Cell

Resource

Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel

Waseem Akhtar,[1,2,5] Johann de Jong,[3,5] Alexey V. Pindyurin,[2,5] Ludo Pagie,[2] Wouter Meuleman,[2,4,6] Jeroen de Ridder,[4] Anton Berns,[1] Lodewyk F.A. Wessels,[3,4,*] Maarten van Lohuizen,[1,*] and Bas van Steensel[2,*]

[1]Division of Molecular Genetics
[2]Division of Gene Regulation
[3]Division of Molecular Carcinogenesis
The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands
[4]Delft Bioinformatics Lab, Delft University of Technology, 2628 CD Delft, the Netherlands
[5]These authors contributed equally to this work
[6]Present address: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: l.wessels@nki.nl (L.F.A.W.), m.v.lohuizen@nki.nl (M.v.L.), b.v.steensel@nki.nl (B.v.S.)
http://dx.doi.org/10.1016/j.cell.2013.07.018

SUMMARY

Reporter genes integrated into the genome are a powerful tool to reveal effects of regulatory elements and local chromatin context on gene expression. However, so far such reporter assays have been of low throughput. Here, we describe a multiplexing approach for the parallel monitoring of transcriptional activity of thousands of randomly integrated reporters. More than 27,000 distinct reporter integrations in mouse embryonic stem cells, obtained with two different promoters, show ~1,000-fold variation in expression levels. Data analysis indicates that lamina-associated domains act as attenuators of transcription, likely by reducing access of transcription factors to binding sites. Furthermore, chromatin compaction is predictive of reporter activity. We also found evidence for crosstalk between neighboring genes and estimate that enhancers can influence gene expression on average over ~20 kb. The multiplexed reporter assay is highly flexible in design and can be modified to query a wide range of aspects of gene regulation.

INTRODUCTION

Control of gene expression in eukaryotes is a complex process regulated at multiple levels, such as the local action of enhancers and other regulatory DNA elements, compartmentalization of the genome into various types of chromatin domains, and spatial positioning of genes within the nucleus (Montavon and Duboule, 2012; Bickmore and van Steensel, 2013). One powerful traditional approach to study the influence of the local environment on gene expression involves the use of a reporter transgene integrated in the genome as a sensor. Activity of such an integrated reporter (IR) depends on its genomic location, which is known as "position effect" (Dobzhansky, 1936). This phenomenon has been exploited extensively to deduce causal relationships in the interplay among DNA sequence, chromatin context, and gene activity. For example, detailed analysis of position effects in yeast and *Drosophila* have contributed to a thorough understanding of heterochromatin (Grewal and Jia, 2007; Girton and Johansen, 2008), and IRs have also been used widely as "enhancer traps" to identify regulatory elements that promote transcription (Weber et al., 1984; Korzh, 2007; Ruf et al., 2011).

To study position effects, reporter genes can be either targeted to selected genomic loci or inserted at random positions. Random integration is achieved by stable transfection or transposon- or virus-based delivery. Even though in the latter approach plenty of random IRs can be obtained at once, the bottleneck is the establishment of clonal lines each harboring a single reporter, followed by the mapping of each integration site. The largest systematic reporter integration studies have yielded dozens to hundreds of characterized clonal lines (Sundaresan et al., 1995; Gierman et al., 2007; Babenko et al., 2010; Ruf et al., 2011; Chen et al., 2013), but these studies were extremely laborious. Furthermore, studies with IRs so far have required the transgene to be expressed at least to some degree, which is necessary to identify integration events. As a consequence, the results may suffer from biases that favor genomic regions that promote gene expression, whereas repressive loci are missed.

Here, we combined the traditional transgene reporter assay with random barcoding technology (Gerlach et al., 2010; Gerrits et al., 2010) and high-throughput sequencing to develop a method, termed Thousands of Reporters Integrated in Parallel (TRIP), that is designed to study position effects genome-wide, without the need to isolate clonal cell lines. We demonstrate the utility of this approach by the analysis of the activity of two different promoters integrated at >27,000 locations (in total) throughout the genome of mouse embryonic stem (mES) cells. Because of

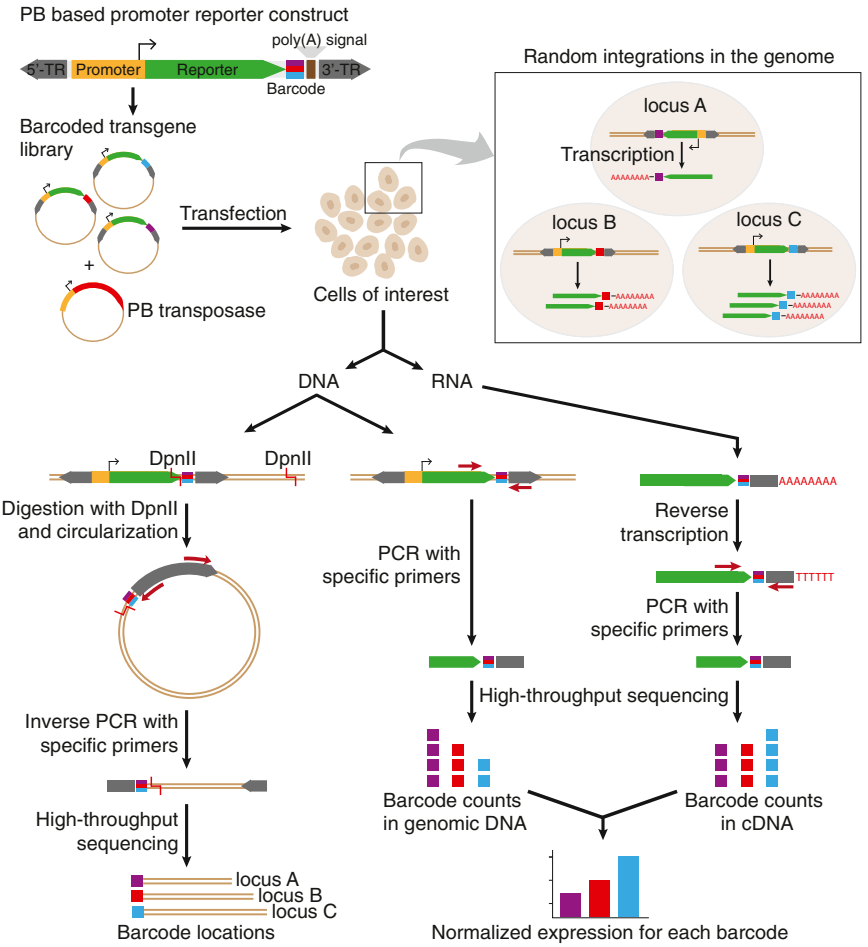


Figure 1. Overview of TRIP

A library of transcription reporters containing short random (16 bp) barcode sequences upstream of the polyadenylation signal is integrated randomly in the genome of cells of interest using piggyBac (PB) transposition. The locations of the IRs are determined by inverse PCR followed by high-throughput sequencing. The expression level of each IR is measured in a pool of cells by high-throughput sequencing of the barcodes in cDNA. These cDNA counts are normalized to the corresponding counts from the genomic DNA. See also Figure S1.

the flexible design of the reporter vector, TRIP is a generally applicable technique to study many facets of gene regulation.

RESULTS

Principle of TRIP

TRIP is based on a large set of reporter genes, which are all identical except for a short random nucleotide "barcode" inserted in the 3′ UTR (Figure 1). These barcodes serve as unique tags used to track each reporter independently. Using a transposable element vector, the reporters are randomly integrated into the genomes of a pool of cells. This pool is then expanded, and the integration sites are identified together with the barcodes by high-throughput sequencing. Next, the expression level of each IR is determined by counting the occurrence of each barcode in mRNA isolated from the cell pool and normalizing these counts to the corresponding barcode representation in the genomic DNA. Combining the mapping and the expression information yields expression variation as a function of genomic position, without the need to derive a clonal cell line for each integration.

Experimental Design

As a proof of concept, we applied TRIP to study how the behavior of two active promoters depends on genomic context in mES cells. We chose the mouse phosphoglycerate kinase (mPGK) promoter, which is a housekeeping promoter containing all the *cis*-regulatory elements required for its full activity (McBurney et al., 1991) and the tet-Off promoter, which offers the advantage that its activity can be tuned by changing the concentration of doxycycline (Dox) in the medium (Gossen et al., 1995). For integration of barcoded reporters, we used the piggyBac (PB) transposition system because of its high efficiency (Cadiñanos and Bradley, 2007) and the relatively small sizes of the essential terminal repeats (TRs) (Meir et al., 2011).

