Biochemistry and Biophysics Reports 2 (2015) 143–152

Contents lists available at ScienceDirect

Biochemistry and Biophysics Reports

journal homepage: www.elsevier.com/locate/bbrep

Chromatin structure analysis enables detection of DNA insertions into the mammalian nuclear genome

Challise J. Sullivan [1], Erik D. Pendleton [1], Rachel E. Abrams [1], David L. Valente, Michelle L. Alvarez, Richard H. Griffey, John Dresios *

Leidos Inc., 10260 Campus Point Drive, San Diego, CA 92121, United States

ARTICLE INFO

Article history:
Received 3 February 2015
Received in revised form
28 May 2015
Accepted 8 June 2015
Available online 10 June 2015

Keywords:
Chromatin
ChIP-Seq
Genetically modified organism
Epigenetics
DNA
Histone modifications

ABSTRACT

Background: Genetically modified organisms (GMOs) have numerous biomedical, agricultural and environmental applications. Development of accurate methods for the detection of GMOs is a prerequisite for the identification and control of authorized and unauthorized release of these engineered organisms into the environment and into the food chain. Current detection methods are unable to detect uncharacterized GMOs, since either the DNA sequence of the transgene or the amino acid sequence of the protein must be known for DNA-based or immunological-based detection, respectively.
Methods: Here we describe the application of an epigenetics-based approach for the detection of mammalian GMOs *via* analysis of chromatin structural changes occurring in the host nucleus upon the insertion of foreign or endogenous DNA.
Results: Immunological methods combined with DNA next generation sequencing enabled direct interrogation of chromatin structure and identification of insertions of various size foreign (human or viral) DNA sequences, DNA sequences often used as genome modification tools (*e.g.* viral sequences, transposon elements), or endogenous DNA sequences into the nuclear genome of a model animal organism.
Conclusions: The results provide a proof-of-concept that epigenetic approaches can be used to detect the insertion of endogenous and exogenous sequences into the genome of higher organisms where the method of genetic modification, the sequence of inserted DNA, and the exact genomic insertion site (s) are unknown.
General significance: Measurement of chromatin dynamics as a sensor for detection of genomic manipulation and, more broadly, organism exposure to environmental or other factors affecting the epigenomic landscape are discussed.

1. Introduction

Recombinant DNA technology allows the modification of particular characteristics of animals, plants, or microbes by introducing selected segments of genetic material from other, sometimes non-related, organisms. According to the World Health Organization, genetically modified organisms (GMOs) are defined as organisms in which the DNA has been altered in a way that does not occur naturally by mating and/or natural recombination [24]. Genetically engineered animals represent a pioneering technology with various applications in biomedicine, through the production of various proteins, drugs, vaccines, and tissues for human use; in agriculture, through the generation of more efficient and disease-resistant livestock; and in diet, through enhancement of the quality and reduction in the cost of food production [25]. Given these benefits of GMOs, it is necessary to develop accurate and sensitive methods to detect, track, and assess the authorized and unauthorized release of GMOs into the environment and into the food chain [7,19]. Additionally, the development of such detection methods is a prerequisite for reliable identification and control of engineered organisms that create risks to the food supply and to human health (*e.g.* agroterrorism).

Commonly used methods cannot detect uncharacterized genetically engineered organisms, since either the DNA sequence of the transgene or the amino acid sequence of the protein must be available for DNA- (*e.g.* PCR, probe magnetic capture/spectroscopy, microarrays) or immunochemical- (*i.e.* monoclonal and polyclonal antibodies) based detection, respectively [15,17,18,2,22,28,5,9]. Similarly, designing DNA microarrays for GMO surveillance would be challenging given the extreme diversity of genomic sequences among organisms and the great variety of molecular tools that can

* Corresponding author.
E-mail address: john.dresios@leidos.com (J. Dresios).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.bbrep.2015.06.002

be used for production of GMOs [8].

We propose that insertion of DNA sequences into a host genome causes remodeling of the chromatin structure by altering the interactions between histone proteins and DNA sequences around the inserted elements. Such chromatin structure changes could influence gene expression by modifying both short- and long-range regulatory interactions, therefore leading to alterations in protein expression, and eventually to a desired physiological outcome. To this end, we applied genome-wide chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) technology to characterize DNA–histone interactions for the identification of molecular signatures corresponding to insertion of endogenous or exogenous DNA elements into the mouse genome. The results provide a proof of concept that chromatin mapping technologies can be used to detect the insertion of DNA sequences into the genome of a higher mammalian organism.

2. Materials and methods

2.1. Sample selection

Muscle tissue samples from wild-type and genetically modified mice (Jackson Laboratory, Bar Harbor, ME, USA) were of FVB/NJ genetic background, were all male, and were ~8 weeks old. Wild-type (stock# 001800) and three GMO samples were selected for testing. Genome alterations present in the GMO mouse samples included insertions of various size endogenous (mouse) or foreign (*e.g.* human, viral) DNA elements that were incorporated into the host genome using various genome modification tools (*e.g.* viral sequences, transposons) and engineering methods (*e.g.* embryonic stem cells transformation or pronuclear injection). A list of the samples used in this study along with their genomic modifications and size of DNA insertion is shown in Table 1. The composition of transgenic insertions is illustrated in Fig. 1.

GMO sample 1 (stock# 018304; [4]) was created *via* the microinjection method and contained an inserted transgene holding the human alpha-skeletal actin (*ACTA1*) promoter sequence, the full-length human tropomyosin-3 cDNA sequence (*TPM3*) and a cassette containing the simian virus 40 (SV40) small t-antigen (tAg) intron and 3′UTR (Fig. 1A). GMO sample 2 (stock# 012460; [6]) was created *via* the microinjection method and contained an insertion of 40 copies of a DNA sequence cassette consisting of the entire coding region of the mouse GTP-binding proteins class Gq protein subunit Gaq (*Gnaq*) gene, under control of a mouse alpha-myosin heavy chain (*Myh6*) promoter, followed by a SV40 intron and a polyadenylation signal (polyA) (Fig. 1B). GMO sample 3 (stock# 017594; [32]) was created *via* co-injection of two transgenes using the Sleeping Beauty (SB) transposon approach. The first transgene contains a left and right inverted repeat/direct repeat sequence (IR/DR) known as the SB transposon recognition site, a mouse tyrosinase (*Tyro*) enhancer sequence, and Tyro minigene (*TyBS*). The second transgene has the mouse protamine 1 (*Prm1*) promoter, a 25 bp linker, SB10 gene, and a rabbit β-globin splice/polyA sequence (Fig. 1C).

