MULTISCALE REPRESENTATION OF GENOMIC SIGNALS

SUPPLEMENT

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]

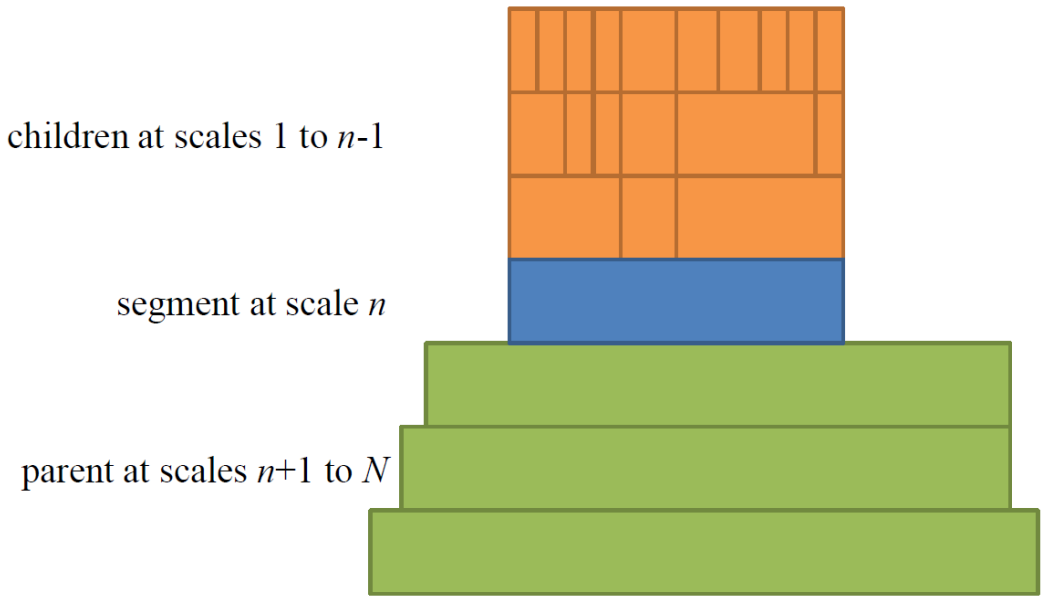
[1] Institute for Systems Biology, Seattle, Washington, USA
[2] Seattle Biomedical Research Institute, Seattle, Washington, USA
[3] Current address: Department of Biomedical Sciences, Oregon State University, Corvallis, Oregon, USA
[4] University of Southern California Epigenome Center, University of 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

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

CONTENTS

SUPPLEMENTARY NOTE 1: PRUNING THE MSR

The MSR of a genomic signal can be used to detect segments enriched (or depleted) in signal across all genomic length scales. Since enrichment of signal intensity is inherited between scales, pruning strategies are necessary to filter out only the most interesting segments. Here, we present a simple pruning strategy that enables the MSR to be used as a multiscale 'peak-calling' algorithm.



Each segment in the MSR is subjected to a test to decide whether the segment will be pruned or not. This test is based on comparing the (SFC) score of the segment with that of its children and its parents. More specifically, a segment at scale n (blue in figure above) is compared to all children at scales 1 to n-1 (orange segments in figure above) and all its parents at scales n+1 to N, where N is the largest scale in the MSR (green segments in the figure above. The segment is only kept (i.e. not pruned) when it has a score higher that all of its children and all of its parents.

To be able to detect segments at different scales a size constraint is introduced: The segment under investigation is not compared to all its children and parents, but only to those children with segment lengths larger than $S \cdot R$ and parents with length smaller than S/R. Here, S is the length of the segment under investigation and R is parameter between 0 and 1. The default setting for R is 0.2, which means that children 5 times smaller than the segment and parents 5 times larger than the segment are not considered. There is one exception: when all children at scales 1 to n-1 have a lower score than the segment under investigation, all children are pruned, also those smaller than $S \cdot R$. When R is 0, there is effectively no size constraint, i.e. the segment under investigation is compared to all its children and parents. In that case, a genomic position is part of at most one segment that is not pruned.

Further, a slack parameter, T, is introduced, which prevents higher-scale segments from breaking up into smaller ones, i.e. it prevents segments from being pruned, because one of the

children has a slightly better score. Specifically, the segment under investigation is kept (i.e. not pruned) when its score (denoted by X) multiplied by T, (i.e. $X \cdot T$) is larger than the scores of all its children, and its (unadjusted) score X is larger than the scores of all its parents. The default setting for T is 1.05.

Parameters R and T were selected manually after visual comparison of the genomic signal and the pruned segments. It is possible to optimize these parameters: for example, for Pol II ChIP-seq R and T can be optimized such that the selected segments maximally cover genes and minimally cover non-genic regions. We have not explored these options here.

Supplementary Fig. X1 depicts an example genomic signal with the results of this pruning scenario for different parameter settings. Also see **Supplementary Fig. X2**, **X3** and **X4** which depict pruned versions of **Fig. 1** and **Fig. 2** from the main text. The code for these pruning algorithms is available at https://github.com/tknijnen/MSR.

SUPPLEMENTARY NOTE 2: A COMPARISON WITH THE PEAK CALLER MACS AND SICER

The widely used peak calling algorithm MACS[1] was applied to the BMM ChIP-seq data described in the main text. MACS (version macs14 1.4.2 20120305) was run with standard settings. Specifically, MACS used the appropriate control ChIP-seq signal, i.e. the signal obtained from three IPs of BMMs with immunoglobulin G derived from rabbits that were not immunized with specific target antigens (same as for MSR), a P-value threshold of 10^{-6} (same as for MSR), and the (standard) effective genome size of 1870 Mb.

Additionally, SICER[2], the widely used algorithm to detect broad histone modifications was also applied to the BMM ChIP-seq data. SICER (version 1.1) was run with standard settings. Specifically, SICER used the appropriate control ChIP-seq signal, i.e. the signal obtained from three IPs of BMMs with immunoglobulin G derived from rabbits that were not immunized with specific target antigens (same as for MSR), a FDR threshold of 0.01, the (standard) effective genome fraction of 1870 Mb / 2655 Mb, a window size of 200 and a gap size of 600.

The detected peaks of MACS and segments of SICER were compared to the pruned MSR with standard settings (T=1.05, R=0.2). **Supplementary Fig. 1** depicts the distribution of the sizes of the detected MACS peaks, the SICER segments and of the enriched segments of the pruned MSR for the six different ChIP targets. It is clear that the MACS peaks are very constrained in their size: The MACS TF peaks are between 100 bp and 1 kb, whereas the MACS peaks for Pol II and the histone modification marks are around 1 kb and not longer than 5 kb. Although, the SICER segments are larger than the peaks from MACS they remain limited to a certain range of segment sizes, approximately between 1 and 10 kb. Because of the window size setting, the size of any SICER segment is a multiple of 200. Thus, SICER segments are granular by design and have a low resolution, which explains the sparsity of small segment sizes in the SICER histograms. Whereas MACS and SICER parameters can be changed (we used the default parameters), it remains the case that the range of the sizes of detected segments is limited by the underlying methodology, which is not based on a multiscale framework.

In contrast, the enriched segments from the pruned MSR show much larger variation in their size. This is because the MSR captures information across all scales, thereby allowing for discovery of segments of all sizes. For Pol II, MSR segments varied from 100 bp to 100 kb, which fits well with the variation of gene sizes throughout the mouse genome. We posit that the pruned MSR resulted in a larger and more reasonable variation in the size of the detected peaks than MACS and SICER. However, we do not make the hard claim that the pruned MSR is 'better' than MACS or SICER, especially when the algorithm in question is used for the appropriate data, such as MACS for TF ChIP-seq. This analysis simply demonstrates the power of the multiscale framework to detect interesting segments at different scales.

Supplementary Fig. 2 depicts the distribution of the MSR segment sizes, not only for the standard pruning settings (T=1.05, R=0.2), but also for different settings discussed in the pruning section above. Additionally, the distribution of segment sizes is shown in the scenario, where SFC scores were computed without the unique mappability map, i.e. assuming a uniform background. This figure shows that the size distribution of the pruned MSR segments does not heavily rely on the parameter settings or the use of the unique mappability map.

SUPPLEMENTARY NOTE 3: EFFECT OF PRUNING PARAMETERS *R* AND *T*

We investigated the effect of R and T on the enriched segments in the pruned MSR. **Supplementary Fig. X5** depicts the results of this analysis, from which we made the following observations:

- **Effect of slack parameter *T***

  Setting T slightly higher than 1 ensures that segments are not dismissed (i.e. pruned) because one (or more) of its children have a slightly higher score. Segments detected without slack (T=1) can be affected in three different ways when slack is applied (T=1.05). First, the detected segment remains unchanged. Second, the segment is dismissed in favor of a larger segment. In this case the original segment (in red) is a child segment of the new segment (in blue). The original segment only has a slightly higher score than the parent segment and is therefore dismissed. Third, multiple original segments are dismissed in favor of one larger segment that covers all of them and which has a score only slightly lower than all of the original segments.

  The original segments detected for the three transcription factors, P65, P50 and ATF3, remain mostly unchanged; i.e. 90-95% of the original segments are unaffected. Between 5-10% of the original segments are replaced by a larger segment. The average size increase of these new segments is approximately 20%. For example, segments of 100 bp are replaced by segments of 120 bp. There are only a handful of cases where multiple original segments are replaced by one larger segment. In almost all of these cases, two original segments were merged into one larger segment, the average size of which is only slightly larger than the boundaries of the two original segments.

  The two histone modification marks, H3K27me3 and H4ac, and Pol II show a larger number of affected original segments, i.e. between 40-50%. For the two histone modification marks,

approximately 30% of the original segments are replaced by a larger segment, which is, on average, about 20% larger than the original segment. The remaining original segments (~15%) are merged into larger segments. The average number of original segments merged into one larger one is 3. For Pol II, 20% of the original segments are replaced by a larger one, which is, on average 15%, larger than the original segment. The remaining original segments (~30%) are merged into larger segments. Between 3 and 8 original segments are merged into one larger segment, the boundaries of which are only slightly wider than the boundaries encapsulating the original segments.

- **Effect of size constraint parameter *R***

Setting *R* larger than 0 enables segments to be detected, i.e. not pruned, even when one of its child or parent segments have a higher score. Setting *R* to 0.2 makes these segments candidates to be detected if their size is at least 5 (1/0.2) times larger or smaller than the higher scoring segment. Thus, the consequence of loosening the size constraint by setting *R* to 0.2 is the detection of additional segments.

The number of segments detected for the three transcription factors increased by about 3% relative to the number of original segments, and by 6% for H3K27me3, H4ac, and Pol II. In all cases the sizes of the additional segments are much larger than the original segment sizes for these ChIP targets.

- **Effect of changing *R* and *T* simultaneously**

In this analysis, we observed that the influence of *T* and *R* are virtually independent of each other. That is, the effect of changing *R* and *T* at the same time can be derived from their individual influences.

- **Summary of observations**

Slack (parameter *T*) favors slightly larger segments or merges multiple segments into one bigger one. The large majority of segments detected without slack remains unchanged or becomes slightly bigger when slack is applied. Merging segments into larger ones is similar to the post processing step of many peak callers, where nearby peaks are clustered. For example, SICER[2] uses the (user defined) gap parameter to merge nearby peaks and ZINBA[3] uses the (user defined) peak refinement threshold to accomplish this.

Loosening the size constraint (parameter *R*) leads to the detection of a moderate number of additional segments. These segments are in general much larger than the original segments and thus point to enrichment of the signal at a much larger scale.

- **Discussion**

Pruning parameters *R* and *T* can be said to control the extent to which a "true" ChIP-seq binding event at some position in the genome will be part of (potentially multiple) enriched segments (containing that position) across multiple scales in the multiscale segmentation. If we select the most aggressive pruning strategy (*R*=0), then our method will call no overlapping peaks at a given location. If instead we select a more minimal pruning strategy (*R*>0), at the genomic location of a "true" ChIP-seq binding event, our method could detect several overlapping segments of different scales, at that location. Whether the user would select an aggressive or minimal pruning strategy depends on how the data will be used. For "traditional" peak-calling, (R=0, T=1.05) are sensible choices, but for correlative analyses of the pruned MSR

with another genomic signal, a minimal pruning strategy (R>0) would preserve more multiscale information.

Finally, we note that the pruning parameters R and T do not control the genome-wide stringency of peak identification. The stringency (across the genome) for peak identification is controlled by the SFC cutoff. The SFC has a clear statistical interpretation by definition, and is directly related to a P-value, which can be set by the user.

SUPPLEMENTARY NOTE 4: DETAILED ANALYSIS THE CORRELATIONS BETWEEN THE MSRS OF GC CONTENT AND REPEAT ELEMENTS

We detected a significant overlap of SINE repeats with GC-enriched segments with approximate sizes from 10 kb to 10 Mb (scale 20 to 40, **Fig. 3b**), consistent with longstanding observations that SINEs are enriched in large-scale GC rich regions in the mouse and other mammalian genomes[4,5]. Surprisingly however, smaller GC-rich regions (scale 10–15, segment sizes around 1 kb) do not seem to overlap with the SINE repeats at all (**Fig. 3b**). Subsequent analysis indicated that at scale 14 (median segment size: 1.2 kb), 60% of the 8,600 GC-enriched regions overlap with the transcription start site (TSS) of a gene, and 85% of the regions are within 1 kb of a TSS. Furthermore, 78% of the GC-rich regions overlap with one of the 16,020 annotated CpG islands in the mouse genome[6]. 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. This observation, however, cannot be complete story, because there is still a significant lack of overlap between these small GC-rich regions and SINEs even if the GC-rich regions that overlap with CpG islands are removed from the analysis.

In contrast, younger long interspersed elements (LINEs) are under-represented both in small and large GC-rich regions. While the small-scale observation may have the same origin as the SINE underrepresentation, the large-scale observation is consistent with reports that LINEs tend to accumulate in AT-rich sections of the genome[4]. The preference of younger LINEs for large-scale AT-rich regions is however not observed for the older LINEs. This may be because the large-scale deletion rate in AT-rich regions is higher than in GC rich regions so that old elements have disappeared faster from AT-rich DNA[7] (**Supplementary Figs. 9 and 10**).

