range of cell types or classes of regulatory elements will permit functional annotation of the genome, and inform the interpretation of histone modifications and other chromatin marks. Tracking changes to CRM function during the differentiation of ESCs will enable the identification of stage-specific elements and their cognate transcription factors, expanding our understanding of transcriptional network dynamics in development.

Online Methods

Cell Culture

E14ESCs were obtained from ATCC (ES-E14TG2a, CRL-1821) and were maintained in 2i/LIFwith the inhibitors CHIR99021 (3μM), and PD0325901 (1 μM) (Axon Medchem BV, The Netherlands) and 100 u/ml LIF in N2B27, or in standard ESC medium (DMEM supplemented with 15% FCS (Stem Cell Technologies,"ES cult"), 0.1mM non-essential amino acids, 0.1mM b mercaptoethanol, 1X Glutamax (Invitrogen), and 1000 u/ml LIFon plates coated with 0.2% Gelatin. 3T3 were maintained in DMEM(Invitrogen), 15% Bovine Calf Serum and Penicillin/Streptomycin.

Preparation of NFR-DNAs

A detailed protocol for the method of extracting DNA from the NFRs of formaldehyde-crosslinked permeabilized cell nuclei has been reported [8]. Briefly, cultures of E14 ESCs were plated in the absence of feeder cells on eight 15 cm gelatinized tissue culture plates, and grown to 70–80% confluency, (approximately 8×10^7 cells). The cells were crosslinked using 1% Formaldehyde in DMEM for 10 minutes at room temperature, quenched using 0.125M Gycine at RT for 10 minutes, washed with cold PBS, and collected using 2 ml/plate of PBS into 15 ml conical polyethylene tubes on ice with cell scraper. The cells were pelleted by centrifugation and stored at −80°C. Nuclei were permeabilized by resuspension in lysis buffer and incubation for 10 minutes on ice with occasional mixing. The suspension was then dounced 10 times (B pestle), and centrifuged at 2 K rpm at 4°C for 10 minutes. The cells were resuspended in 6.4 ml Buffer 2, incubated for 10 minutes at room temperature on a platform rocker, and the nuclei were pelleted by centrifugation for 10 minutes at 2K rpm (4°C).

Pellets of permeabilized nuclei prepared from ~4×10^{7} cells were resuspended in 2.6 ml of NEB2 (New England Biolabs), and distributed as five 500 µl aliquots in eppendorf tubes on ice. 100 units of HaeIII or RsaI restriction enzyme (New England Biolabs) were each added to the NFR DNA samples, 2 tubes for each enzyme. All samples were incubated at 30°C for 1 hour with gentle mixing every 15 minutes. The reaction was stopped with 20mM EDTA and the samples centrifuged. The supernatants were transferred to new Eppendorf tubes and re-centrifuged at maximal speed for 20 seconds. NFR-DNAs in the supernatants were either subjected directly to crosslink reversal or treated to two rounds of phenol:chloroform extraction prior to crosslink reversal.

Construction of Lentiviral reporter plasmid FpG5 and positive control plasmid Fgf4Enh-LV

The DNA plasmid for generation of the self-inactivating lentivirus FUW was obtained from Addgene (Addgene # 14882)[14]. Coding sequences of the Hygromycin resistance gene were generated using PCR amplification of plasmid pCEP4 using primers PR12 and PR13 (Supplementary Table 4). The amplification product, containing BglII sites at each end, was inserted at the unique BamHI site immediately downstream of Ubiquitin promoter DNA sequences within FUW, creating a hybrid BamHI/BglII site that is resistant to digestion by either enzyme. The resulting construct was named FUWH. A DNA cassette containing the *Fgf4* minimal promoter upstream of GFP coding sequences and transcription stop/polyA signals was PCR amplified using primers PR10 and PR11 (Supplementary Table 4) with the −64GFP plasmid DNA as template[7]. In parallel, a cassette containing DNA sequences of the *Fgf4* enhancer, *Fgf4* minimal promoter, and GFP coding sequences and transcription stop/polyA signals was PCR amplified using oligonucleotide primer sequences PR14 and PR11 (Supplementary Table 4) with enhGFP plasmid DNA as template[7]. Due to the design of the primer sequences, these amplification products contain a PacI recognition site at both the 5′ and 3′ ends. Thus the promoter-GFP and enhancer-promoter-GFP cassettes were each cloned, in both orientations, into the single PacI site upstream of the Ubiquitin promoter within FUWH. Assessment of GFP expression following the transduction of these lentiviruses into F9 cells indicated that plasmids containing the colinear orientation of the GFP and Hygromycin units resulted in somewhat better GFP expression (data not shown). FpG5 has BamHI site proximal to the promoter that accepts BglII digested NFR/Adaptor DNA oligos. All lentiviral construct are depicted in Supplementary Note D.

Preparation of NFR DNAs fpr cloning into LV reporter plasmids

The protocol for the preparation of the double-stranded adaptor DNAs and their ligation to NFR-DNAs is detailed elsewhere [8]. Briefly, HaeIII NFR-DNAs and RsaI NFR-DNAs were each subjected to blunt-end ligation to distinct adaptor DNAs that, after ligation, permit the restoration of the HaeIII- or RsaI sites, respectively, at the NFR/Adaptor junction (Supplementary Table 4). The HaeIII and RsaI adaptors were generated by annealing equimolar amounts of the respective “Linker A” and “Linker B” oligos (Supplementary Table 4). The adaptor sequence contains a BglII site that used for cloning. Annealed linkers were ligated to NFR DNAs overnight at 16°C using the following reaction:

$$30\,\mu\mathrm{l\,NFR}-\mathrm{DNA\,supernatant},\ 10\,\mu\mathrm{l\,5X\,Ligation\,Buffer\,(Invitrogen)},\ 6.7\,\mu\mathrm{l\,15\,uM\,annealed\,Linker},2.3\,\mu\mathrm{l\,H_2O},1\,\mu\mathrm{l\,(5\,Units)\,T4\,DNA\,Ligase\,(Invitrogen)}.$$

After ligation, the DNAs were purified using the QiagenMini-Elute PCR cleanup kit and eluted in 30 µl H_2O. 25 µl of the eluted DNA were assembled in a 50 ul reaction for PCR amplification using the appropriate "AMP" oligonucleotide primer (Supplementary Table 4) and the following PCR conditions: 55°C for 2 min to melt the shorter "Linker B" oligo away from the NFR/Adaptor and then 72°C for 5 min, 95°C for two min, followed by fifteen cycles of 95°C 1′, 60°C 1′, and a final extension at 72°C for 5 min. The amplified DNAs were purified using the Qiagen PCR cleanup kit and eluted in 50 µl H_2O.