2.2. Chromatin immunoprecipitation (ChIP-seq)

ChIP-seq experiments were performed in agreement with the guidelines set forth by the ENCODE project [14]. Three histone antibodies were used for the chromatin immunoprecipitation experiment, specifically: H3K4me3 (Millipore, MA) which binds to active promoters, H3K36me3 (AbCam, MA) which binds to active exon boundaries, and H3K4me1 (AbCam, MA) which binds to active enhancers [11]. An optimized protocol for isolation of nuclei from skeletal muscle tissue was developed and was based on a previously published method [26]. Briefly, minced skeletal muscle was cross-linked with formaldehyde and nuclei were prepared for chromatin immunoprecipitation according to the chromatin shearing (sonication) method described by [21]. Chromatin samples were incubated with each histone antibody and immunoprecipitated using Protein A/G Magnetic Beads (Thermo Fisher Scientific, MA) according to the product specifications. Samples were reverse cross-linked and purified using a PCR Purification Kit (Qiagen, MD), followed by quantification using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, CA). The DNA fragment size range was determined using the Bioanalyzer High Sensitivity ChIP (Agilent Technologies, CA).

Validation of the DNA–histone immunoprecipitation reactions was conducted by quantitative PCR using SYBR FAST qPCR Master Mix (KAPA Biosystems, MA) and positive control primers (IDT, IA) designed to bind specific genomic regions known to be immunoprecipitated by the three tested antibodies. Specifically, amplification of the following gene regions was verified before sequencing was initiated: Actg1 and Actb for H3K4me3, Actg1 and Elf1 for H3K36me3, and Actg1, Elf1, and Gapdh for H3K4me1.

Samples were prepared for multiplex sequencing following the Illumina ChIP-seq Library Prep Kit (Illumina, CA). Sequencing libraries were quantified with Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, CA) and Illumina/Universal Quantification Kit (KAPA Biosystems, MA), and DNA fragment sizes were determined using the Bioanalyzer High Sensitivity ChIP (Agilent Technologies,

Table 1
Wild-type and genetically modified organism (GMO) mice samples used in this study.

Sample	Genetic background	Gender	Age	Engineering approach	Endogenous (mouse) insertions	Exogenous (human, viral, etc.) insertions	Total insertion size (~kb)
Wild-type ($n=4$)	FVB/NJ	Male	8 wks	N/A	N/A	N/A	N/A
GMO sample 1 (#018304)	FVB/NJ	Male	8 wks	Microinjection	N/A	Human *ACTA1* promoter Human *TPM3* cDNA SV40 tAg Intron & 3′UTR	40
GMO sample 2 (#012460)	FVB/NJ	Male	8 wks	Microinjection	Mouse *Myh6* promoter Mouse *Gnaq* cDNA	SV40 Intron & poly(A)	200
GMO sample 3 (#017594)	FVB/NJ	Male	8 wks	Sleeping Beauty transposon	Transgene A: Mouse *Tyro* enhancer Mouse *Tyro* minigene Transgene B: Mouse *Prm1* promoter	Transgene A: Left and right IR/DR Transgene B: Linker SB10 gene Rabbit β-globin splice/poly (A)	8 2

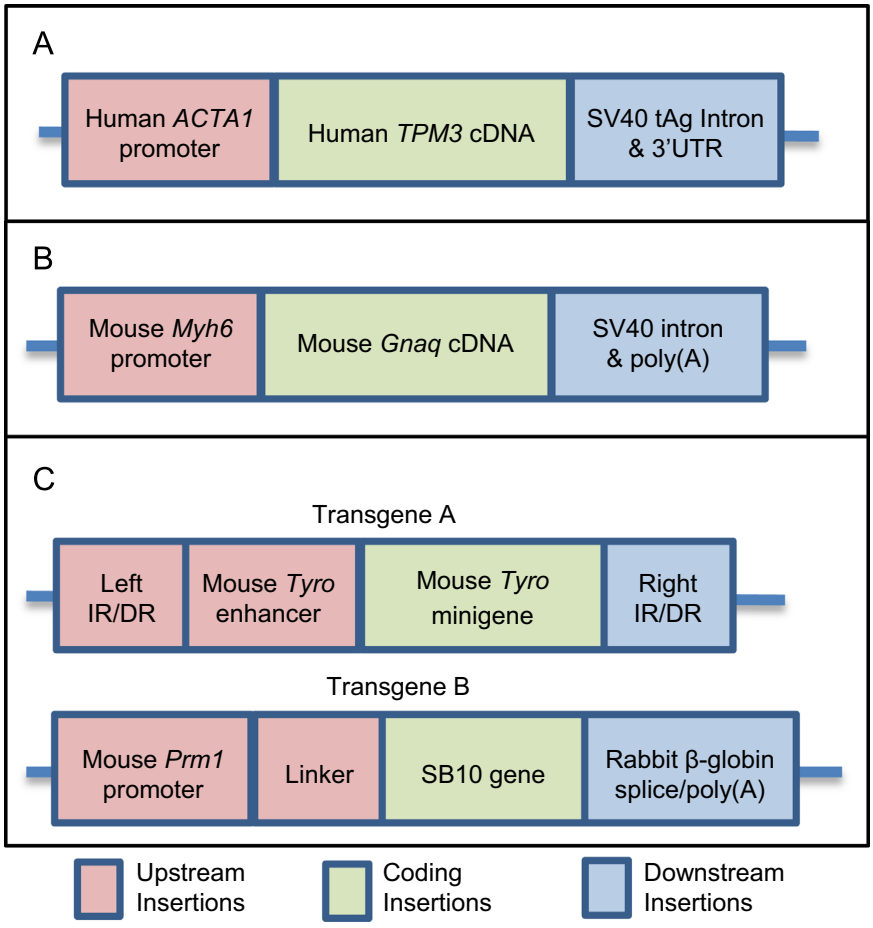

Fig. 1. GMO transgenic insertions. Illustration of inserted cassettes for the three GMO samples tested in this study. (A) GMO sample 1 contains a ~40 kb insertion containing the human *ACTA1* promoter and human *TPM3* cDNA sequence with a downstream viral SV40 element. (B) GMO sample 2 contains a cassette with the mouse *Myh6* promoter and mouse *Gnaq* cDNA sequence followed by a viral 3′ UTR element. (C) GMO sample 3 consists of two transgenes inserted at different genomic locations. Transgene A contains a left and right inverted repeat/direct repeat sequence (IR/DR), a mouse Tyrosine (*Tyro*) enhancer, and Tyro minigene; transgene B contains the mouse protamine 1 (*Prm1*) promoter, linker, SB10 gene, and a rabbit 3′UTR.

CA). Cluster Generation was performed on the cBot (Illumina, CA) and sequencing was performed on the Genome Analyzer IIx (Illumina, CA), generating 35 bp single end reads (Fig. 2a).

2.3. Genomic annotations

Mouse genomic annotations including chromosome, coding strand, and exon boundaries were downloaded from the UCSC RefSeq Genes table using the following settings: Assembly: July 2007, NCBI37/mm9; Group: Gene and Gene Predictions; Track: RefSeq Genes; Table: refGene; and Region: Genome. The entire human *ACTA1* genomic region ±2.5 kb (hg19; chr1; 229564492–229572343) and full length *TPM3* cDNA sequence (NM_152263) were obtained from Genbank and used as reference sequences during alignment. Four SV40 intronic regions were identified at the following positions relative to the SV40 Viral Genome (Genbank NC_001669): pos: 295–434 (late 19s intron), pos: 527–1462 (late 16s intron), pos: 4572–4917 (large TAg intron), and pos: 4572–4636 (small tAg intron).