SUPPLEMENTARY NOTE 5: THE DEGREE OF CORRELATION BETWEEN GENOMIC SIGNALS VARIES WITH SCALE

Having observed that the overlap between functional genomic regions (i.e. genomic annotations, such as genes, exons, LADs and repetitive elements) and genomic signals greatly varies with scale, we asked ourselves whether the correlation between genomic signals themselves would also depend on the length scale. The degree of correlation between a pair of genomic signals was determined as follows. First, enriched segments for both genomic signals at each scale were found by using the score threshold SFC > 1. Second, we computed the overlap between the enriched

segments of the two genomic signals at all combinations of scales. Based on the randomly expected overlap, we derived an SFC score that represents the degree of correlation. Heatmaps were used to depict this correlation across scales for two pairs of genomic signals, of H4ac and Pol II (**Supplementary Fig. 11a**) and H4ac and H3K27me3 (**Supplementary Fig. 11b**). H4ac, which is a mark of open chromatin, and Pol II, which is involved in transcription, are correlated with each other at all scales, except for the very small and very large scales, where no enriched segments are found for these signals (**Supplementary Fig. 11a**,**c**,**e**). In contrast, H4ac and H3K27me3, the latter being a repressive mark involved in gene silencing, are anti-correlated at a smaller scale (scale 20), while they are positively correlated at a larger scale (scale 35) (**Supplementary Fig. 11b**,**d**,**f**). The negative correlation (lack of overlap) at the smaller scale is consistent with current understanding that H3K27me3 is associated with silenced genes, while H4ac is associated with activated genes. Close observation of the significant segments indeed reveals that these epigenetic marks are largely mutually exclusive within a given intergenic region (**Supplementary Fig. 11f**). The positive correlation at the larger scale, 35 (approx. segment size, 1 Mb), was initially more surprising. However, we observed that enriched segments of both H4ac and H3K27me3 at scale 35 show a large overlap with genes (**Supplementary Fig. 4a and 5a**), indicating that both chromatin marks are primarily found in gene-rich regions, which explains their overlap at this large scale.

SUPPLEMENTARY NOTE 6: PREDICTING GENE EXPRESSION USING MULTIPLE GENOMIC SIGNALS

Next, we set out to exploit the fact that genomic signals contain specific information at different scales by formulating predictive models based on the MSR of a genomic signal. Based upon current understanding that covalent histone modifications and polymerase II activity are essential to controlling transcription, we tested whether a multiscale representation of histone acetylation ChIP-seq or polymerase II ChIP-seq data can be used to predict mRNA gene expression levels, for all genes, within murine macrophages.

We created gene-specific MSRs in the genomic region from 1kb upstream to 1kb downstream of each individual gene (**Supplementary Fig. 12a**). 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 to create a normalized gene-specific MSR with 10 scales. This normalization accounts for differences in gene size. Next, for each gene, 50 features were derived for H4ac and Pol II from unstimulated macrophages. These 50 features consisted of the SFC scores of the gene-specific MSR at scale 1 to 10 for five distinct positions within the gene, i.e. 1 kb upstream of the gene (U), the TSS, the middle of the gene (GM), the end of the gene (GE) and 1 kb downstream of the gene (D).

These gene-specific features based on the MSRs of H4ac or Pol II served as input data for a model to predict microarray-derived gene expression data from identical experimental conditions as the ChIP-seq experiments[8]. We employed the Random Forest regression algorithm[9] to predict the gene expression measurements of all non-overlapping genes that are represented on the Affymetrix Mouse Genome 430 2.0 GeneChip. A ten-fold cross-validation scheme was used for model training and evaluation.

Prediction accuracy was assessed using the Pearson correlation coefficient between measured and model-predicted expression levels, for four different models: model T, which used the total signal integrated over the gene plus its 1 kb flanks (i.e., feature data at scale 10); model S, which is based on the original signal (equivalent to the feature data at scale 1); model B, containing features from the best predicting single scale; and model M, which included the multiscale signal levels as features. The multiscale approach (M) outperformed these three single-scale approaches (**Supplementary Fig. 12**). This means that the 'shape' of the genomic signal across the gene, as captured by the MSR, is indicative of the underlying transcriptional state. Pol II outperforms H4ac data most likely because the presence of Pol II on genomic DNA is a more direct indication of active transcription than an acetylation mark on histone proteins associated with "active" chromatin. Analysis of the feature importance scores assigned by the model shows that in addition to the total signal (the multiscale representation at scale 10), feature values derived from segments spanning the TSS between scales 5 and 8 carry information about the expression of the gene (**Supplementary Fig. 12c**).

In a further experiment, 11 genomic signals, including 4 from ENCODE[10], were used independently and jointly in the predictive model. Specifically, we generated gene-specific MSRs (as described in the **Methods** section) for eleven different genomic signals. Seven of these are previously described, namely our BMM ChIP-seq signals of ATF3, p50, p65, H4ac and H3K27me3, and the GC content and conservation signal. Additionally, we used murine BMM ChIP-seq data from ENCODE.

Specifically, we downloaded the raw ChIP-seq data in bigWig format from the ENCODE project at UCSC for the CTCF transcription factor and three histone modifications. The table below lists the accession numbers for these signals.

Antibody or target protein	UCSC Accession	GEO sample accession
CTCF	wgEncodeEM002663	GSM918726
H3K27ac	wgEncodeEM002657	GSM1000074
H3K4me1	wgEncodeEM002658	GSM1000066
H3K4me3	wgEncodeEM002659	GSM1000065

The bigwig files were transformed into wig files using UCSC's BigWigToWig binary utility. The wig files were imported and transformed into genomic signals at 10 bp resolution. The standard workflow was applied to these genomic signals to create the MSRs from which we derived the gene specific MSRs.

The random forest prediction model was run for each of the eleven feature sets as described in the main text. Additionally, the model was run on the combined feature sets. For the multiscale model (M), this means that the model was run with 550 features (11 genomic signals, 5 sampling positions in the gene, 10 scales), whereas model T contained 11 features and model S 55 features.

Supplementary Fig. 13 depicts the results of this experiment. From **(a)** it is clear that the multiscale representations are superior in terms of prediction performance for all of the individual genomic signals compared to feature sets based on either the total signal (T) or the original signal (S). The combined model outperforms the single models, indicating that integrating multiple

genomic signals is beneficial in predicting gene expression. This is in line with previous studies[11,12]. Also for the combined model the multiscale representations have the highest performance, although this is less dramatic than for the individual models. From the importance scores **(b)** it is clear that the predictive features were selected from a range of different scales.

Noteworthy is the fact that, H3K27ac (from ENCODE) was the best individual signal **(a)** and also had the highest importance scores in the combined model **(b)**. This histone modification also ranked highly in the approach of Karlic *et al.*[11], although they focused on human CD4+ cells instead of murine BMMs.

Conclusively, in agreement with previous studies[11,13,14], this experiment showed that gene expression can be predicted by histone modifications with reasonable accuracy, and that certain histone modifications are more important than others. Interestingly, the MSR approach not only identifies which histone modification marks are important, but also at which scales they are most informative, as predictive features were found across the range of scales.

SUPPLEMENTARY NOTE 7: OVERLAP BETWEEN PMDS, MPS AND MSR SEGMENTS

Berman *et al*.[15] used a sliding window approach to find differentially methylated segments. Specifically, the focal methylation prone segments (MPs) were identified using windows of five adjacent CpGs with an average methylation level less than 5% in the adjacent normal tissue and greater than 35% in the tumor. On the other hand, the broad partially methylated domains (PMDs) were identified by scanning all windows of at least 10 kb, where windows with an average methylation between 20-60% were merged into single partially methylated domains of at least 100 kb in length.

We compared the MPs and PMDs from Berman *et al*.[15] with the differentially methylated segments of the MSR. For this we the employed the MSR at P-value threshold $p^{th} = 10^{-6}$. Further, the MSR was pruned (using default settings) to identify:

- The hypermethylated segments (positive differential methylation score)
- The hypomethylated segments (negative differential methylation score)

The hypermethylated MSR segments were compared with the MPs and the hypomethylated MSR segments were compared with the PMDs. **Supplementary Fig. 14** depicts the results of this analysis. There is a large concordance in the sizes of segments found; both when comparing the MPs with the hypermethylated MSR segments and when comparing the PMDs with the hypomethylated MSR segments (**Supplementary Fig. 14a,b**). Additionally, the number of segments identified is comparable and the large majority of MPs and PMDs overlap with at least one of the MSR segments (**Supplementary Fig. 14c,d**). Also, when analyzing the actual genomic overlap between the segments (in basepairs), they largely agree.

Note on the data: The MPs and PMDs were downloaded from http://epigenome.usc.edu/publicationdata/berman20101101/ , and were called:

- Berman2011-shortDomainsTumorHyper.gtf
- PMDsTumorHypo.methylCGsRich_tumorM030510_.wind10000.minOutput100000.minCpg10.meth0.20-0.60.gtf

PMDs that overlapped with the regions in:

- PMDsNormalHypo.methylCGsRich_normalM030510_.wind10000.minOutput100000.minCpg10.meth0.20-0.60.gtf

were discarded.

SUPPLEMENTARY NOTE 8: DETAILED ANALYSIS OF SCALE-DEPENDENT RELATIONSHIP BETWEEN GENE EXPRESSION AND DNA METHYLATION

We compared the differential gene expression between the colon tumor sample and the matched normal tissue with the differential DNA methylation captured by the MSR. Specifically, for all (9111) genes with CpG islands overlapping their TSS, we recorded the differential methylation in the segment overlapping with the TSS for all 50 scales. At each scale, we selected the 20% of the genes with the highest differential DNA methylation and called these hypermethylated. We similarly, created a group of hypomethylated genes for each scale.

Next, 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 in tumor; 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. (The absolute expression levels used for this analysis are from ref[16] and are log2 transformed, i.e. a difference of 1 unit corresponds to a doubling or halving of the gene expression.)

For sets of strongly and moderately up- or downregulated genes (based on the expression data) we examined their membership in the hyper- and hypomethylated groups, the results of which are shown in **Fig. 4b**. The set of 166 genes strongly upregulated in the tumor show a significant depletion for hypermethylated genes at small scales, but enrichment for hypomethylation at these scales. Conversely, the 186 strongly downregulated genes were highly enriched for hypermethylation at the small scales, but not associated with hypomethylation. These observations fit with current understanding of an inverse correlation between promoter methylation and expression[16]. The moderately up- and downregulated genes show an unexpected pattern. Here, 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 characterized by an enrichment of hypermethylation at large scales and hypomethylation at small scales. The 2458 moderately downregulated genes were enriched in hypermethylation at large scales. This analysis clearly demonstrates the scale-dependent relationship between DNA methylation around the TSS and gene expression.

The role of cancer-associated methylation changes at gene bodies is unclear. A number of groups have reported a positive association between methylation state at gene bodies and expression level[17,18], while complete methylome studies have shown that expression corresponds more strongly to LADs, which do not always correspond to gene boundaries and often span multiple genes[16,19]. In order to gain insight into this problem, we repeated the MSR analysis focused on the methylation pattern at the middle of genes (GM) rather than the promoter (**Fig. 4c**) We did not

observe significant hyper- and hypomethylation of strongly up- or down regulated genes. For the moderately differentially expressed genes, we observed the same pattern as for the TSS analysis when investigating the larger scales, i.e. 30 and above. This is not surprising, since at these large scales the segments are far beyond the individual gene scale, and thus virtually all GM segments are identical to the TSS segments. On the smaller scales, 20 and below, where the segments are smaller than 10 kb, there is no pronounced pattern.

SUPPLEMENTARY NOTE 9: CHIP-SEQ PROTOCOL

Sample Preparation

For ChIP–Seq analysis, formaldehyde-fixed cells were sonicated and processed for immunoprecipitation. In brief, 3×10^7 BMMs were fixed for 10 min in 1% formaldehyde in PBS, washed in PBS, and harvested by scraping in PBS. Cells were lysed by resuspending cell pellets in RIPA buffer (10mM Tris-HCl, pH8.0, 140 mM NaCl, 1% Triton X-100, 0.1% SDS, 1% Deoxcholic acid sodium salt) and drawing the cell suspensions three times through a 30 gauge needle. Chromatin was sheared using a probe sonicator (130 W Ultrasonic Processor with a 3mm tip, 5 × 60s at 30% maximum setting). Sonication quality was checked by gel electrophoresis. Protein concentrations in the extracts were determined (BioRad DC Protein Assay Kit I # 500-111) and sonicated cell extracts containing 0.5 mg protein were incubated with antibodies overnight at 4°C. Immune complexes were recovered by incubation with a 50%-50% mix of magnetic beads coated with Protein A or Protein G (Dynabeads Protein A Invitrogen # 100.02D, Dynabeads Protein G Invitrogen # 100.04D). The magnetic beads were washed with RIPA buffer and the chromatin was eluted with 1% SDS in TE buffer (10 mM Tris-HCl pH8.0, 1 mM EDTA pH8.0) at 65°C for 15 min. Eluted chromatin was reverse-cross-linked by adding 226 mM NaCl and incubating at 65°C for > 5 hr. Then 36 mM Tris–HCl pH8.0, 9 mM EDTA, and 1.5 U of Proteinase K (Fermentas # EO0491) was added to the samples and they were incubated at 42°C for > 1 hr. DNA was purified using phenol/chloroform/isoamyl alcohol (25:24:1, v:v:v) extraction. The purified immunoprecipitated DNA was prepared for sequencing with the Illumina ChIP-Seq Sample Prep Kit and processed according to the manufacturer's instructions (Illumina Part # 11257047 Rev. A).

Antibodies

Antibodies used for ChIP-Seq were purchased from the indicated suppliers: ATF3 (Santa Cruz Biotechnology Inc. #sc-188), p50 (eBioscience #14-6732-81), p65 (Santa Cruz Biotechnology Inc. #sc-372), RNA polymerase II (Upstate # 05-623B Anti-RNA polymerase II, clone CTD4H8), Acetylated Histone H4 (Millipore #06-598), Methylated Histone H3K27me3 (Millipore #07-449).

Sequencing

A sequencing library for the Illumina Genome Analyzer was derived from the IP using the Illumina reagent kit (see systemsimmunology.org). Single-ended, 36-cycle sequencing was performed on an Illumina Genome Analyzer, and the raw image data were processed using the Illumina Genome Analysis Pipeline Software on a dedicated sequence data processing system (see Genome Analyzer Pipeline Software User Guide, Illumina, San Diego, CA, USA, v0.3). Reads were

aligned to the mouse genome using eland extended with an ELAND SEED LENGTH value of 25 and an ELAND MAX MATCHES value of 15, and with the 3′-most base excluded. Reads aligned to the same position and strand were counted only once to eliminate duplicates from PCR amplification (consistent with the approach of [20]). For all ChIP-Seq samples, aligned reads were processed into extended fragments (consistent with the approach of [21]) of length 158 bp, the estimated typical insert size in the sequencing library. This estimated size was determined by assaying representative ChIP-Seq samples (after the PCR amplification step) using the Agilent Bioanalyzer to determine the typical fragment size in the sequencing library, and subtracting the combined size of the two Illumina adaptor molecules.