Cloning, bacterial transformation, and isolation of LV plasmid library DNA

Purified PCR-amplified NFR DNAs were digested with BglII restriction enzyme digestion overnight and purified using a Qiagen PCR cleanup column using 50 ul water for elution. The DNA concentration of the samples was determined using Nanodrop. Multiple ligation reactions were assembled containing 200 ng BamHI-digested and phospatase-treated FpG5 LV vector plus 40 ng of BglII-cut, LMPCR-amplified NFR DNAs in a 20 ul total ligation reaction using 5 u T4 DNA ligase (Invitrogen, or Roche). Ligation was performed overnight at 16°C.

Commercially available electro-competent Stbl4 bacteria were used for high efficiency transformation using electroporation according to the parameters suggested by the manufacturer (Invitrogen). After purification of the ligation reactions through QiagenMini-Elute columns and elution in 20 ul of water, 1–2 ul of each reaction were used for the electroporation of 20 ul Stbl4. 700–900 S.O.C. broth was added to the Electroporated cells and, after 1 hour recovery, the sample was divided in three and spread over 3 15 cm Agar plates containing 50 ug/ml Ampicillin. This procedure generally yielded several thousand colonies per electroporated sample.

The ligation efficiency for each reaction was determined by transferring cells from 20 colonies into tubes containing PCR reaction components and primers PR2 and PR21 that are complementary to sequences flanking the LV BamHI cloning site. PCR amplification was performed for 25 cycles (95°C 30″, 58°C 1 min, 72°C 30″) and run in a 2% agarose gel in TAE buffer. The percentage of constructs containing insert was determined by the presence of a PCR product migrating slower than that amplified from FpG5 template. Generally, 80–90% of the constructs contained an NFR DNA insert.

These steps were repeated until approximately 5×10^5 colonies each for the HaeIII and RsaI constructs were obtained. The ampicillin plates were stored at 4°C until collection. To prepare the library DNA, 5 ml of cold LB containing 10% Glycerol were added to the plate, and the colonies were collected without further amplification using a cell scraper across the plate surface. All colonies were collected into a single flask, mixed, and then divided into 2 portions. One tube was stored at −20°C and the other half used for preparing library DNAs.

To prepare library DNA from the pooled colonies, we used a Qiagen maxiprep kit following standard protocols for plasmid DNA isolation.

Preparation and Titre of Lentivirus

NFR-lentiviral libraries were prepared using ViraPower (Invitrogen) following manufacturer's protocol. Briefly, 5×10^6 293FT cells were plated to a p-lysene coated 10 cm dish one day prior to transfection in 10% FCS DMEM without antibiotic. On the day of transfection, the medium was replaced with 5 ml of Opti-DMEM/10% FCS without antibiotic. For each NFR-lentiviral library and control lentivirus DNA-Lipofectamine 2000 complexes were generated as follows. 9 ug of ViraPower Packaging Mix and 3 ug of lentiviral plasmid were diluted into 1.5 ml of Opti-MEM medium without serum. In a separate tube 36 ul of Lipofectamine 2000 was diluted in 1.5 ml Opti-MEM and allowed to incubate at RT for 5 min. After incubation diluted DNA and Lipofectamine 2000 solutions were combined with gentle mixing and allowed to incubate at RT for 20 min. The solution was then added dropwise to a single 10 cm dish of near confluent 293T cells and incubated overnight at 37C. The next day medium was changed to complete ESC medium without LIF. Virus containing media was collected at 48 and 72hrs post transfection and stored at −80C.

Prior to freezing lentiviral titres were determined using p24 Antigen ELISA (ZeptoMetrix). Virus containing media was diluted 10^3 and 10^4 fold with DMEM in 450 μl aliquots. 50 μl of lysis buffer was added to each sample. A six point p24 antigen standard curve was generated by successively diluting 125 pg/ml solution of p24 antigen 1:2 to a final concentration of 7.8 pg/ml. 200 ul of standard sample or lentiviral containing media was added to individual wells of the p24 ELISA microplate, covered with plate sealer, and allowed to incubate for 2hr at 37C. After incubation the wells were aspirated and washed five times with 300 μl wash buffer. 100 μl of HIV-1 p24 Detector Antibody is then added to each and incubated at 37C for 1 hr. Wells are washed as before and 100 μlstreptavidin-peroxidase working solution is added to each well incubated for 30 min at 37C. Wells are washed and 100 ul freshly prepared Substrate Working Solution is added to all wells and incubated uncovered at room temp for 15 minutes. 100 μl of Stop Solution is then added and the optical density, OD, of each well is immediately measured at 450 nm using a Spectromax M5 plate reader. The slope, b, and intercept, m, of the standard curve is determined and the final concentration of lentiviral particles per sample is inferred with the equation [Titer = (OD-b-blank)/m x 100 x dilution factor].