2.4. Data processing: insertion of endogenous mouse sequences

Data workflow was performed following guidelines outlined by [14] (Fig. 2b). ChIP-seq data was demultiplexed using CASAVA 1.8.2 (Illumina, CA) and high quality sequencing data was retained after a per-lane and per-sample data quality check. High quality data was defined as that with a mean quality score of at least 35. For mouse endogenous DNA insertions, the standalone ELAND2 (Illumina, CA) aligner was used for mapping the high quality demultiplexed sequence reads to the mouse genome (mm9). The quality of the alignment was assessed by examining the number of uniquely mapped reads (*i.e.* those reads that align to a single genomic location). For the three antibodies used (H3K4me3, H3K36me3, and H3K4me1), wild-type mice samples averaged ~12.2 ± 1.5 M, ~12.6 ± 1.3 M, and ~12.4 ± 2.2 M uniquely mapped reads, respectively. GMO sample 1 generated ~17.7 M, ~17.0 M, and ~9.1 M; GMO sample 2 generated ~8.2 M, ~6.9 M, and ~9.0 M; and GMO sample 3 generated ~15.0 M, ~18.4 M, and ~17.1 M, uniquely mapped reads, for the three antibodies, respectively.

ChIP-seq peak finding was performed with Model-based Analysis for ChIP-seq (MACS 1.4.2) using the built-in mouse genome-size setting [27]. The total number of called peaks across the mouse genome for the four wild-type samples averaged 21,561 ± 2007 reads for H3K4me3, 58,741 ± 3310 reads for H3K36me3, and 74,988 ± 9,157 reads for H3K4me1. The total number of peaks for GMO sample 1 was 24,647, 56,205, and 72,727 peaks; for GMO sample 2 was 11,420, 66,991, and 76,019 peaks; and for GMO sample 3 were 24,648, 50,402, and 88,566 peaks for the H3K4me3, H3K36me3, and H3K4me1 histone antibodies, respectively. Peaks were examined and graphically represented using MATLAB (R2013b). For each chromosomal region of interest, peak data for all three of the histone antibodies were plotted on a single graph, for each wild-type and each GMO sample. These peak plots were qualitatively examined to identify both similarities and differences between wild-type and GMO samples.

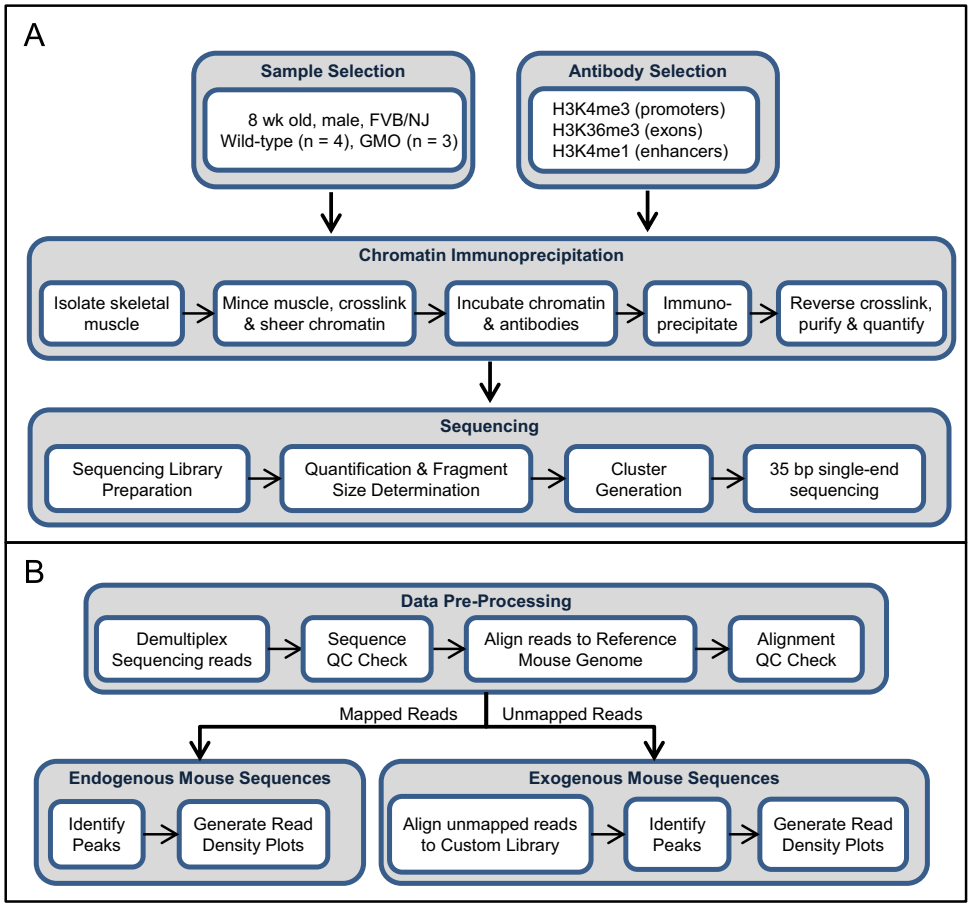

Fig. 2. Laboratory and ChIP-seq data analysis flow diagram. ChIP-seq was performed on wild-type and genetically modified (GMO) mouse samples with three open chromatin histone binding antibodies. Single end sequencing was performed on the Illumina GAIIx and reads demultiplexed with CASAVA. High quality reads were mapped to the mouse reference genome (mm9) and peaks identified with MACS. Unmapped mm9 reads were remapped to a custom reference library and peaks identified with MACS.

2.5. Data processing: insertion of exogenous mouse sequences

The GMO samples used in this research effort each contained an inserted transgene holding some exogenous, non-mouse, genomic sequences; specifically, the human *ACTA1* gene region, the human *TPM3* cDNA sequence, the four SV40 intronic regions, and the SB10 cassette (Table 1, Fig. 1). Accordingly, high quality ChIP-seq reads that had failed to align to the mouse reference genome (*i.e.* unmapped reads) were aligned to a custom reference genomic library containing these exogenous genomic sequences (Fig. 2b). Sequence Alignment/Map tools (SAMtools 0.1.18; [16]) were used to generate FASTQ files for the unaligned reads, which were then aligned to the aforementioned reference library using the ELAND2 standalone aligner. Peak finding was executed on the resulting BAM files using MACS 1.4.2 to generate WIG files with the genome-size setting modified to represent the size of the reference used for alignment. The resulting peak data was then further examined and graphically represented MATLAB (R2013b).

