COMPUTATIONAL EPIGENOMICS IN GENE REGULATION AND CANCER RESEARCH

FINDING THE MUSIC IN THE NOISE

COMPUTATIONAL EPIGENOMICS IN GENE REGULATION AND CANCER RESEARCH

FINDING THE MUSIC IN THE NOISE

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 2 maart 2015 om 12:30 uur

door

Johann DE JONG

doctorandus in de computational science
geboren te Enkhuizen, Nederland.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. L. F. A. Wessels

Copromotor: Dr. J. de Ridder

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. L. F. A. Wessels,	Technische Universiteit Delft, promotor
Dr. J. de Ridder,	Technische Universiteit Delft, copromotor
Prof. dr. A. Berns	The Netherlands Cancer Institute Skolkovo Center for Stem Cell Research
Prof. dr. E. Cuppen	Hubrecht Institute Utrecht University
Assoc. prof. dr. M. Fornerod	Erasmus University Medical Center
Prof. dr. M. Reinders	Delft University of Technology
Prof. dr. B. Snel	Utrecht University

Printed by: Off Page – `http://www.offpage.nl/`

Front & Back: Marco Jeurissen – `http://www.marcojeurissen.nl`

ISBN 978-94-6182-541-4

Molecular biology could read notes in the score, but it couldn’t hear the music.

Carl R. Woese

CONTENTS

1 Introduction 1
1.1 Chromatin . . . 2
1.2 Methods in chromatin biology and epigenomics . . . 6
1.3 Computational epigenomics . . . 10
1.4 Outline of this dissertation . . . 12
References . . . 16

2 Computational identification of insertional mutagenesis targets for cancer gene discovery 27
2.1 Introduction . . . 28
2.2 Results . . . 29
2.3 Discussion . . . 38
2.4 Materials and Methods . . . 40
References . . . 42
2.5 Supplementary Material . . . 45

3 Chromatin landscapes of retroviral and transposon integration profiles 47
3.1 Introduction . . . 48
3.2 Results . . . 50
3.3 Discussion . . . 60
3.4 Materials and Methods . . . 64
References . . . 70
3.5 Supplementary Material . . . 79

4 Chromatin position effects assayed by thousands of reporters integrated in parallel 81
4.1 Introduction . . . 82
4.2 Results . . . 83
4.3 Discussion . . . 93
4.4 Materials and Methods . . . 97
References . . . 105
4.5 Supplementary Material . . . 111

5 Genome-wide profiling of anti-cancer drug effects on chromatin guides rational treatment design 113
5.1 Introduction . . . 114
5.2 Results . . . 116
5.3 Discussion . . . 128
5.4 Materials and Methods . . . 131
References . . . 137

5.5 Supplementary Material 143

6 Discussion **153**

6.1 Chromatin profiling 154

6.2 Analysis of genome-wide sequencing data 158

6.3 Future perspectives 167

References 167

Summary **171**

Samenvatting **173**

Acknowledgments **175**

Curriculum Vitæ **177**

List of Publications **179**

1

INTRODUCTION

1.1. CHROMATIN

Imagine a 30-year old male, 1.72m tall and weighing 70kg. This man contains a total amount of DNA that, when stretched out, would cover the distance from the earth to the moon, almost 200000 times.[1] Wrapping these 7.6×10^{13} meters of DNA into a 1.72m male represents one of the main functions of chromatin.

In addition to wrapping the DNA into the small space of the nucleus of a cell, an important function of chromatin is the regulation of gene activity, e.g. by dynamic repackaging of DNA, by binding of certain proteins to DNA, or by looping of DNA allowing DNA-DNA contacts. Many specifics regarding this regulation are still unknown. Even to what precise extent chromatin structure directly determines gene activity, or the other way around, is still unknown.

Perturbations of chromatin structure can be associated with many diseases [4]. For example, suppose that our 30-year old male has developed some type of cancer. The chromatin structure of the DNA in his tumor cells may show cancer-specific perturbations that led to perturbations of gene activity. Perturbations in gene activity could potentially have provided cells with a certain proliferative or survival advantage eventually leading to tumor formation [5]. He may in fact be treated using drugs that further perturb chromatin, e.g. by deliberately unwrapping and breaking DNA in tumor cells [6]. Just as certain chromatin perturbations may cause cells to proliferate and develop into a tumor, other perturbations may cause cells to die, illustrating the importance of chromatin for normal cell functioning.

Since the structure of chromatin is so essential for cell functioning, artificially perturbing it may give important new insights into its function. In this dissertation, chromatin is perturbed by two means, 1) DNA integrating elements that induce mutations by integrating into host DNA, and 2) anti-cancer drugs that induce unwrapping of chromatin and breaking of DNA. The DNA integrating elements are used for sensing the location-dependent activation potential of chromatin, and the drugs are used for characterizing chromatin in terms of its location-dependent response to these drugs. As such, this dissertation attempts to go deeper into the relations between chromatin, gene activity and cancer.

Since chromatin has many facets that interact in intricate ways, and packages a huge genome, the work presented in this dissertation is strongly based in computational approaches for integrating all these facets across whole genomes. Therefore, a more elaborate introduction into what chromatin actually is, and the computational approaches for analyzing it, would be appropriate.

1.1.1. MULTI-LEVEL DNA PACKAGING

Chromatin can be defined as the complex of DNA and proteins that are used to package this DNA in the nucleus of a eukaryotic cell. The main class of DNA packaging proteins are the histone proteins. At the first level of packaging, ~146 base pairs of DNA wrap in 1.65 turns around an octamer of core histone proteins (H2A, H2B, H3, and H4; Fig-

[1] A 30-year old, 1.72m tall, male, weighing 70kg, contains $\sim 3.72 \times 10^{13}$ human cells [1]. Each cell contains $\sim 6 \times 10^{9}$ base pairs of DNA [2]. The distance between two base pairs is $\sim 0.34 \times 10^{-9}$m [3]. The average lunar distance is 384.4×10^{6}m. Thus, $\frac{6\times10^{9}\times0.34\times10^{-9}\times3.72\times10^{13}}{384.4\times10^{6}} \approx 200000$ times the distance from the earth to the moon.

ure 1.1) [7], thus forming the basic structural unit of chromatin, the nucleosome. The nucleosome is sealed off with the linker histone protein H1, wrapping an additional 20 base pairs, and is connected to neighboring nucleosomes by 10 - 80 base pairs of linker DNA [8]. Higher levels of packaging consist of folding, looping and coiling, further compacting the nucleosomes into a 30 nm fiber, and eventually forming an ultra-compacted chromatid, such as seen during the metaphase of a cell. During any phase of the cell cycle, not all chromatin is compacted to the same degree. There is a large-scale separation between chromatin that is highly condensed (heterochromatin) and chromatin that is more relaxed (euchromatin). Contrary to euchromatin, heterochromatin tends to be associated with low transcriptional activity [9].

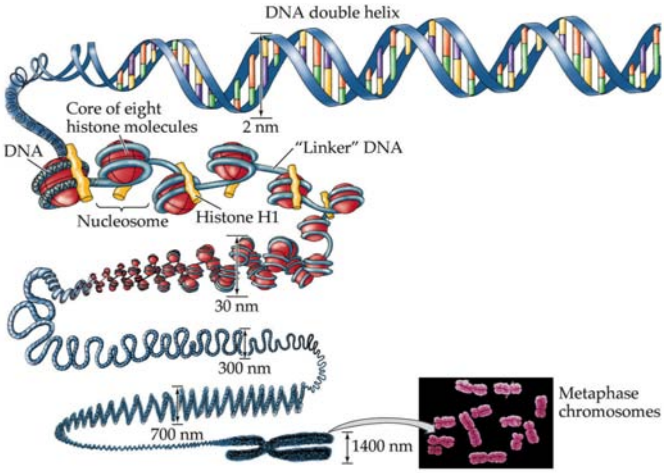


Figure 1.1: [10] The packaging of DNA into chromatin.

1.1.2. EPIGENOMICS

The histone proteins within a nucleosome have outward-facing tails, the histone tails. These tails can be post-translationally modified. An important example of a histone modification is acetylation at lysine residues, where e.g. the addition of an acetyl group on lysine 27 on histone H3 is referred to as H3K27ac. Lysine acetylation is catalyzed by HATs (histone acetyl transferases), such as p300 [11]. Acetyl removal is catalyzed by HDACs (histone deacetylases). Other notable examples of modifications are methylation and phosphorylation (Figure 1.2).

A histone modification is an example of an epigenetic modification. The term epigenetics refers to changes in chromatin composition that can involve anything but changes in the nucleotide sequence. Moreover, these changes need to be functionally relevant, i.e. potentially linked to changes in gene expression. As such, epigenetics is at the basis of cellular differentiation, where different cell types will have the same genome, but a different epigenome. This is studied in the field of epigenomics, as opposed to genomics which focuses strictly on DNA structure and function.

In addition to histone modifications, important epigenetic modifications are methylation of DNA at CpG dinucleotides, which has been linked to cellular differentiation and transcriptional regulation [12], and the binding of transcription factors to DNA. Transcription factors are proteins that bind to DNA in a sequence-specific manner in order to activate or repress transcription of nearby genes[13].

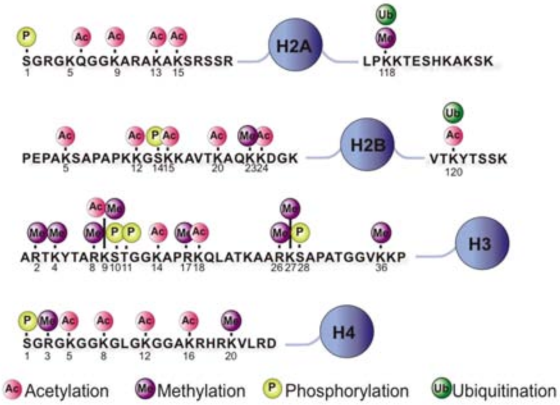


Figure 1.2: [14] Histone proteins H3, H4, H2A and H2B and their post-translationally modified tails.

1.1.3. EPIGENOMICS AND GENE EXPRESSION

For many histone modifications, their distribution across the genome correlates very well with chromatin structure and gene expression, see e.g. [9, 15]. It has been suggested that the reason for this is that histone modifications affect the binding affinities of chromatin-associated proteins [16]. In support of this, it has indeed been shown that histone acetylation can form binding sites for bromodomain proteins [17]. However, histone acetylation can also prevent the compaction of chromatin [18], pointing to another possible mechanism, where modifications exert their effect by changing histone-DNA binding affinity, thus opening (or closing) the chromatin, and making it more (or less) accessible for the transcription machinery.

Some histone modifications are known to act in a combinatorial fashion. These specific combinations of histone modifications ('chromatin states') may be associated with specific gene expression patterns. A typical example of this can be seen at some promoters, where tri-methylation of lysine 27 on histone H3 (H3K27me3) in combination with H3K4me3 signifies a poised promoter [19]. Upon differentiation of a cell, a poised promoter will typically become either active (loss of H3K27me3) or repressed (loss of H3K4me3). Another example concerns enhancers, where the presence of H3K27ac can mark the difference between poised enhancers (only H3K4me1) and active enhancers (H3K27ac and H3K4me1) [20]. Active enhancers can enhance expression of genes across large genomic distances, an effect generally attributed to the looping of DNA which brings the gene in close spatial proximity to the enhancer [21].

1.1.4. THE CHROMATIN POSITION EFFECT

Although differences in chromatin structure strongly associate with differences in gene expression, it is still difficult to unambiguously attribute these differences to epigenomics, since not only the (epi)genomic context is variable (e.g. different levels of protein binding, different DNA sequence context), but also the genes and promoters across the genome. To study the 'chromatin position effect', i.e. the unambiguous influence of the local chromatin environment on gene expression, the gene/promoter variable has to be eliminated. DNA integrating elements, such as retroviruses and transposons can be used to study the chromatin position effect. Retroviruses and transposons have the ability to insert their genetic material into host DNA. When genetically engineered to contain a promoter and reporter gene of choice, they can be used to insert the same gene into different genomic loci, and thus infer causal relationships between local (epi)genomic context and gene expression [22–24].

1.1.5. CHROMATIN REMODELING

Chromatin remodeling refers to dynamically altering the structure of chromatin and is closely linked to transcriptional regulation [25]. The relaxing or condensing of chromatin by post-translational histone modifications is one example of chromatin remodeling (Figure 1.3b). Whereas this type of remodeling is dependent on covalent modification of histone proteins, chromatin can also be remodeled by protein complexes in an ATP-dependent fashion. This includes 1) altering the composition of the nucleosome by replacing histone proteins with variants (Figure 1.3a), repositioning nucleosomes by sliding them across the genome (Figure 1.3c) and eviction of nucleosomes from the chromatin. A notable example of replacement is histone variant H2A.Z, catalyzed by the chromatin remodeler SWR1 [26]. Histone variant H2A.Z is known to be involved in transcriptional regulation [27]. Another example is H2A.X which, when phosphorylated, is a marker for DNA double strand breaks [28].

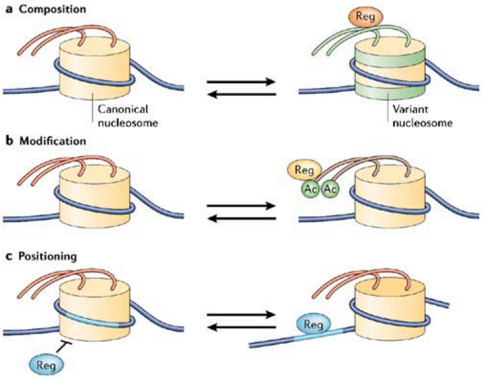


Figure 1.3: [29] Chromatin remodeling induced by regulatory proteins (Reg). (a) Exchange of histone proteins within the octamer core of the nucleosome. (b) Post-translational modification of histone proteins by the addition of e.g. an acetyl group. (c) Sliding of nucleosomes across the DNA.

1.1.6. CHROMATIN PERTURBATION

Chromatin remodeling can be artificially induced, for example by certain drugs. Notable examples are chromatin remodeling by cocaine [30] and psychotropic drugs [31]. More recently, it has also been shown that Doxorubicin, a commonly used anti-cancer drug, partly relies on its ability to evict histones for its chemotherapeutic effects [6].

Evidently, mutations in chromatin remodeling proteins, such as ARID1A, can affect transcription and chromatin structure [32]. However, transcription itself can also induce chromatin remodeling [33]. Therefore, many techniques that induce mutations to influence gene expression, can likely induce remodeling of chromatin. An important example here comes from the field of insertional mutagenesis (IM), where retroviruses and transposons are used for cancer gene discovery. Retroviruses and transposons cause mutations by integrating into host DNA. Once integrated, they can affect the expression of surrounding genes. Integrations that provide a cell with a proliferative advantage can potentially cause tumors, and as such, retroviral and transposon integrations can be used to tag cancer genes, e.g. [34–37]. Regarding effects specifically on chromatin, it has in fact been shown that retroviral integrations can induce methylation of host DNA up to 1kb from the site of integration [38, 39].

1.2. METHODS IN CHROMATIN BIOLOGY AND EPIGENOMICS

The term chromatin was coined by Walther Flemming at the end of the 19th century, after visualizing it using aniline dyes. Since then, many techniques have been developed to gain insight into the structure of chromatin, its modifications, and their association with gene expression. During the last two decades, the development of high-throughput technologies has revolutionized biology by allowing a shift from hypothesis-driven science to a more discovery-driven science. Traditional research focused on relatively small numbers of elements of interest (genes, proteins, etc.) that were hypothesized or known to be relevant in a certain setting. However, high-throughput technologies are much less dependent on such a priori knowledge, allowing a more objective discovery of relevant elements. Consequently, available data and knowledge have exploded during the last two decades. In the following, a brief description will be given of some important high-throughput technologies, with a focus on those that are relevant for the remainder of this dissertation. These technologies can be array-based technologies or sequencing-based, and for most technologies discussed below, both options are available. In array-based technologies, arrays (or biochips) can recognize a limited number of predetermined target sequences, e.g. representing genes, by hybridization of target sequences to complementary probes on the array. Since array-based technologies depend on a limited number of predetermined target sequences, complete coverage of large genomes is difficult. In other words, the resolution of array-based technologies is often limited. On the other hand, sequencing-based technologies rely on determining the order of nucleotides of individual sequences, which are then mapped back to a reference genome. As such, they can potentially recognize any individual sequence in a pool of sequences, and genome coverage is generally higher than in array-based technologies.

1.2.1. GENE EXPRESSION

GENE EXPRESSION MICROARRAYS

Gene expression microarrays were among the first broadly used array-based technologies [40], and are still used today. They typically work by isolating mRNA, reverse transcribing it to cDNA, and then hybridizing it to a microarray. Microarrays contain a predefined set of probes, each of which is designed to bind a certain target sequence of interest. In the case of the human and mouse genomes, a gene expression microarray contains probes representing ~20000 protein coding genes. In addition to a more discovery-based science, as exemplified among others by the development of prognostic profiles in cancer and cancer subtyping [41–43], gene expression microarrays have allowed for the large-scale reconstruction of gene networks, see for example [44, 45].

RNA-SEQUENCING (RNA-SEQ)

In RNA-seq (Figure 1.4), mRNA is isolated and reverse transcribed to cDNA, then fragmented and finally sequenced. Whereas gene expression microarrays and RNA-seq can mostly be used for the same purposes, RNA-seq has the advantage that it is genome-wide. It is not limited to a predefined set of probes representing known genes, and can therefore be used for the discovery of RNA isoforms (splicing variants), gene fusions [46, 47], etc.

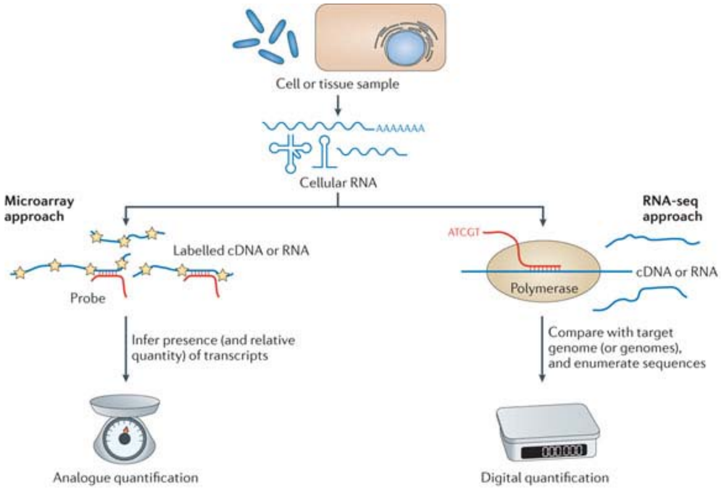


Figure 1.4: [48] Using a microarray, the relative abundance of transcripts is determined by measuring a fluorescent signal that is bounded in intensity. Measuring relative abundance implies there is no natural zero to measure the absence of transcripts. In contrast, in RNA-seq, the individual transcripts are counted, thus providing a natural zero. Furthermore, in principle, there is no upper bound as to the number of transcripts that can be counted.

1.2.2. OPEN CHROMATIN

The technologies described in this section are used to determine which genomic regions are free of nucleosomes, i.e. can be considered open chromatin.

DNASE-SEQUENCING (DNASE-SEQ)

In DNase-seq, DNA is cleaved ('digested') using the Deoxyribonuclease I enzyme (DNase I), and subsequently sequenced. Regions that are preferentially cleaved by DNase I, DNase I hypersensitive sites, are considered to represent open chromatin, and are enriched for gene regulatory elements such as active promoters, active enhancers, etc. [49, 50].

FAIRE-SEQUENCING (FAIRE-SEQ)

Compared to DNase I hypersensitivity assays, Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) is a more recent alternative for probing open chromatin and regulatory regions. In FAIRE-seq, chromatin is first chemically fixed ('cross-linked') using formaldehyde, and then fragmented after which DNA without bound proteins is isolated and sequenced.

There is a strong overlap in regions detected by DNase-seq and FAIRE-seq. However, each assay also identifies unique regions, and a combination of both was shown to be more effective in identifying regulatory elements than using any of the two assays in isolation [51].

1.2.3. PROTEIN-DNA INTERACTIONS

CHIP

Chromatin Immunoprecipitation (ChIP) identifies interactions between DNA and a protein of interest by using antibodies that specifically bind to that protein. Thus, the protein is pulled down ('precipitated') from a solution containing fragmented ('sheared') DNA in which the interactions with proteins were temporarily fixed ('cross-linked') to allow for precipitation. The DNA enriched by ChIP can be hybridized to an array (ChIP-chip) to determine interactions with a predefined set of genomic loci. With first studies appearing at the turn of the century [52, 53], ChIP-chip can be considered the earliest high-throughput application of ChIP, and is still widely used. For example, it has been used to gain insight into the organization of replication timing domains [54] and H3K9me2 domains [55] during cell differentiation.

Alternatively, ChIP can be followed by sequencing (ChIP-seq), to potentially identify all genomic loci that interact with the protein of interest (Figure 1.5). Since the appearance of three landmark studies shortly after another in 2007 [19, 56, 57], ChIP-seq has grown to dominate chromatin biology as the most widely used high-throughput technology. A wide range of transcription factors and histone marks have been mapped by ChIP-seq, enabling numerous breakthroughs in the understanding of chromatin organization, gene regulation, cellular differentiation, etc. [19, 20, 56–62]

Some examples of variations on ChIP-seq are MeDIP-seq for profiling methylation of cytosines [63], hMeDip-seq, for hydroxymethylation of cytosines [64], and Repli-seq for estimating the relative timing of replication of different genomic regions [65].

DAMID

DNA adenine methyltransferase identification (DamID) identifies protein-DNA interactions by fusing a protein of interest to the Dam enzyme. When this protein binds to DNA, the fused Dam protein catalyses methylation of adenine (at position 6) at nearby GATCs.

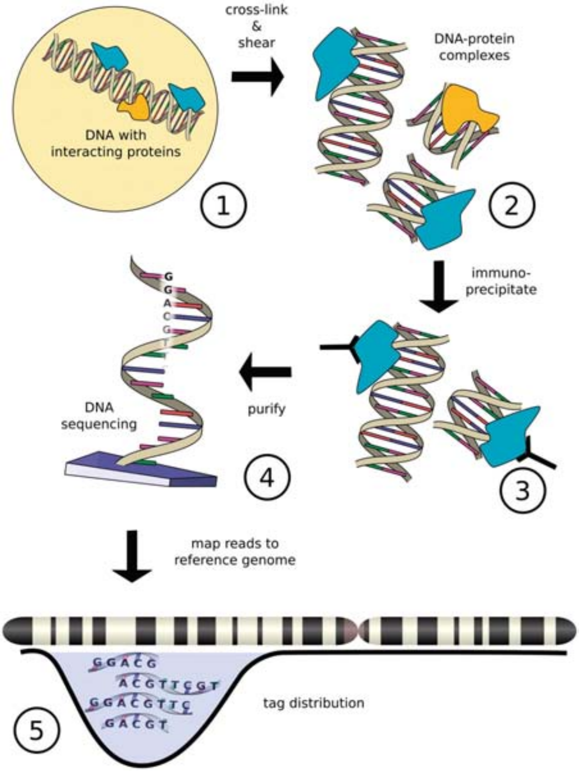


Figure 1.5: [66] Using ChIP-seq, genome-wide profiles of protein binding can be generated. First, DNA-protein interactions are fixed by cross-linking and DNA is fragmented by shearing. From this fragmented DNA, the DNA bound to a protein of interest is pulled down (precipitated) using an antibody specific to that protein. This precipitated DNA is purified and sequenced. Finally, the resulting sequence reads are mapped back to a reference genome, after which the distribution of tags (mapped reads) across the genome can be studied.

Adenine methylation at position 6 generally does not occur in eukaryotes, and can therefore be used as a marker for protein binding. DamID output is typically hybridized to a microarray [67–69], but it can also be followed by sequencing [70]. Important results obtained using the DamID technology include the genome-wide definition of lamina-associated domains [67, 71], which are large genomic regions preferentially located at the nuclear lamina and generally associated with transcriptional repression. DamID and ChIP assays have been shown to give similar results [61, 72, 73].

1.2.4. DNA-DNA INTERACTIONS

HI-C

Hi-C is a technique for the genome-wide detection of 3D chromatin interactions [74]. It works by cross-linking DNA, thus temporarily fixing interacting chromatin segments. Then, the chromatin is fragmented using a restriction enzyme. This leaves sticky ends, which, after being filled with nucleotides marked with biotin, are ligated. Finally, fragments with biotin are pulled down and sequenced from both ends to identify regions of genomes present in spatial proximity

Using Hi-C, it was shown that the log of the contact probability between two loci linearly decays with the log of the distance between these loci, the slope of which was consistent with a type of organization called a fractal globule [74]. In another interesting study, large chromatin interaction domains, termed topological domains, were inferred using Hi-C and suggested to constitute a fundamental organizing principle of metazoan genomes [75].

1.3. COMPUTATIONAL EPIGENOMICS

The development of high-throughput technologies has allowed a shift from hypothesis-driven science to more discovery-based science. With the ever increasing amount of data generated using the techniques described above, new computational techniques for analyzing these data were also required. This section gives a brief overview of some of these techniques, with a focus on sequencing-based techniques that map histone modifications, protein binding and open chromatin, since these are especially relevant for this dissertation.

1.3.1. PREPROCESSING OF GENOME-WIDE SEQUENCING DATA

This section will give a high-level overview of steps typically taken in preprocessing ChIP-seq, FAIRE-seq and DNase-seq data. The starting point for preprocessing is assumed to consist of sequence reads, obtained from a sequencing machine such as built by Illumina or Ion Torrent [76]. Sequence reads are relatively short sequences of nucleotides (a few 10s to a few 100s), which are intended to represent genomic regions e.g. that were bound by a certain protein (ChIP-seq), where histones were modified (ChIP-seq), or where the chromatin was open (FAIRE-seq, DNase-seq).

READ MAPPING

Since the objective is to obtain genome-wide distributions of protein binding, histone modification or open chromatin, the first step is to map the reads to a reference genome.

Many tools are available for this, with notable examples being Bowtie [77] and BWA [78]. After mapping, typically a genome-wide coverage profile will be computed by counting reads in equal-sized consecutive bins across the genome.

Normalization

The steps taken to generate and map the sequence reads induce biases, for example due to PCR amplification biases [79], GC-content [80] and mappability problems [81]. This makes comparing a signal of interest to a control essential in analyzing genome-wide sequencing data. Irrespective of the type of control experiment [82], the eventual problem is how to make read counts in the signal quantitatively comparable to read counts in the control. Normalization techniques can be divided into linear and nonlinear techniques [83]. Nonlinear normalization techniques are often inspired by normalization techniques originally applied to microarray data, such as locally weighted scatterplot smoothing (LOWESS) normalization [84, 85], and quantile normalization [86]. On the other hand, a typical linear technique tries to estimate a factor, the normalization factor, by which the control profile is scaled such that the number of control reads in any genomic window is quantitatively comparable to the number of signal reads in the same window. The most basic, but nevertheless commonly used, estimation of the normalization factor is the ratio of signal sequencing depth and control sequencing depth. Other techniques for estimating normalization factors try to detect regions in the signal that can be considered background noise, and calculate the normalization factor as the ratio of signal reads and controls reads in those background regions, e.g. [87, 88].

Peak calling

Often, one is interested in genomic regions that have significantly more reads mapped to them than expected by chance. In case of ChIP-seq against a transcription factor, these regions, or peaks, would correspond to transcription factor binding sites. In the case of FAIRE-seq, the peaks would correspond to regions of open chromatin. A complicating factor in peak calling is that expected peak widths can differ widely across datasets. Transcription factors often show very narrow peaks, whereas some histone marks (H3K36me3, H3K27me3, H3K9me3, etc.) may be more domain-oriented. There are even chromatin marks such as RNA polymerase II that, depending on post-translational modifications, can display either narrow peaks (phosphorylation of serine 5) or broad domains (phosphorylation of serine 2). A common approach for calling narrow peaks uses a local Poisson background to determine significantly enriched regions, such as implemented in the software tool MACS [89]. Algorithms have also been developed that focus on calling broad domains on ChIP-seq data, such as SICER [90]. SICER uses a global Poisson background to call ’eligible windows’, i.e. potentially significant regions, which it merges to ’islands’, allowing gaps between these windows. A closed form for a background island score distribution is analytically derived, which is then optimized. Both MACS and SICER can optionally call peaks relative to a control dataset, such as input DNA. For scaling the signal and control, both algorithms use the sequencing depth normalization approach as outlined above.

While, depending on parameter settings, many peak calling algorithms can be used both for calling peaks and calling domains, algorithms especially designed for doing both have also been developed. An example is ZINBA [91], which uses zero-inflated negative binomial mixture regression to classify genomic regions into three classes, 1) background, 2) enrichment and 3) artificial zero count. ZINBA can include any covariate in the model, e.g. GC content, mappability to the genome, and control datasets.

1.3.2. ANALYSIS OF GENOME-WIDE SEQUENCING DATA

Using the experimental and computational techniques described above, many interesting results have been obtained, e.g. in distinguishing poised promoters from active promoters [19], integrating signaling pathways with the core transcriptional network in embryonic stem cells [58], and distinguishing poised enhancers from active enhancers [20]. With the increase in compute power, especially interesting has been the recent progress in developing integrative genome segmentation models. These models integrate multiple genome-wide chromatin marks to segment the genome into 'chromatin states'. Chromatin states are characterized by specific combinations of marks occurring simultaneously. For example, by calling peaks on nine chromatin marks in nine cell types and adopting a hidden Markov model approach, ChromHMM, the human genome was segmented into 15 chromatin states. By functionally labeling these chromatin states, the authors systematically characterized regulatory elements, and their cell-type specificities and functional interactions [62]. In another study, 31 ENCODE tracks [92] were used to segment the genome into 25 states using a dynamic Bayesian network method, Segway, results of which positively compared to results obtained using ChromHMM [93]. Whereas the aforementioned studies have relied on ChIP-seq data for segmenting the human genome, a DamID-based genome segmentation of the Drosophila genome has also been published [68]. Here, the authors identified five types of chromatin by segmenting the Drosophila genome using a hidden Markov model approach on 53 DamID-chip profiles, and assessed functional differences between the states. For example, half of the genome was found to be in a repressive state that lacks the characteristic heterochromatic marks.

Moving from single genes and genomic regions in the pre-high-throughput era to arrays and genome-wide sequencing-based profiles, and eventually integrating many genome-wide signals into single models, has made biology and computation increasingly interdependent. It has also illustrated that computation is a key player in moving biology from hypothesis-driven research more towards discovery-driven research, and from a reductionist approach more towards a systems approach. In terms of a muscial analogy, it could be said that biology is starting to hear the music, where previously it could only read the notes in the score [94].

1.4. OUTLINE OF THIS DISSERTATION

1.4.1. CHAPTER 2

As outlined above, DNA integrating elements, such as transposons and retroviruses, are heavily used in many areas of molecular biology, e.g. gene therapy [95, 96], oncogene discovery [35, 37], gene regulation [22, 24], and functional genetics [97, 98]. In Chapter 2 of this dissertation, *Computational identification of insertional mutagenesis targets for cancer gene discovery* (published in Nucleic Acids Research [99]), focus is on the use of transposons and retroviruses for cancer gene discovery from insertional mutagenesis

(IM) screens. In IM screens, transposons and retroviruses perturb the genome and chromatin by randomly integrating into the DNA, e.g. of cancer-predisposed mice. Resulting perturbation of expression of nearby genes can lead to the formation of tumors, which allows to identify putative cancer genes by using the integrations as tags for these genes. In IM screens, putative cancer genes will show more integrations than would be expected based on random chance, and thus may be identified by looking for common integration sites (CISs). For detecting CISs, parametric and nonparametric methods have been used. A parametric method assumes random integration is a stochastic process with a certain underlying analytical probability distribution. For example, it has been assumed that random integration occurrence constitutes a Poisson process [34]. A nonparametric approach does not assume an analytical distribution, but instead estimates a null distribution of integration occurrence by Monte Carlo or permutation-based methods. For example, genome-wide integration probability densities have been estimated using Gaussian kernel convolution, in combination with random permutations to determine significance [100]. After CIS calling, putative cancer genes are identified by manually mapping these CISs to their putative target genes, e.g. [35–37]. These traditional approaches have two main drawbacks. First, the manual mapping of CISs to genes could introduce substantial biases. For example, the mapping may be unintentionally skewed towards well-known cancer genes, thus potentially excluding novel cancer genes. Second, properties of individual integrations are not considered, such as distance to genes and orientation with respect to genes.

Chapter 2 addresses these issues by presenting an alternative to traditional CIS calling, called Kernel Convolved Rule Based Mapping (KC-RBM). The algorithm uses properties of individual integrations, namely distance, orientation and integration density across tumors, to map integrations automatically to putative target genes. We apply our method to two datasets, a Sleeping Beauty transposon (SB) tumor screen and a Murine Leukemia Virus (MuLV) tumor screen, for which same-sample integration data and gene expression data are available. By analyzing the associations between the integrations and gene expression, we gain insight into suitable ranges of parameter values for mapping integrations to putative target genes. By performing an additional aggregation step after the mapping of integrations to genes, putative cancer genes are identified as CTGs (commonly targeted genes). In terms of cancer genes found, the algorithm compares positively to existing approaches for automatically retrieving cancer genes from IM screens.

1.4.2. CHAPTER 3

Integration of elements such as transposons and retroviruses does not occur uniformly random across the genome. There are substantial biases, which may affect research outcomes or complicate interpretation of results in any field using these systems. For example, in cancer gene discovery (Chapter 2), integration biases may be difficult to distinguish from cancer-induced integration hotspots, termed Common integration Sites (CISs). This problem arises from the fact that CIS calling approaches, and also KC-RBM (Chapter 2), often assume a uniform background integration bias in determining the significance of a CIS, e.g. [100–103]. In gene therapy, integration biases are important because therapeutic use of retroviruses can, depending on the locus of integration, cause activation of oncogenes and thus tumor formation [104]. In gene regulation, a strong bias of integrating elements for active chromatin, such as reported for SB, MuLV and the piggyBac transposon (PB) [105–107], may limit studying the chromatin position effect to active chromatin.

In Chapter 3, *Chromatin landscapes of retroviral and transposon integration bias* (published in PLOS Genetics [101]), addresses integration bias for three systems heavily used in many areas of molecular biology: PB, SB, and the mouse mammary tumor virus (MMTV). Multiple studies have reported on biases in one or more of these systems, e.g. [105–111]. While these studies have delivered many new insights into target site selection, they do have some limitations. Examples of these limitations are 1) some datasets were generated using a selectable marker, e.g. [105, 107, 110], 2) datasets were relatively small, e.g. [105, 106, 108], 3) datasets were compared to features in non-matching cell types, e.g. [105], 4) only a limited number of features were analyzed, e.g. [109, 111].

To address these issues, we generated large datasets for these three systems (~120000 to ~180000 integrations), with minimal selective pressure placed upon the integrations, to allow for properly studying the a priori integration bias. The PB and SB datasets were generated in mouse embryonic stem cells (mESCs), enabling us to analyze wide range of (epi)genomic features to characterize, in detail, the preferred chromatin environment of retroviral and transposon integrations in the same cell type. Taking a scale-based approach and using a feature ranking method based on Markov blanket discovery and conditional mutual information, a hierarchical model of integration target site selection is presented: At a domain-oriented scale, target site selection is similar across systems and mostly directed by the same set of (epi)genomic features. At a small scale, notable differences between the systems can be observed, which are characterized by a wide range of features. In addition to characterizing the insertional chromatin environment, the three unselected integration profiles are compared to same-system integrations from IM screens. This shows that 7% - 33% of CISs in these screens may be false positive, i.e. likely the results of a priori integration bias and not of selective pressure (tumor formation).

1.4.3. CHAPTER 4

The DNA integrating elements that were used for perturbing the chromatin in Chapters 2 and 3 can also be used for studying its transcriptional permissiveness, i.e. the chromatin position effect. For this purpose, a DNA integrating element is genetically engineered to contain a promoter and reporter gene of choice. After randomly integrating these elements into the genome, the transcriptional output of the reporter is measured. Although this principle has been used successfully to infer causal relationships between local chromatin context and transcriptional activity [22–24, 112], the studies were generally laborious and of low throughput, tens to hundreds of integration. As such, performing comprehensive genome-wide computational analyses of the chromatin position effect has remained impossible to date.

In Chapter 4, *Chromatin position effects assayed by thousands of reporters integrated in parallel* (published in Cell [113]), a new assay is presented for parallel high-throughput tracking of the expression of thousands of integrated reporter genes. The piggyBac transposon, for which integration bias was studied in Chapter 3, is used as a vector to deliver reporter genes into the genome and study gene regulation by means of the chromatin position effect on a genome-wide scale. Comprehensive computational analyses of ~27000 integrated reporter genes show that lamin associated domains (LADs) repress transcription by attenuation. Limited access of transcription factors to DNA is pointed to as a likely mechanism. Results also show that chromatin compaction is strongly associated with reporter activity, and that on average the influence of enhancers on transcriptional activity extends to ~20kb.

1.4.4. CHAPTER 5

Chapter 5, *Genome-wide profiling of anti-cancer drug effects on chromatin guides rational treatment decisions*, takes a low-level and targeted approach to studying chromatin, by using chemotherapeutic drugs to directly perturb the structure of chromatin. Compared to DNA integrating elements, the mechanistic details by which these drugs perturb chromatin are relatively uniform and well-described. The focus is on drugs that provide their effect by either evicting histones (aclarubicin), inducing DNA breaks (etoposide, topotecan), or doing both of these (daunorubicin). By studying the genomic targeting profiles of these drugs, this chapter provides 1) new insights into chromatin structure, and 2) rationale for improving cancer treatment decisions. More specifically, we observed that the drugs may be used as chemical profiling agents to allow characterization of previously un-annotated genomic regions. Furthermore, all drugs seemed to sense transcriptional activity, and although across drugs similar genomic regions were targeted for DNA breaks, aclarubicin and daunorubicin differed markedly in their genomic preference for histone eviction. Specifically H3K27me3-marked regions were extensively targeted by aclarubicin compared to daunorubicin. As a result, tumor cells with Ezh2 activating mutations, and consequently high levels of H3K27me3, were observed to be highly sensitive to treatment with aclarubicin. Combined, these results show that these anti-cancer drugs are dependent on the local chromatin environment in providing their chemotherapeutic effect, and open up potential for personalized applications of these drugs in the clinic. In order to arrive at the results described above, among others, a new method for normalizing genome-wide sequencing signals with control was developed, and the feature ranking method from Chapter 3 was extended by correcting for the inherent spatial autocorrelation of genome-wide signals.

1.4.5. CHAPTER 6

A central theme of this dissertation is computational epigenomics and data integration. Computational analyses of genome-wide sequencing data, such as those discussed in Section 1.3.2, generally have similar ingredients, e.g. a read mapper, a peak caller, a normalization algorithm, etc. However, these ingredients are often not ideal in the sense that they typically do not give general solutions to the specific problem at hand. Therefore, most analyses are necessarily highly customized. Many of the analyses in this dissertation are also highly customized. Some techniques can be more generally applied, such as a new approach for normalizing genome-wide sequencing data (Chapter 5). These techniques, and some other results, are discussed in more detail in the first part of Chapter 6.

In the second part of Chapter 6, a number of issues are discussed related to the high level of customization of computational analyses of genome-wide sequencing data. For example, some a priori expected characteristics of the genome-wide sequencing profile will bias the choice of peak caller and thus also the results, and read mappers come with many parameters for which it is not always straightforward, or possible, to find a single set of "correct" values. Nevertheless, in order to deal with these issues, taking decisions regarding the analysis pipeline is unavoidable. The careful consideration that must go into taking these decisions is illustrated by demonstrating the substantial impact that these decisions can have on research outcomes.

REFERENCES

[1] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider, *An estimation of the number of cells in the human body.* Ann Hum Biol (2013).

[2] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Harrow, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. J. Searle, *Ensembl 2012,* Nucleic Acids Res **40**, D84 (2012).

[3] J. D. Watson and F. H. C. Crick, *Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid,* Nature **171**, 737 (1953).

[4] K. S. Cho, L. I. Elizondo, and C. F. Boerkoel, *Advances in chromatin remodeling and human disease.* Curr Opin Genet Dev **14**, 308 (2004).

[5] S. Sharma, T. K. Kelly, and P. A. Jones, *Epigenetics in cancer,* Carcinogenesis **31**, 27 (2010).

[6] B. Pang, X. Qiao, L. Janssen, A. Velds, T. Groothuis, R. Kerkhoven, M. Nieuwland, H. Ovaa, S. Rottenberg, O. van Tellingen, J. Janssen, P. Huijgens, W. Zwart, and J. Neefjes, *Drug-induced histone eviction from open chromatin contributes to the chemotherapeutic effects of doxorubicin,* Nature Communications **4**, 1908+ (2013).

[7] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, *Crystal structure of the nucleosome core particle at 2.8 a resolution.* Nature **389**, 251 (1997).

[8] G. Felsenfeld and M. Groudine, *Controlling the double helix,* Nature **421**, 448 (2003).

[9] S. I. Grewal and S. Jia, *Heterochromatin revisited.* Nature reviews. Genetics **8**, 35 (2007).

[10] W. K. Purves, D. Sadava, G. H. Orians, and C. H. Heller, *Life: The Science of Biology, 7th Edition*, 7th ed. (Sinauer Associates and W. H. Freeman, 2003).

[11] H. M. Chan and N. B. La Thangue, *p300/CBP proteins: HATs for transcriptional bridges and scaffolds.* Journal of cell science **114**, 2363 (2001).

[12] M. M. Suzuki and A. Bird, *DNA methylation landscapes: provocative insights from epigenomics,* Nature Reviews Genetics **9**, 465 (2008).

[13] P. J. Mitchell and R. Tjian, *Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins,* Science **245**, 371 (1989).

[14] S. Kato, K. Inoue, and M.-Y. Youn, *Emergence of the osteo-epigenome in bone biology,* IBMS BoneKEy **7**, 314 (2010).

[15] A. J. Bannister and T. Kouzarides, *Regulation of chromatin by histone modifications,* Cell Research **21**, 381 (2011).

[16] T. Jenuwein and C. D. Allis, *Translating the histone code.* Science **293**, 1074 (2001).

[17] X.-J. J. Yang, *Lysine acetylation and the bromodomain: a new partnership for signaling.* BioEssays : news and reviews in molecular, cellular and developmental biology **26**, 1076 (2004).

[18] M. Shogren-Knaak, H. Ishii, J.-M. M. Sun, M. J. Pazin, J. R. Davie, and C. L. Peterson, *Histone H4-K16 acetylation controls chromatin structure and protein interactions.* Science (New York, N.Y.) **311**, 844 (2006).

[19] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O/'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein, *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature **448**, 553 (2007).

[20] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch, *Histone H3K27ac separates active from poised enhancers and predicts developmental state,* P Natl Acad Sci USA **107**, 21931 (2010).

[21] C.-T. Ong and V. G. Corces, *Enhancer function: new insights into the regulation of tissue-specific gene expression,* Nat Rev Genet **12**, 283 (2011).

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

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

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

[25] C. R. Clapier and B. R. Cairns, *The biology of chromatin remodeling complexes,* Annual Review of Biochemistry **78**, 273 (2009).

[26] G. Mizuguchi, X. Shen, J. Landry, W.-H. Wu, S. Sen, and C. Wu, *ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex,* Science **303**, 343 (2004).

[27] M. S. Santisteban, T. Kalashnikova, and M. M. Smith, *Histone H2A.z regulates transcription and is partially redundant with nucleosome remodeling complexes,* Cell **103**, 411 (2000).

[28] E. P. Rogakou, D. R. Pilch, A. H. Orr, V. S. Ivanova, and W. M. Bonner, *DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139,* Journal of Biological Chemistry **273**, 5858 (1998).

[29] A. Saha, J. Wittmeyer, and B. R. Cairns, *Chromatin remodelling: the industrial revolution of DNA around histones,* Nature Reviews Molecular Cell Biology **7**, 437 (2006).

[30] G. Sadri-Vakili, V. Kumaresan, H. D. Schmidt, K. R. Famous, P. Chawla, F. M. Vassoler, R. P. Overland, E. Xia, C. E. Bass, E. F. Terwilliger, R. C. Pierce, and J.-H. J. Cha, *Cocaine-Induced chromatin remodeling increases Brain-Derived neurotrophic factor transcription in the rat medial prefrontal cortex, which alters the reinforcing efficacy of cocaine,* J. Neurosci. **30**, 11735 (2010).

[31] S. S. Newton and R. S. Duman, *Chromatin remodeling: a novel mechanism of psychotropic drug action.* Mol Pharmacol **70**, 440 (2006).

[32] W. Wang, J. Yang, H. Liu, D. Lu, X. Chen, Z. Zenonos, L. S. Campos, R. Rad, G. Guo, S. Zhang, A. Bradley, and P. Liu, *Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog 1,* P Natl Acad Sci USA (2011).

[33] S. D. Farris, E. D. Rubio, J. J. Moon, W. M. Gombert, B. H. Nelson, and A. Krumm, *Transcription-induced chromatin remodeling at the c-myc gene involves the local exchange of histone H2A.z,* J. Biol. Chem. **280**, 25298 (2005).

[34] H. Mikkers, J. Allen, P. Knipscheer, L. Romeijn, A. Hart, E. Vink, A. Berns, and L. Romeyn, *High-throughput retroviral tagging to identify components of specific signaling pathways in cancer.* Nat Genet **32**, 153 (2002).

[35] A. G. Uren, J. Kool, K. Matentzoglu, J. de Ridder, J. Mattison, M. van Uitert, W. Lagcher, D. Sie, E. Tanger, T. Cox, M. Reinders, T. J. Hubbard, J. Rogers, J. Jonkers, L. Wessels, D. J. Adams, M. van Lohuizen, and A. Berns, *Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks.* Cell **133**, 727 (2008).

[36] T. K. Starr, R. Allaei, K. A. T. Silverstein, R. A. Staggs, A. L. Sarver, T. L. Bergemann, M. Gupta, M. G. O'Sullivan, I. Matise, A. J. Dupuy, L. S. Collier, S. Powers, A. L. Oberg, Y. W. Asmann, S. N. Thibodeau, L. Tessarollo, N. G. Copeland, N. A. Jenkins, R. T. Cormier, and D. A. Largaespada, *A Transposon-Based genetic screen in mice identifies genes altered in colorectal cancer,* Science **323**, 1747 (2009).

[37] J. Mattison, J. Kool, A. G. Uren, J. de Ridder, L. Wessels, J. Jonkers, G. R. Bignell, A. Butler, A. G. Rust, M. Brosch, C. H. Wilson, L. van der Weyden, D. A. Largaespada, M. R. Stratton, P. A. Futreal, M. van Lohuizen, A. Berns, L. S. Collier, T. Hubbard, and D. J. Adams, *Novel candidate cancer genes identified by a large-scale cross-species comparative oncogenomics approach.* Cancer Res **70**, 883 (2010).

[38] D. Jahner, H. Stuhlmann, C. L. Stewart, K. Harbers, J. Lohler, I. Simon, and R. Jaenisch, *De novo methylation and expression of retroviral genomes during mouse embryogenesis,* Nature **298**, 623 (1982).

[39] D. Jahner and R. Jaenisch, *Retrovirus-induced de novo methylation of flanking host sequences correlates with gene inactivity,* Nature **315**, 594 (1985).

[40] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray,* Science **270**, 467 (1995).

[41] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark, *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* The New England journal of medicine **351**, 2817 (2004).

[42] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, *Gene expression profiling predicts clinical outcome of breast cancer,* Nature **415**, 530 (2002).

[43] K. J. Kao, K. M. Chang, H. C. Hsu, and A. Huang, *Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization,* BMC Cancer **11**, 143+ (2011).

[44] S. Y. Kim, S. Imoto, and S. Miyano, *Inferring gene networks from time series microarray data using dynamic bayesian networks,* Brief Bioinform **4**, 228 (2003).

[45] Y. Wang, T. Joshi, X.-S. Zhang, D. Xu, and L. Chen, *Inferring gene regulatory networks from multiple microarray datasets,* Bioinformatics **22**, 2413 (2006).

[46] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, *Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation,* Nat Biotechnol **28**, 511 (2010).

[47] H. Ge, K. Liu, T. Juan, F. Fang, M. Newman, and W. Hoeck, *FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution,* Bioinformatics **27**, 1922 (2011).

[48] A. J. Westermann, S. A. Gorski, and J. A. Vogel, *Dual RNA-seq of pathogen and host,* Nat Rev Micro **10**, 618 (2012).

[49] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, *High-resolution mapping and characterization of open chromatin across the genome.* Cell **132**, 311 (2008).

[50] L. Song and G. E. Crawford, *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.* Cold Spring Harbor protocols **2010**, pdb.prot5384+ (2010).

[51] L. Song, Z. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B.-K. K. Lee, N. C. Sheffield, S. Gräf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb, and T. S. Furey, *Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.* Genome research **21**, 1757 (2011).

[52] Y. Blat and N. Kleckner, *Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region,* Cell **98**, 249 (1999).

[53] J. D. Lieb, X. Liu, D. Botstein, and P. O. Brown, *Promoter-specific binding of rap1 revealed by genome-wide maps of protein-DNA association.* Nature genetics **28**, 327 (2001).

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

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

[56] A. Barski, S. Cuddapah, K. Cui, T.-Y. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, *High-resolution profiling of histone methylations in the human genome.* Cell **129**, 823 (2007).

[57] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, *Genome-Wide mapping of in vivo Protein-DNA interactions,* Science **316**, 1497 (2007).

[58] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y.-H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong,

A. Shahab, Y. Ruan, G. Bourque, W.-K. Sung, N. D. Clarke, C.-L. Wei, and H.-H. Ng, *Integration of external signaling pathways with the core transcriptional network in embryonic stem cells,* Cell **133**, 1106 (2008).

[59] M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schubeler, *DNA-binding factors shape the mouse methylome at distal regulatory regions,* Nature **480**, 490 (2011).

[60] W. A. Whyte, S. Bilodeau, D. A. Orlando, H. A. Hoke, G. M. Frampton, C. T. Foster, S. M. Cowley, and R. A. Young, *Enhancer decommissioning by LSD1 during embryonic stem cell differentiation,* Nature **advance online publication** (2012).

[61] L. Handoko, H. Xu, G. Li, C. Y. Y. Ngan, E. Chew, M. Schnapp, C. W. H. W. Lee, C. Ye, J. L. H. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W.-K. K. Sung, Y. Ruan, and C.-L. L. Wei, *CTCF-mediated functional chromatin interactome in pluripotent cells.* Nat Genet **43**, 630 (2011).

[62] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein, *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature **473**, 43 (2011).

[63] G. Ficz, M. R. Branco, S. Seisenberger, F. Santos, F. Krueger, T. A. Hore, C. J. Marques, S. Andrews, and W. Reik, *Dynamic regulation of 5-hydroxymethylcytosine in mouse es cells and during differentiation.* Nature **473**, 398 (2011).

[64] Y. Xu, F. Wu, L. Tan, L. Kong, L. Xiong, J. Deng, A. J. Barbera, L. Zheng, H. Zhang, S. Huang, J. Min, T. Nicholson, T. Chen, G. Xu, Y. Shi, K. Zhang, and Y. G. G. Shi, *Genome-wide regulation of 5hmC, 5mC, and gene expression by tet1 hydroxylase in mouse embryonic stem cells.* Mol Cell **42**, 451 (2011).

[65] R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, and J. A. Stamatoyannopoulos, *Sequencing newly replicated DNA reveals widespread plasticity in human replication timing,* Proceedings of the National Academy of Sciences **107**, 139 (2009).

[66] A. M. Szalkowski and C. D. Schmid, *Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts.* Briefings in bioinformatics **12**, 626 (2011).

[67] D. Peric-Hupkes, W. Meuleman, L. Pagie, S. W. M. Bruggeman, I. Solovei, W. Brugman, S. Gräf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, M. Reinders, L. Wessels, and B. van Steensel, *Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.* Mol Cell **38**, 603 (2010).

[68] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, and B. van Steensel, *Systematic protein location mapping reveals five principal chromatin types in drosophila cells.* Cell **143**, 212 (2010).

[69] J. G. van Bemmel, G. J. Filion, A. Rosado, W. Talhout, M. de Haas, T. van Welsem, F. van Leeuwen, and B. van Steensel, *A network model of the molecular organization of chromatin in drosophila.* Molecular cell **49**, 759 (2013).

