Published in final edited form as:

Nat Methods. 2014 June ; 11(6): 689–694. doi:10.1038/nmeth.2924.

Multiscale Representation of Genomic Signals

Theo A. Knijnenburg[1], **Stephen A. Ramsey**[2,3], **Benjamin P. Berman**[4], **Kathleen A. Kennedy**[2], **Arian F.A. Smit**[1], **Lodewyk F.A. Wessels**[5,6], **Peter W. Laird**[4], **Alan Aderem**[2], and **Ilya Shmulevich**[1]

[1]Institute for Systems Biology, Seattle, Washington, USA [2]Seattle Biomedical Research Institute, Seattle, Washington, USA [4]University of Southern California Epigenome Center, University of Southern California, Keck School of Medicine, Los Angeles, California, USA [5]Division of Molecular Carcinogenesis, Netherlands Cancer Institute, Amsterdam, The Netherlands [6]Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

Abstract

Genomic information is encoded on a wide range of distance scales, ranging from tens of base pairs to megabases. We developed a multiscale framework to analyze and visualize the information content of genomic signals. Different types of signals, such as GC content or DNA methylation, are characterized by distinct patterns of signal enrichment or depletion across scales spanning several orders of magnitude. These patterns are associated with a variety of genomic annotations, including genes, nuclear lamina associated domains, and repeat elements. By integrating the information across all scales, as compared to using any single scale, we demonstrate improved prediction of gene expression from Polymerase II chromatin immunoprecipitation sequencing (ChIP-seq) measurements and we observed that gene expression differences in colorectal cancer are not most strongly related to gene body methylation, but rather to methylation patterns that extend beyond the single-gene scale.

Introduction

In mammalian genomes, information is encoded on a wide range of scales, ranging from 10–100 bases (transcription factor binding sites, microsatellites, exons), to kilobases (CpG islands, genes), to megabases (nuclear lamina associated domains (LADs), heterochromatin). Such information can be detected in patterns in both the genome sequence

Corresponding author: Ilya Shmulevich [ilya.shmulevich@systemsbiology.org].

[3]Current address: Department of Biomedical Sciences, Oregon State University, Corvallis, Oregon, USA

AUTHOR CONTRIBUTIONS: T.A.K. and I.S. conceived the idea for this work. T.A.K. performed the computational experiments. T.A.K., S.A.R. and I.S. analyzed the data. K.A.K. developed methods and protocols for, supervised, and aided in the production of ChIP-Seq data. B.P.B and P.W.L. provided the DNA methylation data and contributed to its analysis and report. A.F.A.S. analyzed and reported on the correlations between the MSR of GC content and repeat elements. L.F.A.W., A.A. and I.S. provided supervision during this project. T.A.K, S.A.R, and I.S. wrote the manuscript. All authors commented on the manuscript.

COMPETING FINANCIAL INTERESTS: The authors declare no competing financial interests.

and the epigenetic state of cells, and these patterns can be represented as quantitative functions of genomic position, or *genomic signals*. Examples of such signals include interspecies sequence conservation, GC content, genome annotations, and (epi-)genomic measurements from RNA sequencing (RNA-seq)[1], chromatin immunoprecipitation sequencing (ChIP-seq)[2] or bisulfite sequencing[3]. In this work, we present a computational approach for systematically and simultaneously interrogating these genomic signals across all length scales.

We are witnessing an explosion in the acquisition of genomic data, but extracting biological insights through integrated analysis of heterogeneous genome assays has proved challenging[4]. Genomic signal analysis methods often filter the data through a sliding window of fixed length scale, necessitating the preselection of a particular length scale for the analysis[5,6]. Analysis methods that are designed to identify localized peaks, such as transcription factor (TF) binding sites, have limited applicability for identifying features that span larger genomic regions, such as epigenetic marks, and RNA polymerase II[7,8]. Hidden Markov Models (HMMs) have also been used to segment genomic signals[9-11], but their use is complicated by the need to predefine the number and type of states. The fundamental challenge of analyzing heterogeneous genomic signals is that it is not known *a priori* which distance scales are the most relevant to a given genomic signal or to a given biological question.

To address this challenge, we have developed the multiscale signal representation (MSR) method, which is adapted from an image segmentation algorithm[12] and inspired by multiscale approaches for classifying image texture patterns[13]. Multiscale techniques have previously been applied to several types of biological data, including insertional mutagenesis data[14], copy number variation data[15], epigenomic data and DNA replication timing domains[16]. The MSR generalizes these approaches by providing information about genomic signal enrichment or depletion at *all* genomic distance scales. The method divides the genome into hierarchically organized segments whose sizes range from basepairs to megabases. The segments are scored for enrichment or depletion of genomic signal intensity. Besides its use in summarizing and visualizing the information content of genomic signals across spatial scales, the MSR presents a novel and powerful way to unravel the biological function of these signals.

Results

Building the Multiscale Representation

In the MSR approach, the genomic signal values are smoothed and then used as a basis for dividing the chromosome into segments (on a succession of increasing length scales), which are then tested for enrichment or depletion of signal intensity. The four steps of the method are (Fig. 1 and **Methods**):

1. Smooth the genomic signal to create the scale space (Fig. 1a).

   The genomic signal is convolved with Gaussian windows of various widths, i.e., length scales. The resulting set of convolved signals at each of the length scales can be described as a Gaussian scale space[17].

2. Create the segmentation tree (Fig. 1b).

   A set of positions in the genomic signal is selected as starting nodes of the *segmentation tree*, which is created by propagating these nodes from the smallest scale in the scale space to the largest scale. This propagation procedure forces branches to follow a signal intensity isoline in the scale space, while ensuring that branches merge, ultimately converging to a root node.

3. Segmenting the genomic signal on multiple scales (Fig. 1c)

   Based on the segmentation tree, the genomic signal is partitioned into segments at successive scales. At scale *n*, each node *i* is mapped to a genomic segment by following the outermost branches originating from that node to the leaf nodes at the smallest scale. The locations where these outermost branches are found on the smallest scale are the boundaries of the segment corresponding to (*n*,*i*). This procedure is carried out for every scale, producing the *multiscale segmentation* of the signal.

4. Scoring the segments (Fig. 1d)

   Segments are assessed for depletion or enrichment of signal intensity using the Significant Fold Change (SFC), a score that combines both the statistical significance and the magnitude of the difference between the variables being compared. The SFC is positive or negative (corresponding to the observed intensity being larger or smaller than expected) in the case where the confidence threshold is met, but is defined as zero otherwise. Importantly, SFC scores can be compared between different scales, i.e., between segments with widely differing sizes.

In summary, the MSR of a genomic signal is a collection of segmentations of the signal at different spatial scales. Each segment in a scale-specific segmentation is scored for signal enrichment or depletion. We used 50 scales, which ensured for all our genomic signals that the largest scale contained only one segment spanning the entire chromosome.

Genomic Signals Distinguished by Multiscale Fingerprints

In order to investigate its ability to reveal patterns of signal enrichment and depletion on diverse distance scales, the MSR was applied to a variety of mouse-derived genomic signals including GC content, interspecies sequence conservation scores, and ChIP-seq data for six chromatin-associated proteins in primary mouse macrophages: three transcription factors (ATF3, NFκB-p50 and NFκB-p65), RNA polymerase II (Pol II) and two covalent histone modifications; acetylation of histone H4 (H4ac) and trimethylation of histone H3 at lysine 27 (H3K27me3) (Fig. 2). As expected for TFs, which bind specific focal locations in the genome, segments for TF ChIP-seq genomic signals were only enriched at smaller scales, with segment sizes between 100 bp and 1 kb (upper-right panel). In contrast, Pol II is enriched across a much broader range of scales, consistent with its function of translocating along the gene in the process of transcribing it. Given that the interquartile range of gene sizes in the mouse genome is 7–46 kb, many large segments would be expected. Large segments are also identified for the chromatin marks H4ac and H3K27me3, consistent with previous observations[7]. In contrast to ChIP-derived genomic signals (which are essentially

convolved over the length-scale of the immunoprecipitated fragment size, ∼100-200 bp), sequence conservation is enriched at even the smallest scales (segment sizes < 100 bp). This reflects the basepair resolution of this signal and the existence of very small, but highly conserved genomic regions. While ChIP- and conservation-derived signals show substantial dynamic range (vertical pattern size in the heatmap), GC content, which ranges between 0 and 1 with a genome-wide average of 0.41, has relatively small fold-changes between observed and expected signal intensity.