3. Results

3.1. Detection of transgene insertions carrying endogenous mouse sequences

The chromatin structure of genetically engineered mouse samples was assessed for the detection of changes caused by the insertion of transgenic DNA sequences into the host genome. These studies employed ChIP-seq methodology for high resolution, genome-wide mapping of DNA fragments associated with selected proteins followed by comparison of those maps between genetically matched transgenic and wild-type mice samples. The ChIP-seq process was initiated with induction of cross-linking between DNA and its associated proteins. Chromatin was then isolated, subjected to sonication for DNA fragmentation, and sheared chromatin was immunoprecipitated using antibodies against specific histones, specifically H3K4me3, H3K4me1, and H3K36me3. H3K4me3 is a hallmark of actively transcribed protein-coding promoters in eukaryotes; H3K4me1 marks active transcriptional enhancers; and H3K36me3, has been associated with the exons of actively transcribed genes [11]. Immunoprecipitated DNA samples were purified, ligated to sequencing adapters, and assayed using the Next Generation Genome Analyzer IIx sequencing platform (Illumina, CA). This step generated short sequence reads of the DNA fragments bound by the protein(s) of interest.

Initially, transgenic samples carrying native mouse elements were examined. Although the genomic location of the transgene was unknown, alignment of ChIP-seq data to the endogenous mouse genomic locations was expected. Specifically, an increase in the number of reads mapped to these known genomic locations should appear similar to a copy number variation (CNV) event. The native genomic positions for the mouse *Myh6* promoter and *Gnaq* coding sequence transfected into GMO sample 2 (#012460) and the *Tyro* enhancer sequence, *TyBS* minigene, and *Prm1* promoter transfected into GMO sample 3 (#017594) were investigated. The transgene injected into GMO sample 1 did not contain any

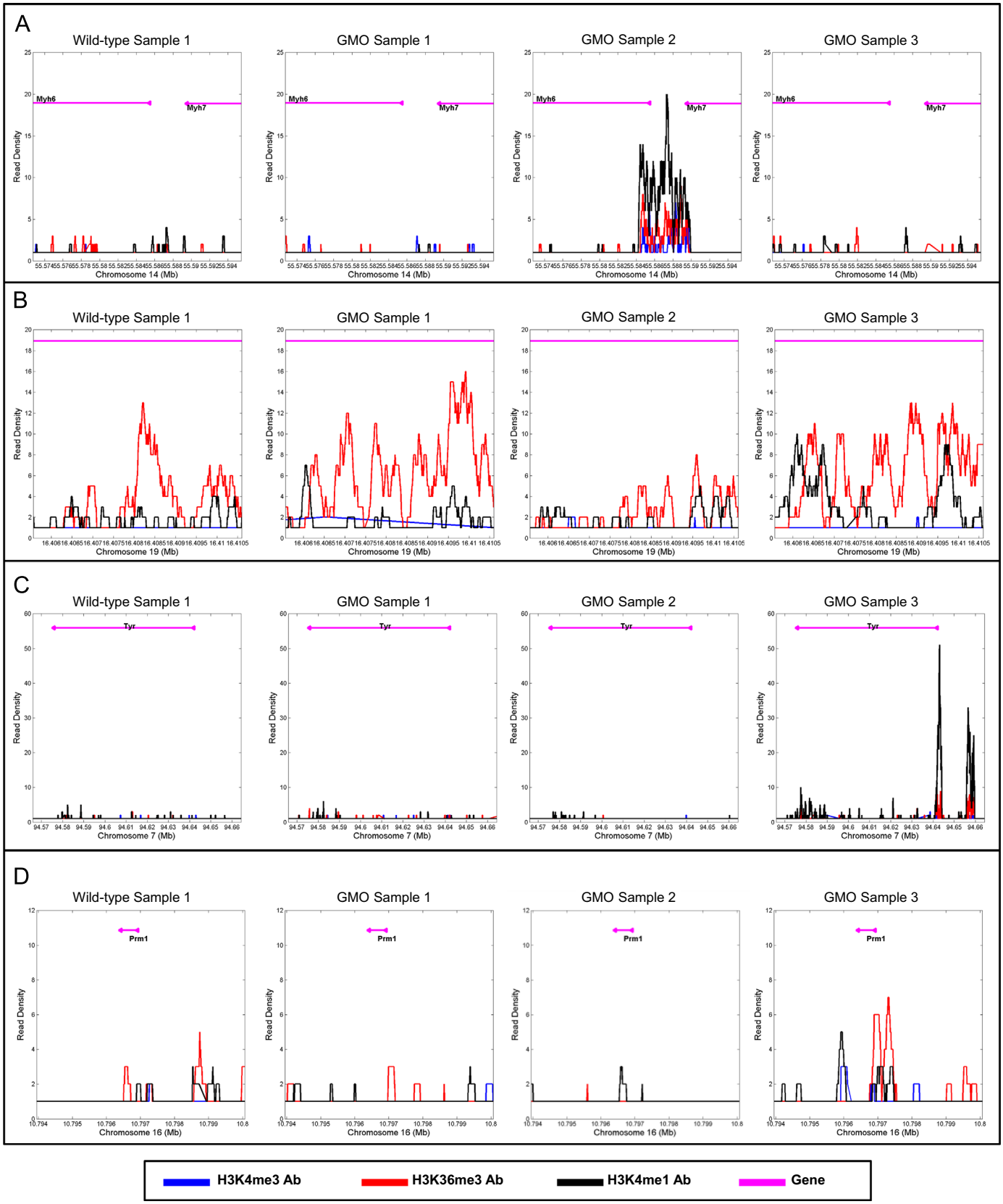

Fig. 3. Endogenous mouse peak plots. Peak read density plots for controls and test samples at various endogenous genomic positions. (A) *Myh6* promoter region in chromosome 14 shows increased signal for GMO sample 2. (B) Exons 4 and 5 in *Gnaq* gene show no distinct variation for any sample tested; (C) T*yro* minigene and *Tyro* enhancer region show increased signal for GMO sample 3; and (D) Protamine promoter region shows increased signal for sample 3.

