Supplementary Information for "EpiSAFARI: Sensitive detection of valleys in epigenetic signals for enhancing annotations of functional elements"

Arif Harmanci[1], Akdes Serin Harmanci[2], Jyothishmathi Swaminathan[3], Vidya Gopalakrishnan[3,4,5,6,7]

1 School of Biomedical Informatics, Center for Precision Health, University of Texas Health Science Center, Houston, TX.
2 School of Biomedical Informatics, Center for Systems Medicine, University of Texas Health Science Center, Houston, TX.
3 Department of Pediatrics, M.D. Anderson Cancer Center, Houston, TX.
4 Department of Molecular and Cellular Oncology, M.D. Anderson Cancer Center, Houston, TX.
5 Brain Tumor Center M.D. Anderson Cancer Center, Houston, TX.
6 Center for Cancer Epigenetics, University of Texas, M.D. Anderson Cancer Center, Houston, TX.
7 M.D. Anderson UTHealth Graduate School of Biomedical Sciences, Houston, TX.

We present the Supplementary Information, Methods, and Supplementary Figures.

Supplementary Methods and Information

Valleys, Troughs, Canyons and Their Length Scales

For clear presentation, we discuss below the nomenclature and definitions of valley shaped signal patterns, namely valleys, troughs, and canyons. We also discuss the length scales of valley shaped signal patterns.

Valleys. Given a functional genomics signal profile (such as from a ChIP-Seq experiment) a valley is basically a region in the genome that shows a "V" shaped signal pattern as shown in Figure 1. At the ends of the valley are the summits (local maxima) of the signal profile. There is also one dip within the valley which harbor the smallest signal within the valley. A "good" valley exhibits

monotonic decrease in the signal level when moving from either of the summit positions to the dip position.

Troughs. We treat the “trough” and “valley” as identical. To the best of our knowledge, troughs and valleys are used interchangeably in the literature (1).

Canyons. The canyons are similar to the valleys except that the basin of the valley is a broad region with depletion of signal. The canyons are studied in the context of DNA methylation data analysis. Unlike valleys, the canyons have much steeper hills compared to valleys when moving from the edge of the canyon to the base of the canyon. Thus, the signal profile over canyons are similar to a broad “U” shaped pattern with large basin (2, 3). A recent publication referred to these broad canyon domains as “nadirs” (4).

Length scales of valleys. As functional genomics signals exhibit dynamic patterns along the genome, almost all of the functional genomics signal profiles exhibit valley patterns. The important question about these valleys is whether the valleys are biologically meaningful. EpiSAFARI represents a general method for sensitive detection of the valleys. In the current manuscript, we have focused on punctate valleys that are generally shorter than 5 kilobases. The main reason for this is that the punctate valleys have been referenced in the literature and they potentially correspond to functional cis-regulatory elements such as promoters, enhancers, and insulators. We believe that the default parameters are sufficient to analyze the punctate cis-regulatory elements and their properties with respect to the valleys. The valleys for other types of assays can have different lengths depending on the assay type and biological context that is being analyzed. For other types of data, it is therefore important to visually evaluate the signal using a genome browser (such as IGV) and make sure that the parameter selection supports the valleys in question.

An example of other type of valleys are the valleys manifesting in RepliSeq experiments (5) where relative replication timing of each position in the genome are measured in a high-throughput manner using a next-generation sequencing based assay. The locations in the genome that replicate early show higher signal than those replicate later. Therefore, evaluation of a RepliSeq signal profiles shows that large valleys correspond to regions in the genome replicate later than the peak regions. These valleys manifest at much larger length scales and they may extend several megabases. We currently did not address detection of the valleys at this length scale in this manuscript.

Yet another example of biologically important valleys are observed in digital genomic foot-printing experiments(6, 7). In these experiments, DNA that is open and accessible is sequenced. This enables probing where the transcription factors can interact with DNA. The functionally important valleys in the DNase signal are extremely narrow regions (5-20 base pairs long) and they are hypothesized to correspond to transcription binding sites. The reason why we observe a valley in the DNase signal is that the bound transcription factor causes a very narrow stretch of the DNA (at the location of the binding motif) to be inaccessible to the assay. Indeed, these extremely punctate valleys are shown to be associated with exact locations of transcription factor binding and positions where proteins interact with DNA upon binding. Related to probing of the open chromatin, the genomewide measurement of the nucleosome positioning, through assays such as MNase-Seq(8, 9) have also been shown to produce valley shaped patterns(10, 11). In essence, these assays measure where the nucleosomes are positioned on the genome. This is an important task for mapping the active open chromatin domains in the genome and for understanding the biological determinants of how nucleosomes are remodeled and managed by the cells. The punctate valley patterns (100-200 base pairs) in these experiments are especially well-studied with regard to their impact on gene regulation stemming from nucleosome positioning at the transcription start sites and transcription end sites(10).

Signal Smoothing by EpiSAFARI by Overlapping Windows

While smoothing the signal, the starting position of the smoothing window is updated using a stepping length that is smaller than the window length. This way, each position on the genome is covered by multiple smoothing windows. For each position, the final smoothed value is assigned as the maximum of the smoothed values assigned by all the windows covering the position. By default, the stepping length is set to l_w, i.e., non-overlapping windows.

Selection of the Spline Parameters

The spline parameters determine the location of the valleys and may impact the accuracy. In addition, the hill scores and valley asymmetry may be affected by the spline smoothing parameters. Thus, we studied the impact of knot numbers, knot positioning, and spline degree on the accuracy of detecting valleys. To determine the effect of spline parameters on the valley accuracy, we used the H3K4me3 data for the K562 cell line from the ENCODE Project. To generate a ground truth set for the H3K4me3 valleys, we used the active genes promoters that overlap with any transcription factor binding peak as detected from ChIP-Seq datasets. The basic motivation for using these regions as ground truth is following: It is generally known that the promoters of active genes are enriched with H3K4me3 histone modification. If we put the additional requirement that these promoters overlap with transcription factor binding, these promoters most likely contain a valley inside them.

To generate the ground truth regions, we first downloaded the gene expression levels for K562 cell line from the ENCODE Project. We then identified the genes whose average replicate expression level is greater than 0.05. We finally extracted the promoters of these active genes and overlapped them with the ChIP-Seq transcription factor peak regions for K562 cell line from the ENCODE project. The intersecting promoters are used as the ground truth. We denote the genomic locations for the ground truth set of promoter regions by $\boldsymbol{P}$. It must be noted that these regions do not necessarily correspond to a complete set of the H3K4me3 valleys for K562 since this mark can also manifest on the enhancers in the intergenic domain. Thus, the valleys that EpiSAFARI detects will most likely contain many valleys that do not overlap with this ground truth. For this reason, we will evaluate the sensitivity of the valleys, i.e., the fraction of the set of ground truth regions that overlap with the detected valleys while evaluating the parameter selection.

After building the ground truth set, we next ran EpiSAFARI to detect the valleys in the H3K4me3 signal profile of K562 cell line with changing spline degree, knot number, and knot placement. To decrease the computation time, we focused only on the chromosome 1 for these analyses.

Knot Locations. To evaluate the effects of knot locations, we evaluated 3 knot placement strategies. First is derivative based knot selection. In this knot selection, we place the knots where the read depth signal shows fast changes along the genome. In this knot placement, EpiSAFARI places the knots at the locations for which the signal has the largest absolute signal derivative. We next implemented the random knot placement where the knots are randomly placed along the domain of the signal (Supplementary Fig.

1). We finally included the uniform knot placement where the knots are placed at equal intervals within the domain of the signal.

Knot Numbers. In order to include a wide range of knots distributed along the domain of the read depth signal, we used between 3 (minimum that we can use) and 15 knots. This way we evaluate both the densely and sparsely positioned knot selections. We denote the knot number with κ.

Spline Degree. For each knot selection, we use spline degrees between 1 and 7. This way, we assess whether the increasing degree of the splines increase the sensitivity of the valley detection. We denote the spline degree with ψ.

We ran EpiSAFARI with the all the knot selection, knot number, and spline degree parameter combinations and computed the sensitivity of the identified valleys from each parameter combination. We next computed the sensitivity of the valleys as:

$$Sensitivity(\boldsymbol{V} \mid \{k_1, k_2, \dots, k_\kappa\}, \psi) = \frac{|\boldsymbol{P} \cap \boldsymbol{V}|}{|\boldsymbol{P}|} \quad (18)$$

where $\boldsymbol{V}$ denotes the set of valleys (i.e., the genomic coordinates of the valleys) that are identified by EpiSAFARI with κ knots positions denoted by $\{k_1, k_2, \dots, k_\kappa\}$ and spline degree ψ. $\boldsymbol{P}$ denotes the set of active promoters (the genomic coordinates) that are bound by transcription factor peaks. $|\boldsymbol{P} \cap \boldsymbol{V}|$ denotes the number of active and TF bound promoters that overlap with the $\boldsymbol{V}$ and $|\boldsymbol{P}|$ denotes the number of promoters in $\boldsymbol{P}$.

We computed the sensitivity of the valleys detected using knot selection and spline degrees. (Supplementary Figure 3a, b, c). When the knot number and spline degree are both small, the sensitivity is smallest at around 0.2. As the number of knots or the spline

degree increases, the sensitivities increase reaches around 0.8. This indicates that the overly simple smoothing is not powerful enough to detect the valleys. However, as we increase the complexity of smoothing, the sensitivity saturates at around 0.80 and starts decreasing as the smoothing is made more complex. This result highlights that increasing complexity of smoothing splines may decrease the sensitivity of valley detection. When different knot selection approaches are compared, the derivative based knot placement does not show improved performance over the uniform and random knot placement strategies. We also evaluated the number of valleys that are detected by EpiSAFARI using different parameter combinations (Supplementary Fig. 3d, e, f). This is important because we want to also compare the number of valleys identified using different parameters. We observed that the number of valleys increases as the knot number and spline degree increases. The number of valleys (and sensitivity) decreases when we use parameter configurations with more than 7 knots and spline degree of 6 and higher.

In summary, we observed that extra complexity does not provide much improvement for our sensitivity analysis and in fact increasing complexity too much may cause overfitting of the data and may decrease the quality of selected valleys. Putting all these considerations together, we decided to use uniform knot selection with number of knots set to 7 and spline degree as 5. This selection is motivated to make balance between the accuracy, the number of valleys, and also the computation time that is required to run the algorithm (Supplementary Fig. 3g, h). The users can change the parameters to make EpiSAFARI run more conservatively or in a relaxed fashion.

It is worth noting that there are knot placement strategies other than the ones that we evaluated here(12). As we discussed before, the knot placement in spline smoothing is

an open problem that is currently not solved in general cases. However, our results show that when we use a set of basis splines that are reasonably complex, i.e. not-very-low spline degree and knot numbers, the placement strategy does not impact the sensitivity considerably.

EpiSAFARI divides the genome into non-overlapping l_w long windows and performs smoothing for each window independently. Within each window, the knots are independently added. We evaluated whether the knot selection impacts the valley detection. For this we analyzed the relative distribution of the valley dips within their corresponding l_w long windows. Supplementary Figure 3i, j, and k show the distribution of the relative valley dip locations for uniform, derivative, and random knot selection procedures. For uniform knot selection, there is a clear periodic pattern in the distribution of the relative dip locations. This pattern stems from the fact that the knots are positioned at the same positions in each window, i.e., uniformly distributed within l_w long window. For derivative and random knot selections, the periodic pattern does not exist. However, there is slight enrichment of dips close to the ends of the windows. These biases are removed when we use overlapping window-based smoothing (Supplementary Figure 3l).

Window Length Selection. Window length parameter, l_w, directly relates to smoothing as it determines the chunk of signals that will be smoothed at every step of smoothing. We computed the sensitivity of the detected valleys with changing l_w parameter (Supplementary Fig. 4). As $l_w < 1000$, the sensitivity increases with increasing window length, after $l_w > 1000$, the sensitivity starts decreasing. The main reason for this is possibly that the spline smoothing is underfit, i.e., the number of knots (and basis functions) is not large enough to reliably smooth the signal. From this observation, we

suggest usage of $l_w = 1000$ for punctate histone modifications. The selection of window length for sparse signals should be increased to increase the number of points of interest in each window so that the smoothing can be performed reliably. In addition, the expected valley lengths must be taken into consideration. For DNA methylation signals, we observed that for $l_w = 5000$ is sufficient with the knot number of 7 and spline degrees of 5.

Impact of Smoothing Parameters on the Hill Score

The hill score is computed for each valley separately using the smoothed signal profiles (Supplementary Figure 5a, b). Thus, the effect of the smoothing parameters on the computed hill scores is important. To compare the hill score estimates from different smoothing parameters, we computed the correlation between the hill scores assigned to valleys detected with different parameters. For this, we ran EpiSAFARI to identify the valleys in H3K4me3 data using the knot numbers, κ, between 4 and 15, and spline degrees, ψ, between 4 and 7. Given two sets of valleys computed by different knot numbers and spline degrees, we identified the valleys that share minima between these valley sets. Next, we computed the correlation between the left hill scores and the correlation between the right hill scores. This correlation computation is performed for all pairwise comparisons of parameters. The distribution of the left and right hill score correlations (Supplementary Fig. 5c, d) show that there is a substantial agreement between the assigned scores such that the correlations are mostly clustered above 0.40 with the most frequent correlations around 0.80.

It should be noted that the maximum allowed error in smoothing was set to a very large value while signal is smoothed in the above computations. This was performed to

compare the impact of the parameters on the hill score without any parameter updates. When we decrease the maximum error in smoothing to the default values, we observed that the correlations between the assigned hill scores increases much. For example, the correlation of left and right hill scores for the most distant parameter sets ($\psi = 4$, $\kappa = 4$) and ($\psi = 7$, $\kappa = 15$) is 0.59 and 0.66, respectively. Whereas, without the parameter updates, the left and right hill score correlations between valleys detected from these parameter sets is 0.49 and 0.44, respectively. This indicates that the hill scores of valleys detected with the parameter updates will exhibit higher consistency.

Selection of hill score threshold with respect to sensitivity and valley redundancy. One of the important parameters is the hill score threshold that is used to filter out topologically low-quality valleys (Supplementary Fig. 5b). In principle, the higher hill scores correspond to valleys that have very good topologies such that hills are monotonically increasing as we move from the valley's dip to the valley's summits. Thus, setting the hill score threshold high enables selecting good valleys. The distribution of left and right hill scores (Supplementary Fig. 5e, f, g) show that there are substantial number of valleys with hill scores very close to 1.