Additional details

The extensive protocols are available at http://portal.systemsimmunology.org/portal/web/guest/chipseq_protocol. 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.

SUPPLEMENTARY NOTE 10: SFC AND OTHER SCORES FOR THE OVERLAP BETWEEN ENRICHED SEGMENTS AND GENOMIC ANNOTATION

Based on the following four numbers, we have computed SFC overlap scores for the overlap (or lack thereof) between enriched segments and genomic annotation (as explained in the Methods section).

I	Total length of the genomic regions
B	Length of the genomic signal
n	Total length of the enriched segments
X	Total length of the overlapping parts of the genomic regions and enriched segments

A standard test for significance of overlap is the hypergeometric test. However, the hypergeometric test is problematic in this case, because these numbers are often very big (as they represent the length of genomic regions and segments). This leads to astronomically small P-values when the null hypothesis is violated (**Supplementary Fig. X6a**). (The rationale behind this is that the segment lengths can be seen as the number of samples in the statistical test, which is easy to understand if one approximates the hypergeometric test (without replacement) by the binomial test (with replacement). This means there is a lot of statistical power to reject the null hypothesis if it is not true, leading to very low P-values.)

An alternative approach is to simply use the fold change between the expected and observed overlap (**Supplementary Fig. X6b**). As explained in the Methods, the SFC can be interpreted as the conservative estimate of the fold change. This is important, because for small sample sizes (and thus for small segments), the fold change can take on large or small values without being statistically significant.

Supplementary Fig. X6c depicts the SFC using a very stringent P-value cut-off. In comparison to the simple fold change, the SFC scores at the smaller scales (smaller segments) are more conservative, i.e. closer to 0, as they have to meet a very stringent statistical cutoff.

SUPPLEMENTARY TABLES

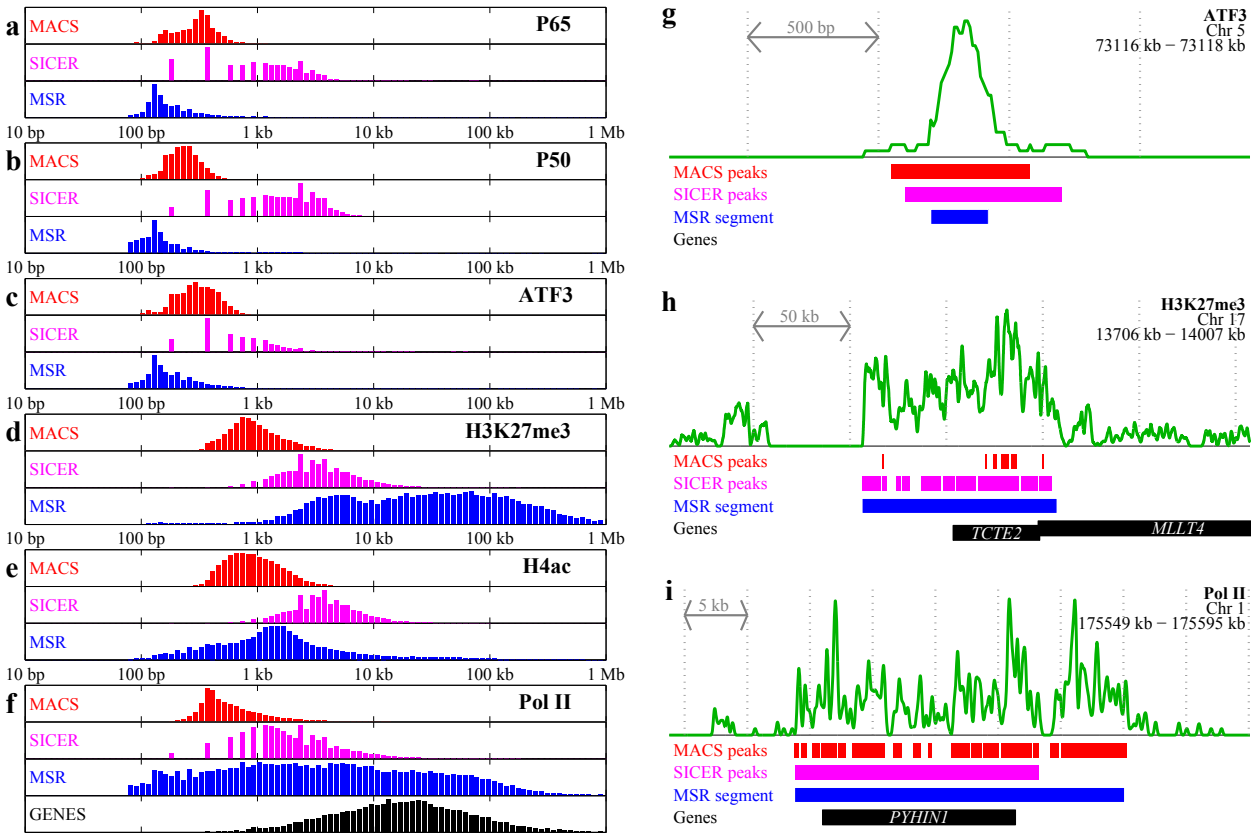
Supplementary Table 1 | Cross-validation results for the random forest regression model. MSRs were created for all (9111) genes with CpG islands overlapping their TSS by recording the differential methylation in the segment overlapping with the TSS for all (50) scales. MSRs were also created for the genes for segments overlapping with the middle of the gene (GM). From these MSRs three different feature sets were created:

1. TSS. The MSR around the TSS from scale 1 to scale 17, since at least 50% of the genes have segments smaller than 2 kb in this range
2. GB (gene body). The MSR around the GM from scale 1 to scale 22, since at least 50% of the genes have segments within the gene boundaries in this range
3. LR (long range). The MSR around the GM from scale 26 to scale 50, since at least 50% of the genes have segments larger than 100 kb in this range.

The random forest regression model was used to predict the differential expression between tumor and normal using different combinations of these feature sets. In every case, the expression of the genes in normal tissue (denoted by E) was added into the model. This table states the Pearson correlation between the actual differential expression and the predicted differential expression. The numbers (mean ± standard deviation) are based on 3 repeats of a 10-fold cross-validation scheme. The model was run for different feature sets (rows) and different P-value thresholds used to compute the MSR (columns).

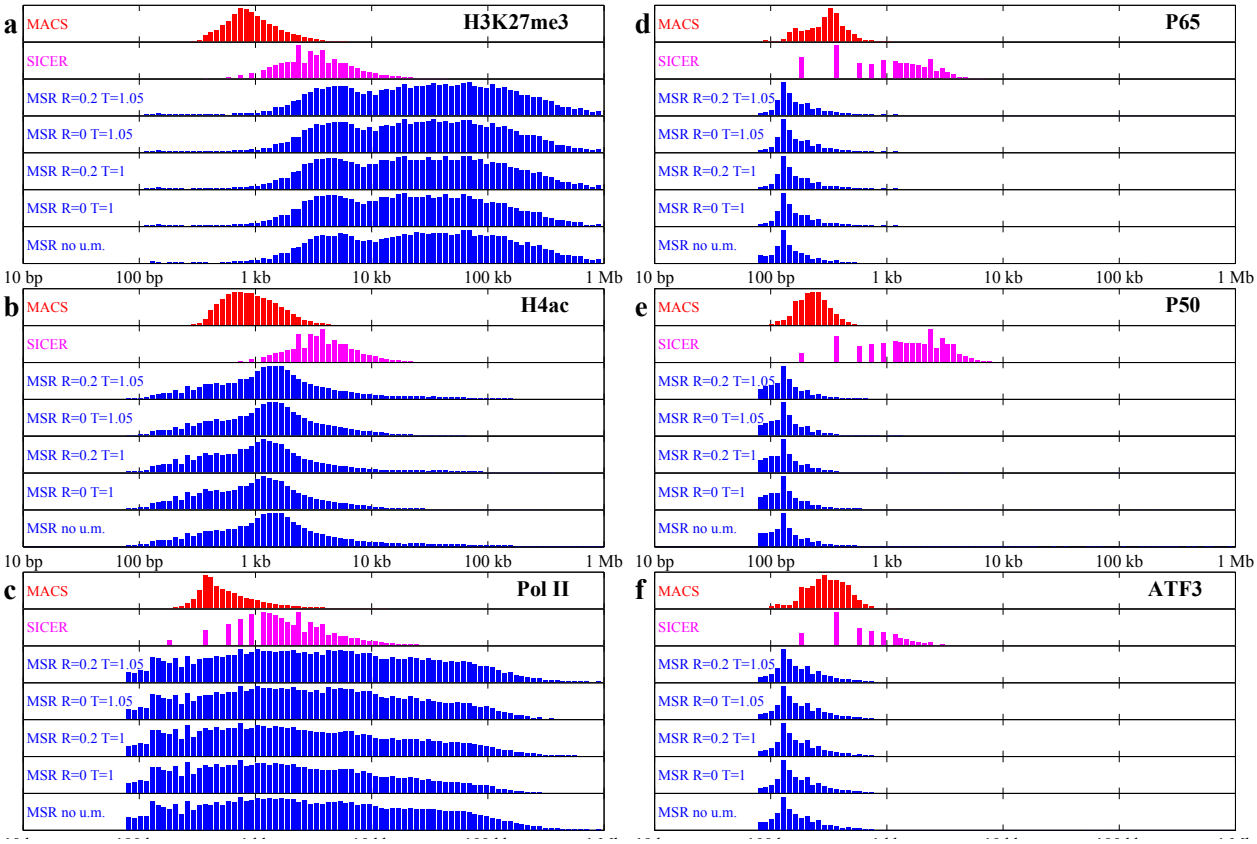
	MSR P-value		
Feature set	**0.5**	**0.05**	**10^{-6}**
E (expression only)	0.0583 ± 0.0054	0.0562 ± 0.002	0.056 ± 0.0024
E + TSS	0.253 ± 0.0053	0.237 ± 0.0014	0.214 ± 0.0016
E + GB	0.176 ± 0.0017	0.158 ± 0.0018	0.903 ± 0.00085
E + LR	0.217 ± 0.0012	0.214 ± 0.003	0.217 ± 0.0036
E + TSS + GB	0.264 ± 0.0038	0.244 ± 0.0019	0.213 ± 0.00081
E + TSS + LR	0.296 ± 0.002	0.279 ± 0.0024	0.266 ± 0.0025
E + GB + LR	0.237 ± 0.0014	0.222 ± 0.0018	0.211 ± 0.0036
All	0.298 ± 0.0013	0.282 ± 0.002	0.264 ± 0.0024

Supplementary Figure 1 | Comparison between the pruned MSR, peak caller MACS and SICER



(**a-f**) Distribution of the size of MACS peaks (red), SICER segments (magenta) and pruned MSR segments (blue) for six different ChIP targets. (**f**) also depicts the size distribution of the genes (black). (**g-i**) Three cases that exemplify the relation between MACS peaks, SICER segments and MSR segments. Each case is focused on one MSR segment (blue). The MACS peaks found in the vicinity are depicted in red; SICER segments in magenta. Genes in the vicinity are depicted in black.

Supplementary Figure 2 | Comparison between the pruned MSR, peak caller MACS and SICER using different settings to prune the MSR

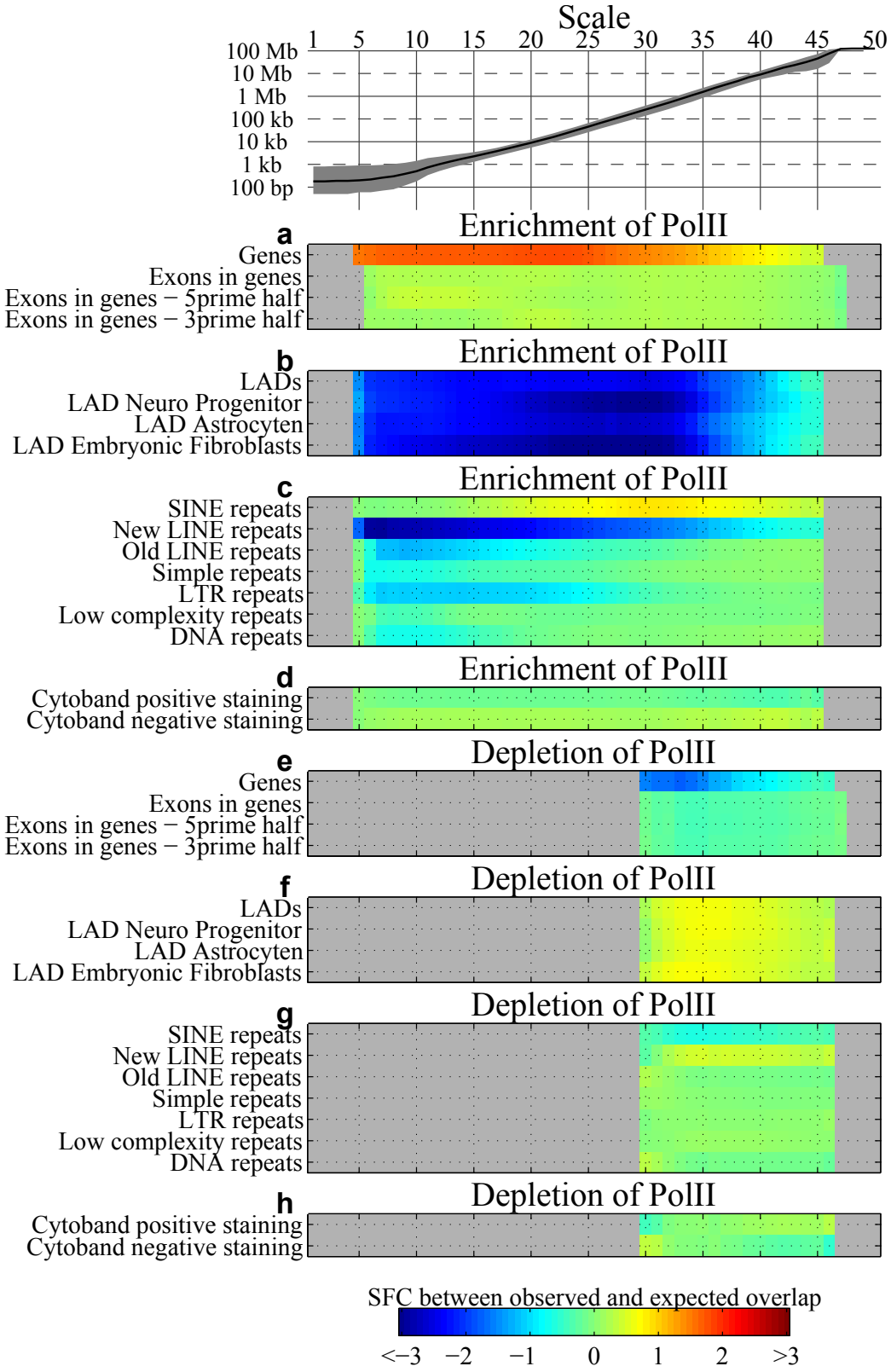


(**a-f**) Distribution of the size of MACS peaks (red), SICER segments (magenta) and pruned MSR segments (blue) for six different ChIP targets. MSR segment size distributions are depicted for different pruning scenarios indicated by the different settings for *R* and *T*. Additionally, a size distribution is shown for the case, where SFC scores were computed without using the unique mappability map, indicate by 'no u.m.'. In the latter case, the standard pruning settings (*T*=1.05, *R*=0.2) were used.