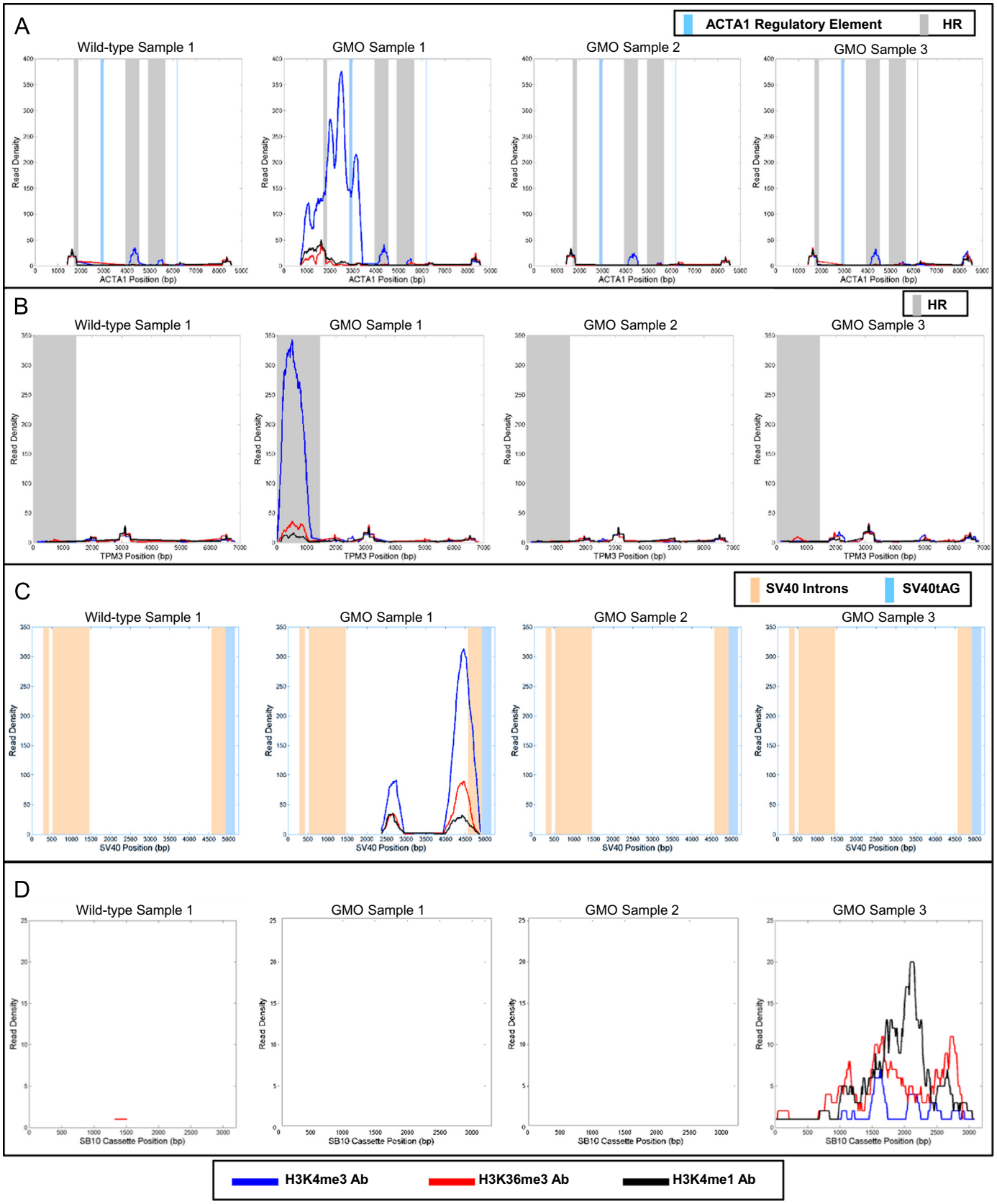

Fig. 4. Exogeneous mouse peak plots. Read density plots for the mouse unmapped sequencing reads. (A) A number of sequencing reads aligned to the human *ACTA1* gene in only the GMO sample 1. The gray vertical bans indicate regions of homology between the human and mouse genomes. (B) Sequencing reads aligned to the human *TPM3* gene in only the GMO sample 1 data. The gray vertical bans indicate regions of homology between the human and mouse genomes. (C) Unmapped reads aligning to the SV40 genomic reference template in GMO sample 1 data. (D) Reads aligning to the SB10 cassette within GMO sample 3 data are shown in relation to the absence of sequencing read alignment with any other sample. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

endogenous mouse sequences and therefore no specific insertions for GMO sample 1 were evaluated at this stage. Qualitative examination of peaks at the endogenous mouse genomic regions for one representative wild-type sample and the three GMO mice is shown in Fig. 3. The four mouse endogenous genomic regions for the *Myh6* (Fig. 3A), *Gnaq* (Fig. 3B), *Tyr* and *TyBS* (chr7 separated by ~12 kb) (Fig. 3C), and *Prm1* (Fig. 3D) genes are shown. A very broad set of peaks at the *Myh6* region for all three antibodies tested can be seen when comparing GMO sample 2 to the wild-type and other GMO samples (Fig. 3A). In contrast, GMO sample 2 showed a decrease in the number of peaks for the H3K36me3 antibody as well as a modest reduction in the number of peaks for the other two antibodies surrounding exons 4 and 5 of the *Gnaq* gene (Fig. 3B). This may indicate an inactivation of the chromatin region surrounding both the endogenous and exogenous *Gnaq* sites therefore causing poor representation in the Chip-seq data. GMO sample 3 contains three native mouse gene sequences: the Tyr enhancer and TyBS minigene (Fig. 3C) and the *Prm1* promoter (Fig. 3D). These regions show an increase in the amount of reads for the H3K4me1 and H3K36me3 antibodies when compared to the wild-type and other GMO samples.

3.2. Detection of transgene insertions carrying exogenous mouse sequences

Next, we examined transgenic samples containing the insertion of elements foreign to the mouse genome. Since these sequences are not native to the mouse genome, the associated ChIP-seq reads were not expected to align to the mouse reference genome. These unmapped reads were thus realigned to a new reference sequence collection corresponding to the exogenous inserted sequence (Section 2). If these reads align to this custom reference sequence, then the presence of the sequence has been identified in the sample of interest.

All three of the test samples for the study contained insertions foreign to the mouse genome (Table 1). The foreign insertions originated from the human genome (*ACTA1* and *TPM3* in GMO sample 1), a viral genome (SV40 in GMO samples 1 and 2), and a common genetic engineering transfection tool (SB10 cassette in GMO sample 3). The unmapped reads were aligned to a reference genome, the detected peaks were plotted, and qualitatively examined for similarities and differences between wild-type and GMO samples (Fig. 4).

3.2.1. Insertion of human sequences into the mouse genome

As shown in Table 1, the foreign elements in GMO sample 1 were two human insertions (*ACTA1* promoter and *TPM3* cDNA) as well as the Simian Virus small tAg (SV40). The sequence for the entire Human *ACTA1* gene was used as the reference genome to align the unmapped reads (see Section 2). Fig. 4A illustrates the detected peaks for a single wild-type sample and the three GMO samples. As shown, GMO sample 1 contains a strong peak difference when compared to the other samples for all three antibodies tested for the *ACTA1* gene. This peak mapped to one of the known *ACTA1* regulatory elements (light blue vertical lines) confirming the insertion of an *ACTA1* promoter into GMO sample 1. For the wild-type controls and the two GMO samples not containing these insertions (*i.e.* GMO samples 2 and 3), only minor peaks (<50 reads) were observed. Further examination revealed that these peaks aligned to the homologous regions of the *ACTA1* gene (*i.e.* those regions of high sequence similarity between the mouse and human genomes) which are denoted by the gray bands in Fig. 4.