We generated a PB transposon plasmid library of reporters for each promoter driving the enhanced GFP (eGFP) transcription unit with one of hundreds of thousands of random DNA barcodes (16 bp each) between the reporter and polyadenylation signal (Figure 1). This library was transfected into mES cells together with a plasmid expressing PB transposase to randomly integrate the reporters throughout the genome. The transfected cells were cultured for 7 days before about 1,000 cells were subcultured to generate a pool of cells (a TRIP pool). We generated six TRIP pools with the mPGK promoter construct (mPGK-A to mPGK-F) and four pools with the tet-Off promoter construct (tet-Off-A to tet-Off-D). Further, each TRIP pool was split into two halves, and each half was separately cultured for an additional week and analyzed independently (Figure S1A available online). These split pools served as technical replicates.

Mapping of Reporter Integration Sites

By quantitative PCR, we estimated that cells in the pools harbor on average 23 ± 3 IRs per cell (mean ± SD across all pools). We mapped the IR integration sites and linked them to the corresponding barcodes by an inverse PCR method coupled to paired-end high-throughput sequencing (Figure 1). Our mapping of the locations of barcodes was highly accurate, because more than 98% of barcodes mapped independently in two technical replicates were located at the same base position in the genome

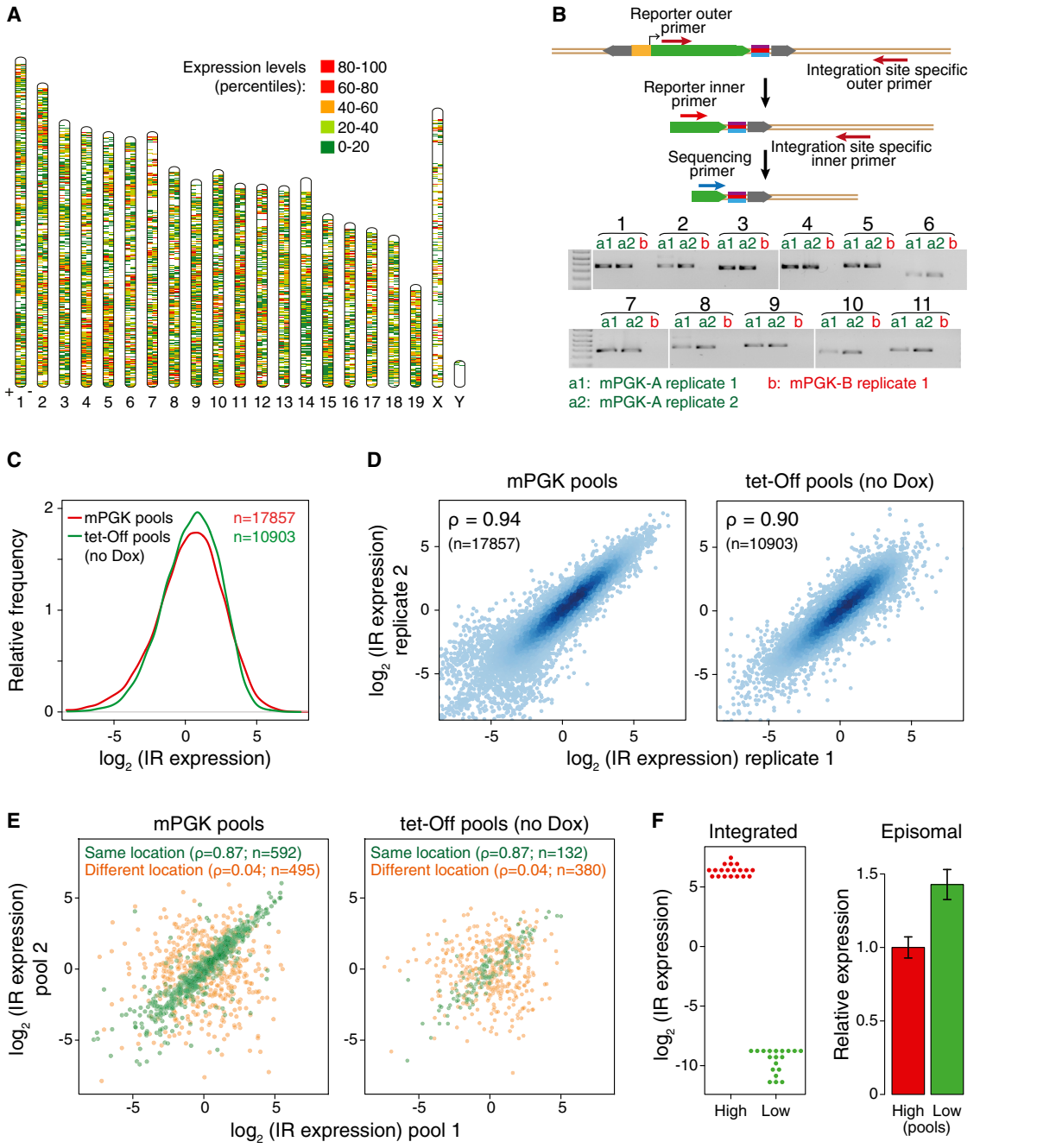


Figure 2. TRIP Works Robustly and Reproducibly

(A) Positions of mapped mPGK IRs along all chromosomes. Each IR is represented as a tick on one of the strands (depending on the orientation of integration), colored by expression level. The mapped IR density on X and Y is lower because these chromosomes occur as a single copy (male mES cells were used) and are relatively repeat dense.

(B) Scheme (top) and results (bottom) of the PCR strategy to validate the locations and barcodes of 11 randomly selected IRs in one of the TRIP pools (mPGK-A). PCR was done with integration site-specific and IR-specific nested primers (see Table S3 for details) on DNA from the two replicates of this TRIP pool (a1 and a2) and a different TRIP pool (b) as a control. Sequence of the barcodes was confirmed in each instance by Sanger sequencing (data not shown).

(C) Distribution of expression values for the entire sets of mPGK and tet-Off (no Dox) IRs.

(D) Correlation of IR expression levels between two technical replicates for mPGK (left) and tet-Off (no Dox) (right) pools. ρ is Spearman's rank correlation coefficient.

(E) Correlation between the expression levels of barcodes that were coincidentally present in two different mPGK pools (left) or tet-Off pools (right). Identical barcodes mapped to the same location in the two distinct pools are shown in green; identical barcodes integrated at different genomic locations are shown in orange.

(legend continued on next page)

(Table S1). After merging of the technical replicates and application of stringent data quality filters, each cell pool yielded roughly 2,300–3,300 mapped IRs (Figures S1B and S2A; Table S1). In total, we unequivocally mapped the locations of 17,857 and 10,903 barcodes in six mPGK and four tet-Off pools, respectively (Figure 2A; Data S1 and S2). We checked the accuracy of the mapping by integration-specific PCR and Sanger sequencing. For all 11 IRs tested, the mapping and the associated barcode were correct, and these integrations were absent in a different TRIP pool (Figure 2B).

PB is known to have a preference for integration near transcription start sites (Huang et al., 2010). We estimate this bias to be ~3-fold; however, the vast majority of integrations occurs in other areas of the genome (Figures S2B and S2C; see also below).

IR Expression Strongly Depends on Integration Site

The expression of the set of mapped barcodes was determined by high-throughput sequencing of the barcodes in the cDNA from corresponding pools. Strikingly, we observed an ~1,000-fold range in expression of the *same* reporter integrated at *distinct* genomic locations (Figure 2C). This large variation is not due to experimental noise, because expression levels of technical replicates were highly correlated (Spearman's ρ = [0.90–0.94]; Figure 2D).

We considered that some barcodes could spuriously contain binding motifs of transcription factors, microRNAs or RNA-binding proteins and thereby affect their own expression. We investigated this using three independent approaches. First, our TRIP pools were made from a single large pool of cells transfected with the reporter library, giving rise to situations where the same barcode sequence was present in different pools either at the same location (essentially the same clonal cell line grown in different pools) or at different locations (the constructs with the same barcode sequences but integrated independently in different cells). Comparison of such barcode pairs showed that the barcodes with identical sequences at the same location were highly correlated (Spearman's ρ = [0.85–0.89]), whereas the sets of identical barcode sequences but integrated at different locations showed no correlation (Figure 2E). Thus, genomic location has a much stronger overall effect on IR expression than barcode sequence.

Second, we searched for any motifs in our barcode sequences that may account for variation in IR expression. Employing the MatrixREDUCE algorithm (Foat et al., 2006), we identified a few motifs that significantly correlate with barcode expression levels of IRs; however, they had an almost negligible contribution to expression. MatrixREDUCE estimates that <10% of the total expression variance can be explained by sequence motifs present in the barcodes (Figure S2D).