We next evaluated how the sensitivity of valleys changes with changing hill score threshold. We detected valleys using hill score parameters between 0.1 and 0.99. It can be seen that the sensitivity decreases as we increase the hill score. While the sensitivity of the valleys is decreasing with increasing hill score, another competing factor is the valley redundancy (Supplementary Figure 5k, l, m). The valley redundancy refers to how many valleys overlap with each other. The valley redundancy will increase with

decreasing hill score because valleys may start engulfing other valleys when the hill score threshold is decreased. We computed the valley redundancy as:

$$Valley\ Redundancy = 1 - \frac{\left|\boldsymbol{V}^{(\eta)}_{merged}\right|}{|\boldsymbol{V}^{(\eta)}|} \quad (19)$$

where η indicates the hill score, $\boldsymbol{V}^{(\eta)}$ denotes the valleys detected using η and $\boldsymbol{V}^{(\eta)}_{merged}$ denotes the set of valleys generated by merging the valleys in $\boldsymbol{V}^{(\eta)}$ where any two valleys with at least 1 base pair overlap are merged into one valley. As expected, the valley redundancy decreases with increasing η because valleys have distinctly uniform shapes. For $\eta = 0.1$, the redundancy is around 40% and decreases to around 3% for $\eta = 0.99$. This result indicates that hill score cutoff of 0.99 enables identification of distinct valleys at a cost of sensitivity. We have decided this is a fair tradeoff to generate high quality valleys and used $\eta = 0.90$.

Impact of Read Depth on Valley Detection

An important question about detection of valleys is to evaluate how many reads are necessary to for robust detection of valleys. To evaluate the impact of sequencing depth, we downloaded a high depth ChIP-Seq sequencing data from another study where 100 million reads are sequenced from H3K4me3 ChIP sample of GM12878 cells(13). For this data, we subsampled reads starting with 5 million reads up to 90 million reads with increments of 5 million reads. We next ran EpiSAFARI using each the reads generated by each subsampling. We observed the number of valleys increases with increasing read depth. Supplementary Figure 4i shows the additional number of valleys detected by each read sampling. While increasing read depth increases the

number of detected valleys, the increase in number of valleys is steady (At around 1000 valleys per 5 million reads) beyond 20 million reads.

We next quantified the increase in the fraction of functional elements (active promoters, transcription factor peaks, and DNase peaks) that overlap with the identified elements (Supplementary Figure 4j) with increasing read depth. As expected, the overlap with functional elements increases with higher read depth. However, the increase in the fraction of identified functional elements is stabilized around 35-40 million reads.

Combining these two observations above, we believe that at least 35-40 million reads are necessary to identify meaningful set of valleys. In comparison with literature, this result is higher than the results of a previous study on the impact of sequencing depth on ChIP-Seq analysis (14). In this study, the authors proposed that around 20-25 million reads are sufficient for detection of the peaks for H3K4me3 marks. We believe this difference is expected since valley identification requires more reads for detecting the detailed patterns associated with valleys. In addition, it should be noted that this estimate will be impacted by the technical factors such as the signal-to-noise ratio in the sample preparation and IP efficiency. In addition, the biological properties of the tested samples (the species, tissue cultures-vs-immortalized cell lines, normal-vs-tumor samples) will also have an impact on the required read depth. As such, these “saturation” analysis are fairly hard to conduct for technical and biological reasons, as the authors of above referenced study also conclude (14). Therefore our estimate should be taken with these factors in mind.

Impact of Smoothing Parameters on Valley Asymmetry

Similar to the hill scores, the smoothing parameters may impact the valley asymmetry, i.e., the imbalance between the left and right summits of the valleys. We first performed correlation of the valley asymmetry between every pairwise set of valleys within the sets of valleys detected using the knot numbers between 3 and 15, and spline degrees

between 3 and 7 (Supplementary Fig. 6a). Most of the correlations are clustered around 0.9, which indicates a high consistency between the asymmetry of valleys detected by set of parameters. We also evaluated the fraction of valleys that “changed direction” when pairwise sets of valleys are compared. To detect valleys that changed direction, we compared pairs of valleys detected using different parameters, then we counted the number of valleys which have turned from a left-to-right valley to a right-to-left valley. By left-to-right (right-to-left) valley, we refer to the valleys whose left (right) summit has higher signal than the right (left) summit. The distribution of the fraction of valleys that changed direction (Supplementary Fig. 6b) shows that directionality changing valley fraction is mostly clustered around less than 5%. These results indicate that the valley asymmetry is affected only slightly by the changing smoothing parameters.

Impact of Search Space and Filtering Parameters on Valley Detection

l_{min} and l_{max} parameters describe the minimum distance and maximum distance between valley dip and valley summits. This enables decreasing search space by evaluating only the summits within certain vicinity of dips to identify valleys. f_{min} sets the minimum ratio between signal at the maxima locations and the signal at the dip. This way the candidate valleys that do not show an expected level of signal depletion at the dip compared to the summits.

In the original manuscript we did not provide a thorough examination of these parameters because we have selected relaxed parameters to enable a sensitive valley detection. The users can choose to change these parameters in case they believe there is a different type of enrichment in the data. For example, the users can evaluate the signal profiles in IGV and get an estimate of the valley sizes that they would like to focus on.

In the revision, we evaluated changing the impact of l_{min}, l_{max}, and f_{min} parameters using H3K4me3 ChIP-Seq data from K562 cell line.

Impact of l_{min}***:*** Supplementary Figure 4d shows the impact of changing l_{min} on the fraction of active promoters of top valleys. As expected, for low l_{min} values (upto 200 base pairs), the accuracy stays constant for top valleys. As l_{min} gets close to 500 base pairs, fraction of detected active promoters decreases substantially close to 0. This provides evidence that the valleys detected by changing l_{min} are fairly robust in terms of detected active promoters. Thus, by default, we suggest using l_{min}=0 for histone modification valleys. We also evaluated the impact of changing l_{min} parameter on DNA methylation valleys. For this, we changed l_{min} parameter and computed the methyl-valleys using the DNA methylation data for H1HESC cell line. We next computed the fraction of the methyl-valleys that overlap with transcription factor peaks. Supplementary Figure 4k shows the changing overlap fraction for l_{min} starting from 0 to 1500 base pairs. It can be seen that lower l_{min} parameter enables highest overlap. When l_{min} is higher than 1000 base pairs, we see a sudden decrease in the overlap. By default, we use l_{min}=0 to enable a sensitive detection of the methyl-valleys.

Impact of $\boldsymbol{l_{max}}$***:*** Impact of changing l_{max} parameter is shown in Supplementary Figure 4e, where we plotted the impact of changing l_{max} in terms of the fraction of active promoters of top valleys. In principle, increase l_{max} increases the search space and should make detected valleys more accurate. For low l_{max} values (upto 1000 base pairs), the accuracy is low and as l_{max} is increased, the detected valleys overlap better with the active promoters. For l_{max}>1000 base pairs, accuracy reaches to a stable value. Thus, we conclude that for l_{max}>1000, valley detection is fairly robust in terms of accuracy measure we use.

We next studied how the changing l_{max} parameter changes the accuracy of DNA methylation valleys. As before, we changed l_{max} parameter and computed the methyl-valleys using the DNA methylation data for H1HESC cell line. We computed the fraction of the methyl-valleys that overlap with transcription factor peaks. Supplementary Figure 4l shows the changing overlap fraction for l_{max} starting from 0 to 5000 base pairs. It can be seen that high l_{max} parameter enables highest overlap. For l_{max} values smaller than 1000 base pairs, we see a sudden decrease in the overlap. By default, we use l_{max}=2000 to enable a sensitive detection of the methyl-valleys.

Impact of f_{min}: Supplementary Figure 4f shows how f_{min} impacts the fraction of active promoters in top valleys. While for most f_{min} selections, the accuracy is fairly stable. Using high f_{min} values (f_{min}>3) makes detected valleys less sensitive for detecting the active promoters. Interestingly, for top 3000, 4000, and 5000 valleys, there is a "sweet spot" at around f_{min}=2.5. Nevertheless, the accuracy for these cases is fairly stable for f_{min}<2.5. These results provide justification for usage of the default parameter that we proposed (f_{min}=1.2) in the comparisons with GM12878 cell line.

Impact of l_{post}: The post filtering is performed to smooth the signal and to alleviate the discontinuities. We evaluated the accuracy with respect to changing l_{post} (Supplementary Figure 4g). The increasing l_{post} smooths the signal after spline smoothing. Increasing l_{post} beyond 100 base pairs slowly decreases the active promoter overlap fraction. We selected l_{post}=50 to ensure a relaxed and sensitive valley detection.

Impact of p-value estimation window length (l_p): The p-value estimation window, which is l_p base pair long, is used for estimating the signal around the maxima and minima while p-value is

being assigned (Supplementary Figure 7a, 7b). EpiSAFARI uses the l_p base pair long vicinity of the summits and the dip and computes the average signal. Then uses these values to assign the p-value. We use alternating values of the p-value window length and evaluate active promoter detection accuracy in K562 cell line data. Supplementary Figure 4h shows the changing fraction of valleys overlapping with active promoters with changing l_p. Increasing l_p decreases the active promoter fraction of valleys. By default, we use l_p=50 base pairs.

Impact of Sequence Content (Minimum CG Content and Minimum CpG Content) on DNAm valleys: A parameter that we used to filter DNA methylation valleys is the minimum CG Content in the methyl-valleys. Supplementary Figure 4m shows the fraction of methyl-valleys as maximum CG nucleotide fraction is changed. For low CG nucleotide fraction, the overlap fraction is steady above 95%. As the minimum CG content threshold is increased above 40%, the overlp fraction starts dropping. We also tested the impact of minimum number of CpG's in the methyl-valleys. This is particularly important for methyl-valleys because the DNA methylation levels are quantified mainly at the positions that have CpG nucleotides at the genome sequence. Supplementary Figure 4n shows the impact of changing minimum CpG dinucleotide count within the valleys. Below 20 CpG dinucleotides and above 60 CpG dinucleotides, the accuracy decreases. As the default parameter, we the minimum CpG dinucleotide count as 20, which enables sensitive detection of methyl-valleys with high overlap with the transcription factor peaks.

Detection of Differential Valleys and Differential Valley Analysis

Before we describe how we identify differential valleys, we would like to first briefly discuss how we describe a differential valley. We describe a differential valley as a region where one sample shows higher signal at one of the summits and/or shows lower signal at the dip. Thus, a differential valley would have a more (or less) pronounced valley shape when two samples are compared.

The details are now included in the Methods Section of the main manuscript and in the Supplementary Information. In addition, Supplementary Figure 13a illustrates the differential valley computation. We describe the differential valley analysis below:

1. ***Pool valleys:*** Differential valley calling starts after the valleys are called for the two samples. We refer to these samples as Sample1 and Sample2. EpiSAFARI pools the the identified valleys from Sample1 and Sample2 without merging.
2. ***Normalization of Profiles:*** Since the read depth of the samples may be different, it is necessary to normalize the signal profiles. To do this, EpiSAFARI uses RPM normalization where the total signal in Sample1 and in Sample2 are computed. Next, the scaling factor is computed by dividing the larger of the total signal values. Finally, the signal profile of the sample with lower total signal is multiplied by this scaling factor. This way, both signal profiles contain the same total signal.
3. ***Computation of Difference Profile:*** Next, for each valley in the pooled list, EpiSAFARI computes the difference signal profile by subtracting the normalized signal profile of Sample2 from the normalized signal profile of Sample1. At any position where the difference is negative, we assign 0 value to the location. Thus, the difference profile reflects the signal within the valleys that are specific to Sample1.
4. ***Significance Assignment to Pooled Valleys:*** Using the difference profile, EpiSAFARI computes the multinomial p-value of all the pooled valleys. The logic of using the difference profile to compute the p-values is that if there is a differential pattern at the tested valley, the difference profile should also look like a valley. Thus, when p-value should provide evidence for a significant differential valley. Note that the multinomial p-value computation was presented in the original manuscript and in the Supplementary Information. The differential p-value for each valley is also computed by using the difference profile computed by subtracting Sample2 profile from the Sample1 profile. After

this, EpiSAFARI assigns two p-values to each of the pooled valleys where first p-value represents the significance of a differential valley in Sample1 compared to Sample2 (Sample1_vs_Sample2 p-value) and second p-value represents the significance of the differential valley in Sample2 compared to Sample1 (Sample2_vs_Sample1 p-value).

5. ***Filtering of Differential Events:*** Since each valley is assigned two p-values, it is necessary to filter out the valleys to identify the final differential values. We set a significance cutoff, by default log(-10), and filter out the valleys for which the Sample1_vs_Sample2 p-value is higher than the threshold. Finally, we ensure that there is no evidence of a significant differential valley behavior in Sample2 by making sure that the Sample2_vs_Sample1 p-value is higher than a relaxed threshold (log(-2)) and that the difference profile shows a valley pattern at the location by making sure the summits in difference profile have higher signal than the dip location.

We applied differential valley analysis by comparing the H3K4me3 valleys detected in GM12878 (Sample1) and K562 (Sample2) cell lines. To evaluate whether the identified valleys are meaningful, we hypothesized that the differential valleys must show differential DNase signal. To test this, we computed the average DNase signal on each of the pooled valleys. For each valley, we divided the DNase signal by the total number of million mapped nucleotides and then by the total length of the valley in kilobases (similar to RPKM normalization). After the DNase signal is computed for all valleys using GM12878 and K562 DNase data, we performed quantile normalization of the signal. This is necessary to remove sample specific global and technical effects and also normalize the distributions of the DNase signal on valleys. We finally computed the difference (in terms of fold change) in the normalized DNase signal on the K562 specific valleys and GM12878 specific valleys, and all the valleys as comparison. Supplementary Figure 13b shows the distribution of the logarithm of the DNase signal ratio between GM12878 and K562 cell line, i.e. log(GM12878 DNase Signal / K562 DNase Signal), on the cell line specific valleys

for both cell lines. From the figure, the log fold change (FC) is almost symmetrically distributed around 0 for all the valleys. On the other hand, FC distribution is significantly positively skewed for GM12878 specific valleys and significantly negatively skewed for K562 specific valleys when they are compared with the FC distribution of all valleys (Wilcoxon test p-value < 2.2x10e-16 for both comparisons). In other words, the cell line specific valleys show concordant differential enrichment of DNase signal in the respective cell line. This analysis presents supporting evidence that the identified differential valleys are enriched in differential DNase signal. We performed the differential valley analysis by comparing the valleys in H1HESC cell line and K562 cell line. Supplementary Figure 13c shows the DNase FC distribution for the cell line specific and all valleys. For the valleys specific to each cell line, we observe higher DNase FC for the corresponding cell line.

In Supplementary Figure 13d, we include an example of a region on chromosome 14 where two differential valleys, one in GM12878 and other in K562, are identified close to each other. Interestingly, this region contains an H3K4me3 peak that manifests on both cell lines. The figure shows the H3K4me3 and DNase signals. The visual examination of the signal profiles alone shows that while there is high signal in both cell line, the valley structure shows considerable changes among cell lines. In addition, the differential valleys (highlighted on the figure) show clear increase in DNase signal for the corresponding cell line. We think that this is an example of how that valley-based analysis can provide novel insight while analyzing functional genomics data.

It must be noted that the valley comparisons can be performed in different ways and we present one way to compare the valleys. For example, another differential valley pattern is that while the valley's summit/dip signal ratio does not change, the signal at the summits may change direction, i.e., the valleys directionality may change. These comparisons can be easily performed using command line tools such as awk to filter out the valleys with respect to their directionality.