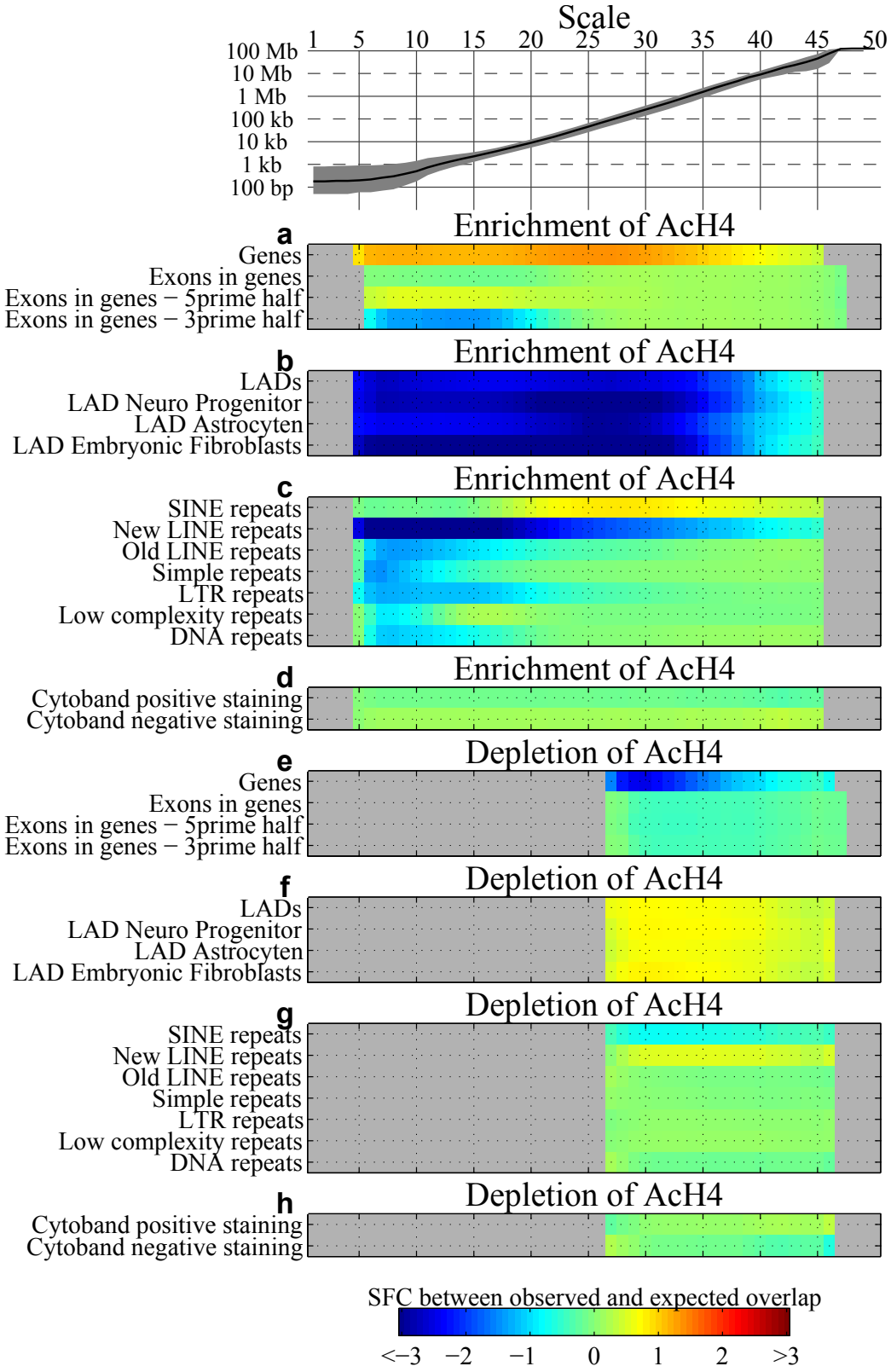
Supplementary Figures 3-10 | Overlap between functional genomic regions and the segments comprising the MSRs of genomic signals.

The heatmaps on the following pages depict the degree of overlap between genomic regions and significant segments of a genomic signal. Each figure represents one of the eight genomic signals. The type of genomic signal and whether the significant segments are enriched or depleted is indicated above the heatmap. The genomic regions are printed to the left of the heatmap. The color within the heatmaps represents the SFC between the observed and randomly expected overlap. A grey color indicates that less than ten significant segments at that scale were found. In that case it is not possible to reliably compute the SFC. The top panel depicts the median (black line) and interquartile range (grey fill) of the segment sizes across the 50 scales. The SFC scores for the exon genomic regions were computed with respect to genes, i.e. not with respect to the whole genome as is the case for the other genomic regions.