Third, we chose 19 barcodes that showed extremely high and 19 barcodes that showed extremely low expression in IRs (Figure 2F). These barcodes were reinserted into the mPGK promoter vector and transiently transfected as two pools of "low" and "high" reporters, in the absence of transposase. Under these conditions, these reporters are not integrated in the genome, allowing us to directly estimate the effects of sequence differences between barcodes on reporter expression. Quantitation of the expression showed no significantly elevated expression of the "high" pool compared to the "low" pool (Figure 2F). We conclude that the effects of the barcode sequences are of such low magnitude that they do not compromise our studies of position effects.

Nonrandom Patterns of IR Expression

Next, we investigated the positional variation in IR expression in detail. We first focused on the mPGK IRs, because this data set is larger than that of the tet-Off IRs. The mPGK IRs have a median interinsertional distance of 65 kb. Besides the somewhat nonhomogeneous spacing of integration sites (a known feature of PB transposition; Huang et al., 2010), we noticed that IRs tend to cluster according to their expression level, with alternating patches of highly and lowly expressed IRs (Figure 3A). Indeed, genome-wide we found a significant autocorrelation of IR expression levels extending over many neighboring IRs (Figure 3B). Thus, IRs landing in the same areas of the genome tend to have similar levels of expression.

To further characterize this domain-like expression pattern, we trained a hidden Markov model (HMM) on the mPGK IR data set to divide the genome into two states, transcriptionally permissive and nonpermissive (Figure S3). This yielded domains with a median size of 1.23 Mb (Figures 3A and 3C), with a striking banding pattern along the chromosomes (Figure S3A). Various approaches to inferring an HMM gave highly similar results (Figures S3B–S3E). In contrast, HMM fitting after random permutation of the expression values (but keeping the IR positions unaltered) resulted in domains of much smaller size (median 0.18 Mb). Therefore, the pattern of large domains cannot be explained by random expression patterns among the IRs. Furthermore, the tet-Off IR expression values were generally high in the mPGK permissive domains and low in the mPGK nonpermissive domains (Figures 3A and 3D), demonstrating that this pattern is overall consistent between the two different reporter constructs.

IR Expression Patterns Reflect Chromatin Domain Organization

We compared the IR expression domain pattern to various chromatin features known to form large domains (Figures 3A and 3E). Interestingly, this revealed that nonpermissive IR domains significantly overlap with lamina-associated domains (LADs), late-replicating domains, and to a lesser extent with regions marked by the histone modification H3K9me2 (Hiratani et al., 2008; Peric-Hupkes et al., 2010; Lienert et al., 2011). These three domain types are known to coincide substantially with one

(F) Barcodes do not affect the reporter expression after transient transfection. Barcodes from mPGK IRs with very high (n = 19; red color dots) or low (n = 19; green color dots) expression (left) were recloned into the reporter plasmid. Plasmids for each group of barcodes were mixed together in equal proportions and transiently transfected. Their pooled expression levels were measured by RT-qPCR of eGFP (right). Error bars (right) represent SD from three transfection experiments.

See also Figure S2.

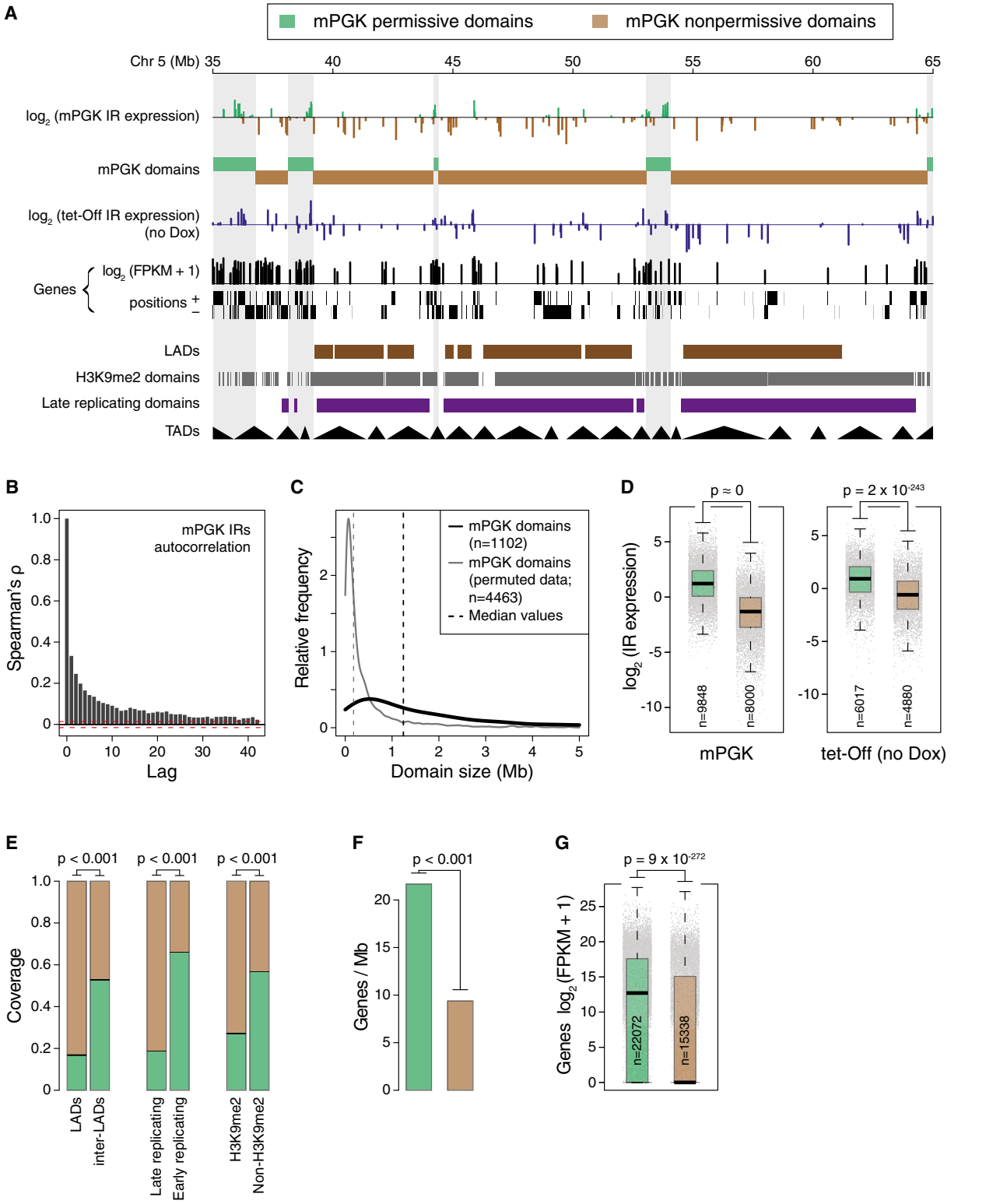


Figure 3. Nonrandom IR Expression Pattern Reflect Chromatin Domain Organization

(A) Segment of chromosome 5 showing expression levels for mPGK and tet-Off (no Dox) IRs, together with tracks showing a two-state HMM of mPGK IR activity (mPGK domains); expression and positions of endogenous genes; and the positions of various types of known chromatin domains (Hiratani et al., 2008; Peric-Hupkes et al., 2010; Lienert et al., 2011; Dixon et al., 2012).

(legend continued on next page)

another, and harbor mostly inactive endogenous genes (Bickmore and van Steensel, 2013). Conversely, permissive IR domains tend to coincide with gene-dense and transcriptionally active segments of the genome (Figures 3A, 3F, and 3G). We found no substantial overlap between the borders of topologically associated domains (TADs) (Dixon et al., 2012; Nora et al., 2012) and borders of IR domains (Figure 3A; data not shown). Although we note that the accuracy of mPGK permissive/nonpermissive HMM domain definitions is compromised by the irregular spacing of IRs, these results nevertheless indicate that IR expression patterns correspond to some known aspects of large-scale domain organization of chromatin.

Attenuated Transcription in LADs

LADs are of particular interest because they are confined at the nuclear periphery and harbor mostly genes that are expressed at very low levels (Guelen et al., 2008; Peric-Hupkes et al., 2010). The IRs in LADs show on average a 5- to 6-fold lower expression compared to IRs in inter-LADs (Figures 4A and 4B). The average profile of IR expression across the borders of LADs shows a sharp transition that is again highly similar to that of endogenous genes (Figure 4C). Thus, LAD positions are predictive of reduced IR expression.

Because LADs and IR expression are both strongly correlated with local gene density, gene activity, H3K9me2 domains, and replication timing (Figures 3A and 3E–3G), these parameters could form confounding factors in linking IR expression to LADs. To resolve this issue, we conducted a partial correlation analysis, taking into account all of these factors. The partial correlation is a conservative approach, because all joint variance between the variables is removed. However, even using this conservative approach, it can be seen that the association between LADs and reduced IR expression cannot be fully explained by the other variables (Figure 4D), suggesting a role for LADs in repression of transcription.