Transduction and FACS of ESCs

ESCs were transduced at a MOI of 7 to ensure that the maximum number of cells is transduced while favoring single copy integration. To increase the likelihood that any given NFR-lentiviral genome would be represented, the number of cells transduced was equivalent to more than ten times each library's complexity. Thus 5×10^6 E14 ESCs were plated in complete ESC medium plus LIF in feeder free conditions on a 10 cm gelatin coated dish one day prior to transduction. The cells were then transduced overnight in 10 ml of complete ESC medium plus 8 ug/ml polybrene, containing 3.5×10^7 virus particles for a MOI of 7. The following day the medium was replaced with fresh ESC medium plus LIF. Hygromycin-selection was initiated four days post transduction in ESC medium/LIF containing 250 ug/ml hygromycin B. Cells were selected for hygromycin B resistance for 5 days, with media changed daily. At least 56% of the input NFR-GFP-LV constructs was estimated to have been transduced into the ESCs using these conditions (Supplementary Note). GFP^+ cells were selected using fluorescence activated cell sorting (FACS) on a iCyt Reflection HAPS2 cell sorter. Cells were treated with propidium iodide at 2 ug/ml prior to sorting to counter-select dead cells. The gate was set relative to the profile of FpG5 transduced cells such that the number GFP+ cells observed was less than 0.5%. Cells transduced by NFR-GFP-LV and expressing GFP at a level higher than this set point were collected using FACS. Collected cells were returned to culture, expanded, and subjected to additional 1–2 rounds of FACS to obtain a population of greater than 90% GFP+ cells. A minimum of 10^6 GFP-positive cells was collected from each sort so as to maintain complexity of the integrated transgene population. Post-sort FACS analysis was performed with a minimum 10^5 cells per 100ul sort buffer on a FACSCalibur flow cytometer (BD Biosciences) and analyzed with FloJo software.

For each NFR-GFP-LV library, i.e. HaeIII and RsaI derived libraries, two independent transductions were performed to generate two biological replicates for each library. Each replicate was transduced, selected for hygromycin resistance, and sorted independently to generate cell lines (HaeIII_BioRep1, RsaI_BioRep2, etc.) comprised of pools of NFR-GFP-LV transduced cells. Each cell line was cultured and independently assayed for copy number and NFR sequences. Downstream informatics analysis was also largely done on independent lines prior to pooling end-result NFR sequence information and analysis.

Determination of Transgene Copy Number

Average copy number of integrated lentivirus was estimated using an adapted qPCR approach [33]. Briefly, genomic DNA from each transduced cell line was obtained from 1×10^6 cells with DNeasy (Qiagen). The number of lentiviral vector genomes per cell was determined by quantitative real-time PCR with primers recognizing the GFP transgene while number of mouse genomes was determined using primers recognizing a unique noncoding region of the genome (Primers “Gen-F” and “Gen-R”). A six point standard curve from 1^8 to 1^2 copies was generated by serial dilution of a single plasmid cloned to contain both the GFP and genomic DNA target elements. Amplification reactions contained 5 ul Sybergreen MasterMix, 2 ul gDNA (100 ng), 2 ul H2O, and 0.5 ul each of 5 uM forward and reverse primer. Reactions consisted of 40 cycles at 95°C (15s) then 60°C (1 min) on a BioRad thermocycler. Data were plotted against and interpreted in the linear portion of the standard curve where regression coefficient was greater than 0.98. The average integrated copy number was determined by dividing the calculated number of lentiviral genomes by the total number of mouse genomes present in the DNA sample of each transduced line and measured in triplicates.

Luciferase Assays

a. Reporter constructs—The pGL3 luciferase reporter plasmid was modified to contain the 162bp minimal Fgf4promoter (fgfprom-luc). This plasmid has a BglII site upstream of the fgf4 promoter sequences used for inserting test DNAs. Oligonucleotide primers used to recover library NFR or FIREWACh DNAs from the lentiviral plasmids and prepare them for InFusion cloning into the fgfprom-luc plasmid were designed as follows: the 5′ portion consisted of 15 bases complementary to the sequence flanking the fgf4prom-luc plasmid BglII site, and the 3′ portion contained sequences complementary to sequences flanking the NFR DNA cloning site within the lentiviral plasmid. PCR amplification was performed using either input lentiviral NFRGFP-library plasmid DNA or the genomic DNA isolated from FACS-sorted GFP+ cells as template. The amplified fragments were used for In-Fusion (Clontech) directional cloning into the fgfprom-luc plasmid. Primers InFusionpGL3R, InFusResFA were used to clone into the proximal BglII site of fgfprom-luc while DisInFusResFA and DisInFuspGL3R were used at the distal BamHI site of TKluc (Supplementary Table 4). Recombinase reactions were assembled according to the manufacture's protocol.

To generate luciferase reporter constructs to assay random genomic DNA fragments, three micrograms of purified gDNA weredigested with either HaeIII or RsaI, and DNA fragments ranging from 100–300bp were gel-purified and cloned into the SmaI site of the fgfprom-luc plasmid.

b. Transfections and luciferase assays—E14 cells grown ESC medium with 1000U/mlLIF were seeded on 0.2% gelatin coated 96 well plates at 5×10^4 cells/well. Cells were transfected using 250ng plasmid DNA and 1.25ul Lipofectamine 2000 (Invitrogen) and supplemented with OPTIMEM and LIF (1000 u/ml) fora total volume of 150 ul/well. 4 hours after transfection the medium was changed to complete ESC medium plus 1000U/ml LIF. 24 hours after transfection, lysates were prepared and luciferase assays were performed as instructed by the manufacturer (Promega). The protein concentration of the lysates was determined (Bio-Rad) and used to normalize the samples. The luciferase activities of all test constructs were calculated relative to the activity displayed by the fgfprom luciferase construct containing only the minimal fgf4 promoter upstream of the luciferase gene.

PCR Rescue of Functionally Selected NFR-DNAs and High-Throughput Sequencing

NFR-DNAs were rescued from either the initial lentiviral plasmid libraries or gDNA of GFP + selected cells using PCR in a method adapted from bacterial rRNA sequencing[34]. In this method, Illumina sequencing adaptors are included in the primers, permitting one step amplification and sequencing library preparation [34]. Primers (termed FMS-F/R, Supplementary Table 4) were designed such that they contain recognition sequences complimentary to lentiviral sequence flanking NFR-DNA and Illumina adaptor sequence for paired-end sequencing in addition to a 6base pair Index sequence. Six PCR reactions (10 ul Phusion Polymerase buffer, 1 μl 10 mMdNTP, 2.5 μl 10 μM forward and reverse primer, 1.5 μl DMSO, 0.5 μl (NEB) 50 ng DNA, 31 μl H_20; 16 cycles with 550C annealing temperature) per plasmid library were pooled and sequenced.