Because the segments at different distance scales are hierarchically related in the MSR, there is a natural way to prune the segment hierarchy to automatically select the scale(s) at which a genomic region is most highly enriched or depleted in signal. This pruning operation provides for enriched segment detection, or "peak calling", in a multiscale context (Supplementary Note 1). We compared the MSR-based pruning approach with the widely used peak caller MACS[18] and with SICER[8] (which detects broad histone modifications), on the ChIP-seq data described above. Consistent with the scale-versatility of the method, we found that the detected peaks of the pruned MSR cover a larger range of length scales than the other peak-calling methods (Supplementary Fig. 1, 2, and Note 2, 3). It can be argued that the MSR segments more faithfully mark enriched regions of appropriate sizes for the various interacting biomolecules. For example, the Pol II MSR segments agree much better with gene lengths than the segments found by either MACS or SICER.

Consistent with our principal goal of investigating the ability of the MSR to enable multiscale comparisons between heterogeneous genomic signals (rather than the specific data reduction step of peak-calling), we used the "unpruned" MSR for all subsequent analyses.

Enriched Segments Overlap with Functional Genomic Regions

We next investigated the relationship between the MSR of the mouse-derived genomic signals and genomic annotations, such as genes, exons, LADs and repetitive elements. We computed the overlap between these annotated regions and the enriched segments (SFC>0) at each scale. Based on the expected overlap computed using a background model in which the annotated regions were randomly placed throughout the genome, we scored the degree of overlap between a genomic signal and a functional genomic element (see **Methods**). By visualizing the overlap scores between the various genomic signals and the various annotated genomic regions as heat maps (Fig. 3, and Supplementary Figs. 3-10), we were able to validate many known associations. For example, both small and large genomic segments that were enriched in sequence conservation tended to overlap with genes (Fig. 3a), and exons exhibited the strongest levels of overlap. On the other hand, LADs (typically 0.1–10 Mb in size) showed a relative depletion of large conserved segments, reflecting the relative paucity of genes in LADs[19].

The MSR method can detect cases where minimal overlap between two genomic signals is observed at one scale while a large overlap is observed at another, such as when comparing SINE (short interspersed elements) repeats with GC content (Fig. 3b). There is significant overlap of SINE repeats with GC-enriched segments with approximate sizes from 10 kb to 10 Mb (scale 20 to 40), consistent with longstanding observations that SINEs are enriched in

large-scale GC rich genomic regions in mammals[20,21]. Surprisingly, smaller GC-rich regions (scale 10–15, segment sizes around 1 kb) do not seem to overlap with the SINE repeats at all. Subsequent analysis revealed that these small GC-rich regions tend to overlap with CpG islands around transcription start sites (Supplementary Note 4). Therefore, a possible explanation of the paucity of SINEs within small GC-rich regions is that the integration of SINEs within the functionally important CpG islands is under negative selection pressure.

These results demonstrate that the associations between annotated genomic regions and genomic signals depend on the scales at which the genomic signals are analyzed, consistent with the fact that genomic signals contain distinct information at different scales. Moreover, correlations between genomic signals themselves also depend on the length scale. For example, we observed differential correlation, i.e., a lack of overlap at small scales and a significant excess of overlap at large scales, between the histone marks H3K27me3 and H4ac (Supplementary Note 5 and Supplementary Fig. 11). This finding likely reflects the mutually exclusive functions of the repressive mark H3K27me3 and the activating mark H4ac on the scale of genes (10 kb), as well as the fact that both chromatin marks are primarily found in gene-rich regions, which is indicated by the positive correlation at the scale of gene islands and deserts (1 Mb).

Multiscale Signatures Predict Gene Expression

Next, we investigated the multiscale information content of genomic signals by integrating them into a model for predicting gene expression. Recent models for predicting gene expression from genomic signals (mainly histone modifications[22-24]) have used a single-scale approach (e.g., a 4 kb bin around the transcription start site (TSS)), to represent the genomic signal data. We created gene-specific MSRs using both the mouse macrophage ChIP-seq data and genomic signals from ENCODE[4] and applied a Random Forest regression algorithm[25] to predict gene expression (see **Methods**). Within this model, using the MSR-derived features improved prediction accuracy and the analysis revealed the variety of distance scales on which specific features have predictive capacity in the model (Supplementary Fig. 12, 13 and Note 6).

Focal and Broad DNA Methylation Linked to Gene Silencing

To evaluate the generality of the MSR we analyzed global DNA methylation changes in human colorectal cancer vs. normal tissue. In tumor cells many CpG islands (which are typically 300-3000 bp in size), often near TSSs, are hypermethylated, whereas the tumor genome as a whole is hypomethylated[26,27]. We recently showed that these two modifications are not independent – in a whole-genome bisulfite-sequencing (WGBS) study of a colon tumor and matched normal mucosa, we found that focal hypermethylation was enriched within large blocks of hypomethylation[28]. Like other similar works[3,29,30] the identification of changes in DNA methylation at such different scales required the prior determination of scales of interest, followed by independent analysis with each of these different scales.

In this work, we applied the MSR method to the colon cancer WGBS dataset, which confirmed the clear scale-dependent differential methylation patterns (Fig. 4a). Additionally, we found substantial concordance between the segments identified by the MSR and the focal and broad regions detected in ref[28] (Supplementary Fig. 14 and Note 7). This demonstrates that a multiscale approach is capable of identifying regions that fixed-scale approaches can only identify when the scale(s) of interest are known in advance.

DNA methylation is important in organism development and plays a central role in oncogenesis[27]. Focal DNA hypermethylation near promoters is associated with the silencing of both developmental genes[31], and, in cancer, of tumor suppressor genes[32]. Hypomethylation, on the other hand, is associated with both gene activation and repression[33]. Based on reports linking hypomethylated domains to chromatin silencing[33] and to nuclear organization of chromatin (via nuclear-lamina associated domains)[28], we hypothesized that the association between gene expression and DNA methylation might be a function of scale.

To test this hypothesis we compared differential gene expression between the cancer and normal tissues with the differential DNA methylation (as captured by the MSR) around the TSS of genes (Fig. 4b and Supplementary Note 8). Consistent with current understanding of an inverse correlation between promoter methylation and expression[28], we found that the 166 genes that are strongly upregulated in the tumor are significantly ($P<10^{-4}$) depleted for small-scale hypermethylation, and the 186 strongly downregulated genes are significantly ($P<10^{-4}$) enriched for small-scale hypermethylation. In contrast, the moderately up- and downregulated genes show an unexpected pattern – differential methylation occurs across scales including the large scales, which extend far beyond the size of individual genes. Particularly, the 2503 moderately upregulated genes were enriched for hypermethylation at large scales and hypomethylation at small scales. This analysis clearly demonstrates the scale-dependent relationship between DNA methylation around the TSS and gene expression in cancer.