We reasoned that LADs could reduce gene expression in at least two distinct ways. First, LAD chromatin could pose a threshold to gene activation that may be overcome only if a promoter reaches a certain minimum strength (which depends on the types of activators and their occupancy). Second, LAD chromatin could act as an attenuator that reduces all transcriptional activity by a roughly constant factor, without a threshold effect and independent of intrinsic promoter strength. To discriminate between these models, we took advantage of the tet-Off IRs. Here, the concentration of Dox controls the occupancy of the promoter by its activator and, as a result, the promoter strength. To test whether the efficacy of LAD repression is dependent on promoter strength, we treated cell pools carrying the tet-Off IRs with four different concentrations of Dox and measured the expression level of all barcodes throughout the genome (Figure S1A).

Quantitative PCR confirmed that the overall expression level of the IRs depended on the Dox level, over an ~50-fold range (Figure S4). However, individual IRs showed substantial differences in induction strengths (Figure 4E). Grouping the IRs by LAD/inter-LAD location revealed that, for all four Dox concentrations, the expression levels of IRs within LADs were systematically lower compared to outside LADs (Figure 4F). Even at the highest induction ([Dox] = 0), the expression level of IRs in LADs was more than 4-fold lower than in inter-LADs. Thus, LADs appear to act primarily as attenuators, although we cannot rule out a modest thresholding effect.

LAD Chromatin Reduces DNA Binding of Activators

We wondered how LADs might cause such a consistent attenuation of gene activity. One possibility is that LAD chromatin is less permissive to the binding of activating factors to their cognate binding motifs. To test this, we used previously published chromatin immunoprecipitation (ChIP) data sets in mES cells (Chen et al., 2008; Marson et al., 2008; Handoko et al., 2011; Li et al., 2012) to analyze the binding of various factors to their motifs inside and outside LADs (Figure 4G). Remarkably, occupancies of all six factors at their binding motifs were consistently lower inside LADs, by 2- to 4-fold. This inefficient binding of transcription factors to their motifs inside LADs may explain in part the reduced expression levels of IRs and endogenous genes that are embedded in LADs.

IR Expression Is Related to Local Chromatin Conformation

A popular model is that gene activity is controlled by the degree of chromatin compaction (Li and Reinberg, 2011). For endogenous genes, this model is, however, difficult to test, because compaction may be the consequence rather than the cause of gene activity. In contrast, with IRs, one can ask whether the local chromatin compaction state prior to integration has predictive value for IR expression levels. A quantitative way to describe chromatin compaction is the rate of decay in contact probability between two loci with increasing genomic distance. This decay

(B) Autocorrelation function showing the similarity (Spearman's ρ) between expression levels of neighboring IRs (lag = nth neighbor, with $0 \leq n \leq 40$). Red dotted lines indicate significance threshold ($p < 0.05$).

(C) Distribution of mPGK HMM domain sizes compared to those obtained after random permutation of mPGK IR expression values.

(D) Distribution of expression levels of mPGK (left) and tet-Off (no Dox) (right) IRs in mPGK permissive and nonpermissive domains. Gray dots show values for individual IRs, colored boxes indicate interquartile range, horizontal line inside each box shows median expression, and the ends of the whiskers extend to the most extreme data points no further than 1.5 times the interquartile range from the box (same applies for G). The p values were determined by Wilcoxon rank sum test. Color legend in (A) also applies to (D)–(G).

(E) Fraction of overlap of known epigenomic domains with mPGK permissive and nonpermissive domains. The p values were determined by circular permutation (n = 1,000) of the mPGK domains, testing the fold difference of the mPGK nonpermissive domain fractions.

(F) Gene density in mPGK permissive and nonpermissive domains. The p value was determined as in (E).

(G) Distribution of expression levels of endogenous genes in mPGK permissive and nonpermissive domains, plotted as in (D). FPKM, fragments per kilobase of exon per million fragments mapped. The p value was determined by Wilcoxon rank sum test.

See also Figure S3.

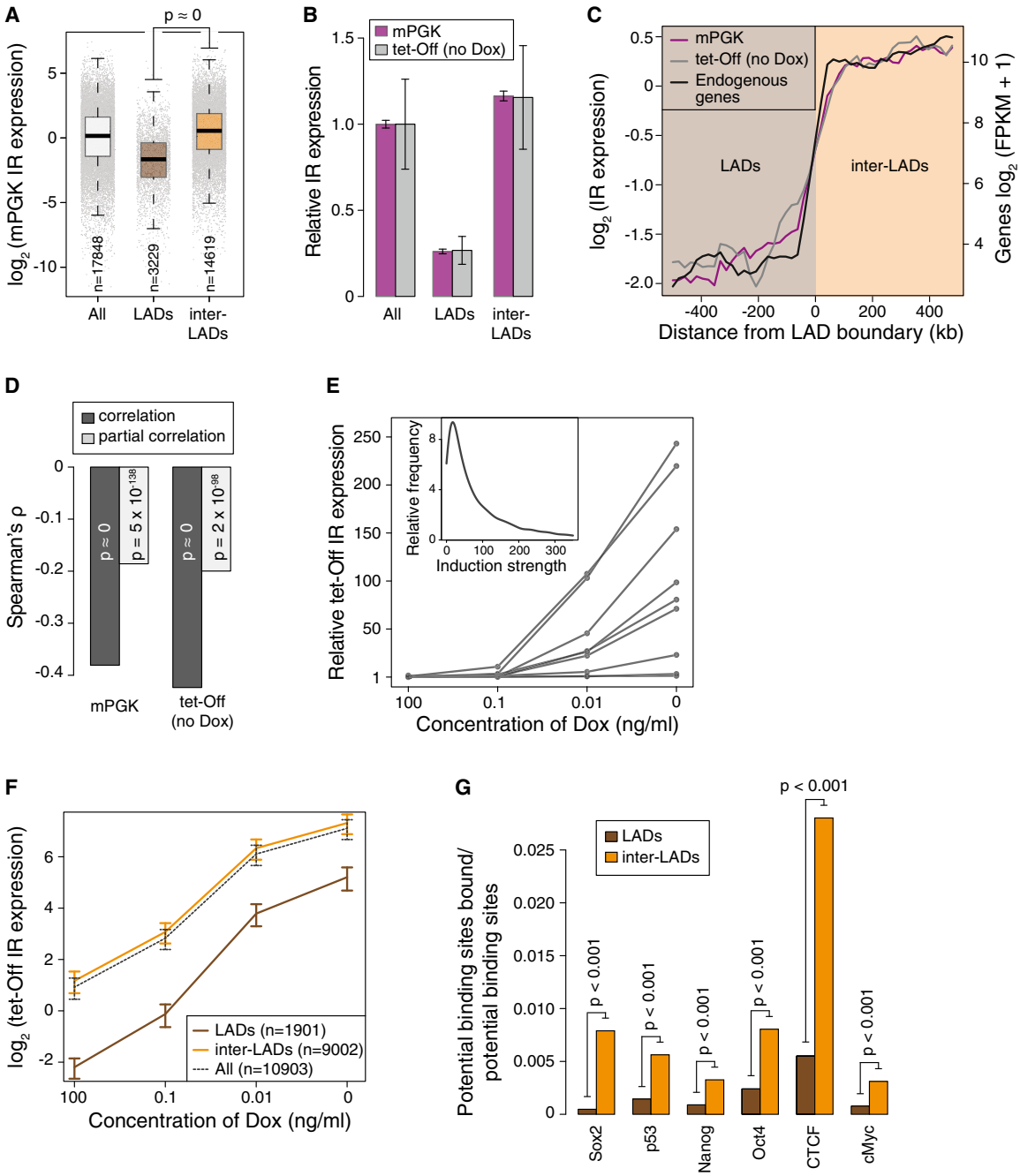


Figure 4. LADs Act as Transcription Attenuators

(A) Expression level distributions for all mPGK IRs and those in LADs and inter-LADs, plotted as in Figure 3D. The p value was determined by Wilcoxon rank sum test.

(B) Biological reproducibility of relative expression of IRs, separated by LAD or inter-LAD location. Error bars represent SEM of median expression values across TRIP pools (i.e., the dispersion around the mean of six pool medians for mPGK and four pool medians for tet-Off IRs). Differences between LADs and inter-LADs are statistically significant ($p = 8 \times 10^{-7}$ and 2.8×10^{-2} for mPGK and tet-Off IRs, respectively; two-sided t test).

(C) Expression levels of IRs and endogenous genes around LAD borders. Lines show average values across 20 kb bins (50 bins in total).

(D) Correlation (dark-gray bars) of Lamin B1 binding with the expression of mPGK and tet-Off (no Dox) IRs, compared to partial correlation (light-gray bars) given H3K9me2, replication timing, and gene proximity.