FIREWACh elements were recovered from the genomic DNA of a least 1×10^6 FACS-sorted GFP+ cells using PCR. In this case 10 PCR reactions were performed using the same conditions as above but 100 ng gDNA and 23 cycles of amplification. The 10 reactions were then pooled.

Each sample was amplified with primers containing Illumina adaptor sequence with 6bp indexing sequence. This allowed us to pool up to six samples within a single lane on the MiSeq machine. Input library derived NFRs were sequenced together using three of the barcodes while FIREWACh NFRS, i.e. NFR's rescued from GFP+ cells, were run with six samples per lane, each sample representing NFRs rescued from an independent biological replicate (i.e. HaeIII_BioRep1 etc). Technical replicates consisting of independent PCR rescued NFR sequencing libraries were sequenced on separate days.

Samples were run on a miSeq sequencer with the miSeq cartridge version 2, as a 2 × 150 bases run, with a 50% PhiX library spiked in to compensatefor potential lowdiversity in the libraries. In order to ensure efficient binding of sequencing primers we designed and used custom Read 1, Index and Read 2 primers (sequences in Supplementary Table 5) of which 17 ul of custom primers at 100 uM were spiked into the Illumina Read 1, Read 2 and index reads positions in the cartridge.

Genomic alignment

IlluminaMiSeq 2x 151 bp data were pre-processed by demultiplexing and trimming of 7 bp from the 5′ end and 44 bp from the 3′ end, yielding a data set of 2 × 100 bp sequence reads. Paired end sequences were aligned to the mouse reference genome (mm9) using BWA[35] software with default settings. Read pairs were filtered from the final data set if either read failed to map to the genome, if both reads did not map in the proper orientation, if the mapping quality score of both reads was less than 25, or if neither read had a unique map location on the genome. Target sites were identified as loci where paired reads both aligned entirely within a 500 bp genomic region.

Each Biological Replicate sample was independently sequenced three times (i.e. three technical replicates) and all sequencing data for all samples were then merged to create the final list of FIREWACh genomic regions (6,364 elements). The final input NFR DNA library dataset was generated by merging the mapped loci from replicate of each enzyme library as well as with all FIREWACh loci generating a list of 84,240 elements.

In Silico Generation of Random genomic DNA Fragments

To create a dataset of random genomic DNAfragments, we generated a list of genomic loci corresponding to digestion of the murine reference genome (mm9) with HaeIII or RsaI. We utilized a script to scan chr19 for pairs of each restriction sites as a regular expression separated by a variable region of DNA up to 500bp in length so that the size distribution of the *in silico* fragments would be comparable to that of the enzymatically derived NFR libraries. The resulting *in silico* DNA fragment dataset was comparable in number to the input NFR library (61,844 random elements versus 84,240 NFR-DNAs) and was then reformatted as genomic loci in a bed file. Random distribution of elements was confirmed in subsequent analysis as this list generated correlative scores expected of a randomly distributed set of loci (e.g. in comparison to DNaseI-HS, the random elements had an AUROC=0.52, typical of random elements, Supplementary Figure 1).

Bioinformatic Data Analysis

To investigate the chromatin status of FIREWACh elements we utilized several publically available sequence files for ChIP-seq and DNaseI-HS sequencing experiments. Data were obtained from the NIH's sequence read archive (SRA) and the UCSC genome browser (for full list of data sets used, see Supplementary Table 5). In the case of the H3K4me1/3, H3K27me3/Ac, H3K9Ac chromatin marks, reads were remapped using the mm9 genome as reference with bowtie[36] (version 2) with the options "-n 1 -k 1 -m 20 --best --strata -p 8 --chunkmbs 1024". Tophat (version 2.0.4,) and cufflinks (version 2.0.2,) with default parameters were used to obtain FPKM values for all genes from RNA-seq data.

a. GREAT Analysis—Bed files of all input library NFR DNAs, HaeIII library elements, or RsaI elements, or FIREWACh DNAs were analyzed using the Genomic Regions Enrichment of Annotations Tool[16] with the settings: Mouse: NCBI build 37, Whole genome background, basal plus extension, Proximal 5kb upstream, 1kb downstream, plus distal up to 100kb. Datasets were analyzed using both the Significance by Both and Significance by Region-based Binomial views.

b. Comparison with genomic regions of DNaseI Hypersensitivty in ESCs—The relation between DNase I HyperSensitivity and the input library NFR DNAs was investigated using Receiver Operating Characteristic (ROC) curves. In particular, we verified that open regions (i.e. with high DNaseI HS coverage) could predict the location of genomic regions covered by input NFR DNA reads. For each dataset (All library, HaeIII NFR DNA library, RsaI NFR DNA library and in silico DNA library), the genome was divided into non overlapping bins of 1kbp and the bin was classified as positive if it contained at least one NFR DNA read or negative if it did not intersect any element. The coverage of DNase I Hypersensitivity Hot Spots (ES-CJ7 Pk1, UCSC genome browser) for each bin were used as classifier in order to build the ROC curve (i.e. DNase I HS coverage is utilized to predict whether a bin would contain any of the elements). The area under the curve of the receiver operating characteristic (AUCROC) was 0.8637 for the combined (All) library, 0.8780 for RsaI-, and 0.8539 for HaeIII DNAs. The AUCROC for the *in silico* reads (0.5267) was not significantly different from 0.5, which is the expected value of random reads. The area under the curve was calculated and plotted in graph form as presented in Supplementary Figure 1.

c. RNA-seq Analysis—For each unique read, we considered the expression of the nearest gene. The expression data for ES cells in 2i medium were obtained from Marks *et al.*[18]. Tophat (version 2.0.4,) and cufflinks (version 2.0.2,) with default parameters were used to obtain FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values for all genes. To aid visualization and analysis, we scaled the FPKM values logarithmically as $\log_2(1+\text{FPKM})$. The significance between different libraries was assessed using the nonparametric Kruskal-Wallis test[37].