We next explored the role of cancer-associated methylation changes within gene bodies. Previous studies have reported a positive association between methylation state at gene bodies and expression level[34,35], while complete methylome studies have shown that expression corresponds more strongly to LADs, which do not always correspond to gene boundaries[28,33]. In this work, we repeated the MSR analysis focused on the methylation pattern at the middle of genes (GM) rather than the promoter (Fig. 4c). For the moderately differentially expressed genes, we observed the same pattern as for the TSS analysis when investigating the larger scales, i.e., segments larger than 100 kb. This is unsurprising, since at these large scales the segments are far beyond the individual gene scale. On the smaller scales, where the segments are smaller than 10 kb, there is no pronounced pattern.

It was not clear whether the observed expression differences were affected specifically by gene body methylation, or whether they were due to methylation patterns at super-genic scales. To investigate this, a Random Forest regression model was trained to predict the differential gene expression using three different feature sets based on the DNA methylation data at multiple scales: 1) focal DNA methylation around the TSS, 2) DNA methylation

within the gene body, and 3) longer-range DNA methylation (at scales larger than 100 kb). We found that gene body methylation does not offer predictive value in explaining expression differences when TSS and long-range methylation are already considered in the predictive model (Supplementary Fig. 15).

Our findings suggest that gene expression is related to methylation patterns that extend beyond the single-gene scale, and further, that the elevated gene body DNA methylation in expressed genes may be attributable to long-range DNA methylation patterns extending beyond the transcribed region (rather than to gene body transcription, as has been suggested for the histone mark H3K36me3[34]).

Discussion

We developed the multiscale signal representation (MSR) framework in order to overcome the limitations inherent to analysis methods based on a fixed genomic length scale. The MSR enables global analysis of genomic data in an unbiased manner with respect to the spatial scales on which biological information is encoded. We used the MSR to analyze measurements of transcription factor binding, covalent histone modifications and DNA methylation, as well as genomic annotations and sequence-derived data such as conservation and GC content.

Multiscale analysis of the macrophage cistrome and epigenome and a colon tumor's DNA methylome revealed distinct patterns of signal enrichment and depletion at different genomic length scales. Moreover, integrating multiscale information in a model to predict gene expression levels showed that understanding the epigenetic basis of gene regulation requires analyzing genomic signals at multiple scales and it provides clues about the interaction between the ChIP target and the gene.

Genome-wide profiling studies (e.g. ENCODE[4]) will undoubtedly broaden the limited gene-centric view of the genome and gene regulation that prevailed in the past, but leveraging such genomic signals to reveal the biological mechanisms and functions of genomic events that occur at a scale much smaller or larger than the scale of genes presents both conceptual and analytical challenges. The MSR approach presented here is a computational tool specifically designed to address these challenges.

The MSR is complementary to other segmentation techniques, such as Segway[11] and ChromHMM[9]: whereas the MSR segments *one* genomic signal at *multiple* scales, these HMM based models use *multiple* genomic signals to provide *one* segmentation, thereby dividing the genome into a number of functional states called chromatin domains. It is an open question whether the MSRs are specific for some of these functional states.

One limitation of the MSR approach is the idealization of the chromosome as a linear entity, i.e., that the physical distance between two genomic locations is related to their relative base pair positions. Future work will be aimed at addressing this limitation, for example, by incorporating contact probability maps between genomic positions[36] into the MSR framework.

Notwithstanding the one-dimensional assumption, the MSR simultaneously solves two important bottlenecks in the biological interpretation of heterogeneous genomic datasets, the problems of (1) *a priori* window size (scale) selection and (2) detecting associations between genomic signals encoding information on disparate genomic scales. As the body of published (epi-)genomic data continues to grow exponentially, the use of such an unbiased analysis method will be increasingly important.

Online Methods

Multiscale Segmentation

The multiscale segmentation algorithm for genomic signals is based on the multiscale image segmentation algorithm described by Vincken *et al*.[12] Here, we describe the algorithm.

We start with a genomic signal $\boldsymbol{x}(i)$, a function of equally spaced genomic positions. A scale space with S scales is created. The first scale in the scale space (s=1) is the signal itself. The subsequent scales are obtained by convolving $\boldsymbol{x}(i)$ with a Gaussian window of increasing width. At smaller scales, the signal is only slightly smoothed, but at increasingly larger scales, the signal is smoothed over greater genomic distances. The standard deviation of this window for scale s is defined to be $\delta_s=\exp((s\text{-}1)\tfrac{1}{2}\ln 2)$. We denote the convolved signal at scale s by $\mathbf{x}_s(i)$.

Next, we choose a set of starting positions, which will serve as starting nodes of the segmentation tree. To constrain computational complexity, not all positions of the genomic signal are selected, but only those for which there is a differential signal intensity, i.e. $\mathbf{x}(i) \neq \mathbf{x}(i+1)$ or $\mathbf{x}(i) \neq \mathbf{x}(i+1)$. Starting nodes are also placed at the first and last position of the signal. The number of starting nodes is on the order of 100,000 for most of the genomic signals analyzed in this work.

A bottom-up linking process is employed to link nodes between two adjacent scales. Basically, for each starting node at scale 1, the best successor node at scale 2 is determined. This process is repeated in an iterative fashion to find the best successor node at scale s based on the nodes at scale s-1. All genomic positions within a window of size r_s centered on the genomic position of the parent node are considered as a potential successor node. The window size is defined as follows.

$$r_s = \sqrt{2\left(\delta_s^2 - \delta_{s-1}^2\right)} \quad (1)$$

Two criteria are considered for each potential successor node: 1) the absolute intensity difference between the parent node at position a on scale s-1 and the potential successor node at position b on scale s, i.e. $|\mathbf{x}_{s\text{-}1}(a)\text{-}\mathbf{x}_s(b)|$, should be as small as possible and 2) the ground volume, which is defined as the number of nodes to which the potential successor node would be connected at scale 1, i.e. the number of starting nodes, should be as large as possible. The two criteria were equally weighted in computing the 'affinity' score between a node and its potential successor. (These two linkage criteria, intensity difference and ground volume, are given in eq. (10) and (11) of Vincken *et al.*[12].) Of all potential successor nodes

in the search volume, the one with the highest 'affinity' score was selected. Due to the two criteria, the branches (connected nodes) have two properties: 1) they tend to follow intensity isolines in the scale space, and 2) they tend to merge with one another, creating a tree. This tree is called the segmentation tree. Thus, the ground volume criterion provides tension countering the intensity difference criterion and ensures that nodes merge to ever fewer nodes, ultimately converging to a root node. In our experiments we set S, the total number of scale, to 50. For the genomic signals used in this work, all branches were found to be merged at the largest scale (50), i.e. there is only one node at the largest scale. In essence, the segmentation tree traces how genomic features, which are represented by differential intensities in the genomic signal, propagate, spread out, and are ultimately combined, when the data are smoothed at increasing distance scales.

The segmentation of a genomic signal is derived from the segmentation tree. The number of nodes at scale s, say n_s, is equal to the number of segments at this scale. For each node at scale s we select the node at scale 1 with the smallest genomic position that is connected to this node at scale s by following the left most branch of the tree down to scale 1. (See Fig. 1b.) This results in n_s genomic positions denoted by t_i with $i = 1,\ldots,n_s$. The boundaries of segment i at scale s are given by $[t_i,t_{i+1}\text{-}1]$, thereby dividing the complete genomic signal at scale s into n_s segments.

Construction of the genomic signals

Most genomic signals analyzed in this work were based on ChIP-seq data of six proteins in primary murine bone marrow macrophage cells (BMMs) under unstimulated and lipopolysaccharide (LPS) stimulated conditions. The BMMs were cultured from female C57BL/6 mice (age 8-12 weeks). Amongst these six proteins were three transcription factors (TFs), ATF3[37], NFκB/p50 and NFκB/p65 (or p50 and p65 for short)[38], all of which are involved in regulating macrophage activation by microbial molecular components such as LPS. The other three ChIP-seq targets were RNA polymerase II (Pol II), and two chromatin modification marks: acetylation of histone H4 (H4ac) and tri-methylation of histone H3 lysine 27 (H3K27me3). A control ChIP-seq signal was also obtained from three IPs of BMMs with immunoglobulin G derived from rabbits that were not immunized with specific target antigens.