Assignment of Statistical Significance

The next step is assignment of statistical significance to the detected valleys (Fig. 1). By statistical significance, we refer to how significant the depletion of the signal at the dip is compared to the signal levels at the summits. Thus, valleys with low p-value correspond to deep valleys. The assigned p-values are used to sort the valleys while performing enrichment analysis.

For a valley at (i, j, k), EpiSAFARI first computes the signal around the vicinity of the dip and the summits using

$$S_i = \sum_{i-\frac{l_p}{2}<a<i+\frac{l_p}{2}} s_a \tag{9}$$

$$S_j = \sum_{j-\frac{l_p}{2}<a<j+\frac{l_p}{2}} s_a \tag{10}$$

$$S_k = \sum_{k-\frac{l_p}{2}<a<k+\frac{l_p}{2}} s_a \tag{11}$$

where S_i, S_j, S_k denote the average signal in the l_p base pair (100 base pairs by default) vicinity of the summits i, j, and the dip k. Next, EpiSAFARI computes the binomial p-value of enrichment of signal around summits compared to the dip:

$$bin(S_i, S_k) = \sum_{a=0}^{S_k} \binom{S_k + S_i}{a} \cdot \left(\frac{1}{2}\right)^{S_k+S_i} \tag{12}$$

$$bin(S_j, S_k) = \sum_{a=0}^{S_k} \binom{S_k + S_j}{a} \cdot \left(\frac{1}{2}\right)^{S_k+S_j} \tag{13}$$

where $\binom{S_k + S_i}{a}$ number of combinations for selecting a items within $S_k + S_i$ items:

$$\binom{S_k + S_i}{a} = \frac{(S_k + S_i)!}{(S_k + S_i - a)! \cdot a!} \tag{14}$$

In order to assign the final p-value to the valley, we combine the p-values that are assigned to enrichment of the signal at the two summits. This process corresponds to combining the null models that are used to assign the two p-values for the observed summit-to-dip signal enrichment. We first use intersection of the null models as the joint null model (Supplementary Fig. 7a). Assuming that the left and right hills are independent, this corresponds to the direct multiplication of the p-values:

$$\log\big(p - value_{\cap}(i,j,k)\big) = \log\Big(bin\big(S_j,\, S_k\big)\Big) + \log\Big(bin\big(S_j,\, S_k\big)\Big). \tag{15}$$

$p - value_{\cap}$ denotes the p-value computed by intersection-based combination of the p-values assigned to observed summit-to-dip signal enrichment. In addition, we use the union of the null models corresponding to null distribution of signal among summits and the dip so as to assign the p-value of the valley. As before, we assume that the p-values assigned to summits are independent from each other. Thus, the p-value estimated from the union of the null models is:

$$\begin{aligned}&\log\big(p - value_{\cup}(i,j,k)\big)\\ &\qquad = \log\Big(bin(S_i,\, S_k) + bin\big(S_j,\, S_k\big) - bin\big(S_j,\, S_k\big) \times bin\big(S_j,\, S_k\big)\Big).\end{aligned} \tag{16}$$

$p - value_{\cup}$ denotes the combined p-value (Supplementary Figure 7a). In (15) and (16), we assumed that the p-values assigned to observed enrichment of the signal at the left and right summits are independent from each other. This assumption may not hold as we see a significant correlation of signals on the left and right summits (Supplementary

Methods, Supplementary Figure 8a). As an alternative significance estimation method, we computed a multinomial distribution-based p-values without the need for combining p-values. The multinomial p-value is computed as:

$$\log\left(p - value_{multin}(i,j,k)\right) = \sum_{a=0}^{S_k} \sum_{b=0}^{a} \frac{(S_i + S_j + S_k)!}{(S_k - a)! \cdot (S_j + b)! \cdot (S_i + a - b)!} \cdot \left(\frac{1}{3}\right)^{S_i + S_k + S_j} \quad (17)$$

where the p-value is computed as the probability for different signal configurations at the summits and the dip such that the configurations are more extreme than what we observed. By more extreme, we mean the signal at one or both of the summits are higher than the observed signals (Supplementary Fig. 7b). We compute the p-value as the total probability of all the signal configurations that correspond to more extreme valleys than the observed valley. In general, the union-based binomial p-value merging is more conservative and exhibits lower sensitivity compared to the intersection-based p-value merging and multinomial based p-values (Supplementary Methods, Supplementary Figure 8b, c). We therefore use intersection-based binomial p-value merging in the benchmarking. After the p-values are assigned, the false discovery rate at which each valley would be deemed significant is estimated using Benjamini-Hochberg procedure(Benjamini, 2010).

The valleys that EpiSAFARI detected may overlap with each other although we generally observed that the overlap between detected valleys tends to be very small. To ensure that a non-redundant set of minima are reported, EpiSAFARI filters out the valleys whose dips are close to each other by selecting the most significant valley (i.e., lowest p-value) around local minima positions.

An important factor about detecting valleys is the required sequencing depth. For analyzing the required read depth, we used a high depth H3K4me3 ChIP-Sequencing data from NA12878 sample (Kasowski *et al.*, 2013) and identified valleys. We next computed the increase in the number of valleys with increasing read depth and the increase in the fraction of identified functional elements (Supplementary Information, Supplementary Figure 4i and 4j). We found that beyond 35-40 million reads, the valley detection does not provide substantial additional information.

Valley Annotation

EpiSAFARI can annotate valleys with respect to genes and transcription factor binding peaks. This step compares the valleys with an annotation file in GFF format and assigns the valley to the promoters, transcripts, and exons. We also created a GFF file from the transcription factor binding peak regions from ENCODE project(Dunham *et al.*, 2012). This GFF file contains the peaks of the transcription factors that are identified by 690 ChIP-Seq experiments performed on cell lines and uniformly processed by the ENCODE Project. EpiSAFARI can use these to annotate the valleys with respect to transcription factor binding. EpiSAFARI generates an extended BED file which contains the valley positions, signal levels, multi-mappability signal, significance, and annotations for each the valley. The smoothed signal profiles can be used for visualizing the signal (Supplementary Fig. 2).

Data Availability and Accession Numbers

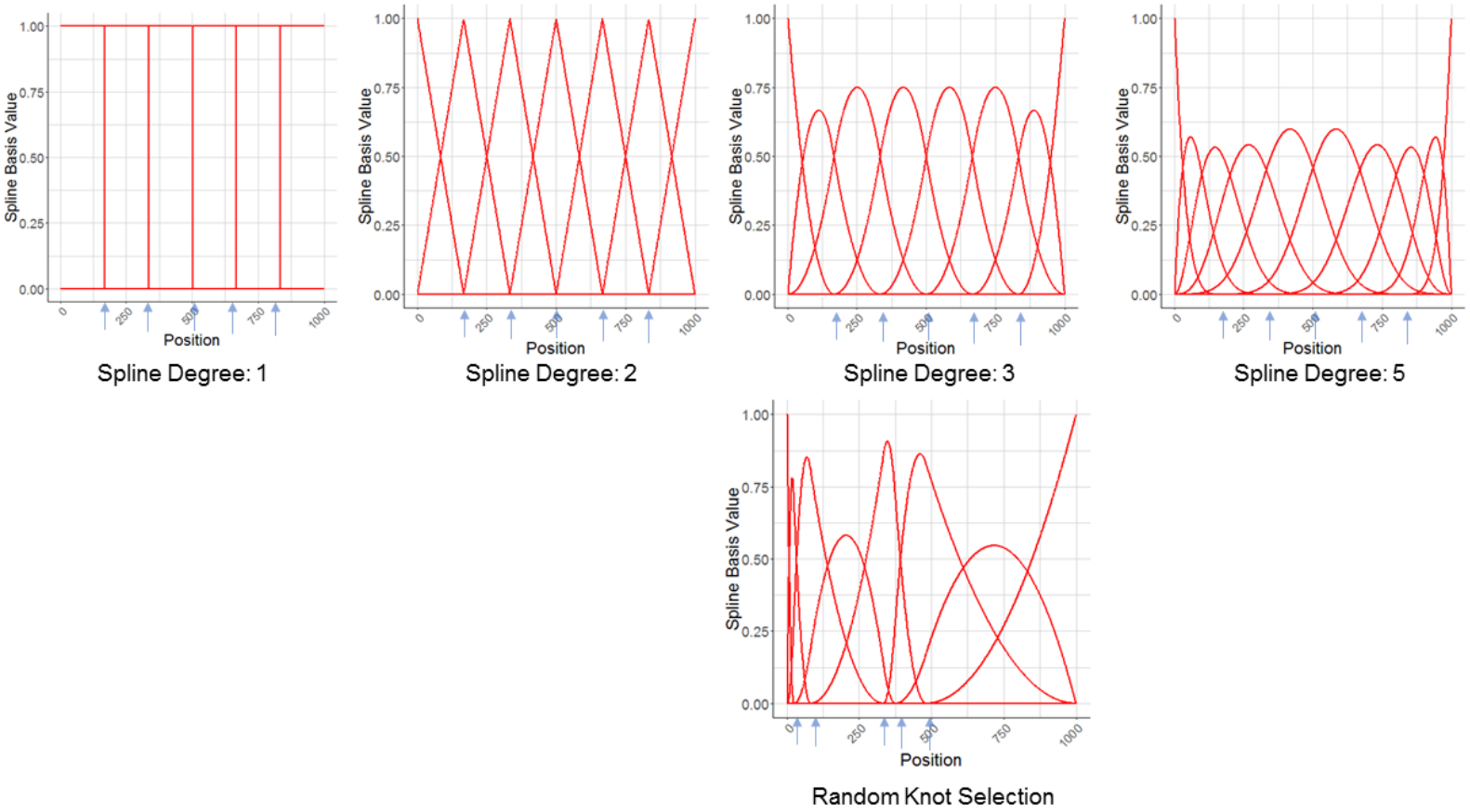
The H3K4me3 histone modification ChIP-Seq and DNase data, and transcript expression quantifications for K562, GM12878, and H1hESC cell lines are downloaded from ENCODE project website (http: //hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC). The

transcription factor binding peaks for K562, GM12878 and H1hESC cell lines are downloaded from the uniformly processed datasets of the ENCODE Project. The conservation scores are downloaded from the 100-way PhyloP track of UCSC Genome Browser. Whole genome bisulfite sequencing-based DNA methylation data for H1hESC data is downloaded from the Roadmap Epigenome Project Data Browser. GRO-Seq data is downloaded from GEO website with accession number GSM1480326. The GO enrichment analysis is performed using DAVID website(15). The random valleys in aggregation analyses and plots are generated by randomly shifting each valley within 1 megabase vicinity of itself. The H3K4me3 peaks are identified using MUSIC(16) algorithm. The whole genome bisulfite sequencing-based DNA methylation data for GM12878 cell line is downloaded from GEO web site with accession number GSM2772524.

The multi-mappability profile is obtained as described in a previous publication(16). In summary, the genomes are fragmented into fragments of the desired read length (denote by l_{read}) and these are mapped back to the reference genome by allowing multimapping reads. After the reads are mapped, we count the number of reads that are mapping to every genomic position. For any genomic position that is uniquely mappable, this computation yields exactly $2 \times l_{read}$ at that position. For any genomic position that is multi-mapped, the number of overlapping reads at the position will be higher than $2 \times l_{read}$, therefore we call this profile the multi-mappability profile. This profile quantifies the multi-mappability of each position in the genome for the given read length of l_{read}.

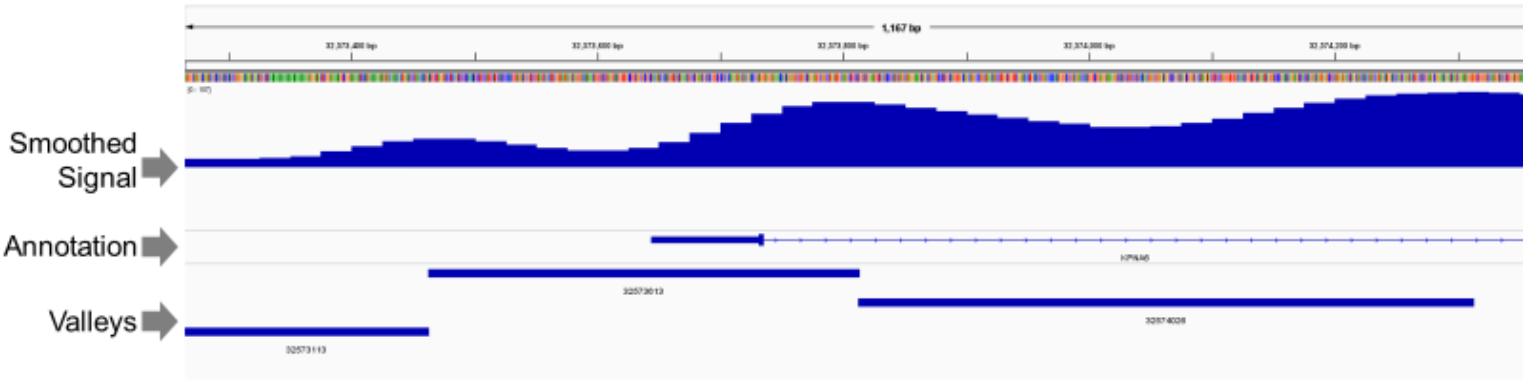
Supplementary Figures

Supplementary Figure 1

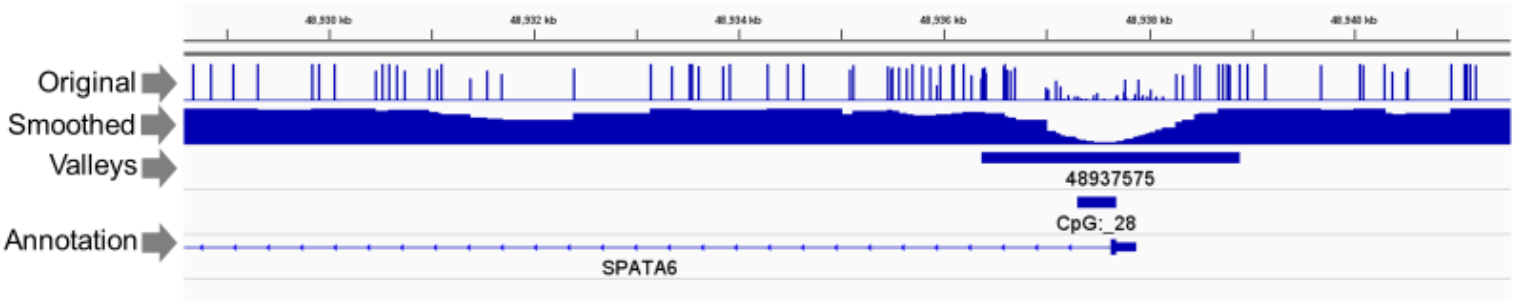


Supplementary Figure 1: Illustration of basis spline functions of different degrees with 7 knots. The x-axis represents the signal domain (between 0 and 1000) and y-axis represents the signal value. Each basis spline is shown with a red curve. The spline degree increases from 1 (leftmost) to 5 (rightmost). For degree 1, smoothing is basically replacing data with its mean between consecutive knots. For degree 2, the smoothing represents the piecewise linear smoothing of the data. Above degree of 2, the splines show more complex patterns that can represent different types of smooth transitions. The knot positions (excluding the first and last knots that are located at the beginning and at the end, respectively) are indicated with light blue arrows on the x-axis. The top row shows the basis splines with uniform knot selection. The bottom figure shows the basis splines generated by the random knot selection.