[70] S. D. Luo, G. W. Shi, and B. S. Baker, *Direct targets of the d. melanogaster dsxf protein and the evolution of sexual development.* Development **138**, 2761 (2011).

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

[72] J. G. van Bemmel, L. Pagie, U. Braunschweig, W. Brugman, W. Meuleman, R. M. Kerkhoven, and B. van Steensel, *The insulator protein SU(HW) Fine-Tunes nuclear lamina interactions of the drosophila genome,* PLoS ONE **5**, e15013+ (2010).

[73] H. Yin, S. Sweeney, D. Raha, M. Snyder, and H. Lin, *A High-Resolution Whole-Genome map of key chromatin modifications in the adult drosophila melanogaster,* PLoS Genet **7**, e1002380+ (2011).

[74] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, *Comprehensive mapping of long-range interactions reveals folding principles of the human genome,* Science **326**, 289 (2009).

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

[76] M. Quail, M. Smith, P. Coupland, T. Otto, S. Harris, T. Connor, A. Bertoni, H. Swerdlow, and Y. Gu, *A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers,* BMC Genomics **13**, 341+ (2012).

[77] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,* Genome Biol **10**, R25 (2009).

[78] H. Li and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform,* Bioinformatics **25**, 1754 (2009).

[79] S. G. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj, and M. F. Polz, *PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample.* Applied and environmental microbiology **71**, 8966 (2005).

[80] Y. Benjamini and T. P. Speed, *Summarizing and correcting the GC content bias in high-throughput sequencing.* Nucleic acids research **40**, e72 (2012).

[81] T. Derrien, J. Estellé, S. M. Sola, D. G. Knowles, E. Raineri, R. Guigó, and P. Ribeca, *Fast computation and applications of genome mappability,* PLoS ONE **7**, e30377+ (2012).

[82] S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder, *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia,* Genome Research **22**, 1813 (2012).

[83] T. Bailey, P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim, and J. Zhang, *Practical guidelines for the comprehensive analysis of ChIP-seq data,* PLoS Comput Biol **9**, e1003326+ (2013).

[84] Z. Shao, Y. Zhang, G.-C. C. Yuan, S. H. Orkin, and D. J. Waxman, *MAnorm: a robust model for quantitative comparison of ChIP-seq data sets.* Genome biology **13**, R16+ (2012).

[85] C. Taslim, J. Wu, P. Yan, G. Singer, J. Parvin, T. Huang, S. Lin, and K. Huang, *Comparative study on ChIP-seq data: normalization and binding pattern characterization,* Bioinformatics **25**, 2334 (2009).

[86] N. U. Nair, A. D. Sahu, P. Bucher, and B. M. E. Moret, *Chipnorm: a statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries.* PLoS One **7**, e39573 (2012).

[87] K. Liang and S. Keles, *Normalization of ChIP-seq data with control,* BMC Bioinformatics **13**, 199+ (2012).

[88] A. Diaz, K. Park, D. A. Lim, and J. S. Song, *Normalization, bias correction, and peak calling for ChIP-seq.* Statistical applications in genetics and molecular biology **11** (2012).

[89] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nusbaum, R. Myers, M. Brown, W. Li, and X. S. Liu, *Model-based analysis of ChIP-seq (MACS),* Genome Biol **9**, R137+ (2008).

[90] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng, *A clustering approach for identification of enriched domains from histone modification ChIP-seq data.* Bioinformatics (Oxford, England) **25**, 1952 (2009).

[91] N. Rashid, P. Giresi, J. Ibrahim, W. Sun, and J. Lieb, *ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions,* Genome Biology **12**, R67+ (2011).

[92] J. Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, E. Giste, A. Johnson, M. Zhang, G. Balasundaram, R. Byron, V. Roach, P. Sabo, R. Sandstrom, A. S. Stehling, R. Thurman, S. Weissman, P. Cayting, M. Hariharan, J. Lian, Y. Cheng, S. Landt, Z. Ma, B. Wold, and J. Dekker, *An encyclopedia of mouse DNA elements (mouse ENCODE),* Genome Biol **13**, 418+ (2012).

[93] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, *Unsupervised pattern discovery in human chromatin structure through genomic segmentation,* Nat Meth **9**, 473 (2012).

[94] C. R. Woese, *A new biology for a new century,* Microbiol Mol Biol Rev **2**, 173 (2004).

[95] N. Cartier, S. Hacein-Bey-Abina, C. C. Bartholomae, G. Veres, M. Schmidt, I. Kutschera, M. Vidaud, U. Abel, L. Dal-Cortivo, L. Caccavelli, N. Mahlaoui, V. Kiermer, D. Mittelstaedt, C. Bellesme, N. Lahlou, F. Lefrere, S. Blanche, M. Audit, E. Payen, P. Leboulch, B. l'Homme, P. Bougneres, C. Von Kalle, A. Fischer, M. Cavazzana-Calvo, and P. Aubourg, *Hematopoietic stem cell gene therapy with a lentiviral vector in X-Linked adrenoleukodystrophy,* Science **326**, 818 (2009).

[96] A. Fischer, S. Hacein-Bey-Abina, and M. Cavazzana-Calvo, *20 years of gene therapy for SCID,* Nat Immunol **11**, 457 (2010).

[97] E. H. Miller, G. Obernosterer, M. Raaben, A. S. Herbert, M. S. Deffieu, A. Krishnan, E. Ndungo, R. G. Sandesara, J. E. Carette, A. I. Kuehne, G. Ruthel, S. R. Pfeffer, J. M. Dye, S. P. Whelan, T. R. Brummelkamp, and K. Chandran, *Ebola virus entry requires the host-programmed recognition of an intracellular receptor,* EMBO J (2012).

[98] P. Bouwman, A. Aly, J. M. Escandell, M. Pieterse, J. Bartkova, H. van der Gulden, S. Hiddingh, M. Thanasoula, A. Kulkarni, Q. Yang, B. G. Haffty, J. Tommiska, C. Blomqvist, R. Drapkin, D. J. Adams, H. Nevanlinna, J. Bartek, M. Tarsounas, S. Ganesan, and J. Jonkers, *53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers,* Nat Struct & Mol Biol **17**, 688 (2010).

[99] J. de Jong, J. de Ridder, L. van der Weyden, N. Sun, M. van Uitert, A. Berns, M. van Lohuizen, J. Jonkers, D. J. Adams, and L. F. A. Wessels, *Computational identification of insertional mutagenesis targets for cancer gene discovery,* Nucleic Acids Res **39**, e105 (2011).

[100] J. de Ridder, A. Uren, J. Kool, M. Reinders, and L. Wessels, *Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens.* PLoS Comput Biol **2** (2006).

[101] J. de Jong, W. Akhtar, J. Badhai, A. G. Rust, R. Rad, J. Hilkens, A. Berns, M. van Lohuizen, L. F. A. Wessels, and J. de Ridder, *Chromatin landscapes of retroviral and transposon integration profiles,* PLoS Genet **10**, e1004250+ (2014).

[102] V. W. Keng, A. Villanueva, D. Y. Chiang, A. J. Dupuy, B. J. Ryan, I. Matise, K. A. Silverstein, A. Sarver, T. K. Starr, K. Akagi, L. Tessarollo, L. S. Collier, S. Powers, S. W. Lowe, N. A. Jenkins, N. G. Copeland, J. M. Llovet, and D. A. Largaespada, *A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma.* Nature biotechnology **27**, 264 (2009).

[103] T. L. Bergemann, T. K. Starr, H. Yu, M. Steinbach, J. Erdmann, Y. Chen, R. T. Cormier, D. A. Largaespada, and K. A. T. Silverstein, *New methods for finding common insertion sites and co-occurring common insertion sites in transposon- and virus-based genetic screens.* Nucleic Acids Res **40**, 3822 (2012).

[104] S. Hacein-Bey-Abina, A. Garrigue, G. P. Wang, J. Soulier, A. Lim, E. Morillon, E. Clappier, L. Caccavelli, E. Delabesse, K. Beldjord, V. Asnafi, E. MacIntyre, L. Dal Cortivo, I. Radford, N. Brousse, F. Sigaux, D. Moshous, J. Hauer, A. Borkhardt, B. H. Belohradsky, U. Wintergerst, M. C. Velez, L. Leiva, R. Sorensen, N. Wulffraat, S. Blanche, F. D. Bushman, A. Fischer, and M. Cavazzana-Calvo, *Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of scid-x1,* J Clin Invest **118**, 3132 (2008).

[105] C. Berry, S. Hannenhalli, J. Leipzig, and F. D. Bushman, *Selection of target sites for mobile DNA integration in the human genome,* PLoS Comput. Biol **2**, e157 (2006).

[106] Q. Liang, J. Kong, J. Stalker, and A. Bradley, *Chromosomal mobilization and reintegration of sleeping beauty and piggybac transposons.* Genesis **47**, 404 (2009).

[107] M. A. Li, S. J. Pettitt, S. Eckert, Z. Ning, S. Rice, J. Cadiñanos, K. Yusa, N. Conte, and A. Bradley, *The piggybac transposon displays local and distant reintegration preferences and can cause mutations at non-canonical integration sites,* Mol Cell Biol (2013).

[108] B. Balu, C. Chauhan, S. Maher, D. Shoue, J. Kissinger, M. Fraser, and J. Adams, *piggyBac is an effective tool for functional analysis of the plasmodium falciparum genome,* BMC Microbiol **9**, 83+ (2009).

[109] D. L. Galvan, Y. Nakazawa, A. Kaja, C. Kettlun, L. J. Cooper, C. M. Rooney, and M. H. Wilson, *Genome-wide mapping of PiggyBac transposon integrations in primary human T cells.* J Immunother **32**, 837 (2009).

[110] W. Wang, C. Lin, D. Lu, Z. Ning, T. Cox, D. Melvin, X. Wang, A. Bradley, and P. Liu, *Chromosomal transposition of PiggyBac in mouse embryonic stem cells,* P Natl Acad of Sci USA **105**, 9290 (2008).

[111] A. Faschinger, F. Rouault, S. Johannes, A. Lukas, B. Salmons, W. H. Günzburg, and S. Indik, *Mouse Mammary Tumor Virus integration site selection in human and mouse genomes,* J Virol **82**, 1360 (2007).

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

[113] W. Akhtar, J. de Jong, A. V. Pindyurin, L. Pagie, W. Meuleman, J. de Ridder, A. Berns, L. F. A. Wessels, M. van Lohuizen, and B. van Steensel, *Chromatin position effects assayed by thousands of reporters integrated in parallel.* Cell **154**, 914 (2013).

2

COMPUTATIONAL IDENTIFICATION OF INSERTIONAL MUTAGENESIS TARGETS FOR CANCER GENE DISCOVERY

Johann DE JONG
Jeroen DE RIDDER
Louise VAN DER WEYDEN
Ning SUN
Miranda VAN UITERT
Anton BERNS
Maarten VAN LOHUIZEN
Jos JONKERS
David J. ADAMS
Lodewyk F.A. WESSELS

Parts of this chapter have been published in Nucleic Acids Research [1].

ABSTRACT

Insertional mutagenesis is a potent forward genetic screening technique used to identify candidate cancer genes in mouse model systems. An important, yet unresolved issue in the analysis of these screens, is the identification of the genes affected by the insertions. To address this, we developed Kernel Convolved Rule Based Mapping (KC-RBM). KC-RBM exploits distance, orientation and insertion density across tumors to automatically map integration sites to target genes. We perform the first genome-wide evaluation of the association of insertion occurrences with aberrant gene expression of the predicted targets in both retroviral and transposon datasets. We demonstrate the efficiency of KC-RBM by showing its superior performance over existing approaches in recovering true positives from a list of independently, manually curated cancer genes. The results of this work will significantly enhance the accuracy and speed of cancer gene discovery in forward genetic screens. KC-RBM is available as R-package.

2.1. INTRODUCTION

Large-scale insertional mutagenesis screens using retroviruses and transposons are of great importance in cancer research. By integration into the host DNA, retroviruses and transposons can mutate the genome. This process is referred to as insertional mutagenesis. Insertional mutagenesis can disrupt cellular processes, alter gene expression and thereby cause cancer. For this reason, large-scale insertional mutagenesis screens have been successfully employed to identify new putative cancer genes, see e.g. [2–7]. In addition, retroviral vectors have been shown to be useful in gene therapy, and transposon based systems also show great potential for this same purpose. However, it is currently still very difficult to predict which surrounding genes will be affected by insertions.

To identify potential cancer genes from an insertional mutagenesis screen, the initial step typically involves the definition of common insertion sites (CISs), see e.g. [2, 4, 8–11]. Insertions are clustered based on inter-insertion distance and clusters that are unlikely to occur by chance are declared CISs. The CISs are then manually mapped to putative target genes. This manual mapping could potentially introduce biases. For example, known cancer genes may be preferred, thus potentially and unintentionally excluding novel cancer genes. An additional drawback of this approach is that in defining CISs, properties of individual insertions, such as distances to genes and orientation relative to genes are disregarded.

In contrast, nearest-gene mapping (NGM), maps each insertion to the nearest gene (e.g. [12]). While this approach does operate on individual insertions, and takes the distance of insertions to genes into account, it still disregards the relative orientation of insertions, and does not aggregate insertion data across tumors.

The orientation of an insertion occurring in the immediate upstream promotor region of a gene is a highly important modulator of the effect of that insertion on the gene. More specifically, if the viral promoter has the same orientation as the host promoter it can take over its function [6]. For larger upstream and downstream distances from genes, relative orientation also plays a role: enhancing insertions are predominantly oriented away from target genes [6]. It is therefore clear that the orientation of an insertion should be taken into account when determining putative target genes. Furthermore, since the nearest gene is not necessarily the only or best target gene, it is important to allow the assignment of multiple target genes to a single insertion.

To address the issues described above, we developed Kernel Convolved Rule Based Mapping (KC-RBM). KC-RBM integrates GKC [9], a method for identifying statistically significant CISs, with rule based mapping of individual insertions to genes. Without user intervention, KC-RBM maps insertions to genes based on orientation-dependent windows defined around transcripts, and exploits the information contained in the repetitive occurrence of insertions at a given locus across tumors, i.e. CIS information. We perform extensive analyses of associations between insertion occurrence and same-sample gene expression to evaluate the parameter choices for KC-RBM. We demonstrate the benefits of KC-RBM in cancer gene discovery through the more accurate identification of target genes from two insertional mutagenesis screens, a Murine Leukemia Virus (MuLV) screen and a Sleeping Beauty (SB) transposon screen. KC-RBM represents the first ever approach for mapping SB transposon insertions to target genes.

2.2. RESULTS

2.2.1. INSERTION OCCURRENCE AND GENE EXPRESSION

Since the orientation of an insertion relative to a target gene and the distance of an insertion to a target gene determine how an insertion may activate that gene, one may expect association between orientation, distance, and gene expression. Figure 2.1 depicts an alignment of all genes. A point represents the average normalized deviation of the gene expression from the mean as a function of the distance between a gene and an insertion. A technical explanation of this figure can be found in the Methods section. For both the MuLV insertions and the SB transposon insertions, it can be seen that association of insertion occurrence with gene expression of nearby genes is dependent on the distance and orientation of the insertion to the gene.

For MuLV, Figure 2.1a shows higher average expression deviation for insertions inside and near genes (demarcated by the two vertical black lines). There is a clear peak in expression levels for antisense insertions (red) just upstream of the gene start site. For sense insertions (green), a slightly less pronounced peak can be seen just downstream of the gene start site. These observations are consistent with mechanisms described in the literature by which retroviral insertions affect their target genes [6, 7, 13]. The insertion density across all aligned genes is plotted in black below the expression values, and demonstrates an explicit preference of MuLV insertions for loci near the gene start site, as previously observed [14].

For the SB transposon, Figure 2.1b suggests that some association does exist, although much less pronounced when compared to the retroviral case. Especially for the sense insertions inside genes there is some elevation in expression. The insertion density (depicted in black, below the binned z-values) shows that SB transposon insertions are predominantly found inside genes.

2.2.2. KC-RBM

Mechanisms described in the literature [6, 7, 13], supported by our own observations (Figure 2.1), suggest that insertions should be mapped to putative target genes using

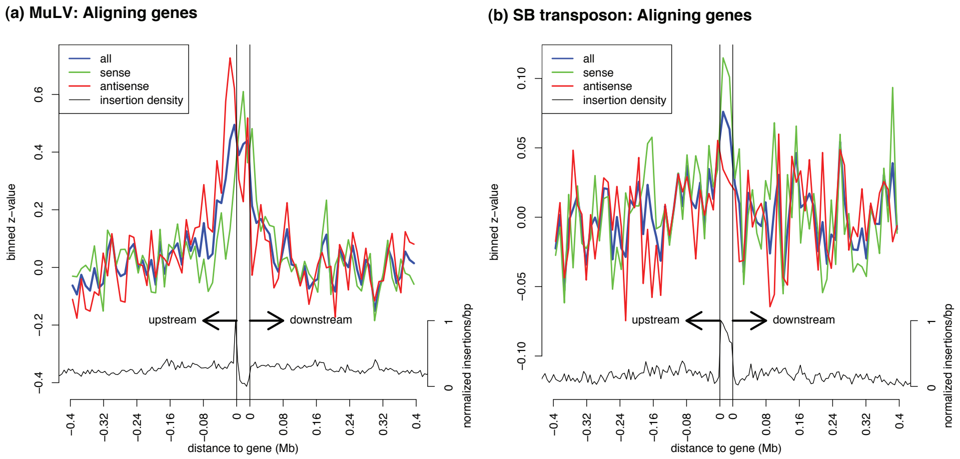


Figure 2.1: Normalized deviation of gene expression from the mean as a function of insertion distance, for (a) MuLV and (b) SB transposon. For all genes, all insertions are identified in a window of 500kb around these genes, from the gene start site and from the gene termination site. All genes are then aligned with respect to location as well as orientation. z-normalized gene expression values are associated with the relative locations of the insertions within the 500kb window. For all genes and insertions taken together, these expression values are binned, and the distinction is made between insertions occurring in sense orientation relative to the gene (green) and in antisense orientation relative to the gene (red). The blue line represents the all insertions taken together. The (aligned) insertion density is plotted in black below the binned z-values. The gene boundaries are represented by two vertical black lines.

orientation dependent windows defined on either side of transcripts. Depending on the orientation and location of an insertion, the insertion will fall within or outside the relevant mapping window. When the insertion falls within a given window, it will be mapped to the associated gene. This approach, which we will call Rule Based Mapping (RBM) is outlined in Figure 2.2a. It uses four window size parameters, for upstream-sense, upstream-antisense, downstream-sense, and downstream-antisense insertions.

The window sizes used by RBM provide strict boundaries outside of which insertions are not mapped to a gene. However, as it is presented in Figure 2.2a, RBM does not directly exploit the fact that information from across tumor samples is available. After all, cancer genes harbor mutations across many independent tumors. Furthermore, it might be that, in an insertion cluster, a minority of insertions occur that contradict the window sizes set for RBM. As an example, consider a cluster of insertions, a CIS. Suppose that a number of these insertions lie outside the mapping window relative to a certain gene, and the other insertions lie within the mapping window. RBM will not map the insertions outside the mapping window to the gene. However, since the insertions constitute a cluster, it is not unreasonable to assume that all these insertions will all target the same gene. As another example, consider again a cluster of insertions. Let us suppose that a large majority of the insertions have a sense orientation relative to a target gene, and just a few insertions are oriented antisense. Here it will again make sense to map the cluster as a whole to the same target gene, thereby disregarding the antisense orientation of a small minority of insertions.

The implication of these two examples is that it is sensible to allow exceptions to the strict application of the rules, when this is suggested by information regarding the frequency and orientation of insertions across tumors. We therefore propose a hybrid approach, involving RBM and GKC, to exploit information from across tumor samples to flexibly apply RBM in a data-driven manner. Recall that GKC [9] detects CISs by estimating the insertion density through a Gaussian kernel convolution and identifying insertion hot spots based on a random permutation approach. The hybrid approach will be referred to as KC-RBM, and is illustrated in Figure 2.2b. First, given an insertion profile, a Gaussian kernel convolution is applied to estimate the insertion density, essentially defining clusters of insertions. Second, all insertions are associated with their nearest peak. This results in a number of insertion clusters, one for each peak. Third, if a cluster is orientation-wise homogeneous enough, all individual insertions are merged into a single orientation cluster, otherwise insertions are separated into a sense and an antisense cluster. The positions of the resulting clusters are taken to be the average position of the insertions constituting that cluster. Fourth, all clusters mean loci are mapped using RBM.

In addition to the four window sizes, KC-RBM depends on two parameters: one for determining the level of smoothing of the positions of the insertions, and one for determining the orientation homogeneity of a cluster. The parameter that determines the smoothing is the standard deviation of the Gaussian kernel, and is called the scale parameter. The parameter that controls the orientation homogeneity of a cluster is defined as the minimal fraction of the insertions constituting a cluster that need to have the same orientation. The kernel width reflects the degree of strictness with which one wishes to enforce the mapping window: the smaller the scale, the less flexibility is allowed in the chosen sizes of the mapping windows. The orientation homogeneity parameter controls the level of noise tolerated in the insertion orientation: the higher the orientation homogeneity fraction, the higher the stringency on the orientation of insertions.

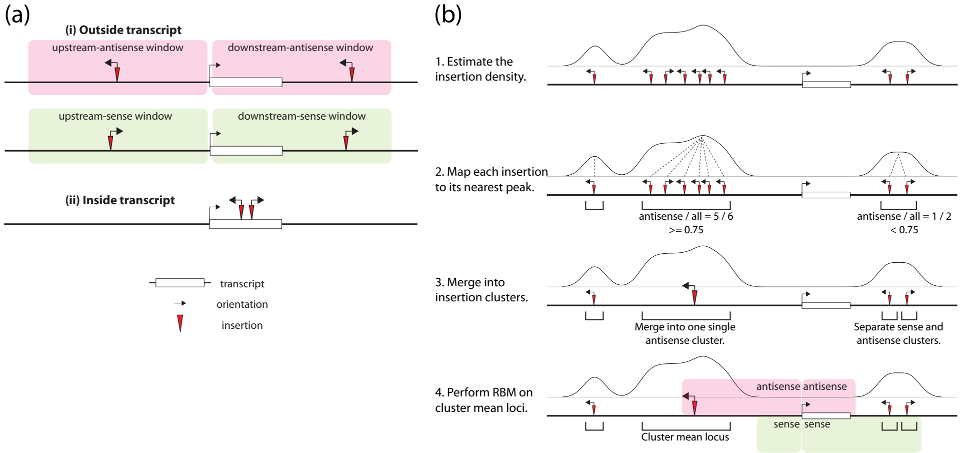

(a)
(i) Outside transcript
upstream-antisense window
downstream-antisense window
upstream-sense window
downstream-sense window
(ii) Inside transcript
transcript
orientation
insertion
(b)
1. Estimate the insertion density.
2. Map each insertion to its nearest peak.
antisense / all = 5 / 6
>= 0.75
antisense / all = 1 / 2
< 0.75
3. Merge into insertion clusters.
Merge into one single antisense cluster.
Separate sense and antisense clusters.
4. Perform RBM on cluster mean loci.
antisense antisense
sense sense
Cluster mean locus

Figure 2.2: (a) RBM for mapping insertions to genes. Distinctions are made based on three properties. Insertions are distinguished by occurrence (i) outside or (ii) inside a transcript, upstream or downstream of a transcript, and in sense or antisense orientation with respect to the orientation of the transcript. (b) KC-RBM for mapping insertions to transcripts. First, given an insertion profile, a Gaussian kernel convolution is applied to estimate the insertion density. Second, all insertions are associated with their nearest peak. This results in a number of insertion clusters, one for each peak. Third, if the cluster is orientation-wise homogeneous enough, all individual insertions are merged into a single-orientation cluster, otherwise insertions are separated into a sense and an antisense cluster. Fourth, all clusters mean loci are mapped using RBM. Finally, a gene is considered a target of an insertion if at least one of its transcripts is a target.

2.2.3. VARYING THE WINDOW SIZES

This section explores the influence of varying the four window size parameters on insertion-expression association, while setting the scale parameter to 0 and the orientation homogeneity parameter to 1.0, i.e. strictly applying the mapping window widths and without smoothing the insertion orientation. When compared to the analysis represented in Figure 2.1, this analysis is more refined in that for a particular value of a window size parameter, a Wilcoxon test is performed to determine whether the median difference between the expression of samples with and without a given insertion is significantly different from zero (for more detail, please refer to the Methods section).

For MuLV, Figure 2.3a shows the influence of varying the window sizes on insertion-expression association, as measured by the fraction of significant associations (true positive rate) and the number of detected significant associations (number of true positives). In Figure 2.3a(i), cases with at least one insertion per gene across samples were taken into account. The insertions oriented away from genes, upstream-antisense and downstream-sense, show the largest association (large fraction of significant genes). This is in concordance with the literature, where these cases are often denoted as enhancer insertions, and can activate genes across large distances [6, 7, 13]. In contrast, the association of upstream-sense insertions is very local. This is also in concordance with the literature, where these insertions are often denoted as promoter insertions [6, 7, 13]. Downstream-antisense insertions show the least association. However, there is a clear association for small window sizes (<10kb).

In contrast to Figure 2.3a(i), in Figure 2.3a(ii) we only included genes with at least two insertions per gene across all samples. This shows that, in general, insertion occurrence associates even better with elevated expression levels, although for small downstream window sizes (<20kb) the data are very sparse. In both these figures the association of downstream antisense insertions is less pronounced than that of downstream sense insertions.

For the SB transposon, Figure 2.3b shows the influence of different window sizes on insertion-expression association. Also in this figure it is evident that the association is far less pronounced than for the retroviral case, although it can be seen that SB transposon insertions are predominantly found inside genes. Requiring at least one insertion per gene across samples gives a noisy result and only shows a slight association for sense insertions. Requiring at least two insertions (Figure 2.3b(ii)) gives a clearer picture, and results in higher fractions of significant genes. It shows that the insertion-expression association for the SB transposon is far more localized and less pronounced when compared to retroviral insertions. Again, the sense insertions associate better with increased expression. For both downstream-antisense insertions and downstream-sense insertions, the data are too sparse to draw meaningful conclusions.

One remark should be made on the visual presentation in Figure 2.3. To allow for a 2D presentation, the four window types and the within-transcript case are treated separately. i.e. when computing the statistic for one specific window size, the other three are set to zero. A more comprehensive view is offered in a more complex 4D visualization in the Supplementary Material (Figures 7, 8, 11 and 12), but does not lead to different observations.

2.2.4. VARYING THE SMOOTHING

This section explores the influence of varying the Gaussian Kernel Convolution scale parameter on insertion-expression association, while keeping the orientation homogeneity parameter and the four window sizes constant. Figure 2.3 showed there are substantial differences in strength of insertion-expression association for the four window size parameters. Therefore, while varying the smoothing, these window sizes are fixed to values reflecting this relative strength of insertion-expression association. This implies that the upstream-antisense window (ua) is the largest window, followed by the downstream-sense (ds) window, the upstream-sense (us) window, and the downstream-antisense (da) window respectively. Furthermore, window sizes are chosen such that the fraction of significant genes never falls below the permutation threshold of 5%, and a particular window size is never too small to retrieve at least one significant gene in that window. This resulted in the window sizes (us,ua,ds,da) = (20kb,120kb,40kb,5kb). Furthermore the orientation homogeneity minimal fraction was set to 0.75.

For these parameter values, the influence of the scale parameter on the number of significant genes and the fraction of significant genes is visualized in Figure 2.4. For

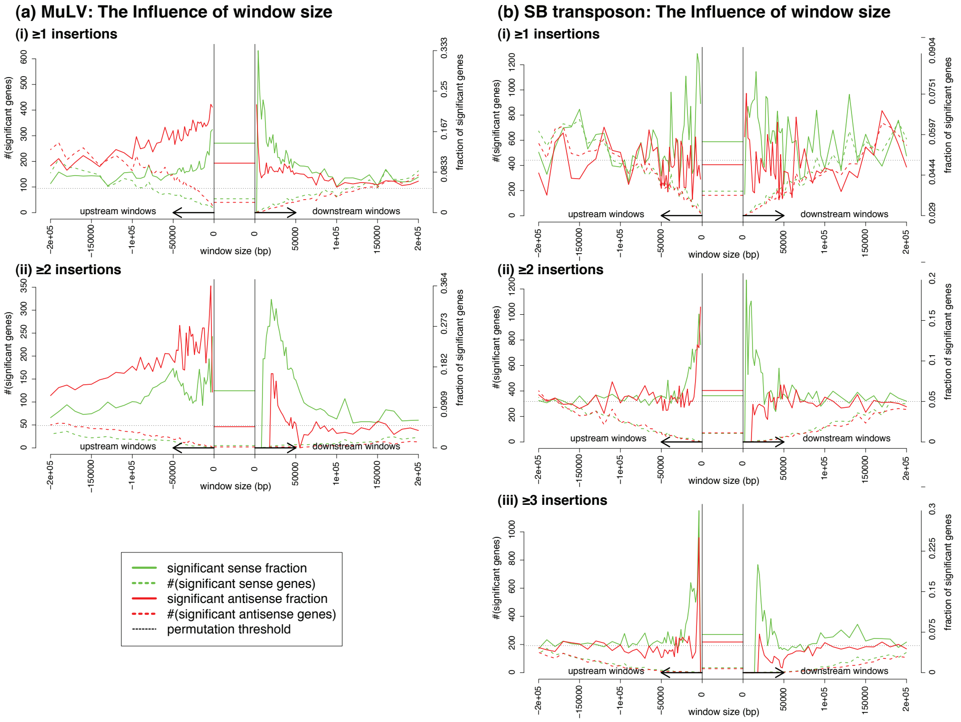


Figure 2.3: The influence of the four window sizes on mapping quality for (a) MuLV and (b) the SB transposon, for requiring (i) at least one insertion per gene across samples, (ii) at least two insertions per gene across samples, or (iii) at least three insertions per gene across samples. A distinction is made between upstream-sense, upstream-antisense, downstream-sense, and downstream-antisense windows. For an explanation of the computation of the fractions and numbers of significant genes (true positive rate and number of true positives respectively) refer to the Methods section. First, computation is done for multiple window sizes. Second, when computing the statistics for one of four window sizes, the other window sizes are set to zero. Third, for all insertions within transcripts association of insertion occurrence with increased expression levels was computed while disregarding the insertions outside transcripts. Permutation thresholds (5%, represented by the dotted black line) were calculated per window size.

the MuLV dataset, KC-RBM achieves stronger associations if smoothing is applied, for all scales larger than 5kb, and especially for scales around 10kb and 35kb (Figure 2.4a). For scales larger than 35kb the performance deteriorates, with the number of significant genes increasing and the fraction of significant genes decreasing.

For the SB transposon, Figure 2.4b again shows that the association of SB transposon insertions with increased gene expression is less pronounced than in the case of MuLV. However, the strongest association is attained for small scales around 2kb. Furthermore, similar to the retroviral case, the number and fraction of significant genes diverges for larger scales. Hence we fixed the the kernel width for MuLV at 10kb and for the SB transposon at 2kb.

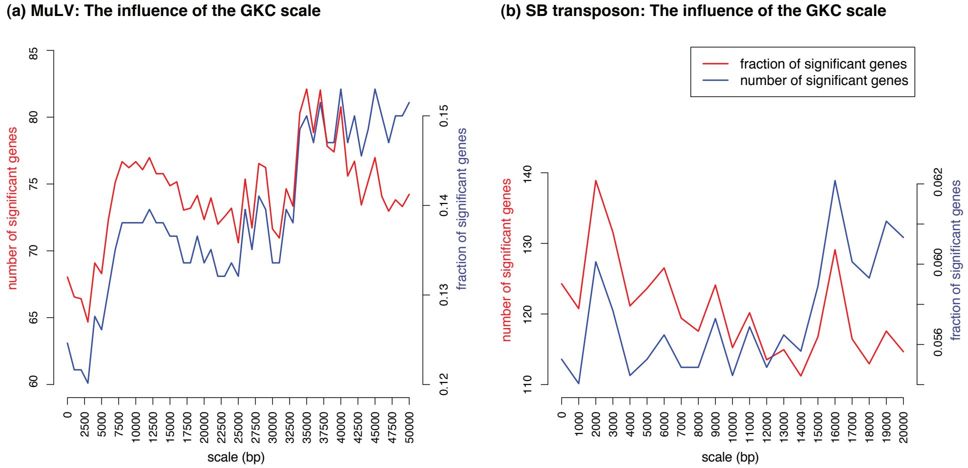


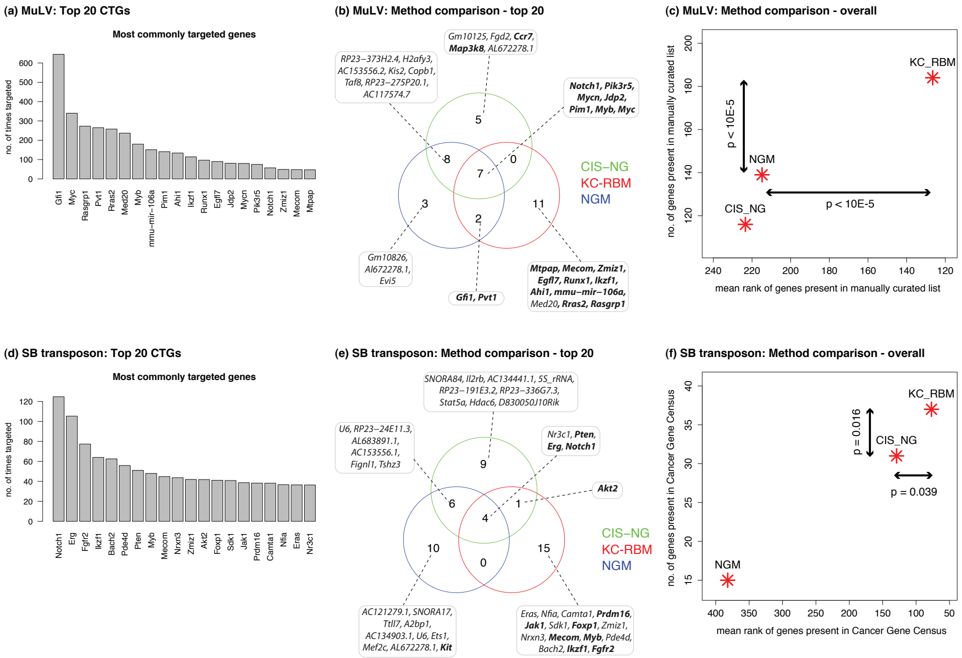
Figure 2.4: The influence of the GKC scale parameter on the number of true positives and the true positive rate, for performing RBM on (a) the MuLV data and (b) the SB transposon data using window sizes (20kb,120kb,40kb,5kb) (MuLV) and (20kb,10kb,25kb,5kb) (SB transposon) for (upstream-sense, upstream-antisense, downstream-sense, downstream-antisense) insertions. The orientation homogeneity parameter was set to 0.75.

2.2.5. Using KC-RBM for cancer gene identification

In previous sections we have investigated how the association between insertion presence and gene expression is modulated by the parameters of KC-RBM (window sizes and kernel width for a fixed homogeneity parameter), and fixed these parameters to appropriate values. Now, we will demonstrate how KC-RBM is employed for the identification of cancer genes based on the insertion data only.

Recall that for each insertion, KC-RBM identifies a list of putative target genes. However, for a given insertion, not all identified targets may be of equal importance. As a first step in extracting interesting genes from a KC-RBM mapping, we will select at most a single target for each insertion. More specifically, among the targets identified per insertion, we rank the genes according to the number of times they were targeted across insertions, and select the top ranking gene as the single target for that insertion. This selects for genes frequently targeted across insertions. The set of all targeted genes can subsequently be ranked by counting for each gene the number of times that gene was targeted by an insertion, resulting in a list of commonly targeted genes (CTGs).

Figure 2.5a shows a top 20 CTG list for the MuLV dataset, obtained by applying the procedure described above. Figure 2.5d shows the results for the SB transposon dataset.


(a) MuLV: Top 20 CTGs

(b) MuLV: Method comparison - top 20

(c) MuLV: Method comparison - overall

(d) SB transposon: Top 20 CTGs

(e) SB transposon: Method comparison - top 20

(f) SB transposon: Method comparison - overall

Figure 2.5: Comparison between genes identified by CTG mapping and by CIS mapping for MuLV and SB transposon. The top 20 CTGs identified by KC-RBM for (a) MuLV and (d) SB transposon; on the x-axis the gene symbol, on the y-axis the number of times a certain gene was identified as a target. KC-RBM was performed using window sizes (20kb,120kb,40kb,5kb) for the MuLV data and (20kb,10kb,25kb,5kb) for the SB transposon data, for upstream-sense, upstream-antisense, downstream-sense, and downstream-antisense windows respectively. The scale parameters were set to 10kb (MuLV) and 2kb (SB transposon). The orientation homogeneity parameter in both cases was set to 0.75. Venn diagrams for both (b) MuLV and (e) SB transposon depicting the overlap between the top 20 CTGs identified by KC-RBM-CTG mapping, and the top 20 CIS-nearest-genes. Gene names in bold face refer to genes that were also identified in the manually curated list of 346 CISs [4] (MuLV) or in the Cancer Gene Census (SB transposon). For (c) MuLV and (f) SB transposon, the complete lists of CTGs are also more extensively compared to the manually curated set (MuLV) and the Cancer Gene Census list (B transposon), again taking the manually curated list as a reference. This shows that KC-RBM performs better than both CIS-NG and NGM.

2.2.6. EVALUATING KC-RBM

In this section we will evaluate the effectiveness of KC-RBM in identifying cancer genes by following the steps described in the previous section, and comparing the results obtained with two other methods for computationally identifying cancer genes from insertional mutagenesis screens. The first method performs a GKC-based CIS analysis [9], and then maps each CIS peak to the nearest gene. We will refer to this method as CIS-nearest-gene mapping (CIS-NG). The second method consists of mapping each insertion to the nearest gene, and then determining the CTGs. We will refer to this method as nearest-gene mapping (NGM). The results obtained by these three methods will be evaluated by comparing them to manually curated lists of cancer genes.

For MuLV, a manually curated list exists, based on a subset of the same MuLV insertion data [4]. In Figure 2.5b this manually curated list is taken as a reference and compared to the top 20 lists for KC-RBM, NGM and CIS-NG mapping. The seven genes identified in all three mappings were also present in the manually curated list. With the exception of *Med20*, all genes in the KC-RBM lists were identified in the manually curated list as well. The presence of *Med20* is explained by its proximity to *Ccnd3*. In the manually curated list, the insertions in the neighborhood of *Med20* were mapped to *Ccnd3*. CIS-NG finds two targets neither of which were present in the other two mappings nor in the manually curated list. These two genes, *Gm10125* and *Fgd2*, lie near *Zhfx1a* (*Zeb1*) and *Pim1* respectively, which are present in the manually curated list. NGM finds three targets neither identified in the other two mapping nor in the manually curated list, *Gm10826*, *Al672278.1*, and *Evi5*. These genes lie near *Myb*, *Runx1*, and *Gfi1* respectively, which were identified in the manually curated list as the target genes of corresponding CISs.

For MuLV, the lists of CTGs are also more extensively compared in Figure 2.5c, not restricting the comparison to only the top 20 CTGs, and again taking the manually curated list as a reference. This shows that, of all three methods, KC-RBM identifies the largest number of genes present in the manually curated list. The difference in the number of genes identified by KC-RBM and the second-best method, NGM, is highly significant ($p < 10^{-5}$, based on a permutation approach). Furthermore, the genes identified by KC-RBM that were also identified in the manually curated list, rank higher (lower average rank in the KC-RBM list), when compared to the other methods. The difference in rank between KC-RBM and the second-best method, NGM, is also highly significant ($p < 10^{-5}$, based on a permutation approach)

No manually curated list exists for the SB transposon insertions. Therefore, the Cancer Gene Census [15] was taken as a reference for evaluating the targets identified by the three approaches based on this dataset. The results of the comparison of KC-RBM, NGM, and CIS-NG mapping are depicted in Figure 2.5. Four genes are present in all three mappings, *Erg*, *Pten*, *Notch1*, and *Nr3c1*. All these genes are known cancer-related genes, see e.g. [16–19], and the first three are also present in the Cancer Gene Census. NGM is the only approach that identifies *Kit* as an additional Cancer Gene Census gene. Both CIS-NG and KC-RBM identify *Akt2* in addition to the Cancer Gene Census jointly detected by all three approaches. However, KC-RBM finds eight additional Cancer Gene Census genes. Similar to the retroviral case, somewhat more obscure proximal targets can score very high in CIS-NG mapping and NGM. Examples are *AC153556.1*, *RP23-336G7.3*, and *RP23-24E11.3*, which are located in the vicinity of *Myb*, *Jak1*, and *Bach2*, respectively. *Myb*, *Jak1*, and *Bach2* are identified exclusively by KC-RBM, and have been shown to be involved in cancer, see e.g. [20–22].

The results in Figure 2.5e are further substantiated by the more extensive comparison in Figure 2.5c, comparing larger lists of genes (refer to the Methods section), and again taking the Cancer Gene Census as a reference. KC-RBM again performs best in terms of the average rank of the Cancer Gene Census genes it identifies as well as the number of Cancer Gene Census genes identified. Furthermore, the differences in presence and rank between KC-RBM and the second-best method (CIS-NG in this case), are again significant ($p = 0.016$ and $p = 0.039$ respectively). Note that the Cancer Gene Census is less relevant for the SB insertion data than the manually curated list is for the MuLV data. It is not based on a SB transposon insertional mutagenesis screen. Furthermore, it is a list of human genes, which additionally have to be mapped to mouse homologs. Consequently, the overlap between the genes identified by the three methods and the Cancer Gene Census genes is much smaller compared to the MuLV case, resulting in higher p-values. Nevertheless, the p-values are significant, and demonstrate clearly that KC-RBM retrieves more cancer-related genes than the other methods.

2.3. DISCUSSION

In this paper we presented KC-RBM, a method for automatically mapping individual retroviral and SB transposon insertions to putative target genes. KC-RBM represents the first ever approach for mapping SB transposon insertions to target genes. In addition, the analyses presented here constitute the first genome-wide analysis of insertion and same-sample gene expression for retroviruses and transposons. Such a comprehensive dataset provides significant power in determining the factors that govern the associations between insertions and neighboring genes.

It is important to emphasize that, while mapping individual insertions to target genes, KC-RBM also exploits cross-sample information in multiple ways. The KC-RBM scale parameter allows for smoothing the positions of insertions based on cross-sample information. The insertion orientations are smoothed based on cross-sample information by setting the orientation homogeneity fraction parameter in KC-RBM. Furthermore, determining CTGs by aggregating insertions also integrates information from across tumor samples.

Although the presence of insertions is associated with increased gene expression (Figure 2.1) the analysis presented in Figure 2.1 does not provide sufficient evidence to distinguish cause from effect. In other words, we cannot conclude from this associative analysis that the presence of an insertion causes higher or lower gene expression. It is certainly possible that insertions are more likely to occur in actively transcribed regions, i.e. near genes with elevated expression levels. This and other factors inducing insertion biases have, in fact, been demonstrated for retroviruses as well as for transposons, e.g. [14, 23–29]. However, regardless of the determinants of insertion bias, it is a fact that insertions cause tumors since there is a statistically significant increase in the tumor incidence in animals infected with MuLV or in animals where SB transposons are activated [4, 30]. It is therefore very likely that these insertions caused, amongst other, changes in gene expression that resulted in oncogenesis. To further explore the causal relationship between insertions and aberrant gene expression, we performed additional analyses, the results of which are presented in the Supplementary Material.

The first analysis involves the distribution of insertion orientation in CTGs. If insertions were simply occurring in regions that were already transcriptionally active, i.e. elevated expression being the cause of insertions, one would expect insertions in these regions to show no preference for orientation. On the other hand, if the insertions were the cause of elevated gene expression, one would expect a strong preference for an orientation consistent with activation of the target gene(s). To test this hypothesis, we performed two tests. First, we randomized the insertion orientation in both the MuLV and SB datasets and regenerated the expression-position-orientation plots as presented in Figure 2.1. As can be seen in Figures S2 and S10, the orientation-dependent effect on gene expression disappears. Second, we analyzed the orientation distribution of insertions assigned to the top 214 MuLV CTGs (only including CTGs with 10 or more insertions). The results show a highly significant preference for orientation, consistent with the known mechanisms employed by retroviruses to activate target genes. (See Figures S3 and S4). Strong preferences for activating orientations were also established for the SB transposon (Figure S13). To further explore the causal relationship between insertions and gene expression, we performed a second analysis where we compared expression data from normal and tumor tissue. If insertions would simply target genes that are already active in normal tissue, there will be little difference between the activity of genes carrying insertions in tumor tissue and the activity of these same genes in normal tissue. To test this hypothesis, we compared gene expression of MuLV CTGs in cancer tissue to gene expression of these same genes in normal tissue. For the cancer tissue, only tumors with insertions in a specific CTG were included in the analysis. These analyses showed that in the cases where the average expression was higher in the cancer tissue compared to normal tissue, the increase was significant in 84.2% of the cases. This implies that in the vast majority of cases a gene carrying an insertion in the tumor tissue was significantly differentially expressed with respect to that same gene in normal tissue. For the cases where the average expression was lower in the tumor tissue with respect to normal tissue, the decrease was significant in 72.7% of the cases (See Figure S5). This strongly suggests that frequently targeted genes showing aberrant expression in tumor tissue, are not already active in normal tissue, hence challenging the aforementioned hypothesis that insertions simply tend to insert in genes that are already active in normal tissue.

Taken together, the strong orientation preference and the strong association of insertion presence with a change in gene expression suggest that a large fraction of insertions play a causal role in aberrant gene expression in tumor samples. Lastly and perhaps most importantly, it should be emphasized that while we employed the association of insertions and gene expression to obtain settings of the four window sizes in KC-RBM, this is not a requirement for the approach. In fact, depending on a researcher's interests, different window sizes may be appropriate.

As the definitive test for the performance of KC-RBM, and regardless of the direction of causality, we evaluated its ability to identify known cancer associated genes from a manually curated list [4], and from the Cancer Gene Census [15]. KC-RBM gives superior results when compared to automated CIS-NG mapping and NGM. Without the need for human intervention, it avoids more obscure proximal targets and finds a clean list of well-known cancer-related genes, as demonstrated by the comparison with a manually curated list [4], and the Cancer Gene Census [15]. This is important, since human interference could cause a bias, for example toward known or expected cancer genes, thus actually preventing the discovery of new or unknown cancer genes. This emphasizes the added value that a reliable automated insertion mapping procedure such as KC-RBM can have for analyzing insertional mutagenesis data and discovering novel oncogenes and tumor suppressor genes. As such, we believe that KC-RBM will significantly increase the efficiency of cancer gene discovery from insertional mutagenesis screens.

2.4. MATERIALS AND METHODS

2.4.1. DATASETS

MURINE LEUKEMIA VIRUS (MULV) DATA.

In total 20312 MuLV insertions were extracted from 8 insertional mutagenesis screens. These screens produced 1020 tumors in total, and were produced in mice with various genetic backgrounds [4, 5]. For a subset of 1986 insertions in 97 samples (p19 ko, p53 ko, and wild-type), Illumina MouseWG-6 v2.0 expression data was available and used.

SLEEPING BEAUTY (SB) TRANSPOSON DATA.

58266 SB transposon insertion loci were extracted from 255 lymphomas collected from T2/Onc2 and Rosa26 SBase mice. For a subset of 26955 insertions in 135 samples, Illumina MouseWG-6 v2.0 expression data was available and used.

2.4.2. LIST OF MAPPINGS

KC-RBM.

KC-RBM maps insertions to multiple putative target genes, using four window sizes, one for upstream-sense insertions, one for upstream-antisense insertions, one for downstream-sense insertions, and one for downstream-antisense insertions (with respect to transcription start site). Per transcript, these window size parameters are flexibly applied using two additional parameters, a GKC scale parameter [9] and a orientation homogeneity parameter. A gene is a target gene of an insertion if at least one of its transcripts is targeted. As an additional step in selecting a single target gene for each insertion, a prioritization can be made among the target genes identified by KC-RBM according to the number of times they were targeted by all insertions taken together. Then select the gene with the highest count to be the single target gene for that insertion.

NEAREST-GENE MAPPING (NGM).

For each insertion, find the nearest gene start site, and select this gene to be the single target gene of that insertion. This method is compared to KC-RBM.

CIS NEAREST-GENE MAPPING (CIS-NG).

Common insertion sites (CISs) are detected using GKC [9]. The peak of each CIS is then mapped to its nearest gene start site. This method is compared to KC-RBM.

2.4.3. METHODS

ALIGNING GENES (FIGURE 2.1).

The set of tumor samples was reduced to the set for which expression data was available ($n = 97$). All gene start sites were aligned with respect to location as well as orientation, and expression values were z-normalized per gene across samples. For all genes, all insertions were identified in a window of 400kb around these genes. All resulting (relative insertion locus, z-normalized gene expression) pairs were regarded as points in the (x,y) plane, and were then binned along the y-axis, making a distinction between insertions occurring in sense orientation relative to the gene and in antisense orientation relative to the gene start site, and normalizing gene length. The insertion density was computed by binning the insertions, and computing the number of insertions per base pair for each bin. These values were then normalized to a scale from 0 to 1.

The influence of window size (Figure 2.3).

The set of tumor samples was reduced to the set for which expression data was available. For each window size value and each gene, the following approach was taken. Tumor samples were divided in two groups. The first group contained the samples for which at least one insertion was mapped to that gene. The second group contained the samples for which no insertion was mapped to that gene. Between these two groups, a Wilcoxon-score was computed for elevated expression in the first group. Having computed this Wilcoxon-score for all genes, a significance threshold was determined per mapping by permuting ($n = 10000$) gene-wise expression profiles across samples with respect to gene-wise insertion profiles across samples, and setting a 5% significance threshold. Per window, each gene with at least one insertion was classified as significant or not significant exactly once. Per gene, each insertion is counted only once. Note that when computing the statistics for one of four window sizes, the other window sizes were set to zero. Furthermore, for all insertions within transcripts, association of insertion occurrence with increased expression levels was computed while disregarding the insertions outside transcripts. Permutation thresholds (5%) were calculated per window size.

The influence of the GKC scale (Figure 2.4).

For each KC-RBM scale, insertions were mapped to genes, and numbers and fractions of significant genes were computed as described above. KC-RBM was performed using window sizes (20kb,120kb,40kb,5kb) (MuLV) and (20kb,10kb,25kb,5kb) (SB transposon) for (upstream-sense,upstream-antisense,downstream-sense, downstream-antisense) insertions, and an orientation homogeneity fraction of 0.75. For each scale, all insertions (for MuLV all insertions from screen 1: p19 ko, p53 ko, and wild-type) were mapped to genes, but insertion-expression association was necessarily only computed for the samples for which expression data was available.

Comparing KC-RBM, RBM, CIS-NG, and CIS-manual mapping (Figure 2.5).

All insertions were mapped using KC-RBM, setting the window sizes to (20kb, 120kb, 40kb, 5kb) (MuLV) and (20kb, 10kb, 25kb, 5kb) (SB transposon) for (upstream-sense, upstream-antisense, downstream-sense, downstream-antisense) insertions. The orientation homogeneity fraction was set to 0.75, and the scale was set to 10kb (MuLV) and 2kb (SB transposon).

For KC-RBM, lists of top 20 CTGs were obtained by counting for each gene the number of times it was targeted, and then sorting this list, based on the number of times a gene was identified as a target. Specifically for SB transposon insertions, the CTGs were corrected for the fact that SB transposons only integrate at TA-sites: In determining CTGs, SB transposon insertions were each weighted by 1 divided by the local TA-density determined using the same kernel width as was used for the mapping of insertions (2kb). The total SB transposon CTG score across all genes was normalized to be equal to the total number of insertions. For NGM, also the top 20 CTGs were determined. CISs were detected using GKC [9], with a scale of 30kb. The 20 CISs with the highest peaks were then mapped to their nearest gene start site.