To create a genomic signal, the (uniquely mappable) 35 bp reads from sequencing the ends of ChIP fragments were aligned to the mouse genome (version: NCBIM37/mm9) and extended to a length of 158 bp, the average length of the sequenced fragments. Then, the extended reads were aggregated and the signal was downsampled to a resolution of 10 bp. Aggregation was accomplished by adding up the overlapping 158 bp reads along the chromosome. Downsampling of the signal was achieved using simple subsampling with a downsampling factor of 10, i.e. picking out every 10^{th} data point. Each ChIP-seq experiment leads to 21 genomic signals, one for each chromosome (19 autosomes and two sex chromosomes). For chromosome 1, the largest chromosome in the mouse, which is about 200 Mb long, the corresponding genomic signal consists of 20 million samples. In Fig. 1 a small part of the genomic signal of chromosome 11 for one of the Pol II ChIP-seq

experiments is depicted. Additional experimental protocols including antibody suppliers and computational details are given in the Supplementary Note 9.

A genome-wide unique mappability map of the mouse genome was created by mapping all ∼3 billion 35 bp sequences of the mouse genome to itself using SOAP 2.0[39], discarding cases, where the mapping was not unique. Similarly to the ChIP-seq data, the uniquely mappable reads were extended, aggregated and downsampled to a resolution of 10 bp. The maximum value of this signal was 316 (158×2). In that case, the genomic position is surrounded on the left side by 158 bp to which 35 bp could be mapped uniquely in the forward (5′ to 3′) direction, and similarly on the right side for mapping in the reverse direction (3′ to 5′). This unique mappability landscape was used in determining enrichment or depletion of segments (as explained below).

Conservation scores for alignments of 29 vertebrate genomes with mouse (based on the phylogenetic hidden Markov model developed by Siepel *et al.*[40]) were downloaded from the UCSC Genome Browser[41] (genome version: NCBIM37/mm9). These scores are real numbers between 0 and 1 or no number is reported in case of alignment gaps. In the genomic signal based on these conservation scores the alignment gaps were set to have a value of zero. A binary background signal (similar the genome-wide unique mappability map) was created by setting all alignment gaps to zero, and all positions for which a conservation score was reported to one. Both the conservation genomic signal and its background signal were downsampled to a resolution of 10 bp, after they were smoothed by convolution with a rectangular window of 10 bp.

GC content scores for the mouse genome were downloaded from the UCSC UCSC Genome Browser[41] (genome version: NCBIM37/mm9). The GC content is a binary signal, which is one when the basepair is a “G” or a “C”, and zero otherwise. The corresponding background signal is also a binary signal, which is zero for undetermined basepairs (“N”), and one otherwise. Again, both signals were downsampled to a resolution of 10 bp after convolution with a 10 bp rectangular window.

Significant Fold Change

The MSR of a genomic signal contains segments ranging from basepairs to megabases, i.e. the segment lengths span several orders of magnitude. This frustrates scoring the segments for enrichment or depletion using standard enrichment tests (such as the hypergeometric test or the Z-test[42,43]), since the range of P-values output by such procedures is heavily influenced by the segment lengths[44]. Specifically, the segment length corresponds to the number of samples in the statistical test. Thus, larger segments will lead to much lower P-values when the null hypothesis is not true as the larger number of samples leads to a higher confidence in rejecting the null hypothesis. Moreover, the P-value does not directly provide information on the effect size, which is biologically more relevant than the statistical significance reported by the P-value[45,46].

Here, we introduce the significant fold change (SFC), a statistical score that combines the P-value and effect size, to score the enrichment and depletion of segments in the MSR. The SFC is basically the effect size that is significant at a predefined level of statistical

confidence (P-value). It can be interpreted as the lower bound on the effect size. In this work, the SFC computation is based on the Z-test. However, the SFC can also be derived from other statistical tests, such as the t-test and the hypergeometric test.

The observed intensity of a segment is simply the summed signal intensity in the segment, denoted by X. The expected intensity of a segment follows a normal distribution $N(np, np(1-p))$ with $p=I/B$. Here, I and B are the total summed signal intensity and total background signal intensity across the complete genomic signal (i.e. across the complete chromosome) respectively, and n is the summed intensity of the background signal of the segment under investigation. As explained above, for each type of genomic signal there is a background signal. For the ChIP-seq signals this is the unique mappability map. The genomic signal is per definition never larger than the background signal. As such, the background signal represents the potential signal and accommodates normalization. The normal approximation of the expected intensity is based on the central limit theorem under the null assumption that segments consisted out of randomly selected positions in the genomic signals[42,43]. We empirically tested this assumption by creating random segments in this manner and inspecting the Q-Q plots. For segments with $n<10$ the central limit theorem does not sufficiently hold. In that case, we inflate the variance by setting $n=10$ in the variance term of the normal distribution (not in the mean term) to avoid spuriously significant results for very small segments.

Next, we picked a P-value threshold p^{th}. All analyses except the gene-specific MSRs (see below) have been performed with $p^{th}=10^{-6}$. The P-value threshold is converted to a Z-score using the inverse error function; $Z^{th}=-\sqrt{2}\cdot\mathrm{erf}^{-1}(2p^{th})$, i.e. $Z^{th}=-4.75$ in this work. Then we solve the equations 2 and 3 (below) to find p^e and p^d. Here $n^*=\max(n,10)$.

$$Z^{th}=\frac{X-np^e}{\sqrt{n^*p^e(1-p^e)}} \quad (2)$$

$$-Z^{th}=\frac{X-np^d}{\sqrt{n^*p^d(1-p^d)}} \quad (3)$$

Thus, the value of p^e defines a normal distribution with mean np^e (and variance $n^*p^e(1-p^e)$), for which the intensities equal or greater than observed intensity X are expected with probability p^{th}, the P-value. Similarly, the value of p^d defines a normal distribution with mean np^d, for which the intensities equal or smaller than observed intensity X are expected with probability p^{th}. Given that np is the expected mean background intensity of the segment, if $np<np^e$, there is significant enrichment, i.e. SFC > 0, and if $np^d<np$, there is significant depletion, i.e. SFC < 0. Specifically, the SFC is computed as follows:

$$\mathrm{SFC}=\begin{cases}\log_2\left(\frac{np^e}{np}\right) & if \quad np<np^e \quad \text{Enrichment (SFC}>0)\\ 0 & if \quad np^e\le np\le np^d \quad \text{No significant difference}\\ \log_2\left(\frac{np^d}{np}\right) & if \quad np^d<np \quad \text{Depletion(SFC}<0)\end{cases} \quad (4)$$

Supplementary Fig. 16 presents a visual explanation of the SFC computation.

The control ChIP-seq signal was employed to avoid spurious enrichment: For each segment enriched in the signal, i.e. SFC > 0, the corresponding SFC was also computed for the same segment using the control ChIP-seq signal. If the segment was also enriched in the signal for the control IP, the SFC of the target signal was set to 0.

Two additional notes on the SFC: First, in the case that very large segments are enriched, i.e. segments that cover a significant portion of the chromosome, there must also be large segments that are depleted in the signal. This phenomenon is visible in Fig. 2 by the large scale enrichment and depletion of Pol II and the histone modifications beyond scale 30. This phenomenon only occurs when a segment is assessed for enrichment within the genomic signal, but not when the segment is compared to the segment of another genomic signal as in Eq. 5. Second, although in Fig. 2 no signals show depletion at small scales, this is not a property of the SFC, but a property of the genomic signals themselves. For example, the inverse of a peaky signal (such as the inverse of a TF ChIP-seq signal), which is characterized by sharp and deep valleys, would show depletion on small scales.