Supplementary Fig. 2a

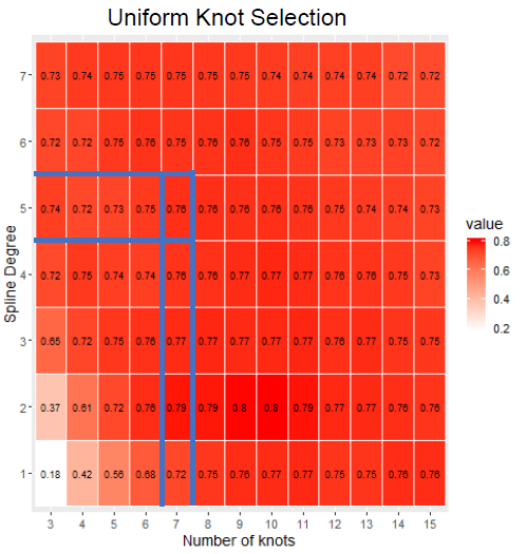


Supplementary Fig. 2b

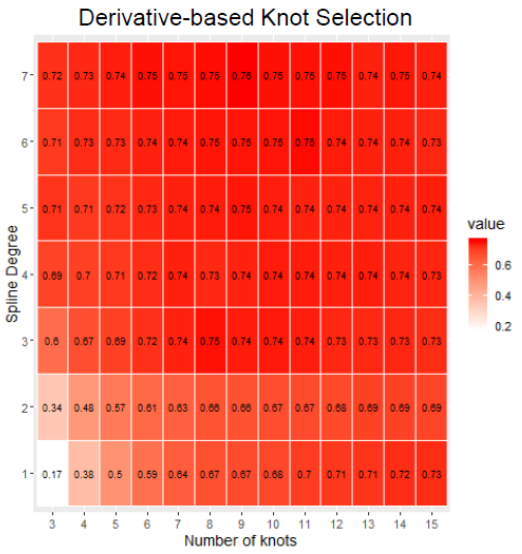


Supplementary Figure 2: a) Screenshot of valleys detected by EpiSAFARI as visualized using IGV. a) The top panel shows the smoothed signal profile. The valleys are shown in the bottom panel. The middle panel shows the gene annotations. The gene has a beginning where EpiSAFARI detected a valley. Other valleys are identified neighboring this valley. b) Example region containing a valley in DNA methylation signal profile. The top track, titled 'Original Signal' shows the original signal profile from the WGBS experiment. Note the sparseness of the signal as it is measured only at the CpG dinucleotides. The second track shows the spline smoothed signal profile generated by EpiSAFARI. The signal is turned into a continuous profile that is used to identify the valley that is shown in the following track. The gene and CpG island annotations are shown in the bottom two tracks.

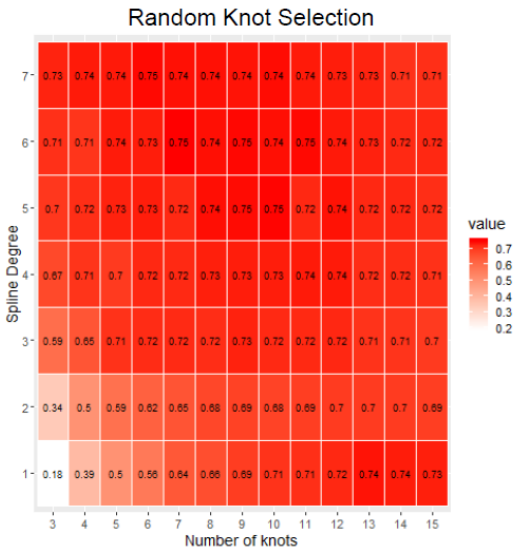
Supplementary Figure 3a



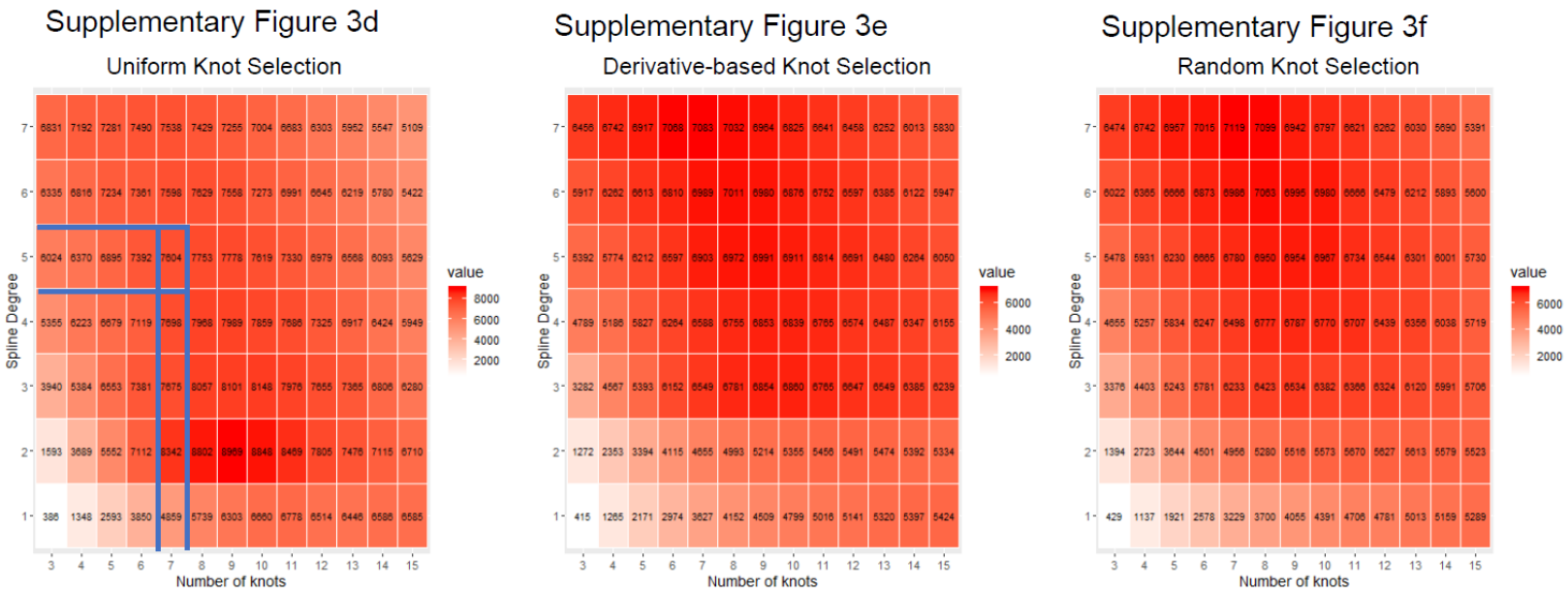
Supplementary Figure 3b



Supplementary Figure 3c

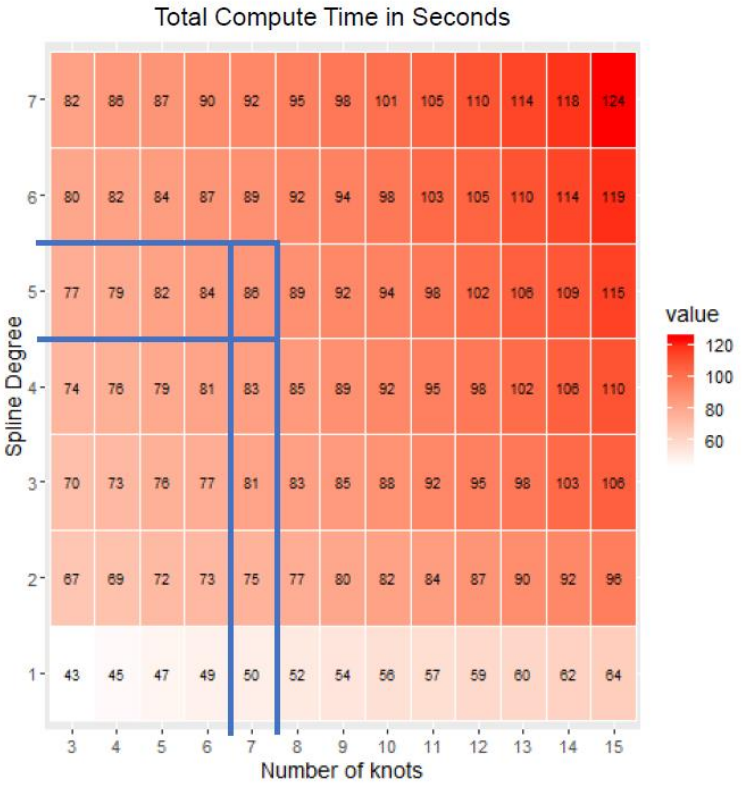


Supplementary Figure 3: The sensitivity of EpiSAFARI with changing number of knots (x-axis, between 3 and 15) and spline degrees (y-axis, between 1 and 7) for 3 different knot placement approaches: uniform (a), derivative-based (b), and random (c). The colors indicate sensitivity. The exact sensitivity value is included in each cell for clarification. The blue lines indicate the accuracy at the default spline degree (5) and knot number (7).

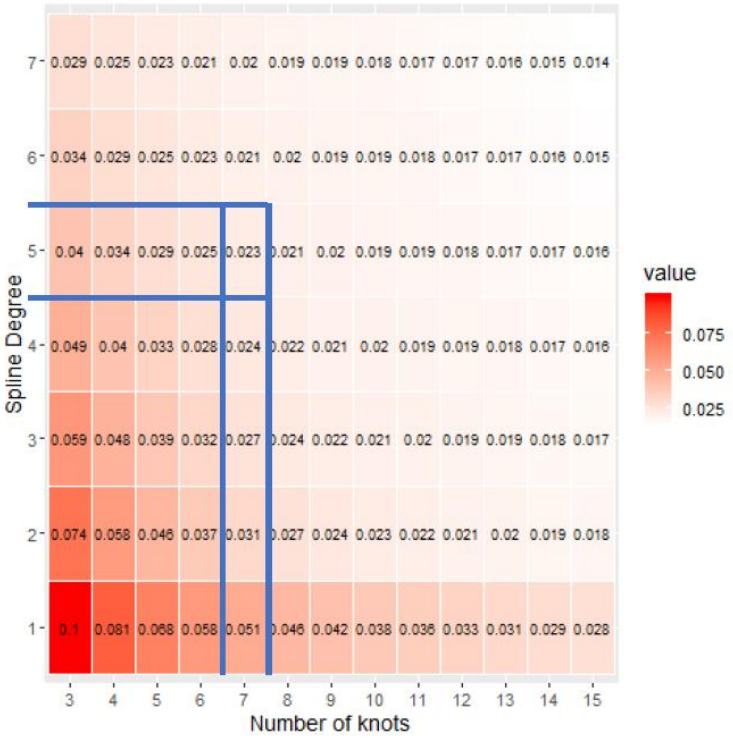


Supplementary Figure 3d, e, f: Number of valleys detected by EpiSAFARI for different number of knots (x-axis), spline degrees (y-axis), and knot placements. The exact number of detected valleys is included in each cell.

Supplementary Figure 3g



Supplementary Figure 3h



Supplementary Figure 3g, h: Illustration of compute time in seconds (g) and average smoothing error (h) versus knot number and spline degree.