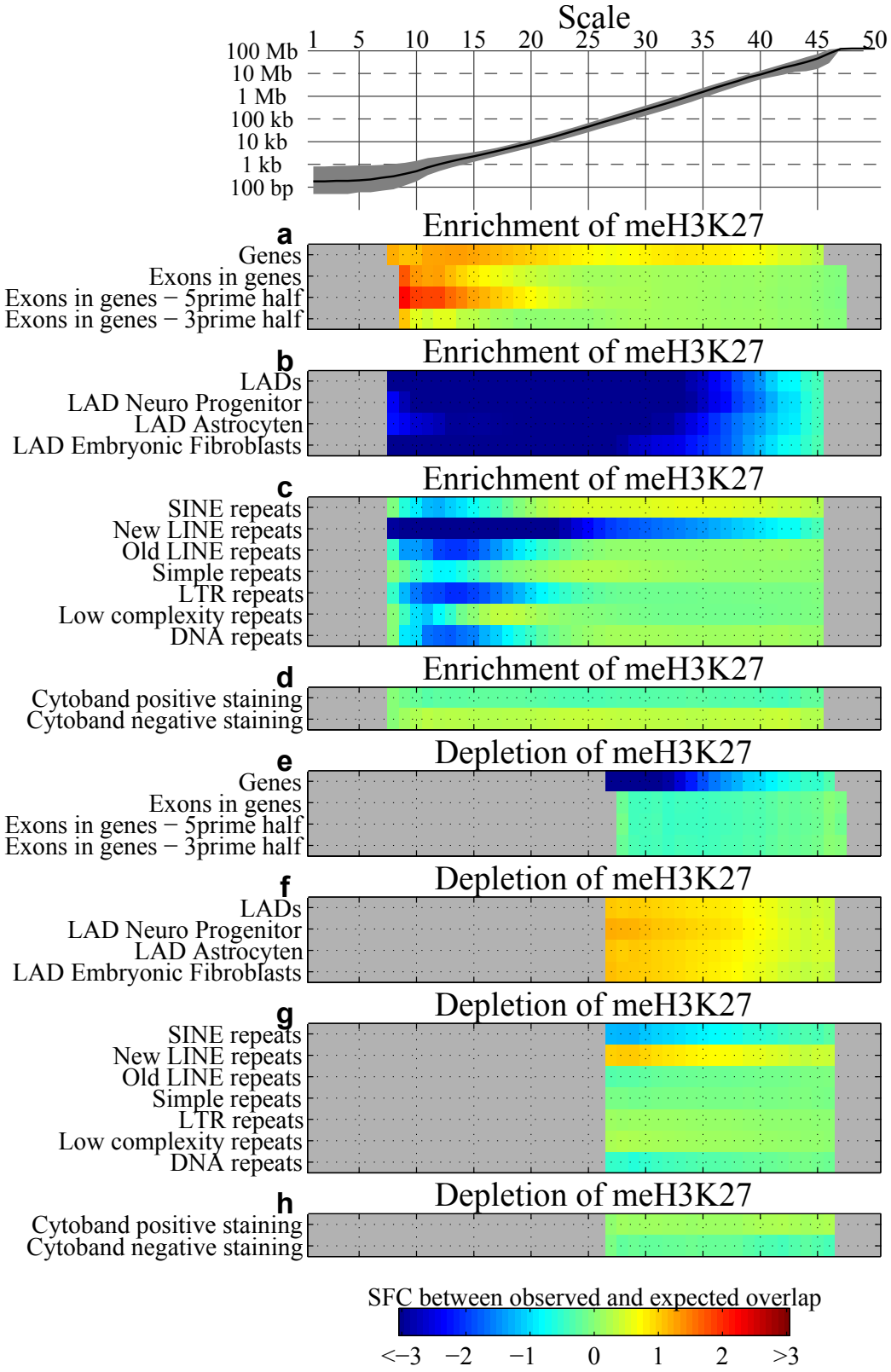
Supplementary Figure 3



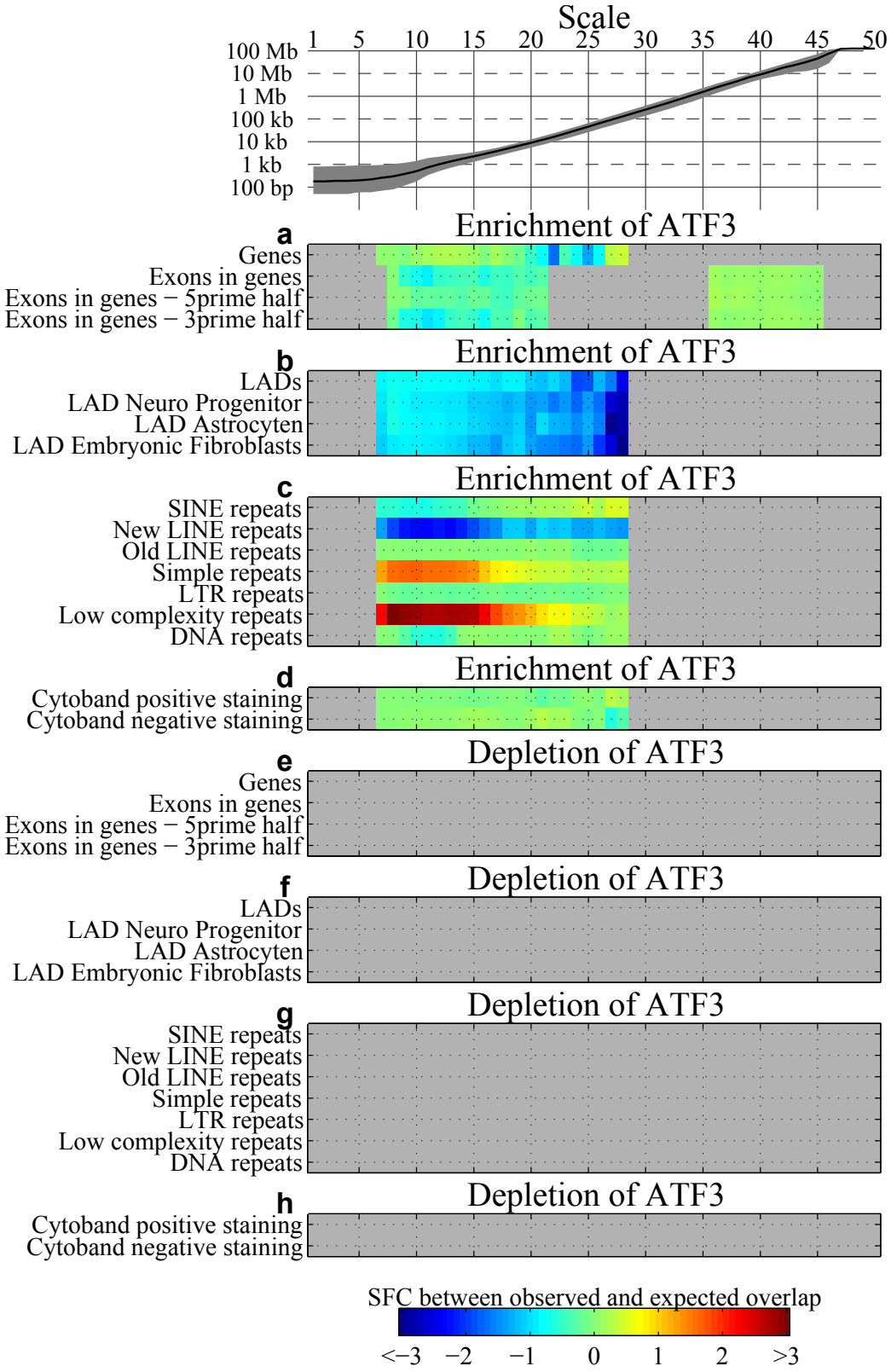
Supplementary Figure 4



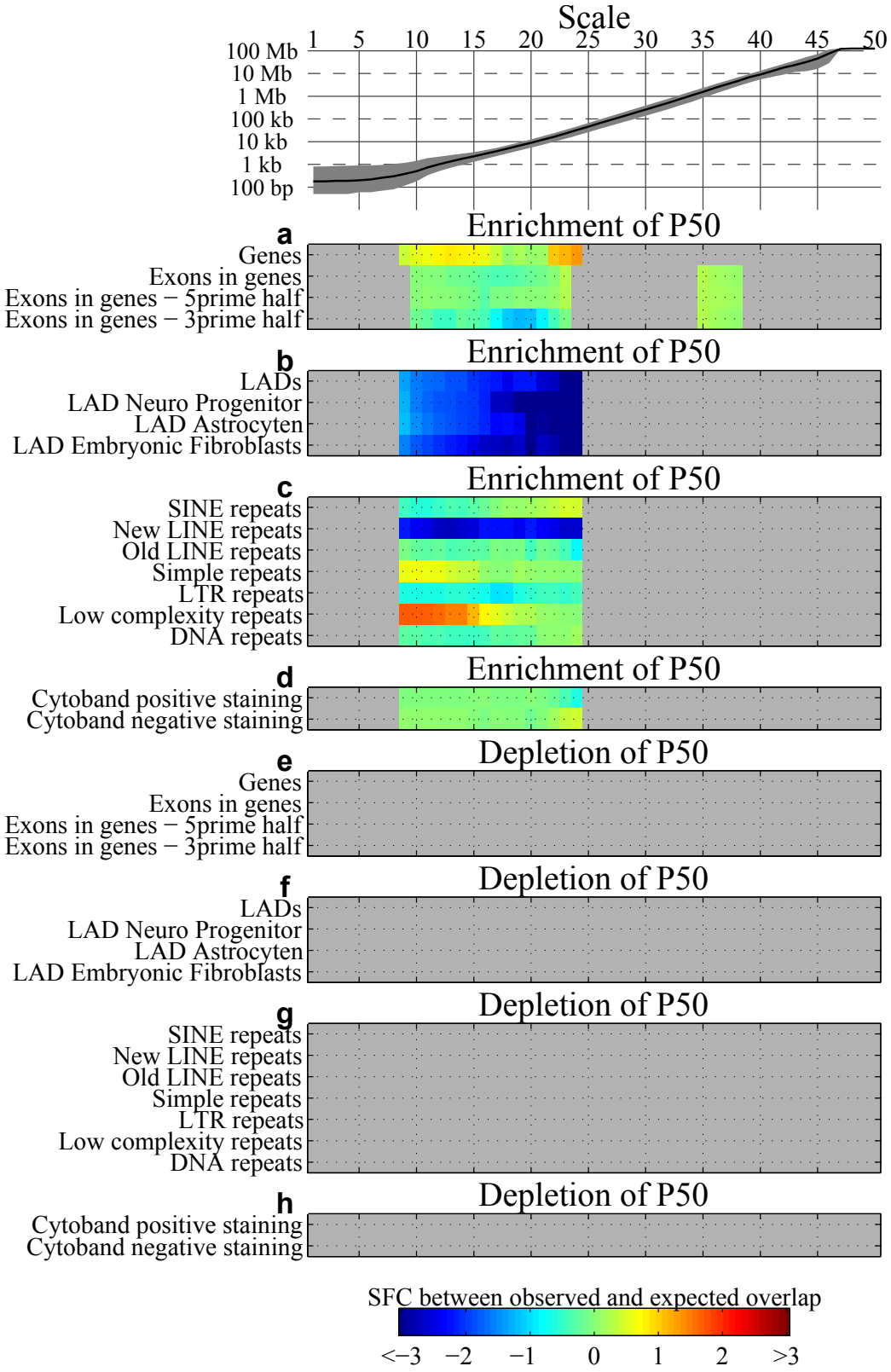
Supplementary Figure 5



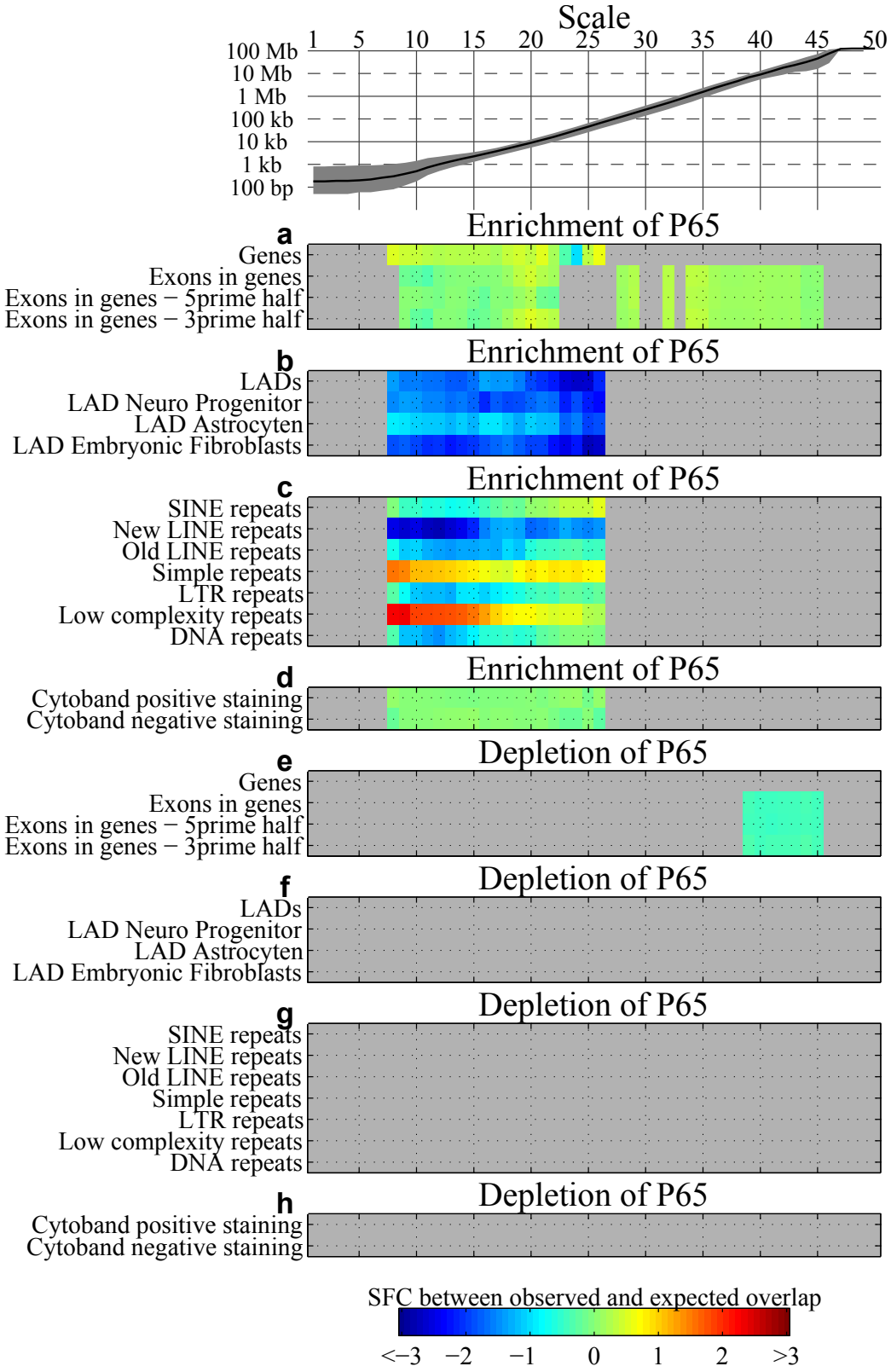
Supplementary Figure 6



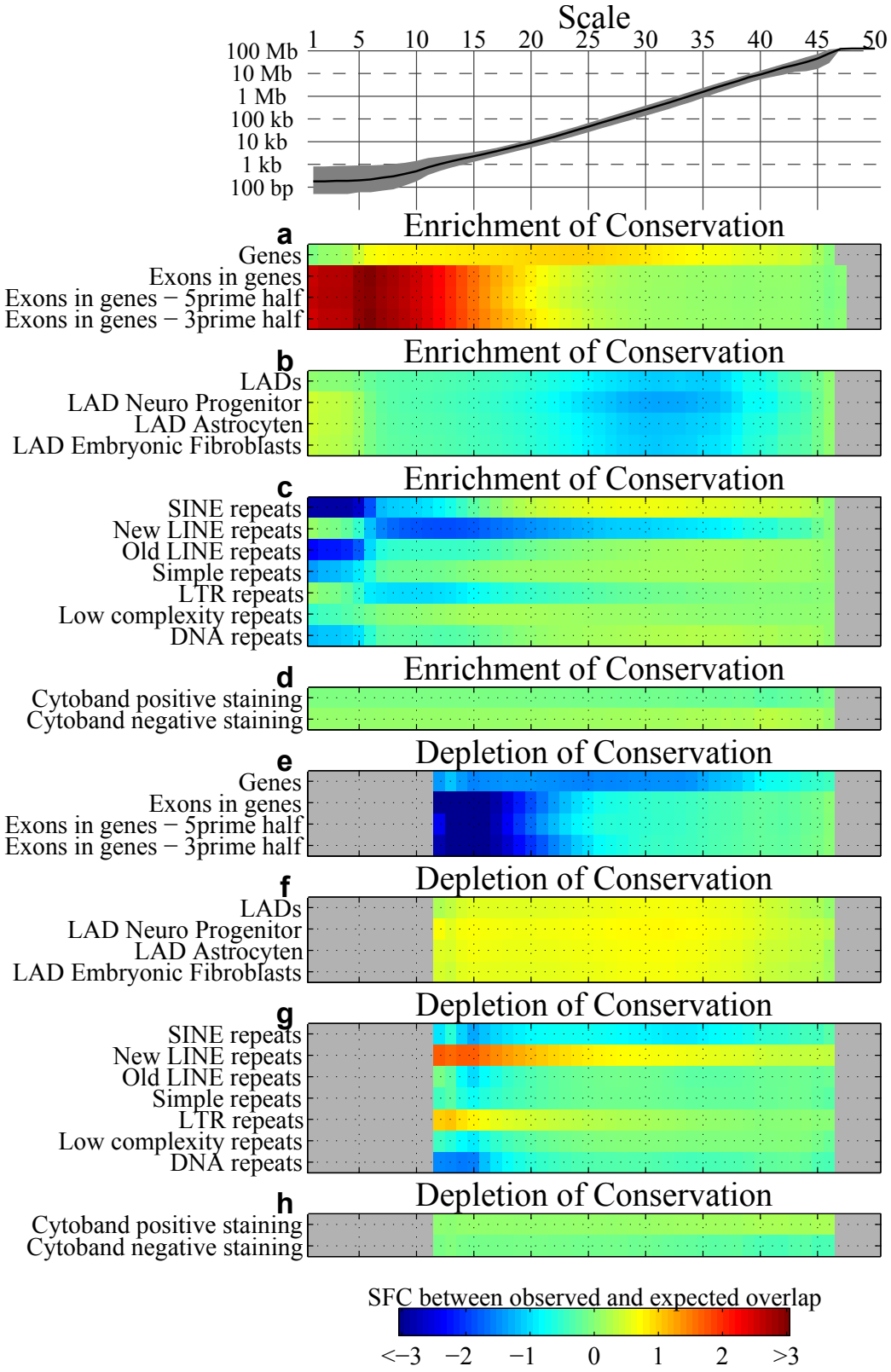
Supplementary Figure 7



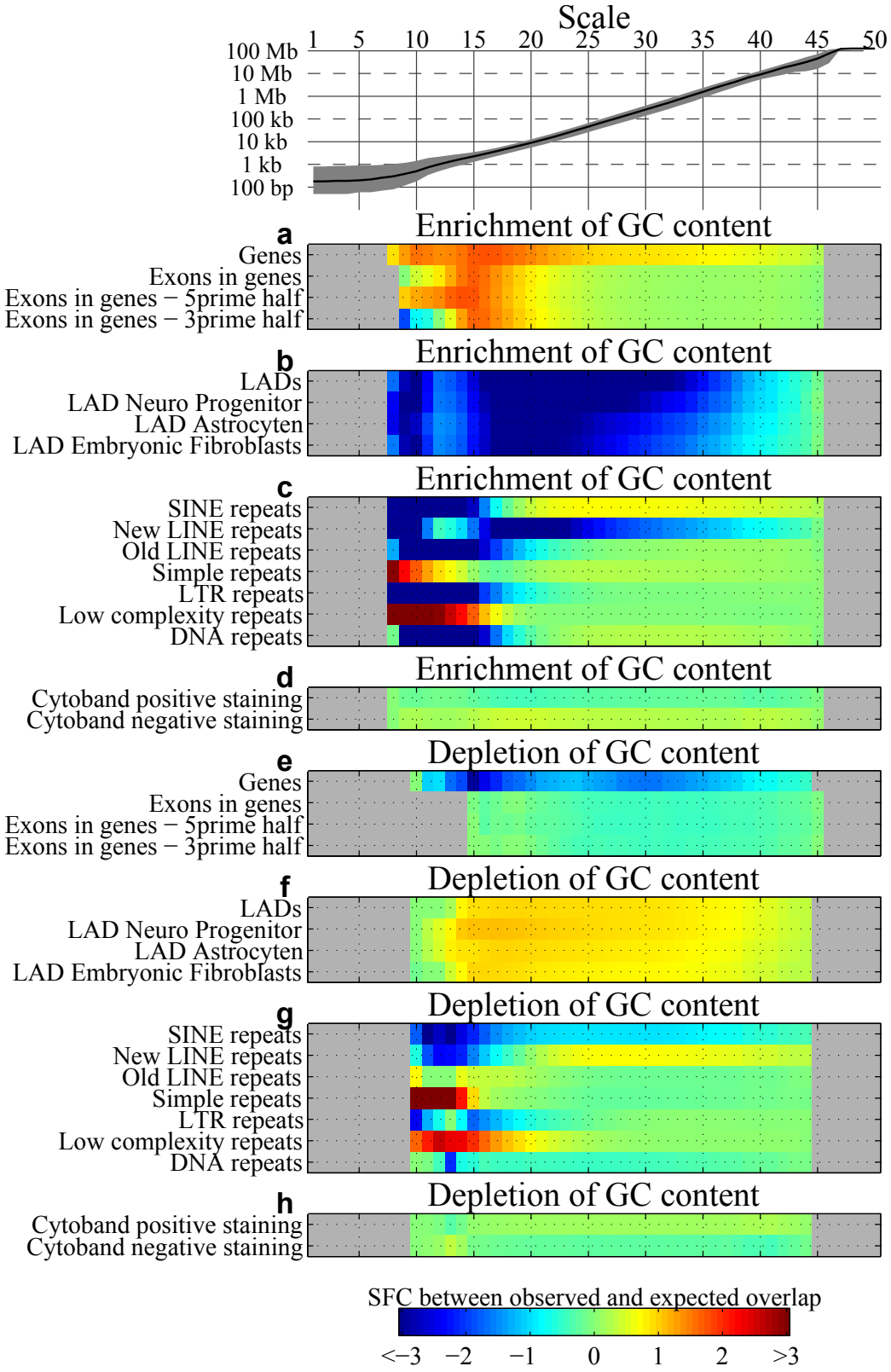
Supplementary Figure 8



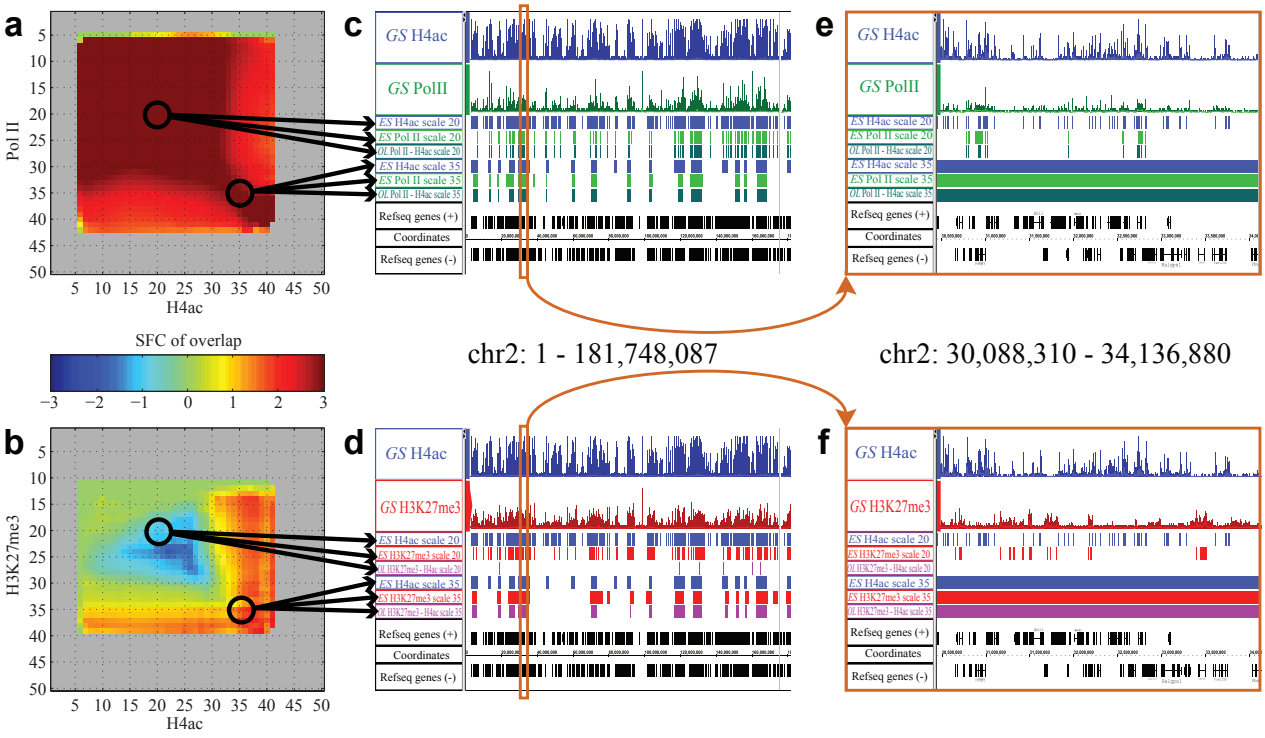
Supplementary Figure 9



Supplementary Figure 10

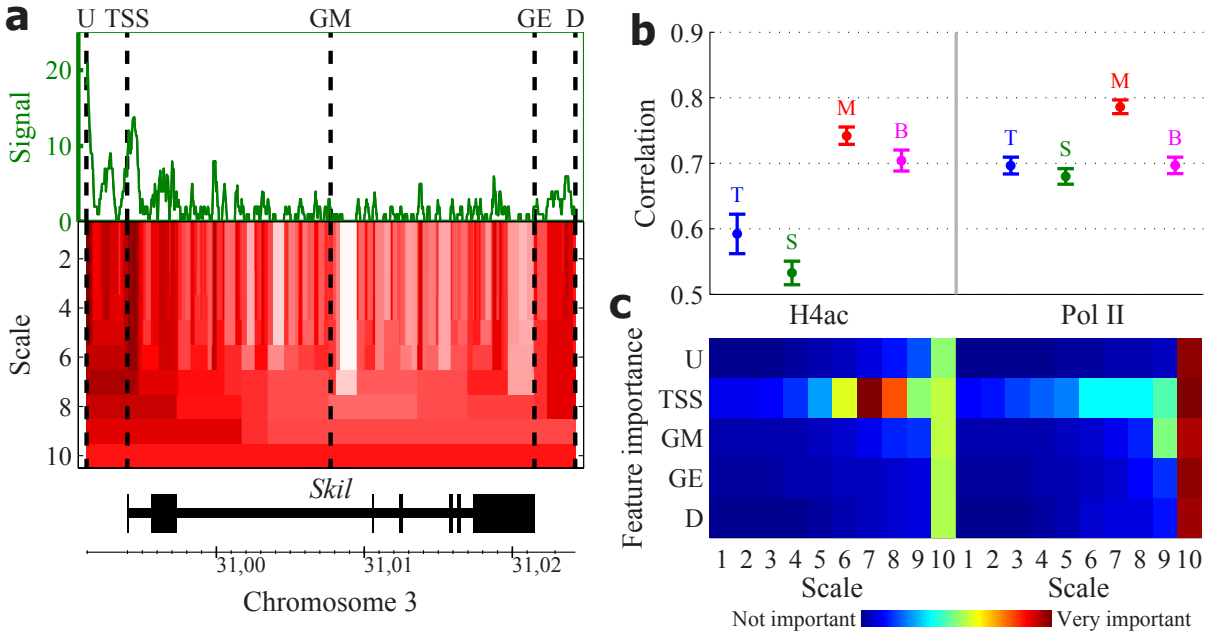


Supplementary Figure 11 | Overlap between the enriched segments of two pairs of ChIP targets.



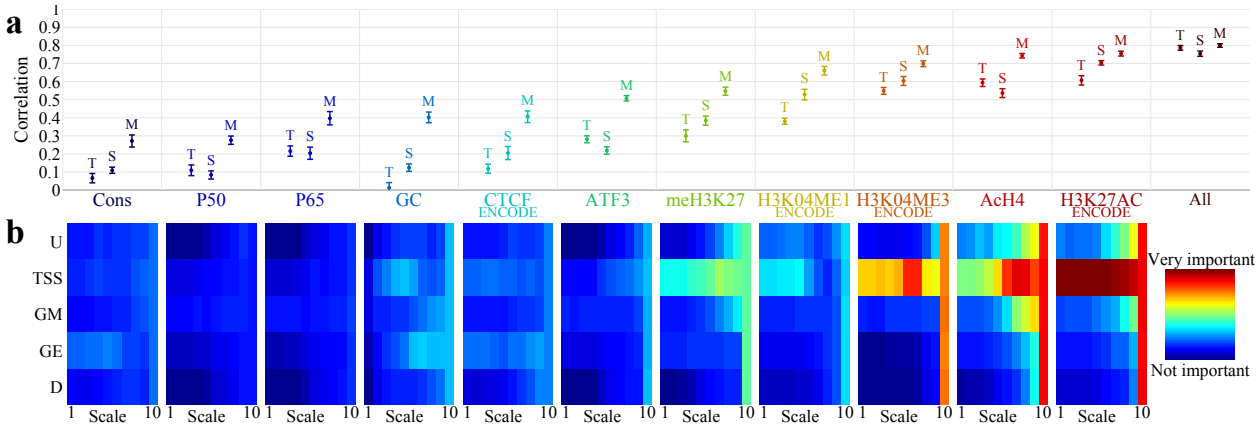
(**a**,**b**) The heatmaps depict the overlap between the enriched segments of H4ac and Pol II and of H4ac and K3K27me3, respectively, across the 50 scales. A grey color indicates that less than ten significant segments were found for one or both signals. (**c**) A genome browser view of the genomic signals (*GS*) of H4ac and Pol II for Chromosome 2 including the enriched segments (*ES*) of these signals at scale 20 and 35 as well as their overlap (*OL*) at both these scales. (**e**) A zoom-in of (**c**). (**d**,**f**) Similar to (**c**,**e**), but for the pair of H4ac and H3K27me3.

Supplementary Figure 12 | Predicting expression using gene specific multiscale representations.



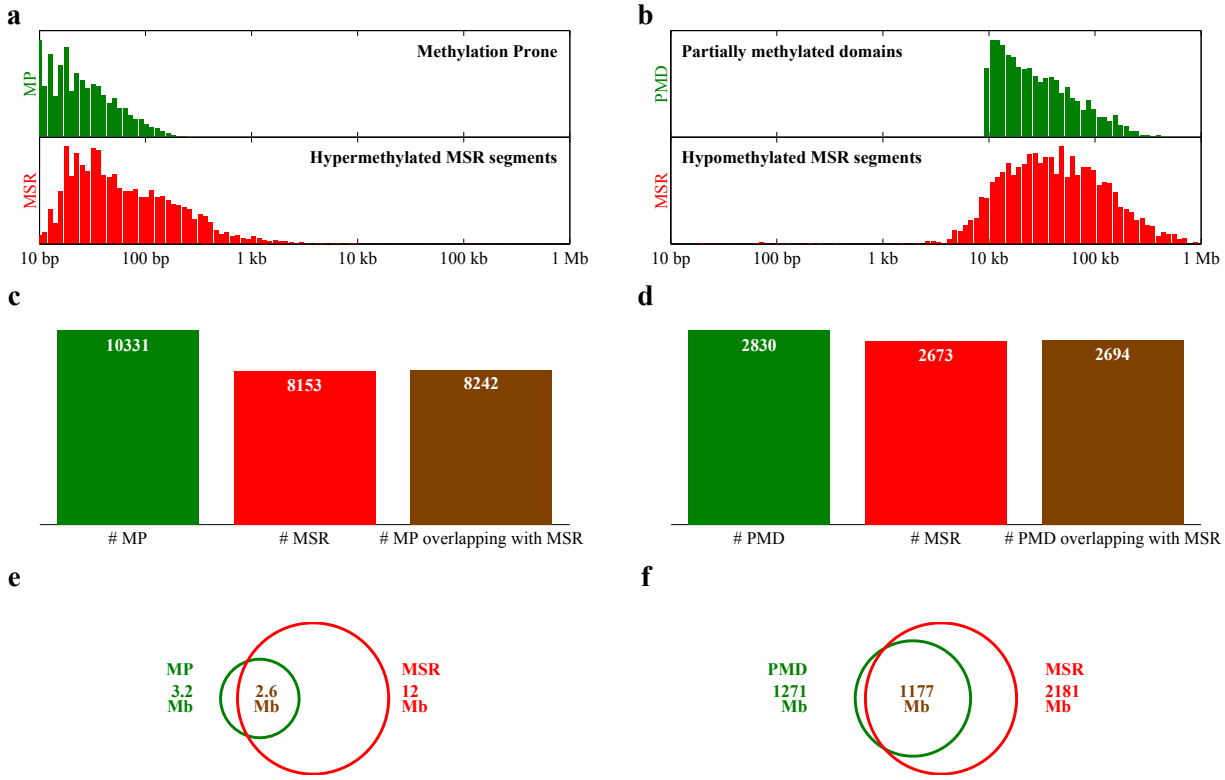