d. Carpet methods—High-density maps of coverage of chromatin marks around FIREWACh loci was visualized as previously described[38]. Each horizontal line represents the center of a unique FIREWACh DNA. The expression of the nearest gene is color-coded from red (expressed) to green (not expressed) and the expression values are used to sort the horizontal values. The ChIP-seq signal in the ±1 kb region around each FIREWACh locus was determined for H3K4me1 H3K4me3, H3K27me3, H3K27Ac, and DNaseI hypersensitivity. The ChIP-seq and DNaseI HS signals were normalized by total number of reads, The gene expression data was quantile-normalized over all genes. In case of identical, overlapping or nearby FIREWACh loci (< 100 b), the profile of only one read was used in the high-density map.

e. Correlation Coefficients—The correlation coefficients for technical or biological replicates were calculated by binning the genome into windows of 100bp and computing the Pearson correlation of the genomic coverage between all pairs of the coverage vectors, which represent our sequencing datasets. The calculation was done using an in-house Java code. Technical replicates consist of independent sequencing library preparations from a common template (e.g. GFP$^+$ cells transduced with HaeIII NFR-GFP LV), and were generated to assess the reproducibility of our sequencing library preparation protocol. Three independent runs of each biological replicate were compared pairwise and the average of all taken for a given enzyme-derived NFR library (eg. 0.86–0.98 for HaeIII_BioRep1). These replicates did not correlate with random NFRs generated in silico (Average of 0.001 for HaeIII and RsaI both).

For measuring the correlation between biological replicates, the total reads from all three technical replicates of FIREWACh-seq elements were combined into a single file. For example, HaeIII_BioRep1, contains three technical sequencing replicates generated from the recovery of cloned NFR DNAs from the integrated LV vectors within HaeIII_BioRep1 transduced GFP+ ESC. The HaeIII_BioRep1 sequences were combined with RsaI_BioRep1 sequences into a single file (Rep1). Rep2 was similarly generated from HaeIII_BioRep2 and Rsa_BioRep2. Comparison of Rep1 and Rep2 generated the correlation between biological replicates (0.61).

f. Motif Analysis—Motif enrichment analysis was performed for the distal FIREWACh elements using the AME module in the MEME suite[26] with the following command line options: "--method mhg --scoring totalhits --length-correction". Random gDNA elements from /in silico/ digestions were used as background model. We also analyzed distal elements using input NFRs as background. This allowed us to determine which motifs were enriched above those obtained from open chromatin alone. P-values were calculated based on the multi-hypergeometric distribution and corrected for multiple hypothesis testing. Analysis was performed using a database of known motifs that covers approximately 50% of mouse TFs (Supplementary Figure. 8 and Supplementary Table 7). This curated compendium of motifs can be accessed via the Timothy Hughes lab webpage (http://cisbp2.ccbr.utoronto.ca/) and is derived from protein binding microarray data, HT-SELEX, and ChIP-seq. Most of the motifs used are also redundantly availiable in the JASPER and TRANSFAC databases[39,40]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank A. Heguy, E. Mazzoni, M. Stadtfeld, and U. Basu-Roy for their critical reading of this manuscript. We also thank A. Mansukhani and members of the Skok lab for helpful discussions, and the Genome Technology Center and Flow Cytometry and Cell Sorting Center at New York University Langone Medical Center for technical assistance. This work was supported by an Empire State Stem Cell Board grant through the New York State Department of Health (NYSTEM Contract #CO24322) to L.D., and a postdoctoral fellowship from NIH NCI institutional training grant 1T32CA160002 awarded to M.M.

References

1. Landolin JM, et al. Sequence features that drive human promoter function and tissue specificity. Genome Research. 2010; 20:890–898. [PubMed: 20501695]
2. Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nature Reviews Genetics. 2012; 13:469–483.
3. Patwardhan RP, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012; 30:265–270. [PubMed: 22371081]
4. Melnikov A, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012; 30:271–277. [PubMed: 22371084]
5. Akhtar W, et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. Cell. 2013; 154:914–927. [PubMed: 23953119]
6. Mogno I, Kwasnieski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. Genome Research. 2013; 23:1908–1915. [PubMed: 23921661]
7. Yaragatti M, Basilico C, Dailey L. Identification of active transcriptional regulatory modules by the functional assay of DNA from nucleosome-free regions. Genome Research. 2008; 18:930–938. [PubMed: 18441229]
8. Murtha M, Wang Y, Basilico C, Dailey L. Isolation and analysis of DNA derived from nucleosome-free regions. Methods Mol Biol. 2013; 977:35–51. [PubMed: 23436352]
9. Cockerill PN. Structure and function of active chromatin and DNase I hypersensitive sites. FEBS J. 2011; 278:2182–2210. [PubMed: 21501387]
10. Jaenisch R, Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell. 2008; 132:567–582. [PubMed: 18295576]
11. Bernstein BE, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006; 125:315–326. [PubMed: 16630819]
12. Chen X, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell. 2008; 133:1106–1117. [PubMed: 18555785]
13. Shen Y, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012; 488:116–120. [PubMed: 22763441]
14. Lois C, Hong EJ, Pease S, Brown EJ, Baltimore D. Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors. Science. 2002; 295:868–872. [PubMed: 11786607]
15. Yuan H, Corbi N, Basilico C, Dailey L. Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. Genes Dev. 1995; 9:2635–2645. [PubMed: 7590241]
16. McLean CY, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010; 28:495–501. [PubMed: 20436461]
17. Voigt P, Tee WW, Reinberg D. A double take on bivalent promoters. Genes Dev. 2013; 27:1318–1338. [PubMed: 23788621]
18. Marks H, et al. The Transcriptional and Epigenomic Foundations of Ground State Pluripotency. Cell. 2012; 149:590–604. [PubMed: 22541430]
19. Rada-Iglesias A, et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2010:1–7.
20. Visel A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009; 457:854–858. [PubMed: 19212405]

21. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010; 28:817–825. [PubMed: 20657582]

22. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007; 39:311–318. [PubMed: 17277777]

23. Whyte WA, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013; 153:307–319. [PubMed: 23582322]

24. Chen CY, Morris Q, Mitchell JA. Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features. BMC Genomics. 2012; 13:152. [PubMed: 22537144]

25. Visel A, Minovitsky S, Dubchak I. VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic acids. 2007

26. Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research. 2009; 37:W202–8. [PubMed: 19458158]

27. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009; 461:199–205. [PubMed: 19741700]