For both the MuLV and the SB transposon dataset, the top 20 results as well as the overall results were compared to a reference list. For MuLV a manually curated list based on the same dataset exists ([6]). The complete lists of genes identified by the three methods KC-RBM, NGM, and CIS-NG were compared to this manually curated list (CIS-M) with respect to presence and rank in this list. Regarding the presence of genes in either of the three methods in the CIS-M list, the three lists were made the same size by taking the top N of each list (where N is the length of the shortest list), to allow for a fair comparison. For each resulting list, the number of genes in that list also present in the CIS-M list was counted. For the comparison between the top two methods, KC-RBM and NGM, significance of the difference in numbers present in the CIS-M list was determined by permutation ($n = 100000$). Regarding rank, the following steps were taken. First, all lists were restricted to genes also occurring in CIS-M. Then, the three lists were made the same size by selecting only the top N from each list (where N is the length of the shortest list). This is necessary since the highest ranking CISs and CTGs are the easiest to retrieve, which may negatively affect the average rank of longer lists. Then, for each of the three lists, the average rank in that list of the genes also present in the manually curated list was calculated. For the comparison between the top two methods, KC-RBM and NGM, significance of the difference in average rank was determined by permutation ($n = 100000$).

For the SB transposon insertions a similar approach was taken, using as a reference the Cancer Gene Census [15], a list of human cancer-related genes. Mouse homologs were identified by mapping the human EntrezGene identifiers to mouse EntrezGene and Ensembl identifiers using the Bioconductor biomaRt 2.2.0 package [31].

REFERENCES

[1] J. de Jong, J. de Ridder, L. van der Weyden, N. Sun, M. van Uitert, A. Berns, M. van Lohuizen, J. Jonkers, D. J. Adams, and L. F. A. Wessels, *Computational identification of insertional mutagenesis targets for cancer gene discovery,* Nucleic Acids Res **39**, e105 (2011).

[2] H. Mikkers, J. Allen, P. Knipscheer, L. Romeijn, A. Hart, E. Vink, A. Berns, and L. Romeyn, *High-throughput retroviral tagging to identify components of specific signaling pathways in cancer.* Nat Genet **32**, 153 (2002).

[3] A. H. Lund, G. Turner, A. Trubetskoy, E. Verhoeven, E. Wientjens, D. Hulsman, R. Russell, R. A. DePinho, J. Lenz, and M. van Lohuizen, *Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice.* Nat Genet **32**, 160 (2002).

[4] A. G. Uren, J. Kool, K. Matentzoglu, J. de Ridder, J. Mattison, M. van Uitert, W. Lagcher, D. Sie, E. Tanger, T. Cox, M. Reinders, T. J. Hubbard, J. Rogers, J. Jonkers,

L. Wessels, D. J. Adams, M. van Lohuizen, and A. Berns, *Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks.* Cell **133**, 727 (2008).

[5] J. Kool, A. G. Uren, C. P. Martins, D. Sie, J. de Ridder, G. Turner, M. van Uitert, K. Matentzoglu, W. Lagcher, P. Krimpenfort, J. Gadiot, C. Pritchard, J. Lenz, A. H. Lund, J. Jonkers, J. Rogers, D. J. Adams, L. Wessels, A. Berns, and M. van Lohuizen, *Insertional mutagenesis in mice deficient for p15Ink4b, p16Ink4a, p21Cip1, and p27Kip1 reveals cancer gene interactions and correlations with tumor phenotypes,* Cancer Res **70**, 520 (2010).

[6] A. G. Uren, J. Kool, A. Berns, and M. van Lohuizen, *Retroviral insertional mutagenesis: past, present and future.* Oncogene **24**, 7656 (2005).

[7] J. Kool and A. Berns, *High-throughput insertional mutagenesis screens in mice to identify oncogenic networks.* Nat Rev Cancer **9**, 389 (2009).

[8] T. Suzuki, H. Shen, K. Akagi, H. C. Morse, J. D. Malley, D. Q. Naiman, N. A. Jenkins, and N. G. Copeland, *New genes involved in cancer identified by retroviral tagging.* Nat Genet **32**, 166 (2002).

[9] J. de Ridder, A. Uren, J. Kool, M. Reinders, and L. Wessels, *Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens.* PLoS Comput Biol **2** (2006).

[10] M. Sauvageau, M. Miller, S. Lemieux, J. Lessard, J. Hebert, and G. Sauvageau, *Quantitative expression profiling guided by common retroviral insertion sites reveals novel and cell type specific cancer genes in leukemia.* Blood **111**, 790 (2008).

[11] J. Mattison, J. Kool, A. G. Uren, J. de Ridder, L. Wessels, J. Jonkers, G. R. Bignell, A. Butler, A. G. Rust, M. Brosch, C. H. Wilson, L. van der Weyden, D. A. Largaespada, M. R. Stratton, P. A. Futreal, M. van Lohuizen, A. Berns, L. S. Collier, T. Hubbard, and D. J. Adams, *Novel candidate cancer genes identified by a large-scale cross-species comparative oncogenomics approach.* Cancer Res **70**, 883 (2010).

[12] S. J. Erkeland, R. G. Verhaak, P. J. M. Valk, R. Delwel, B. Löwenberg, and I. P. Touw, *Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia.* Cancer Res **66(2)**, 622 (2006).

[13] J. Jonkers and A. Berns, *Retroviral insertional mutagenesis as a strategy to identify cancer genes.* Biochim Biophys Acta **1287**, 29 (1996).

[14] X. Wu, Y. Li, B. Crise, and S. M. Burgess, *Transcription Start Regions in the Human Genome Are Favored Targets for MLV Integration,* Science **300**, 1749 (2003).

[15] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, *A census of human cancer genes,* Nat Rev Cancer **4**, 177 (2004).

[16] G. Sashida, E. Bazzoli, S. Menendez, Y. Liu, and S. D. Nimer, *The oncogenic role of the ETS transcription factors MEF and ERG.* Cell Cycle **9** (2010).

[17] L. Salmena, A. Carracedo, and P. P. Pandolfi, *Tenets of PTEN tumor suppression.* Cell **133**, 403 (2008).

[18] E. J. Allenspach, I. Maillard, J. C. Aster, and W. S. Pear, *Notch signaling in cancer.* Cancer Biol Ther **1**, 466 (2002).

[19] G. E. Lind, K. Kleivi, G. I. Meling, M. R. Teixeira, E. Thiis-Evensen, T. O. Rognum, and R. A. Lothe, *ADAMTS1, CRABP1, and NR3C1 identified as epigenetically deregulated genes in colorectal tumorigenesis.* Cell Oncol **28**, 259 (2006).

[20] R. G. Ramsay and T. J. Gonda, *MYB function in normal and cancer cells,* Nat Rev Cancer **8**, 523 (2008).

[21] A. Ono, K. Kono, D. Ikebe, A. Muto, J. Sun, M. Kobayashi, K. Ueda, J. V. Melo, K. Igarashi, and S. Tashiro, *Nuclear positioning of the BACH2 gene in BCR-ABL positive leukemic cells.* Gene Chromosome Canc **46**, 67 (2007).

[22] A. Verma, S. Kambhampati, S. Parmar, and L. C. Platanias, *Jak family of kinases in cancer.* Cancer Metast Rev **22**, 423 (2003).

[23] P. Hematti, B.-K. Hong, C. Ferguson, R. Adler, H. Hanawa, S. Sellers, I. E. Holt, C. E. Eckfeldt, Y. Sharma, M. Schmidt, C. von Kalle, D. A. Persons, E. M. Billings, C. M. Verfaillie, A. W. Nienhuis, T. G. Wolfsberg, C. E. Dunbar, and B. Calmels, *Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells,* PLoS Biol **2** (2004).

[24] R. S. Mitchell, B. F. Beitzel, A. R. Schroder, P. Shinn, H. Chen, C. C. Berry, J. R. Ecker, and F. D. Bushman, *Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences,* PLoS Biol **2**, e234 (2004).

[25] F. Bushman, M. Lewinski, A. Ciuffi, S. Barr, J. Leipzig, S. Hannenhalli, and C. Hoffmann, *Genome-wide analysis of retroviral DNA integration,* Nat Rev Microbiol **3**, 848 (2005).

[26] C. Berry, S. Hannenhalli, J. Leipzig, and F. D. Bushman, *Selection of target sites for mobile DNA integration in the human genome,* PLoS Comput. Biol **2**, e157 (2006).

[27] M. K. Lewinski, M. Yamashita, M. Emerman, A. Ciuffi, H. Marshall, G. Crawford, F. Collins, P. Shinn, J. Leipzig, S. Hannenhalli, C. C. Berry, J. R. Ecker, and F. D. Bushman, *Retroviral DNA integration: viral and cellular determinants of target-site selection.* PLoS Pathog **2** (2006).

[28] A. Ambrosi, C. Cattoglio, and C. Di Serio, *Retroviral integration process in the human genome: Is it really non-random? a new statistical approach,* PLoS Comput. Biol **4**, e1000144 (2008).

[29] J. Plachy, J. Kotab, P. Divina, M. Reinisova, F. Senigl, and J. Hejnar, *Proviruses selected for high and stable expression of transduced genes accumulate in broadly transcribed genome areas.* J Virol (2010).

[30] T. K. Starr, R. Allaei, K. A. T. Silverstein, R. A. Staggs, A. L. Sarver, T. L. Bergemann, M. Gupta, M. G. O'Sullivan, I. Matise, A. J. Dupuy, L. S. Collier, S. Powers, A. L. Oberg, Y. W. Asmann, S. N. Thibodeau, L. Tessarollo, N. G. Copeland, N. A. Jenkins, R. T. Cormier, and D. A. Largaespada, *A Transposon-Based genetic screen in mice identifies genes altered in colorectal cancer,* Science **323**, 1747 (2009).

[31] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber, *BioMart and bioconductor: a powerful link between biological databases and microarray data analysis.* Bioinformatics **21**, 3439 (2005).

2.5. Supplementary Material

http://www.ncbi.nlm.nih.gov/pubmed/21652642

3

CHROMATIN LANDSCAPES OF RETROVIRAL AND TRANSPOSON INTEGRATION PROFILES

Johann DE JONG*
Waseem AKHTAR*
Jitendra BADHAI
Alistair G. RUST
Roland RAD
John HILKENS
Anton BERNS
Maarten VAN LOHUIZEN
Lodewyk F.A. WESSELS
Jeroen DE RIDDER

Parts of this chapter have been published in PLOS Genetics [1].

* These authors contributed equally to this work.

Abstract

The ability of retroviruses and transposons to insert their genetic material into host DNA makes them widely used tools in molecular biology, cancer research and gene therapy. However, these systems have biases that may strongly affect research outcomes.

To address this issue, we generated very large datasets consisting of ~120000 to ~180000 unselected integrations in the mouse genome for the Sleeping Beauty (SB) and piggyBac (PB) transposons, and the Mouse Mammary Tumor Virus (MMTV). We analyzed ~80 (epi)genomic features to generate bias maps at both local and genome-wide scales. MMTV showed a remarkably uniform distribution of integrations across the genome. More distinct preferences were observed for the two transposons, with PB showing remarkable resemblance to bias profiles of the Murine Leukemia Virus. Furthermore, we present a model where target site selection is directed at multiple scales. At a large scale, target site selection is similar across systems, and defined by domain-oriented features, namely expression of proximal genes, proximity to CpG islands and to genic features, chromatin compaction and replication timing. Notable differences between the systems are mainly observed at smaller scales, and are directed by a diverse range of features.

To study the effect of these biases on integration sites occupied under selective pressure we turned to insertional mutagenesis (IM) screens. In IM screens, putative cancer genes are identified by finding frequently targeted genomic regions, or Common Integration Sites (CISs). Within three recently completed IM screens, we identified 7% - 33% putative false positive CISs, which are likely not the result of the oncogenic selection process. Moreover, results indicate that PB, compared to SB, is more suited to tag oncogenes.

3.1. Introduction

DNA integrating elements, such as transposons and retroviruses, are an important tool in many areas of molecular biology, e.g. gene therapy [2, 3], oncogene discovery [4, 5], gene regulation [6, 7], and functional genetics [8, 9]. A current limitation to the use of retroviruses and transposons is that, even without selective pressure, integration loci are not uniformly distributed across the genome. There are significant biases, the molecular determinants of which are still largely unknown. Such biases can pose problems, for example in the discovery of novel cancer genes by insertional mutagenesis (IM), because it can be difficult to distinguish clusters of integrations arising purely through integration bias from those giving a selective growth advantage to the cell. More insight into target site selection would also benefit gene therapy, where adverse integrational activation of oncogenes resulting from treatment with retroviral vectors has been observed [10].

Three of the main integrating elements currently used in the fields mentioned above are the Sleeping Beauty transposon (SB), the piggyBac transposon (PB), and the mouse mammary tumor virus (MMTV). During the last decade, some studies have reported on integration biases in the mouse genome for one or more of these systems. SB does not integrate randomly on a micro-scale, since it is dependent on local DNA deformability, and the presence of a TA dinucleotide at the site of integration [11, 12]. At larger scales integration target site selection was found to be relatively random [13, 14], although (sometimes conflicting) associations have been observed for CpG islands, gene density, and actively transcribed loci [15, 16]. PB integration is TTAA-specific, although slight variations on this target sequence have been observed [17]. PB was found to be biased towards transcriptional units, CpG islands and transcription start sites (TSSs) and actively transcribed loci, and in general marks of open chromatin [15, 18–20]. MMTV is the least well-characterized of the three. In mouse and human cell lines, no bias was detected with respect to genes, TSSs and CpG islands, and MMTV was suggested to be the retrovirus least biased in its target site selection [21].

While the studies mentioned above have provided valuable insights into retroviral and transposon target site selection for SB, PB and MMTV in the mouse genome, they do have some limitations with respect to gaining insight into de novo integration target site selection. These limitations can be subdivided into three categories. First, there are limitations regarding the individual integration datasets. For example, some datasets were generated using cells that were enriched with a selectable marker, e.g. [16, 19, 20]. Also, considering only the datasets that were not under selective pressure, the sample sizes were fairly small compared to current standards, mostly in the range of several tens to several hundreds of integrations, e.g. [15–17]. Note that having large numbers of integrations is important for gaining sufficient statistical power to detect even relatively weak biases. Second, some limitations complicate the comparison between integration datasets. For example, integration datasets have been compared that differed substantially in the cell lines used, as well as the degree of selection imposed on those cell lines, e.g. [16]. Third, other limitations concern the features used to analyze the integration datasets. For example, integration datasets have been compared to features in non-matching cell types [16], while for example Murine Leukemia Virus target site selection within the human genome has been suggested [22] and shown [23] to have a cell type dependent component. Interesting to note here is that for a resurrected human endogenous retrovirus, no cell type specific integration into the human genome could be detected [24]. Also, many studies focused only on a limited number of features, e.g. genomic features [18, 21], or genomic features and DNase I hypersensitivity [16]. Moreover, the features, such as ChIP-seq profiles, were not necessarily preprocessed in similar ways, for example in terms of sequence alignment, e.g. [19]. This complicates the comparison of features across different systems.

To address these questions, we generated large datasets of SB and PB integrations in mouse embryonic stem cells (mESC). In order to directly compare the two transposons, they were mobilized from the same construct containing inverted repeats (IRs) for both PB and SB. This eliminates any other possible cause than the IR (or the transposon-specific transposase) for the observed differences between the two systems. In addition, we generated a large dataset of MMTV integrations in normal murine mammary gland epithelial (NMuMG) cells. All three datasets were generated under minimal selective pressure, and are henceforward referred to as unselected integration profiles. They are considerably larger than previously published datasets of unselected integrations, ~120000, ~130000, and ~180000 integrations respectively.

We associated the three integration profiles with a large number of genomic and epigenomic features. In particular, for SB and PB, the recent explosive growth in publicly available ChIP-seq datasets [25] enabled us to analyze a large number of epigenomic features (~70), all in mESCs. To allow for a better comparison between these datasets, they were preprocessed from the raw sequence reads in exactly the same way.

Additionally, the impact of selective pressure on an unselected integration profile, which is important in using IM for cancer gene discovery, has never been addressed extensively. Previous work can be classified as either knowledge-based or data-driven. The knowledge-based approaches use modeling of previously described integration biases to avoid CIS calls that can be explained by these biases, such as SB TA sequence specificity [26], or γ-retroviral TSS specificity and lentiviral gene specificity [27]. Alternatively, by assuming that a genic region harboring a true positive CIS should contain significantly more integrations than its flanking genes, three out of nine CISs from gene therapeutic clinical trials were labeled false positive [28]. These knowledge-based approaches are necessarily limited in their modeling of integration bias, for example in the number of features that are considered. Conversely, data-driven approaches treat integration bias as a black box, and compare integration datasets that were under substantial selective pressure to integration datasets that lacked this pressure. Using this approach, one study suggested a 47% false positive rate for a MuLV tumor screen [29], and another observed 6 control CISs from SB integrations present in mouse tail DNA, where 79 CISs could be found in the corresponding SB tumor screen [30]. To analyze the impact of selective pressure on integration bias profiles, we take the data-driven approach and compare our three unselected datasets with CIS integration profiles from three previously published tumor screens [4, 31–33].

Taken together, this allows us to present the most extensive analysis of SB, PB and MMTV target site selection to date, the results of which include previously undetected biases and differences between selected and unselected integration profiles. Another focal point of the analysis is the influence of scale. By analyzing differences between small-scale (within ~800bp from integrations) and large-scale (~800bp or further from integrations) associations of genomic and epigenomic features with the proximity of integrations, we reveal a hierarchical organization in integration bias. On a global scale, target sites of different systems are selected in similar ways, whereas differences mainly exist in fine-tuning on a local scale.

3.2. Results

The integration datasets used in this study are described in Table 1. For each tumor screen, the numbers of singleton integrations and CIS integrations is given. Here, a CIS is a genomic region with more integrations across tumors than expected by chance (see Material and Methods section and [4, 34]). An integration falling (not falling) within a CIS is referred to as a CIS (singleton) integration.

The features for which integration bias was studied are summarized in Table 2. Statistical procedures used for each figure are listed in Table S1.

3.2.1. SB, PB, and MMTV exhibit unique sequence and gene specificity

We started our analysis by studying the sequence specificity at integration sites of the three systems. As expected, PB and SB mostly, but not exclusively, integrate at TTAA

Table 3.1: Integration datasets used in this chapter.

System	Cell type	Size	Selected	#CIS insertions	#Singleton insertions	Reference
MuLV	B/T-cell tumors	20312	Yes	8707	11605	[4, 35]
SB	Mouse ES cells	131594	No	NA	NA	This work
SB	B/T-cell tumors	58266	Yes	1945	56321	[31]
PB	Mouse ES cells	122667	No	NA	NA	This work
PB	Haematopoietic tumors	5590	Yes	306	5284	[33]
MMTV	Mouse mammary cells	180469	No	NA	NA	This work
MMTV	Mouse mammary tumors	34753	Yes	2834	31919	[32]

Table 3.2: Genome and chromatin profiling data used in this chapter.

Description	Technique	Cell type	Reference
Gene expression mESC	RNA-seq	mESC	This work
Gene expression NMuMG	Microarray	NMuMG	[36]
Replication timing	ChIP-chip	mESC	[37]
H3K9me2	ChIP-chip	mESC	[38]
LaminB1	DamID	mESC	[39]
Hi-C	Hi-C	mESC	[40]
mC; hmC	Bisulfite sequencing; hMeDIP	mESC	[41]
CpG island proximity; gene proximity	NA	NA	[42]
Dnase I hypersensitivity	ChIP-seq	mESC	[43]
H3K4me3; H4K20me3	ChIP-seq	mESC	[44]
H3K9me3	ChIP-seq	mESC	[45]
H3K27me3; H3K36me3; H2AZ	ChIP-seq	mESC	[46]
Atrx	ChIP-seq	mESC	[47]
BrgJ1	ChIP-seq	mESC	[48]
cMyc; E2f1; Esrrb; Klf4; nMyc; Oct4; Smad1; Sox2; Stat3; Suz12; Tcfcp2l1; Zfx	ChIP-seq	mESC	[49]
H3K79me2; Nanog; Tcf3	ChIP-seq	mESC	[50]
Ezh2	ChIP-seq	mESC	[51]
SetDB1	ChIP-seq	mESC	[52]
Jarid2; Mtf2	ChIP-seq	mESC	[53]
Tbx3	ChIP-seq	mESC	[54]
Ctr9; Pol2-Ser2P; Pol2-Ser5P	ChIP-seq	mESC	[55]
Luzp1	ChIP-seq	mESC	[56]
Chd7	ChIP-seq	mESC	[57]
Med1; Med12; Smc1; Smc3	ChIP-seq	mESC	[58]
H3K27ac; H3K4me1	ChIP-seq	mESC	[59]
Yy1	ChIP-seq	mESC	[60]
CTCF; p300	ChIP-seq	mESC	[61]
H3K9ac	ChIP-seq	mESC	[62]
Taf3	ChIP-seq	mESC	[63]
Jaridb1	ChIP-seq	mESC	[64]
Smad2/3; Smad3	ChIP-seq	mESC	[65]
Kap1	ChIP-seq	mESC	[66]
H3K4me2	ChIP-seq	mESC	[67]
macroH2A1	ChIP-seq	mESC	[68]
p53; p53S18P	ChIP-seq	mESC	[69]
Cbx7; Ring1b	ChIP-seq	mESC	[70]
CoREST; Hdac1; Hdac2; Lsd1; Mi-2b	ChIP-seq	mESC	[71]
Dpy30	ChIP-seq	mESC	[72]
Mbd3	ChIP-seq	mESC	[73]
Mcaf1	ChIP-seq	mESC	[74]
Pol2-Ser7P	ChIP-seq	mESC	[75]
Tbp	ChIP-seq	mESC	[76]
Ell	ChIP-seq	mESC	[77]

and TA sites respectively, accounting for 93% (PB) and 94% (SB) of the integrations (Figure S1). The remaining integrations show sequences that are relatively similar to these motifs and often differ by only one nucleotide (Figure S1). MMTV has little integration site sequence specificity. Focusing instead on integration-flanking sequences (50bp on either side), de novo motif discovery using the HOMER software [78] revealed no enrichment of non-trivial motifs (i.e. not TTAA for PB, and not TA for SB), except for two motifs in the case of MMTV (binding TFAP2A and Tcfap2e respectively; Figure S2).

The bias of an integrating element with respect to genes is of particular interest in IM and gene therapy. In IM screens, more integrations near genes are desired whereas in gene therapy integrations proximal to genes pose a potential threat to the patient, since such integrations may give rise to cancer. We, therefore, compared the integration density of the three systems in and around genes, by aligning all genes (Figure 1A). PB shows a strong bias for TSSs. The PB bias profile with respect to genes is remarkably similar to that of MuLV, see e.g. [79, 80], as well as Figure S3 which shows singleton MuLV profiles, based on a previously published tumor screen [4, 35]. While PB has a strong bias for TSSs, SB is enriched uniformly along the body of genes. MMTV shows the weakest bias, although it does slightly prefer TSSs, and has a mild bias against gene bodies.

The significance of the observations made above was assessed using the binomial test, and visualized at a 5% FDR threshold in Figures 1B and 1C, for different gene and transcript related regions. Refer to Table S2 for the raw *p*-values associated with these statistical tests. It confirms the strong bias of PB for TSSs, and indicates that PB prefers integrating with its transcription unit oriented towards the TSS. When landing within genes, MMTV prefers a sense orientation relative to the host gene, and SB and PB an antisense orientation. In general, MMTV shows the least biased profile, with weak but significant biases for TSSs, and against genic regions. PB integrations are mainly enriched in the 5'UTR, weakly enriched in exons and the 3'UTR, and biased against introns. This pattern is highly similar to that observed for singleton MuLV integrations (Figure S4). Regarding orientation biases within transcripts, MMTV shows sense orientation biases for exons and introns, whereas the two transposons show only antisense orientation biases, SB in introns, and PB in 3'UTRs and exons.

3.2.2. Integration profiles of SB and PB are shaped by endogenous gene expression

Next, we analyzed the influence of expression status of endogenous genes on integration bias. This revealed interesting differences between the unselected integration profiles of the three systems (Figure 2), significance of which was assessed by the Cochran-Armitage trend test, unless mentioned otherwise. Across genes and TSSs, PB is strongly influenced by the a priori gene expression levels ($p < 10^{-7}$ for genic, TSS_upstream and TSS_downstream). For SB this same positive trend is only apparent for intragenic integrations ($p < 10^{-7}$), whereas around TSSs, the numbers of integrations decrease with increasing gene expression ($p = 1.2 \times 10^{-6}$ and $p < 10^{-7}$ for TSS_upstream and TSS_downstream respectively). Within weakly expressed genes, there is a depletion of PB integrations (binomial test; $p < 10^{-7}$). MMTV target site selection is largely independent from the expression levels of endogenous genes ($p = 1.5 \times 10^{-3}$, $p = 0.58$ and $p = 0.97$ for genic, TSS_upstream and TSS_upstream respectively). Although it is evident

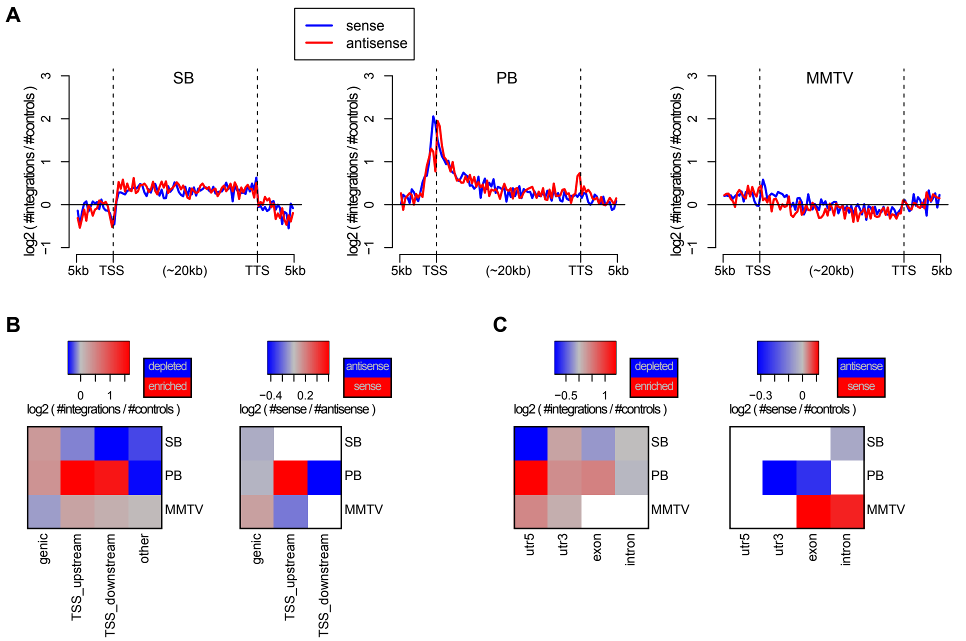


Figure 3.1: **Biases with respect to genes and transcripts.** A) Gene alignment plots showing the distribution of integrations across genes from 5kb upstream to the transcription start site (TSS), transcription termination site (TTS), and 5kb downstream. The red line depicts the integrations with sense orientation relative to the gene, blue depicts antisense. B) Biases with respect to genes for the unselected integration datasets. We distinguish between integrations within genes (genic), within 1kb upstream of the TSS (TSS_upstream), within 1kb downstream of the TSS (TSS_downstream), and other integrations (other). On the left, the color scale blue-gray-red represents increasing numbers of integrations, relative to expected. On the right, the color scale blue-gray-red represents the integration orientation bias relative to genes, from antisense to sense. Associations that are not significant (binomial test; FDR-corrected $p > 0.05$) are white. C) Biases with respect to transcripts, distinguishing between integrations in 5’UTRs, 3’UTRs, exons and introns.

that there are more MMTV integrations in TSS regions than within genes (Figures 1 and 2), this preference is clearly independent from gene expression.

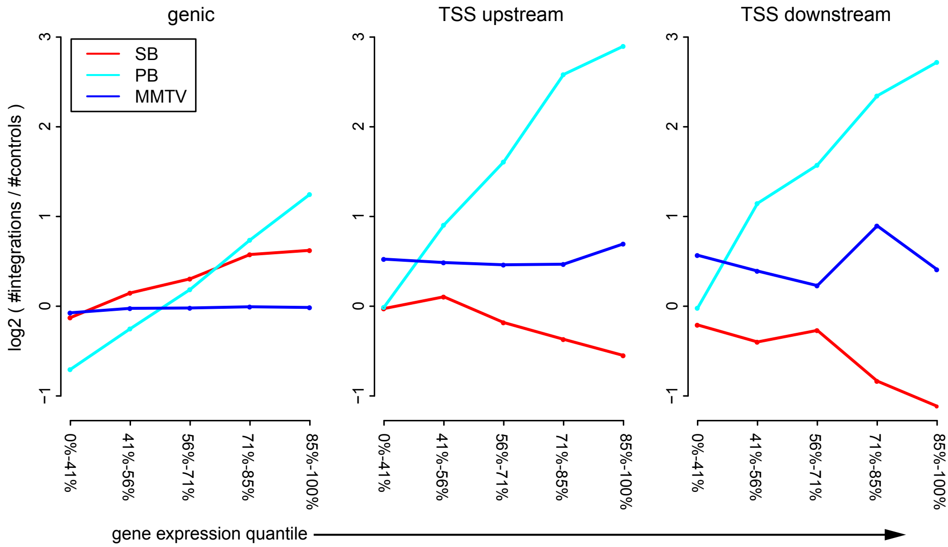


Figure 3.2: **Influence of gene expression on integration bias.** For each of the systems SB, PB, and MMTV, the unselected integrations are divided into genic integrations, integrations occurring within 1kb upstream of the TSS (TSS_upstream), and integrations occurring within 1kb downstream of the TSS (TSS_downstream). Genes are divided into 5 groups, based on expression level. The sizes of these groups are indicated on the *x*-axis. For each pair of gene expression level and system, the number of observed integrations is counted, and compared to the number of expected integrations.

3.2.3. TOPOLOGICAL DOMAIN INTERFACES ARE HOTSPOTS OF INTEGRATION

In addition to gene structure, the integration site selection of an integrating element can also be influenced by other features such as organization of the genome, state of chromatin compaction and transcription factor binding events, as well as by epigenetic modifications. Considering that an important barrier for integration of viral or transposon DNA into host DNA can be how tightly the DNA is packed in chromatin, we looked at the influence of the a priori chromatin organization on the unselected integration profiles. Hi-C [81] is a technique for studying chromatin compaction and organization by determining interaction frequencies between different genomic loci on a genome-wide scale. Analysis of Hi-C data has suggested that the genome is organized into chromatin modules, called topologically associated domains (TADs), which are stable across different cell types [40]. TADs are separated by less organized (showing fewer 3D interactions) regions called TAD boundaries. It is conceivable that chromatin is relatively less compact in and near TAD boundaries as compared to TADs. We asked if this 3D organization of the genome has any influence on the integration bias of systems. In general, we found that more integrations are close to the interface between TADs and their boundaries (Figure 3), i.e. all systems have a preference for inserting at the border of TADs, which are tightly organized, but not necessarily in the less organized chromatin of boundary regions. It is interesting to note that for MMTV, which is generally the least biased system, the bias for the TAD - TAD boundary interface is stronger than that of SB (Cochran-Mantel-Haenszel test in a window of 10kb on either side of the interface; $p < 10^{-7}$).

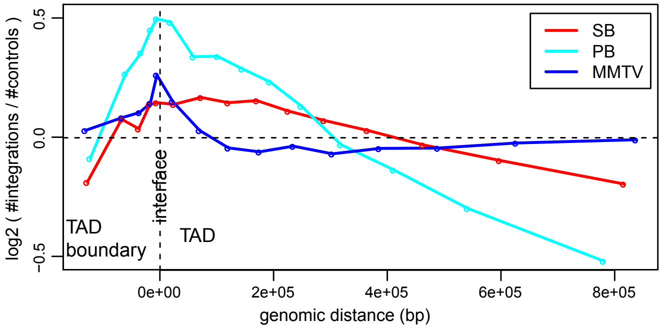


Figure 3.3: **Unselected integration profiles with respect to TAD - TAD boundary interface.** The x-axis represents genomic distance from the interface. The y-axis represents the $\log_2$ ratio of observed number of integrations versus the expected number of integrations.

3.2.4. TRANSPOSONS SHOW HIGHLY DIVERGENT BEHAVIOR IN INTEGRATIVE (EPI)GENOMIC CONTEXT

In the previous two sections, we demonstrated that there are strong biases of the unselected integration profiles with respect to genes, transcripts, gene expression, and genome organization. However, these features themselves have strong spatial ties with other features, such as histone marks and transcription factor binding. Therefore, we asked the following two questions. First, how do these features associate with integration proximity? Second, do they provide extra information with respect to integration bias, in addition to what gene proximity and gene expression provide? The features we analyzed are listed in Table 2. To maximize comparability between the features, the ChIP-seq datasets were preprocessed from the raw sequencing reads in exactly the same way. Since these features are not available in NMuMG cells, which were used for generating the MMTV integrations, we restricted all analyses based on these data to SB and PB.

First, we analyzed the orientation biases with respect to these features (Figures S5 and S6). This showed that the two transposons preferably integrate with the transcription cassette cloned in them oriented towards regions of high feature signal. Although for individual marks this bias is not very substantial, it is highly consistent across different marks, especially for SB. It is important to note that an orientation bias of these systems relative to genes cannot explain this bias for SB, and only partly for PB (Figures S5 and S6).

Using a limited number of mostly genomic features, it has been observed before that associations of integration occurrence with these features depend on the scale chosen for the analysis [16]. Therefore, we analyzed our genomic and epigenomic features across different scales, by comparing feature scores at the site of integration with feature scores at increasing distances (scales) from the integration site. Features were then clustered based on their association profiles across scales (Figure 4A). Resulting associations can be positive, i.e. higher feature scores at integration sites compared to their neighborhood, or negative, i.e. lower feature scores at integration sites compared to their neighborhood.

A clustering of the association profiles results in four groups of features generally associated with activation (Clusters 1, 2, 3 and 5), and three groups associated with repression (Clusters 6, 7 and 8). The remaining Cluster 5 is more mixed (Figure 4C).

Another characterization of the resulting clusters is into groups of features for which the behavior is either fairly similar (Clusters 2, 3, 6 and 8), or groups for which SB and PB behave very differently (Clusters 1, 4, 5, and 7). Especially striking when observing the differences is that PB is positively associated (Clusters 1, 4 and 5) with far more features than SB (Cluster 7). Since for both PB and SB, association with the many gene-related features in Cluster 3 is mostly positive, this indicates that SB does prefer gene-rich regions and active genes over heterochromatin, but in these regions, compared to PB, generally avoids regulatory units such as histone modifications and transcription factor bound regions. Interestingly, the single cluster positively associating with SB but negatively with PB (Cluster 7) contains mostly repressive features. Combined, these observations suggest that PB is much more biased to active chromatin than SB, whereas SB is also partly biased towards more repressed chromatin.

The scale-based approach reveals that the sign of association can change across different scales. For example, SB shows negative association with some of the features in Cluster 3 on a small scale, but a positive association on larger scales. This implies that SB has a bias for larger scale domains containing these features such as Ctr9, H3K79me2 and 5hmC, but within these domains integration sites will generally avoid overlap with these marks.

Conversely, association changing from positive to negative for increasing scales is seen for example for PB and Stat3 in Cluster 5. This indicates that PB prefers domains relatively devoid of these features. However, given a PB integration in such a domain, it will be mildly biased towards Stat3.

The above observations suggest a hierarchy in target site selection, which is further illustrated by the fact that some features, for both SB and PB, are consistently non-significant at smaller scales (Figures 4A and 4B). For example, at smaller distances from integrations, associations with features such as gene proximity and expression, CpG island proximity, replication timing and H3K9me2 are not significant for both SB and PB. They are however consistently significant on larger scales. On the other hand, most transcription factors and other histone marks show strong associations already at small scales. Henceforward, features that are significant only at larger scales will be referred to as ’macrofeatures’, as opposed to ’microfeatures’, which are significant already at smaller scales (refer to Table S3 for a list of macrofeatures and microfeatures). For a selected set of macrofeatures that were available for NMuMG cells, we performed a similar analysis showing that these features also behave as macrofeatures in MMTV, an unrelated system (Figure S7).

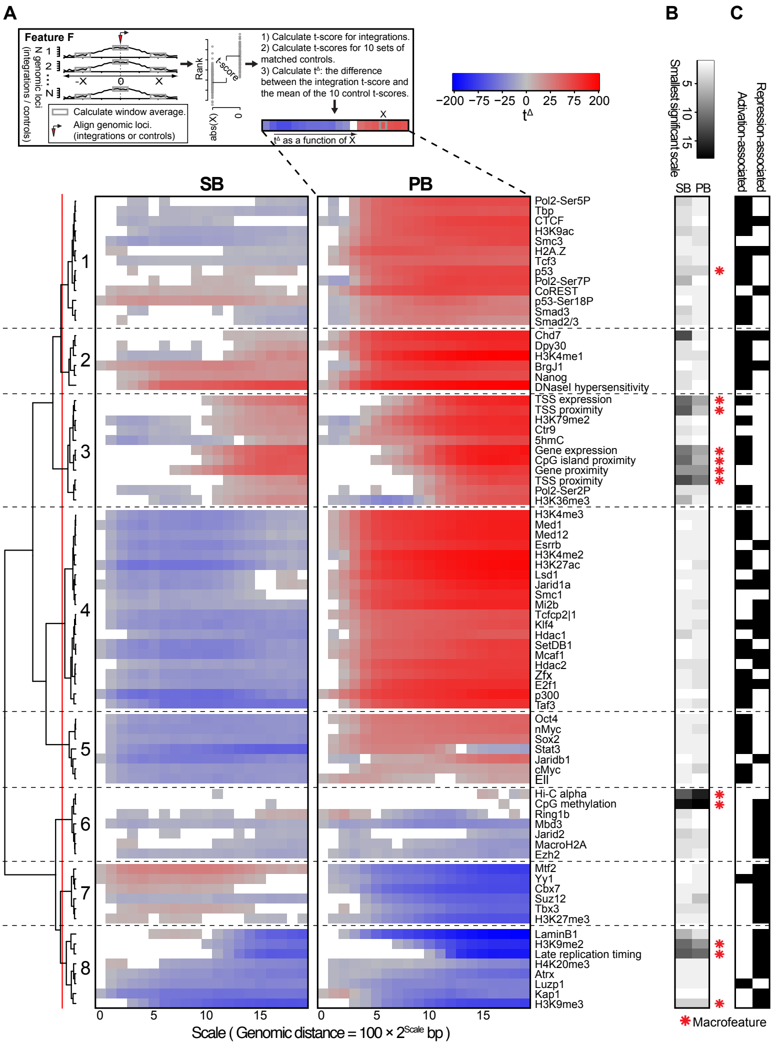


Figure 3.4: **Scale-based analysis of integration bias.** A) Association of the unselected integration profiles with various genome-wide features across different scales. Measure of association is a normalized t-score (see Material and Methods), computed on rank-normalized feature values, visualized on a blue-gray-red scale from negative to positive t-scores. Associations that are not significant (FDR-corrected $p > 0.05$) are white. A positive (negative) t-score for a certain scale x and feature y means that for that particular feature, the mean values in a 200bp window around the integrations are on average higher (lower) than the mean values in a 200bp window around the points at a distance of 100×2^{x}bp upstream and downstream from the integration (see Material and Methods). The dendrogram shows a hierarchical clustering of the profiles using the euclidean distance measure and ward linkage. B) The rank-transformed smallest scale at which significance is achieved, with a scale going from white (small scale) to black (large scale). A feature is called a 'macrofeature' if its smallest significant scale is larger than the mean rank-normalized smallest significant scale across features, in both systems. C) Features associated with transcriptional repression and/or activation, based on published literature.

3.2.5. Integration site selection is directed at multiple levels

Biases of unselected integration profiles with respect to the macrofeatures are similar across the systems and scales. This suggests that on a large scale, integration bias is regulated in similar ways for both systems, and that this large scale bias is mainly determined by the macrofeatures. However, within a distance of ~800bp, macrofeatures provide no information with regard to integration locus, contrary to the microfeatures. This indicates that microfeatures may in fact be determinants of integration bias at a higher resolution, which prompted us to ask the following question: Are macrofeatures needed at all to explain integration proximity, or are microfeatures sufficient for this purpose?

To address this question, we needed to take into account that the features in Figure 4 show a high degree of multicollinearity. Multicollinearity implies that a strong association between a certain feature A and integration proximity may potentially be explained by the association of A with another feature B that also strongly associates with integration proximity, i.e. integration proximity may be conditionally independent from A, given B. Then, rephrasing the question above, for each system we wanted to identify a set of features such that integration proximity is conditionally independent from all other features, given this set of features.

BANJO [82] is designed to identify such conditional independencies in the form of Bayesian networks [83], and thus allowed us to determine for each feature its importance for integration proximity. For this, we used two measures derived from the Bayesian networks. The first measure (’log$_1$0 % bootstraps’; see Material and Methods) represents the confidence that a feature is truly relevant for integration proximity. The second measure (’log$_1$0 mean CMI’, or conditional mutual information; see Material and Methods) represents the strength of association between integration proximity and a feature.

Interestingly, the results show that seven macrofeatures are consistently of great importance, i.e. of high-confidence and strongly associating, in both systems (Figure 5). These features are gene / TSS / TTS proximity, TSS expression (the expression of the gene with the nearest TSS), replication timing, CpG island proximity, and Hi-C alpha (a measure of chromatin compaction; see Material and Methods section). This shows that in addition to microfeatures for explaining local differences between systems (Figure 4), macrofeatures are needed to explain integration bias in each of these systems on large scales (Figure 5). Furthermore, it indicates that on a large scale, biases of the two systems are similar, and that differences between the systems are mainly found in the microfeatures.

3.2.6. Integration bias is a potential cause of spurious common integration sites

Insertional mutagenesis (IM) using retroviruses and transposons is an important tool in the discovery of new putative cancer genes. These elements mutate the genome by inserting into the host DNA. Mutations providing cells with a proliferative and/or survival advantage can cause tumors. Because the integration loci can be retrieved using sequencing, these mutations can act as cancer gene tags, allowing discovery of novel cancer genes, e.g. [4, 5, 30, 35, 84–86]. However, integration biases can pose problems because they can be difficult to distinguish from the accumulation of integrations in cells that are under selective pressure to retain these integrations. Therefore, we compared

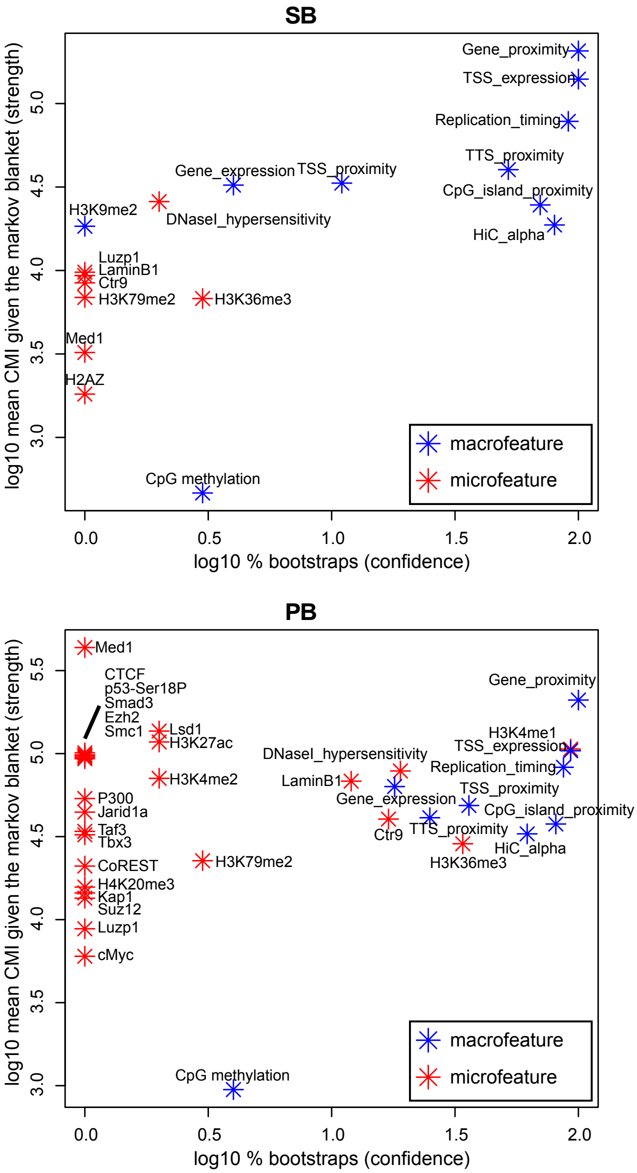


Figure 3.5: **Bootstrapped Markov blanket discovery.** Bayesian network inference (BNI) is performed on 400 bootstraps of size 20000. The x-axis represents the fraction of bootstraps that a feature occurs in the Markov blanket of integration proximity in a resulting Bayesian network, i.e. the confidence we have in an edge. The y-axis represents the mean conditional mutual information (CMI) of integration proximity with a feature across all Markov blankets in which this feature occurs, i.e. the strength of an edge. Note that features that do not occur in the Markov blanket of any bootstrap, i.e. are never considered relevant for integration proximity by the BNI approach, are not shown in this figure.

the unselected integration profiles with CIS integration profiles [4, 31–33].

Generally, the orientation biases of CIS integrations for genes and transcripts are much stronger than those of the unselected integrations (Figure S4). This indicates that in tumors, the orientation bias is mainly the result of selective pressure. For all systems and especially for PB and SB, there are significantly more unselected integrations than CIS integrations in regions other than genes and TSSs (Figure 6A; visualized at a 5% FDR threshold based on the binomial test. Refer to Table S2 for the raw *p*-values). Additionally, biases of unselected integrations for intergenic CIS regions (>100kb from genes) are relatively strong, compared to the biases for genic CIS regions (+/- 100kb) (Figure 6C; visualized at a 5% FDR threshold based on the binomial test. Refer to Table S2 for the raw *p*-values). Combined, these observations show that in regions far from genes, unselected integration profiles correlate relatively strongly with CIS profiles, compared to regions close to genes. Therefore, to avoid calling spurious CISs in IM screens, higher statistical stringency should be required for CISs found far from genes.

Combined, these observations suggest that CISs found far away from genes are more likely to be spurious, i.e. arise from the integration bias of the system. Conversely, the observations suggest that true CISs, i.e. CISs arising from selective pressure, are more often found in the vicinity of genes than would be expected based on an unselected integration profile.

To identify potentially spurious CISs, we tested for all CISs if the corresponding CIS regions contained significantly more CIS integrations than unselected integrations. This revealed a number of potentially spurious CISs, 13%, 33% and 7.4% of all CISs for SB, PB and MMTV respectively (Figure 6D, Tables S4, S5 and S6). For MMTV, it could be confirmed that potentially spurious CISs tended to be relatively far away from genes (One-sided Mann-Whitney U test; $p = 1.0 \times 10^{-2}$). For SB and PB, when ranking CISs according to increasing *p*-value, the potentially spurious CISs consisted mainly of lower ranking CISs (Mann-Whitney U test; $p = 4.7 \times 10^{-4}$ and $p < 10^{-7}$ for SB and PB respectively).

Next, we asked whether integration bias has an influence on the types of CISs that are found in screens. For this purpose, we separated CISs into activating and repressing CISs, based on orientation homogeneity and occurrence within or outside genes. We observed more activating CISs for PB than for SB (Figure 6B). Considering that the constructs used for the SB and PB tumor screens are similar [33, 87], this indicates that for use in IM screens, PB is more efficient at finding oncogenes, whereas SB would find more tumor suppressor genes.

3.3. Discussion

In this study, we have analyzed the integration biases of unselected integrations of a retrovirus and two transposons. For generating these sets of integrations, cells were grown in culture for three to four weeks. This implies that a few of our integration loci could potentially have been selected for. However, non-acute retroviruses, such as MMTV, induce tumors only very slowly (months to years) due to the absence of oncogenes in their genome [88]. Similarly, our transposon constructs can be described as non-acute in the sense that they do not carry oncogenes. Moreover, they do not contain any gene-trap or enhancer-trap elements, limiting the potential of disrupting endogenous gene expression. Hence, three or four weeks of cell culturing is a very short time

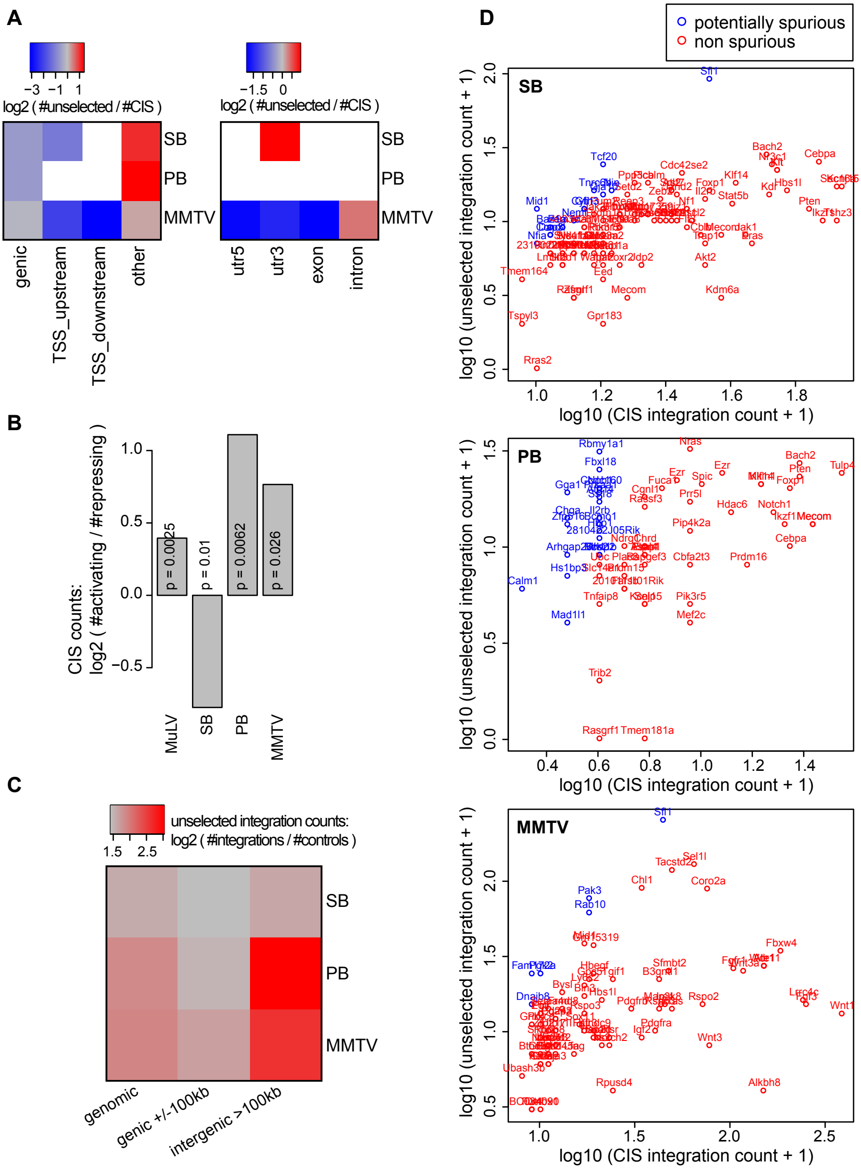


Figure 3.6: **Unselected integration profiles and CIS designation.** A) The bias of unselected integrations relative to CIS integrations, on a scale from blue (more CIS integrations) to red (more unselected integrations). B) $\log_2$ ratio of activating CISs and repressing CISs. A CIS is activating if it is not within a gene, or within a gene and 90% homogeneous with regard to orientation relative to that gene. Otherwise it is repressive. C) Bias of unselected integrations for CIS regions in a (i) genome-wide background, (ii) genic background (+/- 100kb), and (iii) intergenic background (whole genome except genes +/- 100kb), as measured by the $\log_2$ ratio of observed (unselected integrations) and expected (matched controls). D) CIS integration counts vs. unselected integration counts. CISs are annotated with the nearest TSS. Note that a single gene can be associated with multiple CISs. Spurious CISs were determined by a one-sided binomial test to determine if the CIS contained more CIS integrations than unselected integrations ($p < 0.01$, FDR-corrected).

frame compared to the latency to integration-induced tumor formation, and the influence of selected integration loci will be minimal at best. This is also supported by the observations that 1) we find a large number of unique integration sites, whereas in the case of substantial selection a relatively small number of (selected) integrations would be expected, and 2) while both PB and SB were mobilized from the same construct, we do obtain completely different insertion profiles for each transposon.