Supplementary Figure 3i

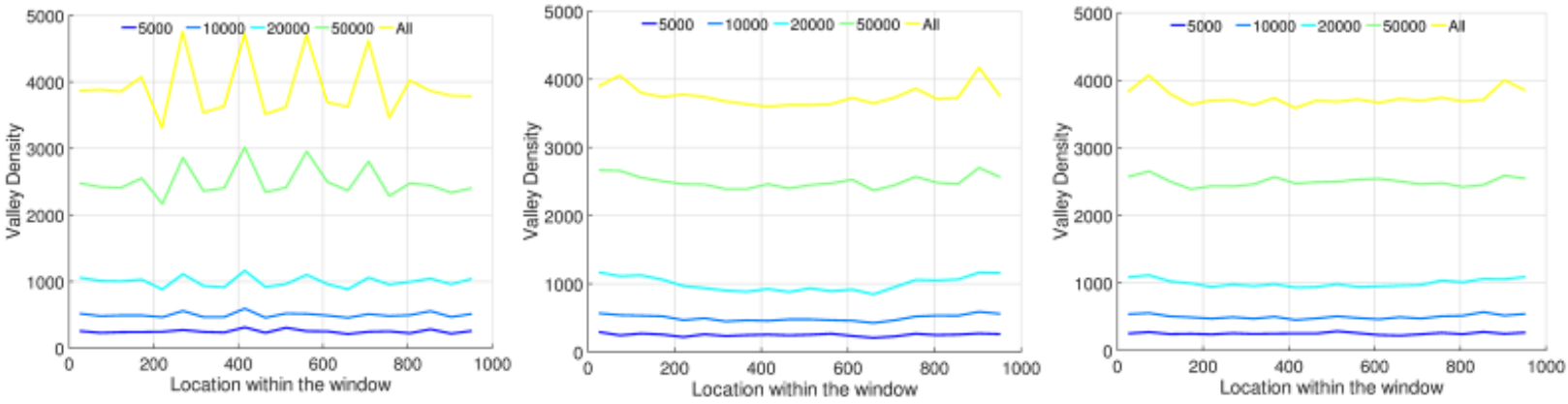
Uniform Breakpoints

Supplementary Figure 3j

Derivative Breakpoints

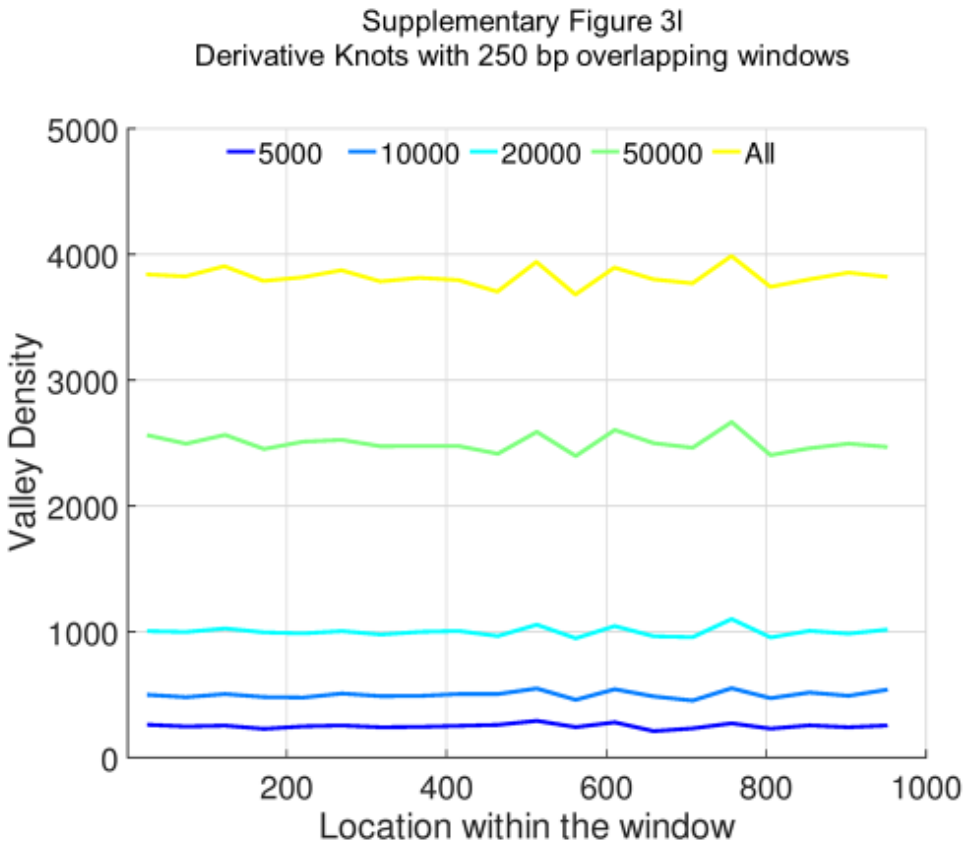
Supplementary Figure 3k

Random Breakpoints



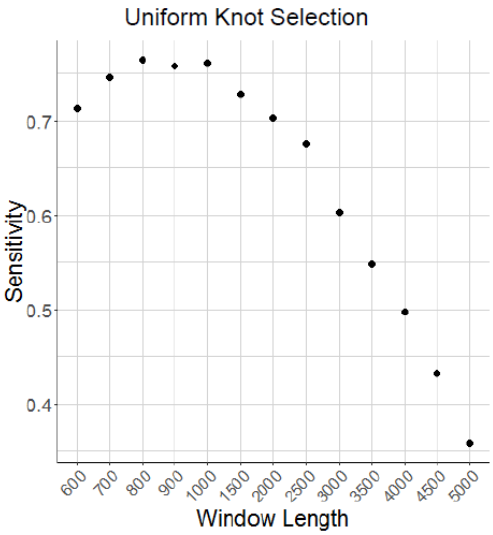
Distribution of the relative location of valley dip within smoothing windows

Supplementary Figure 3i, j, k: The distribution of relative position of valley dips within l_w=1000 base pair windows for K562 H3K4me3 ChIP-Seq data with different breakpoint selection strategies. X-axis shows the relative position with respect to the start position of the window. Y-axis shows the number of valley dips whose relative location is observed at the corresponding location on X-axis. The uniform breakpoint (left plot) selection shows periodic bias for the valley dip locations. For derivative and random breakpoint selections (middle and right plots, respectively), slight bias is observed at the ends of the windows. The top 5000 (blue), 10,000 (light blue), 20,000 (cyan), 50,000 (green), and all valleys (yellow) are plotted in each plot.

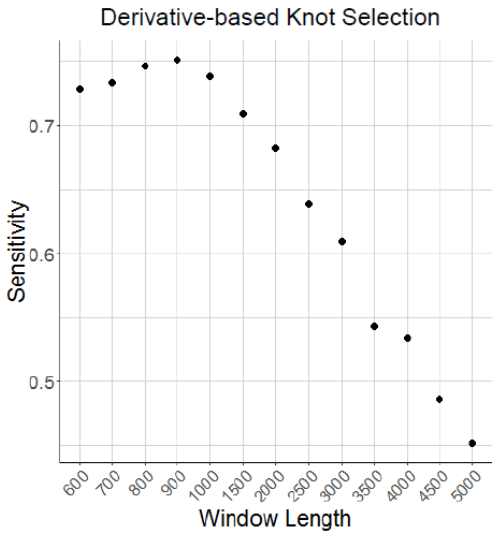


Supplementary Figure 3l: The distribution of relative position of valley dips within l_w=1000 base pair windows for K562 H3K4me3 ChIP-Seq data when smoothing is performed with sliding window with 250 base pair steps. X-axis shows the relative position with respect to the start position of the window. Y-axis shows the number of valley dips whose relative location is observed at the corresponding location on X-axis. The top 5000 (blue), 10,000 (light blue), 20,000 (cyan), 50,000 (green), and all valleys (yellow) are plotted in each plot.

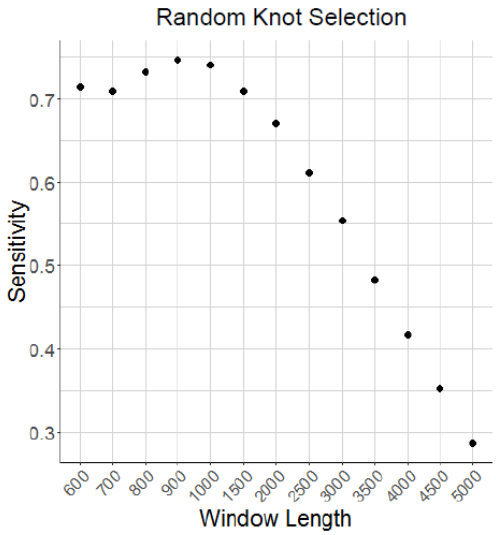
Supplementary Figure 4a



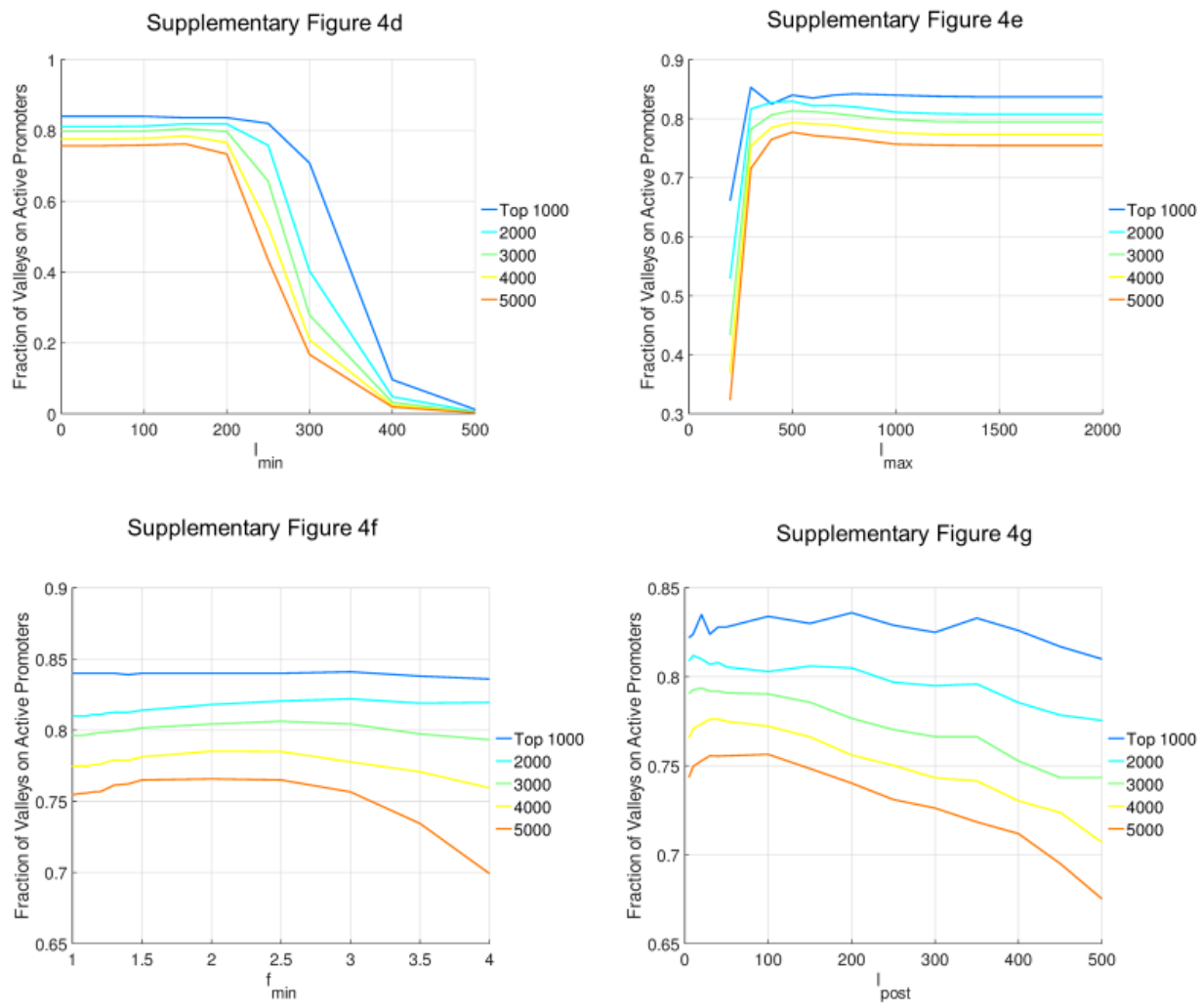
Supplementary Figure 4b



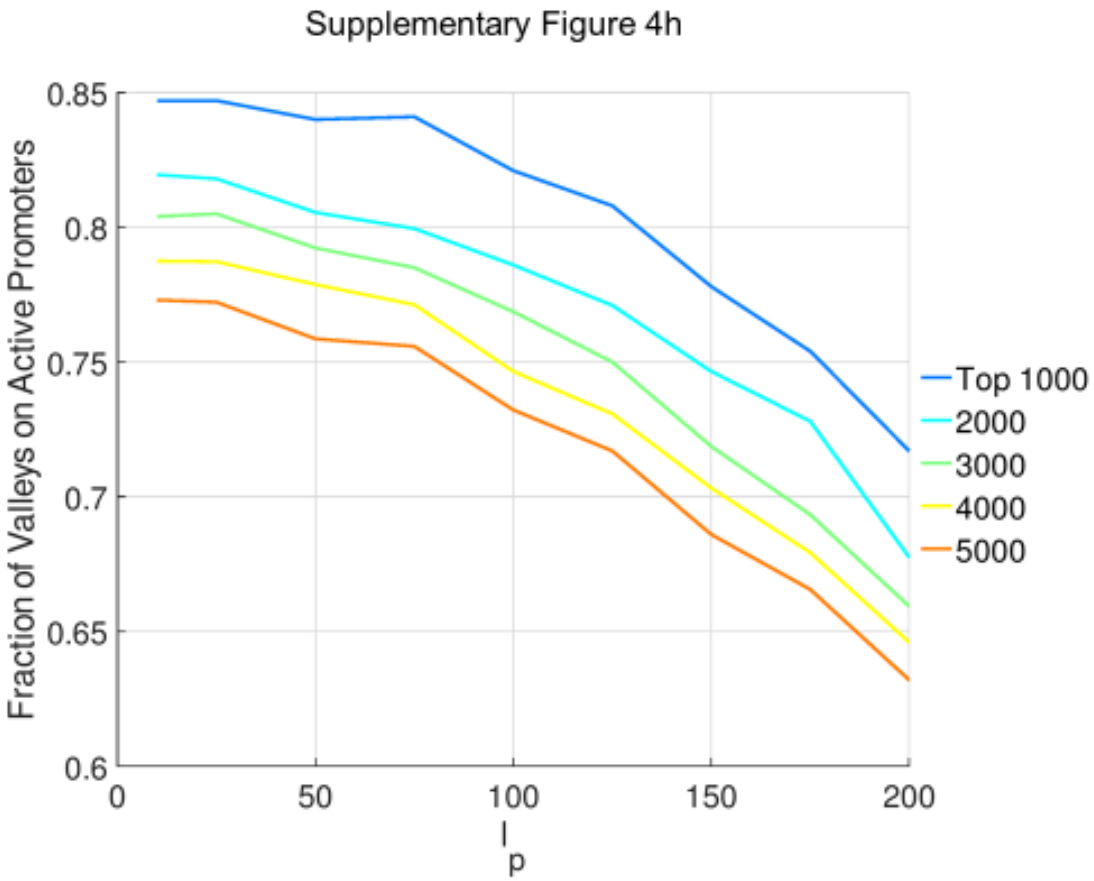
Supplementary Figure 4c



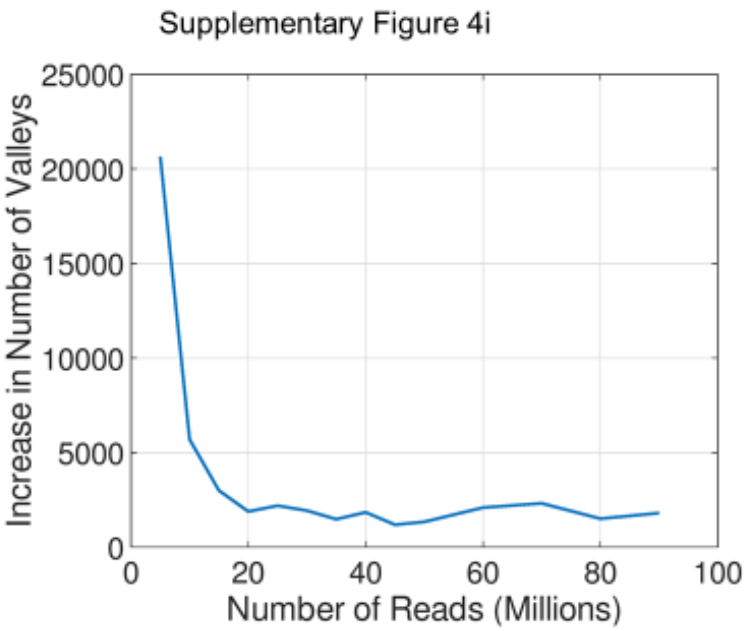
Supplementary Figure 4a, b, c: Sensitivity of EpiSAFARI with changing window length parameter, l_w, for three different knot placement approaches (a) Uniform, (b) Derivative-based, (c) Random. X-axis shows the window length and y-axis shows the sensitivity.