28. Pennacchio LA, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006; 444:499–502. [PubMed: 17086198]

29. Nam J, Dong P, Tarpine R, Istrail S, Davidson EH. Functional cis-regulatory genomics for systems biology. Proc Natl Acad Sci USA. 2010; 107:3930–3935. [PubMed: 20142491]

30. Arnold CD, et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013; 339:1074–1077. [PubMed: 23328393]

31. Gisselbrecht SS, et al. Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. Nat Methods. 2013; 10:774–780. [PubMed: 23852450]

32. Herold M, Bartkuhn M, Renkawitz R. CTCF: insights into insulator function during development. Development. 2012; 139:1045–1057. [PubMed: 22354838]

33. Charrier S, et al. Quantification of lentiviral vector copy numbers in individual hematopoietic colony-forming cells shows vector dose-dependent effects on the frequency and level of transduction. Gene Ther. 2011; 18:479–487. [PubMed: 21160533]

34. Caporaso JG, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA. 2011; 108 (Suppl 1):4516–4522. [PubMed: 20534432]

35. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26:589–595. [PubMed: 20080505]

36. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009; 10:R25. [PubMed: 19261174]

37. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Journal of the American statistical. 1952

38. Gao Z, et al. PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Mol Cell. 2012; 45:344–356. [PubMed: 22325352]

39. Mathelier A, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Research. 201310.1093/nar/gkt997

40. Matys V, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Research. 2006; 34:D108–10. [PubMed: 16381825]

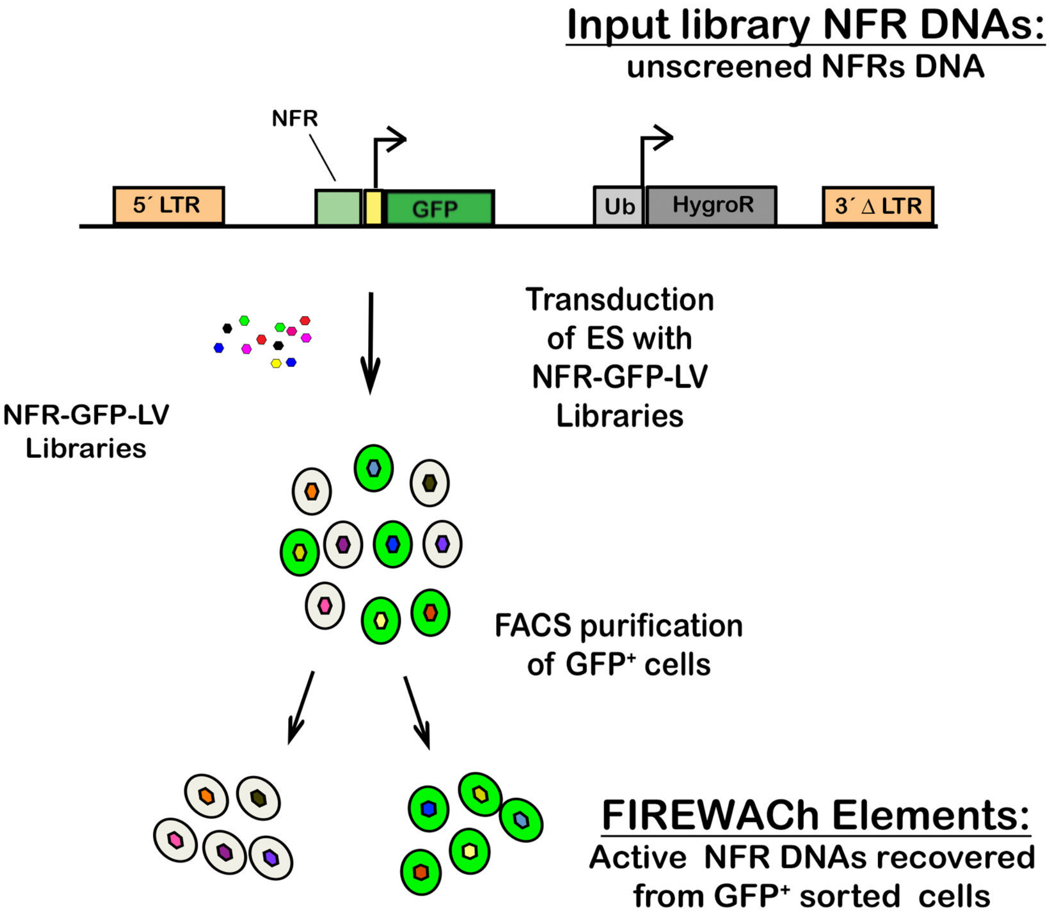

Input library NFR DNAs:
unscreened NFRs DNA
NFR
5´ LTR
GFP
Ub
HygroR
3´ Δ LTR
Transduction of ES with NFR-GFP-LV Libraries
NFR-GFP-LV Libraries
FACS purification of GFP+ cells
FIREWACh Elements:
Active NFR DNAs recovered from GFP+ sorted cells

Figure 1. Overview of FIREWACh

LV reporter plasmids a contain cloning site for NFR DNAs (NFR) upstream of minimal Fgf4 promoter sequences (yellow) and transcription start site (arrow), Ubiquitin promoter (Ub), and Hygromycin Resistance gene (Hygro^R). Small colored circles represent lentiviral (LV) particles; large circles represent GFP$^+$ (green) or GFP$^-$ (white) transduced cells.