The main differences between the three systems regarding integration bias are summarized in Table 3. Generally, MMTV was observed to be the system least biased in its integration profile. Although many associations were found to be significant, they were generally not very substantial. In this context, it is surprising that only a small set of oncogenes has been tagged with this virus in IM screens [32, 89]. This could be due to activation of a limited set of promoters by the MMTV enhancer. Alternatively, certain unknown aspects of murine mammary tissue biology might allow only a limited number of tumorigenic mechanisms. In any case, our data rules out integration bias as a reason for the limited potential of IM by MMTV.

Recent availability of data describing the three-dimensional architecture of the genome [40] has allowed us to identify TAD - TAD boundary interfaces as hotspots of integration. Of the three systems, PB is most strongly affected, from a strong enrichment at TAD interfaces to a strong depletion towards the inner regions of TADs. While MMTV is largely indifferent regarding these inner TAD regions, its bias for TAD - TAD boundary interfaces is relatively strong. Although the TADs were defined in a different cell type (Tables 1 and 2), they have been shown to be stable across different cell types [40]. Altogether, our data show that integration target site selection is strongly associated with the topological organization of the genome.

The two transposons were found to have very different integration profiles. Generally, it is unknown to what extent certain sequences cloned into an IM construct affect the integration bias of that construct, which complicates the interpretation of differences observed between systems. However, the construct that was used in this study contained both the SB and PB IRs. Therefore, any difference between the two profiles can only be explained by the IR or the transposon-specific transposase, indicating that the IR and transposase are major defining elements of integration bias.

Although the SB and PB profiles are very different, they were shown to share a bias for activating macrofeatures and a bias against repressive macrofeatures. Analysis of a subset of macrofeatures for MMTV suggested that these may also operate as macrofeatures in a wider range of systems. Differences between SB and PB were mostly seen for the microfeatures. Together, these observations support a model where integration sites are selected at two levels (Figure 7). On larger scales, both systems target the same type of domains, determined by the macrofeatures. In particular, gene / TSS / TTS proximity, TSS expression, replication timing, CpG island proximity, and Hi-C alpha were found consistently indispensable for integration site selection in both systems (Figure 5). Once these domains have been selected, fine tuning of integration site selection is dependent on different microfeatures for each system. These microfeatures appear to be indispensable for integration site selection as well (Figure 4).

For the current study, the mESC model system was selected because it is the most thoroughly studied model system, with the broadest availability of (epi)genomic

Table 3.3: Summary of main observations.

	SB	PB	MMTV
Sequence	TA	TTAA	Very weak bias
Genes	Whole gene; bias against TSS	TSS; whole gene	TSS; weakly intergenic; bias against gene body
Transcripts	3’UTR; bias against 5’UTR / exon	Exon; 5’/3’UTR	5’/3’UTR
Orientation	Antisense within genes (introns)	Towards TSS	Sense within genes; upstream away from TSS
CIS targets	Tumor suppressor genes	Oncogenes	Oncogenes
spurious CISs	Lower ranking CISs	Lower ranking CISs	Farther away from genes
gene expression	Intermediate bias	Strong bias	Weak bias
TAD interface	Weak bias	Strong bias	Intermediate bias
Orientation w.r.t. chromatin marks	Towards chromatin marks (weak but consistent)	Towards chromatin marks (weak but consistent)	NA
macrofeatures	Relatively strong biases, consistent across systems: TSS expression, replication timing, Hi-C alpha, CpG island / TSS / TTS / gene proximity	Relatively strong biases, consistent across systems: TSS expression, replication timing, Hi-C alpha, CpG island / TSS / TTS / gene proximity	NA
microfeatures	Relatively many associations are negative	Strong positive associations with many features	NA

datasets. While limited cell type specificities have been demonstrated for retroviral integration profiles [23], earlier studies have used epigenomic features in non-matching cell types to analyze retroviral integration profiles [16, 24], noting that differences due to experimental error are generally greater than differences due to cell type [24]. We do not expect our transposon integration profiles to be highly cell type specific. This is supported by a supplementary analysis comparing the SB and PB integration profiles to a selected set of epigenomic features available for both mESC and mouse embryonic fibroblasts (mEF) [44, 58, 59, 90, 91], which shows that the mEF associations are highly similar to the mESC associations (Figure S8). This strong similarity suggests that cell type specificities are relatively weak. Nevertheless, it is interesting to note that the mEF associations are consistently slightly weaker than the mESC associations, indicating that cell type specificities, while weak, do exist.

In large scale IM screens, identification of overwhelming numbers of CISs is a serious impediment in distinguishing true CISs from spurious ones. In such screens, a true CIS arises through tumorigenic selection, whereas a spurious CIS is defined by the a priori integration bias of the IM system that was used. Our large datasets of unselected integrations can be used as a valuable resource for prioritizing candidate cancer genes emerging from IM screens. As a proof of principle, we showed that a substantial number of CIS regions in three recent IM screens do not contain more integrations than would be expected based on the unselected integration profiles. Although the cell types between PB and SB tumor screens and their corresponding unselected profiles are different, this nevertheless indicates that these CISs can potentially be explained by an a priori integration bias, and therefore likely represent passenger mutations.

In conclusion, the large numbers of integrations for three of the main systems used in IM, unselected integrations from cell lines and selected integrations from tumor screens, as well as the wide range of publicly available datasets, has enabled us to assess integration bias at unprecedented resolution, and assess its relation to CIS designation.

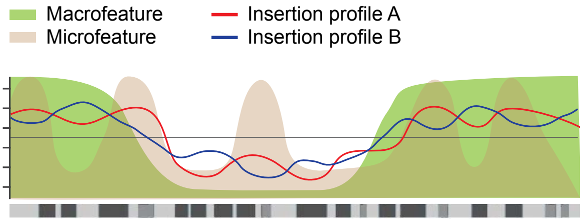


Figure 3.7: **Hierarchical model of integration target site selection.** On a large scale, target site selection is directed by macrofeatures for all three systems in similar ways. Differences between the systems are determined by microfeatures.

3.4. MATERIALS AND METHODS

3.4.1. DATA GENERATION

All integration data generated in this study are available on http://mutapedia.nki.nl.

PB AND SB INTEGRATION SITE DATA

mES cells EBRTcH3 expressing the tetracycline-controlled transactivator (tTA) from the endogenous ROSA26 promoter [92] were cultured in 60% BRL cell-conditioned medium in the presence of leukemia inhibitory factor, MEK inhibitor PD0325901 and GSK-3 inhibitor CHIR99021 [93]. pPB-SB-CMV-GFP was constructed by cloning GFP CDS in PB-MSCV [94] at Nru-I and BstXI sites. 4 hr before transfection, 6×10^6 EBRTcH3 cells were seeded on a 10cm dish. The cells were transfected with 12.5μg of pPB-SB-CMV-GFP and either 5μg of mPB transposase plasmid [95] or 12.5μg of SB100X transposase plasmid [96] using Lipofectamine 2000 (Invitrogen). Mock transfected and non-transfected controls were included. After 48 hr, 60000-80000 cells were isolated and further propagated (Figure S9). After three weeks of culturing post-transfection the genomic DNA was isolated using Qiagen DNeasy Blood & Tissue kit. 2μg of genomic DNA was digested with 20 units of Dpn-II (New England Biolabs) at 37°C for overnight in a 100μl reaction. 1μg of purified digested DNA was ligated with 0.8μM of splinkerette adapter using 10 units of T4 DNA ligase (Roche Applied Science) in a 50μl reaction. The splinkerette adapter was prepared by annealing equimolar amounts (40μM each) of Universal US[1] and Sau-3A-1 LS [2] oligos. The ligation reactions were amplified in two (SB) or three (PB) rounds of PCR to generate libraries for high throughput sequencing (for details see Table S7).

Sequencing was done on an Illumina HiSeq 2000 instrument to obtain single 100bp reads. The reads contained ends of IRs and the neighboring genomic DNA. The genomic DNA sequences were extracted from sequencing reads, and aligned against mouse genome assembly mm9 using Bowtie 2 [97] to determine the sites and orientation of integrations, using parameter '–very-sensitive-local'. Only those positions which were represented by five or more reads in the data, were retained and used for subsequent analyses.

[1] GTTCCCATGGTACTACTCATATAATACGACTCACTATAGG

[2] GATCCCTATAGTGAGTCGTATTATAATTTTTTTTTCAAAAAAA

MMTV integration site data

Mm5MT (MMTV producing cells) and NM-Pbabe/2 (NMuMG cells harboring PuroR transgene) were cultured in DMEM/F-12 + GlutaMAX-I medium supplemented with serum (10%) and Insulin (10μg/ml). For infection 0.5 million Mm5MT cells were plated in a T25 flask in the presence of 1.0μM Hydrocortisone. Next day the cells were treated with 25μg/ml of Mitomycin C in serum-free medium for two hours. Then 0.5 million NMuMG cells were cultured on top of the Mitomycin C treated Mm5MT cells in the presence of 1.0μM Hydrocortisone. Three days later the mixed cell culture was treated with 4.0μg/ml of Puromycin to remove Mm5MT cells. The remaining cells (NMuMG) were grown till passage 8 before the isolation of genomic DNA for integration site mapping (Figure S9). By using a primer pair, which was specific to MMTV in Mm5MT and did not amplify endogenous MMTV sequences in NMuMG cells, it was confirmed that NMuMG cells got infected. The integration sites were measured by two methods: either shearing the DNA with sonication and blunt end ligation of adapters as described previously [32] or cutting the DNA with Nla-III and ligation of adapters with sticky ends. 2μg of genomic DNA was digested with 20 units of Nla-III (New England Biolabs) at 37° C for overnight in a 100μl reaction. 1.0μg of purified digested DNA was ligated with 0.8μM of splinkerette adapter using 10 units of T4 DNA ligase (Roche Applied Science) in a 50μl reaction. The splinkerette adapter was prepared by annealing equimolar amounts (40μM each) of Universal LS[3] and Nla-III US[4] oligos. The ligated DNA was cut with Dra-I (New England Biolabs). The ligation reactions were amplified in two rounds of PCR to generate high throughput sequencing libraries (for details see Table S7). Sequencing was done on an Illumina HiSeq 2000 instrument to obtain single 100bp reads. The reads contained ends of MMTV LTR and the neighboring genomic DNA. The genomic DNA sequences were extracted from sequencing reads, and aligned against mouse genome assembly mm9 using Bowtie [98] to determine the sites and orientation of integrations.

3.4.2. Data preprocessing

Matched random controls

Given an integration dataset (either one of SB, PB, or MMTV), each integration in that dataset was matched to 10 random controls. These random controls were subject to a number of criteria. First, specifically for SB and PB, matched controls were restricted to loci containing the system-specific integration motif (TA and TTAA respectively). Second, the distance of the matched control to the nearest restriction site upstream of the integration was required to be the same as that of the integration itself. Third, matched controls were not allowed to fall within ’unmappable’ regions. Here, unmappable regions were defined in a dataset-dependent manner. Given an integration dataset (either one of SB, PB, or MMTV), the sequence read length n was determined. Then, the mouse genome (mm9) was cut up into all possible sequences of length n. These artificial reads were mapped to the mm9 genome using the same tool and parameter settings as used to generate the integration datasets (see above). Unmappable regions were then defined as regions that did not have any reads mapped to them using this approach, and controls

[3] CCTATAGTGAGTCGTATTATAATTTTTTTTTCAAAAAAA

[4] GTTCCCATGGTACTACTCATATAATACGACTCACTATAGGCATG

were excluded from these regions.

ChIP-seq

To maximize the comparability of the ChIP-seq datasets, they were processed from the sequence read archives as obtained from GEO [99], where possible, in exactly the same way. Sequence read archives were converted to FASTA format and then aligned against mouse genome assembly mm9 using Bowtie 0.12.7 [98], allowing at most 2 mismatches in end-to-end alignment (the following settings were used: -M 1 –best –tryhard -v 2 –chunkmbs 1024). Duplicate reads were removed, and a 25bp coverage was computed by counting the number of reads in 25bp consecutive bins. These coverage profiles were normalized to a sequencing depth of 10^9, smoothed (running mean with window n = 6), and sampled (to 100bp spacing). Then, all available input DNA datasets (9 in total) were collected and clustered based on the coverage profile. A cluster of 6 input DNA datasets with a correlation of at least 0.97 was selected to be pooled and used as control for all non-histone mark features. This was done because 1) not for all datasets controls were available, 2) not for all datasets controls of the same type (input DNA, GFP, mock IP, etc.) were available. The 6 selected input DNA datasets were used as a control dataset. First, after normalizing to total read count, they were averaged to obtain a pooled control coverage profile, and then used to normalize all non-histone mark features by computing $\log_2$(signal/control). For the histone marks, a similar approach was taken, using all available pan-H3 datasets (2 in total).

Bisulfite sequencing

Bisulfite sequencing reads [41] were processed using Methylcoder [100], with "–mismatches=0", and Bowtie to align the reads, with "-M 1 –best –tryhard -v 2 –chunkmbs 1024". A coverage profile was computed by selecting only methylation context 'CG' for CpG methylation, and counting in 25bp consecutive bins the numbers of unconverted Cs c, converted Cs t, and methylable basepairs n, and calculating $\frac{c}{n(c+t)}$. The resulting profile was smoothed and sampled as explained above.

RNA-seq

The RNA-seq reads were processed using Cufflinks [101] to compute the $\log_2$(FPKM+1) for each gene, where FPKM refers to the number of fragments per kilobase of exon per million fragments mapped.

Preprocessing of microarray datasets

The H3K9me2, late replication timing data, and LaminB1 data were downloaded from GEO and processed as in the corresponding publications [37–39].

3.4.3. Data analysis

Gene alignments (Figure 1A)

Gene locations were retrieved from the Ensembl database (release 66). Partially overlapping genes were removed (40%). In case of complete overlap, the larger gene was retained. The remaining genes (n = 22822) were aligned with respect to transcription start sites and transcription termination sites. For each integration dataset, integrations and controls (see above) were counted in equal-sized bins outside genes, and gene length dependent bins within genes. Then, for each bin a ratio was computed of integration counts versus control counts. This ratio was normalized by multiplication with the ratio of control dataset size and integration dataset size. Then the base 2 logarithm was taken.

BIAS WITH RESPECT TO GENES AND TRANSCRIPTS (FIGURE 1B)

For genes (Ensembl release 66), integrations and controls (see above) were counted within genic regions, TSS upstream regions (defined as the union of those regions within 1kb upstream of a TSS), TSS downstream regions (defined as the union of those regions within 1kb downstream of a TSS), and everything else. Note that these classes can overlap. Then, the ratio was computed of integration counts versus control counts. This ratio was normalized by multiplication with the ratio of control dataset size and integration dataset size. Then the base 2 logarithm was taken. A similar approach was taken for the transcript-related classes 5'UTR, 3'UTR, exon, and intron.

BIAS WITH RESPECT TO GENE EXPRESSION (FIGURE 2)

Genes were divided into five quantiles, based on their expression level. For SB and PB, Group 1 consisted of all genes with an FPKM of zero in the RNA-Seq dataset. The remaining genes were divided across four equal quantiles. The NMuMG microarray expression dataset for MMTV was divided according to the same expression quantiles. Subsequently, the same approach as above was taken for determining the numbers of integrations within each of these subsets of genes.

TOPOLOGICALLY ASSOCIATED DOMAINS (FIGURE 3)

Domain definitions were adopted from [40]. Interfaces between TADs and TAD boundaries were aligned, and integrations and controls (see above) were counted until halfway into the TAD as well as halfway into the TAD boundary region. Then, a $\log_2$ ratio between the two was calculated.

ASSOCIATION OF INTEGRATION OCCURRENCE WITH GENOME-WIDE FEATURES (FIGURE 4)

For each feature and integration, a feature score was computed at exponentially increasing distances (scales) from that integration. For all deep sequencing based features, this was done by taking the mean normalized read count within a 200bp window, from the genome-wide binned read count profiles computed as outlined above. For CpG islands, genes, and TSSs, this score was calculated by taking the negative $\log_2$ transformed distance (+ 1) to the nearest CpG island, gene, or TSS. For the microarray features, this score represented the value of the nearest probe. The Hi-C score was computed as follows. Normalized Hi-C contact frequency matrices (20kb bins) were downloaded from (http://chromosome.sdsc.edu/mouse/hi-c/mESC.norm.tar.gz). For each locus, defined by an integration and a scale, average contact frequencies as a function of distance from that locus, were calculated within a window of 400kb. The Hi-C α was computed as the slope of a linear regression fit to the $\log_1 0$ transformed distances and $\log_1 0$ transformed contact frequencies. Once the feature scores for all triples (feature, scale, and integration) were calculated, feature scores were rank-normalized on a per-feature basis, and a t-score, $t_{i,j}^{\text{integration}}$, was computed for each scale i and feature j as follows:

$$t_{i,j}^{\text{integration}} = \frac{\overline{x}_{0,j} - \overline{x}_{i,j}}{s_{0,i,j} \times \sqrt{\frac{1}{n} + \frac{1}{2n}}} \tag{3.1}$$

where

$$s_{0,i,j} = \sqrt{\frac{(n-1)\,s_{0,j}^2 + (2n-1)\,s_{i,j}^2}{3n-2}} \tag{3.2}$$

Here, $\overline{x}_{0,j}$ represents the mean of all scores of feature j at the sites of integration, $\overline{x}_{i,j}$ the mean of all scores of feature j at scale i, and $s_{0,j}^2$ and $s_{i,j}^2$ their respective variances. n represents the number of integrations.

The same was done for the 10 sets of control loci, since we have 10 matched controls for each integration. This resulted in 1 set of integration t-scores, $t_{i,j}^{\text{integration}}$, and 10 sets of control t-scores, $t_{i,j}^{\text{control}_k}$. Then, the difference between the set of integration t-scores and the mean of the control t-scores was computed (and plotted in Figure 4):

$$t_{i,j}^{\Delta} = t_{i,j}^{\text{integration}} - \frac{\sum_{k=1}^{10} t_{i,j}^{\text{control}_k}}{10} \tag{3.3}$$

To compute p-values for the $t_{i,j}^{\Delta}$, we can take advantage of the fact that for large degrees of freedom, the t-distribution converges to the standard normal distribution. Thus, the calculated t-scores can be interpreted as normally distributed with mean 0 and standard deviation 1, i.e. as 1 set of integration z-scores, $z_{i,j}^{\text{integration}}$, and 10 sets of control z-scores, $z_{i,j}^{\text{control}_k}$:

$$z_{i,j}^{\Delta} = z_{i,j}^{\text{integration}} - \frac{\sum_{k=1}^{10} z_{i,j}^{\text{control}_k}}{10} \tag{3.4}$$

Now, note that if:

$$\begin{aligned} x &\sim N(\mu_x, \sigma_x^2) \\ y &\sim N(\mu_y, \sigma_y^2) \\ z &= x + y \end{aligned}$$

Then

$$z \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

Note furthermore, that if a is some constant and

$$x \sim N(\mu_x, \sigma_x^2)$$
$$z = a \times x$$

Then

$$z \sim N(a \times \mu_x, a^2 \times \sigma_x^2)$$

Therefore

$$z_{i,j}^{\Delta} \sim N(0, 11/10) \tag{3.5}$$

The p-values for this distribution are readily computed.

CONDITIONAL INDEPENDENCIES (FIGURE 5)

For each integration, the feature values were normalized by division with the average of the 10 corresponding control feature values (or subtraction in case of log-transformed feature values). Then, all data was discretized into three quantiles. To identify conditional independencies in the discretized data we took the approach of Bayesian networks [83]. Bayesian networks specify for each feature a set of other features, the Markov blanket. A Markov blanket of a feature represents a minimal set of features sufficient to characterize the distribution of that feature. More formally, given its Markov blanket, a feature is conditionally independent from all other features. In addition to modeling conditional independencies, Bayesian networks can capture nonlinear effects, such as the changes of sign of association across scales in Figure 4. For inferring Bayesian networks, we used BANJO [82]. Since BANJO relies on stochastic optimization (simulated annealing), we inferred Bayesian networks for 400 bootstraps of size 20000. For each bootstrap, the Markov blanket of the integration proximity node was determined, and the importance of a feature was represented in two dimensions: First, the fraction of bootstraps a feature occurred in a resulting Markov blanket. This conveys the degree of confidence we have that this feature is truly relevant. Second, the conditional mutual information between integration proximity and a feature from its Markov blanket, given the other features in its Markov blanket, averaged across all inferred Markov blankets. This coveys the strength of association between integration proximity and this feature, where this strength could not be explained by any of the other features that were inferred to be relevant.

UNSELECTED VS. CIS INTEGRATIONS (FIGURE 6A)

For each of the three systems, CISs were called on the tumor screens, using the approach outlined in [34], with a 30kb kernel width and a 5% Bonferroni-corrected p-value threshold. CIS regions were defined as those regions where the Gaussian smoothing kernel exceeded the significance threshold, extended on either side with 30kb (kernel width used for calling the CISs). CIS integrations were defined as those integrations from the tumor screen that fell within a CIS region, and $\log_2$ ratios of unselected integrations and CIS integrations were calculated as described above, replacing control loci with CIS integrations.

Activating/repressing CISs (Figure 6B)

CISs were called as outlined above. A CIS was defined to be an activating CIS if its peak location was either not within a gene, or within a gene and orientation-wise homogeneous, requiring 90% of integrations falling within a CIS to be of the same orientation.

CIS region bias of unselected integrations (Figure 6C)

For the union of genic regions (+/- 100kb) it was counted how many unselected integrations were found within CIS regions, how many outside CIS regions, and the corresponding numbers of expected unselected integrations in those regions; p-values were calculated based on a binomial test. This procedure outlined for genic regions was repeated for the whole genome, and for intergenic regions.

Potentially spurious CISs (Figure 6D)

For all CISs, it was determined how many CIS integrations and unselected integrations fell within the corresponding CIS region. CIS regions were defined as above. To determine whether a CIS contained significantly more CIS integrations than unselected integrations, a one-sided binomial test was performed, testing the significance of n_s successes in $(n_s + n_u)$ trials, and corrected for multiple testing (FDR). Here, n_s is the number of tumor screen integrations within a CIS region, and n_u is the number of unselected integrations within a CIS region. Low (high) p-values correspond to true (spurious) CISs. The probability of success for each binomial test was defined as $\frac{s}{s+u}$, where s is the tumor screen dataset size, and u is the unselected integration dataset size.

CISs were annotated with the name of the gene of nearest TSS, where the TSSs were restricted to Ensembl IDs that had corresponding UCSC, EntrezGene, MGI, and UniGene IDs. To determine whether the non-significant CISs tended to be farther away from genes, the genome-wide TSS density was estimated using a Gaussian smoothing kernel (standard deviation 1Mb), and sampled at 1kb intervals. Integrations were then mapped to the nearest sampled density estimation point, and a Mann-Whitney U test was performed on selected integrations within CIS regions versus unselected integrations within CIS regions.

Acknowledgments

We would like to acknowledge Bas van Steensel, Alexey Pindyurin, Hayssam Soueidan, Sander Canisius and David Adams for their critical readings of the manuscript, and their thoughtful comments.

References

[1] J. de Jong, W. Akhtar, J. Badhai, A. G. Rust, R. Rad, J. Hilkens, A. Berns, M. van Lohuizen, L. F. A. Wessels, and J. de Ridder, *Chromatin landscapes of retroviral and transposon integration profiles,* PLoS Genet **10**, e1004250+ (2014).

[2] N. Cartier, S. Hacein-Bey-Abina, C. C. Bartholomae, G. Veres, M. Schmidt, I. Kutschera, M. Vidaud, U. Abel, L. Dal-Cortivo, L. Caccavelli, N. Mahlaoui, V. Kiermer, D. Mittelstaedt, C. Bellesme, N. Lahlou, F. Lefrere, S. Blanche, M. Audit, E. Payen, P. Leboulch, B. l'Homme, P. Bougneres, C. Von Kalle, A. Fischer,

M. Cavazzana-Calvo, and P. Aubourg, *Hematopoietic stem cell gene therapy with a lentiviral vector in X-Linked adrenoleukodystrophy,* Science **326**, 818 (2009).

[3] A. Fischer, S. Hacein-Bey-Abina, and M. Cavazzana-Calvo, *20 years of gene therapy for SCID,* Nat Immunol **11**, 457 (2010).

[4] A. G. Uren, J. Kool, K. Matentzoglu, J. de Ridder, J. Mattison, M. van Uitert, W. Lagcher, D. Sie, E. Tanger, T. Cox, M. Reinders, T. J. Hubbard, J. Rogers, J. Jonkers, L. Wessels, D. J. Adams, M. van Lohuizen, and A. Berns, *Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks.* Cell **133**, 727 (2008).

[5] J. Mattison, J. Kool, A. G. Uren, J. de Ridder, L. Wessels, J. Jonkers, G. R. Bignell, A. Butler, A. G. Rust, M. Brosch, C. H. Wilson, L. van der Weyden, D. A. Largaespada, M. R. Stratton, P. A. Futreal, M. van Lohuizen, A. Berns, L. S. Collier, T. Hubbard, and D. J. Adams, *Novel candidate cancer genes identified by a large-scale cross-species comparative oncogenomics approach.* Cancer Res **70**, 883 (2010).

[6] W. Akhtar, J. de Jong, A. V. Pindyurin, L. Pagie, W. Meuleman, J. de Ridder, A. Berns, L. F. A. Wessels, M. van Lohuizen, and B. van Steensel, *Chromatin position effects assayed by thousands of reporters integrated in parallel.* Cell **154**, 914 (2013).

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

[8] E. H. Miller, G. Obernosterer, M. Raaben, A. S. Herbert, M. S. Deffieu, A. Krishnan, E. Ndungo, R. G. Sandesara, J. E. Carette, A. I. Kuehne, G. Ruthel, S. R. Pfeffer, J. M. Dye, S. P. Whelan, T. R. Brummelkamp, and K. Chandran, *Ebola virus entry requires the host-programmed recognition of an intracellular receptor,* EMBO J (2012).

[9] P. Bouwman, A. Aly, J. M. Escandell, M. Pieterse, J. Bartkova, H. van der Gulden, S. Hiddingh, M. Thanasoula, A. Kulkarni, Q. Yang, B. G. Haffty, J. Tommiska, C. Blomqvist, R. Drapkin, D. J. Adams, H. Nevanlinna, J. Bartek, M. Tarsounas, S. Ganesan, and J. Jonkers, *53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers,* Nat Struct & Mol Biol **17**, 688 (2010).

[10] S. Hacein-Bey-Abina, A. Garrigue, G. P. Wang, J. Soulier, A. Lim, E. Morillon, E. Clappier, L. Caccavelli, E. Delabesse, K. Beldjord, V. Asnafi, E. MacIntyre, L. Dal Cortivo, I. Radford, N. Brousse, F. Sigaux, D. Moshous, J. Hauer, A. Borkhardt, B. H. Belohradsky, U. Wintergerst, M. C. Velez, L. Leiva, R. Sorensen, N. Wulffraat, S. Blanche, F. D. Bushman, A. Fischer, and M. Cavazzana-Calvo, *Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of scid-x1,* J Clin Invest **118**, 3132 (2008).

[11] G. Liu, A. M. Geurts, K. Yae, A. R. Srinivasan, S. C. Fahrenkrug, D. A. Largaespada, J. Takeda, K. Horie, W. K. Olson, and P. B. Hackett, *Target-site preferences of sleeping beauty transposons.* J Mol Biol **346**, 161 (2005).

[12] A. M. Geurts, C. S. Hackett, J. B. Bell, T. L. Bergemann, L. S. Collier, C. M. Carlson, D. A. Largaespada, and P. B. Hackett, *Structure-based prediction of insertion-site preferences of transposons into chromosomes.* Nucleic acids research **34**, 2803 (2006).

[13] N. G. Copeland and N. A. Jenkins, *Harnessing transposons for cancer gene discovery.* Nat Rev Cancer **10**, 696 (2010).

[14] T. Vandendriessche, Z. Ivics, Z. Izsvak, and M. K. Chuah, *Emerging potential of transposons for gene therapy and generation of induced pluripotent stem cells,* Blood **114**, 1461 (2009).

[15] Q. Liang, J. Kong, J. Stalker, and A. Bradley, *Chromosomal mobilization and reintegration of sleeping beauty and piggybac transposons.* Genesis **47**, 404 (2009).

[16] C. Berry, S. Hannenhalli, J. Leipzig, and F. D. Bushman, *Selection of target sites for mobile DNA integration in the human genome,* PLoS Comput. Biol **2**, e157 (2006).

[17] B. Balu, C. Chauhan, S. Maher, D. Shoue, J. Kissinger, M. Fraser, and J. Adams, *piggyBac is an effective tool for functional analysis of the plasmodium falciparum genome,* BMC Microbiol **9**, 83+ (2009).

[18] D. L. Galvan, Y. Nakazawa, A. Kaja, C. Kettlun, L. J. Cooper, C. M. Rooney, and M. H. Wilson, *Genome-wide mapping of PiggyBac transposon integrations in primary human T cells.* J Immunother **32**, 837 (2009).

[19] M. A. Li, S. J. Pettitt, S. Eckert, Z. Ning, S. Rice, J. Cadiñanos, K. Yusa, N. Conte, and A. Bradley, *The piggybac transposon displays local and distant reintegration preferences and can cause mutations at non-canonical integration sites,* Mol Cell Biol (2013).

[20] W. Wang, C. Lin, D. Lu, Z. Ning, T. Cox, D. Melvin, X. Wang, A. Bradley, and P. Liu, *Chromosomal transposition of PiggyBac in mouse embryonic stem cells,* P Natl Acad of Sci USA **105**, 9290 (2008).

[21] A. Faschinger, F. Rouault, S. Johannes, A. Lukas, B. Salmons, W. H. Günzburg, and S. Indik, *Mouse Mammary Tumor Virus integration site selection in human and mouse genomes,* J Virol **82**, 1360 (2007).

[22] B. Felice, C. Cattoglio, D. Cittaro, A. Testa, A. Miccio, G. Ferrari, L. Luzi, A. Recchia, and F. Mavilio, *Transcription factor binding sites are genetic determinants of retroviral integration in the human genome.* PLoS One **4**, e4571 (2009).

[23] F. A. Santoni, O. Hartley, and J. Luban, *Deciphering the code for retroviral integration target site selection,* PLoS Comput Biol **6**, e1001008 (2010).

[24] T. Brady, Y. N. Lee, K. Ronen, N. Malani, C. C. Berry, P. D. Bieniasz, and F. D. Bushman, *Integration target site selection by a resurrected human endogenous retrovirus,* Genes & Development **23**, 633 (2009).

[25] Y. Kodama, M. Shumway, R. Leinonen, and International Nucleotide Sequence Database Collaboration, *The sequence read archive: explosive growth of sequencing data.* Nucleic acids research **40** (2012).

[26] T. L. Bergemann, T. K. Starr, H. Yu, M. Steinbach, J. Erdmann, Y. Chen, R. T. Cormier, D. A. Largaespada, and K. A. T. Silverstein, *New methods for finding common insertion sites and co-occurring common insertion sites in transposon- and virus-based genetic screens.* Nucleic Acids Res **40**, 3822 (2012).

[27] U. Abel, A. Deichmann, A. Nowrouzi, R. Gabriel, C. C. Bartholomae, H. Glimm, C. von Kalle, and M. Schmidt, *Analyzing the number of common integration sites of viral vectors – new methods and computer programs,* PLoS ONE **6**, e24247 (2011).

[28] A. Biffi, C. Bartolomae, D. Cesana, N. Cartier, P. Aubourg, M. Ranzani, M. Cesani, F. Benedicenti, T. Plati, E. Rubagotti, S. Merella, A. Capotondo, J. Sgualdino, G. Zanetti, C. von Kalle, M. Schmidt, L. Naldini, and E. Montini, *Lentiviral vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection,* Blood **117**, 5332 (2011).

[29] X. Wu, B. Luke, and S. Burgess, *Redefining the common insertion site,* Virology **344**, 292 (2006).

[30] T. K. Starr, R. Allaei, K. A. T. Silverstein, R. A. Staggs, A. L. Sarver, T. L. Bergemann, M. Gupta, M. G. O'Sullivan, I. Matise, A. J. Dupuy, L. S. Collier, S. Powers, A. L. Oberg, Y. W. Asmann, S. N. Thibodeau, L. Tessarollo, N. G. Copeland, N. A. Jenkins, R. T. Cormier, and D. A. Largaespada, *A Transposon-Based genetic screen in mice identifies genes altered in colorectal cancer,* Science **323**, 1747 (2009).

[31] J. de Jong, J. de Ridder, L. van der Weyden, N. Sun, M. van Uitert, A. Berns, M. van Lohuizen, J. Jonkers, D. J. Adams, and L. F. A. Wessels, *Computational identification of insertional mutagenesis targets for cancer gene discovery,* Nucleic Acids Res **39**, e105 (2011).

[32] M. J. Koudijs, C. Klijn, L. van der Weyden, J. Kool, J. ten Hoeve, D. Sie, P. R. Prasetyanti, E. Schut, S. Kas, T. Whipp, E. Cuppen, L. Wessels, D. J. Adams, and J. Jonkers, *High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors,* Genome Res **21**, 2181 (2011).

[33] R. Rad, L. Rad, W. Wang, J. Cadinanos, G. Vassiliou, S. Rice, L. S. Campos, K. Yusa, R. Banerjee, M. A. Li, J. de la Rosa, A. Strong, D. Lu, P. Ellis, N. Conte, F. T. Yang, P. Liu, and A. Bradley, *PiggyBac transposon mutagenesis: A tool for cancer gene discovery in mice,* Science **330**, 1104 (2010).

[34] J. de Ridder, A. Uren, J. Kool, M. Reinders, and L. Wessels, *Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens.* PLoS Comput Biol **2** (2006).

[35] J. Kool, A. G. Uren, C. P. Martins, D. Sie, J. de Ridder, G. Turner, M. van Uitert, K. Matentzoglu, W. Lagcher, P. Krimpenfort, J. Gadiot, C. Pritchard, J. Lenz, A. H.

Lund, J. Jonkers, J. Rogers, D. J. Adams, L. Wessels, A. Berns, and M. van Lohuizen, *Insertional mutagenesis in mice deficient for p15Ink4b, p16Ink4a, p21Cip1, and p27Kip1 reveals cancer gene interactions and correlations with tumor phenotypes,* Cancer Res **70**, 520 (2010).

[36] C. Chang, X. Yang, B. Pursell, and A. M. Mercurio, *Id2 complexes with the snag domain of snai1 inhibiting snai1-mediated repression of integrin beta-4.* Mol Cell Biol (2013).

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

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

[39] D. Peric-Hupkes, W. Meuleman, L. Pagie, S. W. M. Bruggeman, I. Solovei, W. Brugman, S. Gräf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, M. Reinders, L. Wessels, and B. van Steensel, *Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.* Mol Cell **38**, 603 (2010).

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

[41] Y. Xu, F. Wu, L. Tan, L. Kong, L. Xiong, J. Deng, A. J. Barbera, L. Zheng, H. Zhang, S. Huang, J. Min, T. Nicholson, T. Chen, G. Xu, Y. Shi, K. Zhang, and Y. G. G. Shi, *Genome-wide regulation of 5hmC, 5mC, and gene expression by tet1 hydroxylase in mouse embryonic stem cells.* Mol Cell **42**, 451 (2011).

[42] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Harrow, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. J. Searle, *Ensembl 2012,* Nucleic Acids Res **40**, D84 (2012).

[43] J. Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, E. Giste, A. Johnson, M. Zhang, G. Balasundaram, R. Byron, V. Roach, P. Sabo, R. Sandstrom, A. S. Stehling, R. Thurman, S. Weissman, P. Cayting, M. Hariharan, J. Lian, Y. Cheng, S. Landt, Z. Ma,

B. Wold, and J. Dekker, *An encyclopedia of mouse DNA elements (mouse ENCODE),* Genome Biol **13**, 418+ (2012).

[44] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O/'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein, *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature **448**, 553 (2007).

[45] M. M. Karimi, P. Goyal, I. A. Maksakova, M. Bilenky, D. Leung, J. Xin, Y. Shinkai, D. L. Mager, S. Jones, M. Hirst, and M. C. Lorincz, *DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs.* Cell Stem Cell **8**, 676 (2011).

[46] S. Xiao, D. Xie, X. Cao, P. Yu, X. Xing, C.-C. Chen, M. Musselman, M. Xie, F. D. West, H. A. Lewin, T. Wang, and S. Zhong, *Comparative epigenomic annotation of regulatory DNA.* Cell **149**, 1381 (2012).

[47] M. J. Law, K. M. Lower, H. P. J. Voon, J. R. Hughes, D. Garrick, V. Viprakasit, M. Mitson, M. De Gobbi, M. Marra, A. Morris, A. Abbott, S. P. Wilder, S. Taylor, G. M. Santos, J. Cross, H. Ayyub, S. Jones, J. Ragoussis, D. Rhodes, I. Dunham, D. R. Higgs, and R. J. Gibbons, *ATR-x syndrome protein targets tandem repeats and influences Allele-Specific expression in a Size-Dependent manner,* Cell **143**, 367 (2010).

[48] L. Ho, R. Jothi, J. L. Ronan, K. Cui, K. Zhao, and G. R. Crabtree, *An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network.* P Natl Acad Sci USA **106**, 5187 (2009).

[49] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y.-H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W.-K. Sung, N. D. Clarke, C.-L. Wei, and H.-H. Ng, *Integration of external signaling pathways with the core transcriptional network in embryonic stem cells,* Cell **133**, 1106 (2008).

[50] A. Marson, S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love, N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch, and R. A. Young, *Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells.* Cell **134**, 521 (2008).

[51] M. Ku, R. P. Koche, E. Rheinbay, E. M. Mendenhall, M. Endoh, T. S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie, A. S. Chi, M. Adli, S. Kasif, L. M. Ptaszek, C. A. Cowan, E. S. Lander, H. Koseki, and B. E. Bernstein, *Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains,* PLoS Genet **4**, e1000242+ (2008).

[52] S. Bilodeau, M. H. Kagey, G. M. Frampton, P. B. Rahl, and R. A. Young, *SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state,* Gene Dev **23**, 2484 (2009).

[53] G. Li, R. Margueron, M. Ku, P. Chambon, B. E. Bernstein, and D. Reinberg, *Jarid2 and PRC2, partners in regulating gene expression,* Gene Dev **24**, 368 (2010).

[54] J. Han, P. Yuan, H. Yang, J. Zhang, B. S. Soh, P. Li, S. L. Lim, S. Cao, J. Tay, Y. L. Orlov, T. Lufkin, H.-H. Ng, W.-L. Tam, and B. Lim, *Tbx3 improves the germ-line competency of induced pluripotent stem cells,* Nature **463** (2010).

[55] P. B. Rahl, C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, P. A. Sharp, and R. A. Young, *c-Myc regulates transcriptional pause release,* Cell **141**, 432 (2010).

[56] A. R. Krebs, J. Demmers, K. Karmodiya, N.-C. Chang, A. C. Chang, and L. Tora, *ATAC and mediator coactivators form a stable complex and regulate a set of non-coding RNA genes,* EMBO Rep **11**, 541 (2010).

[57] M. P. Schnetz, L. Handoko, B. Akhtar-Zaidi, C. F. Bartels, C. F. Pereira, A. G. Fisher, D. J. Adams, P. Flicek, G. E. Crawford, T. Laframboise, P. Tesar, C.-L. L. Wei, and P. C. Scacheri, *CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression,* PLoS Genet **6** (2010).

[58] M. H. Kagey, J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young, *Mediator and cohesin connect gene expression and chromatin architecture.* Nature **467**, 430 (2010).

[59] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch, *Histone H3K27ac separates active from poised enhancers and predicts developmental state,* P Natl Acad Sci USA **107**, 21931 (2010).

[60] E. M. Mendenhall, R. P. Koche, T. Truong, V. W. Zhou, B. Issac, A. S. Chi, M. Ku, and B. E. Bernstein, *GC-rich sequence elements recruit PRC2 in mammalian ES cells,* PLoS Genet **6**, e1001244+ (2010).

[61] L. Handoko, H. Xu, G. Li, C. Y. Y. Ngan, E. Chew, M. Schnapp, C. W. H. W. Lee, C. Ye, J. L. H. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W.-K. K. Sung, Y. Ruan, and C.-L. L. Wei, *CTCF-mediated functional chromatin interactome in pluripotent cells.* Nat Genet **43**, 630 (2011).

[62] H. Hezroni, B. S. Sailaja, and E. Meshorer, *Pluripotency-related, valproic acid (VPA)-induced genome-wide histone h3 lysine 9 (H3K9) acetylation patterns in embryonic stem cells,* J Biol Chem **286**, 35977 (2011).

[63] Z. Liu, D. R. Scannell, M. B. Eisen, and R. Tjian, *Control of embryonic stem cell lineage commitment by core promoter factor, TAF3.* Cell **146**, 720 (2011).

[64] S. U. Schmitz, M. Albert, M. Malatesta, L. Morey, J. V. Johansen, M. Bak, N. Tommerup, I. Abarrategui, and K. Helin, *Jarid1b targets genes regulating development and is involved in neural differentiation.* EMBO J (2011).

[65] A. C. Mullen, D. A. Orlando, J. J. Newman, J. Lovén, R. M. Kumar, S. Bilodeau, J. Reddy, M. G. Guenther, R. P. DeKoter, and R. A. Young, *Master transcription factors determine cell-type-specific responses to TGF-β signaling.* Cell **147**, 565 (2011).

[66] S. Quenneville, G. Verde, A. Corsinotti, A. Kapopoulou, J. Jakobsson, S. Offner, I. Baglivo, P. V. Pedone, G. Grimaldi, A. Riccio, and D. Trono, *In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions,* Mol Cell **44**, 361 (2011).

[67] M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schubeler, *DNA-binding factors shape the mouse methylome at distal regulatory regions,* Nature **480**, 490 (2011).

[68] C. Creppe, P. Janich, N. Cantarino, M. Noguera, V. Valero, E. Musulen, J. Douet, M. Posavec, J. Martin-Caballero, L. Sumoy, L. Di Croce, S. A. Benitah, and M. Buschbeck, *Macroh2a1 regulates the balance between self-renewal and differentiation commitment in embryonic and adult stem cells.* Mol Cell Biol **32**, 1442 (2012).

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

[70] L. Tavares, E. Dimitrova, D. Oxley, J. Webster, R. Poot, J. Demmers, K. Bezstarosti, S. Taylor, H. Ura, H. Koide, A. Wutz, M. Vidal, S. Elderkin, and N. Brockdorff, *RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3.* Cell **148**, 664 (2012).

[71] W. A. Whyte, S. Bilodeau, D. A. Orlando, H. A. Hoke, G. M. Frampton, C. T. Foster, S. M. Cowley, and R. A. Young, *Enhancer decommissioning by LSD1 during embryonic stem cell differentiation,* Nature **advance online publication** (2012).

[72] H. Jiang, A. Shukla, X. Wang, W.-y. Chen, B. E. Bernstein, and R. G. Roeder, *Role for dpy-30 in ES Cell-Fate specification by regulation of H3K4 methylation within bivalent domains,* Cell **144**, 825 (2011).

[73] O. Yildirim, R. Li, J.-H. H. Hung, P. B. Chen, X. Dong, L.-S. S. Ee, Z. Weng, O. J. Rando, and T. G. Fazzio, *Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells.* Cell **147**, 1498 (2011).

[74] R. Young, *ChIP-seq for Mcaf1 (unpublished),* (2011), GEO accession: GSE26680.

[75] R. Young, *ChIP-seq for Tbp (unpublished),* (2010), GEO accession: GSE22303.

[76] R. Young, *ChIP-seq for PolII-Ser18P (unpublished),* (2010), GEO accession: GSE21917.

[77] E. R. Smith, C. Lin, A. S. Garrett, J. Thornton, N. Mohaghegh, D. Hu, J. Jackson, A. Saraf, S. K. Swanson, C. Seidel, L. Florens, M. P. Washburn, J. C. Eissenberg, and A. Shilatifard, *The little elongation complex regulates small nuclear RNA transcription,* Mol Cell **44**, 954 (2011).

[78] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, *Simple combinations of Lineage-Determining transcription factors prime cis-Regulatory elements required for macrophage and b cell identities,* Mol Cell **38**, 576 (2010).

[79] R. S. Mitchell, B. F. Beitzel, A. R. Schroder, P. Shinn, H. Chen, C. C. Berry, J. R. Ecker, and F. D. Bushman, *Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences,* PLoS Biol **2**, e234 (2004).

[80] X. Wu, Y. Li, B. Crise, and S. M. Burgess, *Transcription Start Regions in the Human Genome Are Favored Targets for MLV Integration,* Science **300**, 1749 (2003).

[81] J.-M. M. Belton, R. P. P. McCord, J. H. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, *Hi-C: A comprehensive technique to capture the conformation of genomes.* Methods **58**, 268 (2012).

[82] A. J. Hartemink, *Reverse engineering gene regulatory networks,* Nat Biotechnol **23**, 554 (2005).

[83] J. Pearl, *Bayesian networks: A model of self-activated memory for evidential reasoning,* in *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine* (1985) pp. 329–334.

[84] H. Mikkers, J. Allen, P. Knipscheer, L. Romeijn, A. Hart, E. Vink, A. Berns, and L. Romeyn, *High-throughput retroviral tagging to identify components of specific signaling pathways in cancer.* Nat Genet **32**, 153 (2002).

[85] A. H. Lund, G. Turner, A. Trubetskoy, E. Verhoeven, E. Wientjens, D. Hulsman, R. Russell, R. A. DePinho, J. Lenz, and M. van Lohuizen, *Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice.* Nat Genet **32**, 160 (2002).

[86] A. G. Uren, J. Kool, A. Berns, and M. van Lohuizen, *Retroviral insertional mutagenesis: past, present and future.* Oncogene **24**, 7656 (2005).

[87] L. S. Collier, C. M. Carlson, S. Ravimohan, A. J. Dupuy, and D. A. Largaespada, *Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse,* Nature **436**, 272 (2005).

[88] E. S. Robertson, *Cancer Associated Viruses,* Current Cancer Research (Springer, 2011).

[89] V. Theodorou, M. A. Kimm, M. Boer, L. Wessels, W. Theelen, J. Jonkers, and J. Hilkens, *MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer,* Nature Genetics **39**, 759 (2007).

[90] S. Bilodeau and R. Young, *ChIP-seq for H3K79me2 (unpublished),* (2010), GEO accession: GSE26680.

[91] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren, *A map of the cis-regulatory sequences in the mouse genome.* Nature **488**, 116 (2012).

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

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

[94] W. Wang, J. Yang, H. Liu, D. Lu, X. Chen, Z. Zenonos, L. S. Campos, R. Rad, G. Guo, S. Zhang, A. Bradley, and P. Liu, *Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog 1,* P Natl Acad Sci USA (2011).

[95] J. Cadinanos and A. Bradley, *Generation of an inducible and optimized piggybac transposon system.* Nucleic Acids Res **35**, e87 (2007).

[96] L. Mates, M. K. L. Chuah, E. Belay, B. Jerchow, N. Manoj, A. Acosta-Sanchez, D. P. Grzela, A. Schmitt, K. Becker, J. Matrai, L. Ma, E. Samara-Kuko, C. Gysemans, D. Pryputniewicz, C. Miskey, B. Fletcher, T. VandenDriessche, Z. Ivics, and Z. Izsvak, *Molecular evolution of a novel hyperactive sleeping beauty transposase enables robust stable gene transfer in vertebrates,* Nat Genet **41**, 753 (2009).

[97] B. Langmead and S. L. Salzberg, *Fast gapped-read alignment with bowtie 2,* Nat Meth **9**, 357 (2012).

[98] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,* Genome Biol **10**, R25 (2009).

[99] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva, *NCBI GEO: archive for functional genomics data sets–10 years on,* Nucleic Acids Res **39**, D1005 (2010).

[100] B. Pedersen, T.-F. Hsieh, C. Ibarra, and R. L. Fischer, *MethylCoder: software pipeline for bisulfite-treated sequences,* Bioinformatics **27**, 2435 (2011).

[101] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, *Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation,* Nat Biotechnol **28**, 511 (2010).

3.5. SUPPLEMENTARY MATERIAL

http://www.ncbi.nlm.nih.gov/pubmed/24721906

4

Chromatin position effects assayed by thousands of reporters integrated in parallel

Waseem Akhtar*
Johann de Jong*
Alexey V. Pindyurin*
Ludo Pagie
Wouter Meuleman
Jeroen de Ridder
Anton Berns
Lodewyk F.A. Wessels
Maarten van Lohuizen
Bas van Steensel

Parts of this chapter have been published in Cell [1].
* These authors contributed equally to this work.

Abstract

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 27000 distinct reporter integrations in mouse embryonic stem cells, obtained with two different promoters, show ~1000-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 cross-talk 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.

4.1. 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 [2, 3]. 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" [4]. 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 [5, 6], and IRs have also been used widely as "enhancer traps" to identify regulatory elements that promote transcription [7–9].

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 [9–13], 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 [14, 15] 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 >27000 locations (in total) throughout the genome of mouse embryonic stem (mES) cells. Because of the flexible design of the reporter vector, TRIP is a generally applicable technique to study many facets of gene regulation.

4.2. RESULTS

4.2.1. 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.

4.2.2. 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 [16] 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 [17]. For integration of barcoded reporters, we used the piggyBac (PB) transposition system because of its high efficiency [18] and the relatively small sizes of the essential terminal repeats (TRs) [19].

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 1000 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.

4.2.3. 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

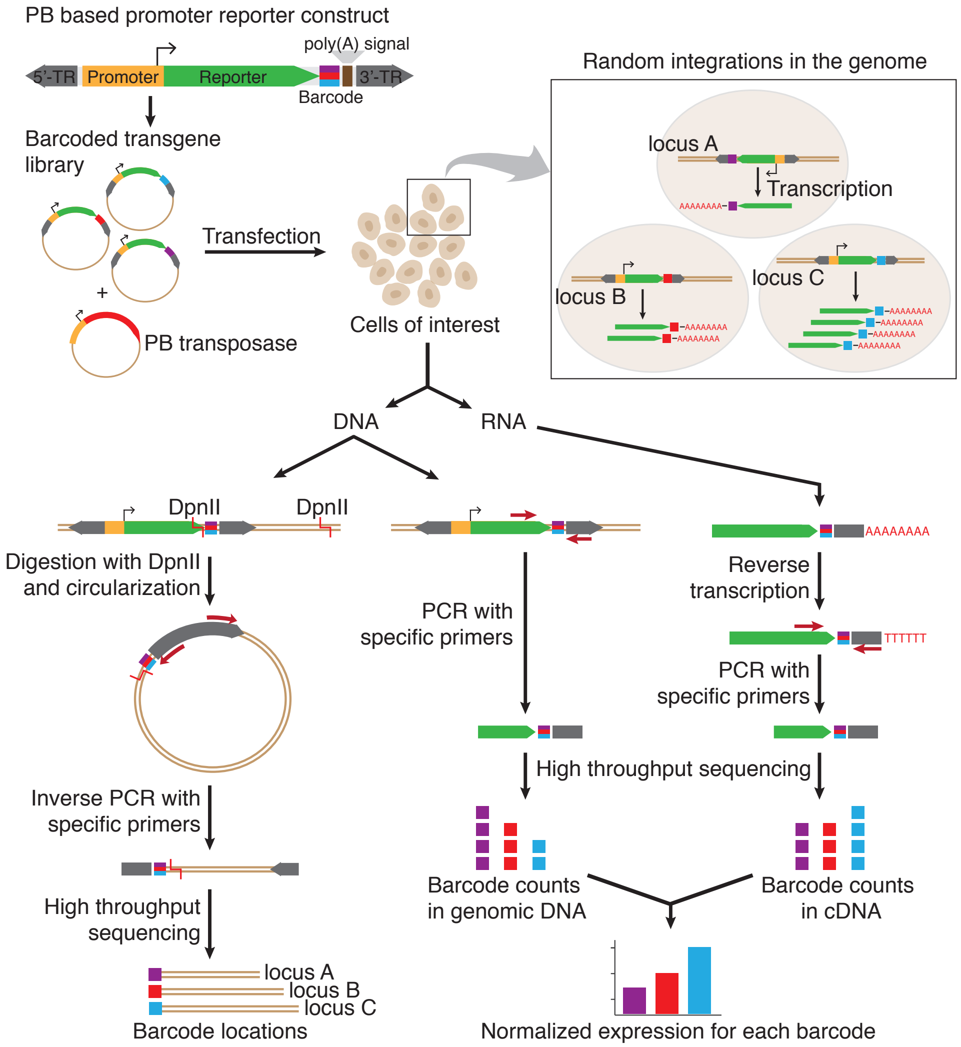

PB based promoter reporter construct
poly(A) signal
5'-TR
Promoter
Reporter
3'-TR
Barcode
Barcoded transgene library
+
PB transposase
Transfection
Cells of interest
Random integrations in the genome
locus A
Transcription
AAAAAAAA
locus B
AAAAAAAA
locus C
AAAAAAAA
DNA
RNA
DpnII
DpnII
Digestion with DpnII and circularization
Inverse PCR with specific primers
High throughput sequencing
locus A
locus B
locus C
Barcode locations
PCR with specific primers
AAAAAAAA
Reverse transcription
TTTTTT
PCR with specific primers
High throughput sequencing
Barcode counts in genomic DNA
Barcode counts in cDNA
Normalized expression for each barcode

Figure 4.1: **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.

replicates were located at the same base position in the genome (Table S1). After merging of the technical replicates and application of stringent data quality filters, each cell pool yielded roughly 2300-3300 mapped IRs (Figures S1B and S2A; Table S1). In total, we unequivocally mapped the locations of 17857 and 10903 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).