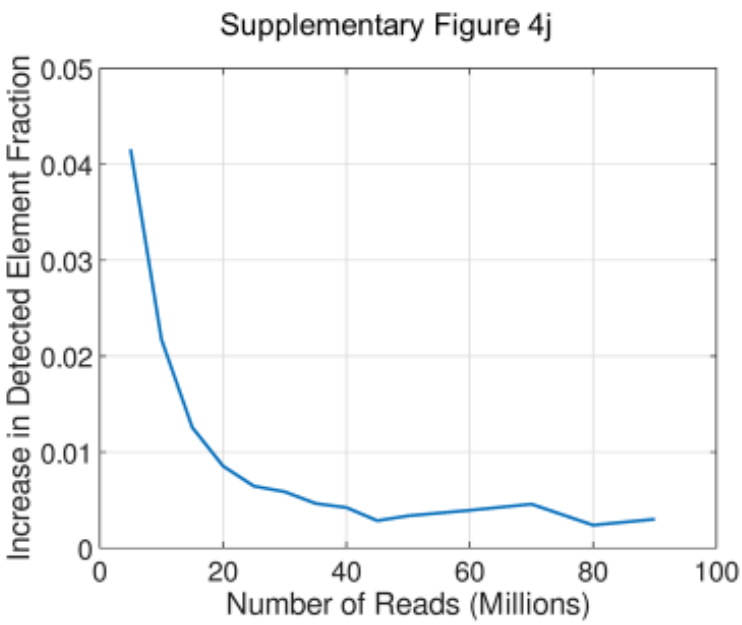


Supplementary Figure 4d, e, f, g: Active promoter sensitivity of EpiSAFARI with changing window length parameter, l_{min} (d), l_{max} (e), f_{min} (f), and l_{post} (g). Each plot contains the fraction of active promoters that overlap with 200 base pair vicinity of the top 1000 (blue), 2000 (cyan), 3000 (Green), 4000 (Yellow), and 5000 (Red) valleys. X-axis shows the changing parameters.

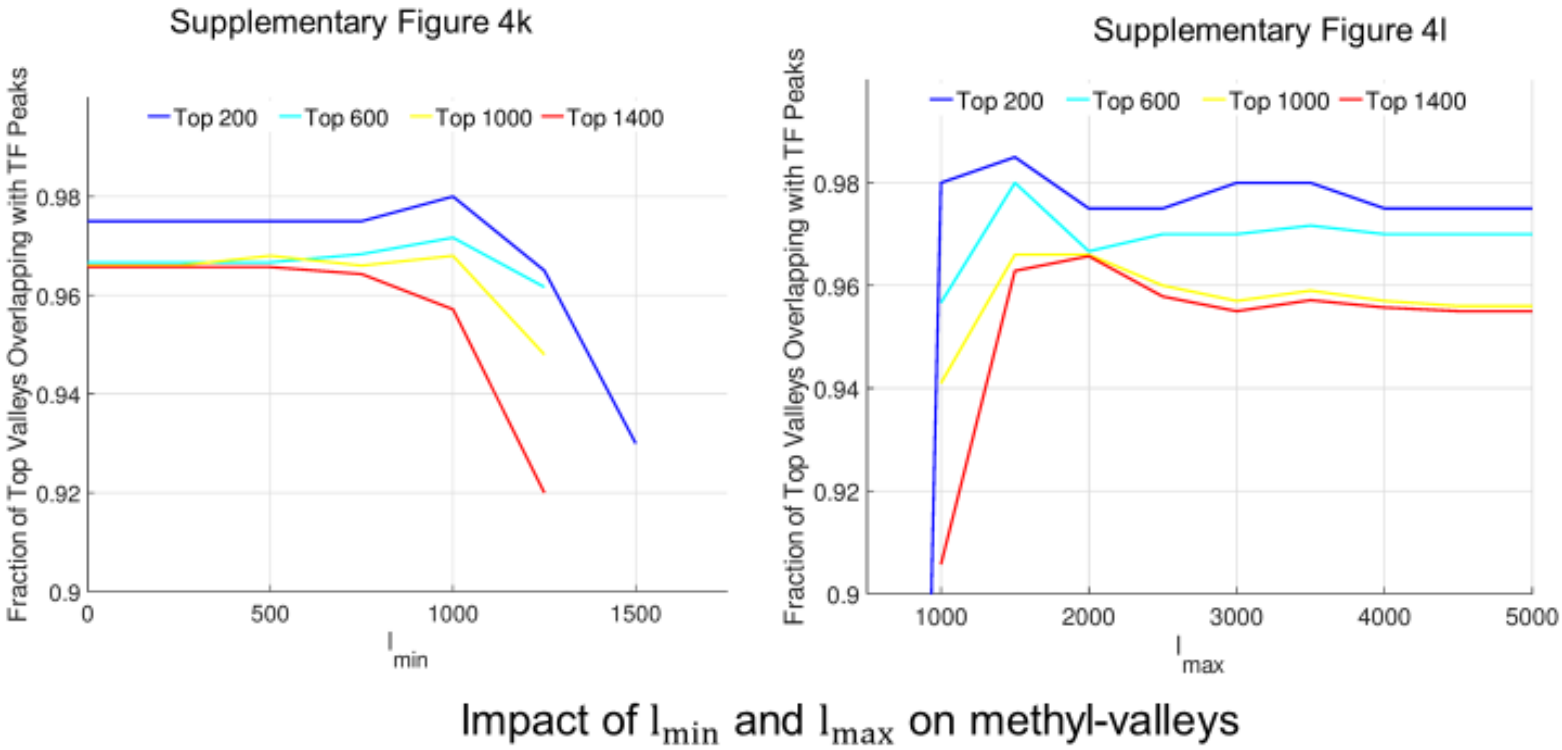


Supplementary Figure 4h: Active promoter sensitivity of EpiSAFARI with changing p-value signal estimation window length, l_p. Each plot contains the fraction of active promoters that overlap with 200 base pair vicinity of the top 1000 (blue), 2000 (cyan), 3000 (Green), 4000 (Yellow), and 5000 (Red) valleys. X-axis shows the changing p-value signal estimation window length.

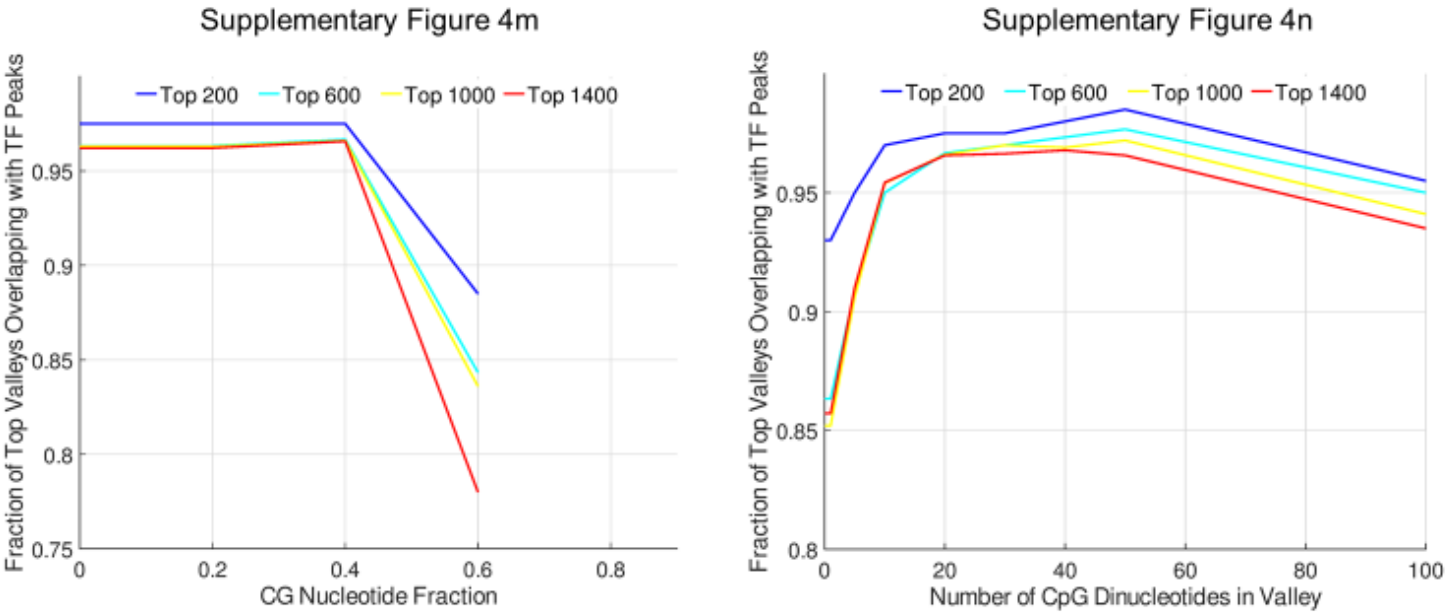




Supplementary Figure 4i, j: Increase in the detected number of valleys (i), and increase in the fraction of valleys that overlap with functional elements (j) as the number of reads are changed. X-axis shows the number of reads.



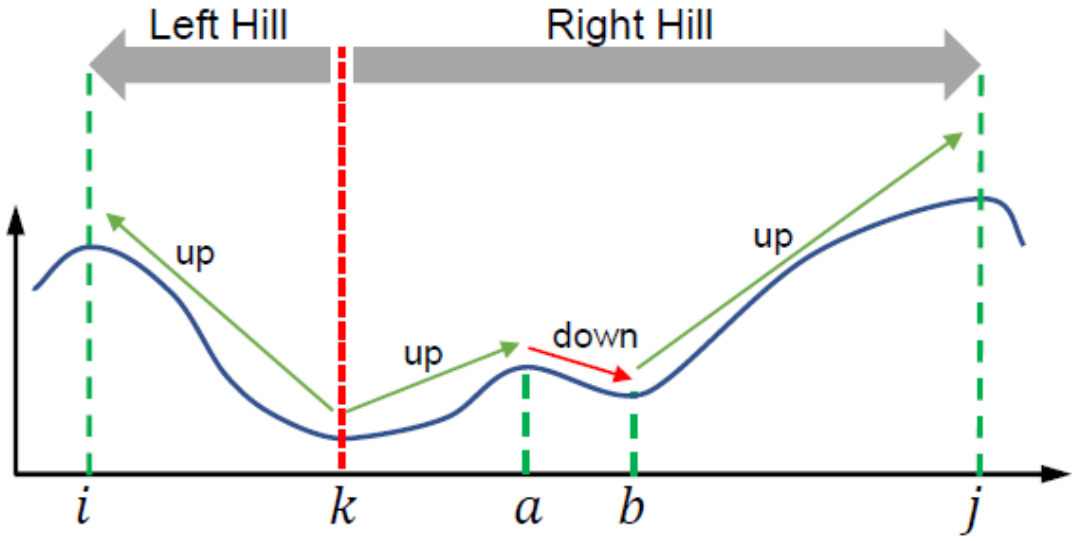
Supplementary Figure 4k, l: Impact of l_{min} (k), l_{max} (l) on the fraction of top methyl-valleys overlapping with transcription factor peaks.



Impact of Sequence Content on methyl-valleys

Supplementary Figure 4m, n: Impact of CG nucleotide fraction (m), number of CpG dinucleotides in valleys (l) on the fraction of top methyl-valleys overlapping with transcription factor peaks.

Supplementary Fig. 5a

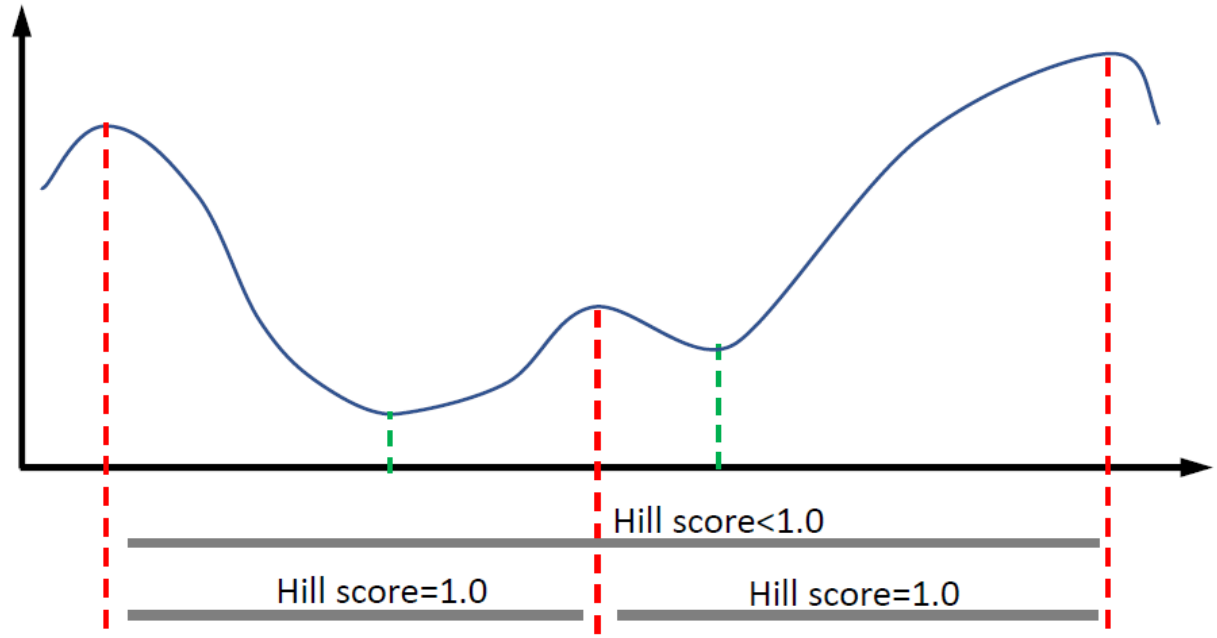


$$\text{Left Hill Score} = \frac{Total\ \#\ uphill\ positions}{left\ hill\ length} = \frac{k-i}{k-i} = 1.0$$

$$\text{Right Hill Score} = \frac{Total\ \#\ uphill\ positions}{right\ hill\ length} = \frac{(a-k)+(j-b)}{j-k} < 1.0$$

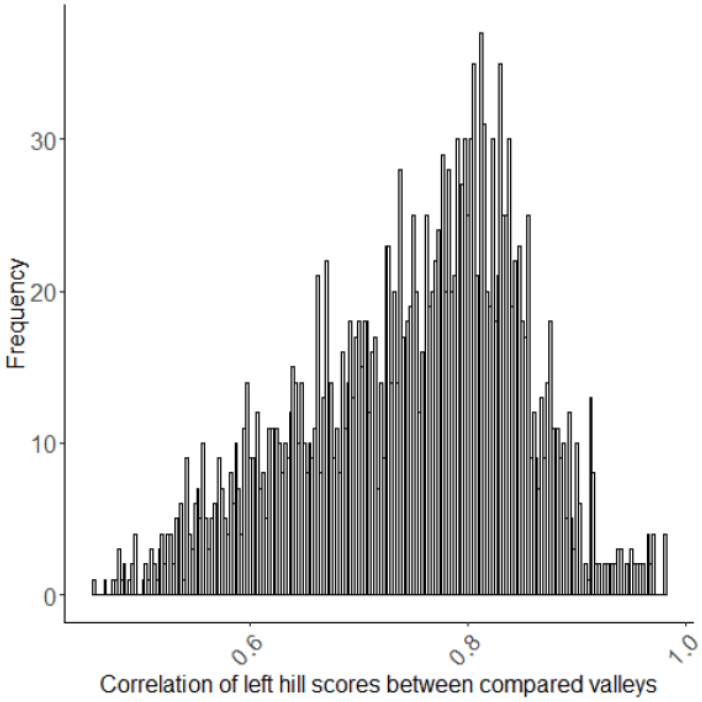
Supplementary Figure 5a: Illustration of hill score. A valley is defined as a dip (k) surrounded by two summits (i and j). The left (right) hill is the region between the left (right) summit and the dip. The left (right) hill score is counted as the number of genomic coordinates that go up while traveling from the dip to the left (right) summit. The coordinates that are "up-hill" and "down-hill" are indicated with green and red arrows, respectively. In this example, left hill score is 1 because there are no coordinates that are down-hill. Right hill score is smaller than 1 because the region between coordinates a and b are down-hill.