Average run time for a genome wide signal

The average run time to compute the MSR (segmentation and SFC) for all chromosomal signals in a genome is about 2 hours using 8 cores (Intel(R) Xeon(R) CPU X5472 3.00GHz) with the default parameters. The algorithm has a large memory requirement peaking at about 13GB necessary for the convolution of very large signals. The size of the genome (chromosomes), number of scales, the density of starting positions and the sampling resolution of the signal are the most important parameters that affect computation time and memory requirement.

Score for the overlap between enriched segments and genomic annotation

Genomic annotation was gathered from different data sources for the mouse genome (version: NCBIM37/mm9). Specifically, annotation of genes and exons was obtained from Refseq[47], annotation of LADs (4 types) was obtained from ref[48], annotation of repeat regions (7 types) was obtained from http://repeatmasker.org/species/musMus.html (version 3.2.8)[49], and cytoband annotation was downloaded from UCSC Genome Browser[41] (original data generated by ref[50]).

Each genomic annotation type consists of a set of genomic regions. These genomic regions were compared to the enriched segments of a genomic signal at a particular scale. Specifically, the SFC computation (explained above) was used to assess the degree of overlap between the genomic regions and the enriched segments when compared to the randomly expected overlap. (In the main text, we abstain from using the term ‘SFC’ for this overlap score to avoid confusion.) The parameters to compute this SFC are: I, the total length of the genomic regions, B, the length of the genomic signal, n, total length of the enriched segments, and X, the total length of the overlapping parts of the genomic regions and enriched segments.

Thus, the SFC represents the fold-change ($\log_2$ ratio) between observed and randomly expected overlap statistically significant at $P=10^{-6}$. (Note that all parameters are scaled with the background signal of the corresponding genomic signal to avoid artificial in- or deflation of the SFC.)

Color values in Fig. 3, and Supplementary Figs. 3-11 are SFCs averaged across chromosomes (an SFC is computed per chromosome) and experimental conditions (each genomic signal, e.g. Pol II, is measured under different experimental conditions).

See Supplementary Note 10, where we compare this approach to the standard hypergeometric test and the fold change.

Gene-specific MSRs

A gene-specific MSR is the MSR at the genomic region from 1 kb upstream to 1 kb downstream of the gene normalized to ten scales. Given that scale *s* is the smallest scale at which the region is spanned by one segment, the MSR is sampled at 10 equidistant steps from scale 1 to scale *s* using nearest-neighbor interpolation to create the normalized MSR with ten scales. We decided to use 10 scales based on visual inspection of the gene specific MSRs, where we sought enough resolution to capture the intensity difference from the base pair to the whole-gene scale, however avoiding unnecessary redundancy of two consecutive scales showing the same information.

The P-value threshold for these MSRs is set to $p^{th}=0.5$, such that $Z^{th}=0$ (in Eq. 2 and 3) and the SFC simply becomes the fold change between observed and expected signal intensity. Thus, here we don't use a hypothesis testing framework to find significant segments, but let the machine learning framework ‘decide’ which segments are important for prediction.

In general, the number of scales to use for a MSR is the scale, at which the segmentation tree consists of only one node, corresponding to a segment spanning the entire genomic signal. When generating MSRs specific to particular genomic regions, this number might differ between these regions (mostly as a function of their length). However, the use of a fixed number of scales might be convenient or necessary for certain applications. In order to capture the multiscale information content with sufficient resolution, this number of scales can be set as the mean number of scales necessary to span each of the genomic regions. The interpolation strategy described above can then be used to derive the region specific MSRs with a fixed number of scales.

Random Forest

We employed the Random Forest implementation for MATLAB v0.02 downloaded from http://code.google.com/p/randomforest-matlab/. The Random Forest regression models were run with 5000 trees each and default settings for the other parameters. The reported importance scores represent the mean decrease in accuracy[25].

Differential DNA methylation MSR

Three genomic signals were used to construct the MSR of differential DNA methylation: 1) a CpG binary signal, which is one for CpG sites, and zero otherwise, 2) the tumor

methylation signal, which contains the average methylation of the CpG sites in the tumor sample represented by scores between 0 and 1, and 3) the normal methylation signal, which contains the which contains the average methylation of the CpG sites in the adjacent tissue sample. The signals were downsampled to a resolution of 10 bp.

The multiscale segmentation was performed on the CpG signal. The differential DNA methylation score (DM) is computed as the SFC as explained before, where the total summed signal intensity of a segment (X) is derived from the tumor signal and the background signal is the CpG track. However, instead of a random background model, the normal methylation signal is used. Specifically, we employed Eq. 5 instead of Eq. 4.

$$\mathrm{DM}=\begin{cases} \frac{np^u-Y}{n} & \wedge \quad np^u>Y \\ 0 & \wedge \quad np^u \le Y \le np^l \\ \frac{Y-np^l}{n} & \wedge \quad np^l<Y \end{cases} \quad (5)$$

Here, Y is the total summed signal intensity of the normal methylation signal in the corresponding segment. The differential DNA methylation score (DM) is between -1 and 1, where positive values indicate hypermethylation of the tumor sample with respect to the normal sample, and negative values indicate hypomethylation. DM was computed for three different P-value thresholds p^{th}=10^{-6}, p^{th}=0.05 and p^{th}=0.5. In the latter case (p^{th}=0.5), DM is simply the differential methylation between tumor and normal, since DM becomes (X-Y)/n. Fig. 4a depicts results with p^{th}=10^{-6}, whereas Fig. 4b,c depicts results with p^{th}=0.5. Cross-validation results of the random forest model are given for the three P-value thresholds in Supplementary Table 1.

Code and data

MATLAB code for the multiscale segmentation, SFC computation, visualization and pruning are available through https://github.com/tknijnen/MSR. This code repository also contains a publicly downloadable MATLAB runtime environment that will allow users who do not have access to a full MATLAB installation to use the MSR software.

All BMM ChIP-seq data used in this project can be found under GEO accession number GSE54414, which can be accessed at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54414.

The sequence data and alignments from the DNA methylation study are available under the dbGaP accession code PHS000385. However, we downloaded these data from http://epigenome.usc.edu/publicationdata/berman20101101/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

T.A.K. thanks J. de Ridder, J. de Ruiter and V. Thorsson for helpful discussions. The authors thank G. Glusman and H. Dinh for careful reading and commenting on the manuscript. ChIP-seq data was produced by T. Stolyar, G. S.

Navarro and C. D. Johnson. T.A.K. thanks H. Rovira and L. Amon for technical assistance. The authors thank the anonymous reviewers for their helpful and constructive comments.

Funding: This work was supported by NIH grant U54-AI54253, NIH contract HHSN272200700038C from the National Institute of Allergy and Infectious Diseases, NIH/NIAID award U19AI100627, NIH/NIAID award R01AI025032 to A.A., NIH award K25HL098807 to S.A.R, and NIH award R01HG002939 to A.F.A.S..