4.2.4. 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 ~1000-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 [20], 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.

4.2.5. 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 [21], 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.

4.2.6. 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 [22–24]. These three domain types are known to coincide substantially with one another, and harbor mostly inactive endogenous genes [3]. 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) [25, 26] 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.

4.2.7. 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 [23, 27]. The IRs in LADs show

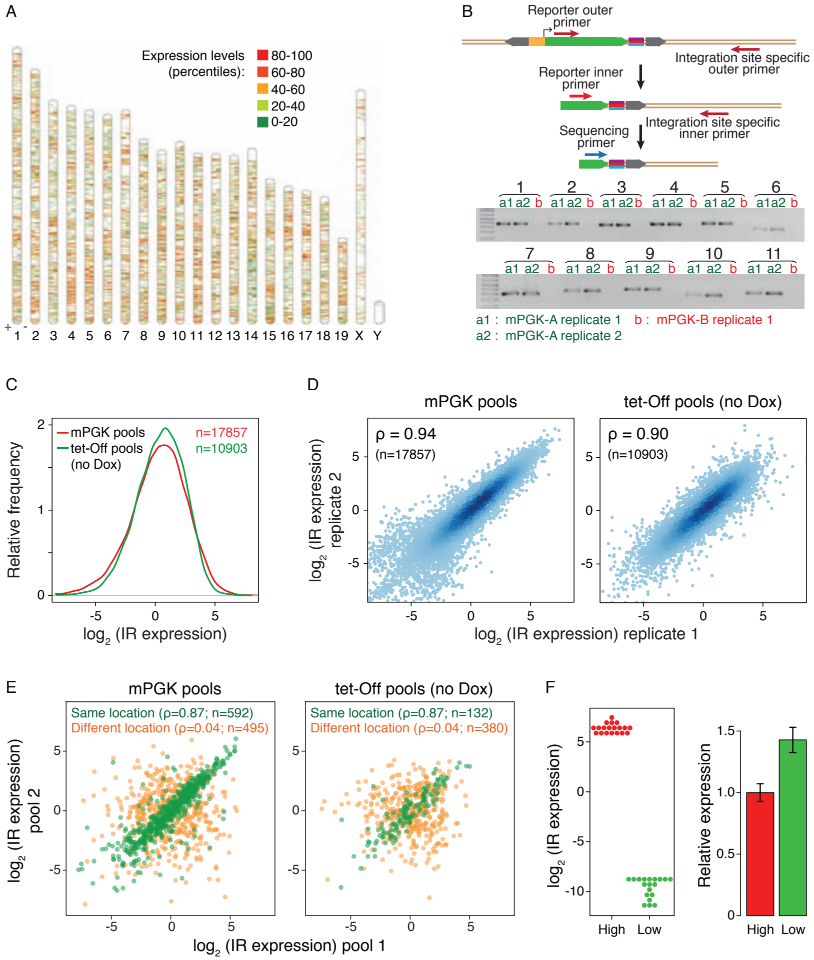


Figure 4.2: **TRIP works robustly and reproducibly.** (A) Positions of mapped mPGK IRs along all chromosomes. Each IR is represented as a tick on one of the strands (depending on the orientation of integration), colored by expression level. The mapped IR density on X and Y is lower because these chromosomes 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. (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.

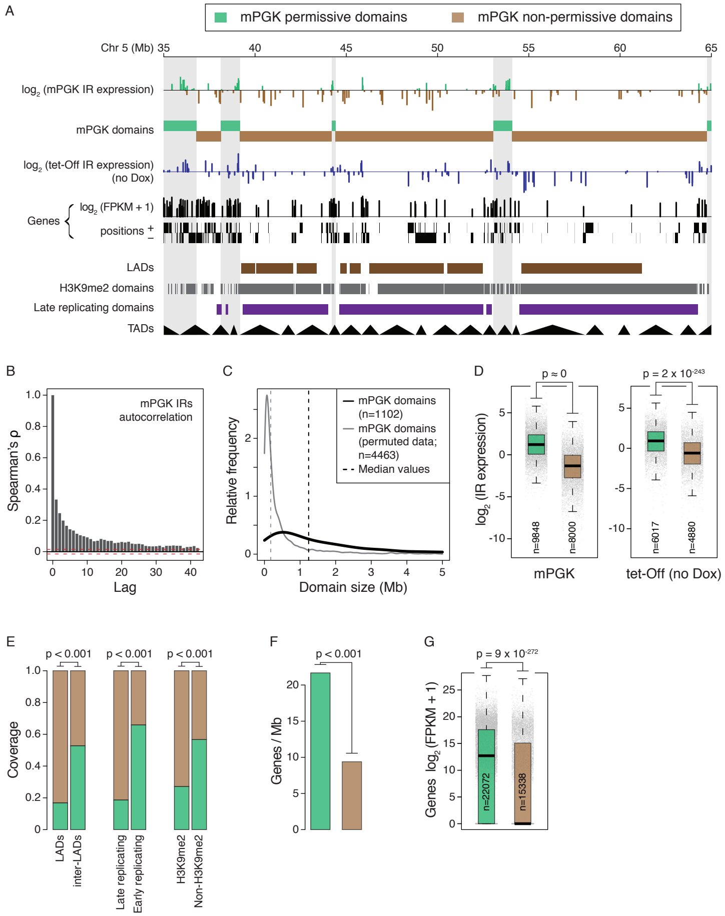


Figure 4.3: **Nonrandom IR expression patterns reflect chromatin domain organization.** (A) Segment of chromosome 5 showing expression levels for mPGK and tet-Off (no Dox) IRs, together with tracks showing a two-state HMM of mPGK IR activity (mPGK domains); expression and positions of endogenous genes; and the positions of various types of known chromatin domains [22–25]. (B) Autocorrelation function showing the similarity (Spearman's ρ) between expression levels of neighboring IRs (lag = nth neighbor, with $0 \le n \le 40$). 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 = 1000) 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.

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.

4.2.8. 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 [28–31] 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.

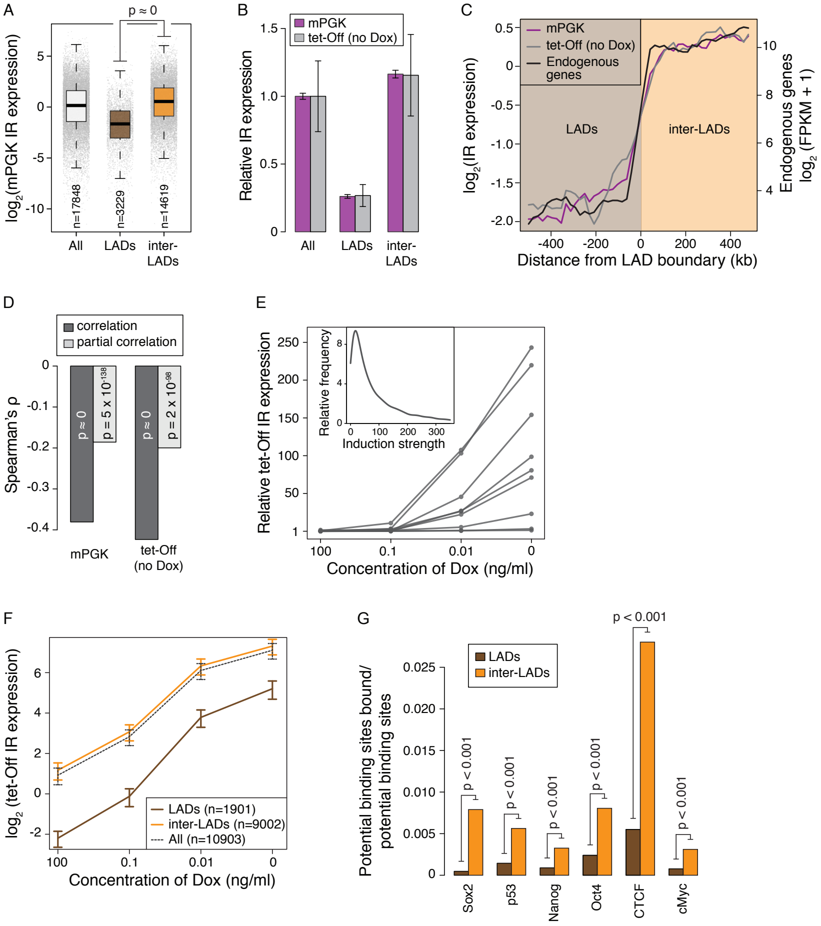


Figure 4.4: **LADs act as transcription attenuators.** (A) Expression level distributions for all mPGK IRs and those in LADs and inter-LADs, plotted as in Figure 3D. The p value was determined by Wilcoxon rank sum test. (B) Biological reproducibility of relative expression of IRs, separated by LAD or 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 $p = 2.8 \times 10^2$ 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 Materials and Methods) 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). (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 [28–31]. The p values were determined by circular permutation (n = 1000) of LADs, testing the fold difference of the inter-LAD fraction and the LAD fraction. See also Figure S4.

4.2.9. IR expression is related to local chromatin conformation

A popular model is that gene activity is controlled by the degree of chromatin compaction [32]. 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 function can be inferred from Hi-C data and approximated by a power law with a scaling exponent α [33, 34]. 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 [25], 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 ρ = -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.

4.2.10. Proximity effects of active genes and enhancers

Although LADs and chromatin compaction explain part of the variation in IR expression, much of the 1000-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 [35, 36]. 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" [37] and thereby promote recruit-

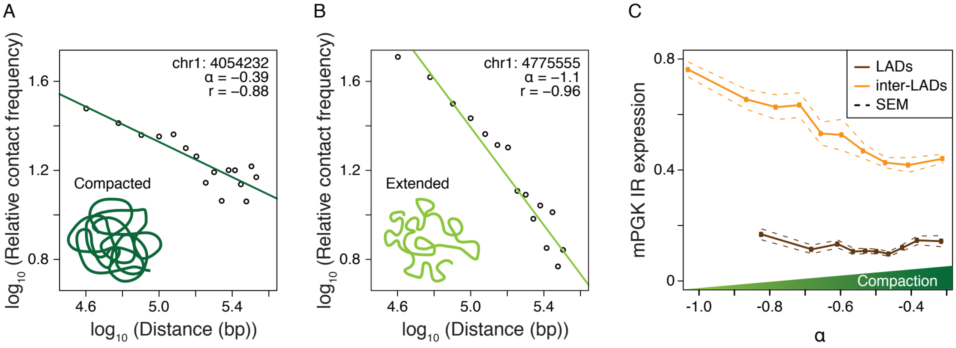


Figure 4.5: **Local chromatin conformation partially predicts IR expression levels.** (A and B) Examples of the dependency of relative contact frequency (as determined by Hi-C; Dixon et al., 2012) on genomic distance in 400 kb windows around two mPGK IRs. Note the difference in the slope (α) of the fitted 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.

ment 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 [38]. To test whether these enhancers might stimulate expression of nearby 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.

4.2.11. 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 [30, 39–42] 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).

4.3. Discussion

4.3.1. 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 and 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.

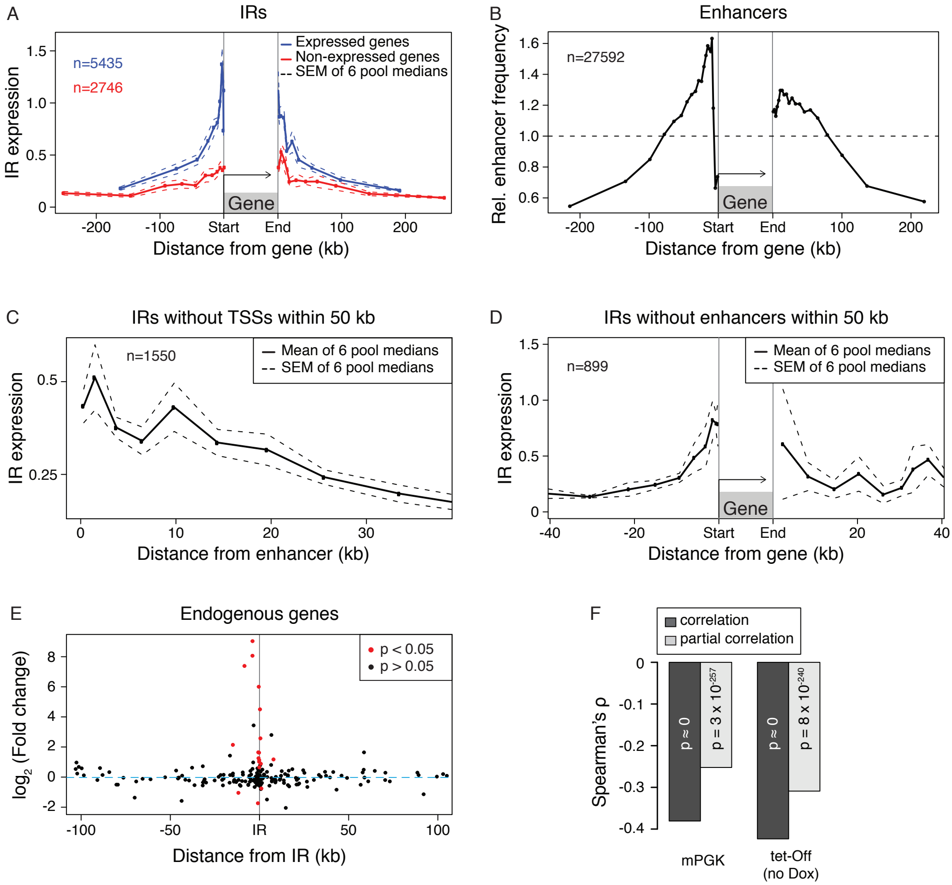


Figure 4.6: **Proximity effects of genes and enhancers.** (A) Intergenic mPGK IR expression as a function of their distance from the nearest endogenous gene. The endogenous genes are divided into two categories: expressed (blue) and not detectably expressed (red). The solid lines show the mean of median 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 Materials and Methods) 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.

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).

4.3.2. 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.

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 [43–45] 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.

4.3.3. 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 [23, 27]. 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 inte-

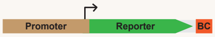
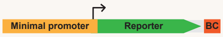
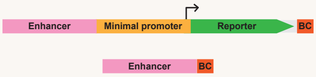
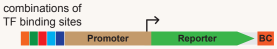
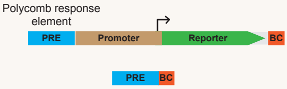
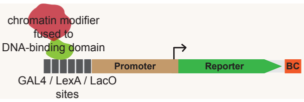
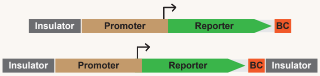
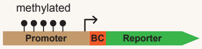
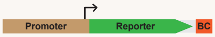
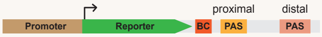
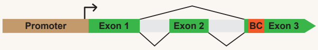
Process	Design of TRIP construct	Assays	Expected insights
Transcription		barcode-RNA-seq, ChIP/DamID	Effects of chromatin context on the activity of a promoter
		barcode-RNA-seq, ChIP/DamID	Enhancer activities in different cell types
	 Enhancer Minimal promoter Reporter BC Enhancer BC	barcode-RNA-seq, ChIP/DamID	Effects of chromatin context on enhancer activity
	 combinations of TF binding sites Promoter Reporter BC	barcode-RNA-seq, ChIP/DamID	Interaction between different TFs in varying epigenetic enviroments
Chromatin dynamics	 Polycomb response element PRE Promoter Reporter BC PRE BC	barcode-RNA-seq, ChIP/DamID	Dynamics of establishment and maintenance of polycomb domains in different epigenomic contexts
	 chromatin modifier fused to DNA-binding domain Promoter Reporter BC GAL4 / LexA / LacO sites	barcode-RNA-seq, ChIP/DamID	Behavior of a chromatin modifier in a variety of chromatin states
	 Insulator Promoter Reporter BC Insulator Promoter Reporter BC Insulator	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/ polyadenlylation	 Promoter Reporter BC proximal PAS distal PAS	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 4.7: **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.

gration. 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 [46]. 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 [47, 48]. Previous experimental studies noted a transcription "ripple effect" between neighboring genes [35] and activation of IRs nearby active gene clusters [11], 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 (~1000-fold) in IR expression levels. Further computational modeling of the data may uncover additional features that determine gene expression.

4.4. Materials and Methods

4.4.1. Construction of barcoded piggyBac plasmid libraries

The pPTK-Gal4-mPGK-Puro-IRES-eGFP-sNRP-pA ("mPGK") and pPTK-Gal4-tet-Off-Puro-IRES-eGFP-sNRP-pA ("tet-Off") piggyBac (PB) reporter constructs (GenBank accession numbers KC710227 and KC710228) contain the following elements, some of which were not used in this study. As a transcription unit, we used a bicistronic gene consisting of the Puromycin resistance cassette (PuroR) and enhanced green fluorescent protein (eGFP), linked by an encephalomyocarditis virus internal ribosome entry site (IRES). The PuroR cassette was not used in this study. The transcription unit is driven by either the mouse phosphoglycerate kinase (mPGK) promoter or tetracycline-dependent (tet-Off) promoter and ends with a short polyA signal from the soluble neuropilin-1 (sNRP-1) gene [49]. We also placed 14 repeats of the Gal4 binding site upstream of the promoters; this element was not used in this study but may serve to target Gal4 fusion proteins. All elements were cloned between the terminal repeats (TRs) of the PB element in the 5'-PTK-3' plasmid [18].

To generate the barcoded plasmid libraries, the 3'-TR of the PB was amplified with primers PB-barcode-long-7[1], which contains a 16 base long random barcode sequence with the preceding DpnII site) and PB-barcode-short-7[2], using the pPTK-Gal4-tet-Off-Puro-IRES-eGFP-sNRP-pA-trim1 plasmid (a modified version of "tet-Off" with 3'-TR of PB shortened to only the last 67 bp; GenBank accession number KC710229) as a template. The PCR product was digested with PstI and MluI (the underlined sequences in the primers), ligated into the same sites into "mPGK" and "tet-Off" vectors and transformed into E. coli cells. For each construct, ~250000 transformed bacterial colonies were grown on plates and pooled before plasmid DNA purification. This resulted in the preparation of the "mPGK" and "tet-Off" PB plasmid libraries with random barcodes (GenBank accession numbers KC710230 and KC710231). The barcode cloning step reduced the length of the 3'-TR from 242 bp down to 67 bp, which is sufficient for PB transposition (Meir et al., 2011) and lacks DpnII sites.

After completion of the TRIP experiments we discovered that ~16% of the plasmid molecules in all libraries carry a point deletion (at positions 4002 and 3914 in GenBank sequences KC710230 and KC710231, respectively) in the 3'TR. Transient transfection assays indicate that this mutation causes a 2-fold increase in reporter expression. Because the barcodes are randomly linked to either the unmutated or the mutated 3'TR, this causes some noise (~2-fold) in the estimates of expression levels. Future builds of the libraries without this mutation should therefore allow for even more accurate measurements.

4.4.2. Mouse embryonic stem cell culture and transfection

mES cells EBRTcH3 expressing the tetracycline-controlled transactivator from the endogenous ROSA26 promoter [50] were cultured in 60% BRL cell-conditioned medium in the presence of leukemia inhibitory factor, MEK inhibitor PD0325901, and GSK-3 inhibitor CHIR99021 [51]. 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 [18] 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

[1] 5'-GTGACACCTGCAGGATCA(N)16CTCGAGTTGTGGCCGGCCCTTGTGACTG-3'

[2] 5'-GACATAACGCGTATACTAGATTAACCCT-3'

measurements (see below). After sorting, the medium-eGFP population was grown for 5 days before several aliquots of ~1000 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 ~7000-11000 barcodes (Table S1) of the expected ~23000 (1000 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).

4.4.3. 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.

4.4.4. 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.

4.4.5. Estimation of average numbers of IRs per cell in TRIP pools

To estimate the average copy number of IRs in TRIP pools, gDNA from each pool was subjected to qPCR using primers against eGFP (eGFP-qPCR-for[3] and eGFP-qPCR-rev[4]), murine lamin B receptor gene (mLBR-2-for[5] and mLBR-2-rev[6]) and ampicillin (qAmp-3[7] and qAmp-4[8]). A plasmid harboring one copy each of the three amplicons (construction details are available upon request) was used as a reference for quantification.

4.4.6. Analysis of barcode high-throughput sequencing data

Sequencing was done on an Illumina HiSeq 2000 instrument and resulted in two types of sequencing reads, (1) the "expression" and "normalization" reads (single reads of 100 bp) corresponding to the quantification of barcodes in cDNA and gDNA samples, respectively, and (2) the "mapping" reads (paired-end reads of 100 bp) from inverse PCR fragments. The latter reads contain the barcode and the genomic sequence next to DpnII site in the forward read, and the end of the PB transposon and the neighboring genomic DNA sequence in the reverse read. The scheme of data analysis pipeline is shown in Figure S1B.

Because the identity of integrated barcodes was not known a priori and some barcodes might be absent in the expression and mapping reads (due to the full repression or an inappropriate location of the DpnII site in flanking genomic DNA region, respectively), we focused the initial analysis on the normalization reads. The barcode sequences (15-17 base long) were extracted from sequencing reads using R (http://www.R-project.org) and Bioconductor [52] packages. For each replicate of the normalization reads, only the barcodes with at least 5 counts were taken for the subsequent analysis. To eliminate aberrant barcodes arising from mutations during PCR and sequencing, we used the following algorithm. First, we sorted barcodes according to their counts. Then, for each barcode (starting from the most frequent one), we identified and removed all its mutant versions, defined as barcodes within a Hamming distance of 2. The remaining sequences were regarded as "genuine" barcodes and demonstrated a large overlap between the replicates (Figure S2A). The barcodes identified only in one technical replicate were mostly with low counts and frequently were lost in the other replicate due to the threshold of 5 counts (data not shown).

The set of genuine barcodes found in both replicates was used as a reference to identify and count the barcodes in the expression and normalization reads independently for each replicate. For the mPGK integrations, a pseudocount of 1 was added to all expression and normalization counts to account for zeros in the expression data (fully repressed IRs), and expression and normalization counts for all replicates were normalized to a total read count of 1. Subsequently, expression values were normalized by dividing by corresponding normalization values, and replicates were averaged to obtain the expression for each barcode. For the tet-Off data set, a similar approach was taken for each

[3] 5'-CGACAACCACTACCTGAGCA-3'

[4] 5'-GAACTCCAGCAGGACCATGT-3'

[5] 5'-CTCCACTTCCCTCCACCTCT-3'

[6] 5'-GAGAGCTTGAACGAAAAACCA-3'

[7] 5'-CGATCGTTGTCAGAAGTAAGTTGG-3'

[8] 5'-CACAGAAAAGCATCTTACGGATGG-3'

of the induction levels. Additionally, the resulting expression values were multiplied by the relative level of eGFP mRNA for each induction level as determined by RT-qPCR (Figure S4), to make the different levels of induction comparable. For qPCR, we used 1.0 μl of the same cDNA that was used for amplifying the barcodes for high-throughput sequencing. The primers used were eGFP-qPCR-for and eGFP-qPCR-rev (see above). Signal in each sample was normalized to TBP levels for which following primers were used: TBP-forward[9] and TBP-reverse[10].

The genomic regions associated with genuine barcodes were extracted from the mapping reads and aligned against mouse genome assembly mm9 using Bowtie2 [53] independently for each replicate. It should be noted that during the inverse PCR step, there is a slight chance of intermolecular ligation, resulting in a small fraction of fragments where the barcode is associated with an unexpected genomic region. Additionally, one barcode might be present at two distinct genomic locations either because of random chance or as a result of re-transposition of the reporter in one of the daughter cells during early days of culturing when PB transposase is still present. Therefore we applied a stringent set of criteria to select only those barcodes which are unambiguously associated with unique genomic locations. Specifically, only those barcodes were retained that had at least 3 reads in the mapping data and more than 90% of the reads were associated with the most frequent genomic location and less than 2.5% were associated with the second most frequent location.

During the mapping process, a number of genuine barcodes was filtered out at different steps (Table S1). About 28% were lost, as they did not have any reads in the mapping data, primarily for two reasons. First, sequencing in mapping data is not saturating. This is because preparation of inverse PCR samples is complex and the chance to lose a rare barcode is very high. This is also evident from the fact that there is a direct correlation between the abundance of a barcode (counts in normalization data) and the reads of that barcode in the mapping data (Spearman's ρ = 0.631 for mPGK pools and Spearman's ρ = 0.415 for tet-Off pools; data not shown). Second, some barcodes are integrated at places where the DpnII site is not at an optimal distance from the site of integration. A little over 2% were associated with DNA that could not be aligned at all to the genome. About 30% could not pass the stringent criteria (described above) for being associated to a single genomic location. About 15% of barcodes were integrated in the repetitive parts of the genome. A small fraction of barcodes was redundant being present in more than one pool (same location in each pool) and thus only one of the entries was retained for the analysis. Despite these losses about 35% of the genuine barcodes could be mapped unequivocally. Ideograms of mouse chromosomes with the positions of mapped IRs and HMM domains were generated with the help of idiographica (http://www.ncrna.org/idiographica) [54].

4.4.7. BARCODE EXPRESSION LEVELS AFTER TRANSIENT TRANSFECTION

To assess the effect of barcode sequences on the IR expression, two sets of 19 barcodes each were randomly selected from 30 least and 30 most expressed IRs in mPGK TRIP pools and individually cloned into the pPTK-Gal4-mPGK-Puro-IRES-eGFP-sNRP-pA

[9] 5'-CTGGAATTGTACCGCAGCTT-3'

[10] 5'-TCCTGTGCACACCATTTTTC-3'

vector at the PstI and MluI sites as described above. Next, equal amounts of resulting constructs were pooled to make small libraries (19 barcodes each) of the "low" and "high" expressed barcodes. 4 μg of each 19-plasmid library was transiently transfected (in triplicate; without the mPB transposase plasmid) into 1.5 × 106 EBRTcH3 cells using Lipofectamine 2000 (Invitrogen). RNA was isolated 24 hr posttransfection and analyzed by RT-qPCR using eGFP and TBP primers as described above.

4.4.8. MPGK DOMAIN CALLING

To show domain-oriented behavior, the Spearman autocorrelation function was computed for different lags, using per-pool mean-centered $\log_2$(expression) values. The confidence bounds were determined by the following formula:

$$\pm\frac{f_z(1-\frac{\alpha}{2})}{\sqrt{N}} \tag{4.1}$$

where α is the significance level (5%), f_z is the standard normal quantile function and N is the number of IRs. For each chromosome individually, a Hidden Markov Model (HMM) was inferred on the sequence of mean-centered expression values using the Baum-Welch algorithm assuming Gaussian distribution of the emission, and states were called using the Viterbi algorithm as implemented in the R package RHmm (http://www.r-project.org/), assuming equidistant spacing between the integrations. Domain boundaries were placed halfway between the flanking IRs. The probability density of the resulting domain sizes was estimated by convolving the domain sizes with a Gaussian kernel, while automatically selecting the standard deviation of this smoothing kernel with Silverman's "Rule of Thumb" [55]. The nonhomogeneous HMM was inferred using the BioHMM algorithm [56], as implemented in the Bioconductor package snapCGH [52]. Several domain-oriented epigenetic features were tested for their fractional overlap with these mPGK domains. LADs were taken from [23]. For late replication timing [22] and H3K9me2 [24], domains were called using the RHmm package, assuming equidistant spacing between the probes. For computing significance of overlap, the mPGK permissive domains were circularly permuted genome-wide, starting with a shift the size of the largest mPGK permissive domain for the first permutation, and equally distributing the remaining permutations across the remainder of the genome. The test-statistic used was the fold difference of the mPGK nonpermissive domain fractions.

4.4.9. PREPROCESSING OF PREVIOUSLY PUBLISHED CHIP-SEQ DATA

To maximize the comparability of the ChIP-seq data sets, they were processed from the sequence read archives as obtained from GEO [57]. Sequence read archives were converted to FASTA format and then aligned against mouse genome assembly mm9 using Bowtie [53], allowing at most two mismatches in end-to-end alignment (the following settings were used: "-M 1 –best –tryhard -v 2 –chunkmbs 1024").

Additionally, duplicate reads were removed. Peaks were called on the aligned reads with MACS [58] using default settings, and corresponding GFP, whole-cell extract, pan-H3 or input DNA data sets as background. Additional filtering of peaks was performed by requiring a minimum fold enrichment of 3 with respect to the background.

For the HMM chromatin state classification (Figure S6), we used the following ChIP-seq data sets: H3K4me1, H3K27ac [40], H3K4me2 [42], H3K4me3, H3K9me3, H3K27me3, HeK36me3, H4K20me3 [39], H3K9ac [41], CTCF [30], and input DNA pooled from [30, 31, 59–62]. The multivariate HMM was trained using the ChromHMM software [63]. First, aligned sequences, as described above, were binarized using the BinarizeBed function at a binsize of 200 bp. Second, the HMM was trained and states were called using the LearnModel function. The number of 15 states was chosen to correspond with [64]. Third, all integrations were mapped to the resulting states.

4.4.10. Analysis of LADs

To visualize the behavior of transgenes around LAD boundaries, all LAD boundaries were aligned. mPGK, tet-Off IRs, and endogenous transcriptional start sites (TSSs) were located around each LAD boundary, within a window of 500 kb, but no further than halfway toward the next LAD boundary. Average expression values were computed in 20 kb bins. Per-pool mean-centered $\log_2$(expression) values of the mPGK and tet-Off IRs and the $\log_2$(FPKM+1) values of the endogenous genes were used. Here, FPKM refers to the number of fragments per kilobase of exon per million fragments mapped, as determined by Cufflinks (Trapnell et al., 2010).

For the partial correlation analysis, the ordinary Spearman correlation was computed between mPGK and tet-Off expression on one hand, and the Lamin B1 signal (the probe values) [23] on the other hand. These correlation coefficients were then compared to the partial Spearman correlation coefficients of mPGK and tet-Off expression, with the Lamin B1 signal, given H3K9me2, replication timing (for both variables using the probe values), and TSS proximity. TSS proximity was defined as the negative distance from an IR to the nearest TSS.

To get an impression of the variation in induction strengths, a linear model was fitted to the non-$\log_2$-transformed expression values across the four different induction levels, for each individual tet-Off IR. For the independent variable, the Dox concentrations for the different induction levels were taken. The highest concentration of 100 ng/ml was approximated by 1 ng/ml, since the results for 1 ng/ml and 100 ng/ml are highly similar (data not shown). The absence of Dox was approximated by 0.001 ng/ml. The probability density of the slope coefficients (induction strengths) was estimated by a Gaussian kernel convolution, and Silverman's "Rule of Thumb" for automatically selecting the bandwidth of the kernel.

Binding of transcription factors within LADs and inter-LADs was assessed by comparing position weight matrix (PWM) scores to transcription factor (TF) binding sites determined by ChIP-seq. For the six TFs, PWMs were downloaded from TRANSFAC [65]. The mouse genome assembly mm9 was scanned and a potential binding site was called for a locus if, centered on that locus, the PWM score was at least 80% of information content of motif. Actual TF binding sites were determined by peak-calling of ChIP-seq data as outlined above. Subsequently, the fraction of PWM hits covered by a ChIP-seq peak was computed for each TF, for LADs as well as inter-LADs. For computing significance of overlap, LADs were circularly permuted genome-wide, starting with a shift the size of the largest LAD for the first permutation, and equally distributing the remaining permutations across the remainder of the genome. The test-statistic used was the fold difference of the inter-LAD fraction and the LAD fraction.

4.4.11. CHROMATIN INTERACTIONS

To compare mPGK domains and topologically associated domains (TADs) [25], the overlap of boundary areas between the data sets was assessed. mPGK boundary areas were defined as the areas between two adjacent IRs in a different HMM state. TAD boundaries were defined as in [25] with a 40 kb margin added to each side to account for resolution uncertainty.

For each mPGK IR, the local chromatin contact probability as a function of genomic distance was fitted to a power-law curve. For this we used normalized Hi-C contact frequency matrices (20 kb bins) from mES cells downloaded from the RenLab website [25]. Average contact frequencies were calculated as a function of genomic distance within a window of 400 kb centered around each IR. Genomic distance and contact frequencies were then $\log_1 0$-transformed and α value was calculated as the slope of a linear regression fit. To visualize the relationship between α value and IR expression, IRs were divided into LAD and inter-LAD groups, and for each the α values were binned into 10 quantiles. For each mPGK pool and bin, the median expression of the IRs was then computed. This resulted in six medians per bin, which were then averaged, and the standard error of the mean of the six pool medians was computed.

4.4.12. ANALYSIS OF GENE PROXIMITY AND ENHANCER PROXIMITY

The association between gene proximity and mPGK IR expression was analyzed by matching each IR with its nearest endogenous gene start or gene end (depending on which one was closest). Here, endogenous genes were separated according to their expression level (FPKM = 0, and FPKM > 0). Distances between IRs and nearest gene start or gene end were binned into 10 quantiles, and expression values were binned accordingly. Per-pool median expression values were computed, and their mean and standard error of the mean were determined. Similar analyses were performed for gene starts and gene ends excluding IRs with an enhancer within 50 kb, as well as for enhancers excluding IRs with a gene start within 50 kb. Enhancers were defined as regions with both an H3K4me1 peak and an H3K27ac peak within 500 bp, or alternatively a p300 peak. Regions that were close to (<5 kb) an H3K4me3 peak were removed to avoid overlap with transcription start sites, and remaining regions were called enhancers (median size ~1 kb).

Enhancer density around genes was analyzed by determining for each enhancer its nearest gene start or gene end (depending on which one was closest). The distance between a gene start or gene end and the midpoint of the enhancer, and distances were binned and counted. The same was done for a set of 10^6 randomly selected genomic coordinates, and within bins a ratio of observed versus expected was calculated.

4.4.13. EXPRESSION OF ENDOGENOUS GENES NEAR IRS

mES cells were transfected with barcoded PB mPGK plasmid library and mPB transposase plasmid [18]. The transfected cells with very high levels of eGFP were selected by FACS sorting. The sorted cells were grown for 5 additional days before they were plated sparsely on 10-cm dishes. After two weeks, colonies were picked and further propagated to generate 11 clones with multiple IRs. RNA-seq was performed for each of these cell lines, using an Illumina HiSeq 2000 instrument and standard protocol. To assess the change in the expression of genes near an IR in one sample, the other 10 samples were used as reference. The change in gene expression was thus defined as the $\log_2$ ratio of the raw read count for a gene and the mean read count for the same gene in the 10 reference samples. p-values of differential expression were calculated using the R package DESeq [66], and were corrected for multiple testing after pooling the p-values of all 11 comparisons (Holms' method). The distance between a gene and an IR was defined as the distance between the closest of the gene's transcription start site or transcription termination site, and the position of the IR. IRs located within genes were discarded.

4.4.14. 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, accession number GSE48606.

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.).

REFERENCES

[1] W. Akhtar, J. de Jong, A. V. Pindyurin, L. Pagie, W. Meuleman, J. de Ridder, A. Berns, L. F. A. Wessels, M. van Lohuizen, and B. van Steensel, *Chromatin position effects assayed by thousands of reporters integrated in parallel.* Cell **154**, 914 (2013).

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

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

[4] T. Dobzhansky, *Position effects on genes.* Biol Rev **11**, 364 (1936).

[5] S. I. Grewal and S. Jia, *Heterochromatin revisited.* Nature reviews. Genetics **8**, 35 (2007).

[6] J. R. Girton and K. M. Johansen, *Chapter 1 chromatin structure and the regulation of gene expression: The lessons of PEV in drosophila,* (Elsevier, 2008) pp. 1–43.

[7] F. Weber, J. de Villiers, and W. Schaffner, *An sv40 enhancer trap incorporates exogenous enhancers or generates enhancers from its own sequences.* Cell **36**, 983 (1984).

[8] V. Korzh, *Transposons as tools for enhancer trap screens in vertebrates,* Genome Biology **8**, S8+ (2007).

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

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

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

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

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

[14] C. Gerlach, J. W. van Heijst, E. Swart, D. Sie, N. Armstrong, R. M. Kerkhoven, D. Zehn, M. J. Bevan, K. Schepers, and T. N. Schumacher, *One naive t cell, multiple fates in cd8+ t cell differentiation.* J Exp Med **207**, 1235 (2010).

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

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

[17] M. Gossen, S. Freundlieb, G. Bender, G. Muller, W. Hillen, and H. Bujard, *Transcriptional activation by tetracyclines in mammalian cells.* Science **268**, 1766 (1995).

[18] J. Cadinanos and A. Bradley, *Generation of an inducible and optimized piggybac transposon system.* Nucleic Acids Res **35**, e87 (2007).

[19] Y. J. Meir, M. T. Weirauch, H. S. Yang, P. C. Chung, R. K. Yu, and S. C. Wu, *Genome-wide target profiling of piggybac and tol2 in hek 293: pros and cons for gene discovery and gene therapy.* BMC Biotechnol **11**, 28 (2011).

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

[21] X. Huang, H. Guo, S. Tammana, Y.-C. Jung, E. Mellgren, P. Bassi, Q. Cao, Z. J. Tu, Y. C. Kim, S. C. Ekker, X. Wu, S. M. Wang, and X. Zhou, *Gene transfer efficiency and genome-wide integration profiling of sleeping beauty, tol2, and piggybac transposons in human primary t cells.* Mol Ther **18**, 1803 (2010).

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

[23] D. Peric-Hupkes, W. Meuleman, L. Pagie, S. W. M. Bruggeman, I. Solovei, W. Brugman, S. Gräf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, M. Reinders, L. Wessels, and B. van Steensel, *Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.* Mol Cell **38**, 603 (2010).

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

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

[26] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, and E. Heard, *Spatial partitioning of the regulatory landscape of the x-inactivation centre.* Nature **485**, 381 (2012).

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

[28] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y.-H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W.-K. Sung, N. D. Clarke, C.-L. Wei, and H.-H. Ng, *Integration of external signaling pathways with the core transcriptional network in embryonic stem cells,* Cell **133**, 1106 (2008).

[29] A. Marson, S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love, N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch, and R. A. Young, *Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells.* Cell **134**, 521 (2008).

[30] L. Handoko, H. Xu, G. Li, C. Y. Y. Ngan, E. Chew, M. Schnapp, C. W. H. W. Lee, C. Ye, J. L. H. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W.-K. K. Sung, Y. Ruan, and

C.-L. L. Wei, *CTCF-mediated functional chromatin interactome in pluripotent cells.* Nat Genet **43**, 630 (2011).

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

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

[33] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, *Comprehensive mapping of long-range interactions reveals folding principles of the human genome,* Science **326**, 289 (2009).

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

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

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

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

[38] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren, *Histone modifications at human enhancers reflect global cell-type-specific gene expression,* Nature **459**, 108 (2009).

[39] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O/'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein, *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature **448**, 553 (2007).

[40] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch, *Histone H3K27ac separates active from poised enhancers and predicts developmental state,* P Natl Acad Sci USA **107**, 21931 (2010).

[41] H. Hezroni, B. S. Sailaja, and E. Meshorer, *Pluripotency-related, valproic acid (VPA)-induced genome-wide histone h3 lysine 9 (H3K9) acetylation patterns in embryonic stem cells,* J Biol Chem **286**, 35977 (2011).

[42] M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schubeler, *DNA-binding factors shape the mouse methylome at distal regulatory regions,* Nature **480**, 490 (2011).

[43] M. J. Vogel, D. Peric-Hupkes, and B. van Steensel, *Detection of in vivo protein-dna interactions using damid in mammalian cells.* Nat Protoc **2**, 1467 (2007).

[44] F. Mohn, M. Weber, D. Schubeler, and T. C. Roloff, *Methylated dna immunoprecipitation (medip).* Methods Mol Biol **507**, 55 (2009).

[45] T. S. Furey, *Chip-seq and beyond: new and improved methodologies to detect and characterize protein-dna interactions.* Nat Rev Genet **13**, 840 (2012).

[46] J. Kind and B. van Steensel, *Genome-nuclear lamina interactions and gene regulation.* Current opinion in cell biology **22**, 320 (2010).

[47] L. D. Hurst, C. Pal, and M. J. Lercher, *The evolutionary dynamics of eukaryotic gene order.* Nat Rev Genet **5**, 299 (2004).

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

[49] T. J. McFarland, Y. Zhang, L. O. Atchaneeyaskul, P. Francis, J. T. Stout, and B. Appukuttan, *Evaluation of a novel short polyadenylation signal as an alternative to the sv40 polyadenylation signal.* Plasmid **56**, 62 (2006).

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

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

[52] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang, *Bioconductor: open software development for computational biology and bioinformatics.* Genome biology **5**, 1 (2004).

[53] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,* Genome Biol **10**, R25 (2009).

[54] T. Kin and Y. Ono, *Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat.* Bioinformatics **23**, 2945 (2007).

[55] B. W. Silverman, *Density Estimation for Statistics and Data Analysis,* 1st ed., Chapman and Hall/CRC Monographs on Statistics and Applied Probability (Chapman and Hall/CRC, 1986).

[56] J. C. Marioni, N. P. Thorne, and S. Tavare, *Biohmm: a heterogeneous hidden markov model for segmenting array cgh data.* Bioinformatics **22**, 1144 (2006).

[57] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva, *NCBI GEO: archive for functional genomics data sets–10 years on,* Nucleic Acids Res **39**, D1005 (2010).

[58] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nusbaum, R. Myers, M. Brown, W. Li, and X. S. Liu, *Model-based analysis of ChIP-seq (MACS),* Genome Biol **9**, R137+ (2008).

[59] J. Han, P. Yuan, H. Yang, J. Zhang, B. S. Soh, P. Li, S. L. Lim, S. Cao, J. Tay, Y. L. Orlov, T. Lufkin, H.-H. Ng, W.-L. Tam, and B. Lim, *Tbx3 improves the germ-line competency of induced pluripotent stem cells,* Nature **463** (2010).

[60] M. J. Law, K. M. Lower, H. P. J. Voon, J. R. Hughes, D. Garrick, V. Viprakasit, M. Mitson, M. De Gobbi, M. Marra, A. Morris, A. Abbott, S. P. Wilder, S. Taylor, G. M. Santos, J. Cross, H. Ayyub, S. Jones, J. Ragoussis, D. Rhodes, I. Dunham, D. R. Higgs, and R. J. Gibbons, *ATR-x syndrome protein targets tandem repeats and influences Allele-Specific expression in a Size-Dependent manner,* Cell **143**, 367 (2010).

[61] E. R. Smith, C. Lin, A. S. Garrett, J. Thornton, N. Mohaghegh, D. Hu, J. Jackson, A. Saraf, S. K. Swanson, C. Seidel, L. Florens, M. P. Washburn, J. C. Eissenberg, and A. Shilatifard, *The little elongation complex regulates small nuclear RNA transcription,* Mol Cell **44**, 954 (2011).

[62] L. Tavares, E. Dimitrova, D. Oxley, J. Webster, R. Poot, J. Demmers, K. Bezstarosti, S. Taylor, H. Ura, H. Koide, A. Wutz, M. Vidal, S. Elderkin, and N. Brockdorff, *RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3.* Cell **148**, 664 (2012).

[63] J. Ernst and M. Kellis, *Chromhmm: automating chromatin-state discovery and characterization.* Nat Methods **9**, 215 (2012).

[64] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein, *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature **473**, 43 (2011).

[65] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender, *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic acids research **34**, D108 (2006).

[66] S. Anders and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol **11**, R106 (2010).

4.5. SUPPLEMENTARY MATERIAL

http://www.ncbi.nlm.nih.gov/pubmed/23953119

5

Genome-wide profiling of anti-cancer drug effects on chromatin guides rational treatment design

Baoxu Pang[*]
Johann de Jong[*]
Xiaohang Qiao[*]
Lodewyk F.A. Wessels
Jacques Neefjes

Parts of this chapter have been submitted for publication.

[*] These authors contributed equally to this work.

Abstract

Many anti-cancer drugs induce DNA breaks to eliminate tumor cells. The anthracycline topoisomerase II inhibitors additionally cause histone eviction. Here, we performed genome-wide high-resolution mapping of chemotherapeutic effects of various topoisomerase I and II inhibitors and integrated this mapping with a wide range of established (epi)genomic features. These anti-cancer drugs appeared to have a marked selectivity for defined areas in the genome. The topoisomerase I inhibitor topotecan and the topoisomerase II inhibitor etoposide induce DNA damage at the same transcriptionally active regions in the genome. The anthracycline daunorubicin induces DNA breaks and evicts histones from active chromatin, thus quenching local DNA damage responses. Another anthracycline, aclarubicin, has a different genomic specificity and evicts histones from H3K27me3-marked heterochromatin with consequences for diffuse large B-cell lymphoma cells with elevated levels of H3K27me3. The chemical profiling of the genome with different anti-cancer drugs reveals their unique genomic selectivity including previously un-annotated regions, and provides information critical for treatment decisions of specific tumors with altered epigenetics.

5.1. Introduction

The anthracycline topoisomerase II (TopoII) inhibitors daunorubicin (Daun) and doxorubicin (Doxo; also called hydroxydaunorubicin) have been cornerstones of oncology for many decades. Millions of cancer patients have been treated with these drugs with varying success [1]. Given the serious side effects of Daun and Doxo, alternative compounds have been developed. These include anthracycline variants such as aclarubicin (Acla; now only used in Japan and China for cancer treatment), and the chemically unrelated TopoII inhibitor etoposide (Etop) and topoisomerase I (TopoI) inhibitor topotecan (TPT). These drugs all intercalate into the DNA to trap their target topoisomerases, but have different effects in the clinic. Although the mechanism of Etop to induce DNA breaks is assumed to be similar to that of the anthracyclines, it is far less effective in the clinic and also presents with fewer side effects to patients. This suggests that the effectiveness of anthracyclines depends on additional mechanisms of action. Indeed, the anthracycline anti-cancer drugs were recently shown to also have the ability to evict histones from loose chromatin structures [2] and enhance nucleosome turnover in promoter regions [3]. Another anthracycline variant, Acla, does not induce DNA breaks but still evicts histones from chromatin [2]. Thus, histone eviction and DNA break formation are independent mechanisms. Although Acla does not induce DNA breaks, it is an effective anti-cancer drug [4]. This implies that DNA break formation is not necessarily the only anti-cancer mechanism provided by the anthracyclines. As histone H2AX is also evicted by Doxo and Daun, this histone is not available for phosphorylation by ATM/ATR following DNA breaks. As a result, DNA repair is attenuated after Doxo and Daun treatment when compared to Etop that does not evict histones [2]. In addition, since the evicted histones are replaced by new ones after anthracycline exposure, the epigenetic code is altered, as is the transcription [2]. As such, these anti-cancer drugs can also be considered epigenetic modifiers, rather than merely DNA damaging agents.

The recent development of genome-wide sequencing in combination with chromatin immunoprecipitation (ChIP-seq) technologies has enabled detailed profiling of the genome. It has been estimated that 80% of the genome can now be annotated as transcriptionally active, repressed or silent, or related to other biochemical functions, as illustrated by the presence of various histone marks and transcription factors or repressors [5]. Closely related to transcriptional activity is the degree of compaction of chromatin: chromatin should be more 'open' in order to facilitate transcription. The transcriptional machinery recruits topoisomerases to unwind DNA [6, 7], which can be targeted by TopoI and TopoII inhibitors to generate DNA double-strand breaks [8, 9]. These breaks are assumed to induce mitotic catastrophe, which may selectively eliminate cancer cells as opposed to normal, more slowly dividing, cells.

While it is known that the different anti-cancer drugs target the genome for DNA damage and/or histone eviction, it is usually assumed that the targeting is more or less uniformly random across the genome. However, it is possible that different anti-cancer drugs have distinct genomic specificities, which can have important implications for the use of these drugs in the treatment of cancer. Distinct genomic targeting specificities can also affect the epigenome, which is increasingly recognized as an important determinant in tumor formation. In many tumors, the epigenome is perturbed, including in gliomas, prostate tumors, breast tumors and various lymphomas and leukemias [10, 11]. Consequently, targeting the epigenome through the use of HDAC and histone methyltransferase inhibitors is explored as a new strategy in cancer treatment [12, 13]. Since it was recently realized that the anthracyclines also affect the epigenome by virtue of histone eviction [2] and enhancement of nucleosome turn-over around open chromatin [3], a deeper insight into the epigenomic landscape preferentially targeted by these drugs has become even more relevant [14].

Despite these recent advances in understanding the mechanisms of action of topoisomerase inhibitors, the analyses of drug targeting regions were limited to the cis-regulatory elements such as promoters or gene regions, since comprehensive data on the chromatin structure and (epi)genetic modifications were unavailable in the systems used in these studies. As a result, a number of key issues still need to be addressed. First, do anti-cancer drugs have a random genome-wide activity or are DNA damage and histone eviction restricted to particular areas in the genome? Second, while anthracycline drugs induce both DNA damage and histone eviction and then attenuate phosphorylation of H2AX (γ-H2AX), is the DNA repair attenuated throughout the genome or more restricted to particular sites in the genome? To address these questions, we generated a high-resolution genomic profile of the regions of action of various TopoI and TopoII anti-cancer drugs. We used the human erythroleukemia cell line K562 because this cell line has been extensively profiled for (epi)genomic modifications as part of the ENCODE project [5, 15]. The genomic regions of action of the TopoI inhibitor TPT, as well as those of the TopoII inhibitors Etop, Acla and Daun, were determined by ChIP-seq with γ-H2AX antibodies for the DNA damage response, and by FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements) for the regions of histone eviction. Based on these data, we show that each anti-cancer drug has a distinct (epi)genomic profile of histone eviction and/or DNA damage induction. The clinical observations that some drugs (particularly Etop and TPT) provide no additional effect when used in combination therapies [16] may be explained by overlapping (epi)genomic targeting profiles of these drugs. In addition, the selectivity of these drugs for specific types of chromatin predicts selective sensitivity for specific tumors. We illustrate this with diffuse large B-cell lymphoma (DLBCL) cells. These cells are characterized by high levels of H3K27me3 [17, 18], and are –as predicted from our chemical profiling– highly sensitive to Acla, which targets chromatin regions marked with H3K27me3. This demonstrates the potential of chemical profiling of the genome using DNA intercalating anti-cancer drugs for finding new applications of single or combination therapies in cancer.