The sequence for the entire human *TPM3* gene was also used as a reference template for unmapped read mapping. GMO sample 1 contained a large peak located at the 5′ end of the *TPM3* gene which is absent for all other samples (Fig. 4B). Similar to the results for the human *ACTA1* gene, minor background peaks (<50 reads) for all three tested antibodies were observed for all other samples.

3.2.2. Insertion of viral sequences into the mouse genome

The entire SV40 sequence was used as the reference genome to investigate the insertion of the SV40 small tAg present in GMO sample 1 and SV40 intron present in GMO sample 2. As shown in Fig. 4C, GMO sample 1 showed two separate locations of peaks, one spanning a region centered around ~2.7 kb and another centered around ~4.4 kb, where each location had a peak for each antibody tested. The region at ~4.4 kb is adjacent to the small tAg, the inserted element. The area around ~2.7 kb is in a region that exhibits sequence similarity to the mouse genome and most likely represents either an artifact of the alignment or an additional SV40 sequence present in the transgene but not originally annotated. In contrast to the peaks identified in GMO sample 1, the other two GMOs and all four wild-type samples had no peaks aligning to the SV40 genome. It is not unexpected that GMO sample 2 failed to have peaks aligning to SV40 even though it contained an intron of the SV40 genome since the histone antibodies used in these experiments are not known to bind to intronic regions.

3.2.3. Insertion of transposase gene elements into the mouse genome

As shown in Table 1 and in Fig. 1, the exogenous inserted element in GMO sample 3 was the Sleeping Beauty (SB) cassette. This cassette contains multiple sequence components (*e.g.* IR/DR, SB10, linker), which were concatenated to form a complete reference sequence of the transgene for alignment of the unmapped reads. As expected, the wild-type and GMO samples 1 and 2 had no peaks aligned to the SB10 cassette sequence. GMO sample 3, which contained the SB10 insertion, showed a response for all antibodies tested across the reference sequence (Fig. 4D). It should be noted that the alignment to the SB10 cassette sequence was executed in two different ways. First, as previously described, the elements of the cassette were concatenated to form a single reference sequence used for alignment. Second, the alignment was done using each SB10 cassette element as its own reference sequence. The peaks detected were the same regardless if the alignment was done to the synthesized sequence or each component separately.

4. Discussion

It has been established that the dynamics of chromatin structure are modulated by epigenetic factors, such as DNA methylation and histone modifications (acetylation, methylation and ubiquitylation), nucleoprotein remodeling complexes, and tissue-specific proteins [10]. Histone modifications in particular, affect chromatin structure and subsequent gene expression through local changes of nucleosome structure and recruitment of chromatin remodeling complexes [23]. We postulated that insertion of DNA sequences into a host eukaryotic organism may alter the patterns of interaction of nuclear DNA with a variety of proteins, such as histones, transcription factors and other regulatory, DNA-associated transfactors. These differences would result from: a) direct interactions between the inserted DNA sequences and nuclear proteins, b) changes in the chromatin structure owing to transcriptional activation of the inserted gene(s), and c) DNA methylation changes that may in turn affect DNA–protein association patterns. However, to our knowledge, there has been neither comprehensive characterization of the effects on chromatin structure of the insertion of DNA sequences into the mammalian genome nor exploitation of such effects for GMO detection. The objective of the present study was therefore to test the hypothesis that chromatin mapping methods can be used to detect the footprint of DNA

insertions into the mammalian genome, irrespectively of the method of modification used, the exact sequence of the inserted DNA or its insertion position in the host genome.

In this work, immunological methods combined with DNA next generation sequencing enabled direct interrogation of chromatin structure and identification of genetic modifications into the nuclear genome of a model animal organism. These studies used publicly available muscle tissue samples from genetically modified mice generated *via* multiple DNA insertion methods and carrying a variety of sequence alterations. These alterations included insertions of various size foreign (human or viral) DNA sequences, DNA sequences often used as genome modification tools (*e.g.* viral sequences, transposon elements), or endogenous DNA sequences resulting in gene duplications or CNVs in the host genome. To minimize the intrinsic chromatin structure variability between animals, all mice used in this study were of the same genetic background (FVB/NJ), were all male, and were approximately 8 weeks of age. Muscle tissue control and engineered mice were subjected to ChIP-seq, a technique that enriches DNA fragments to which a specific protein is bound, and enables genome-wide profiling of protein–DNA interactions and identification of potential differences between transgenic and wild-type organisms. For this study, we employed antibodies against posttranslationally modified histone molecules known to associate with different active chromatin regions; H3K4me3, which marks active promoters, H3K4me1 that binds to active enhancers, and H3K36me3, which marks exons of actively transcribed genes [11]. Presumably, depending on the appropriate selection of antibodies against different DNA binding proteins, a similar approach can be used to detect inserted genes that are either in an active (transcribed) or inactive (silenced) state at the time of harvesting, depending on the type of tissue source and/or gene temporal expression pattern. Such antibodies associated with transcribed or silenced parts of the genome have been demonstrated and validated previously [29,30,31,11].

The insertion of an endogenous sequence into the host genome, irrespective of where in the genome it has been inserted, would result in increased copies of that sequence, which upon alignment of the ChIP-seq data is expected to appear as a change in the enrichment level in the endogenous location of that element. For example, Fig. 3A illustrates this scenario in the case of insertion of mouse *Myh6* promoter sequences, depicting an increased signal response. Interestingly, this response was not limited to only H3K4me3, which marks active promoters but also to H3K4me1 and H3K36me3, which mark active enhancers and exons of actively transcribed genes, respectively. This result is in agreement with a proposed model of enhancer activity and long-distance gene activation [3]. Additionally, there are reports showing enhancers can be marked by histone H3K4me3 [1], further suggesting physical interactions between H3K4me3-marked promoters and H3K4me1-marked enhancers. Signatures of DNA insertions were also detected in the case of insertion of multiple endogenous transgenes at different locations of a mouse genome. A sample carrying insertion of endogenous sequences containing the Tyrosinase enhancer and Tyrosinase minigene exhibited increased response with the H3K4me1 and H3K36me3 antibodies (Fig. 3C), which as described above, mark active enhancers and exons of actively transcribed genes, respectively.

A different approach was taken for analysis of ChIP-seq data originating from samples isolated from animals containing insertion of elements foreign to the host genome, such as human and viral sequences or sequences of commonly used genetic engineering tools (*e.g.* the SB10 cassette). Since such ChIP-seq reads did not align to the mouse genome, they were tested for alignment to a collection of reference sequences to assess the potential presence of a foreign sequence into the mouse genome. A signal response for the presence of human *ACTA1* and *TPM3* gene insertions was detected with all three antibodies tested (Fig. 4A and B, respectively). Similarly, all three antibodies tested showed a signal for the presence of SV40 viral sequences in the expected sample (Fig. 4C). Interestingly, a sample containing an intron of the SV40 genome had no peak reads align to the viral sequence (Fig. 4C), which may be explained by the fact that the antibodies used for this study are known to correspond to active exons, enhancers or promoters, but not intronic regions of the genome. Using the aforementioned approach, we were also able to identify the presence of engineered SB10 transposase elements; alignment of pulled-down sequences to a compiled reference sequence led to successful allocation of the signals generated by any of the three antibodies tested to only the particular sample that had been modified with these sequences (Fig. 4D).