(E) Expression levels of nine randomly selected tet-Off IRs at different concentrations of Dox. Inset shows the distribution of induction strengths (see Extended Experimental Procedures) in the whole data set.

(F) tet-Off IR expression levels in LADs and inter-LADs depending on the Dox concentration. Error bars represent SEM of mean expression values across TRIP pools (i.e., the dispersion around the mean of six pool means for mPGK and four pool means for tet-Off IRs).

(legend continued on next page)

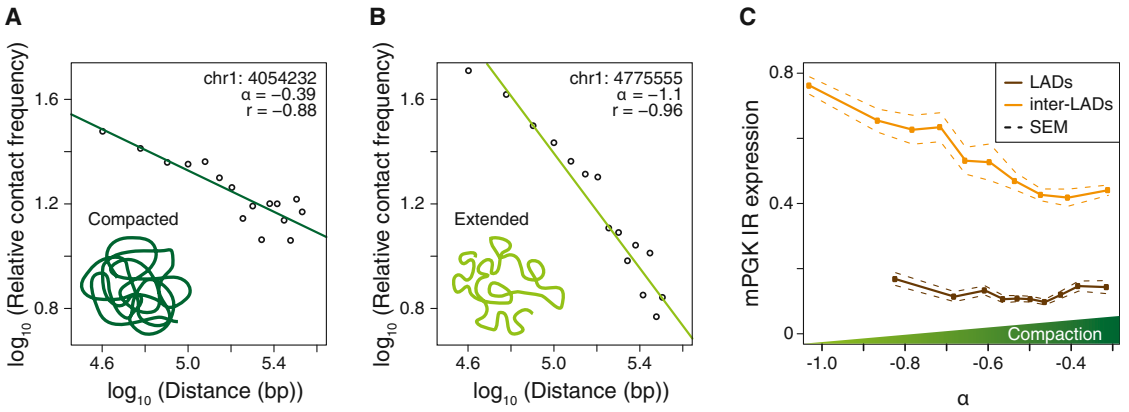


Figure 5. Local Chromatin Conformation Partially Predicts IR Expression Levels

(A and B) Examples of the dependency of relative contact frequency (as determined by Hi-C; Dixon et al., 2012) on genomic distance in 400 kb windows around two mPGK IRs. Note the difference in the slope (α) of the fitted line, which reflects a difference in local compaction. r denotes Pearson correlation coefficient.

(C) Expression of IRs as a function of α, for LADs and inter-LADs. The solid lines refer to the mean of median expression values across six mPGK pools, for ten equally sized bins; the dotted lines represent error bands (±SEM), computed in the same way as the error bars in Figure 4B.

See also Figure S5.

function can be inferred from Hi-C data and approximated by a power law with a scaling exponent α (Lieberman-Aiden et al., 2009; Sexton et al., 2012). Low (i.e., more negative) α values correspond to a steep decay function, which reflects decondensed chromatin, whereas α values close to 0 correspond to a flat decay function, reflecting a more compacted chromatin configuration (Figures 5A and 5B). Using published Hi-C data for mES cells (Dixon et al., 2012), we found that for most integration sites the local decay function fitted a power law reasonably well if a window size of 400 kb was used (Figures 5A, 5B, and S5A), with highly reproducible α values between replicate Hi-C data sets (Figure S5B). The α values of integration sites ranged from -1.0 to -0.31 (5th and 95th percentile; Figure S5C). We then investigated the relationship between IR expression and the local α value.

Strikingly, in integration sites that do not overlap with LADs, we found a significant inverse correlation (Spearman's $\rho = -0.80$; $p < 2.2 \times 10^{-16}$) between local α values and IR expression (Figure 5C). This result suggests that the local chromatin configuration contributes to the regulation of IR activity, with IRs being more active in more decompacted regions. In contrast, integration sites that overlap with LADs have a very narrow distribution of α values that is centered around -0.5 (Figure S5C), suggesting that they tend to share a particular chromatin configuration. The IR expression levels in LADs are another 2- to 3-fold lower compared to inter-LAD IRs with similar α values (Figure 5C). Together, these results indicate that the local chromatin compaction state is partially predictive for IR expression levels, but chromatin compaction alone (as measured by Hi-C) cannot fully explain the difference in IR expression between LADs and inter-LADs.

Proximity Effects of Active Genes and Enhancers

Although LADs and chromatin compaction explain part of the variation in IR expression, much of the 1,000-fold range in IR activity remained unaccounted for. This prompted us to study the possible contribution of smaller elements in the genome. Previous correlative analyses of genome-wide expression data sets have suggested regulatory crosstalk between neighboring genes in mammals (Ebisuya et al., 2008; De et al., 2009). In line with these studies, we found that IRs proximal to genes are on average ~10-fold more active than those located far from any gene. This effect is similar in magnitude for IRs upstream and downstream of genes, decreases gradually with distance, but is still detectable at ~100–200 kb from genes. Splitting the data according to the expression level of the endogenous genes indicates that active genes contribute much more to this effect than inactive genes (Figure 6A).

The remarkably long distance over which IRs appear to be affected by neighboring active genes could have several explanations. One possibility is that active transcription units themselves promote the activity of neighboring transcription units, for example, because they are tethered to a "transcription factory" (Sutherland and Bickmore, 2009) and thereby promote recruitment of *cis*-linked genes into the same factory. Alternatively, active genes may be surrounded over a long-distance range by multiple enhancers, which could be responsible for the activation of IRs. Consistent with the latter model, we find that active enhancers —as identified by occupancy of H3K4me1, H3K27ac, and p300—are distributed around genes over an ~200 kb range (Figure 6B), which is in agreement with observations in human cells (Heintzman et al., 2009). To test whether these enhancers might stimulate expression of nearby

(G) Reduced binding site occupancy by six DNA-binding factors in LADs compared to inter-LADs. Bars show the fraction of cognate binding motifs for each factor that is occupied by this factor in mES cells according to ChIP-seq data (Chen et al., 2008; Marson et al., 2008; Handoko et al., 2011; Li et al., 2012). The p values were determined by circular permutation (n = 1,000) of LADs, testing the fold difference of the inter-LAD fraction and the LAD fraction.

See also Figure S4.