5.2. Results

5.2.1. Anthracyclines induce histone eviction with distinct preferences

In oncology, compounds that induce DNA damage are widely used. These compounds include TopoI and TopoII inhibitors, which generate DNA breaks [8, 9]. Anthracyclines present an additional mechanism of action that enhances the effect of DNA breaks: eviction of histones and enhancement of nucleosome turnover from particular open areas of the chromatin [2, 3]. Histone eviction by anthracyclines was not observed in mitotic cells when the chromatin was fully condensed, as illustrated in Figure 1a. This simple experiment suggested that anthracycline-induced histone eviction may be dependent on specific properties of the local chromatin structure. To obtain a genome-wide, high-resolution mapping of histone eviction by different anthracyclines, we exposed K562 cells, a model cell line in the ENCODE project [15], with either the anthracyclines Daun (induces histone eviction and DNA breaks [2]) or Acla (induces histone eviction only [2]), or the chemically unrelated TopoII inhibitor Etop (induces DNA breaks only [2]). FAIRE-seq was performed to profile histone eviction in a genome-wide fashion (Figure 1b and Supplementary Table 1). The resulting profiles were normalized against the histone eviction profile of untreated K562 cells, using an autocorrelation-based approach for detecting background regions in the FAIRE-seq data (see Methods). We first assessed genome-wide pairwise correlations between the normalized FAIRE-seq profiles (Figure 1c). As expected [2], Etop did not induce histone eviction, as illustrated by the strong correlation of its FAIRE-seq profile with that of untreated control cells (Figure 1c). In contrast, the relatively weak correlation of the Daun and Acla FAIRE-seq profiles with the untreated and Etop-exposed cells indicated that Daun and Acla did induce widespread histone eviction. Note that the correlation between Daun and Acla was substantially weaker than that between Etop and untreated cells (Figure 1c; $\rho = 0.73$ for Daun and Acla; $\rho = 0.87$ for Etop and Control; $p < 10^{-10}$ for their difference using a Fisher z-transformation). This suggests that the anthracyclines Daun and Acla target overlapping, but not identical genomic regions for histone eviction. Because promoter regions are preferentially targeted by both anthracyclines [2, 3], we first determined the average drug-induced histone eviction around transcription starting sites (TSS) (Supplementary Figure 1). Although both anthracycline drugs induced histone eviction in these regions, there was also a marked difference. Daun appeared to sense the transcriptional activity in these regions, as illustrated by a preference for evicting histones at and around TSSs associated with the most strongly expressed genes (Supplementary Figure 1b). Acla, however, also induced histone eviction around TSSs of intermediately expressed genes (Supplementary Figure 1a). This indicates that different anthracycline anti-cancer drugs selectively evict histones in different genomic regions.

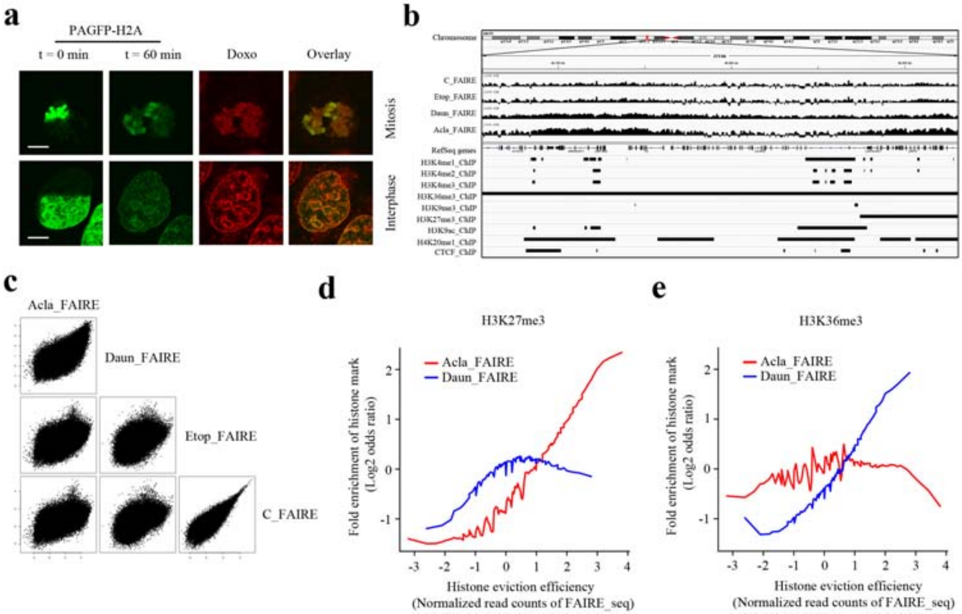


Figure 5.1: **Different anthracycline drugs induce histone eviction from unique chromatin regions.** (a) MelJuSo cells expressing PAGFP-H2A were pre-treated with nocodazole for 6 hrs to arrest cells in M phase. Part of the nucleus of a mitotic (top panel) or interphase cell (bottom panel) in the same treatment condition was photoactivated (left) followed by 9μM Doxo exposure for another 1 hr. Green color represents photoactivated histone H2A tagged with PAGFP. Redistribution of activated PAGFP-H2A in the interphase cell indicates histones were evicted by Doxo treatment, whilst activated PAGFP-H2A in the mitotic cell remained bound to the chromatin. Red color visualizes Doxo intercalated into chromatin; see also [2]. Scale bar, 10μm. (b) A snapshot of an area on chromosome 11, showing FAIRE-seq data after normalization against input DNA, aligned with selected ChIP-seq profiles from the ENCODE consortium, including H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K9me3, H3K36me, H3K9ac, H4K20me1, CTCF. (c) Genome-wide comparison of the FAIRE-seq profiles of untreated, Daun-, Acla- or Etop-treated cells, normalized to input DNA. (d and e) Enrichment of two selected histone marks (H3K27me3 and H3K36me3) in FAIRE-seq quantiles from Daun (blue line) or Acla (red line) exposed cells. The FAIRE-Seq profile was normalized to untreated cells and then sorted based on the level of histone eviction. The relative enrichment of a selected set of (epi)genomic marks was calculated within these regions as a $\log_2$ odds ratio (see Methods for further details). The x-axis represents FAIRE-seq signal strength, and the y-axis represents the enrichment of (d) H3K27me3 and (e) H3K36me3.

5.2.2. ANTHRACYCLINE DRUGS TARGET DISTINCT HISTONE MODIFICATIONS AND CHROMATIN BINDING FACTORS FOR HISTONE EVICTION

Daun and Acla target promoter regions with distinct transcriptional activity. Since transcriptional activity is associated with specific histone modifications, these two drugs may thus target different types of modified histones for eviction. Therefore, we compared the genome-wide distributions of a wide range of (epi)genomic marks to our histone eviction profiles (Supplementary Figure 2). In Etop-treated cells, no specific enrichment of any histone modifications or other chromatin-associated factors was observed (Supplementary Figure 2c), as Etop does not induce histone eviction [2]. With respect to Daun and Acla, the distributions of two histone marks were of particular interest, namely H3K36me3 and H3K27me3, both determined by ChIP-seq as part of the ENCODE project [5]. H3K36me3 is generally associated with transcriptionally active gene bodies [19], whereas H3K27me3 is associated with transcriptional repression [20, 21]. Indeed Daun preferentially evicted histones in H3K36me3-marked regions, while leaving H3K27me3-marked regions relatively untouched (Figure 1d and e). Moreover, its preference extended to a wide range of marks generally associated with transcriptional activation (Supplementary Figure 2 and 3). This suggests that Daun has the ability to sense active and relatively loose chromatin, as reported previously for its close homologue Doxo [2, 3]. Additionally, Daun also targeted regions enriched in H3K9me1 and H4K20me1 (Supplementary Figure 2), features that are associated with both transcriptional activation and repression [22–24].

Whereas Daun preferentially evicted H3K36me3-marked histones over H3K27me3-marked histones, the reverse was observed for Acla (Figure 1d and e). Acla evicted H3K27me3-marked histones, typically considered facultative heterochromatin [20, 21], but failed to evict histones from H3K9me3-marked constitutive heterochromatin (Supplementary Figure 2).

Thus, our FAIRE-seq data specify different patterns of histone eviction for different anthracyclines. Daun-induced histone eviction is more strongly associated with transcriptional activation, and Acla-induced histone eviction with transcriptional repression (Supplementary Figure 2 and 3).

5.2.3. A fine-mapping of chromatin states selectively targeted by Daun or Acla for histone eviction

The previous analyses provided important insights into the (epi)genomic marks that are present in the specific regions targeted by the different drugs. However, the (epi)genomic features used for the analyses (Figure 1d and e and Supplementary Figure 2) are strongly correlated. Hence, based on these pairwise associations, it is difficult to establish which of the features are most likely the ‘true’ determinants of the histone eviction profile. To find these determinants of histone eviction by the respective anthracyclines, we employed a feature ranking method [25] (for details, see Methods). This method ranks features according to two criteria: 1) the confidence in the relevance of a feature for predicting histone eviction, and 2) the strength of association of a feature with histone eviction. The results reveal that H3K27me3, compared to the other features, is a likely and strong determinant of Acla-induced histone eviction (Figure 2a). For Daun, the most likely and strongest determinants are early replication timing and the binding of Nsd2, which is the methyltransferase for H3K36 [26] (Figure 2b). This is in line with the analysis shown in Supplementary Figure 2. The association strength score of H3K27me3 with the Acla profile is much stronger than that of early replication timing and Nsd2 with Daun. This indicates that Acla is more specific in inducing H3K27me3-associated histone eviction than Daun in evicting histones associated with early replication timing and Nsd2 binding.

By means of these analyses (Figure 2A and B), the strongest (epi)genomic determinants of histone eviction induced by the anthracycline drugs could be inferred in a genome-wide fashion. On a genome-wide scale, this demonstrated interesting differences between the drugs. However, it provided only limited insight into the local differences and similarities between the different drugs, i.e. which genomic regions are different or shared between the different drugs, and which combinations of (epi)genomic features are found in these regions. To address these questions, histone eviction profiles were binarized to segment the genome into regions where histones are either evicted or not evicted, as compared to untreated K562 cells (see Methods). After binarization, the genome was segmented into regions that can be defined by specific combinations ("states") of histone eviction induced by different drugs. By including Etop in the analysis, in total three drugs gave rise to eight states. An additional state was included to represent histone-free DNA in untreated K562 cells, which defines the background signal. The relative fold enrichment of a selected set of (epi)genomic features was then computed for each state (Figure 2c; for the complete set of features, see Supplementary Figure 4). Regions where only Daun evicted histones, and not Acla, were highly enriched in H3K36me3 (Figure 2c, State F5 and Figure 2d; for other states see Supplementary Figure 5), in line with the analyses shown in Figure 1e. H3K36me3 marks active gene bodies, as reflected by a strong enrichment of Pol2b and H3K79me2, features that associate with transcriptional elongation. This shows that histone eviction by Daun is more selective for transcriptionally active chromatin, as was confirmed by RNA-seq data monitoring the transcriptional activity in the respective states (Supplementary Figure 3). These data may provide an additional explanation for the relative selectivity of Daun for cancer cells over healthy cells. Many cancer cells can show global transcriptional amplification in response to, for example, c-Myc over-expression [27]. This would make their DNA more susceptible to the anthracycline drugs like Daun, thus providing a therapeutic window over normal healthy cells.

The genomic targeting profile of Acla differed from that of Daun. Acla preferentially evicted histones in H3K27me3-marked genomic regions (Figure 2c, State F4 and Figure 2d). As expected, histone modification writers of H3K27me3, such as SUZ12 and EZH2 [28], were also relatively enriched in this state. H3K27me3 represents facultative heterochromatin, which is relatively more condensed, and linked to cell-type specific transcriptional silencing (Supplementary Figure 6). Despite the more condensed nature of these regions, H3K27me3-marked histones can still be evicted by Acla. This is in contrast to State F8 in Figure 2c, where no histone eviction was observed following either Acla or Daun treatment. Here, chromatin is decorated with H3K9me3, a mark of constitutive heterochromatin. Although H3K9me3 and H3K27me3 are both considered marks of heterochromatin, H3K9me3-marked heterochromatin may be more tightly compacted than H3K27me3-marked heterochromatin, thus preventing intercalation and/or the eviction process. Yet, these two types of chromatin regions can be distinguished on the basis of chemical profiling with the two anthracycline drugs.

While Daun and Acla can be distinguished by their preferential eviction of histones in H3K36me3-marked and H3K27me3-marked regions respectively, some other regions targeted by Daun or Acla displayed canonical loose chromatin structures (H3K4me3, H3K27ac, H3K9ac, CpG islands and early replication) (Figure 2c, States F1 and F2; Supplementary Figure 5, State F1 and F2). Some genomic regions targeted by both Daun and Acla displayed also a non-typical loose chromatin composition (Figure 2c, State F3; Supplementary Figure 5, State F3). For instance, the activating marks H3K27ac and Pol2b in combination with the repressive mark H3K27me3 were relatively enriched in this state. This indicates that Daun and Acla also target a more bivalent type of chromatin that may be poised for transcription [29], which is different from the more facultative type of heterochromatin observed in state F4 in Figure 2c. This also illustrates how chemical profiling allows a further dissection of the different chromatin states in cells.

To facilitate a functional interpretation of the drug-specific histone eviction preferences, we labeled genomic regions based on combinations of chromatin marks (Figure 2e; see Methods). Although both Daun and Acla evicted histones in most of the active promoter regions, Daun was more selective for active gene bodies, in concordance with the preferential eviction of H3K36me3 by Daun. Acla additionally targeted relatively silent and more condensed regions for histone eviction, such as poised promoters and some heterochromatin (Figure 2e), as expected based on its preference for the eviction of H3K27me3. The Etop enrichment profile (Figure 2e, State F7) was highly similar to the control case (Figure 2e, State F9), confirming that no specific histone eviction was induced by Etop [2].

In conclusion, the genome can be chemically profiled through the similar but not identical histone eviction effects of two members of the anthracycline class of anti-cancer drugs, Daun and Acla.

5.2.4. Drug-induced histone eviction may allow annotation of previously un-annotated genomic regions

The analyses of the histone eviction profiles, as measured by FAIRE-seq in K562 cells exposed to Daun or Acla, also revealed some histone-evicted genomic regions that could not be annotated by any of the (epi)genomic marks that were included for analyses in this study. Based on this extensive set of (epi)genomic features, we defined 13% of the genome as currently un-annotated. Within these un-annotated regions, some regions were selectively sensed for histone eviction by the different anthracyclines. Notably, regions sensed by Acla differed from those sensed by Daun. To characterize these regions, we analyzed their genomic sequence specificity. The regions targeted for histone eviction by Acla were enriched in G and C bases, and also showed high GC-content (Figure 3a and b). This differed markedly from the regions in the un-annotated genome that were targeted by Daun for histone eviction. These were rich in A and T bases, and also showed high AT-content (Figure 3c and d). This indicates that mapping the genome with chemical compounds may allow annotation of genomic regions that were previously un-annotated by any of the (epi)genomic marks tested in genome-wide analyses.

5.2.5. Mapping γ-H2AX signaling following DNA damage induced by topoisomerase I or II inhibitors

Contrary to Acla, the anthracycline Daun also induces DNA damage, as do the chemically unrelated TopoII inhibitor Etop and the TopoI inhibitor TPT. Although these drugs have been used in the clinic for decades, it is unknown whether DNA damage occurs uni-

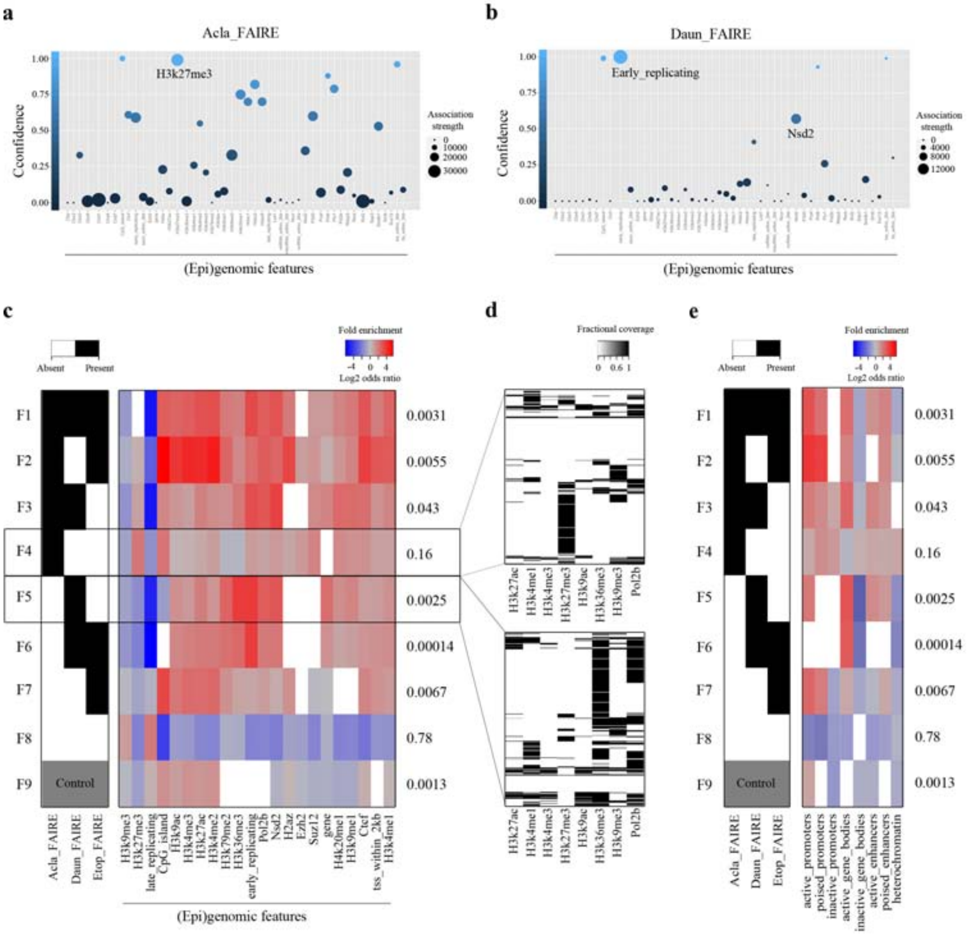


Figure 5.2: **Unique chromatin features and states are associated with drug-induced histone eviction.** (a and b) Determinants of the genome-wide Acla and Daun FAIRE-seq profiles. A Bayesian network based feature ranking method24 is employed to rank features according to two measures, 1) the confidence that a feature is relevant for explaining the FAIRE-seq profile (y-axis and color scale), calculated as bootstrap fraction; and 2) the strength of association of a feature with the FAIRE-seq profile (indicated by the size of the bubbles), calculated as the mean conditional mutual information given the Markov blanket (see Methods for further details). (c) Segmentation of the genome into states based on histone eviction induced by different drugs. The black-and-white heatmap relates the different drug effects to the different states, with black and white indicating histone eviction and no histone eviction, respectively. The heatmap shows the association of chromatin features with the different states, ranging from blue (weak association) to red (strong association), expressed as a $\log_2$ odds ratio (for details, see Methods). Significance was determined using a permutation-based approach (n=1000; see Methods). Non-significant associations are depicted as white boxes. The labels on the y-axis right represent the fraction of the genome that is covered by each state. (d) Histone modifications in two selected drug-defined genomic states, F4 and F5. Within each state, 10^4 genomic regions were randomly sampled (size-weighted) and visualized proportional to their size. For each region, the fraction of that region covered by a certain histone modification was visualized on a scale from white (no overlap) to black (complete coverage). (e) Association of the drug-defined genomic states with a range of functionally annotated (epi)genomic elements. For a definition of these elements, see Methods.

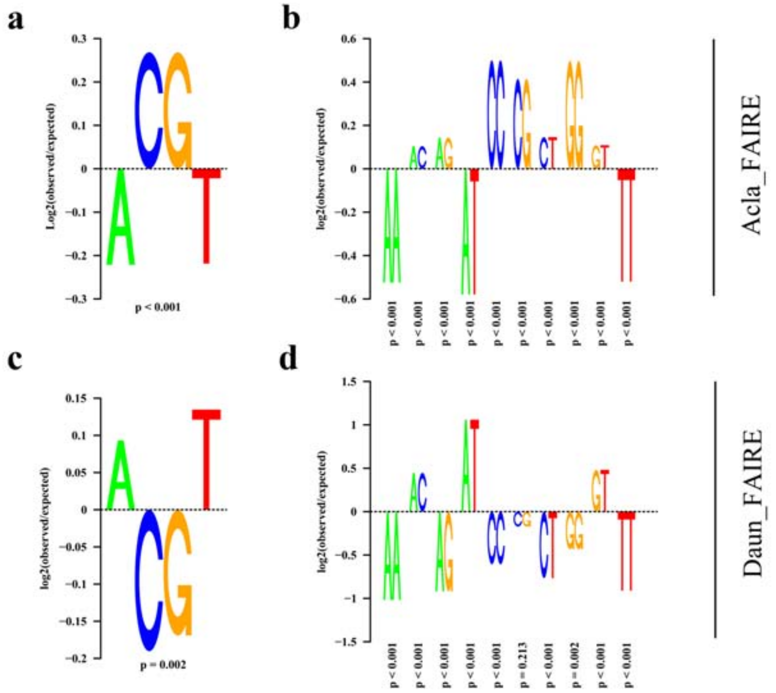


Figure 5.3: **Definition of unique regions targeted by Acla or Daun in the un-annotated parts of the genome.** Nucleotide distribution of regions showing Acla-induced histone eviction (a and b) and Daun-induced histone eviction (c and d) that were not covered by any other (epi)genomic features. For (a and c) mononucleotides and (b and d) dinucleotides are shown. Enrichment of (di)nucleotides is represented by the $\log_2$ ratio of observed number of (di)nucleotides to expected number of (di)nucleotides. Here, a positive score indicates more than expected by chance, and a negative score indicates less than expected by chance.

formly across the genome, or only at distinct genomic sites. A canonical DNA damage response includes the phosphorylation of histone variant H2AX by ATM/ATR kinases [30]. To profile the DNA damage response and relate this to histone eviction, we performed γ-H2AX ChIP-seq experiments on K562 cells after exposure to the various anti-cancer drugs (Supplementary Table 1). Similar to Etop, TPT did not induce histone eviction, as shown by a cell biology experiment (Supplementary Movie 1). γ-H2AX ChIP-seq profiles of drug-exposed cells generally showed stronger signals compared to untreated cells (Figure 4a). Overall, Daun-treated cells induced less γ-H2AX than Etop- and TPT-treated cells (Figure 4a). This can be the result of eviction of histone H2AX during anthracycline treatment. After its eviction, H2AX is no longer available for phosphorylation by the ATM/ATR kinases [2].

Similar to the FAIRE-seq signals, the γ-H2AX signals showed a diffuse domain-oriented profile, as ATR/ATM phosphorylates H2AX in relatively broad areas around the initial sites of DNA damage [31]. The computational analyses of the ChIP-seq data were performed analogous to those of the FAIRE-seq data (Figure 1 and 2). First, the γ-H2AX profiles were normalized to the γ-H2AX profiles as obtained from untreated K562 cells. Then, the feature ranking method was used to extract those features that are most predictive of the γ-H2AX profiles, in a genome-wide fashion. Compared to the other features, H3K27me3 is a likely and strong determinant of the Daun-induced γ-H2AX profile (Figure 4b). Nsd2, H3K27me3 and replication timing are likely and strong determinants of Etop-induced DNA damage signals (Figure 4c). The DNA damage response of the TopoI inhibitor TPT is determined by many of the same (epi)genomic features as the Etop profile (Figure 4d), with the exception of H3K27me3, which is more specific for Etop and Daun. This suggests that the various anti-cancer drugs induce overlapping, but also distinct, local responses to DNA damage, as visualized by phosphorylation of histone H2AX during the DNA repair process.

In order to identify which specific genomic regions are different or shared between the different drugs, and to characterize these regions in terms of combinations of (epi)genomic features, we binarized all γ-H2AX ChIP-seq profiles by applying a cut-off based on the signal from untreated cells after normalization to input DNA (see Methods). After binarization, the genome was segmented based on the three binarized profiles from drug-exposed cells (Figure 4e; for the complete set of features, see Supplementary Figure 7 and 8), similar to our strategy with the FAIRE-seq profiles in Figure 2. Both Etop and TPT elicited strong γ-H2AX responses in transcriptionally active regions, as shown by co-localization of γ-H2AX with histone marks H3K36me3 and H3K79me2, and with Pol2b (Figure 4e, States C1, C2 and C3). This was further confirmed by an analysis of gene expression levels in these states (Supplementary Figure 9), and was in line with the observation that topoisomerases TopoI and TopoII often target transcriptionally active regions, especially surrounding promoter regions [7, 33]. Yet, the other TopoII inhibitor Daun is not associated with these transcriptionally active regions. Possibly, upon eviction of H2AX by Daun in these regions, no H2AX was available for phosphorylation by ATM/ATR and hence no γ-H2AX was detected by the subsequent ChIP-seq procedure (visualized in more detail in Supplementary Figure 10). This indicates that the loss of γ-H2AX is locally coupled to drug-induced histone eviction. In addition, there were regions where Daun treatment induced a γ-H2AX

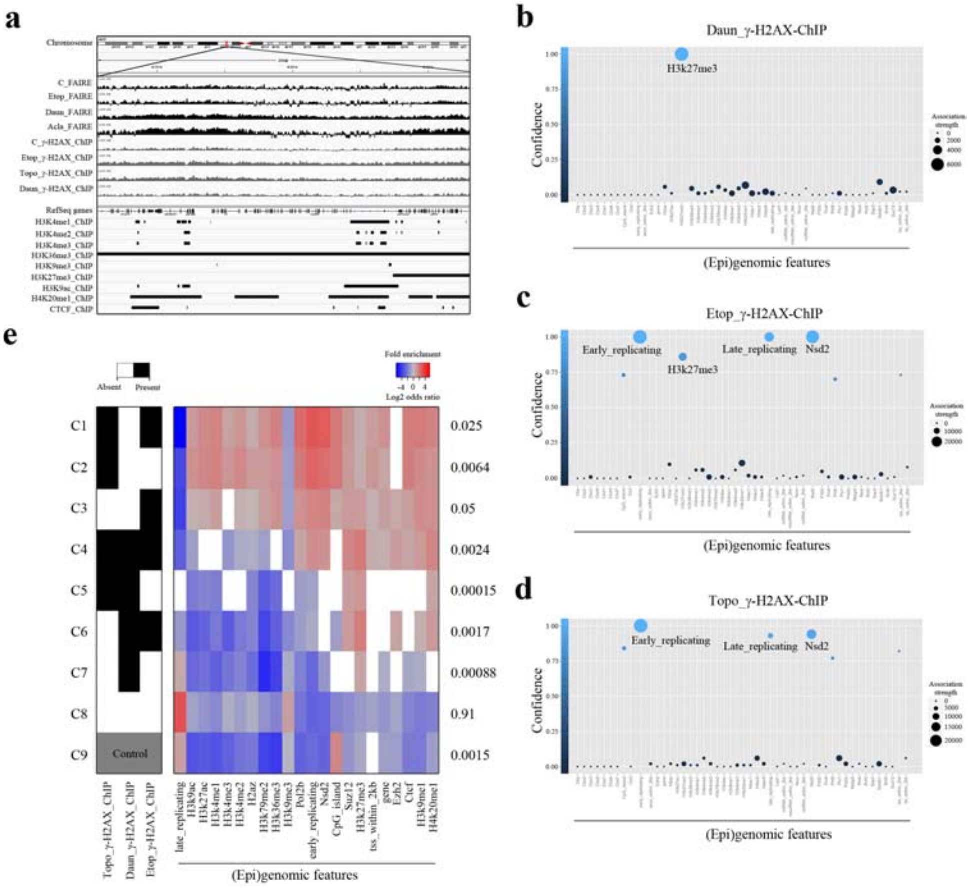


Figure 5.4: **Chromatin features and states associate with γ-H2AX induced by TopoI and TopoII inhibitors.** (a) A snapshot of an area on chromosome 11, showing FAIRE-seq and γ-H2AX ChIP-seq profiles after normalization against input DNA, aligned with selected ChIP-seq profiles from the ENCODE consortium, including H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K9me3, H3K36me, H3K9ac, H4K20me1, CTCF. (b to d) Determinants of the genome-wide γ-H2AX profiles from cells following exposure to Daun, Etop or TPT. A Bayesian network based feature ranking method [32] was employed to rank features according to two measures, 1) the confidence that a feature is relevant for explaining the γ-H2AX profile (y-axis and color scale), calculated as bootstrap fraction; and 2) the strength of association of a feature with the γ-H2AX profile (indicated by the size of the bubbles), calculated as the mean conditional mutual information given the Markov blanket (see Methods for further details). (e) Segmentation of the genome into states based on DNA damage induced γ-H2AX following exposure of K562 cells to the different drugs. The black-and-white heatmap relates the states to drugs, with black and white indicating γ-H2AX and no γ-H2AX, respectively. The heatmap shows the association of chromatin features with these states, ranging from blue (weak association) to red (strong association), expressed as a $\log_2$ odds ratio (for details, see Methods). Significance was determined using a permutation-based approach (n=1000; see Methods). Non-significant associations are depicted as white boxes. The labels on the y-axis represent the fraction of the genome covered by each state.

response (Figure 4e, States C6 and C7). These were likely regions where Daun induced DNA breaks but failed to induce histone eviction, such as the H3K27me3-marked regions (Figure 2). Since Daun induced DNA breaks in these regions, as detected by γ-H2AX, it apparently successfully intercalated into the genome, but failed to additionally induce histone eviction. Consequently, phosphorylation of H2AX was still possible. This explains why the primary γ-H2AX response following Daun exposure concentrated in these H3K27me3-marked regions (see also Supplementary Figure 10).

The sensing of transcriptional activity by the various anti-cancer drugs may have important implications for the tissue and tumor selectivity of these drugs. Promoter regions are strongly affected by anthracycline drugs [2, 3], as was also shown above (Supplementary Figure 1). Therefore, we analyzed regions surrounding TSSs for enrichment of the FAIRE-seq and γ-H2AX ChIP-seq signals. TSSs were categorized based on transcript levels as determined by RNA-seq of untreated K562 cells, in order to relate DNA damage responses to local histone eviction and gene expression (Figure 5). We included FAIRE-seq and ChIP-seq profiles from untreated K562 cells as a control. Unlike Daun (Figure 5b and e), the TopoII inhibitor Etop (Figure 5c) showed no histone eviction. The γ-H2AX signal was markedly increased –both upstream and downstream of TSSs– following Etop as well as TPT exposure in a manner strongly associated with the transcriptional activity of genes (Figure 5f and g). Daun treatment markedly reduced the γ-H2AX signal around the TSS of highly expressed genes (Figure 5e), probably following Daun-induced histone eviction around highly expressed genes (Figure 5b).

The precise application of the different DNA damaging anti-cancer drugs in treatment of different tumors is mainly based on clinical experience. Genome-wide chemical profiling of their effects on chromatin showed distinct selectivity both related to histone eviction and to DNA damage response processes, which may have important consequences for single as well as combination therapies of cancer.

5.2.6. CHEMICAL PROFILING OF THE GENOME FOR PROGNOSIS OF ANTI-CANCER DRUG SENSITIVITY

Acla is an anthracycline that was considered in the treatment of cancer in Europe and the USA, but was later abandoned due to disappointing sales (P. Brown, pers. comm.). However Acla is still actively used in China and Japan, mainly in the treatment of AML [4, 34]. Rational use of the different drugs would require a better understanding of their distinct specificities of action. We have shown above that Acla preferentially induces histone eviction in genomic regions marked by H3K27me3 (Figure 2), unlike Daun, which instead induces a γ-H2AX response in these regions (Figure 4). Given this selectivity, we wondered whether these drugs could be more effective in the treatment of tumors relying on H3K27me3 modifications as a driving feature. A large fraction of diffuse large B-cell lymphomas (DLBCL) have elevated levels of H3K27me3, due to activating mutations of the histone methyltransferase EZH2 [17], a subunit of the Polycomb Repressive Complex 2 (PRC2) [35]. Increased H3K27me3 levels drive survival of these tumors, and silencing EZH2 or chemically inhibiting EZH2 eliminates these tumors in experimental mouse models [18]. Since Acla evicted histones in H3K27me3-marked regions, we wondered whether Acla showed different efficacy compared to Daun on DLBCL cell lines harboring EZH2 activating mutations with corresponding elevated levels of H3K27me3.

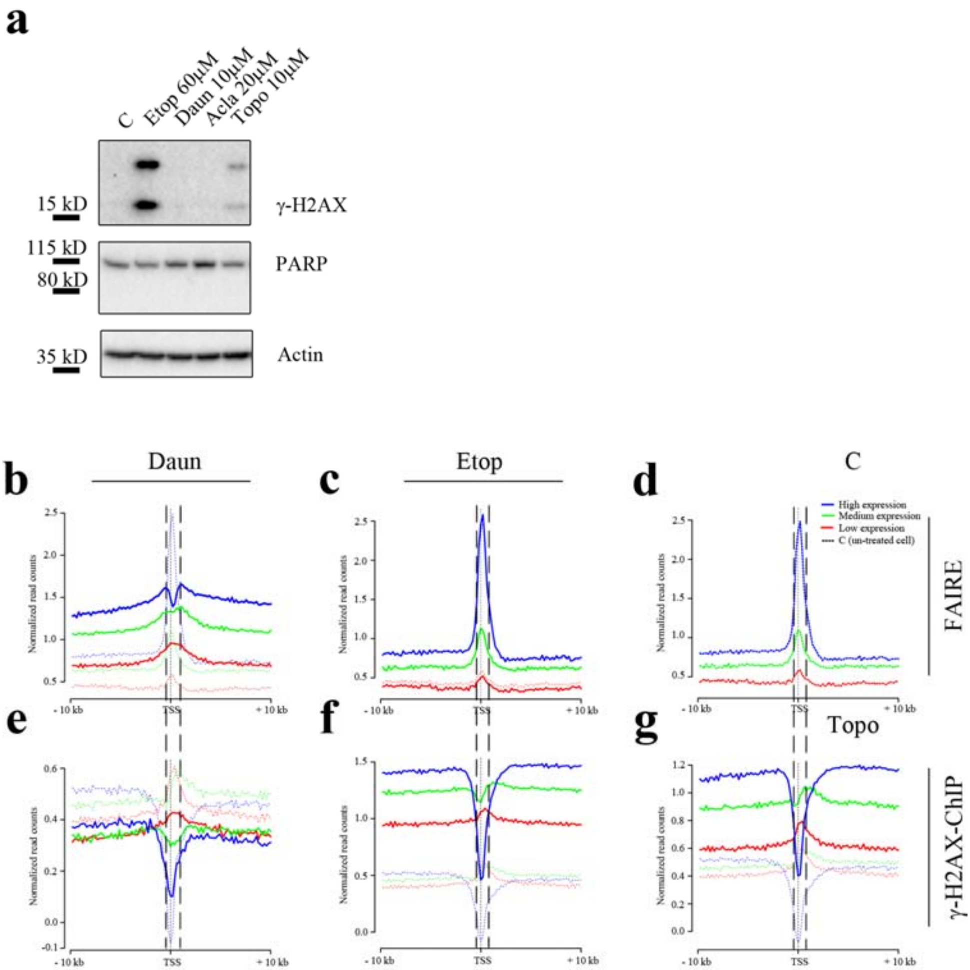


Figure 5.5: **Loss of γ-H2AX DNA damage signal is related to histone eviction during Daun treatment.** (a) K562 cells were treated with the respective drugs for 4 hrs. Equal amounts of cell lysates were analyzed by 4%-12% gradient SDS-PAGE and Western blot. No cleavage of PARP was observed, indicating that no apoptosis was initiated during treatment. Actin was used as a loading control. Left, the position of protein weight markers; right, the different proteins probed. (b to d) Average FAIRE-seq signal (normalized to input DNA) in a region 10 kb around transcription start sites (TSSs) upon treatment with Daun (b) and Etop (c), as well as without treatment (d). TSSs were grouped based on high (blue), medium (green) and low (red) expression of the associated gene, based on RNA-seq data from K562 cells. The dashed line shows the FAIRE-seq signal in untreated conditions. (e to g) Average γ-H2AX signal (normalized to input DNA) in a region 10 kb around transcription start sites (TSSs) upon treatment with Daun (e), Etop (f) or TPT (g), as indicated. TSSs were grouped based on high (blue), medium (green) and low (red) expression of the associated gene, based on RNA-seq data from K562 cells. The dashed line shows the FAIRE-seq signal in untreated conditions.

We analyzed four different lymphoma and leukemia cell lines: the erythroleukemia cell line K562, the ALL cell line HPB-ALL and two DLBCL cell lines SU-DHL-4 and Pfeiffer. The two DLBCL cell lines showed high levels of H3K27me3 due to activating mutations in EZH2 [18] (Figure 6a). On the other hand, HPB-ALL has an activating mutation in Nsd2 and thus increased levels of H3K36me3 [36], but has almost no H3K27me3 (Figure 6a). The four cell lines were treated with the different drugs, as well as with a chemical inhibitor of EZH2. Induction of apoptosis was determined by the cleavage of PARP [37]. Acla was most effective in inducing apoptosis in the DLBCL cells (Figure 6b). To further decipher differences in selectivity of the different anti-cancer drugs, the four cell lines were exposed to different concentrations of Acla, Daun or Etop, and viability was measured 24 hrs after drug exposure. The DLBCL cells were about 10-fold more sensitive to Acla than to Daun, while Daun and Acla were equally active in the other cells (Figure 6c). The DLBCL cells were not sensitive to the different concentrations of Etop that were tested (Figure 6c). This suggests that Acla induced histone eviction from H3K27me3-marked genomic regions was more efficient in eliminating DLBCL than the DNA damage induction by Daun or Etop. In conclusion, different anti-cancer drugs show selectivity for distinct genomic regions. This selectivity can potentially be translated into personalized applications of these drugs in oncology, as illustrated for DLBCL.

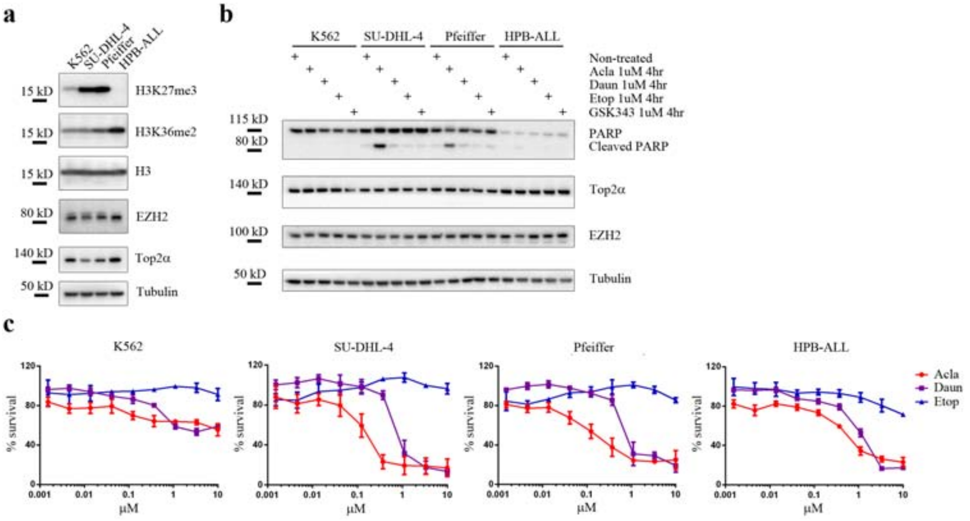


Figure 5.6: **DLBCL tumors accumulating H3K27me3 modifications are more sensitive to Acla exposure.** (a) SU-DHL-4 and Pfeiffer cells that harbor EZH2 activating mutations, HPB-ALL cells with an NSD2 activating mutation, and K562 cells were lysed and analyzed by 4%-12% SDS-PAGE and Western blotting for the proteins indicated. Total H3 and tubulin were used as loading control. Left, the position of the protein markers; right, the different proteins probed. (b) K562, SU-DHL-4, Pfeiffer and HPB-ALL cells were treated with drugs for 4 hrs at the concentrations indicated. Equal amounts of protein in cell lysates were analyzed by 4%-12% SDS-PAGE and Western blotting for the proteins indicated. Tubulin was used as loading control. The position of the marker proteins is indicated. (c) K562, SU-DHL-4, Pfeiffer and HPB-ALL cells were exposed to serial dilutions of different drugs (as indicated) for 24 hrs. Cell viability was determined by a metabolic cell titer blue assay. The data points were related to the values of untreated cells that were set at 100% survival. Data points represent the average of three independent experiments with SD.

5.3. DISCUSSION

Millions of cancer patients have been, or are currently being, treated with the anti-cancer drugs studied here [1]. These drugs are known to intercalate in the patient's DNA and, after trapping TopoI or TopoII, induce DNA damage [38, 39]. As tumor cells are likely more sensitive to DNA damage than normal cells, this is supposed to provide a therapeutic window. Whether these drugs have specificity for distinct genomic regions, and, if so, whether these specificities have any clinical consequences, are unknown. We performed a chemical profiling of the human genome with anti-cancer drugs, considering their histone eviction and DNA damage induction properties. Using the K562 cell line as a reference cell line, we profiled, at high resolution, the chromatin structure following treatment with different drugs. The computational analyses of the FAIRE-seq and ChIP-seq data showed that the different anti-cancer drugs manipulated the genome in distinct ways, with various consequences.

Chemical profiling may distinguish regions within un-annotated genomic regions. Histone eviction requires no enzymes, but only the presence of the drug at an eviction site [2]. It is dependent on the aminosugar group attached to the tetracycline group of the anthracycline, which is different between Daun and Acla. This structural difference is likely the cause of the observed differences in genomic targeting profiles of Daun and Acla. Differences are also observed within the previously un-annotated parts of the genome. Here, segments can be identified as targeted for histone eviction by Daun or Acla. These segments do not have any characteristics other than being rich in A, T and AT (Daun), or C, G and GC (Acla). It is currently unknown what structural differences between the two drugs cause this sequence bias. However it has been suggested that poly(dA:dT) sequence is less favorable for nucleosome binding [39, 40], which may be energetically easier for Daun to evict histones. This illustrates that profiling the genome with chemical compounds may allow further annotation of the human genome, as illustrated here with anti-cancer drugs.

Topoisomerase I and II inhibitors sense transcriptional activity. TopoI and TopoII inhibitors are supposed to provide a therapeutic window by virtue of the higher sensitivity of cancer cells to DNA damage. This would certainly be expected from tumors that grow faster than normal cells. Whether this sensitivity provides the only explanation for this therapeutic window in the clinic, is unclear. The analyses of the sites of action of the various inhibitors used here show that these inhibitors typically act in regions that contain active promoters that are enriched for histone marks such as H3K4me3, H3K27ac and H3K9ac, as well as enriched for CpG islands. Indeed, transcript levels, as determined by RNA-seq, associated strongly with DNA damage response signals and histone eviction around transcription start sites, as measured by γ-H2AX ChIP-seq and FAIRE-seq, respectively. This suggests that the different anti-cancer drugs tested here sense transcriptional activity. This could reflect the de-compacted nature of transcriptionally active chromatin, which would facilitate histone eviction. Transcriptional activity may be an important factor for tumor selectivity by the different TopoI and TopoII inhibitors. This is best exemplified by the c-Myc-oncogene, which is found over-expressed in many tumors and amplifies transcription of thousands of other genes [27]. As a result, c-Myc over-expression may thus render tumor cells more sensitive to anti-cancer drugs sensing transcriptional activity, such as identified in our study.

Different anthracyclines evict histones from different genomic regions. The drugs from the anthracycline class of TopoII inhibitors trap TopoII after the formation of transient DNA double-strand breaks, as such preventing ligation of the DNA breaks. Although, as an exception in this class of inhibitors, Acla is not able to induce DNA breaks, it is active in cancer treatment4. In addition to inducing DNA breaks, anthracyclines evict histones [2] and enhance nucleosome turnover [3]. The evicted histones will be replaced by other mostly nascent histones, as such changing the epigenetic code. Analyses of FAIRE-seq data show that the histone eviction profiles of the two anthracyclines only partially overlap. Acla efficiently targets genomic regions marked by H3K27me3, which typically contain repressed genes, while Daun targets more transcriptionally active regions for histone eviction. The different specificities of these homologous drugs suggest that new anthracycline variants can be developed that would target other unexplored regions of the genome, which would provide new options for selectively targeting tumor cells.

The similarity of topotecan and etoposide DNA damage targeting profiles, and its consequences for combination therapies. The genome-wide chemical profiling of the γ-H2AX response induced by TPT or Etop shows that DNA damage responses locate in transcriptionally active regions. Etop induces additional DNA damage signals in transcriptionally repressed regions marked by H3K27me3. Whether this provides an additional advantage of Etop over TPT in the clinic is unclear. Nevertheless, the fact that TopoI inhibitor TPT and TopoII inhibitor Etop generally target highly similar genomic regions may have been expected, since both enzymes are required in the unwinding of genomic regions for allowing transcription to occur [7]. Still, this observation has an interesting implication for the use of these inhibitors in combination treatments, as it indicates that combining Etop and TPT in treatment may not yield additional anti-cancer effects and only additional toxicity, which was indeed observed in the clinic [16, 41].

The DNA repair following daunorubicin exposure. Compared to Etop and TPT, the activity of the anthracycline Daun on DNA break formation is more complex because of the additional histone eviction activity that quenches phosphorylation of histone H2AX and DNA repair. Indeed, Daun targets genomic regions for histone eviction that are highly similar to the regions targeted for DNA damage by Etop and TPT. As a result, γ-H2AX is not detected in genomic regions where histones are evicted by Daun. However, γ-H2AX is induced by Daun in genomic regions marked by H3K27me3 because Daun fails to evict histones in these regions, as was observed in the FAIRE-seq experiments. This indicates that Daun does successfully intercalate into these regions for inducing DNA breaks, but fails to additionally evict histones, as was also observed in the FAIRE-seq data. Consequently, histone H2AX can still be phosphorylated, and DNA damage responses continue in these regions.

Conventional anti-cancer drugs act as epigenetic modifiers for personalized applications in the clinic. Recently, it has been recognized that many tumors can be specified by their epigenetic make-up, as they activate, over-express or repress histone modifying enzymes [11]. Our results suggest that the conventional anti-cancer drugs tested here elicit DNA damage signals and/or evict histones in genomic regions characterized by specific epigenetic make-ups, while ignoring other types of chromatin. Since histones are modified and evicted following drug treatment, the local epigenetic code may be (temporarily) reset. As such, the different anthracycline TopoII inhibitors may alter the epigenetic state of cells, with major effects on tumors dependent on a particular epigenetic state. We illustrate this by using DLBCL cell lines, where PRC2-subunit EZH2 is frequently mutated yielding a hyper-activated H3K27 histone methyltransferase and thus high levels of H3K27me3 [17]. EZH2 inhibitors are currently under development for specifically targeting these tumors [18]. Yet, the conventional compound Acla, by virtue of preferentially evicting histones in H3K27me3-marked regions, may be equally effective in eliminating these types of tumors. The current treatment for DLBCL patients involves Doxo in combination with four other drugs, known as the R-CHOP therapy [42]. We showed that DLBCL cells with EZH2 activating mutations are some 10-fold more sensitive to Acla than to Daun, while insensitive to Etop. This suggests that the local eviction of histones from H3K27me3-marked genomic regions by Acla may have more potent anti-cancer effects than the local DNA break formation by Etop, and that Acla could be a more effective additional drug in the R-CHOP therapy than the current alternatives Doxo or Daun. As increased H3K27me3 levels are also observed in other human tumors such as esophageal squamous cell carcinomas [43], hepatocellular carcinomas [44] and breast tumors [45], reconsidering Acla for such patients may provide new and personalized treatment opportunities.

5

The high-resolution genome-wide chemical profiling of the sites of action of a selected set of anti-cancer drugs shows that these drugs have strong selectivity for distinct types of chromatin and are able to sense different levels of transcriptional activity. Our analyses may provide rationale for the failure of particular combination therapies. Furthermore, they provide a basis for new combinatorial applications of conventional chemotherapeutic drugs, and for new applications of anthracycline drugs in treating specific types of cancer by sensing and manipulating their epigenetic make-up.

Author contributions

BP and JN designed experiments. XQ and BP performed experiments. JdJ designed and implemented the computational methods. JdJ and BP analyzed the data. BP, JdJ and JN drafted the manuscript with input from all authors. JN and LW supervised the study and analyses.

Acknowledgments

We thank Ron Kerkhoven, Marja Nieuwland and Arno Velds of the NKI Genomic Core Facility, Lennert Janssen for support in cloning, and Waseem Akhtar, Bas van Steensel and Piet Borst for their critical reading of the manuscript. This work was supported by an ERC grant to JN. JN is a member of the Institute of Chemical Immunology (ICI).

5.4. Materials and Methods

5.4.1. Data generation

Cell Culture

K562 cells were maintained in RPMI 1640 media with penicillin/streptomycin and 8% FCS, according to the ENCODE recommended culture conditions. Cell density reached around 0.5×106/ml before treatment. MelJuSo/PAGFP-H2A cells were maintained in IMDM with penicillin/streptomycin, G-418 and 8% FCS.

Reagents

Doxorubicin and etoposide were obtained from Pharmachemie (The Netherlands). Daunorubicin was obtained from Sanofi-Aventis (The Netherlands). Topotecan was obtained from GlaxoSmithKline (UK). Aclarubicin was obtained from Santa Cruz (US) and dissolved in dimethylsulphoxide at 20 mg/ml concentrations, aliquoted and stored at -20°C for further use.

Western blotting

Cells were lysed directly in RIPA buffer (50 mM Tris-HCl pH 7.4, 1% NP-40, 0.5% Na-deoxycholate, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA). All buffers were supplemented with a protease inhibitor cocktail (Roche). Lysates were quantified and equal amounts of total proteins were analyzed by SDS–polyacrylamide gel electrophoresis for subsequent Western blotting analysis. The following primary antibodies were used: γ-H2AX(1/1000; Millipore), poly(ADP-ribose) polymerase (1/1000; Cell Signaling), actin and tubulin (1/2000; Sigma), EZH2 (1/1000; Millipore), Top2α (1/1000; Bethyl).

Cell viability assay

Cells were seeded in 96-well plate at the density of 5×10^4 cells/well. Then drugs were added to the cells at the indicated final-concentration. Drug-treated cells were further cultured for 24 hours, and CellTiter-Blue® (Promega) was used for the quantification of viable cells.

FAIRE-seq and ChIP-seq

For FAIRE-seq, 5×10^7 K562 cells in two 15cm dishes were treated with 10 μM Daun, 60 μM Etop or 20 μM Acla for 4 h, fixed and processed as described [46]. For ChIP-seq, 5×10^7 K562 cells in two 15 cm dishes were treated with 10 μM Daun, 60 μM Etop or 10 μM Topo for 4 h, fixed and processed as previously descrived [47]. Anti-γ-H2AX (10 μg per ChIP sample; Millipore05-636) antibodies were used for ChIP. All DNA samples were processed and sequenced on Illumina HiSeq 2000 platforms, according to the provider's kits and protocols for ChIP samples. Two biological replicates were included for each treatment and for all ChIP-seq or FAIRE-seq experiments. All samples were quality controlled and processed in the same way before further analyses.

5.4.2. Computational analyses

Normalization

Normalization of a high-throughput signal x with a control y representing the background, such as input DNA or whole-cell extract, is commonly performed by estimating a normalization factor f to linearly scale y such that it can be quantitatively compared to x. For our data, we divided the reference genome into consecutive 500bp bins, and computed a normalized signal as follows:

$$x_i^{\text{norm}} = \log_2\left(\frac{x_i + 1}{y_i \times f + 1}\right) \tag{5.1}$$

Here, x_i^{norm} is the normalized signal in bin i, x_i represents the number of signal reads in bin i, y_i represents the number of control (whole-cell extract) reads in bin i, and we added a pseudocount of 1 to account for bins with no reads. Note that if and only if $x_i > y_i \times f$, then $x_i^{\text{norm}} > 0$, implying there is enrichment of our signal of interest over the background.