(**a**) The original genomic signal (top-panel) and MSR (bottom-panel, heatmap) of Pol II ChIP-seq data, from unstimulated macrophages, in the vicinity of the gene *Skil*. Feature values for the predictive model were derived from five positions (indicated by the black dashed lines): 1 kb upstream of the gene (U), the transcription start site (TSS), the middle of the gene (GM), the end of the gene (GE) and 1 kb downstream of the gene (D). (**b**) Mean ± standard deviation of the Pearson correlation of the test sets in the 10- fold cross-validation between the microarray expression and the predicted expression based on the total signal (T), the original signal (S), the MSR (M) and the best individual scale in the MSR (B) for genomic signals H4ac and Pol II. The best individual scale is 7 for H4ac and 10 for Pol II. (**c**) Importance of the MSR features for H4ac and Pol II in the random forest regression models.

Supplementary Figure 13 | Predicting expression using gene specific multiscale representations derived from multiple genomic signals.



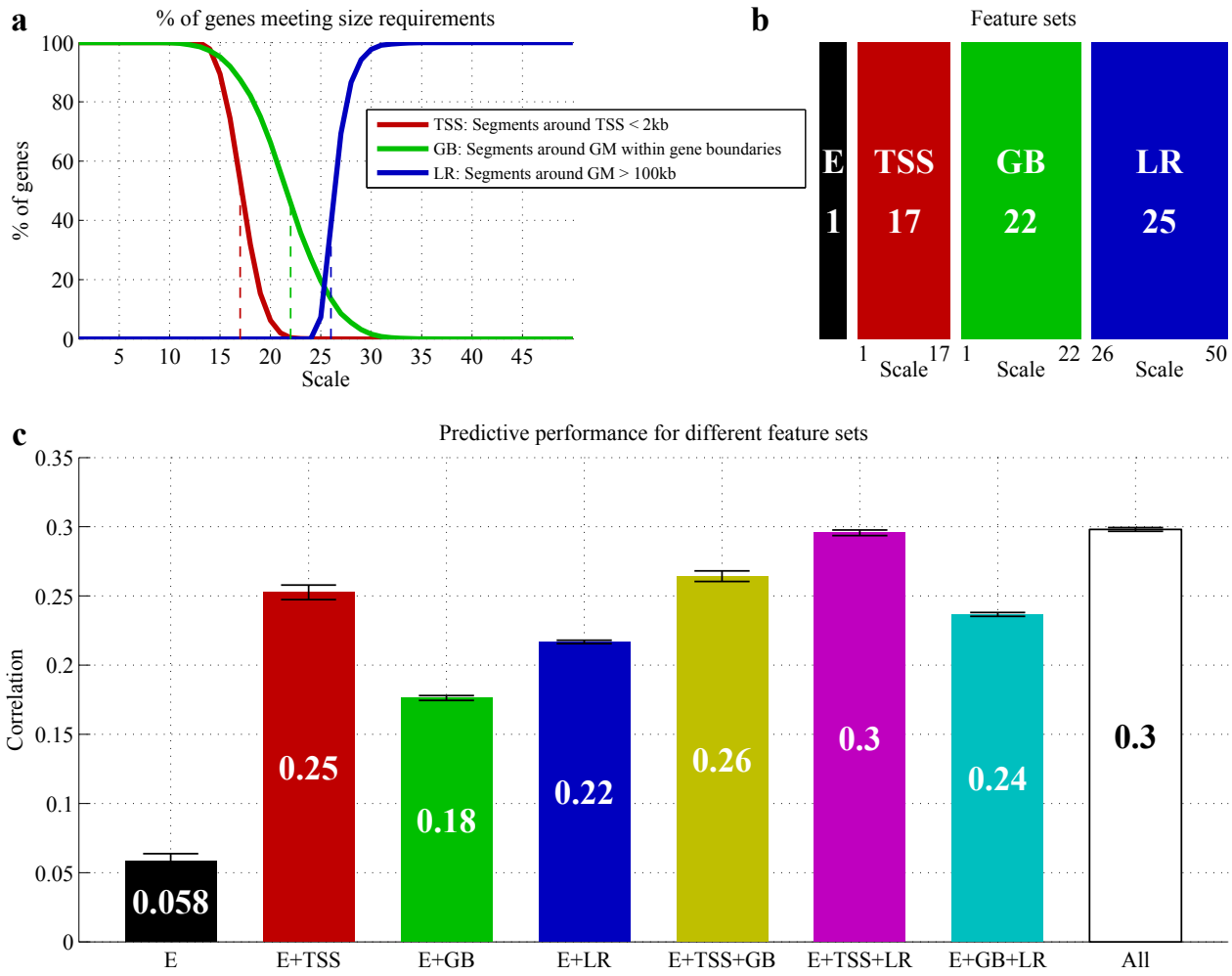
(**a**) Mean ± standard deviation of the Pearson correlation of the test sets in the 10- fold cross-validation between the microarray expression and the predicted expression based on the total signal (T), the original signal (S) and the MSR (M) for 11 different genomic signals and for the joint model of all 11 signals, called 'All'. The models are ranked based on performance. Genomic signals obtained from ENCODE are labeled correspondingly. (**b**) Importance scores of the MSR features for the 11 genomic signals in the joint ('All') model.