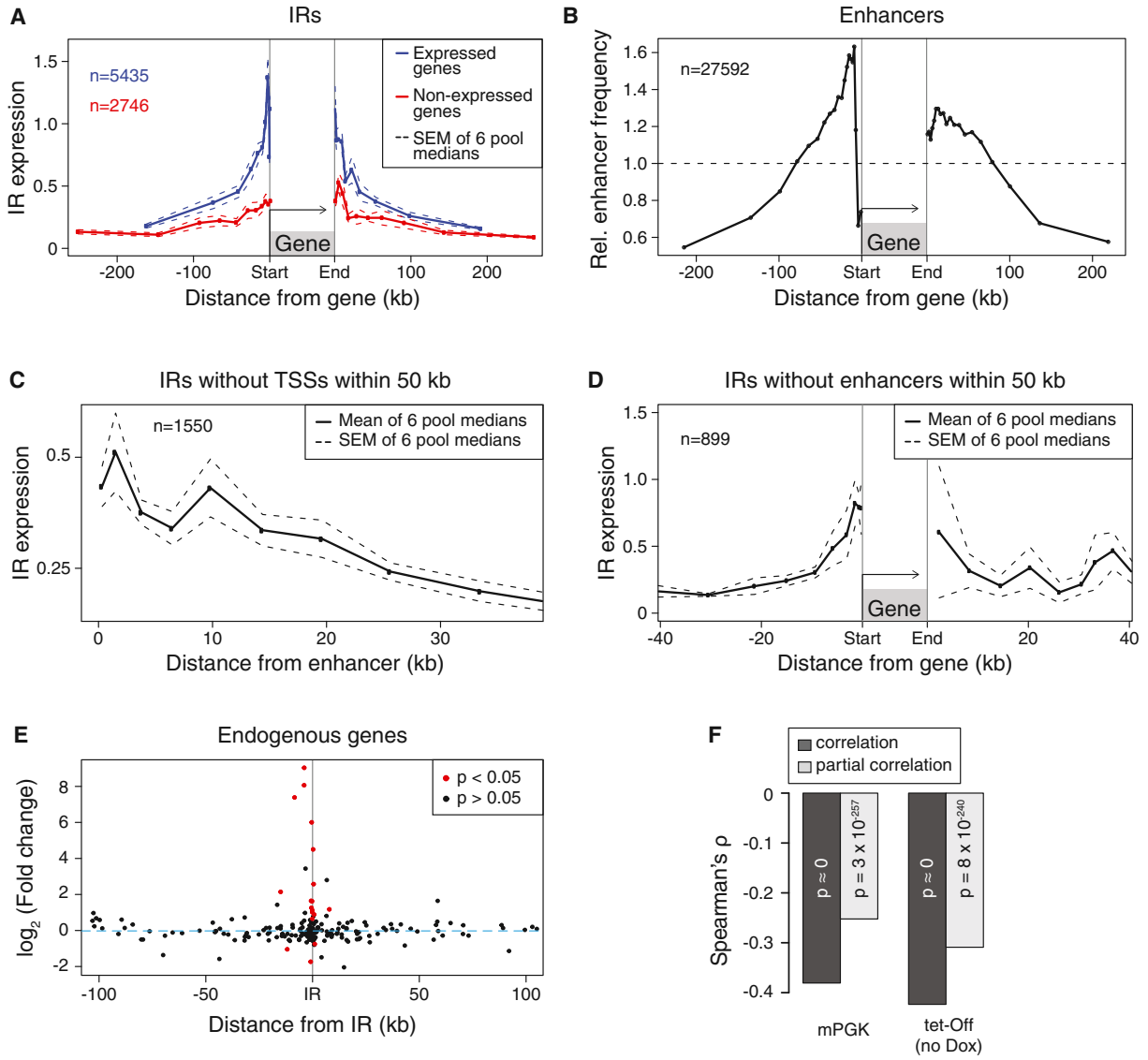


Figure 6. Proximity Effects of Genes and Enhancers

(A) Intergenic mPGK IR expression as a function of their distance from the nearest endogenous gene. The endogenous genes are divided into two categories: expressed (blue) and not detectably expressed (red). The solid lines show the mean of median expression values across six mPGK pools, for ten equally sized bins on each side of genes; the dotted lines represent error bands (±SEM), computed in the same way as the error bars in Figure 4B (same applies for C and D).

(B) Relative frequency of active intergenic enhancers (for definition, see Extended Experimental Procedures) around endogenous genes. Values above the dashed horizontal line imply the presence of more enhancers than expected by chance.

(C) Expression of intergenic mPGK IRs as a function of distance from the nearest active enhancer. To avoid confounding effects of neighboring genes, only enhancers >50 kb away from any endogenous transcription start site were considered.

(D) Expression of intergenic mPGK IRs, without an active enhancer within 50 kb, as a function of distance to the nearest gene.

(E) Change in the expression levels of nearest endogenous genes in 11 monoclonal mPGK cell lines as a result of reporter integration.

(F) Correlation (dark-gray bars) of Lamin B1 binding with the expression of mPGK and tet-Off (no Dox) IRs, compared to partial correlation (light-gray bars) given gene proximity and enhancer proximity.

See also Figure S6.

IRs, we plotted IR expression versus the distance to the nearest enhancer, while excluding IRs within 50 kb from genes in order to remove confounding effects of transcription units (Figure 6C). This revealed a significant correlation between enhancer proximity and IR expression, with the effect extending over ~20 kb. Similarly, plotting IR expression versus the distance to the nearest gene after removal of all IRs with an enhancer within 50 kb showed a significant residual effect of gene proximity, again over ~20 kb (Figure 6D). These data indicate that enhancers as well as transcription units individually promote the activity of IRs over a distance of ~20 kb. We propose that their collective action results into transcription-promoting regions

that cover on average ~100–200 kb on each side of active genes.

We investigated whether IRs might reciprocally affect the expression of neighboring genes. For this purpose, we established a set of 11 clonal cell lines that each carry 11–131 mPGK IRs of which the genomic location could be mapped. We subjected each cell line to mRNA sequencing (RNA-seq) to determine the expression levels of the nearest flanking genes (Data S3 and S4). We focused our analysis on the 264 IRs that were intergenic. The expression levels of 178 of the 197 endogenous genes located within 100 kb from these IRs were not significantly altered, whereas 16 genes were significantly upregulated and 3 were significantly downregulated. Interestingly, all 19 misregulated genes reside within 20 kb distance from IRs (Figure 6E). However, only a minority (19/118) of the genes within this distance is significantly affected. Together, these data indicate that the transcription of one gene can affect the activity of some neighboring genes, and these effects are mostly limited to a range of ~20 kb.

Based on these results, we considered that the low expression levels of IRs in LADs may be explained by a lack of nearby enhancers and active genes. However, partial correlation analysis indicates a significant residual correlation when taking into account the local density of these features (Figure 6F), suggesting the presence of an active repressive mechanism inside LADs.

Histone Modification States and IR Expression

Finally, we investigated how IR expression is linked to the local histone modification state. We used published mES cell chromatin immunoprecipitation sequencing (ChIP-seq) data sets for 11 histone modifications as well as CTCF (Mikkelsen et al., 2007; Creyghton et al., 2010; Handoko et al., 2011; Hezroni et al., 2011; Stadler et al., 2011) to identify the 15 most prevalent combinations ("chromatin states") in mES cells (Figures S6A and S6B) by applying a classification algorithm that was previously reported (Ernst et al., 2011). H3K9me2 was not included because a matching ChIP-seq data set was not available. Between the 15 states, average IR expression varied over more than 10-fold (Figures S6C–S6F). For the mPGK IRs, highest expression was observed in the states (#2 and #3) enriched in H3K4me1 and H3K27ac, which are characteristic of enhancer regions. Lowest expression occurred in a highly prevalent state (#12) that lacks any of the mapped histone marks, and in a state (#15) marked by H3K9me3 and H4K20me3. State #8, which is enriched exclusively for H3K27me3, showed moderate IR expression levels. A similar expression pattern was observed for the tet-Off IRs except that the highest expression was detected in the bivalent state (#9). Except for two rare states of unclear biological relevance (#13 and #14), all states were covered by dozens or hundreds of IRs, providing sufficient statistical power to compare their expression distributions (Figures S6G and S6H).

DISCUSSION

Genome-wide Surveys of Position Effects by TRIP

We combined random reporter integration with barcoding and deep sequencing to develop TRIP, a method to measure position effects in a high-throughput mode. TRIP helps to establish causal relationships, because it directly tests the functional consequence of integration into a certain chromatin environment. At the same time, the thousands of IRs provide enough statistical power to infer general, genome-wide relationships. TRIP thus bridges a gap between reductionist mechanistic studies of single loci on the one hand, and descriptive genome-wide mapping approaches such as ChIP, DamID, and RNA sequencing (Southall and Brand, 2007; Hawkins et al., 2010; Furey, 2012) on the other hand. Because all IRs are identical (except for the short barcode) and can be custom designed, TRIP is more suited for the systematic decoding of regulatory mechanisms than genome-wide studies of endogenous gene expression, where every gene is different and cannot be easily manipulated.

Although PB integrations exhibit some preference for transcriptional start sites (TSSs) and genes, the thousands of integrations elsewhere provide sufficient statistical power to determine the correlation of IR expression with most genomic features. Naturally, for TRIP studies of rarer features (or combinations of features) it may be necessary to generate larger data sets in order to probe these features sufficiently frequently. Other delivery vehicles, e.g., Sleeping Beauty, which has a more random integration profile (Huang et al., 2010), could further reduce any bias issues.

The cells used in this study harbored about two dozen IRs on average. Because each barcode is unique, each IR could nevertheless be tracked individually. Although some IRs could potentially interrupt the genome sequence at critical sites, cells with such IRs would likely be lost during culture. We note that the 11 clonal lines with 11–131 IRs show highly similar RNA-seq profiles (pairwise genome-wide correlation coefficients 0.96–0.99; data not shown), suggesting that the IRs in the established cell pools rarely cause major changes in the genome-wide expression program. We cannot completely rule out interference between IRs in the same cell, e.g., because they compete for limiting amounts of certain transcription factors, but this seems unlikely because most transcription factors are sufficiently abundant to occupy thousands of sites in the genome (Kind and van Steensel, 2010).

Future Applications of TRIP

The design of TRIP vectors is highly flexible. The only essential components are the short PB TRs and a random barcode of 16–20 bp. A variety of sequence elements in many arrangements can be added to study the influence of chromatin context on a wide range of processes (Figure 7). In the present study, we placed the barcode in the 3′ UTR of the reporters as a transcriptional readout. This approach can also be used to study how chromatin context affects the regulatory activity of other elements such as enhancers, silencers, insulators, and synthetic transcription factor binding sites, alone or in combination. The barcode can also be put in other locations of a transcription unit; with only minor modifications in the experimental design, it will then be possible to explore links between chromatin context and pre-mRNA processing events, such as mRNA alternative splicing and polyadenylation.