References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–628. [PubMed: 18516045]
2. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007; 316:1497–1502. [PubMed: 17540862]
3. Lister R, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315–322. [PubMed: 19829295]
4. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
5. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009; 10:669–680. [PubMed: 19736561]
6. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. Nat Methods. 2009; 6:S22–32. [PubMed: 19844228]
7. Barski A, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007; 129:823–837. [PubMed: 17512414]
8. Zang C, et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009; 25:1952–1958. [PubMed: 19505939]
9. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010; 28:817–825. [PubMed: 20657582]
10. Stadler MB, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011; 480:490–495. [PubMed: 22170606]
11. Hoffman MM, et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012; 9:473–476. [PubMed: 22426492]
12. Vincken KL, Koster ASE, Viergever MA. Probabilistic multiscale image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1997; 19:109–120.
13. Marr, D. Vision: A computational investigation into the human representation and processing of visual information, Henry Holt and Co. Inc.; New York, NY: 1982.
14. de Ridder J, Uren A, Kool J, Reinders M, Wessels L. Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. PLoS Comput Biol. 2006; 2:e166. [PubMed: 17154714]
15. Klijn C, et al. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. Nucleic Acids Res. 2008; 36:e13. [PubMed: 18187509]
16. Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. Identification of higher-order functional domains in the human ENCODE regions. Genome Res. 2007; 17:917–927. [PubMed: 17568007]
17. Lindeberg, T. Scale-space theory in computer vision. Kluwer Academic Print on Demand; 1993.
18. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]
19. Guelen L, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008; 453:948–951. [PubMed: 18463634]
20. Yang S, et al. Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. Genome Res. 2004; 14:517. [PubMed: 15059992]
21. Meunier-Rotival M, Soriano P, Cuny G, Strauss F, Bernardi G. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc Natl Acad Sci USA. 1982; 79:355–359. [PubMed: 6281768]

22. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci USA. 2010; 107:2926–2931. [PubMed: 20133639]
23. Cheng C, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biology. 2011; 12:R15. [PubMed: 21324173]
24. McLeay RC, Lesluyes T, Partida GC, Bailey TL. Genome-wide in silico prediction of gene expression. Bioinformatics. 2012; 28:2789–2796. [PubMed: 22954627]
25. Breiman L. Random forests. Machine learning. 2001; 45:5–32.
26. Ehrlich M. DNA methylation in cancer: too much, but also too little. Oncogene. 2002; 21:5400–5413. [PubMed: 12154403]
27. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012; 13:484–492. [PubMed: 22641018]
28. Berman BP, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet. 2012; 44:40–46. [PubMed: 22120008]
29. Lister R, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature. 2011; 471:68–73. [PubMed: 21289626]
30. Hansen RS, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci USA. 2010; 107:139–144. [PubMed: 19966280]
31. Meissner A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008; 454:766–770. [PubMed: 18600261]
32. Esteller M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene. 2002; 21:5427–5440. [PubMed: 12154405]
33. Hon GC, et al. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome Res. 2012; 22:246–258. [PubMed: 22156296]
34. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics. 2012; 13:484–492.
35. Ball MP, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol. 2009; 27:361–368. [PubMed: 19329998]
36. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289. [PubMed: 19815776]
37. Gilchrist M, et al. Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. Nature. 2006; 441:173–178. [PubMed: 16688168]
38. Hoffmann A, Baltimore D. Circuitry of nuclear factor B signaling. Immunol Rev. 2006; 210:171–186. [PubMed: 16623771]
39. Li R, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009; 25:1966–1967. [PubMed: 19497933]
40. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15:1034–1050. [PubMed: 16024819]
41. Kent WJ, et al. The human genome browser at UCSC. Genome Res. 2002; 12:996–1006. [PubMed: 12045153]
42. Kim SY, Volsky D. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics. 2005; 6:144. [PubMed: 15941488]
43. Knijnenburg T, Wessels L, Reinders M. Creating gene set activity profiles with time-series expression data. International Journal of Bioinformatics Research and Applications. 2008; 4:306–323. [PubMed: 18640906]
44. Panagiotakos DB. The value of p-value in biomedical research. The open cardiovascular medicine journal. 2008; 2:97. [PubMed: 19430522]
45. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews. 2007; 82:591–605. [PubMed: 17944619]
46. Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: problems, prevalence, and an alternative. The journal of wildlife management. 2000:912–923.

47. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35:D61–D65. [PubMed: 17130148]

48. Peric-Hupkes D, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell. 2010; 38:603–613. [PubMed: 20513434]

49. Smit AF, Hubley R, Green P. RepeatMasker Open-3.0. 1996

50. Furey TS, Haussler D. Integration of the cytogenetic map with the draft human genome sequence. Hum Mol Genet. 2003; 12:1037–1044. [PubMed: 12700172]

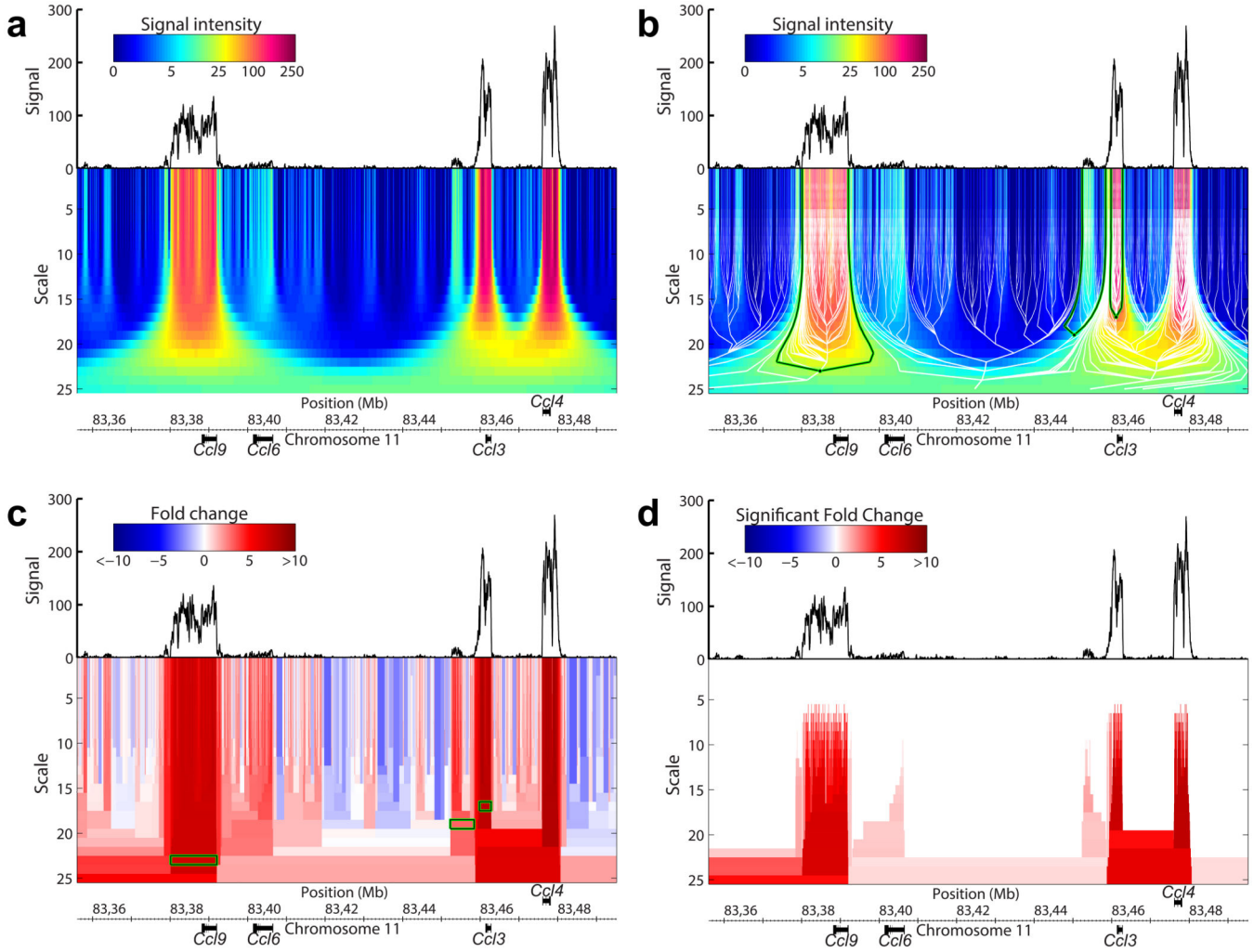

Figure 1.