Supplementary Figure 5b

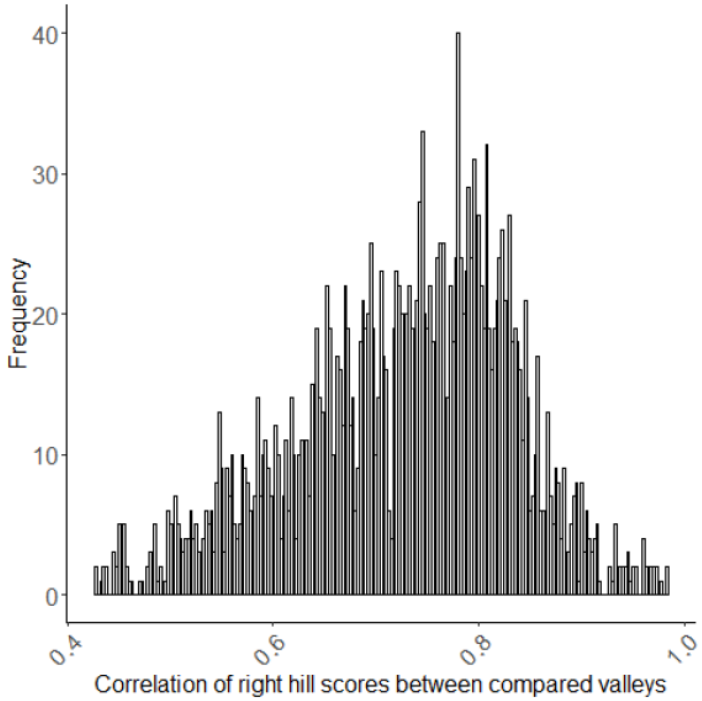


Supplementary Figure 5b: Illustration of valley merging with decreasing hill score threshold. The left and right valleys have good shapes and are assigned hill scores of 1.0. When we combine them and compute the hill score for the merged valley, the hill score of this valley is subpar, i.e., smaller than 1.0 because of the summit in the middle (See Supplementary Figure 4a). Thus, decreasing the hill score cutoff may create merged valleys that have lower topological quality.

Supplementary Figure 5c

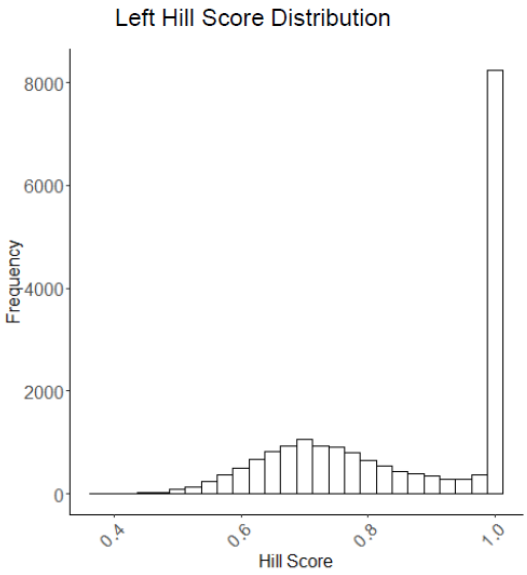


Supplementary Figure 5d

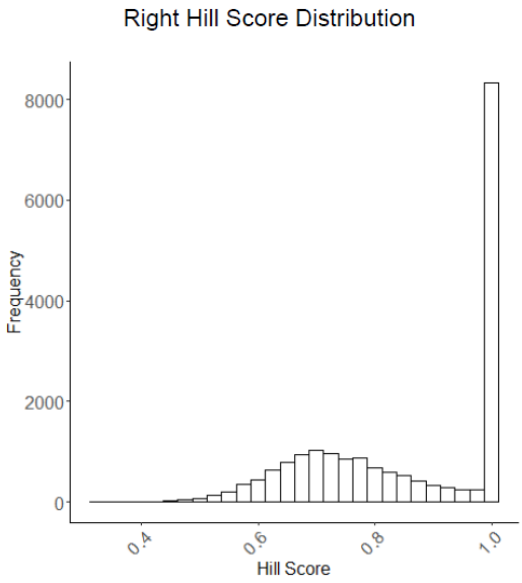


Supplementary Figure 5c, d: Distribution of Pearson Correlation between left hill scores (a), and right hill scores (b) assigned by different parameter combinations, (knot number, spline degree). X-axis shows the correlation and y-axis shows the frequency that the correlation is observed among pairwise comparisons of the parameter sets. The correlations are clustered around 0.8.

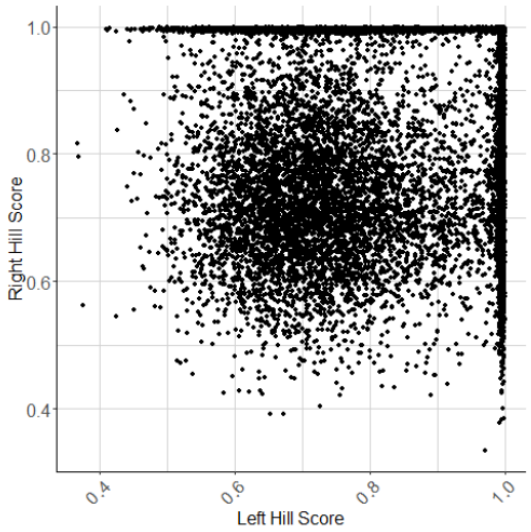
Supplementary Figure 5e



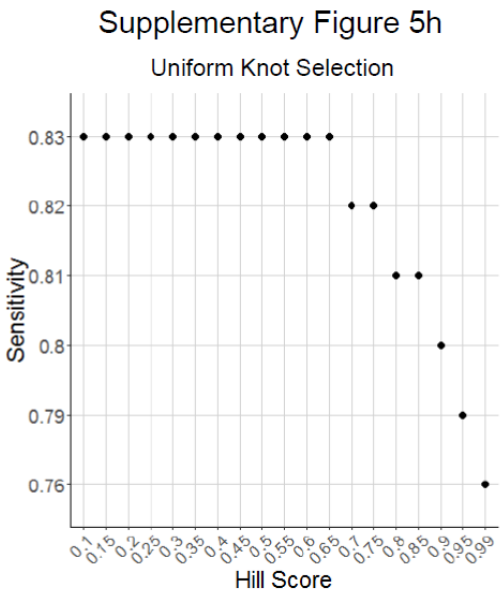
Supplementary Figure 5f

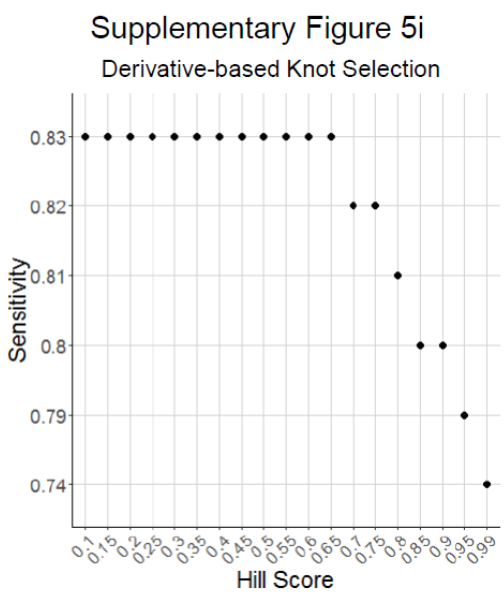


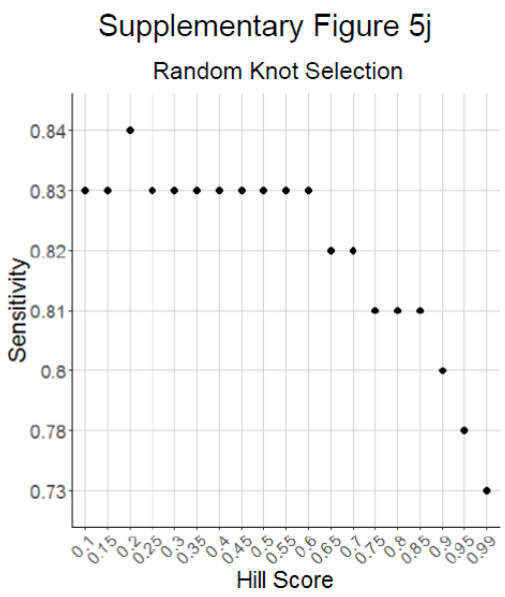
Supplementary Figure 5g



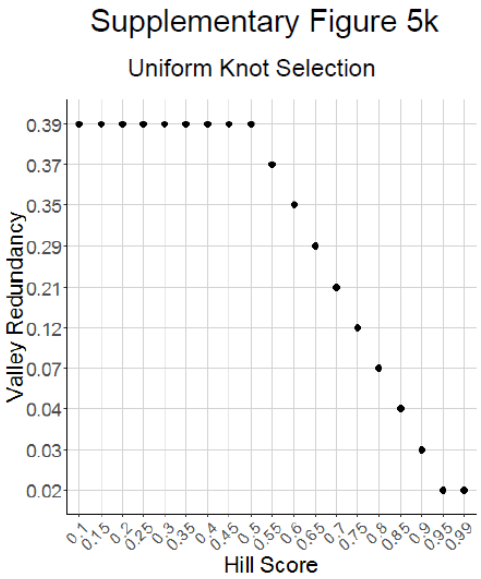
Supplementary Figure 5e, f, g: Distribution of left hill scores (e), distribution of right hill scores (f), and the scatter plot showing the left (x-axis) and right hill scores (y-axis) where each point represents a valley and x and y coordinates are the left and right hill scores, respectively. Note the large amount of clustering of valleys with hill scores equal to 1.

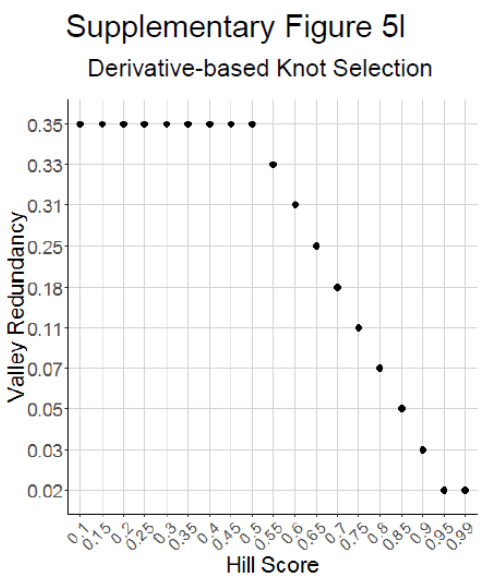


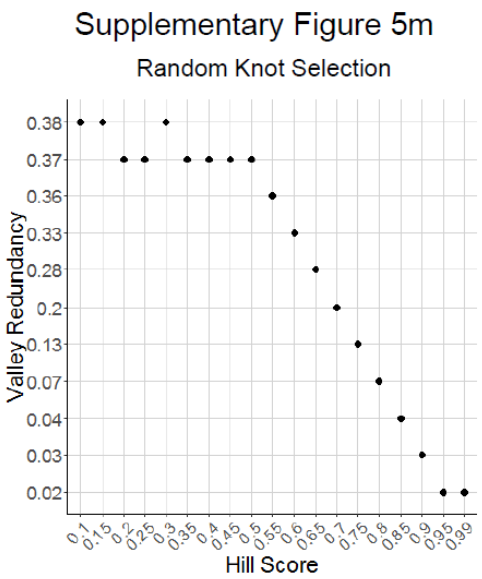




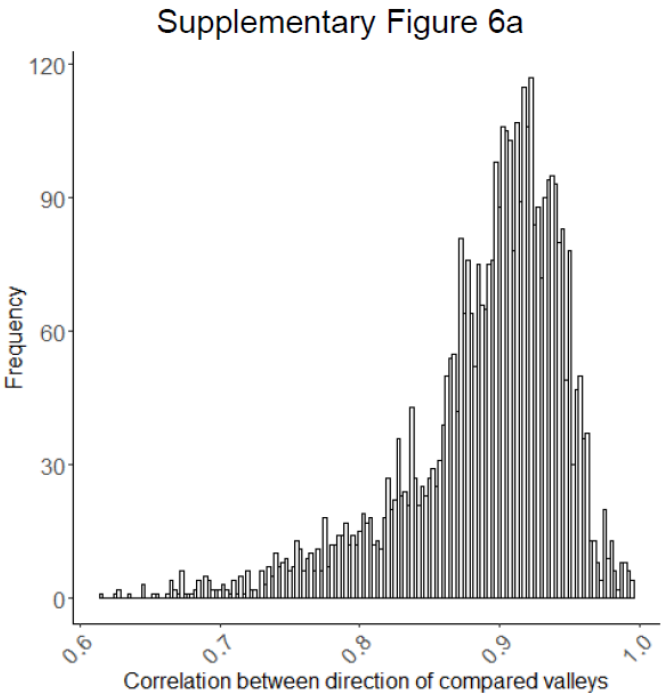
Supplementary Figure 5h, i, j: The sensitivity of the valleys detected by EpiSAFARI with changing hill score cutoff (x-axis) for uniform knot selection (a), derivative-based knot selection (b), and random knot selection (c). X-axis shows the hill score cutoff and y-axis shows the sensitivity of detected valleys.