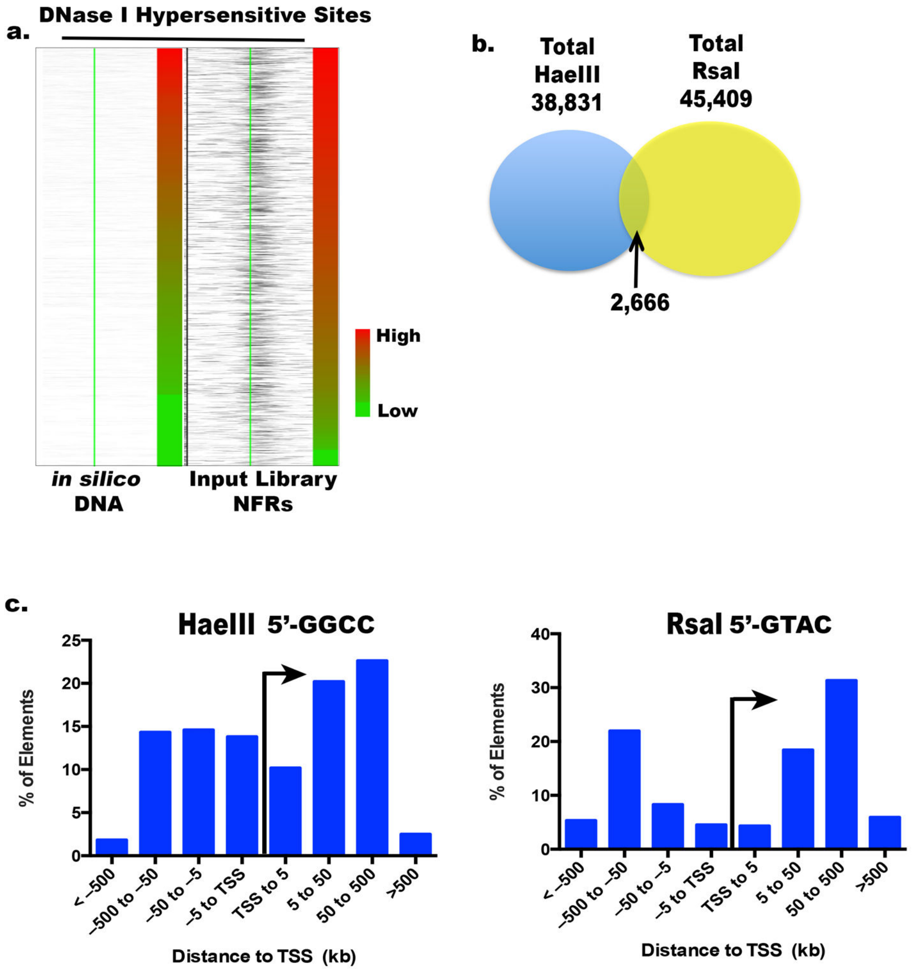
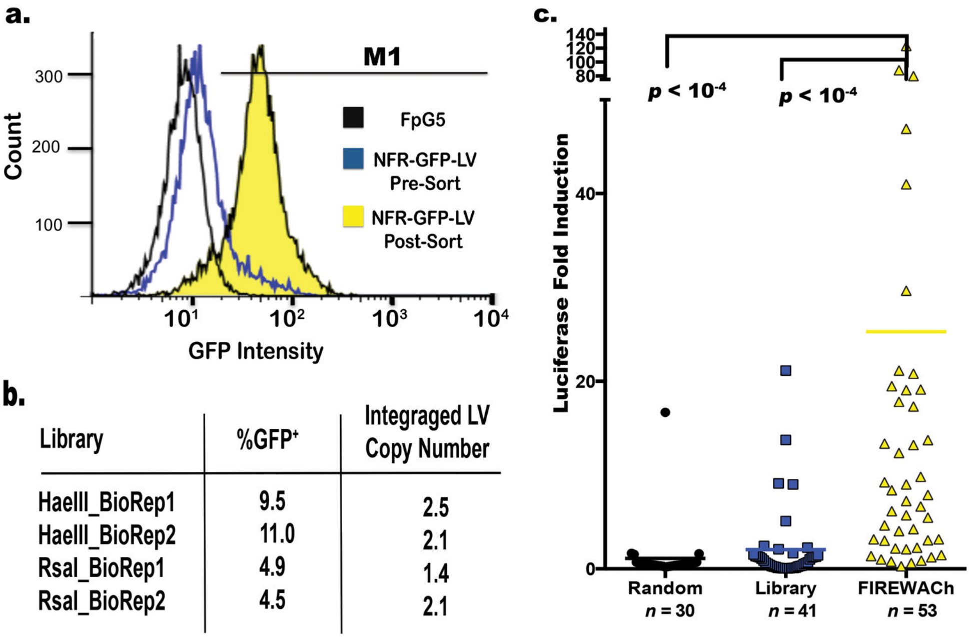


Figure 2. NFR-derived DNAs correspond to accessible chromatin regions located throughout the genome

(a) Carpet plots depicting the correspondence of *in silico*-generated genomic DNA fragments (*n* = 61,844) or Input Library NFR-DNAs (*n* = 84,241) with DNaseI hypersensitive sites (HS) in ESC chromatin[9]. The DNAs in each dataset were ranked according to the expression level of their associated gene (s) in ESCs (bar to the right of each panel; red, high expression, green, low expression). The presence of a DNaseI HS site (black) was assessed for a region corresponding to the genomic interval ±1 kb of the center (green vertical line) of the DNA fragments within Library NFR- or in silico-generated random DNA fragments. **(b)** Venn diagram examining the relatedness of genomic regions present in the HaeIII- and RsaI-NFR DNA libraries. The total number of elements in each library is indicated at the top of each circle. **(c).** Genomic distribution of HaeIII- or RsaI-NFR-input library DNA populations relative to annotated transcription start sites (TSS, black arrow) determined using the GREAT analysis tool.



Library	%GFP$^+$	Integraged LV Copy Number
HaeIII_BioRep1	9.5	2.5
HaeIII_BioRep2	11.0	2.1
RsaI_BioRep1	4.9	1.4
RsaI_BioRep2	4.5	2.1

Figure 3. NFR-GFP-LVs detect active CRMs

(a) Histogram depicting relative GFP expression profiles for FpG5-transduced ESCs (black line), and NFR-GFP-LV library-transduced ESCs before- (blue line), or after- FACS purification (yellow). The gate for GFP expression is set as depicted (M1). **(b)** The percentage GFP$^+$ cells, determined using quantitative flow cytometry and FloJo software, for ESCs transduced with each of the indicated NFR-GFP-LV libraries. BioRep1 and BioRep2 are Biological replicate samples resulting from two independent transductions for each of the NFR-GFP-LV libraries (See Supplementary Fig. 2). Integrated LV copy number represents the average number of transgenes/ cell for each GFP$^+$ transduced cell population, *n* = 2. (Supplementary Note I) **(c)** Luciferase reporter plasmids harboring Random genomic DNAs, or DNAs recovered from the Input NFR-GFP- LV library or FACs-sorted GFP$^+$ cells were transfected into ESCs and assayed for luciferase activity. Each plasmid was tested in duplicate in independent experiments. Data points show the mean; *p* values were calculated using unpaired t-test (also see Supplementary Fig. 6).