The normalization factor f can be computed by identifying those bins x_j that can be considered devoid of signal, i.e. containing only background reads, and then computing the ratio of the sums of signal reads and control reads in these bins:

$$f = \frac{\sum_j x_j}{\sum_j y_j} \tag{5.2}$$

To find these background bins, consider that in a normalized signal, a relatively large genomic region a containing only background bins should show up as white noise with a mean of 0. Suppose a contains n bins, namely bins m to $m+n-1$, then white noise can be identified by computing the autocorrelation in this genomic region a, for a certain number of lags k:

$$r(k) = \frac{c(k)}{c(0)} \tag{5.3}$$

Here, $c(k)$ is the autocovariance function in our genomic region a:

$$c(k) = \frac{1}{n} \sum_{i=m}^{m+n-k-1} (x_i^{\text{norm}} - \overline{x}_m^{\text{norm}})(x_{i+k}^{\text{norm}} - \overline{x}_m^{\text{norm}}) \tag{5.4}$$

Where $\overline{x}_m^{\text{norm}}$ is the mean value of x^{norm} in a:

$$\overline{x}_m^{\text{norm}} = \frac{1}{n} \sum_{i=m}^{m+n-1} x_i^{\text{norm}} \tag{5.5}$$

If the theoretical autocorrelation $\rho(k) = 0$, then the sampling autocorrelation $r(k)$ is roughly normally distributed, which can be used to determine significance:

$$r(k) \sim N\left(-\frac{1}{n}, \frac{1}{n}\right) \tag{5.6}$$

Note that until now, the assumption was that our signal was properly normalized, although the normalization factor f was still unknown. Therefore, to actually identify background regions, we temporarily set $f = 1$. We normalized x using $f = 1$ and y, and slid a 5Mb window across the genome, at 100kb steps. Since our bin size was 500bp, we had 2500 data points in each 5Mb window, and for each window w we computed the mean autocorrelation across 10 lags:

$$\overline{r}_w = \frac{1}{10} \sum_{k=1}^{10} r(k) \tag{5.7}$$

Bearing in mind that $r_w \sim N(-\frac{1}{2500}, \frac{1}{2500})$, the autocorrelation in a 5Mb window was deemed significant if the above statistic was larger than the 99% upper confidence bound:

$$\overline{r}_w > z_{1-\frac{0.01}{2}} \times \frac{1}{50} - \frac{1}{2500} \tag{5.8}$$

Using the above approach, we estimated which genomic bins could be considered background, based on setting $f = 1$. Now, let I be the set of indices representing the bins that fall within 5Mb windows called as background regions. We then computed our final estimate of the normalization factor f as follows:

$$f = \frac{\sum_{i \in I} x_i}{\sum_{i \in I} y_i} \tag{5.9}$$

UCSC gaps

From all FAIRE-seq and ChIP-seq profiles, those 500bp bins were removed that overlap with the genomic regions defined by the UCSC gaps track.

Binarization

Considering that a bit over 1% of the genome represents open chromatin in K562 cells [48], the 99th percentile of the untreated FAIRE-seq profile was used as a threshold to binarize the drug-treated FAIRE-seq profiles. For a drug-treated FAIRE-seq profile, a 500bp bin was called as enriched if the normalized read count exceeded the 99th percentile of the untreated FAIRE-seq profile, and if the normalized read count of the untreated FAIRE-seq itself did not exceed this threshold. The same approach was taken for the γ-H2AX profiles.

(Epi)genomic data

Gene-related data was downloaded from the Ensembl website (GRCh37; release 70). Repli-seq data was downloaded from the ENCODE website and processed as in [49]. RNA-seq data was downloaded from the ENCODE website, and processed using TopHat v2.0.9 [50], Bowtie v2.1.0 [51] and Cufflinks 2.1.1 [52] with default parameter settings. All other transcription factor and histone modification related data were downloaded as processed peak sets from the ENCODE website. Gene sets (MSigDB, gene symbols, v4.0) were downloaded from the Broad Institute website, and were restricted to those gene symbols that could be mapped by name to gene symbols in the Ensembl release 70.

PAIRWISE SCATTER PLOTS

Data were normalized against whole-cell extract, as described above, and 10^5 points were randomly selected and plotted.

TSS ALIGNMENT PLOTS

Data were normalized against whole-cell extract, as described above, and the average normalized read count was calculated in 50 equal-sized bins around the TSS.

GENOME-WIDE QUANTILE PLOTS

All drug-treated FAIRE-seq profiles were normalized against untreated FAIRE-seq by subtracting the untreated FAIRE-seq that was normalized against whole-cell extract from the drug-treated FAIRE-seq profile that was normalized against whole-cell extract. For each individual profile, the bins were divided into 100 quantiles, and for each quantile and each (epi)genomic mark, the overlap was computed as a $\log_2$ odds ratio. As such, a $\log_2$ odds ratio of 2 for a certain feature and a certain FAIRE-seq quantile indicates that a twice as large fraction of genomic regions marked by that feature was covered by that FAIRE-seq bins of that quantile, compared to genomic regions not marked by that feature.

GENOME SEGMENTATION BASED ON DRUG-TREATED FAIRE-SEQ OR γ-H2AX PROFILES

The FAIRE-seq profiles were binarized as explained above. Having three binarized FAIRE-seq profiles (Aclarubicin, Daunorubicin and Etoposide) thus resulted in eight possible configurations or states for each 500bp bin. For each of these states, the chromatin composition was determined by computing overlap with (epi)genomic marks as a $\log_2$ odds ratio. As such, a $\log_2$ odds ratio of 2 for a certain feature and a certain state indicates that a twice as large fraction of genomic regions marked by that feature was covered by that state, compared to genomic regions not marked by that feature. Significance of the resulting associations was determined by circular permutation ($p = 10^3$ equidistant permutation shifts) of states across genomic 500bp bins. To avoid contamination of the null distribution by artificial autocorrelation, the first permutation shift was set equal to the size of the largest region of consecutive bins defining a single state. A log-odds ratio for an (epi)genomic mark was called significant if none of the permutation $\log_2$ odds ratios were more extreme (i.e. $0 < p < 0.001$).

GSEA

All drug-treated FAIRE-seq profiles were normalized against untreated FAIRE-seq by subtracting theuntreated FAIRE-seq that was normalized against whole-cell extract from the drug-treated FAIRE-seq profile that was normalized against whole-cell extract. For each gene (GRCh37; Ensembl release 70), the mean FAIRE-seq normalized read count was calculated, by averaging the read counts in those 500bp bins that overlapped the gene. For each pair of drug-treated FAIRE-seq profile and MSigDB gene set, a GSEA score was computed by performing a Mann-Whitney U-test of genes in the gene set against genes not in the gene set. A similar GSEA was performed using the gene expression values, instead of the FAIRE-seq normalized read counts.

Sequence composition

From the set of eight binarized FAIRE-seq and ChIP-seq profiles (see above), those bins were removed that 1) overlap with the genomic regions defined in the UCSC gaps track and 2) overlap with any of the (epi)genomic marks used in our analyses, or with more than one of our binarized FAIRE-seq or ChIP-seq profiles. Therefore, the remaining regions are strictly unique for a single drug, and cannot be characterized in terms of other (epi)genomic marks. For each of these sets of drug-profiled regions, mononucleotide counts as fractions of total mononucleotide count, i.e. discrete probability distributions, were determined:

$$f(x) = \frac{N_x}{\sum_{y \in \{\mathrm{A,C,G,T}\}} N_y}, x \in \{\mathrm{A,C,G,T}\} \tag{5.10}$$

Here, N_x and N_y represent the counts of a certain nucleotide x or y. Then, corresponding frequencies were determined for the whole genome:

$$g(x) = \frac{N_x}{\sum_{y \in \{\mathrm{A,C,G,T}\}} N_y}, x \in \{\mathrm{A,C,G,T}\} \tag{5.11}$$

Normalized mononucleotide frequencies used for plotting were computed by taking the $\log_2$ ratio of the drug-profiled region frequencies with the whole-genome frequences:

$$\log_2 \frac{f(x)}{g(x)} \tag{5.12}$$

For determining statistical significance of the observed nucleotide distribution $f(x)$, the base 2 Shannon entropy of this distribution was computed:

$$H_f = -\sum_{x \in \{\mathrm{A,C,G,T}\}} f(x) \times \log_2(f(x)) \tag{5.13}$$

Then, a null distribution of expected frequencies $f(x)$ was generated by circularly permuting the identified bins across the genome ($p = 10^4$ equidistant permutation shifts), counting mononucleotide occurrences, and determining their Shannon entropies. To avoid contamination of the null distribution by artificial autocorrelation, the first permutation shift was set equal to the size of the largest consecutive drug-profiled region.

Dinucleotide frequencies were determined in similar fashion. However, in the dinucleotide case, significance was inferred by comparing each of the individual dinucleotide frequencies to their corresponding null distribution, resulting in a single p-value for each dinucleotide.

Markov blanket / conditional mutual information analysis

Broadly, the approach as previously applied to retroviral and transposon integration target site selection [25] was used. Genome-wide FAIRE-seq and γ-H2AX profiles were binarized as outlined above, resulting in a value of 0 or 1 for each 500bp bin. Resulting binarized profiles were normalized with the untreated profiles by setting to 0 each bin for which the untreated showed a 1. For the (epi)genomic features, a 500bp bin was set to 1 if a particular feature overlapped with that bin, and 0 otherwise. For each drug-treated profile in combination with the binarized (epi)genomic profiles, we inferred Bayesian networks using BANJO [32]. Since BANJO employs simulated annealing for determining the network structure, we performed 400 bootstraps of size 40000. In each resulting network, we identified the Markov blanket of the drug-treated node. Using the Markov blankets, the importance of an (epi)genomic feature for drug-specific histone eviction or γ-H2AX was determined in two dimensions. First, the fraction of times was determined that a certain feature occurred in the Markov blanket. This represented the confidence that this feature is truly relevant. Second, the conditional mutual information was computed of this feature with the drug-treated profile given the remaining features in the Markov blanket, taking the average across all Markov blankets. This measure represented the unique strength of association of this feature with the drug-treated profile, that can not be explained by any of other (epi)genomic features in the analysis.

To take into account the strong autocorrelation along the genome of any (epi)genomic profile, we included an autocorrelation node for each feature in the Bayesian network, which represented the ceiling of the average binarized feature values of this (epi)genomic mark in the neighboring bins. During the network structure learning using BANJO, these autocorrelation nodes were fixed parents of their corresponding (epi)genomic feature node, and were allowed no connections to other nodes.

DEFINITION OF REGULATORY ELEMENTS

1. Active promoters: An H3K4me3 or H3K9ac peak within 3kb upstream of a TSS or 1kb downstream of a TSS, but no H327me3 or H3K9me3 peaks.

2. Poised promoters: An H3K4me3 and H3K27me3 peak within 3kb upstream of a TSS or 1kb downstream of a TSS.

3. Inactive promoters: An H3K9me3 or H3K27me3 peak within 3kb upstream of a TSS or 1kb downstream of a TSS, but no H3K4me3 or H3K9ac peaks.

4. Active gene bodies: An H3K36me3 peak within a gene body, but no H327me3 or H3K9me3 peaks.

5. Inactive gene bodies: An H3K36me3 or H3K9me3 within a gene body, but no H3K36me3.

6. Active enhancers: An H3k4me1 peak and an H3k27ac spaced at most 500bp apart, or alternatively a p300 peak alone suffices. All regions within 5kb of an H3K4me3 peak are removed.

7. Poised enhancers: An H3k4me1 peak with a 250bp slack. All regions within 5kb of an H3K4me3, p300 or H3K27ac peak are removed.

8. Heterochromatin: An H3K27me3 or H3K9me3 peak.

REFERENCES

[1] F.-M. Arcamone, *Fifty years of chemical research at farmitalia.* Chemistry **15**, 7774 (2009).

[2] B. Pang, X. Qiao, L. Janssen, A. Velds, T. Groothuis, R. Kerkhoven, M. Nieuwland, H. Ovaa, S. Rottenberg, O. van Tellingen, J. Janssen, P. Huijgens, W. Zwart, and J. Neefjes, *Drug-induced histone eviction from open chromatin contributes to the chemotherapeutic effects of doxorubicin,* Nature Communications **4**, 1908+ (2013).

[3] F. Yang, C. J. Kemp, and S. Henikoff, *Doxorubicin enhances nucleosome turnover around promoters.* Curr Biol **23**, 782 (2013).

[4] J. Jin, J.-X. Wang, F.-F. Chen, D.-P. Wu, J. Hu, J.-F. Zhou, J.-D. Hu, J.-M. Wang, J.-Y. Li, X.-J. Huang, J. Ma, C.-Y. Ji, X.-P. Xu, K. Yu, H.-Y. Ren, Y.-H. Zhou, Y. Tong, Y.-J. Lou, W.-M. Ni, H.-Y. Tong, H.-F. Wang, Y.-C. Mi, X. Du, B.-A. Chen, Y. Shen, Z. Chen, and S.-J. Chen, *Homoharringtonine-based induction regimens for patients with de-novo acute myeloid leukaemia: a multicentre, open-label, randomised, controlled phase 3 trial.* Lancet Oncol (2013).

[5] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B.-K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, J. B. Brown, C. Cheng, S. Djebali, X. Dong, J. Ernst, T. S. Furey, B. Giardine, M. Greven, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, P. Kheradpour, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. J. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, C. Gunter, M. J. Pazin, R. F. Lowdon, L. A. L. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, P. J. Good, E. A. Feingold, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, M. L. Eaton, A. Dobin, A. Tanzer, J. Lagarde, W. Lin, C. Xue, B. A. Williams, C. Zaleski, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, D. Robyr, X. Ruan, M. Sammeth, K. Singh Sandhu, L. Schaeffer, L.-H. See, A. Shahab, J. Skancke, A. Maria Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, Y. Hayashizaki, A. Reymond, S. E. Antonarakis, G. J. Hannon, Y. Ruan, P. Carninci, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, L. L. Grasfeder, P. G. Giresi, A. Battenhouse, N. C. Sheffield, K. A. Showers, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. Ki Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. Jae Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, V. R. Iyer, M. Zheng, P. Wang, J. Gertz, J. Vielmetter, E. Christopher Partridge, K. E. Varley, C. Gasper,

A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, M. A. Muratet, N. S. Davis, K. McCue, T. Eggleston, K. I. Fisher-Aylor, G. DeSalvo, S. K. Meadows, S. Balasubramanian, A. S. Nesmith, J. Scott Newberry, K. M. Newberry, S. L. Parker, B. Pusey, S. E. Levy, D. M. Absher, W. H. Wong, M. J. Blow, A. Visel, L. A. Pennachio, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, C. Davidson, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. Manuel Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, T. Hunt, I. Jungreis, M. Kay, E. Khurana, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. Manuel Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, E. Tapanari, M. L. Tress, M. J. van Baren, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler, A. Valencia, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, D. Raha, L. Ramirez, B. Reed, M. Shi, T. Slifer, H. Witt, L. Wu, X. Xu, K.-K. Yan, X. Yang, K. Struhl, S. M. Weissman, L. O. Penalva, S. Karmakar, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, A. Victorsen, T. Auer, L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, J. Wittbrodt, B. Wittbrodt, G. Jain, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. Scott Hansen, L. Boatman, E. Haugen, R. Humbert, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, M. E. Sanchez, R. S. Sandstrom, A. O. Shafer, A. B. Stergachis, S. Thomas, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, K. Beal, A. Brazma, P. Flicek, N. Johnson, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, W. Miller, P. J. Bickel, B. Banfai, N. P. Boley, H. Huang, J. Jessica Li, W. Stafford Noble, J. A. Bilmes, O. J. Buske, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, L. Lochovsky, B. E. Bernstein, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigó, R. C. Hardison, T. J. Hubbard, M. Kellis, W. James Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, and E. Birney, *An integrated encyclopedia of DNA elements in the human genome,* Nature **489**, 57 (2012).

[6] N. Mondal and J. D. Parvin, *DNA topoisomerase IIα is required for RNA polymerase II transcription on chromatin templates,* Nature **413**, 435 (2001).

[7] A. Sperling, K. Jeong, T. Kitada, and M. Grunstein, *Topoisomerase ii binds nucleosome-free dna and acts redundantly with topoisomerase i to enhance recruit-*

5

ment of rna pol ii in budding yeast, Proceedings of the National Academy of Sciences of the United States of America **108**, 12693 (2011).

[8] J. L. Nitiss, *Targeting dna topoisomerase ii in cancer chemotherapy,* Nature Reviews Cancer **9**, 338 (2009).

[9] Y. Pommier, *Topoisomerase i inhibitors: camptothecins and beyond.* Nature reviews. Cancer **6**, 789 (2006).

[10] M. Rodríguez-Paredes and M. Esteller, *Cancer epigenetics reaches mainstream oncology.* Nature medicine **17**, 330 (2011).

[11] W. Timp and A. P. Feinberg, *Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host,* Nature Reviews Cancer **13**, 497 (2013).

[12] N. Azad, C. A. Zahnow, C. M. Rudin, and S. B. Baylin, *The future of epigenetic therapy in solid tumours[mdash]lessons from the past,* Nat Rev Clin Oncol **10**, 256 (2013).

[13] M. Roschewski, L. M. Staudt, and W. H. Wilson, *Diffuse large b-cell lymphoma—treatment approaches in the molecular era,* Nature Reviews Clinical Oncology (2013).

[14] R. Rodriguez and K. M. Miller, *Unravelling the genomic targets of small molecules using high-throughput sequencing,* Nat Rev Genet **15**, 783 (2014).

[15] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O/'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and M. Snyder, *Architecture of the human regulatory network derived from ENCODE data,* Nature **489**, 91 (2012).

[16] J. Sehouli, D. Stengel, G. Oskay-Oezcelik, A. G. Zeimet, H. Sommer, P. Klare, M. Stauch, A. Paulenz, O. Camara, E. Keil, and W. Lichtenegger, *Nonplatinum topotecan combinations versus topotecan alone for recurrent ovarian cancer: results of a phase iii study of the north-eastern german society of gynecological oncology ovarian cancer study group.* J Clin Oncol **26**, 3176 (2008).

[17] C. J. Sneeringer, M. P. Scott, K. W. Kuntz, S. K. Knutson, R. M. Pollock, V. M. Richon, and R. A. Copeland, *Coordinated activities of wild-type plus mutant ezh2 drive tumor-associated hypertrimethylation of lysine 27 on histone h3 (h3k27) in human b-cell lymphomas.* Proc Natl Acad Sci U S A **107**, 20980 (2010).

[18] M. T. McCabe, H. M. Ott, G. Ganji, S. Korenchuk, C. Thompson, G. S. Van Aller, Y. Liu, A. P. Graves, A. D. Pietra, E. Diaz, L. V. LaFrance, M. Mellinger, C. Duquenne, X. Tian, R. G. Kruger, C. F. McHugh, M. Brandt, W. H. Miller, D. Dhanak, S. K. Verma, P. J.

Tummino, and C. L. Creasy, *EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations,* Nature **492**, 108 (2012).

[19] P. Kolasinska-Zwierz, T. Down, I. Latorre, T. Liu, X. S. Liu, and J. Ahringer, *Differential chromatin marking of introns and expressed exons by H3K36me3,* Nat Genet **41**, 376 (2009).

[20] A. Barski, S. Cuddapah, K. Cui, T.-Y. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, *High-resolution profiling of histone methylations in the human genome.* Cell **129**, 823 (2007).

[21] P. Trojer and D. Reinberg, *Facultative heterochromatin: is there a distinctive molecular signature?* Molecular cell **28**, 1 (2007).

[22] J. Gupta, S. Kumar, J. Li, R. Krishna Murthy Karuturi, and K. Tikoo, *Histone h3 lysine 4 monomethylation (H3K4me1) and h3 lysine 9 monomethylation (H3K9me1): Distribution and their association in regulating gene expression under hyperglycaemic/hyperinsulinemic conditions in 3T3 cells,* Biochimie (2012).

[23] D. B. Beck, H. Oda, S. S. Shen, and D. Reinberg, *Pr-set7 and h4k20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription.* Genes Dev **26**, 325 (2012).

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

[25] J. de Jong, W. Akhtar, J. Badhai, A. G. Rust, R. Rad, J. Hilkens, A. Berns, M. van Lohuizen, L. F. A. Wessels, and J. de Ridder, *Chromatin landscapes of retroviral and transposon integration profiles,* PLoS Genet **10**, e1004250+ (2014).

[26] A. J. Kuo, P. Cheung, K. Chen, B. M. Zee, M. Kioi, J. Lauring, Y. Xi, B. H. Park, X. Shi, B. A. Garcia, W. Li, and O. Gozani, *NSD2 links dimethylation of histone h3 at lysine 36 to oncogenic programming,* Molecular Cell **44**, 609 (2011).

[27] C. Y. Lin, J. Lovén, P. B. Rahl, R. M. Paranal, C. B. Burge, J. E. Bradner, T. I. Lee, and R. A. Young, *Transcriptional amplification in tumor cells with elevated c-Myc,* Cell **151**, 56 (2012).

[28] R. Cao and Y. Zhang, *Suz12 is required for both the histone methyltransferase activity and the silencing function of the eed-ezh2 complex.* Mol Cell **15**, 57 (2004).

[29] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander, *A bivalent chromatin structure marks key developmental genes in embryonic stem cells.* Cell **125**, 315 (2006).

[30] T. Misteli and E. Soutoglou, *The emerging role of nuclear architecture in DNA repair and genome maintenance,* Nature Reviews Molecular Cell Biology **10**, 243 (2009).

[31] S. Polo and S. Jackson, *Dynamics of dna damage response proteins at dna breaks: a focus on protein modifications,* Genes & development **25**, 409 (2011).

[32] J. Pearl, *Bayesian networks: A model of self-activated memory for evidential reasoning,* in *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine* (1985) pp. 329–334.

[33] S. Thakurela, A. Garding, J. Jung, D. Schuebeler, L. Burger, and V. K. Tiwari, *Gene regulation and priming by topoisomerase iiα; in embryonic stem cells.* Nat Commun **4**, 2478 (2013).

[34] G. Wei, W. Ni, J. Chiao, Z. Cai, H. Huang, and D. Liu, *A meta-analysis of cag (cytarabine, aclarubicin, g-csf) regimen for the treatment of 1029 patients with acute myeloid leukemia and myelodysplastic syndrome,* Journal of hematology & oncology **4**, 46 (2011).

[35] A. Sparmann and M. van Lohuizen, *Polycomb silencers control cell fate, development and cancer.* Nature reviews. Cancer **6**, 846 (2006).

[36] J. D. Jaffe, Y. Wang, H. M. Chan, J. Zhang, R. Huether, G. V. Kryukov, H.-e. C. Bhang, J. E. Taylor, M. Hu, N. P. Englund, F. Yan, Z. Wang, E. Robert McDonald, L. Wei, J. Ma, J. Easton, Z. Yu, R. deBeaumount, V. Gibaja, K. Venkatesan, R. Schlegel, W. R. Sellers, N. Keen, J. Liu, G. Caponigro, J. Barretina, V. G. Cooke, C. Mullighan, S. A. Carr, J. R. Downing, L. A. Garraway, and F. Stegmeier, *Global chromatin profiling reveals NSD2 mutations in pediatric acute lymphoblastic leukemia,* Nat Genet **45**, 1386 (2013).

[37] B. Pang, J. Neijssen, X. Qiao, L. Janssen, H. Janssen, C. Lippuner, and J. Neefjes, *Direct antigen presentation and gap junction mediated cross-presentation during apoptosis.* J Immunol **183**, 1083 (2009).

[38] B. L. Staker, K. Hjerrild, M. D. Feese, C. A. Behnke, A. B. Burgin, and L. Stewart, *The mechanism of topoisomerase i poisoning by a camptothecin analog.* Proceedings of the National Academy of Sciences of the United States of America **99**, 15387 (2002).

[39] K. Struhl and E. Segal, *Determinants of nucleosome positioning,* Nature Structural & Molecular Biology **20**, 267 (2013).

[40] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom, *A genomic code for nucleosome positioning,* Nature **442**, 772 (2006).

[41] N. Vey, H. Kantarjian, M. Beran, S. O'Brien, J. Cortes, C. Koller, and E. Estey, *Combination of topotecan with cytarabine or etoposide in patients with refractory or relapsed acute myeloid leukemia: results of a randomized phase i/ii study.* Invest New Drugs **17**, 89 (1999).

[42] L. H. Sehn, *A decade of r-chop,* Blood **116**, 2000 (2010).

[43] L.-R. He, M.-Z. Liu, B.-K. Li, H.-L. Rao, Y.-J. Liao, X.-Y. Guan, Y.-X. Zeng, and D. Xie, *Prognostic impact of H3K27me3 expression on locoregional progression after chemoradiotherapy in esophageal squamous cell carcinoma,* BMC Cancer **9**, 461+ (2009).

[44] M. Y. Cai, J. H. Hou, H. L. Rao, R. Z. Luo, M. Li, X. Q. Pei, M. C. Lin, X. Y. Guan, H. F. Kung, Y. X. Zeng, and D. Xie, *High expression of h3k27me3 in human hepatocellular carcinomas correlates closely with vascular invasion and predicts worse prognosis in patients.* Mol Med **17**, 12 (2011).

[45] F. Cottini, T. Hideshima, C. Xu, M. Sattler, M. Dori, L. Agnelli, E. T. Hacken, M. T. Bertilaccio, E. Antonini, A. Neri, M. Ponzoni, M. Marcatti, P. G. Richardson, R. Carrasco, A. C. Kimmelman, K.-K. Wong, F. Caligaris-Cappio, G. Blandino, W. M. Kuehl, K. C. Anderson, and G. Tonon, *Rescue of hippo coactivator YAP1 triggers DNA damage-induced apoptosis in hematological cancers,* Nature Medicine **20**, 599 (2014).

[46] K. J. Gaulton, T. Nammo, L. Pasquali, J. M. Simon, P. G. Giresi, M. P. Fogarty, T. M. Panhuis, P. Mieczkowski, A. Secchi, D. Bosco, T. Berney, E. Montanya, K. L. Mohlke, J. D. Lieb, and J. Ferrer, *A map of open chromatin in human pancreatic islets.* Nature genetics **42**, 255 (2010).

[47] D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom, *ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions,* Methods **48**, 240 (2009).

[48] L. Song, Z. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B.-K. K. Lee, N. C. Sheffield, S. Gräf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb, and T. S. Furey, *Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.* Genome research **21**, 1757 (2011).

[49] R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, and J. A. Stamatoyannopoulos, *Sequencing newly replicated DNA reveals widespread plasticity in human replication timing,* Proceedings of the National Academy of Sciences **107**, 139 (2009).

[50] S. Y. Kim, S. Imoto, and S. Miyano, *Inferring gene networks from time series microarray data using dynamic bayesian networks,* Brief Bioinform **4**, 228 (2003).

[51] B. Langmead and S. L. Salzberg, *Fast gapped-read alignment with bowtie 2,* Nat Meth **9**, 357 (2012).

[52] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, *Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation,* Nat Biotechnol **28**, 511 (2010).

5.5. SUPPLEMENTARY MATERIAL

5.5.1. SUPPLEMENTARY FIGURES

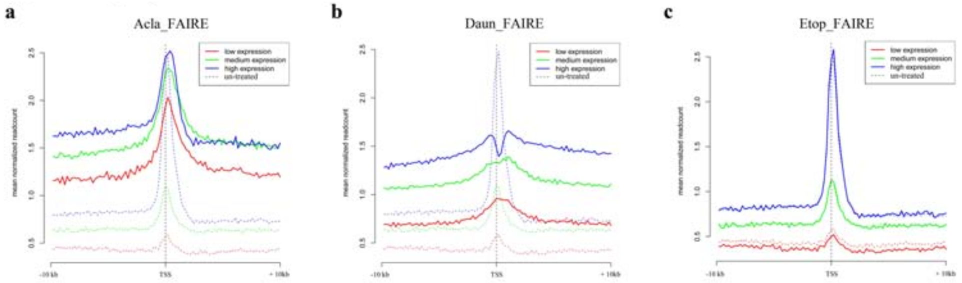


Figure S 5.1: **Drug-induced histone eviction around promoter regions.** Average FAIRE-seq signal (normalized to input DNA) around transcription start sites (TSSs) upon treatment with a) Acla, b) Daun and c) Etop. The dashed line shows the FAIRE-seq signal in untreated samples.

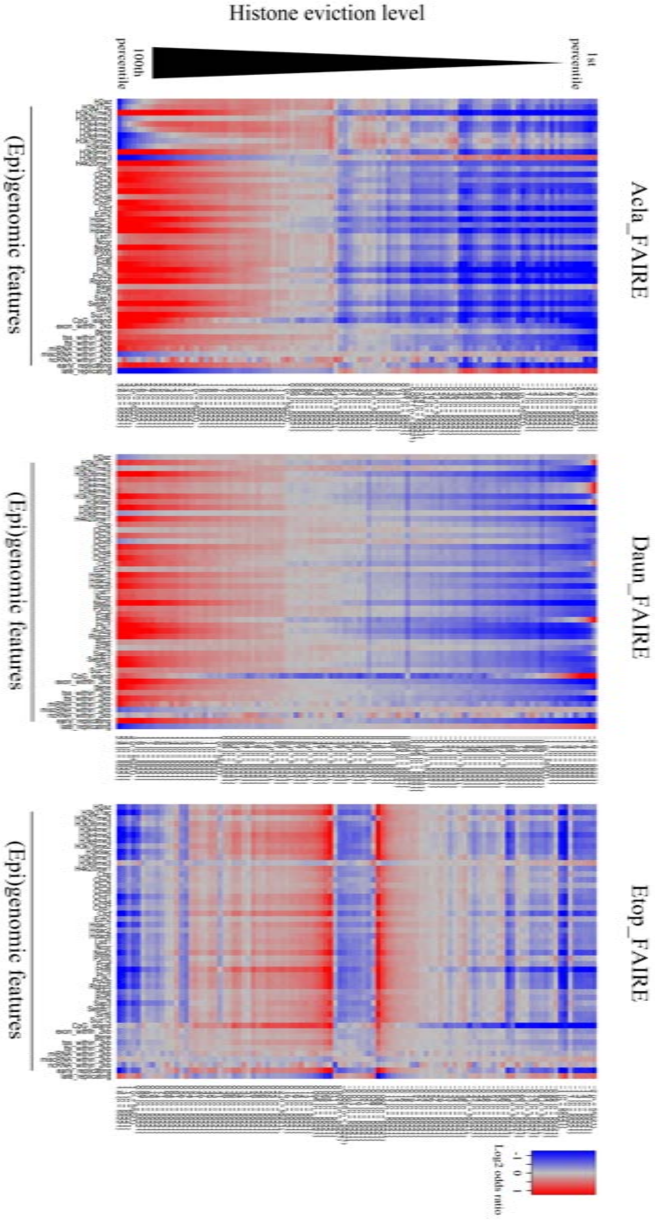


Figure S 5.2: **(Epi)genomic features enriched in drug-induced histone eviction regions.** Association of the FAIRE-seq profile from Etop-treated cells (normalized to FAIRE-seq in untreated samples) with various (epi)genomic features. The genome-wide FAIRE-seq profile is divided into 100 quantiles, and sorted along the y-axis. For each pair of chromatin feature and FAIRE-seq quantile, the overlap is calculated between the genomic bins representing that FAIRE-seq quantile, and the genomic regions marked by that chromatin feature, ranging from blue (weak overlap) to red (strong overlap).

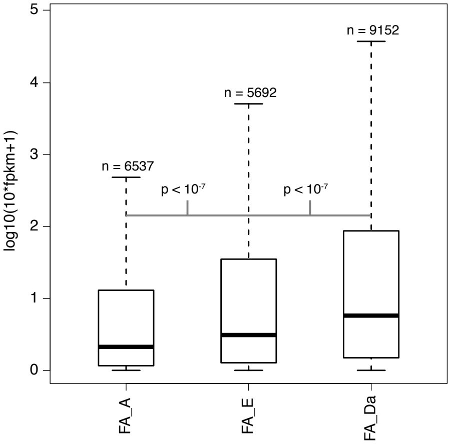


Figure S 5.3: **Transcriptional activity in drug-induced histone eviction regions.** Association of FAIRE-seq profiles (normalized to FAIRE-seq in untreated condition) with RNA-seq. A gene was associated with a drug if it was located within 2kb of one of the top 5% FAIRE-seq bins for that drug. Then, a Mann-Whitney test was performed to test the difference in expression levels between genes associated with one drug and genes associated with another drug.

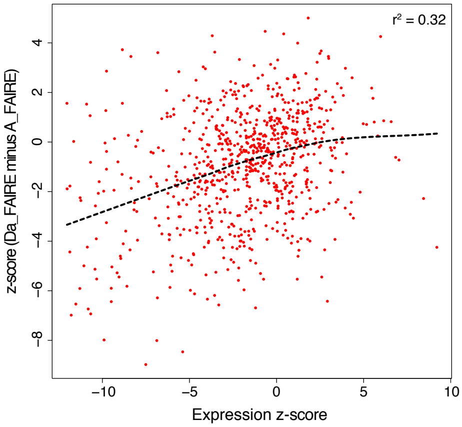


Figure S 5.4: **Comparison of gene sets targeted by different drugs for histone eviction.** Gene set enrichment analysis (GSEA) based on the Mann-Whitney U test, using Daun and Acla FAIRE-seq read counts (normalized to untreated samples), and RNA-seq RPKM (see Methods). Each point represents a gene set. On the x-axis, gene sets are scored by GSEA expression $z-$score. On the y-axis gene sets are scored by the difference of the GSEA Daun FAIRE-seq score and the GSEA Acla FAIRE-seq score. The correlation is positive, showing that Daun targets more highly expressed gene sets for histone eviction, compared to Acla.

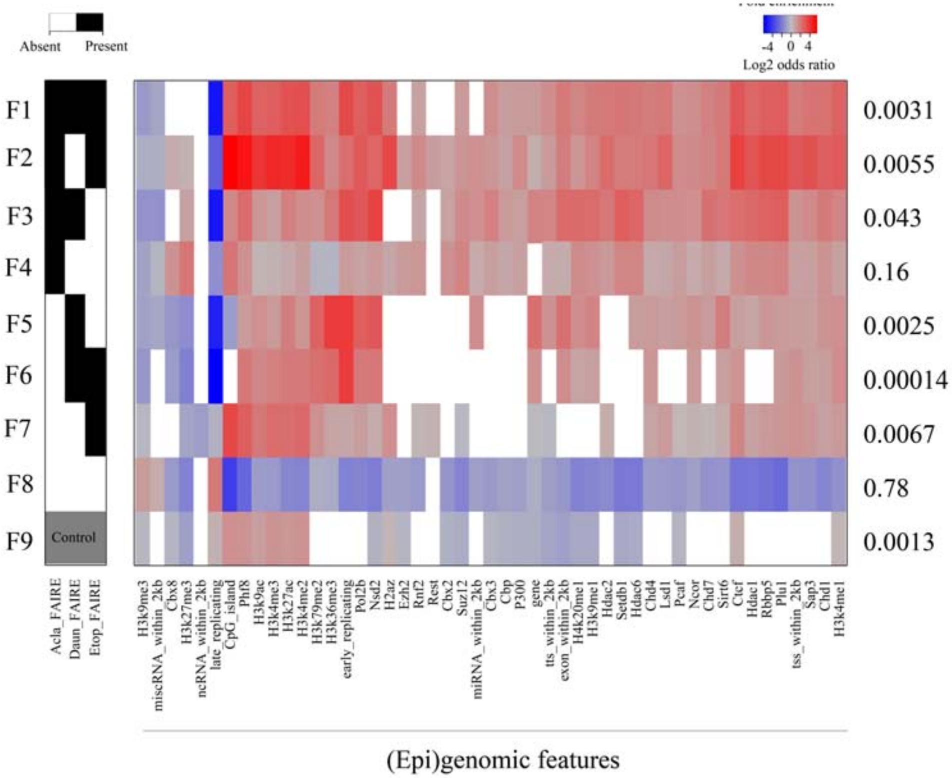


Figure S 5.5: **(Epi)genomic features targeted differently or similarly by distinct drugs for histone eviction.** Segmentation of the genome into states based on histone eviction induced by different drugs. The black-and-white heatmap relates the different drug effects to the different states, with black and white indicating histone eviction and no histone eviction, respectively. The blue-to-red heatmap shows the association of chromatin features with the different states, ranging from blue (weak association) to red (strong association). The measure of association is the $\log_2$ odds ratio (see Methods for details), where for example a value of 2 for a certain feature and a certain state indicates that a twice as large fraction of genomic regions marked by that feature was covered by that state, compared to genomic regions not marked by that feature. Significance was determined using a permutation-based approach (p =1000; see Methods). Non-significant associations are depicted as white boxes. The labels on the y-axis represent the fraction of the genome that is covered by each state.

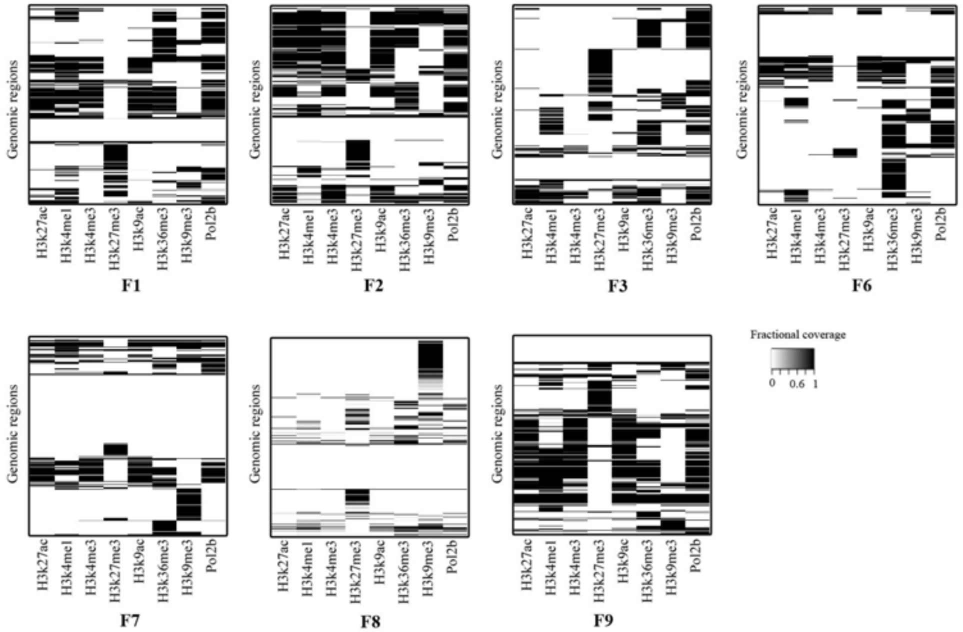


Figure S 5.6: **Representative of major histone modifications in drug-defined genomic states (FAIRE-seq).** Within each state, 10^4 genomic regions were randomly sampled (size-weighted), and visualized proportional to their size. For each region, the fraction of that region that was covered by a certain histone modification was visualized on a scale from white (no overlap) to black (complete coverage).

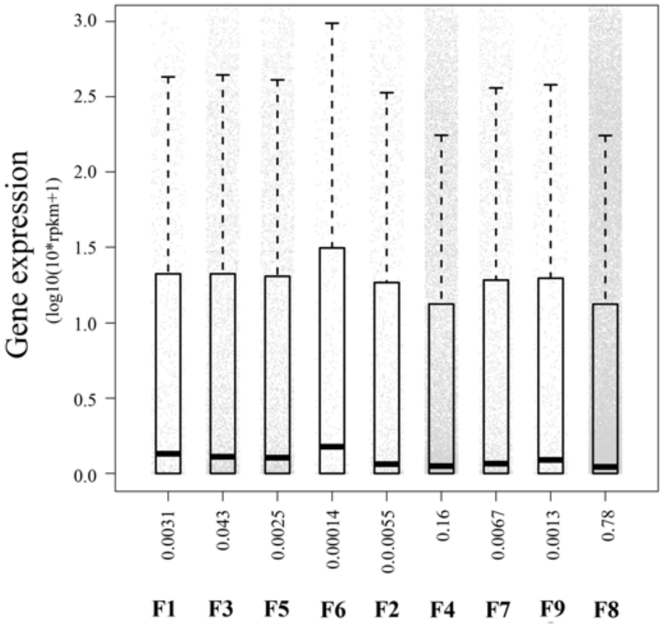


Figure S 5.7: **Transcriptional activity in drug-defined genomic states based on drug-induced histone eviction.** Gene expression in drug-defined genomic states (FAIRE-seq). Genes were mapped to states if their respective TSS (+/-2kb) overlapped a region called as a certain state.

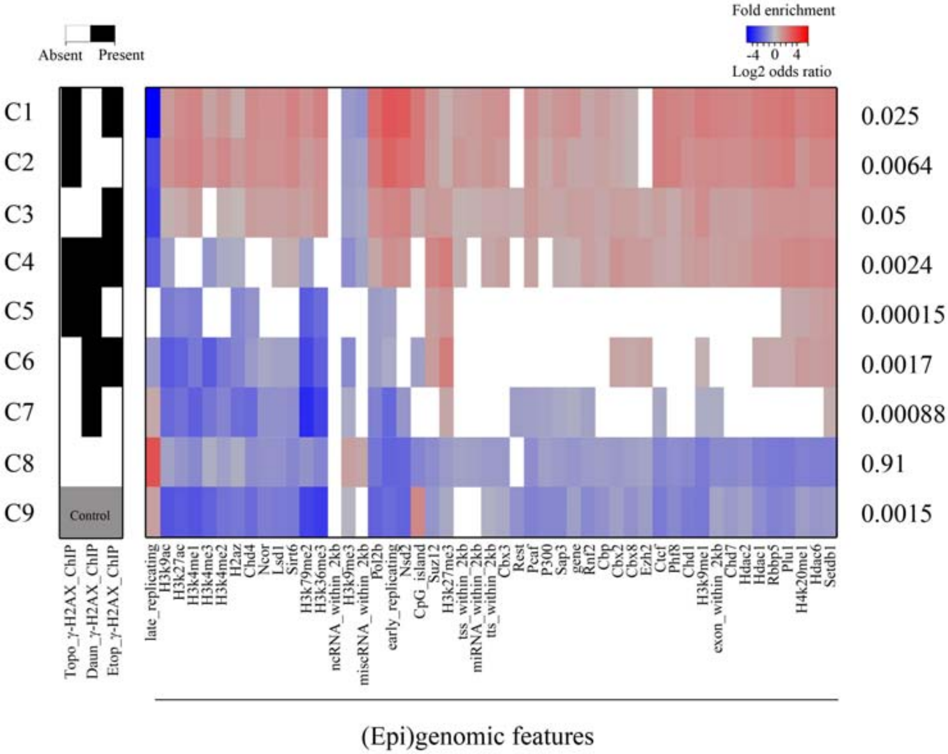


Figure S 5.8: **(Epi)genomic features targeted differently or similarly by distinct drugs for DNA damage response.** Segmentation of the genome into states based on DNA damage induced γ-H2AX following exposure of K562 cells to the different drugs. The black-and-white heatmap relates the states to drugs, with black and white indicating γ-H2AX and no γ-H2AX, respectively The blue-to-red heatmap shows the association of chromatin features with these states, ranging from blue (weak association) to red (strong association). The measure of association is the $\log_2$ odds ratio (see Methods for details), where for example a value of 2 for a certain feature and a certain state indicates that a twice as large fraction of genomic regions marked by that feature was covered by that state, compared to genomic regions not marked by that feature. Significance was determined using a permutation-based approach (p =1000; see Methods). Non-significant associations are depicted as white boxes. The labels on the y-axis represent the fraction of the genome covered by each state.

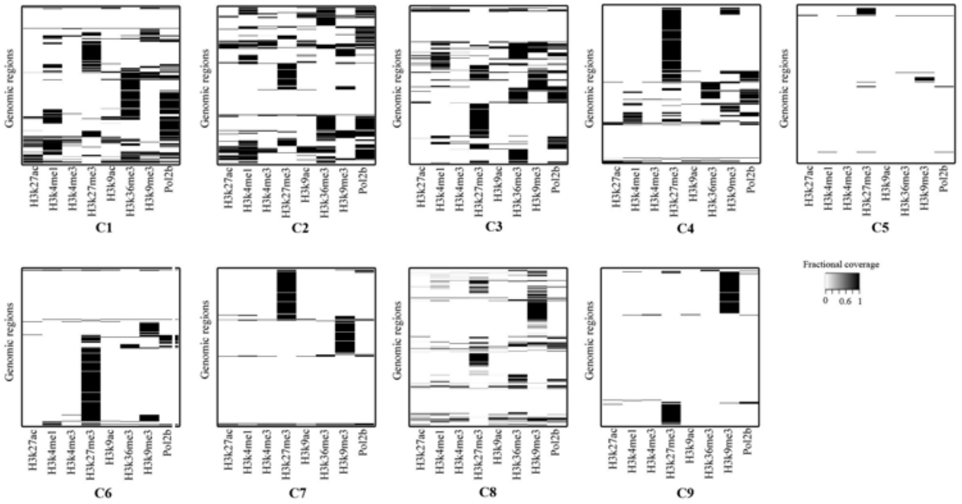


Figure S 5.9: **Representative of major histone modifications in drug-defined genomic states (γ-H2AX ChIP-seq).** Within each state, 10^4 genomic regions were randomly sampled (size-weighted), and visualized proportional to their size. For each region, the fraction of that region that was covered by a certain histone modification was visualized on a scale from white (no overlap) to black (complete coverage).

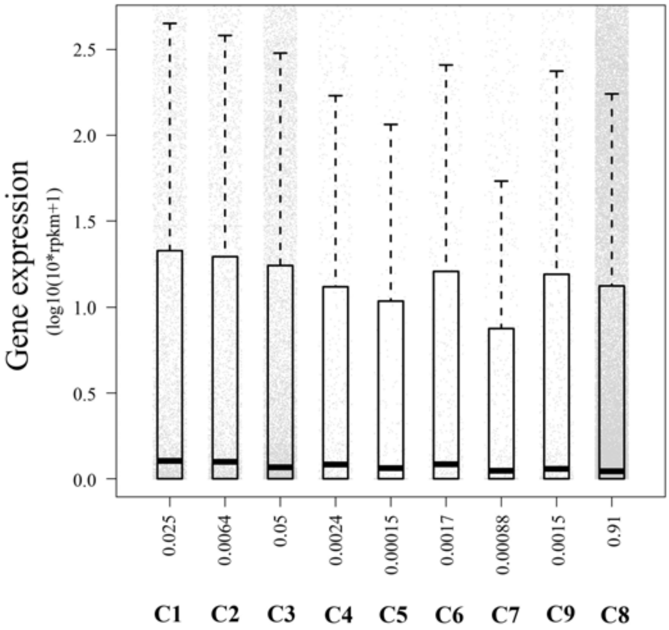


Figure S 5.10: **Transcriptional activity in drug-defined genomic states based on drug-induced DNA damage response.** Gene expression in drug-defined genomic states (γ-H2AX ChIP-seq). Genes were mapped to states if their respective TSS (+/-2kb) overlapped a region called as a certain state.

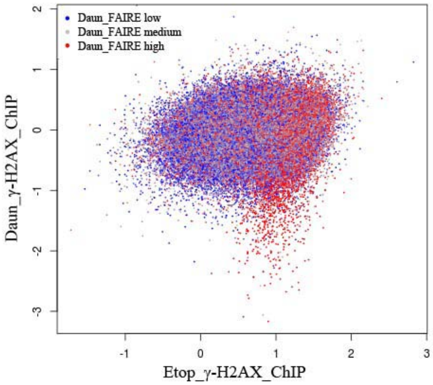


Figure S 5.11: **Attenuated γ-H2AX response correlated to drug-induced histone eviction in Daun-treated cells.** Comparison of Daun γ-H2AX ChP-seq, Etop γ-H2AX ChIP-seq, and Daun FAIRE-seq (normalized to untreated samples of the respective experiment procedure). This shows that there are genomic regions where the Etop-treated profile shows DNA damage signaling, but Daun-treated one does not, likely because of the high levels of histone eviction induced by Daun in these same genomic regions.

6

DISCUSSION

Parts of this chapter have been accepted for publication in Mobile Genetic Elements.

6.1. CHROMATIN PROFILING

In this dissertation, we studied DNA integrating elements and anti-cancer drugs for the purpose of perturbing chromatin. DNA integrating elements were used for inducing mutations and studying chromatin position effects, and as such for improving cancer gene discovery and gaining insights into gene regulation. Anti-cancer drugs were used for histone eviction and inducing DNA breaks, and as such for gaining insights into chromatin structure and opening up potential for new cancer treatment designs. In perturbing chromatin using these techniques, monitoring the response of chromatin to these perturbations can be considered chromatin profiling, as the response is dependent on physical properties of the chromatin.

6.1.1. INSERTIONAL MUTAGENESIS

Insertional mutagenesis (IM; Chapter 2) can be considered genome-wide profiling of oncogenic potential. While oncogenic potential can be seen as a genome-wide variable, eventually researchers are interested in specific genes involved in tumorigenesis, e.g. [1–3]. As was discussed in Chapter 2, a typical approach for identifying genes involved in tumorigenesis from IM screens is to find the signficant peaks (CISs) in an integration profile, and manually map these peaks to putative target genes, e.g. [4, 5]. Not only does this manual mapping introduce bias, also disregarding properties of individual integrations describing their tumorigenic selection bias, such as distance to genes, orientation relative to genes, expression of nearby genes and chromatin environment, can potentially lead to high numbers of false positive CISs. Chapter 2 made a first step in addressing these issues by presenting KC-RBM, a tool for automatically mapping integrations to putative target genes. An additional aggregation step allowed for retrieving putative cancer-related genes from IM screens by finding commonly targeted genes (CTGs). KC-RBM used orientation and distance relative to genes for mapping integrations, and furthermore gene expression was analyzed in conjunction with integration profiles to gain insight into the genes targeted by integrations.

In Chapter 3, the influence of integration bias on cancer gene discovery was demonstrated using paired unselected and selected integration datasets, which led to the conclusion that 7%-33% of CISs retrieved from IM screens could well be false positive. This is likely no different for CTGs, and is therefore also a potential problem for KC-RBM. In KC-RBM, integration bias could be addressed explicitly by drawing inspiration from Chapter 3. Due to integration bias, some genes will a priori have more integrations assigned to them, not strictly due to selective pressure. As was shown in Chapter 3, integration bias can be described in detail using a wide range of (epi)genomic features. Thus, a nonlinear classifier, such as a randomForest classifier, could be trained to distinguish (unselected) integration loci from random control loci. This classifier could estimate, for each base pair in the genome, the likelihood of it being a potential integration locus given the chromatin context. In essence, this would provide a genome-wide a priori integration probability distribution. By comparing the CTG counts to integration counts expected based on this a priori integration probability distribution, CTGs could be corrected for a priori integration bias.

6.1.2. INTEGRATION BIAS

Genome-wide integration profiling (Chapter 3) can be considered chromatin profiling of integration sensitivity. Integration sensitivity may provide insights into chromatin structure. In Chapter 3, the existence of a hierarchy in integration target site selection was inferred by virtue of comparing the integration profiles of different DNA integrating systems, in terms of a large amount of data describing the chromatin. It was shown that on large genomic scales, target site selection is similar across systems, whereas on small genomic scales, integration target site selection is system-specific. It is not unlikely that this hierarchy in target site selection represents a hierarchy in genome organization as well, as some genomic regions are inherently more accessible, not only for DNA integrating elements, but also for DNA binding proteins. This is illustrated by the observation that this large-scale division into two classes of genomic regions roughly corresponds to the distinction between heterochromatin and euchromatin, between LADs and iLADs, and between early and late replicating domains. These are chromatin features that are all strongly associated with genome organization and chromatin accessibility [6–8]. On a smaller scale, the integrating systems each recognize distinct types of chromatin, as such profiling the chromatin on two levels that could be distinguished by studying multiple DNA integrating elements in parallel.

In studying the integration targeting profiles among different systems, MMTV was found to be remarkably unbiased, compared to SB and PB. One possible explanation could be given by cell phase dependent integration. For another species of retrovirus, MuLV, it has been shown that cells must pass through mitosis in order for MuLV to be able to enter the nucleus [9]. If this requirement would also hold for MMTV, and additionally, for some reason, MMTV would only be active during this same phase, it would find chromatin in a highly compacted state. In this highly compacted state, perhaps most genomic regions equally (in)accessible to integration, possibly with the exception of some "fundamental regulatory and structural building blocks of chromosomes" [10] such as the genomic loci that define TADs by virtue of their strong 3D interactions in only one direction along a chromosome [11]. Indeed, in Chapter 3, in contrast to other (epi)genomic features, these TAD boundary interfaces were shown to be hotspots of MMTV integration.

6.1.3. CHROMATIN POSITION EFFECTS

TRIP (Chapter 4) provides a way of profiling transcriptional permissiveness. TRIP can be adapted to study many other processes that are dependent on local chromatin state. A number of possibilities were outlined in the Discussion of Chapter 4. Given the recent advances and growing interest in CRISPR-based genome editing [12], an interesting additional possibility is to assess the efficiency of the CRISPR/Cas9 system in different chromatin states. It has been suggested that this efficiency is dependent on the local chromatin state [13, 14]. Using TRIP, the efficiency could be studied at the exact same site in different chromatin states, which is not possible in an endogenous setting.

TRIP can be considered genome-wide in the sense that it can generate expression values for thousands of randomly integrated reporter genes, without pre-imposed restrictions to integrating into certain genomic regions. However, as was shown in Chapter 3, integration of reporter genes into the genome does come with substantial biases.