It is envisioned that a similar approach can be followed for identification of GMOs created through modifications by foreign DNA sequences based on an automated screening of ChIP sequences against a library of reference data consisting of sequences derived from published genetic engineering tools as well as genomes from multiple organisms different than the host leveraging available reference files (*e.g.*, NCBI). For compound insertions, as shown with the example of the SB10 cassette, it is sufficient for the library to contain the building blocks of the cassette independently of each other. Advances in *de novo* assembly of very short reads (*e.g.*, [13] could be explored to enable assembly of unaligned reads to agnostically identify exogenous inserts. This approach would render the need for a custom reference sequences library obsolete but may require additional analysis upon sequence assembly as well as potential modifications to the experimental parameters (*e.g.*, increase read lengths or paired-end reads).

These studies successfully determined signatures in the ChIP-Seq data for both endogenous and exogenous insertions as a result of the modifications. Our approach of initially aligning to the host genome followed by an analysis of the unaligned reads is recommended when examining an unknown sample suspected of having a modification. Analyzing these results requires knowledge of the expected variability relative to the host genome as well as the statistics of unmapped reads for unmodified organisms. Hence, one of the greatest challenges toward a practical application of the approach outlined in this study is the intrinsic variation in chromatin states between samples. It is evident that the cellular chromatin structure depends on the underlying genetic background and is constantly changing in response to physiological and environmental factors, playing an integral role in regulation of gene expression and phenotype. In addition, the algorithms developed here will need to be extended to further normalize chromatin structure data and generate baseline epigenetic patterns accounting for variations in the genetic background, age, and environmental factors affecting chromosome dynamics. Continuation of the extraordinary progress in performance and cost of high resolution/high-throughput sequencing of large genomes, array technologies, single-cell level characterization platforms, and software data analysis tools achieved in the recent years [20], in combination with the vast amount of information being generated from the human and model organism Encyclopedia of DNA Elements (ENCODE) projects, should allow chromatin profiling across numerous tissues, cell types, and conditions. Furthermore, application of an integrative systems biology approach using data originating from ChIP-seq combined with other genome-wide interrogation platforms (methylation arrays, gene expression and proteomic data, etc.) in concert with Bayesian statistics could allow separation of the effect on chromatin structure originating from genomic modifications from that caused due to population epigenetic variability and stochastic/environmental effects. Such approach may also capture secondary effects of endogenous

and/or exogenous genomic insertions on distant chromosomal regions through long range chromatin interactions and/or activation of metabolic pathways related to the inserted elements. Importantly, a well-defined, probabilistic model of chromatin structure should have applications beyond GMO detection, allowing comprehensive mapping of the structural and functional elements in the genome and enabling characterization of their changes in response to a variety of internal or external stimuli. These stimuli may include various environmental and/or physiological factors, exposure to which may cause a long-lasting signature on chromatin and DNA structure.

The described approach represents a general methodology for detection and subsequent sequencing of inserted DNA elements into a host genome introduced by conventional or emerging gene editing tools, such as the CRISPR (clustered regularly interspaced short palindromic repeats)/Cas9 (CRISPR-associated protein 9) system, Zinc finger nucleases (ZNFs), or transcription activator-like effector nucleases (TALENs). Although the proposed epigenetics approach is not suitable to pinpoint the use of genetic engineering tools that do not result in incorporation of their DNA sequences into the host genome, it may allow detection of chromatin pattern changes that result from collateral genome alterations caused by those technologies (*e.g.*, CRISPR/Cas9), such as off-target double stranded breaks and insertion or deletion mutations (indels). In addition, analysis of chromatin structure may shed light on how the epigenetic environment might affect the frequency of such off-target events.

An alternative approach for the detection of GMOs could involve whole genome sequencing as a direct method for interrogation of inserted elements. However, current technologies for whole genome sequencing require long read lengths, resulting in high reagent cost per sample. In addition, coverage requirements in whole genome sequencing limit the number of samples that can be multiplexed in a single sequencing run. Furthermore, the large quantity of data generated by whole genome sequencing increases the required processing resources, complexity and required time for analysis. The need for higher read length, sequence coverage, sample size limitations and volume of data would make this approach less efficient with respect to cost, speed and ease of analysis than the proposed epigenetics-based methodology described in this work. In this study, by focusing on epigenetic marks, we limit the amount of sequencing data required to detect a GMO, considerably reducing the cost of these endeavors. Moreover, the proposed approach provides additional information with respect to the chromatin structure and its activity status resulting from the inserted elements.

Overall, this study provides proof-of-concept for the application of genome-wide interrogation methods for detection of epigenetic signatures of genetic manipulation. Effective detection of GMOs should enable control and security of adversely genetically manipulated organisms and detection of infectious agents (*e.g.* prion, viruses) genetically engineered into zoonotic species, affecting the food chain. It should be noted that the findings of this work should have broader applications not only for GMO detection but also for improving transgene expression. Although transgenic technology holds great promise for advancing our basic understanding of gene expression and regulation as well as practical applications, including production of pharmaceutical proteins and food quality improvement, such efforts are generally hampered due to the fact that transgenes expression is unpredictable and often much lower than anticipated [12]. Transgene expression is associated with regulatory elements and their epigenetic modifications upon transformation. Better understanding of chromatin biology and remodeling upon gene transfer should enable prediction of gene expression and more efficient production of transgenic animals for biomedical and agricultural applications.

5. Conclusion

This study provides a proof-of-concept that epigenetic approaches can be used to detect the insertion of endogenous and exogenous sequences into the genome of higher organisms where the method of genetic modification, the sequence of inserted DNA, and the exact genomic insertion site(s) are unknown. In a broader sense, characterizing chromatin dynamics can be applied beyond genomic manipulation to include organism exposure to environmental or other factors affecting the epigenomic landscape.

Acknowledgments

We thank Dr. Eric B. Haas and Scott D. Stewart for critical reading of the manuscript and useful insights. This work was supported by funding from Leidos Inc. The authors declare no competing financial interests.