Supplementary Figure 14 | Comparing methylation prone segments (MPs) and partially methylated domains (PMDs) with the MSR segments



(**a**) Distribution of the sizes of the MPs and the hypermethylated MSR segments. (**b**) Distribution of the sizes of the PMDs and the hypomethylated MSR segments. (**c**,**d**) Number of identified segments as well as the overlap. (**e**,**f**) Overlap between the segment sets in base pairs.

Supplementary Figure 15 | Random forest model predicts differential expression of genes between tumor and normal based on their DNA methylation MSR

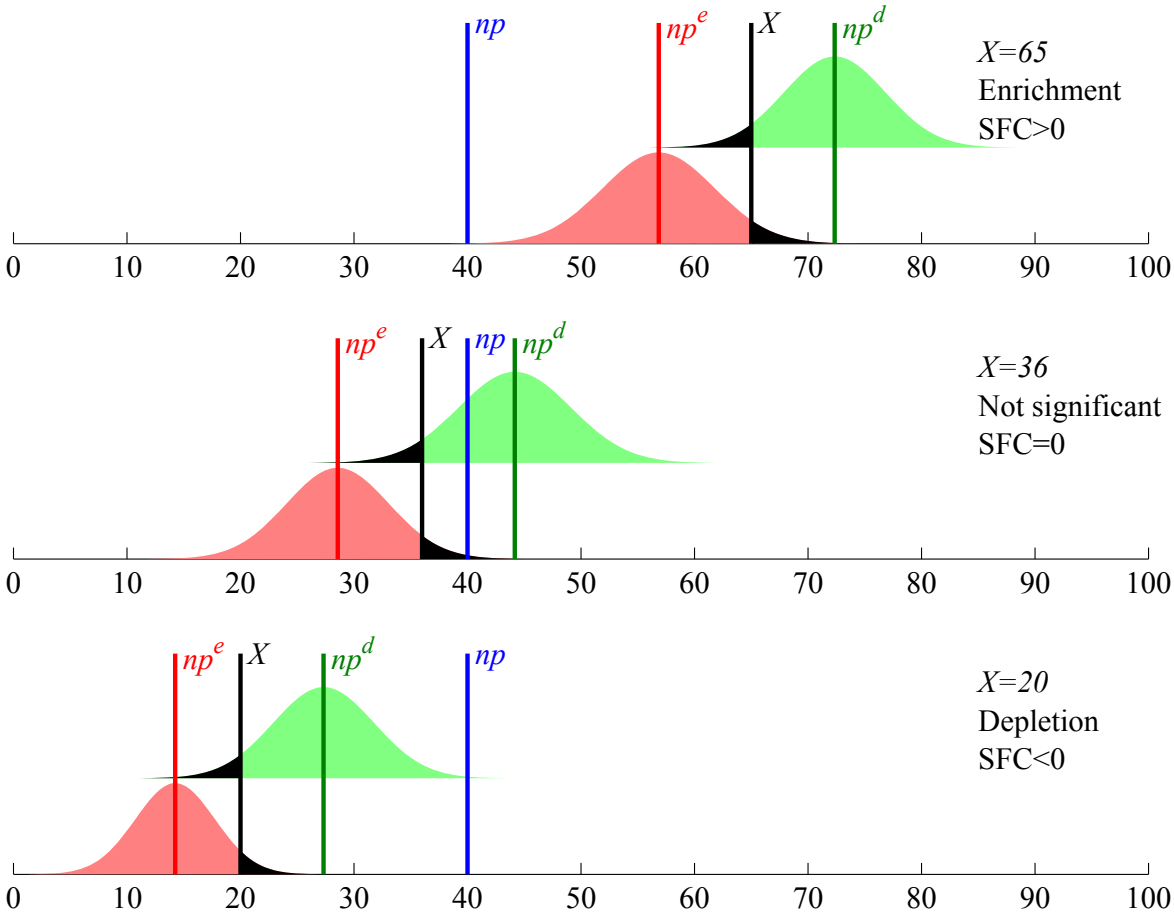


(**a**) Three feature sets were created based on the MSRs that capture the differential methylation of segments overlapping with the TSS or the GM across all (50) scales for all (9111) genes:

4. TSS. The MSR around the TSS from scale 1 to scale 17, since at least 50% of the genes have segments smaller than 2 kb in this range
5. GB (gene body). The MSR around the GM from scale 1 to scale 22, since at least 50% of the genes have segments within the gene boundaries in this range
6. LR (long range). The MSR around the GM from scale 26 to scale 50, since at least 50% of the genes have segments larger than 100 kb in this range.

(**b**) Visual representation of the feature sets used in the random forest regression model to predict the differential expression between tumor and normal. E: the gene expression of normal tissue (1 feature); TSS (17 features); GB (22 features); and LR (25 features). (**c**) Pearson correlation of the test sets in the 10- fold cross-validation between the differential expression and the predicted differential expression based on combinations of different feature sets. The numbers (mean ± standard deviation) are based on 3 repeats of the 10-fold cross-validation scheme.

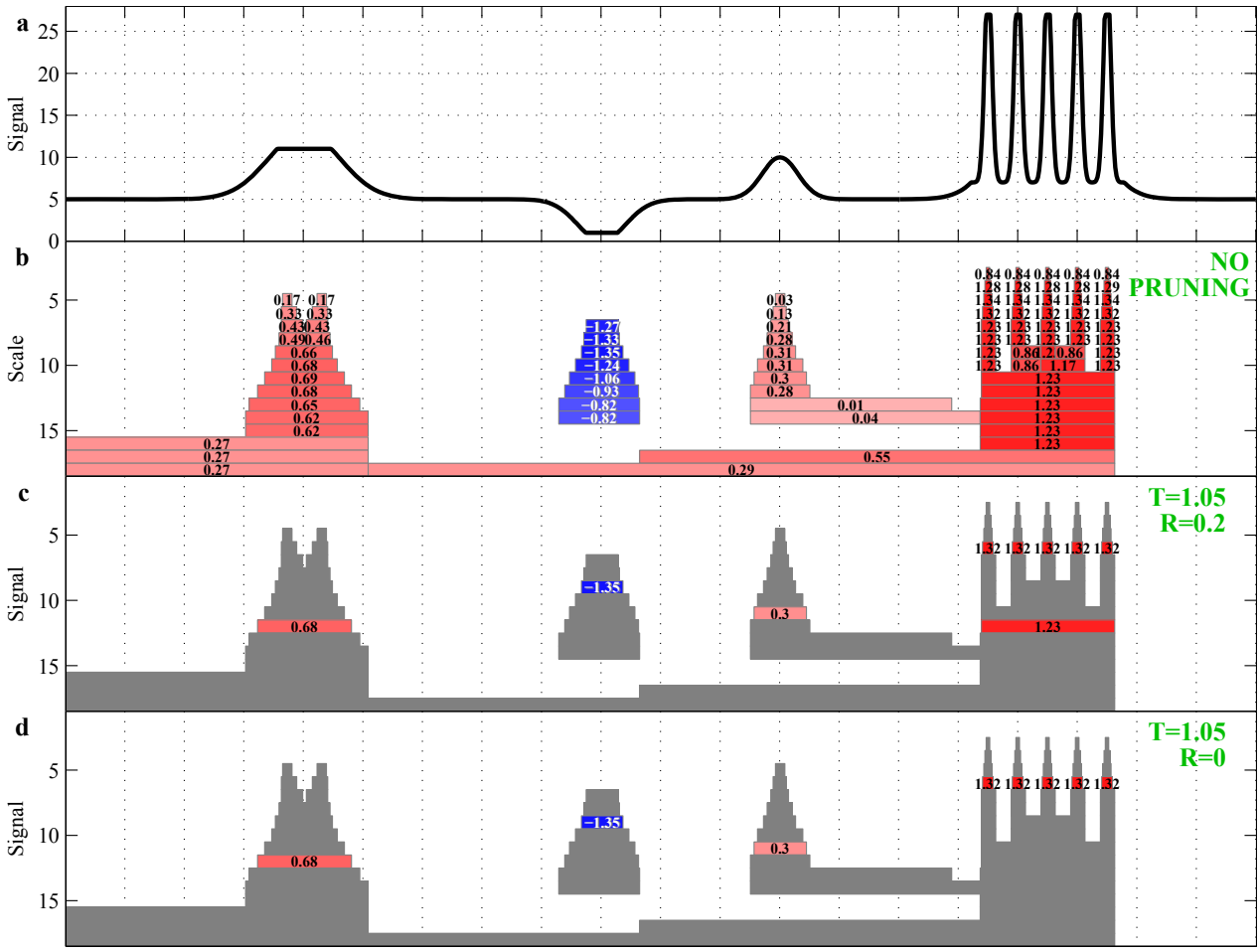
Supplementary Figure 16 | Visual explanation of SFC for enrichment or depletion of signal intensity



This figure depicts the three different scenarios from equation (3) in the main text to compute the SFC. In all cases, the expected mean background intensity, np, is 40. The actual observed intensity X is 65, 36 and 20 for the top, middle and bottom panel, respectively. Probabilities p^e and p^d are computed as in equations (1) and (2) in the main text. np^e represents the mean observed intensity, such that X is the upper bound of the normal distribution with mean np^e at the P-value threshold. This normal distribution is depicted in red. The black area to the right of X is equal to p^{th}, the user defined P-value threshold. If $np < np^e$ there is significant enrichment, i.e. SFC>0 (top panel). np^d represents the mean observed intensity, such that X is the upper bound of the normal distribution with mean np^d at the P-value threshold. This normal distribution is depicted in green. The black area to the left of X is equal to p^{th}. If $np^d < np$ there is significant depletion, i.e. SFC<0 (lower panel).

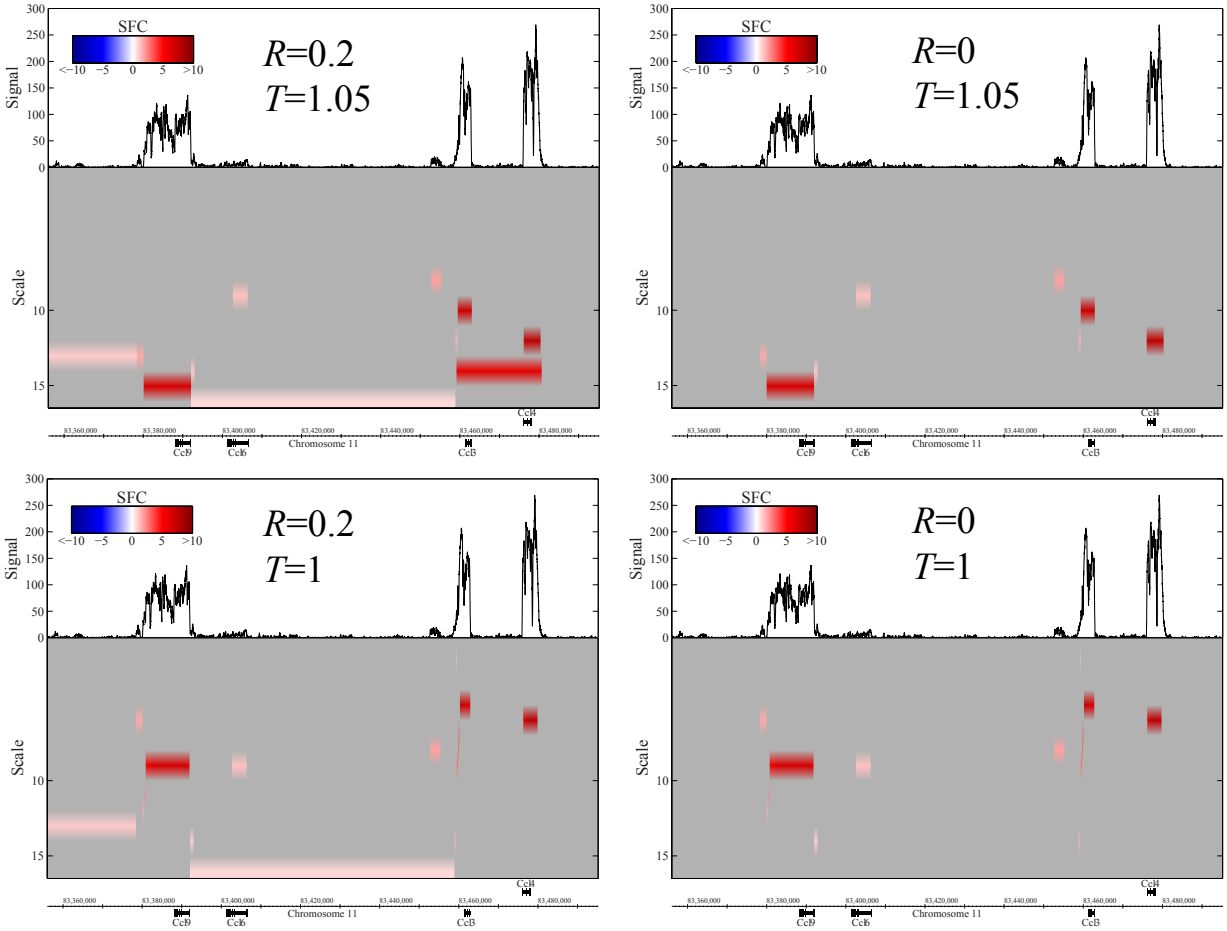
SUPPLEMENTARY FIGURE **X** LEGENDS (ONLY REFERENCED IN SUPPLEMENT)