Four-step procedure for the multiscale segmentation of genomic signals. The depicted genomic signal is a part of a Pol II ChIP-seq signal derived from primary murine bone marrow macrophage cells after 1 hour of lipopolysaccharide stimulation, mapped to genome assembly mm9. The genomic coordinates are in Mb.

(**a**) Smoothing of the genomic signal at different scales results in the Gaussian scale space. The scale space is represented as a heatmap below the original signal. (**b**) A segmentation tree is created by propagating nodes from the smallest scale to the largest scale. This tree is visualized by the white lines. The three pairs of green and black branches are examples of how to derive segments from the tree, as is explained in the text. (**c**) Segments at multiple scales are derived from the segmentation tree. The three green and black rectangles (segments) are derived from the green and black pairs of branches in **b**. The different segments are colored according to the (log2 transformed) fold change between observed and expected signal intensity within the segments. (**d**) The segments are scored for enrichment using a statistical testing procedure that outputs the fold change that is statistically significant at a predefined confidence level (‘Significant Fold Change’, SFC).

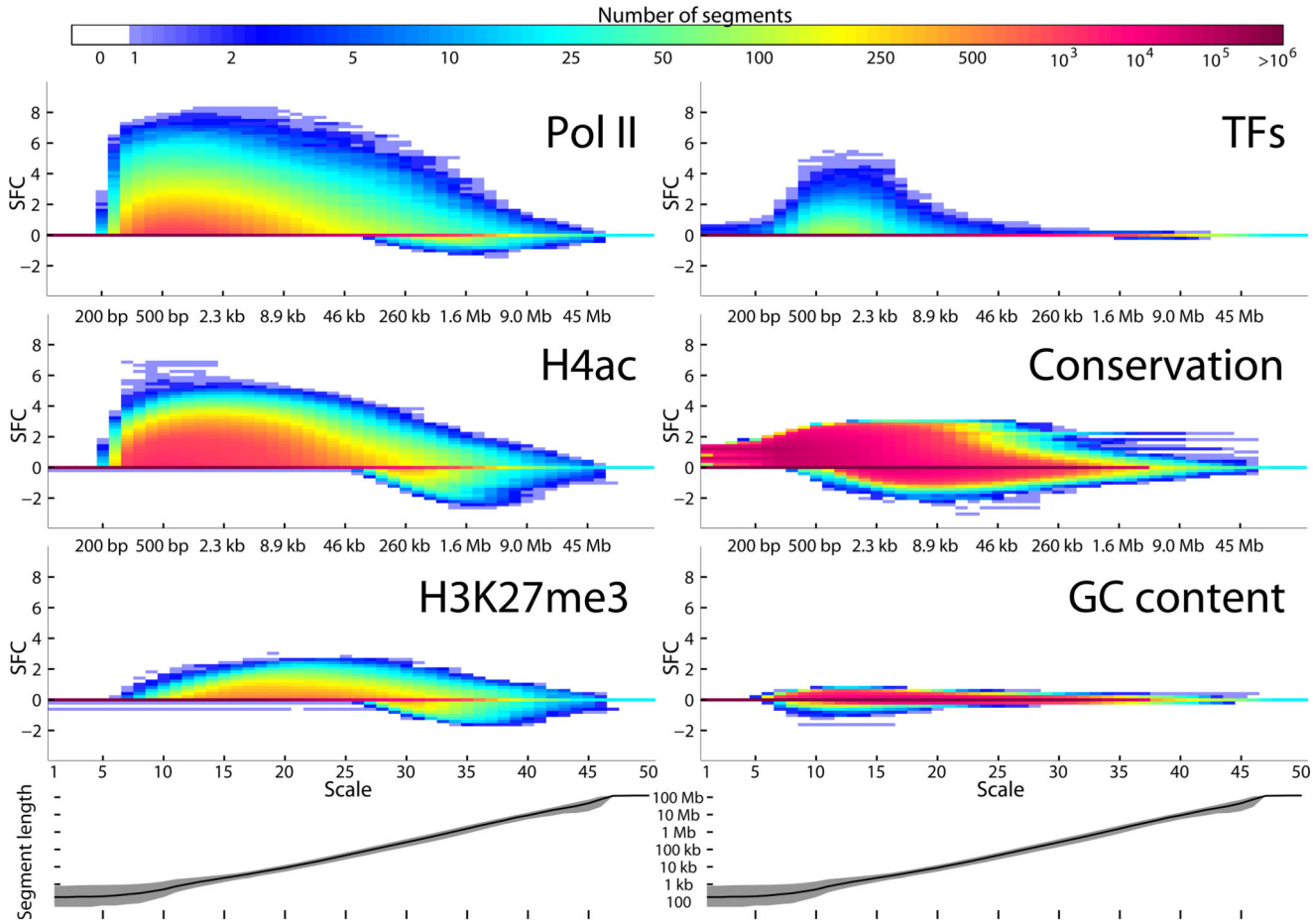

Figure 2.

Significant fold change (SFC) of segments across multiple scales for different ChIP targets, conservation scores and GC content.

Each of the six panels displays the results of a genome-wide multiscale segmentation analysis. In each panel, the heat map diagram shows a two-dimensional histogram that is created by binning the segments based on their scale (x-axis) and on their SFC (y-axis). The color indicates the number of segments.

The three panels on the left represent specific ChIP targets; their two-dimensional histograms are averages across multiple histograms derived from ChIP-seq experiments performed under different experimental conditions. The upper-right histogram is an average across the histograms corresponding to the ChIP experiments of three TFs; ATF3, p50 and p65 (also based on multiple experimental conditions). The histograms were averaged, as they were very similar for the three TFs. The middle and lower-right panels represent the multiscale signatures derived from the 29-way vertebrate conservation scores and the GC content signal, respectively. In the bottom of the figure the median (black line) and interquartile range (grey fill) of the segment sizes at different scales is shown. The median segments sizes for scales 5,10,…,45 are stated between the heat maps.