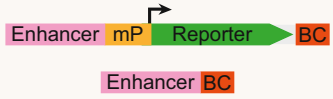
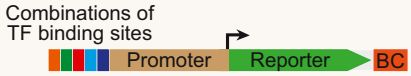
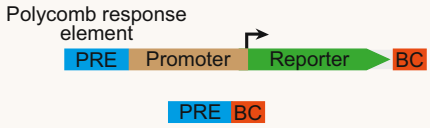
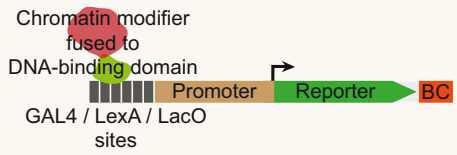
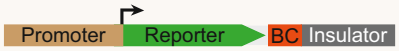
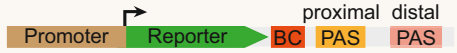
Process	Design of TRIP construct	Assays	Expected insights
Transcription		barcode-RNA-seq, ChIP/DamID	Effects of chromatin context on the activity of a promoter
		barcode-RNA-seq, ChIP/DamID	Enhancer activities in different cell types
		barcode-RNA-seq, ChIP/DamID	Effects of chromatin context on enhancer activity
		barcode-RNA-seq, ChIP/DamID	Interaction between different transcription factors (TFs) in varying epigenetic enviroments
Chromatin dynamics		barcode-RNA-seq, ChIP/DamID	Dynamics of establishment and maintenance of polycomb domains in different epigenomic contexts
		barcode-RNA-seq, ChIP/DamID	Behavior of a chromatin modifier in a variety of chromatin states
		barcode-RNA-seq, ChIP/DamID	The potential of (putative) insulator sites in different chromatin contexts
DNA methylation		barcode-RNA-seq, MeDIP, ChIP/DamID	How DNA methylation is established/maintained in different epigenomic enivronments and how it affects transcription
RNA stability		RNA labelling followed by barcode-RNA-seq	The connection between chromatin and RNA stability
RNA cleavage/ polyadenlyation		barcode-RNA-seq with alternative primers	The connection between chromatin and RNA polyadenilation
RNA alternative splicing		barcode RNA-seq with alternative primers	The connection between chromatin and RNA splicing

Figure 7. Potential Applications of TRIP

Barcodes (red boxes labeled "BC") can be combined in many configurations with reporter genes or regulatory elements to determine the effects of local chromatin context on a variety of molecular processes as indicated. PAS, polyadenylation signal.

Furthermore, the barcode may be placed outside of the transcribed region, for example, next to a promoter or enhancer. In this case, ChIP, DamID, and MeDIP methods (Vogel et al., 2007; Mohn et al., 2009; Furey, 2012) could be used to investigate how the binding of specific transcription factors and the deposition of histone modifications, chromatin proteins, and DNA methylation near the barcode is affected by different chromatin environments. We anticipate that TRIP may also be applicable to study other genome-related functions, such as DNA replication and DNA repair.

Gene Regulatory Patterns across the Genome

The expression pattern of IRs across the genome is not random and correlates partially with the previously described LADs and inter-LADs (Guelen et al., 2008; Peric-Hupkes et al., 2010). In part, the reduced activity of IRs in LADs may be explained by the low density of functional enhancers and active genes in LADs. Partial correlation analysis indicates that another aspect of chromatin architecture at LADs contributes to attenuated transcription. How this attenuation is achieved is not clear, but it is likely to involve

reduced binding of transcription factors to their cognate binding sites.

IR expression also correlates with the local compaction of chromatin prior to integration. We note that we calculated the α values over a 400 kb window, which is large compared to the size of the IRs; estimates of α values in smaller windows will require Hi-C data of yet higher resolution. We do not know whether the differences in chromatin conformation are a direct determinant of IR expression, or merely reflective of another key feature of chromatin, such as the presence of various repressive or activating proteins. Interestingly, the IR expression in LADs is consistently lower compared to inter-LAD regions with similar α value. This indicates that chromatin compaction alone does not fully explain the attenuation of transcription in LADs; other features such as their contacts with the nuclear lamina or their distinct histone modification state may render LADs less permissive to transcription (Kind and van Steensel, 2010). The lack of a clear relationship between IR expression patterns and TADs may be attributed to the relatively low precision at which both the IR expression domains and TADs are currently defined; alternatively, TADs and IR expression domains may be biologically distinct aspects of chromosome organization.

Our data reveal that IRs are generally more active when located within ~200 kb from active genes. This substantial crosstalk suggests that the linear order and spacing of genes along chromosomes is of importance for gene regulation. Indeed, bioinformatics studies have shown that neighboring genes tend to be coexpressed (Hurst et al., 2004; Michalak, 2008). Previous experimental studies noted a transcription "ripple effect" between neighboring genes (Ebisuya et al., 2008) and activation of IRs nearby active gene clusters (Gierman et al., 2007), but these studies lacked the statistical power needed to identify the origin of the activating signals. Our analysis suggests that the crosstalk arises in part from the active transcription units themselves, and in part from enhancers that surround active genes. Which component of active transcription units is responsible for the observed crosstalk remains to be determined. Reciprocal effects of the IRs on neighboring genes are also limited to a range of ~20 kb, but only a minority of neighboring genes appears sensitive. It will be interesting to further investigate the basis of this differential sensitivity of genes.

Although our initial data analyses point to regulatory contributions of LADs, chromatin states that differ in the degree of compaction, neighboring genes, and enhancers, we note that these features do not fully explain the large dynamic range (~1,000-fold) in IR expression levels. Further computational modeling of the data may uncover additional features that determine gene expression.

EXPERIMENTAL PROCEDURES

Plasmid Libraries

Construction of the barcoded piggyBac plasmid libraries is described in the Extended Experimental Procedures.

Mouse Embryonic Stem Cell Culture and Transfection

mES cells EBRTcH3 expressing the tetracycline-controlled transactivator from the endogenous *ROSA26* promoter (Masui et al., 2005) were cultured in 60% BRL cell-conditioned medium in the presence of leukemia inhibitory factor, MEK inhibitor PD0325901, and GSK-3 inhibitor CHIR99021 (Ying et al., 2008). Four hours before transfection, 6×10^6 EBRTcH3 cells were seeded on a 10 cm dish. The cells were transfected with 22.5 μg of barcoded PB plasmid library and 2.5 μg of mouse codon-optimized version of PB transposase (mPB) plasmid (Cadiñanos and Bradley, 2007) using Lipofectamine 2000 (Invitrogen). Mock-transfected and nontransfected controls were included. After 36–48 hr, the cells were sorted with fluorescence-activated cell sorting (FACS) into three populations with respect to eGFP signal. We discarded cells without any detectable eGFP signal, because they most likely failed to take up any plasmid. We also discarded cells with very high eGFP signals because typically these cells have a large number of integrations per cell. The cells with medium levels of eGFP expression were used to establish the cell pools with IRs. Note that the sorting of cells was done within a time window when most eGFP expression is coming from free plasmid; hence, a possible bias caused by this selection step is most likely minor. Furthermore, a significant number (>1%) of IRs had undetectable level of expression according to our measurements (see below). After sorting, the medium-eGFP population was grown for 5 days before several aliquots of ~1,000 cells were subcultured to establish the "biological replicate" mES cell pools, each with a different collection of integrated transgenes. Because sequencing of each pool identified ~7,000–11,000 barcodes (Table S1) of the expected ~23,000 (1,000 cells times ~23 IRs/cell on average according to quantitative PCR), it is possible that we overestimated the number of cells subcultured, that not all cells survived the subculturing step, or that barcodes were missed in the sequencing (which is less likely considering large overlap and strong correlation between the technical replicates). Two weeks after transfection, each cell pool was split into two "technical replicates," which were grown independently for another week before the isolation of total RNA and genomic DNA (gDNA) (Figure S1A).

Preparation of Samples for High-Throughput Illumina Sequencing

Mapping of the barcoded PB insertion sites was done by inverse PCR (Ochman et al., 1988) coupled with high-throughput sequencing. Briefly, 2 μg of gDNA was digested with 20 units of DpnII (New England Biolabs) overnight at 37°C in a volume of 100 μl. Subsequently, 600 ng of purified digested DNA was self-ligated with 40 units of high-concentration T4 DNA ligase (Promega) overnight at 4°C in a volume of 400 μl (two times for each technical replicate of the TRIP pool). The ligation reactions were phenol/chloroform/isoamylalcohol extracted and ethanol precipitated. DNA pellets were dissolved in 30 μl of water. Five microliters of each sample was used as a template for amplification of fragments containing both the barcodes and flanking genomic DNA regions. PCR was performed in three rounds (for details, see Table S2), and purified products were directly used for high-throughput Illumina paired-end sequencing.