Supplementary Figure X1 | Effect of pruning on example genomic signal



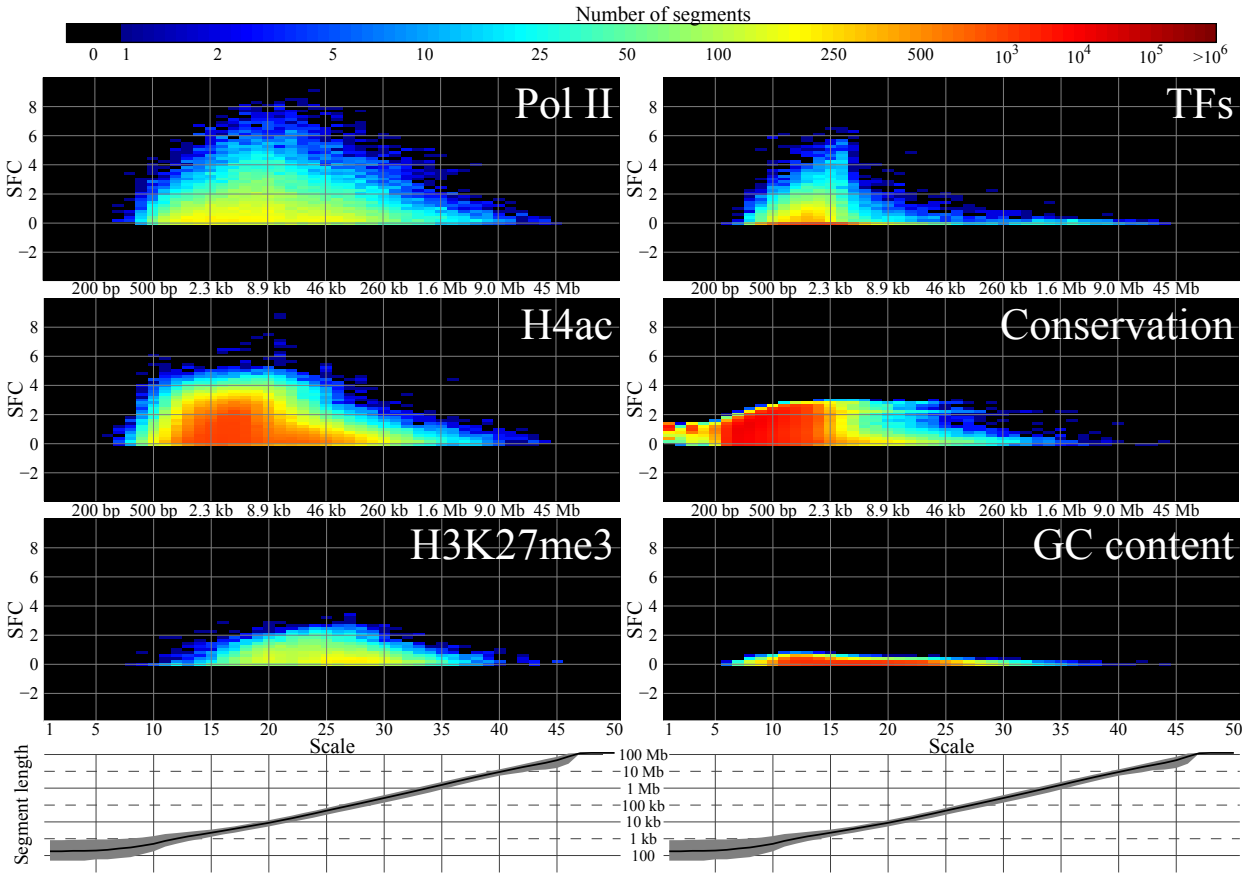
(**a**) An example genomic signal. (**b**) The MSR of the genomic signal. Segments with a non-zero SFC score are depicted in blue (SFC<0) and red (SFC>0). The SFC score is printed in the segments. (**c**) Segments removed by pruning using default settings (T=1.05, R=0.2) are depicted in grey. Note that the size of the remaining segments agrees well with the genomic signal. Further, the five little peaks on the right of the signal lead to five little segments on scale 6 and a larger segment on scale 12. (**d**) Segments removed by pruning without size constraint (R=0) are depicted in grey. With these settings the larger segment on scale 12 is pruned, because the smaller segments corresponding to the five little peaks have a higher score. When R=0 a genomic location can only belong to at most one significant segment.

Supplementary Figure X2 | Pruned MSR of genomic signal from **Fig. 1**



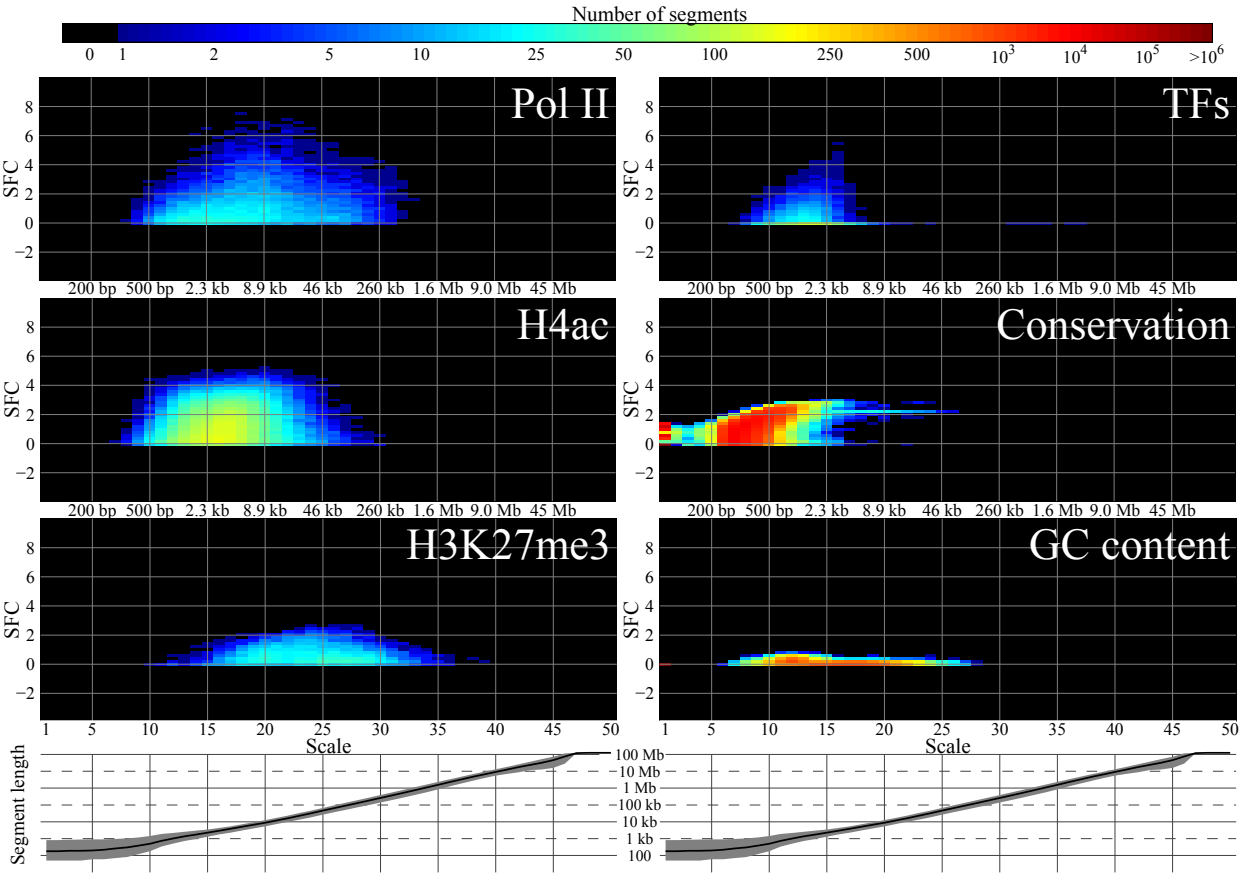
(**top-left**) Pruning with default settings (T=1.05, R=0.2); pruned segments are depicted in grey. (**top-right**) Pruning without size constraints (T=1.05, R=0). (**bottom-left**) Pruning without slack (T=1, R=0.2). (**bottom-right**) Pruning without size constraints and without slack (T=1, R=0). The genomic signal in this region was selected specifically to explain the MSR and the effect of pruning parameters R and T. In this region, these parameters have a large effect on which segments are pruned. This is, however, not representative for these genomic signals in general, where the changes are typically much smaller. See also **Supplementary Fig. X5**.

Supplementary Figure X3 | Pruned version of the MSR signatures from **Fig. 2**



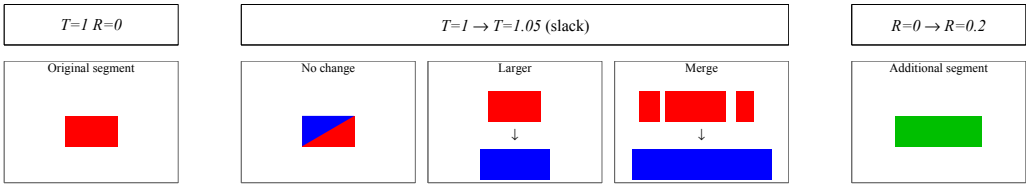
The heat map diagrams show the two-dimensional histograms created by binning the segments based on their scale and on their SFC. Before creating these histograms, pruning with default settings (T=1.05, R=0.2) was applied to the MSRs. This means that segments that were pruned are not used in creating the histograms. Note that the pruning strategy was only applied to detect enriched segments, i.e. depleted segments (SFC<0) were all pruned.

Supplementary Figure X4 | Stringently pruned version of the MSR signatures from **Fig. 2**



Same as **Suppl. Fig. X3** except pruning without size constraints (T=1.05, R=0) was used to create the histograms.

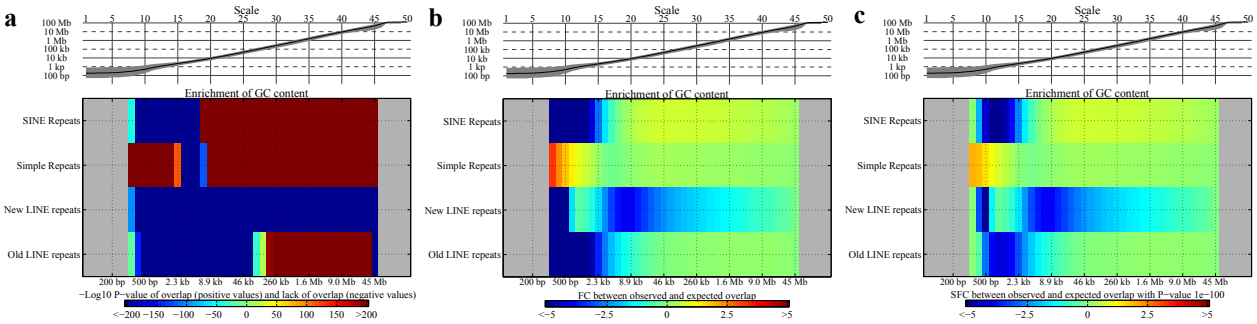
Supplementary Figure X5 | Breakdown of the effect of pruning parameters *R* and *T* on the MSR segments.



P65	903 [0.13−0.28kb]	95%	4.9% [10−34%]	0.089% [0.31−1.7%] [2−2]	34 (3.8%) [517−4344kb]
P50	566 [0.11−0.19kb]	90%	10% [15−33%]	0.051% [0−0%] [2−2]	17 (2.9%) [2−23kb]
ATF3	1958 [0.13−0.26kb]	95%	4.5% [15−36%]	0.1% [6.2−25%] [2−2]	36 (1.8%) [9−1236kb]
H3K27me3	2854 [5.8−75kb]	57%	30% [4.9−36%]	13% [0.097−12%] [2−3]	175 (6.1%) [112−478kb]
H4ac	16862 [0.48−2kb]	53%	33% [8.3−40%]	14% [3.8−23%] [2−4]	975 (5.8%) [24−100kb]
Pol II	4936 [0.48−8.5kb]	51%	20% [5.4−37%]	29% [0.97−8.7%] [3−8]	341 (6.9%) [28−190kb]
Explanation	# of detected segments [IQR of segment sizes]	% of unchanged segments	% of enlarged segments [IQR of % enlargement]	% of merged segments [IQR of % enlargement] [IQR of % of merged segments]	# of new segments (%) [IQR of segment sizes]

This figure lists for each of 6 different ChIP targets the number and size of enriched segments when no slack (*T*=1) and size constraints (*R*=0) are applied (**left**), how these segments change when slack is applied (*T*=1.05) (**middle**), and how many additional segments are detected when the size constraint is applied (*R*=0.2) (**right**). All these numbers are averages across the multiple experimental conditions under which these targets were measured. IQR stands for interquartile range.

Supplementary Figure X6 | Overlap between functional genomic regions and the segments using the SFC and other scores.



(**a**) The heatmap depicts the –log10 hypergeometric test P-value for enrichment of overlap (positive scores) and lack of overlap (negative scores). The data used for this figure are the same as in **Fig. 3b** in the main text. (**b**) The heatmap depicts the log2 of the fold change (FC) between the observed and expected overlap. (**c**) The heatmap depicts the SFC score between the observed and expected overlap. This is the same heatmap as **Fig. 3b**, except that a much lower P-value was used to compute SFC score.

CITATIONS

1. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
2. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952-1958 (2009).
3. Rashid, N.U., Giresi, P.G., Ibrahim, J.G., Sun, W., and Lieb, J.D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol* **12**, R67 (2011).
4. Yang, S. *et al.* Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res.* **14**, 517 (2004).
5. Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F., and Bernardi, G. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. *Proceedings of the National Academy of Sciences* **79**, 355-359 (1982).
6. Gardiner-Garden, M. and Frommer, M. CpG Islands in vertebrate genomes* 1. *Journal of molecular biology* **196**, 261-282 (1987).
7. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
8. Ramsey, S.A. *et al.* Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. *PLoS Computational Biology* **4**, e1000021 (2008).
9. Breiman, L. Random forests. *Machine learning* **45**, 5-32 (2001).
10. Bernstein, B.E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
11. Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K., and Vingron, M. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2926-2931 (2010).
12. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13**, R53 (2012).
13. Cheng, C. *et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology* **12**, R15 (2011).
14. McLeay, R.C., Lesluyes, T., Partida, G.C., and Bailey, T.L. Genome-wide in silico prediction of gene expression. *Bioinformatics* 2012).
15. Berman, B.P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature genetics* **44**, 40-46 (2011).
16. Berman, B.P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* **44**, 40-46 (2012).
17. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**, 484-492 (2012).
18. Ball, M.P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361-368 (2009).
19. Hon, G.C. *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* **22**, 246-258 (2012).
20. Robertson, A.G. *et al.* Genome-wide relationship between histone H3 lysine 4 mono-and tri-methylation and transcription factor binding. *Genome Res.* **18**, 1906-1917 (2008).
21. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651-657 (2007).