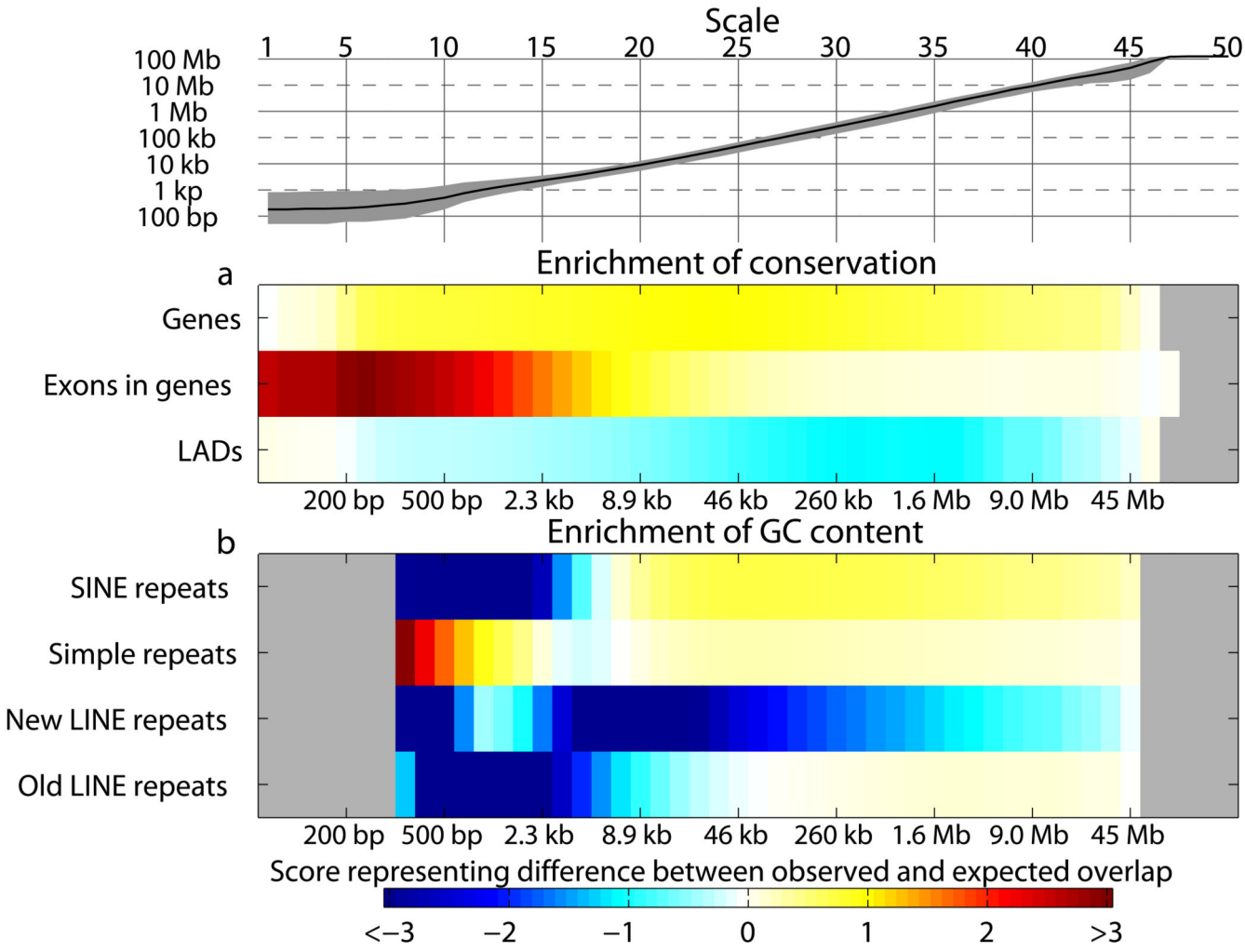

Figure 3.

Overlap between functional genomic regions and the segments comprising the MSRs of genomic signals.

(**a**,**b**) The heatmaps depict the degree of overlap between genomic regions and enriched segments (SFC>0) of the conservation (in **a**) and GC content (in **b**) genomic signals. The genomic regions are shown on the left of the heatmaps. Positive scores represent a larger overlap than expected by chance; negative numbers a smaller overlap. A grey color indicates that fewer than ten enriched segments at that scale were found. In that case, the overlap score was not computed. The top panel depicts the median (black line) and interquartile range (grey fill) of the segment sizes across the 50 scales. The median segments sizes for scales 5,10,…,45 are stated between the heat maps.

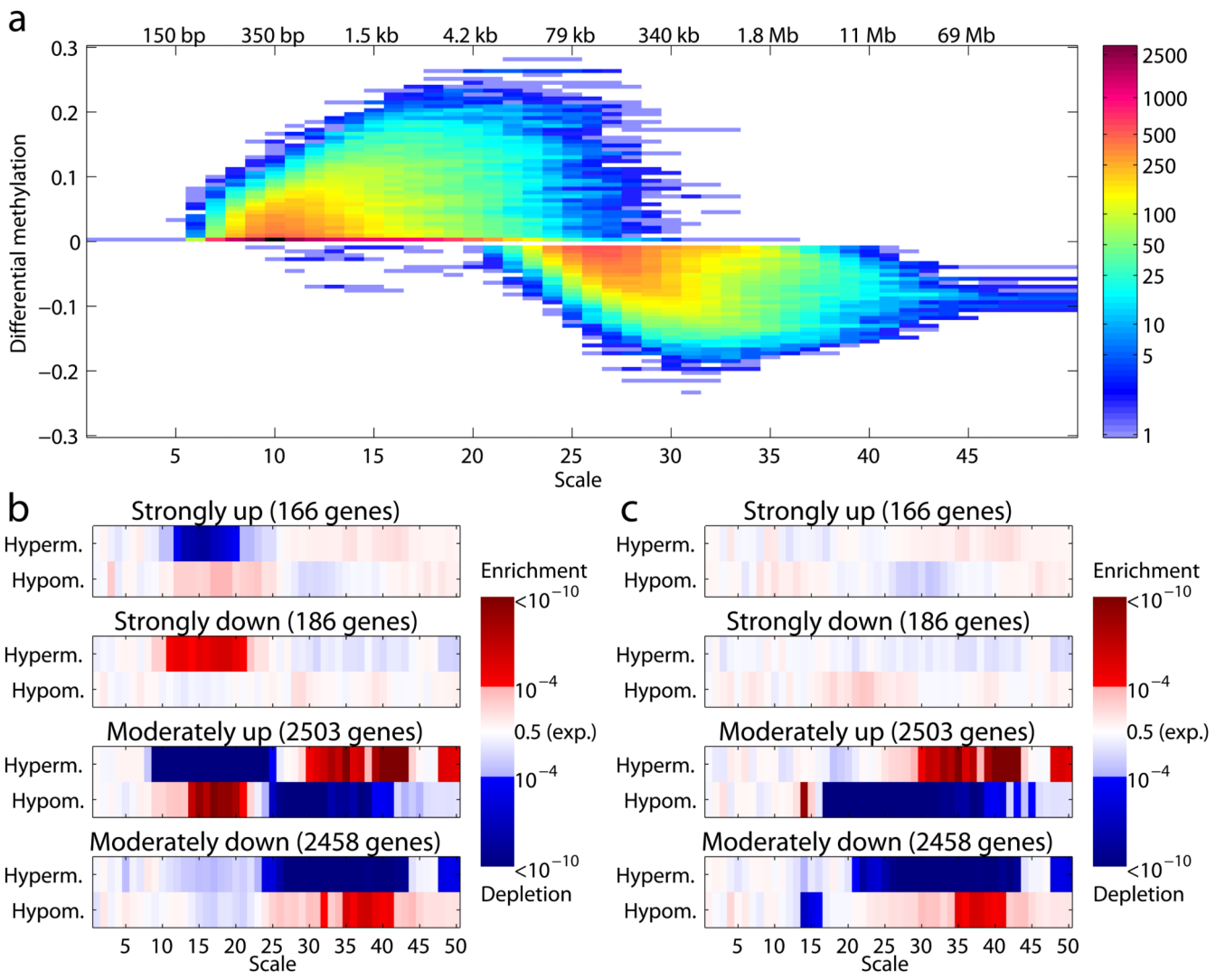

Figure 4.

DNA methylation MSR of differentially expressed colorectal cancer genes

(**a**) A two-dimensional histogram created by binning the segments based on their scale and on their differential DNA methylation score (DM) between a primary human colorectal tumor and adjacent normal colon tissue. The color indicates the number of segments. The x-axis and y-axis represent the scale and the DM, respectively. Positive DM scores indicate hypermethylation in the tumor, whereas negative DM scores represent hypomethylation. DM scores in the bin around zero were removed. This figure is created similarly to Fig. 2. (**b**) Four groups of genes were created based on the differential expression between tumor and normal: 1) the strongly upregulated set of genes have at least 1 unit more expression in the tumor than in the normal tissue; 2) strongly downregulated genes have at least 1 unit less expression; 3) moderately upregulated genes have between 0.1 and 1 higher expression in tumor; and 4) moderately downregulated genes have between 0.1 and 1 lower expression. (A difference of 1 unit corresponds to a doubling or halving for these log2 transformed gene expression values.) These groups were compared with the genes that had the largest hyper-

or hypomethylation values around the TSS at a particular scale. The heatmaps depict the P-values of the hypergeometric overlap test, where red indicates significant overlap (enrichment), blue indicates a significant lack of overlap (depletion) and white represents the randomly expected overlap. (**c**) Similar to **b**, but for segments overlapping the gene middle (GM).