To measure the barcode expression levels, 2 μg of total RNA was reverse transcribed in a 50 μl reaction containing 50 ng of oligo(dT) primer and 1 μl of Superscript II (Invitrogen). One microliter of cDNA was used as a template for amplification of barcode sequences. PCR was performed in two rounds (for details, see Table S2), and purified products were directly used for high-throughput Illumina single-read sequencing. To quantify the barcode abundances for normalization, 100 ng of gDNA instead of cDNA was used as a template.

Validation of Mapped piggyBac Insertions

For the validation of mapping of insertion sites by inverse PCR, 11 IRs were randomly chosen from the pool mPGK-A. gDNA (100 ng) from each technical replicate of mPGK-A was used as a template for amplification with a nested set of the reporter-specific and the location-specific primers (Figure 2B; Tables S3 and S4). The PCR products were run on a 1.5% agarose gel for visualization. To verify the barcode sequence, the PCR products were Sanger sequenced using the primer PB-Valid.Gen.Seq-1 (Table S3). The gDNA from pool mPGK-B was used as a negative control.

Processing and Analysis of TRIP Data

Detailed descriptions of the processing and analysis of TRIP data are provided in the Extended Experimental Procedures.

ACCESSION NUMBERS

The GenBank accession numbers for the TRIP vectors and libraries are KC710227–KC710231. TRIP and RNA-seq data are available from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), accession number GSE48606.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, four tables, and four data sets and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.07.018.

ACKNOWLEDGMENTS

We thank the NKI Genomics Core Facility for sequencing support, Guillaume Filion for insightful suggestions, Mario Amendola for providing the reference plasmid for IR copy number quantification, and members of our laboratories for helpful discussions and critical reading of the manuscript. This work was supported by the Netherlands Consortium for Systems Biology (L.F.A.W., M.v.L., B.v.S.) and EURYI, NWO-ALW VICI and ERC Advanced grant 293662 (B.v.S.).

Received: February 19, 2013
Revised: May 31, 2013
Accepted: July 12, 2013
Published: August 15, 2013

REFERENCES

Babenko, V.N., Makunin, I.V., Brusentsova, I.V., Belyaeva, E.S., Maksimov, D.A., Belyakin, S.N., Maroy, P., Vasil'eva, L.A., and Zhimulev, I.F. (2010). Paucity and preferential suppression of transgenes in late replication domains of the D. melanogaster genome. BMC Genomics *11*, 318.

Bickmore, W.A., and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

Cadiñanos, J., and Bradley, A. (2007). Generation of an inducible and optimized piggyBac transposon system. Nucleic Acids Res. *35*, e87.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117.

Chen, M., Licon, K., Otsuka, R., Pillus, L., and Ideker, T. (2013). Decoupling epigenetic and genetic effects through systematic analysis of gene position. Cell Rep. *3*, 128–137.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. USA *107*, 21931–21936.

De, S., Teichmann, S.A., and Babu, M.M. (2009). The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. Genome Res. *19*, 785–794.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dobzhansky, T. (1936). Position effects on genes. Biol. Rev. Camb. Philos. Soc. *11*, 364–384.

Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008). Ripples from neighbouring transcription. Nat. Cell Biol. *10*, 1106–1113.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Foat, B.C., Morozov, A.V., and Bussemaker, H.J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics *22*, e141–e149.

Furey, T.S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet. *13*, 840–852.

Gerlach, C., van Heijst, J.W., Swart, E., Sie, D., Armstrong, N., Kerkhoven, R.M., Zehn, D., Bevan, M.J., Schepers, K., and Schumacher, T.N. (2010). One naive T cell, multiple fates in CD8+ T cell differentiation. J. Exp. Med. *207*, 1235–1246.

Gerrits, A., Dykstra, B., Kalmykowa, O.J., Klauke, K., Verovskaya, E., Broekhuis, M.J., de Haan, G., and Bystrykh, L.V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood *115*, 2610–2618.

Gierman, H.J., Indemans, M.H., Koster, J., Goetze, S., Seppen, J., Geerts, D., van Driel, R., and Versteeg, R. (2007). Domain-wide regulation of gene expression in the human genome. Genome Res. *17*, 1286–1295.

Girton, J.R., and Johansen, K.M. (2008). Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila. Adv. Genet. *61*, 1–43.

Gossen, M., Freundlieb, S., Bender, G., Müller, G., Hillen, W., and Bujard, H. (1995). Transcriptional activation by tetracyclines in mammalian cells. Science *268*, 1766–1769.

Grewal, S.I., and Jia, S. (2007). Heterochromatin revisited. Nat. Rev. Genet. *8*, 35–46.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet. *43*, 630–638.

Hawkins, R.D., Hon, G.C., and Ren, B. (2010). Next-generation genomics: an integrative approach. Nat. Rev. Genet. *11*, 476–486.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Hezroni, H., Sailaja, B.S., and Meshorer, E. (2011). Pluripotency-related, valproic acid (VPA)-induced genome-wide histone H3 lysine 9 (H3K9) acetylation patterns in embryonic stem cells. J. Biol. Chem. *286*, 35977–35988.

Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.W., Lyou, Y., Townes, T.M., Schübeler, D., and Gilbert, D.M. (2008). Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol. *6*, e245.

Huang, X., Guo, H., Tammana, S., Jung, Y.C., Mellgren, E., Bassi, P., Cao, Q., Tu, Z.J., Kim, Y.C., Ekker, S.C., et al. (2010). Gene transfer efficiency and genome-wide integration profiling of Sleeping Beauty, Tol2, and piggyBac transposons in human primary T cells. Mol. Ther. *18*, 1803–1813.

Hurst, L.D., Pál, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. *5*, 299–310.

Kind, J., and van Steensel, B. (2010). Genome-nuclear lamina interactions and gene regulation. Curr. Opin. Cell Biol. *22*, 320–325.

Korzh, V. (2007). Transposons as tools for enhancer trap screens in vertebrates. Genome Biol. *8*(*Suppl 1*), S8.

Li, G., and Reinberg, D. (2011). Chromatin higher-order structures and gene regulation. Curr. Opin. Genet. Dev. *21*, 175–186.

Li, M., He, Y., Dubois, W., Wu, X., Shi, J., and Huang, J. (2012). Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. Mol. Cell *46*, 30–42.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al.

(2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lienert, F., Mohn, F., Tiwari, V.K., Baubec, T., Roloff, T.C., Gaidatzis, D., Stadler, M.B., and Schübeler, D. (2011). Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. PLoS Genet. *7*, e1002090.

Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J., et al. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521–533.

Masui, S., Shimosato, D., Toyooka, Y., Yagi, R., Takahashi, K., and Niwa, H. (2005). An efficient system to establish multiple embryonic stem cell lines carrying an inducible expression unit. Nucleic Acids Res. *33*, e43.

McBurney, M.W., Sutherland, L.C., Adra, C.N., Leclair, B., Rudnicki, M.A., and Jardine, K. (1991). The mouse Pgk-1 gene promoter contains an upstream activator sequence. Nucleic Acids Res. *19*, 5755–5761.

Meir, Y.J., Weirauch, M.T., Yang, H.S., Chung, P.C., Yu, R.K., and Wu, S.C. (2011). Genome-wide target profiling of piggyBac and Tol2 in HEK 293: pros and cons for gene discovery and gene therapy. BMC Biotechnol. *11*, 28.

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics *91*, 243–248.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Mohn, F., Weber, M., Schübeler, D., and Roloff, T.C. (2009). Methylated DNA immunoprecipitation (MeDIP). Methods Mol. Biol. *507*, 55–64.

Montavon, T., and Duboule, D. (2012). Landscapes and archipelagos: spatial organization of gene regulation in vertebrates. Trends Cell Biol. *22*, 347–354.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Ochman, H., Gerber, A.S., and Hartl, D.L. (1988). Genetic applications of an inverse polymerase chain reaction. Genetics *120*, 621–623.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol. Cell *38*, 603–613.

Ruf, S., Symmons, O., Uslu, V.V., Dolle, D., Hot, C., Ettwiller, L., and Spitz, F. (2011). Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. Nat. Genet. *43*, 379–386.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Southall, T.D., and Brand, A.H. (2007). Chromatin profiling in model organisms. Brief. Funct. Genomics Proteomics *6*, 133–140.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D., Dean, C., Ma, H., and Martienssen, R. (1995). Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. Genes Dev. *9*, 1797–1810.

Sutherland, H., and Bickmore, W.A. (2009). Transcription factories: gene expression in unions? Nat. Rev. Genet. *10*, 457–466.

Vogel, M.J., Peric-Hupkes, D., and van Steensel, B. (2007). Detection of in vivo protein-DNA interactions using DamID in mammalian cells. Nat. Protoc. *2*, 1467–1478.

Weber, F., de Villiers, J., and Schaffner, W. (1984). An SV40 "enhancer trap" incorporates exogenous enhancers or generates enhancers from its own sequences. Cell *36*, 983–992.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.