References

[1] A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, High-resolution profiling of histone methylations in the human genome, Cell 129 (2007) 823–837.
[2] M. Buh Gasparic, T. Tengs, J.L. La Paz, A. Holst-Jensen, M. Pla, T. Esteve, J. Zel, K. Gruden, Comparison of nine different real-time PCR chemistries for qualitative and quantitative applications in GMO detection, Anal. Bioanal. Chem. 396 (2010) 2023–2029.
[3] M. Bulger, M. Groudine, Looping versus linking: toward a model for long-distance gene activation, Genes Dev. 13 (1999) 2465–2477.
[4] M.A. Corbett, C.S. Robinson, G.F. Dunglison, N. Yang, J.E. Joya, A.W. Stewart, C. Schnell, P.W. Gunning, K.N. North, E.C. Hardeman, A mutation in alpha-tropomyosin(slow) affects muscle strength, maturation and hypertrophy in a mouse model for nemaline myopathy, Hum. Mol. Genet. 10 (2001) 317–328.
[5] M. Chaouachi, G. Chupeau, A. Berard, H. McKhann, M. Romaniuk, S. Giancola, V. Laval, Y. Bertheau, D. Brunel, A high-throughput multiplex method adapted for GMO detection, J. Agric. Food Chem. 56 (2008) 11596–11606.
[6] D. D'Angelo, Y. Sakata, J. Lorenz, G. Boivin, R. Walsh, S. Liggett, G. Dorn, Transgenic Gαq overexpression induces cardiac contractile failure in mice, Proc. Natl. Acad. Sci. USA 94 (1997) 8121–8126.
[7] F. DeLeo, F. DeLeo, Traceability and detection of genetically modified organisms in the labelling of food production chain, in: Proceedings of the 9th ICABR International Conference on Agricultural Biotechnology (2005).
[8] E. Dugat-Bony, E. Peyretaillade, N. Parisot, C. Biderre-Petit, F. Jaziri, D. Hill, S. Rimour, P. Peyret, Detecting unknown sequences with DNA microarrays: explorative probe design strategies, Environ. Microbiol. 14 (2012) 356–371.
[9] D.S. Elenis, D.P. Kalogianni, K. Glynou, P.C. Ioannou, T.K. Christopoulos, Advances in molecular techniques for the detection and quantification of genetically modified organisms, Anal. Bioanal. Chem. 392 (2008) 347–354.
[10] B. Enkhmandakh, D. Bayarsaihan, Chromatin topology and long-range genomic interactions, J. Regen. Med. 2 (2013) 2.
[11] G.C. Hon, R.D. Hawkins, B. Ren, Predictive chromatin signatures in the mammalian genome, Hum. Mol. Genet. 18 (R2) (2009) R195–R201.
[12] L.M. Houdebine, The methods to generate transgenic animals and to control transgene expression, J. Biotechnol. 98 (2002) 145–160.
[13] Y. Ji, Y. Shi, G. Ding, Y. Li, A new strategy for better genome assembly from very short reads, BMC Bioinform. 12 (2011) 493.
[14] 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, M. Snyder, ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia, Genome Res. 22 (2012) 1813–1831.
[15] D. Lee, M. La Mura, T.R. Allnutt, W. Powell, Detection of genetically modified organisms (GMOs) using isothermal amplification of target DNA sequences, BMC Biotechnol. 9 (2009) 7, http://dx.doi.org/10.1186/1472-6750-9-7.
[16] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079.
[17] N. Marmiroli, E. Maestri, M. Gullì, A. Malcevschi, C. Peano, R. Bordoni, G. De Bellis, Methods for detection of GMOs in food and feed, Anal. Bioanal. Chem. 392 (2008) 369–384.
[18] E. Michelini, P. Simoni, L. Cevenini, L. Mezzanotte, A. Roda, New trends in bioanalytical tools for the detection of genetically modified organisms: an

update, Anal. Bioanal. Chem. 392 (2008) 355–367.
[19] D. Morisset, T. Demsar, K. Gruden, J. Vojvoda, D. Stebih, J. Zel, Detection of genetically modified organisms – closing the gaps, Nat. Biotechnol. 27 (2009) 700–701.
[20] P.J. Park, ChIP-seq: advantages and challenges of a maturing technology, Nat. Rev. Genet. 10 (2009) 669–680.
[21] K. Tachibana, Y. Kobayashi, T. Tanaka, M. Tagami, A. Sugiyama, T. Katayama, C. Ueda, D. Yamasaki, K. Ishimoto, M. Sumitomo, Y. Uchiyama, T. Kohro, J. Sakai, T. Hamakubo, T. Kodama, T. Doi, Gene expression profiling of potential peroxisome proliferator-activated receptor (PPAR) target genes in human hepatoblastoma cell lines inducibly expressing different PPAR isoforms, Nucl. Recept. 3 (2005) 3.
[22] M. Querci, M. Van den Bulcke, J. Zel, G. Van den Eede, H. Broll, New approaches in GMO detection, Anal. Bioanal. Chem. 396 (2010) 1991–2002.
[23] T. Vavouri, B. Lehner, Human genes with CpG island promoters have a distinct transcription-associated chromatin organization, Genome Biol. 13 (11) (2012) R110.
[24] C. Verma, S. Nanda, R.K. Singh, R.B. Singh, S. Mishra, A review on impacts of genetically modified food on human health, Open Nutraceuticals J. 4 (2011) 3–11.
[25] M.B. Wheeler, Production of transgenic livestock: promise fulfilled, J. Anim. Sci. 81 (Suppl. 3) (2003) S32–S37.
[26] C.L. Wu, S.C. Kandarian, R.W. Jackman, Identification of genes that elicit disuse muscle atrophy via the transcription factors p50 and Bcl-3, PLoS One 6 (1) (2011) e16171.
[27] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li, X.S. Liu, Model-based analysis of ChIP-seq (MACS), Genome Biol. 9 (9) (2008) R137.
[28] X. Zhou, D. Xing, Y. Tang, W.R. Chen, PCR-free detection of genetically modified organisms using magnetic capture technology and fluorescence cross-correlation spectroscopy, PLoS One 4 (2009) e8074.
[29] R.K. Arthur, L. Ma, M. Slattery, R.F. Spokony, A. Ostapenko, N. Negre, K.P. White, Evolution of H3K27me3-marked chromatin is linked to gene expression evolution and to patterns of gene duplication and diversification, Genome Res. 24 (2014) 1115–1124.
[30] N.C. Riddle, A. Minoda, P.V. Kharchenko, A.A. Alekseyenko, Y.B. Schwartz, M.Y. Tolstorukov, A.A. Gorchakov, J.D. Jaffe, C. Kennedy, D. Linder-Basso, S.E. Peach, G. Shanower, H. Zheng, M.I. Kuroda, V. Pirrotta, P.J. Park, S.C.R. Elgin, G.H. Karpen, Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin, Genome Res. 21 (2010) 147–163.
[31] A. Vielle, J. Lang, Y. Dong, S. Ercan, C. Kotwaliwale, A. Rechtsteiner, A. Appert, Q.B. Chen, A. Dose, T. Egelhofer, H. Kimura, P. Stempor, A. Dernburg, J.D. Lieb, S. Strome, J. Ahringer, H4K20me1 Contributes to Downregulation of X-Linked Genes for C. Elegans Dosage Compensation, PLoS Genet. 8 (9) (2012), e1002933.
[32] Overbeek PA. Direct Data Submission for Overbeek Lentiviral Transgenic Lines MGI Direct Data Submission, 2011. [MGI Ref ID J:175597].