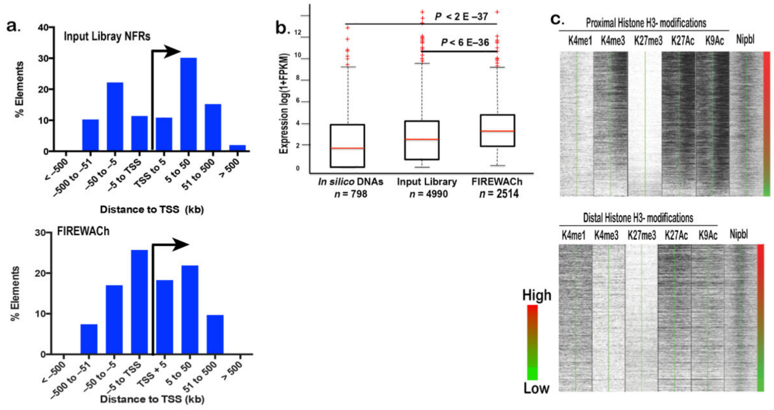


Figure 4. FIREWACh Elements are associated with expressed genes and correlate with chromatin marks of active promoters and enhancers

(a) Distribution of NFR-derived DNAs with respect to annotated TSSs. These datasets each represent the combined populations of HaeIII-and RsaI DNAs for Library (**top)** and FIREWACh elements (**bottom). (b)** Boxplot of the expression status of genes associated with the promoter-proximal DNAs (i.e 2 kb+/− of an annotated TSS) in each dataset. Median values of the distribution are denoted by a red line; whiskers show tails of the distribution and outliers are marked by red crosses. *P*-values were determined by nonparametric Kruskal-Wallis test. **(c)** Carpet plots assessing genomic loci corresponding to Promoter-proximal FIREWACh DNAs **(top panel)** or Distal FIREWACh DNAs **(bottom panel)** for the presence of histone modifications H3K4me4, H3K27ac, H3K27me3, or cohesin complex protein Nipbl. Each row corresponds to 1kb +/− from the center (green line) of a single FIREWACh DNA. Read density for the histone modification or feature indicated on the top of the column is depicted in black. FIREWACh DNAs were ranked according to the expression level of their associated gene(s) in ESCs (bar to the right of the panel; red, high expression, green, low expression).

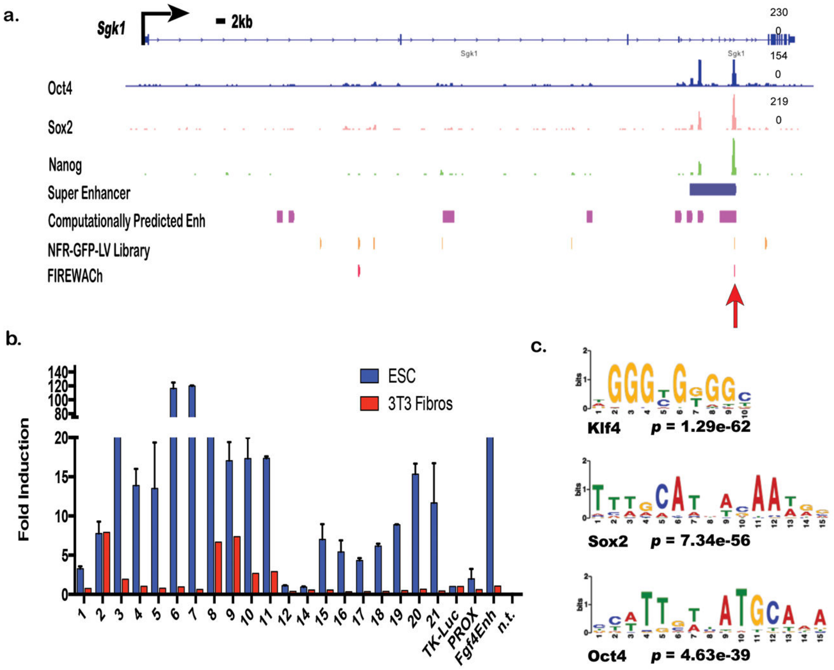


Figure 5. Distal FIREWACh DNAs correlate with previously predicted ESC enhancers and act as cell-specific enhancers that can be used for TF prediction

(a) Integrated Genome View screenshot of the genomic regions surrounding the *Sgk1* gene associated with distal FIREWACh DNAs and 'Super Enhancers[23], computationally predicted ESC enhancers[24], or a functionally validated enhancer (red arrow)[12]. The TSS is depicted by the black arrow at the top of the panel. Scale bar= 2 kb. **(b)** Testing enhancer function of distal FIREWACh DNAs. 20 distal FIREWACh DNAs were each inserted 2 kb downstream of the TK promoter within the pGL3TK luciferase reporter plasmid and tested in duplicate for their ability to activate transcription from this distal position after transfection in ESCs (blue bars) or 3T3 mouse fibroblasts (red bars). Shown is the average fold induction for each test plasmid compared to pGL3TK. Error bars= standard deviation. Fgf4Enh= positive control plasmid harboring the ESC-specific *fgf4* enhancer. TK-luc= basal reporter plasmid, PROX= proximal FIREWACh element control cloned at the distal location in pGL3TK. **(c)** Representative motifs for ESC TFs discovered after analysis of distal FIREWACh DNAs. A comparison of the top eight motifs discovered in the analysis of Distal FIREWACh Elements and Distal Elements of the Input library are presented in Supplementary Figure 10; the complete list of enriched motifs for these datasets, their associated p values and PWMs are presented in Supplementary Table 7. P-values were calculated based on the multi-hypergeometric distribution and corrected for multiple hypothesis testing.