One may wonder what the importance of these integration biases is for computing associations of reporter gene expression with any (epi)genomic features. For example, in the TRIP study [15] we computed the association of reporter gene expression with a number of binarized (epi)genomic features, such as lamina-associated domains (LADs) [7, 16]. It is known that there is an integration bias against LADs, i.e. there are relatively few integrations within LADs [17]. To demonstrate the influence of this bias on the association of reporter gene expression with LADs we ran a simple simulation. We randomly generated 10^4 integrations *in silico*, and distributed these in an increasingly uneven fashion across two classes, e.g. LADs and inter-LADs [7, 16], from completely even (i.e. 5000 in one class and 5000 in the other) to highly uneven (i.e. 9998 in one class and 2 in the other). For each integration, depending on the class of an integration, we simulated expression values by sampling from a certain class-specific expression distribution, i.e. a normal distribution with mean 0.1 and standard deviation 1 for Class 1, and a normal distribution with mean 0 and standard deviation 1 for Class 2. Then, for each distribution, we performed Welch's t-test to distinguish between the two classes. The results of the simulation are shown in Figure 6.1. It shows two measures, 1) the statistical significance of the t-test, expressed as a z-normalized t-statistic, and 2) the effect size, expressed as the difference in mean reporter gene expression between the two classes. As could be expected, it clearly illustrates that with an increasingly uneven distribution, the expected effect size remains the same. However, the variance in the effect size increases, and with it the statistical significance reduces. In other words, given a certain distribution of a number of integrations across two classes, a more asymmetric distribution will require a larger total number of integrations to detect a certain effect size as statistically significant.

Not for all questions it is equally straightforward to determine the (lack of) influence of integration bias. In these cases, it may be needed to regularize the genome-wide integration profile. We provided one such example when inferring PGK domains reflecting genome-wide domains of transcriptional permissiveness, using a hidden Markov model (HMM) [15]. Since by inferring an HMM, equidistant spacing of integrations on the genome was assumed, we asked to what extent integration bias affected the eventual domain calling. For this purpose, a non-homogeneous HMM was additionally inferred, with the HMM transition probabilities depending on the distance between IRs [18]. The domains inferred using both approaches were highly similar.

In conclusion, while in the case of interpreting TRIP results it should always be kept in mind that integration is random but biased, the impact of these biases on results seems often limited. However, an important drawback of integration bias, is that it reduces statistical power.

6.1.4. HISTONE EVICTION AND DNA DAMAGE INDUCTION

On a large scale, the three types of profiles mentioned above, oncogenic potential, integration sensitivity and transcriptional permissiveness are all strongly associated. For example, all profiles somehow relate to chromatin compaction, a basic characterization of chromatin structure. In Chapter 5, we used anti-cancer drugs to directly manipulate this structure, for example by drug-induced eviction of histones from chromatin. Monitoring the response of chromatin to these drugs can be considered profiling of drug-dependent

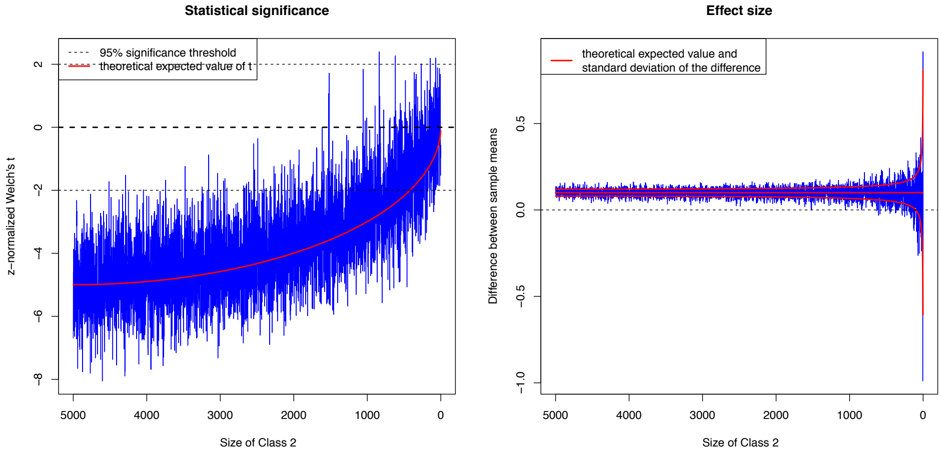


Figure 6.1: 5×10^3 integrations were generated in silico, and distributed across two classes (Class 1 and Class 2) in an increasingly uneven fashion. Depending on the assigned class, reporter gene expression was simulated by drawing from a class-specific distribution. Then, (A) the significance of the difference between the two classes was determined by Welch's t-test as a function of the size of Class 2 (dashed red line: two-sided 5% significance threahold; black solid line: theoretical expected value of the z-normalized t-statistic), and (B) the effect size was determined as the difference in means between the two classes as a function of the size of Class 2 (black solid lines: theoretical expected value and standard deviation of the sample distribution of the difference). The x-axes represents the size of Class 2 which indicates how uneven the distribution across the two classes is.

nucleosome integrity.

For proper interpretation of the profiles in Chapter 5, an appropriate technique for normalizing the profiles with respect to input DNA was essential. The problem of normalizing genome-wide sequencing data is as yet unsolved in the general case. Many approaches exist, which generally work well only in a limited set of cases. There are nonlinear approaches, for example based on LOWESS regression [19] and quantile normalization [20], which assume that genome-wide distributions of tag counts are comparable between the two samples. Another nonlinear method focuses only on peaks, and normalizes based on the relative intensities of peaks shared between the two samples [21], assuming that similar binding mechanisms underlie shared peaks. In our case of normalizing diffuse, genome-wide, profiles against more peaky control profiles, these nonlinear approaches were not suitable. In addition to nonlinear approaches, there are linear approaches for normalizing genome-wide sequencing data. As a justification, it has been shown that, in its background regions, a ChIP-seq profile shows linear correlation with its control [22, 23]. A typical linear approach estimates the background regions in the ChIP-seq signal and scales the control by a constant, the ratio of total signal read count and total control read count in these background regions [22, 24]. On our data, these approaches performed relatively poorly. The reason for this is likely twofold. First, for our data it could not be assumed that the vast majority of the genome represents background. Second, both approaches try to estimate the complete set of background regions, by identifying a break point in a certain representation of the genome-wide profile. For more diffuse, domain-oriented profiles these break points can be difficult to unambiguously detect. Therefore, we implemented a custom-designed autocorrelation-based approach for normalizing our data. Our approach bears some similarity to the NCIS method [22]. However, it does not assume that the large majority of the genome can be considered background, and furthermore exploits the spatial dependence between adjacent genomic bins. Furthermore, contrary to NCIS, it does not try to estimate the complete set of background regions. Instead, it finds a limited set of genomic regions, a subset of the complete set of background regions, for which there is sufficient statistical evidence (see Chapter 5: Methods) to call these regions background. As such, this algorithm fills a gap in the current repertoire of normalization techniques.

6.2. ANALYSIS OF GENOME-WIDE SEQUENCING DATA

Regarding the results presented in this dissertation, it is important not to forget that conclusions drawn are often strongly dependent on the quality and processing of the data. In studying biology through sequencing technology, there can be many issues that complicate the interpretation of results, especially when analyzing large numbers of datasets, potentially from different labs, simultaneously. In the analyses that we performed as part of this dissertation, we observed many such issues, some of which have received relatively little attention in the literature to date. This section will present a number of these issues, and the impact these can have on the results.

6.2.1. Normalization: the pseudocount

As was explained in Chapter 5, a commonly used approach for normalizing a binned genome-wide sequencing profile with a corresponding input DNA profile, is the following:

$$x_i^{\text{norm}} = \log_2 \frac{x_i + 1}{c y_i + 1} \tag{6.1}$$

Here, x_i^{norm} is the normalized signal in bin i, x_i represents the number of signal reads in bin i, y_i represents the number of control (whole-cell extract) reads in bin i. c is the normalization constant used to make signal and control quantitatively comparable. Often, the ratio of sequencing depths is used, but in Chapter 5, an alternative autocorrelation-based method was presented. To account for bins with no reads, a pseudocount of 1 is typically added, as such avoiding division by zero. However, adding a pseudocount can give surprising, and questionable, results. As an example, consider the hypothetical ChIP-seq signal and ChIP control in Figure 6.2(A). On each position, the control has exactly twice as many reads as the signal. Therefore, calculating an ordinary $\log_2$ ratio without pseudocount results in a flat profile. However, adding a pseudocount of 1 results in a profile that is negatively correlated with both signal and control. While this may seem a contrived example, in our analyses for Chapter 4 we have observed exactly this on a genome-wide scale. After normalization of a Nanog binned coverage with input DNA, the resulting normalized signal was negatively correlated with both the original Nanog binned coverage and the input DNA (Figure 6.2(B)). Additionally, the signal negatively correlated with the PGK IRs (Chapter 4). This is unexpected given the demonstrated role of Nanog as a transcriptional activator [25–27], and further suggests that the normalized profile may not be representative of the actual Nanog binding profile. The underlying problem may be a combination of pseudocount, sequencing depth and binsize: Combined, the replicates constituted 17802070 reads, which is below the ENCODE standard of 20×10^6 [28]. Note that in case of low sequencing depth and fixed, relatively small, bin size, the contribution of each pseudocount to the normalized signal is relatively large. As such, this example emphasizes the need to find an optimal compromise between the resolution of an analysis, i.e. the bin size, and the relative contribution of the pseudocount. The nature of this compromise is dependent on sequencing depth, and on the distribution of the available reads across the binned coverage.

6.2.2. Calling peaks on ChIP-seq data

The fold-enrichment threshold

For the analyses of the TRIP datasets, summarized in Chapter 4, one topic we were interested in was the genome-wide associations of IR expression status with a wide range of chromatin marks, as profiled using ChIP-seq or one of its variants. A typical final step in processing ChIP-seq data is to determine regions of significantly enriched signal, i.e. to call peaks. In addition to setting a significance threshold, a fold-enrichment threshold is commonly applied to select from all significantly enriched regions those regions that are considered substantially enriched relative to a control. The choice of fold-enrichment threshold can sometimes have a surprising impact on the results of an analysis. Figure

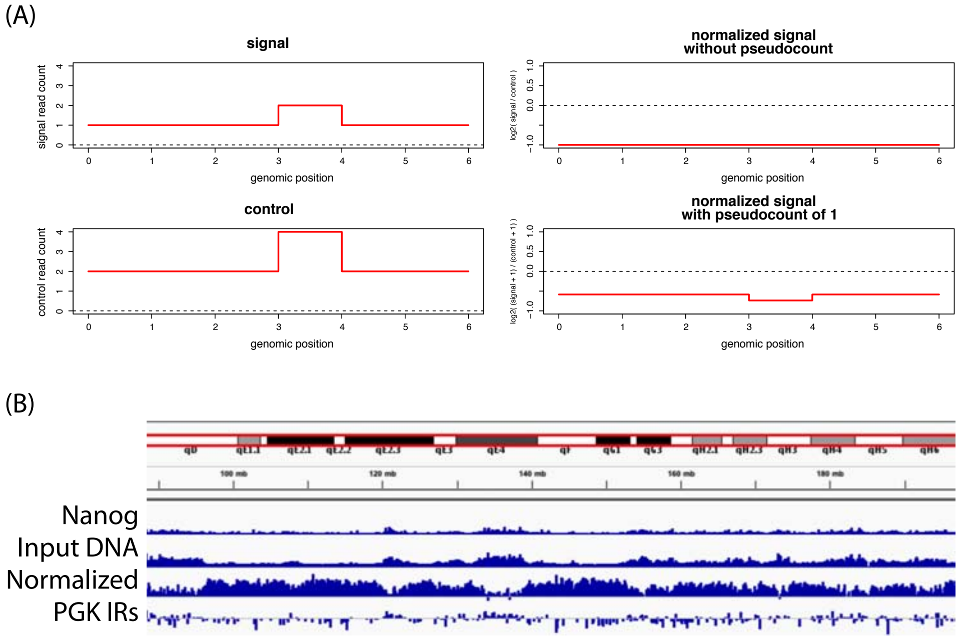


Figure 6.2: The influence of the pseudocount on normalizing a ChIP-seq signal with a control. (A) Normalization of an example signal with an example control, with and without pseudocount. (B) Normalization of a Nanog ChIP-seq binned coverage with input DNA.

6.3 shows the results of applying 5 different fold-enrichment thresholds on a set of significant Kap1 ChIP-seq peaks. For each threshold, the Spearman correlation was computed of IR expression status with negative distance of an IR to the nearest peak. Strikingly, it can be seen that, depending on the fold-enrichment threshold, the correlation can vary from significantly positive (taking all peaks) to significantly negative (taking only top 20% of peaks). Given that Kap1 is mostly associated with transcriptional repression [29] and that a background signal such as input DNA will generally positively correlate with transcriptional activation, a possible explanation for this change of sign is that many of the weaker peaks still represent background signal, even though peaks were called relative to a control sample. The underlying cause for this could be the sequencing depth scaling of signal and control as performed internally by the MACS peak calling algorithm [30], which could be addressed by the normalization approach proposed in Chapter 5. Alternatively, it may be caused by an expression bias inherent to the ChIP procedure which can be strong enough for transcriptional repressors to appear as activators [31, 32].

Although the correlations are not very strong, in conventional statistics these results would be considered highly significant. As such, these results stress the importance of a strong focus on effect size, rather than statistical significance [33], which in the case of Figure 6.3 is relatively small.

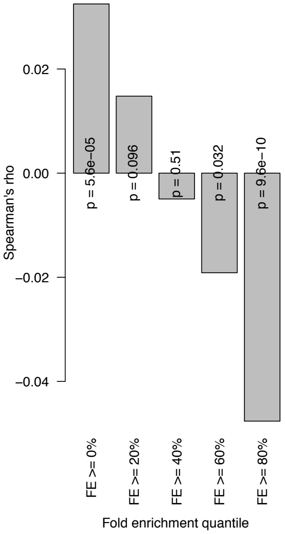


Figure 6.3: Spearman correlation of the PGK IR expression level with a set of Nanog ChIP-seq peaks, as a function of fold-enrichment treshold on the Nanog ChIP-seq peaks.

COMPUTING GENOME-WIDE ASSOCIATIONS USING PEAKS

After calling peaks on ChIP-seq profiles, one may be interested in computing genome-wide associations, e.g. the Spearman correlation, between different (epi)genomic features, such as between Kap1 binding sites and IR expression status above. For computing correlations between (epi)genomic features and IR expression, first each IR has to be assigned a score for the particular feature of interest. There are many ways in which such a score can be computed. In Figure 6.4(A), the results of seven of these are given, after calling peaks on genome-wide sequencing profiles, and applying a typical threshold of 1.5 fold-enrichment over the background:

1. presence_vs_absence: 1 if a peak is found within 1kb of the IR, otherwise 0.
2. coverage: The fraction of base pairs that is covered by the feature in a window of 1kb on either side of the IR.
3. peak_weighted_coverage: As above, and additionally multiply the contribution of each individual peak within the window with its peak height.
4. max_peak: The height of the highest peak within a 1kb window on either side of the IR.
5. gaussian_weighted_coverage: The surface area of a Gaussian ($\sigma = 1000$bp) centered at the IR, restricted to those genomic regions that are covered by peaks.
6. peak_and_gaussian_weighted_coverage: As above, and additionally multiply the contribution of each individual peak within the window with its peak height.
7. peak_proximity: Negative distance towards the nearest peak.

Some observations can be made from Figure 6.4(A). First, it does not really seem to matter which method for scoring IRs is used, since the correlations are very comparable across different methods. However, generally, the strongest associations are found for approaches that do not binarize distance by taking a fixed window around the IR (methods 5, 6 and 7). Second, in addition to using the genomic region that is covered by a peak, taking into account peak height hardly leads to any differences (methods 3 and 4, as well as 5 and 6), which suggests that at least a substantial part of the variation in peak heights is due to noise. Third, the genome-wide correlations are almost exclusively positive, which is unexpected given that Figure 6.4(A) also includes many repression-associated features. In this light, it is interesting to compare the peak-based correlations with an approach based on a genome-wide binned coverage approach (Figure 6.4(B)). In this approach, the average of a genome-wide binned coverage (normalized to input DNA) within a window of 1kb on either side of the IR is computed, and used to estimate the correlation between the IR expression status and the feature of interest. It can be directly seen that more features are negatively correlated with IR expression status than when adopting a peak-based approach. In fact, the features that are negatively correlated with IR expression are mostly features associated with repression. One possible explanation is the usage of a pseudocount in computing the normalized profile. However, the changes of sign mostly conform to what can be expected based on the literature. Therefore, a more likely explanation is that the local strength of a particular signal is just a better measure than a measure that is based on a significancly enriched region (i.e. peak) at an arbitrary distance. Another explanation, as was similarly suggested in the case of Figure 6.3, is that many significant peaks still represent background signal, even when calling peaks using a control such as input DNA. One of the underlying causes may be that many peak calling algorithms, such as MACS and SICER, scale the signal and control proportional to the ratio of their sequencing depths, which may not necessarily be appropriate. To address this issue, an approach such as the one presented in Chapter 5 for computing a normalization factor between two sequencing datasets, could be implemented in these tools. This example demonstrates that the decision whether to call peaks on the ChIP-seq profile is not straightforward.

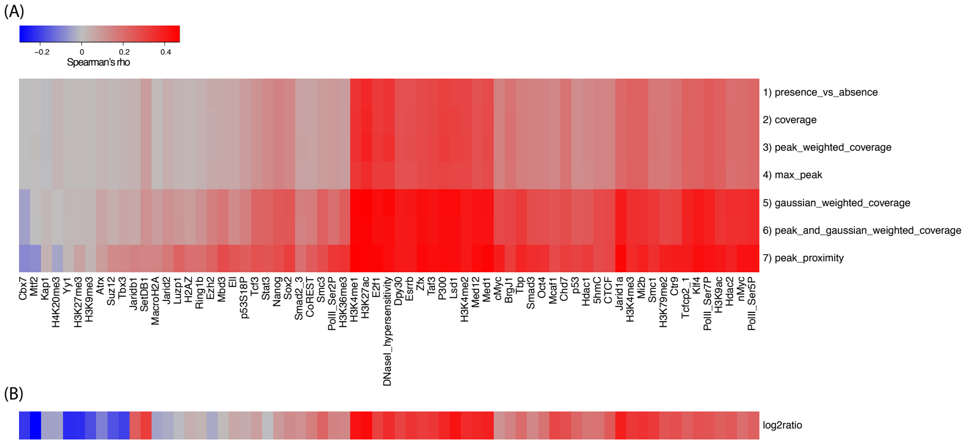
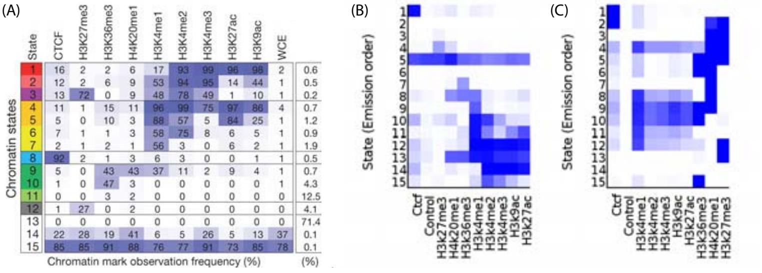


Figure 6.4: (A) Using different methods to compute the Spearman correlation between the PGK IR expression levels and sets of ChIP-seq peaks. (B) Using a genome-wide binned coverage approach (normalized to input DNA) to compute the Spearman correlation between the PGK IR expression levels and set of ChIP-seq datasets.

The impact of peak caller choice on the analysis

From the above it seems that, once peaks are called on ChIP-seq data, the eventual approach for computing genome-wide associations is of secondary importance. However, the initial choice of peak caller can be of crucial importance to the outcome of analyses, as can be demonstrated by segmenting the genome using the ChromHMM software [34]. The input to ChromHMM is a series of binarized genome-wide profiles of features of interest. After providing a resolution, i.e. specifying consecutive bins of e.g. 200bp, ChrommHMM models the observed combinations of features as a product of independent Bernoulli random variables, and learns a multivariate hidden Markov model (HMM) [34] that identifies a given number of states. For each state, the HMM specifies 1) the emission probability distribution: the probability of observing a certain combination of features in that state, and 2) the transition probability distribution: the probability of that state being adjacent to a certain other state. The emission probability distribution represents the chromatin composition of each state, and in combination with the transition probability distribution it is used to segment the genome by assigning each bin to the most likely state. In the original application of this approach [35], ChIP-seq profiles of ten features in nine human cell types from the ENCODE project were binarized assuming a Poisson background distribution on the number of reads in a sliding window of fixed size. For each feature, the resulting nine binarized profiles were concatenated, and an HMM was inferred from the concatenated profiles [35] (Figure 6.5(A)). These results can be fairly accurately reproduced when restricting the analysis to only one cell type, the K562 cell type, and using the same Poisson background distribution approach for the initial calling of peaks (Figure 6.5(B)): All states but States 4, 13 and 15 can be said to have close relatives in Figure 6.5(A). However, these results change dramatically when using the same K562 ENCODE data, but now binarized as provided on the ENCODE website, using Scripture [36] (Figure 6.5(C)): In this case, it is safe to say that less than half of the states (States 1, 6, 11, 13, 14, 15) have analogues in Figure 6.5(A). It is interesting to note that Scripture also uses a Poisson background for detecting significant enrichment. However, the discrepancy between Figures 6.5(A) and 6.5(B) and Figure 6.5(C) is likely caused by the fact that Scripture uses a wide range of window sizes to slide across the genome, in order to also detect more diffuse enrichment, whereas the approach as applied in the original paper mainly finds focal points of enrichment [35]. Furthermore, in addition to using a wide range of window sizes, substantial post-processing of the called peaks was performed, as reported on the ENCODE website. This is reflected by the observation that in Figure 6.5(C), only State 14 can be considered devoid of any chromatin marks, and accounts for roughly 55% of the genome. However, in Figure 6.5(B), these percentages are roughly 86% (States 2, 3, 6) or 70% (States 2, 6). Finally, in Figure 6.5(A), it is even higher: Roughly 88% (States 10, 11, 12) or 84% (States 10, 12) of the genome is almost entirely devoid of any features, and as such, the analyses presented [35] are mostly based on 12%–16% of the genome.

This example demonstrates the importance of thoroughly considering what approach to use for determining enrichment. However, it should also be noted that, to a certain extent, choosing a peak caller can be a self-fulfilling prophecy. If a signal is expected to show mainly focal enrichment, then applying a peak caller suitable for detecting focal enrichment will mainly detect focal enrichment. The same principle applies to detecting more diffuse enrichment. For this reason, approaches for detecting enrichment simultaneously across multiple scales, such as Scripture and a more recent approach for multiscale segmentation of genomic signals [37], are crucial for making peak calling from ChIP-seq data more robust and unbiased.



State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	(%)
1	16	2	2	6	17	93	99	96	98	2	0.6
2	12	2	6	9	53	94	95	14	44	1	0.5
3	13	72	0	9	48	78	49	1	10	1	0.2
4	11	1	15	11	96	99	75	97	86	4	0.7
5	5	0	10	3	88	57	5	84	25	1	1.2
6	7	1	1	3	58	75	8	6	5	1	0.9
7	2	1	2	1	56	3	0	6	2	1	1.9
8	92	2	1	3	6	3	0	0	1	1	0.5
9	5	0	43	43	37	11	2	9	4	1	0.7
10	1	0	47	3	0	0	0	0	0	1	4.3
11	0	0	3	2	0	0	0	0	0	0	12.5
12	1	27	0	2	0	0	0	0	0	0	4.1
13	0	0	0	0	0	0	0	0	0	0	71.4
14	22	28	19	41	6	5	26	5	13	37	0.1
15	85	85	91	88	76	77	91	73	85	78	0.1

Figure 6.5: Using ChromHMM to learn HMMs from ENCODE data [34]. (A) Emission probabilities resulting from learning an HMM on 10 chromatin features across nine cell types [35], using the peak calling specified by ChromHMM. (B) Emission probabilities resulting from learning an HMM on 10 chromatin features for only one cell type, the K562 cell type, using the peak calling specified by ChromHMM. (C) Emission probabilities resulting from learning an HMM on 10 chromatin features for only one cell type, the K562 cell type, using the set of ENCODE peaks.

6.2.3. Variability between data from different labs

As was shown above, processing a sequencing dataset in different ways can occasionally lead to apparently inconsistent results. However, even a comparison between datasets representing the same epigenomic feature, and processed in exactly the same way, can present with inconsistent results. Whereas the demonstrated substantial between-lab variability of ChIP-chip [38] is a good indication that substantial variability also exists for ChIP-seq, no such comprehensive study for ChIP-seq has been published to date.

Figure 6.6(A) shows three epigenomic features associated with the Polycomb repressive complex number 2 (PRC2). PRC2, represented by Ezh2 and Suz12, is involved in the trimethylation of lysine 27 on histone H3 (H3K27me3) through the presence of Ezh2 [39]. For each of these features, the Spearman correlation with PGK IR expression (Chapter 4) was computed, using the genome-wide binned coverage approach as it was also used for Figure 6.4(B). Note that in the current situation, the 0.01 significance level is to be found at a correlation coefficient of about 0.019, rendering all correlation coefficients highly significant according to that commonly used standard. It is interesting to see that although the two H3K27me3 datasets and the two Suz12 datasets were processed in exactly the same way, the results are contradicting in terms of correlation.

Another good indication of the variability between different datasets can be given by comparing different mESC input DNA datasets by correlating their genome-wide binned coverage profiles (Figure 6.6(B)). It is immediately evident that there is a high degree of variability between the different datasets. The dataset most dissimilar to the others is GSM706675, which can be expected since this is a 5hmC control. Contrary to the other controls, there is no cross linking step in generating the 5hmC control. The strong dissimilarity between this control and the others suggests that cross linking induces a very strong bias. In other words, for any arbitrary ChIP-seq profile, a large portion of the variation in the profile can likely be explained by the cross linking step. While the remaining controls were all generated in comparable ways (i.e. cross linking, fragmenting, reverse cross linking, sequencing) there is still substantial variation across the datasets. In part, this may be explained by the specific strain of mESC used, indicated between brackets in Figure 6.6(B). This is reflected by the single cluster of four V6.5 (C57BL/6-129) datasets correlating relatively strongly. However, the same strain can also be found mixed with other strains such as E14 in other clusters. Considering that the differences between the input DNA datasets are fairly large, and that the differences partly associate with the specific strain of mESC that was used, this suggests that part of the differences in correlations between different features observed in Figure 6.4 can be attributed to strain specificities as well.

The two examples presented in this section demonstrate that between-lab variability can be substantial to the point of leading to contradicting conclusions. This makes a strong case for setting, and adhering to, clear standards and quality controls, such as initiated by the ENCODE consortium [28].

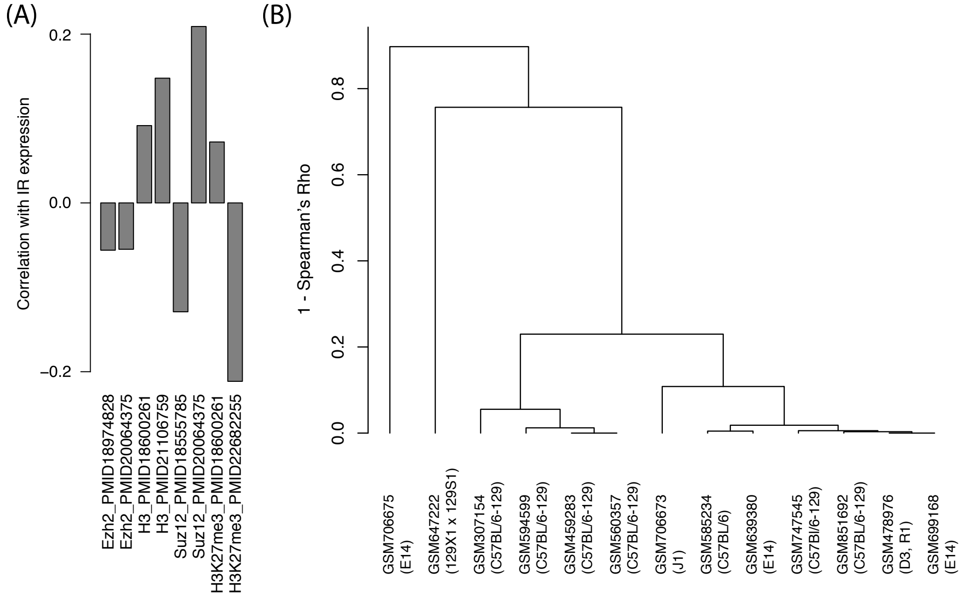


Figure 6.6: Variability between data from different labs. (A) Spearman correlation of PGK IR expression levels with genome-wide binned coverage profiles (normalized to input DNA) for four different features, each represented by two different labs. (B) Spearman correlation based clustering of a range of input DNA datasets, after computing their genome-wide binned coverage profiles.

6.3. FUTURE PERSPECTIVES

The computational examples discussed in this chapter show that specific decisions anywhere in the data analysis pipeline can have major implications for research outcomes. Consequently, for the quality of the research results, it is detrimental to carefully consider the available options, determine whether suitable approaches exist, and if needed develop tailor-made computational solutions. Specifically, with the increase in availability and variety of data, there has also been an increasing need for tools that integrate different datasets, either for direct comparison by normalization or for more complex modeling to study biology at a systems level. This integration is not trivial, and often it is difficult to develop generic solutions, i.e. solutions applicable to more than just the specific problem at hand. Future applications will no doubt become computationally even more challenging, with Hi-C data of increasing resolution to study the 3D architecture of the genome in more detail, time-resolved epigenomics datasets to study chromatin dynamics, etc.

The work presented in this dissertation dealt substantially with data integration, and provided many tailor-made solutions. However, some are also more generally applicable. For example, the feature ranking method (Chapter 3) can be used for many types of data. Similarly, the autocorrelation-based normalization algorithm (Chapter 5) is not limited to FAIRE-seq data, but can be applied to a wide range of genome-wide sequencing data. However, the last example on variability between data from different labs has also shown that not all problems can be solved using computational tools. Therefore, elaborate standardization, such as has been initiated by the ENCODE consortium [28], is essential.

REFERENCES

[1] A. G. Uren, J. Kool, K. Matentzoglu, J. de Ridder, J. Mattison, M. van Uitert, W. Lagcher, D. Sie, E. Tanger, T. Cox, M. Reinders, T. J. Hubbard, J. Rogers, J. Jonkers, L. Wessels, D. J. Adams, M. van Lohuizen, and A. Berns, *Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks.* Cell **133**, 727 (2008).

[2] J. Kool, A. G. Uren, C. P. Martins, D. Sie, J. de Ridder, G. Turner, M. van Uitert, K. Matentzoglu, W. Lagcher, P. Krimpenfort, J. Gadiot, C. Pritchard, J. Lenz, A. H. Lund, J. Jonkers, J. Rogers, D. J. Adams, L. Wessels, A. Berns, and M. van Lohuizen, *Insertional mutagenesis in mice deficient for p15Ink4b, p16Ink4a, p21Cip1, and p27Kip1 reveals cancer gene interactions and correlations with tumor phenotypes,* Cancer Res **70**, 520 (2010).

[3] J. Mattison, J. Kool, A. G. Uren, J. de Ridder, L. Wessels, J. Jonkers, G. R. Bignell, A. Butler, A. G. Rust, M. Brosch, C. H. Wilson, L. van der Weyden, D. A. Largaespada, M. R. Stratton, P. A. Futreal, M. van Lohuizen, A. Berns, L. S. Collier, T. Hubbard, and D. J. Adams, *Novel candidate cancer genes identified by a large-scale cross-species comparative oncogenomics approach.* Cancer Res **70**, 883 (2010).

[4] H. Mikkers, J. Allen, P. Knipscheer, L. Romeijn, A. Hart, E. Vink, A. Berns, and

L. Romeyn, *High-throughput retroviral tagging to identify components of specific signaling pathways in cancer.* Nat Genet **32**, 153 (2002).

[5] J. de Ridder, A. Uren, J. Kool, M. Reinders, and L. Wessels, *Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens.* PLoS Comput Biol **2** (2006).

[6] S. I. Grewal and S. Jia, *Heterochromatin revisited.* Nature reviews. Genetics **8**, 35 (2007).

[7] D. Peric-Hupkes, W. Meuleman, L. Pagie, S. W. M. Bruggeman, I. Solovei, W. Brugman, S. Gräf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, M. Reinders, L. Wessels, and B. van Steensel, *Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.* Mol Cell **38**, 603 (2010).

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

[9] P. F. Lewis and M. Emerman, *Passage through mitosis is required for oncoretroviruses but not for the human immunodeficiency virus.* J Virol **68**, 510 (1994).

[10] G. J. H and D. Job, *Connecting the genome: dynamics and stochasticity in a new hierarchy for chromosome conformation,* Molecular Cell **49**, 773 (2013).

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

[12] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang, *Multiplex genome engineering using CRISPR/cas systems,* Science **339**, 819 (2013).

[13] M. H. Larson, L. A. Gilbert, X. Wang, W. A. Lim, J. S. Weissman, and L. S. Qi, *Crispr interference (crispri) for sequence-specific control of gene expression.* Nat Protoc **8**, 2180 (2013).

[14] L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin, and W. A. Lim, *Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression.* Cell **152**, 1173 (2013).

[15] W. Akhtar, J. de Jong, A. V. Pindyurin, L. Pagie, W. Meuleman, J. de Ridder, A. Berns, L. F. A. Wessels, M. van Lohuizen, and B. van Steensel, *Chromatin position effects assayed by thousands of reporters integrated in parallel.* Cell **154**, 914 (2013).

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

[17] J. de Jong, W. Akhtar, J. Badhai, A. G. Rust, R. Rad, J. Hilkens, A. Berns, M. van Lohuizen, L. F. A. Wessels, and J. de Ridder, *Chromatin landscapes of retroviral and transposon integration profiles,* PLoS Genet **10**, e1004250+ (2014).

[18] J. C. Marioni, N. P. Thorne, and S. Tavare, *Biohmm: a heterogeneous hidden markov model for segmenting array cgh data.* Bioinformatics **22**, 1144 (2006).

[19] C. Taslim, J. Wu, P. Yan, G. Singer, J. Parvin, T. Huang, S. Lin, and K. Huang, *Comparative study on ChIP-seq data: normalization and binding pattern characterization,* Bioinformatics **25**, 2334 (2009).

[20] N. U. Nair, A. D. Sahu, P. Bucher, and B. M. E. Moret, *Chipnorm: a statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries.* PLoS One **7**, e39573 (2012).

[21] Z. Shao, Y. Zhang, G.-C. C. Yuan, S. H. Orkin, and D. J. Waxman, *MAnorm: a robust model for quantitative comparison of ChIP-seq data sets.* Genome biology **13**, R16+ (2012).

[22] K. Liang and S. Keles, *Normalization of ChIP-seq data with control,* BMC Bioinformatics **13**, 199+ (2012).

[23] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein, *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls,* Nature Biotechnology **27**, 66 (2009).

[24] A. Diaz, K. Park, D. A. Lim, and J. S. Song, *Normalization, bias correction, and peak calling for ChIP-seq.* Statistical applications in genetics and molecular biology **11** (2012).

[25] G. Pan and D. Pei, *The stem cell pluripotency factor nanog activates transcription with two unusually potent subdomains at its c terminus.* J Biol Chem **280**, 1401 (2005).

[26] Y.-H. Loh, Q. Wu, J.-L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K.-Y. Wong, K. W. Sung, C. W. H. Lee, X.-D. Zhao, K.-P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C.-L. Wei, Y. Ruan, B. Lim, and H.-H. Ng, *The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells.* Nat Genet **38**, 431 (2006).

[27] L. Zhang, Y.-B. Luo, G. Bou, Q.-R. Kong, Y.-J. Huan, J. Zhu, J.-Y. Wang, H. Li, F. Wang, Y.-Q. Shi, Y.-C. Wei, and Z.-H. Liu, *Overexpression nanog activates pluripotent genes in porcine fetal fibroblasts and nuclear transfer embryos.* Anat Rec (Hoboken) (2011).

[28] S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu,

L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder, *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia,* Genome Research **22**, 1813 (2012).

[29] S. Iyengar and P. J. Farnham, *KAP1 protein: An enigmatic master regulator of the genome,* Journal of Biological Chemistry **286**, 26267 (2011).

[30] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nusbaum, R. Myers, M. Brown, W. Li, and X. S. Liu, *Model-based analysis of ChIP-seq (MACS),* Genome Biol **9**, R137+ (2008).

[31] L. Teytelman, D. M. Thurtle, J. Rine, and A. van Oudenaarden, *Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins,* Proceedings of the National Academy of Sciences **110**, 18602 (2013).

[32] D. Park, Y. Lee, G. Bhupindersingh, and V. R. Iyer, *Widespread misinterpretable ChIP-seq bias in yeast,* PLoS ONE **8**, e83506+ (2013).

[33] R. Nuzzo, *Scientific method: Statistical errors,* Nature **506**, 150 (2014).

[34] J. Ernst and M. Kellis, *Chromhmm: automating chromatin-state discovery and characterization.* Nat Methods **9**, 215 (2012).

[35] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein, *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature **473**, 43 (2011).

[36] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev, *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nature biotechnology **28**, 503 (2010).

[37] T. A. Knijnenburg, S. A. Ramsey, B. P. Berman, K. A. Kennedy, A. F. A. Smit, L. F. A. Wessels, P. W. Laird, A. Aderem, and I. Shmulevich, *Multiscale representation of genomic signals,* Nat Meth **11**, 689 (2014).

[38] D. S. Johnson, W. Li, D. B. Gordon, A. Bhattacharjee, B. Curry, J. Ghosh, L. Brizuela, J. S. Carroll, M. Brown, P. Flicek, C. M. Koch, I. Dunham, M. Bieda, X. Xu, P. J. Farnham, P. Kapranov, D. A. Nix, T. R. Gingeras, X. Zhang, H. Holster, N. Jiang, R. D. Green, J. S. Song, S. A. McCuine, E. Anton, L. Nguyen, N. D. Trinklein, Z. Ye, K. Ching, D. Hawkins, B. Ren, P. C. Scacheri, J. Rozowsky, A. Karpikov, G. Euskirchen, S. Weissman, M. Gerstein, M. Snyder, A. Yang, Z. Moqtaderi, H. Hirsch, H. P. Shulha, Y. Fu, Z. Weng, K. Struhl, R. M. Myers, J. D. Lieb, and X. S. Liu, *Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets,* Genome Research **18**, 393 (2008).

[39] R. Margueron and D. Reinberg, *The polycomb complex PRC2 and its mark in life,* Nature **469**, 343 (2011).

SUMMARY

DNA is packaged together with proteins, such as histones, in the nucleus of a cell to form a fiber called chromatin. The nature of this packaging, the *chromatin structure*, is essential for proper cell functioning. This is illustrated by the fact that perturbating chromatin can be associated with many diseases. Hence, artificial perturbation of chromatin may give important new insights into its function. In this dissertation, we have perturbed chromatin by 1) inducing mutations by integrating retroviruses and transposons into DNA, and 2) evicting histones from chromatin and inducing DNA breaks, by the application of anti-cancer drugs.

By virtue of their ability to integrate into foreign DNA, DNA integrating elements such as retroviruses and transposons are used in gene regulation and cancer research, among others. In cancer research, DNA integrating elements are used for detecting cancer genes in tumor screens. We presented a novel algorithm that fully automates this detection, thus removing any potential for bias induced by manual analysis. In gene regulation, DNA integrating elements can be used for studying the chromatin position effect by the location-dependent activation of transgenes present within randomly integrated transposons. We presented a high-throughput method for studying the chromatin position effect using DNA integrating elements, and studied genome-wide transgene expression values generated using this method, especially in relation to enhancers and domains associated with the nuclear lamina. For both applications of DNA integrating elements, it was important to realize that integrations are randomly, but not uniformly randomly, distributed across the genome. To further investigate this, we generated large datasets of integrations that were under minimal selective pressure, for two transposons and one retrovirus. We compared the integration profiles with a wide range of (epi)genomic features to generate bias maps across multiple genomic scales. This revealed a hierarchical organization in target site selection, and showed that a substantial fraction of cancer genes retrieved from tumor screens may be false positives.

The application of anti-cancer drugs to directly perturb chromatin structure allowed us to take a very low-level approach in studying chromatin. We showed that different drugs target different types of chromatin in evicting histones from chromatin and/or inducing DNA breaks, which can have implications for their chemotherapeutic efficacy.

Central themes throughout this dissertation were computational epigenomics and data integration. Due to the complexity of the biology and the data, many of the computational methods were highly customized. Some are more generally applicable, including a method for the normalization of genome-wide sequencing data with control, and a feature ranking method. However, in general, high levels of customization are unavoidable. Therefore, as a conclusion, the careful consideration that must go into decisions regarding this customization was illustrated by demonstrating the substantial impact that these decisions can have on research outcomes.

SAMENVATTING

DNA wordt samen met eiwitten, zoals histonen, verpakt in de celkern in een complex dat chromatine wordt genoemd. De aard van dit verpakken, ofwel de *chromatinestructuur*, is essentieel voor het normaal functioneren van een cel, en veel ziektebeelden gaan dan ook samen met verstoring van deze structuur. Daarom kan bewuste perturbatie ervan belangrijke nieuwe inzichten opleveren in de functie van chromatine. In dit proefschrift perturbeerden we chromatine door 1) retrovirussen en transposons te laten integreren in het DNA, en 2) het uitstoten van histonen uit chromatine, en het veroorzaken van DNA-breuken met behulp van geneesmiddelen tegen kanker.

Het vermogen van retrovirussen en transposons om te integreren in het DNA van een gastheercel is van belang in het onderzoek naar, onder andere, genregulatie en kanker. In het kankeronderzoek worden retrovirussen en transposons gebruikt om genen te detecteren die betrokken zijn bij het ontstaan van kanker, op basis van de analyse van grote aantallen tumoren. In dit proefschrift beschreven we een nieuw algoritme dat deze detectie volledig automatiseert. In het onderzoek naar genregulatie worden retrovirussen en transposons gebruikt om te bestuderen hoe de genomische locatie van een geïntegreerd transgen (het gen in een retrovirus of transposon) zijn activiteit beïnvloedt. In dit proefschrift beschreven we een methode waarbij dit voor grote aantallen transgenen tegelijkertijd kan worden gedaan, en konden zo hun activiteit genoomwijd bestuderen, voornamelijk in relatie tot enhancers en gebieden geassocieerd met het nucleaire lamina. Voor beide bovenstaande toepassingen was het belangrijk te beseffen dat de integraties niet geheel willekeurig verdeeld zijn over het genoom. Om meer inzicht te krijgen in integratievoorkeuren, genereerden we grote aantallen integraties voor twee transposons en een retrovirus, onder minimale selectiedruk. Deze datasets vergeleken we met een groot aantal (epi)genomische kenmerken. Hieruit bleek dat integratievoorkeur hierarchisch is georganiseerd, en dat bij de detectie van kanker-gerelateerde genen zich veel fout-positieven kunnen voordoen.

Met behulp van bepaalde geneesmiddelen tegen kanker konden we de chromatinestructuur direct beïnvloeden, en zo chromatine op zeer laag niveau bestuderen. We toonden aan dat verschillende geneesmiddelen zich richten op verschillende types chromatine, in hun vermogen om histonen uit te stoten en DNA-breuken te veroorzaken, wat implicaties heeft voor hun chemotherapeutische doeltreffendheid.

Centrale thema's in dit proefschrift waren computational epigenomics en data-integratie. Vanwege de complexiteit van de biologie en de data waren veel van de gebruikte computationele methoden maatwerk. Enkele van de methoden zijn ook meer algemeen toepasbaar, zoals de methode voor het normaliseren van genoomwijde sequencing data met controle data, en de feature-ranking methode. Echter, in het algemeen vereisen kwesties zoals beschreven in dit proefschrift vaak maatwerk. Daarom illustreerden we, ter conclusie, het belang van het maken zorgvuldige afwegingen in het ontwikkelen van maatwerk met behulp van enkele voorbeelden.

ACKNOWLEDGMENTS

During five years of research, there were many people who have directly or indirectly contributed to the completion of this dissertation.

I would like to thank my promotor Lodewyk Wessels for always striving for the best, and for his ever objective, contructive and to-the-point criticism. He provided me with the right amount of freedom in my research, while conscientiously remaining up to date with what I was working on. His comprehensive professional expertise still amazes me, and was a key factor in shaping my research.

By his side, my co-promoter Jeroen de Ridder allowed me to benefit from his years of research experience at the NKI in the field of insertional mutagenesis. I highly value his creative insights and stay-close-to-the-data mentality, which have contributed substantially to giving work in this dissertation a competitive edge.

Due to the collaborative nature of my PhD research, I had the opportunity to benefit from the experience of many highly accomplished research professionals from different research groups within and outside the NKI, including Anton Berns, Sjaak Neefjes, Maarten van Lohuizen, Bas van Steensel, Jos Jonkers, John Hilkens, David Adams, Louise van der Weyden, Alistair Rust, Roland Rad, Jaap Kool, Jitendra Badhai, Wouter Meuleman, Jan-Hermen Dannenberg and Ludo Pagie, all of whom I would like to thank for their valuable discussions, feedback, and the lessons I learned from them.

Specifically, for all but the first half a year of my PhD research, I was privileged to collaborate with Waseem Akhtar, which was not only very enjoyable, but also very productive. For this, much credit should go to his patience in teaching me about biology, his eagerness to learn about the computational sciences, his intelligence and creativity in coming up with envyingly simple and brilliant ideas, and his great sense of humour to put things into perspective even when not all goes according to plan. I am very happy he agreed to be one of the two paranymphs at my dissertation defense.

In much of the work done together with Waseem, also Aleksey Pindyurin played an important role. Within our science trio, his precise nature, being highly critical while always remaining friendly and fair, was often crucial in shaping bomb proof results.

During the second half of my PhD research, my collaboration with Baoxu Pang and Xiaohang Qiao allowed me to get more involved in actual cancer research, while staying true to my interest in epigenomics. I enjoyed a lot, and learned a lot from, our lively discussions and Baoxu's persistence and determination, continuously challenging me to convince him.

In addition to my co-authors mentioned above, this dissertation would not have been the same without many colleagues with whom I discussed my work, worked on side projects, or who provided feedback and/or an inspiring and pleasant working environment. These include the current and past members of the Computational Cancer Biology Group and my office mates on H4, where I spent one day a week (Andi, Bram, Bram, Christiaan, Christine, Daniel, Denise, Evert, Ewald, Gergana, Guillem, Hayssam, Jelle, Joana, Jorma, Julian, Katja, Lorenzo, Magali, Marlous, Martijn, Michael, Nanne, Nicola, Nicos, Nienke, Sander, Santiago, Sergio, Theo, Tycho), and office managers Patti Lagerweij and Tom de Knegt.

I would like to thank my friend Marcel Reuter for being my second paranymph. While not directly involved in my research, our discussions on science in general during our hiking and climbing trips have always been very stimulating.

Finally, I am greatly indebted to my parents, as well as my sisters Nienke and Hanneke, who have taught me that you should do what you enjoy doing, and to Vera, my better half for more than 14 years, with whom I have shared the end of a career in music and the beginning of one in science, and in whose company I enjoy everything so much more.

Curriculum Vitæ

Johann de Jong

20-06-1977	Born in Enkhuizen, The Netherlands.

Education

1989–1995	Pre-university education (VWO) Regionale Scholen Gemeenschap, Enkhuizen, The Netherlands
1990–1995	Special program for gifted children Conservatory of Amsterdam, The Netherlands
1995–2002	Bachelor / Master of Music (piano) Conservatory of Amsterdam, The Netherlands
1997–1998	Exchange student University of the Arts Graz, Austria
2003–2006	Bachelor of Science in computer science The Dutch Open University, Heerlen, The Netherlands
2007–2009	Master of Science in computational science, *cum laude* University of Amsterdam, Amsterdam, The Netherlands

Professional career

1990–2009	Career in music performance and teaching
2009–2014	PhD student The Netherlands Cancer Institute, The Netherlands
2014–2015	Postdoctoral scientist The Netherlands Cancer Institute, The Netherlands

LIST OF PUBLICATIONS

(* indicates equal contribution)

15. **Pindyurin A, De Jong J, Akhtar W**, *High-throughput dissection of chromatin regulatory mechanisms by the TRIP approach*, review commisioned by Genomics: submitted.

14. **Botman D, Jansson F, Röttinger E, Martindale M, De Jong J, Kaandorp JA**, *Analysis of a spatial gene expression database for sea anemone Nematostella vectensis during early development*, submitted.

13. **Pang B*, De Jong J*, Qiao X*, Wessels LFA, Neefjes J**, *Genome-wide profiling of anti-cancer drug effects on chromatin guides rational treatment design*, submitted.

12. **De Jong J, Van Lohuizen M, De Ridder J, Wessels LFA, Akhtar W**, *Applications of DNA integrating elements: facing the bias bully*, review commissioned by Mobile Genetic Elements: in press.

11. **Babaei S, Akhtar W, De Jong J, Reinders M, De Ridder J**, *Long-range chromatin interactions of cancer-causing insertions in mouse tumors*, Nature Communications, in press.

10. **De Vries N, Hulsman D, Akhtar W, De Jong J, Blom M, Miles D, Van Tellingen O, Jonkers J, and Van Lohuizen M**, *Prolonged Ezh2 Depletion in Glioblastoma Causes a Robust Switch in Cell Fate Resulting in Tumor Progression*, Cell Reports, 2015, doi:10.1016/j.celrep.2014.12.028.

9. **Botman D, Roettinger E, Martindale M, De Jong J, Kaandorp JA**, *A computational approach towards a gene regulatory network for the developing Nematostella vectensis gut*, PLOS ONE, 2014, doi:10.1371/journal.pone.0103341.

8. **De Jong J*, Akhtar W*, Badhai J, Rust AG, Rad R, Adams DJ, Hilkens J, Berns A, Van Lohuizen M, Wessels LFA, De Ridder J**, *Chromatin landscapes of retroviral and transposon integration profiles*, PLoS Genetics, 2014, doi:10.1371/journal.pgen.1004250.

7. **Akhtar W, Pindyurin A, De Jong J, Pagie L, Berns A, Wessels LFA, Van Steensel B, Van Lohuizen M**, *Using TRIP for genome-wide position effect analysis in cultured cells*, Nature Protocols, 2014, doi:10.1038/nprot.2014.072.

6. **Bard-Chapeau E, Nguyen A-T, Rust A, Sayadi A, Lee P, Chua B, New L-S, De Jong J, Ward J, Chin C, Chew V, Dr. Toh H-C, Abastado J-P, Benoukraf T, Soong R, Bard F, Dupuy A, Johnson R, Radda G, Wessels L, Chan E, Adams D, Jenkins N**, *Transposon mutagenesis identifies genes driving hepatocellular carcinoma in a chronic hepatitis B mouse model*, Nature Genetics, 2014, doi:10.1038/ng.2847.

5. **Akhtar W*, De Jong J*, Pindyurin A*, Pagie L, Meuleman W, De Ridder J, Berns A, Wessels LFA, Van Lohuizen M, Van Steensel B**, *Chromatin position effects assayed by thousands of reporters integrated in parallel*, Cell, 2013, doi:10.1016/j.cell.2013.07.018.

4. **De Ridder J, Kool J, Uren AG, Bot J, De Jong J, Rust AG, Berns A, Van Lohuizen M, Adams DJ, Wessels LFA, Reinders M**, *Mutational genomics for cancer pathway discovery*, Pattern Recognition in Bioinformatics, 2013, doi:10.1007/978-3-642-39159-0.

3. **Heideman RM, Wilting RH, Yanover E, Proost N, Velds A, De Jong J, Jacobs H, Wessels LFA, Dannenberg JH**, *Gene dosage dependent tumor suppression by histone deacetylases 1 and 2 through regulation of c-Myc collaborating genes and p53 function*, Blood, 2013, doi:10.1182/blood-2012-08-450916.

2. **De Jong J, De Ridder J, Van der Weyden L, Sun N, Van Uitert M, Berns A, Van Lohuizen M, Jonkers J, Adams DJ, Wessels LFA**, *Computational identification of insertional mutagenesis targets for cancer gene discovery*, Nucleic Acids Research, 2011, doi:10.1093/nar/gkr447.

1. **Tamulonis C, Postma M, Marlow HQ, Magie CR, De Jong J, Kaandorp JA**, *A cell-based model of Nematostella vectensis gastrulation including bottle cell formation, invagination and zippering*, Developmental Biology, 2010, doi:10.1016/j.ydbio.2010.10.017.