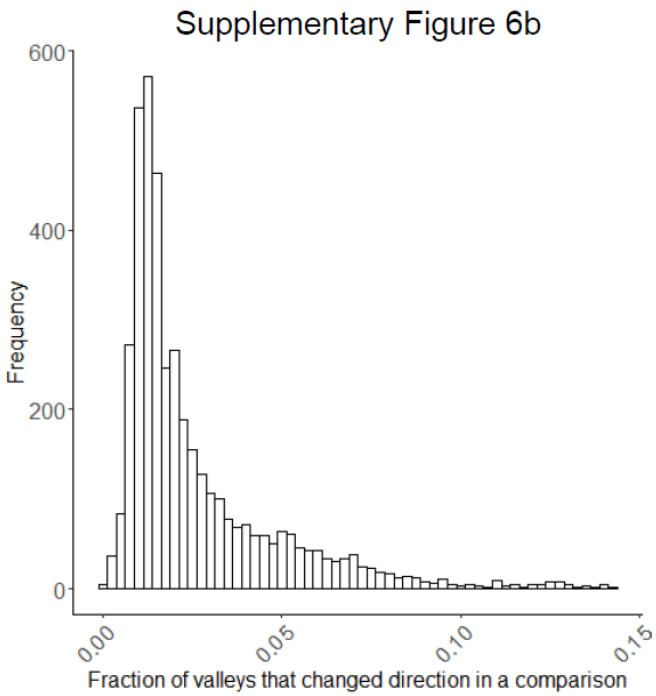






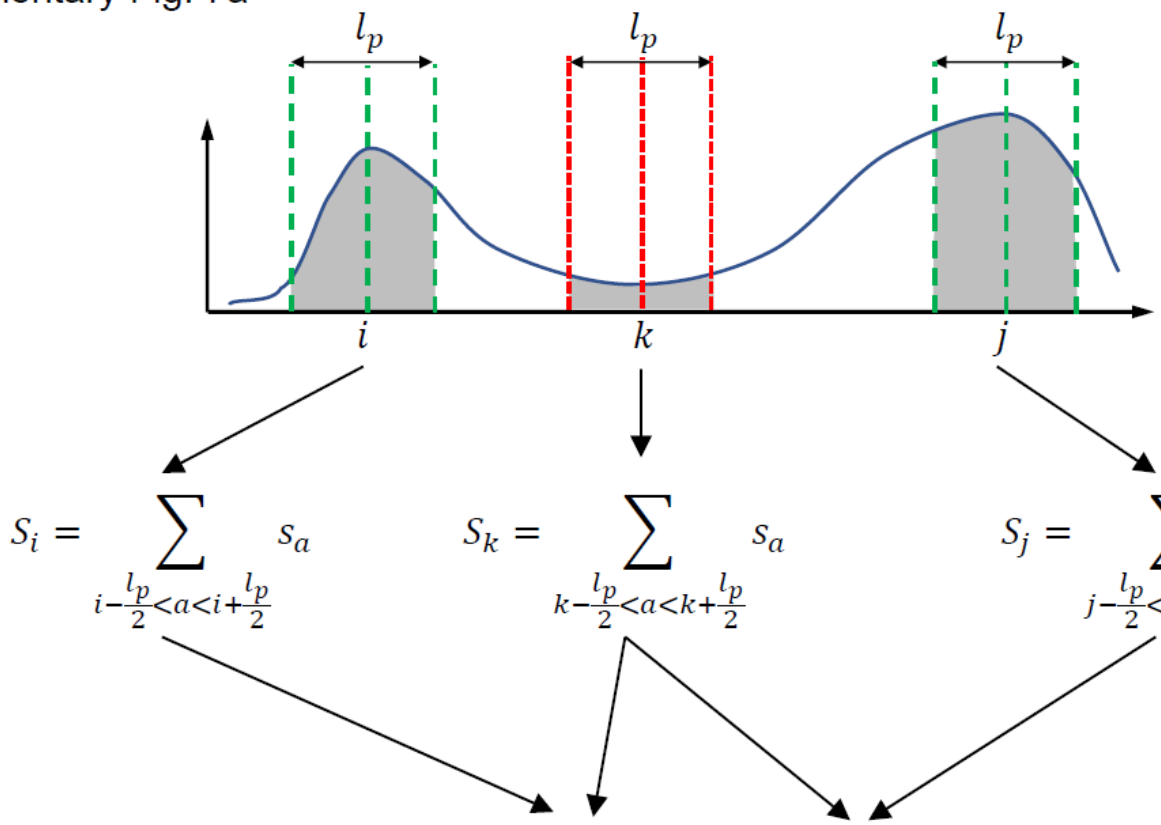
Supplementary Figure 5k, l, m: Illustration of valley redundancy (y-axis) with changing hill score cutoff (x-axis). The changing redundancy with uniform knot selection (a), derivative-based knot selection (b), and random knot selection (c) are plotted.





Supplementary Figure 6a, b: Effect of smoothing parameters on valley asymmetry. a) The distribution of Pearson Correlation of valley asymmetry values between different parameter combinations (knot number, spline degree). X-axis shows the correlation between the valley direction in comparisons. b) The distribution of fraction of valleys that change direction. X-axis shows the fraction of valleys that changed direction in pairwise comparisons. Y-axis shows the frequency of direction changing valleys.

Supplementary Fig. 7a

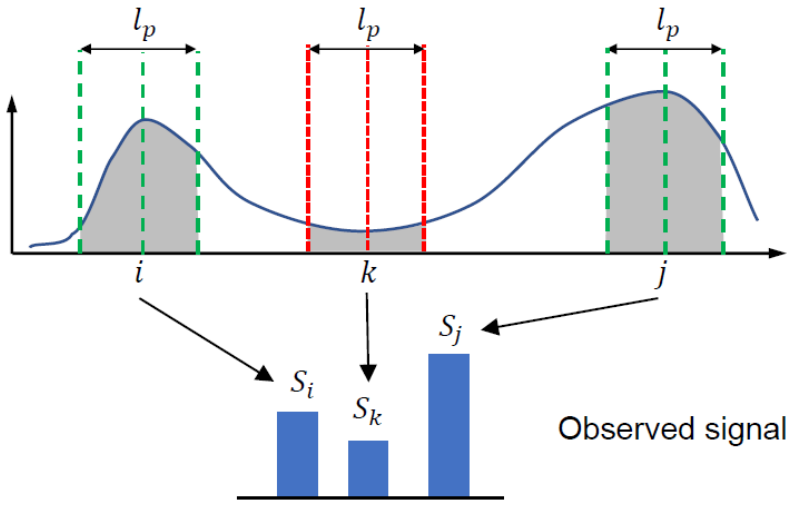


$$\log\left(p-value_{\cap}(i,j,k)\right) = \log\left(bin\left(S_j,S_k\right)\right) + \log\left(bin\left(S_j,S_k\right)\right).$$

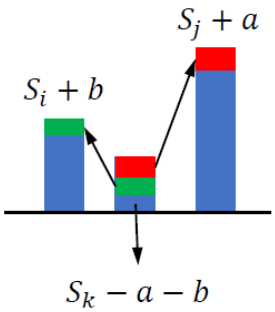
$$\log\left(p-value_{\cup}(i,j,k)\right) = \log\left(bin(S_i,S_k) + bin\left(S_j,S_k\right) - bin\left(S_j,S_k\right) \times bin\left(S_j,S_k\right)\right)$$

Supplementary Figure 7a: Illustration of statistical significance estimation for a valley at (i,j,k) using combination of binomial p-values of summit-to-dip enrichments for left and right summits. The total signal within l_p base pair vicinity of the dip, k, and the summits, i and j, are computed. These are denoted by S_i, S_j, and S_k. The statistical significance of enrichment of signal at summits (S_i, S_j) compared to the dip (S_k) is estimated using binomial test, which are denoted by $bin(S_i, S_k)$ and $bin\left(S_j, S_k\right)$. The p-values are combined using two methods. First is based on intersection of the null models. In this case, the combined p-value is computed by multiplication of the p-values for each hill, or by summing the logarithms of these binomial p-values. Second is based on taking the union of the null models, which corresponds to combining the p-values under the assumption that the p-values are not correlated with each other.

Supplementary Fig. 7b



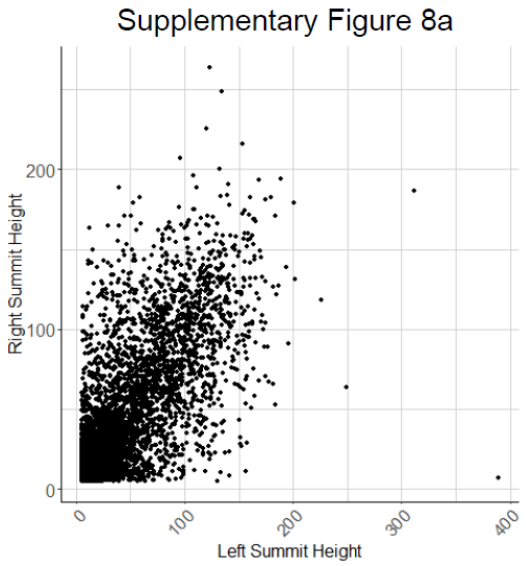
Observed signal configuration

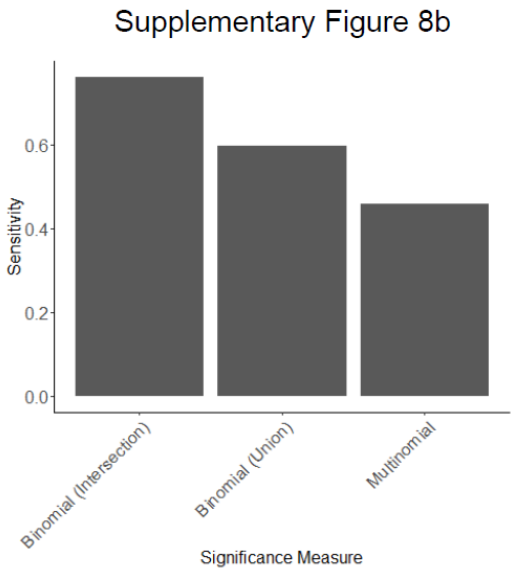


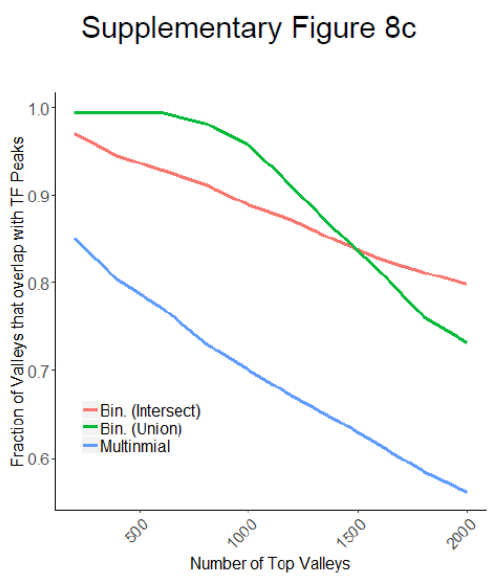
Signal configurations representing more "extreme" valleys

$$Probability = \frac{(S_i + S_j + S_k)!}{(S_k - a - b)! \cdot (S_j + \mathrm{a})! \cdot (S_i - b)!} \cdot \left(\frac{1}{3}\right)^{S_i + S_k + S_j}$$

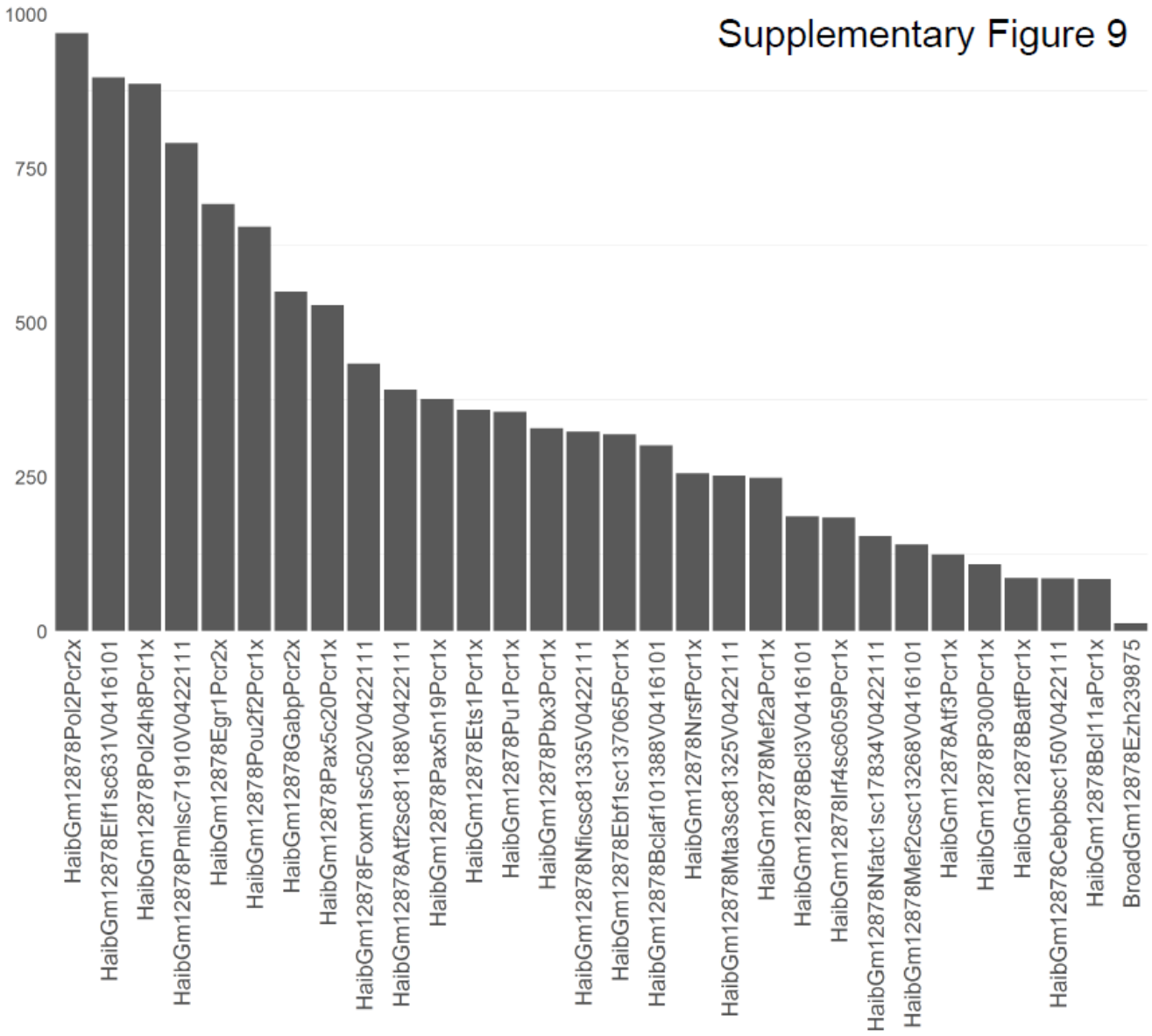
Supplementary Figure 7b: Illustration of the signal configurations in multinomial p-value computation. The signal levels in the vicinity of summits and the dip are computed. In order to compute the multinomial p-value, we enumerate all the signal configurations (illustrated at the bottom) that corresponding to valleys that are equal or more "extreme" than the observed configuration illustrated in the middle. We define more extreme valleys as valleys that have higher signal in either of the summits and lower signal at the dip. Thus, we enumerate all the ways that the signal at the dip, S_k, can be "partitioned" to the summits. In the figure, a and b represent the portions of signal from the dip that is partitioned to right and left summits, respectively. After partitioning, left summit has total signal of $S_i + b$, right summit has $S_j + a$, and the dip has $S_k - a - b$ signal. We next compute the probability of this configuration under null hypothesis that the signal is equally likely distributed to the summits and the dip. The p-value is computed by enumerating all a and b such that $a + b = S_k$.





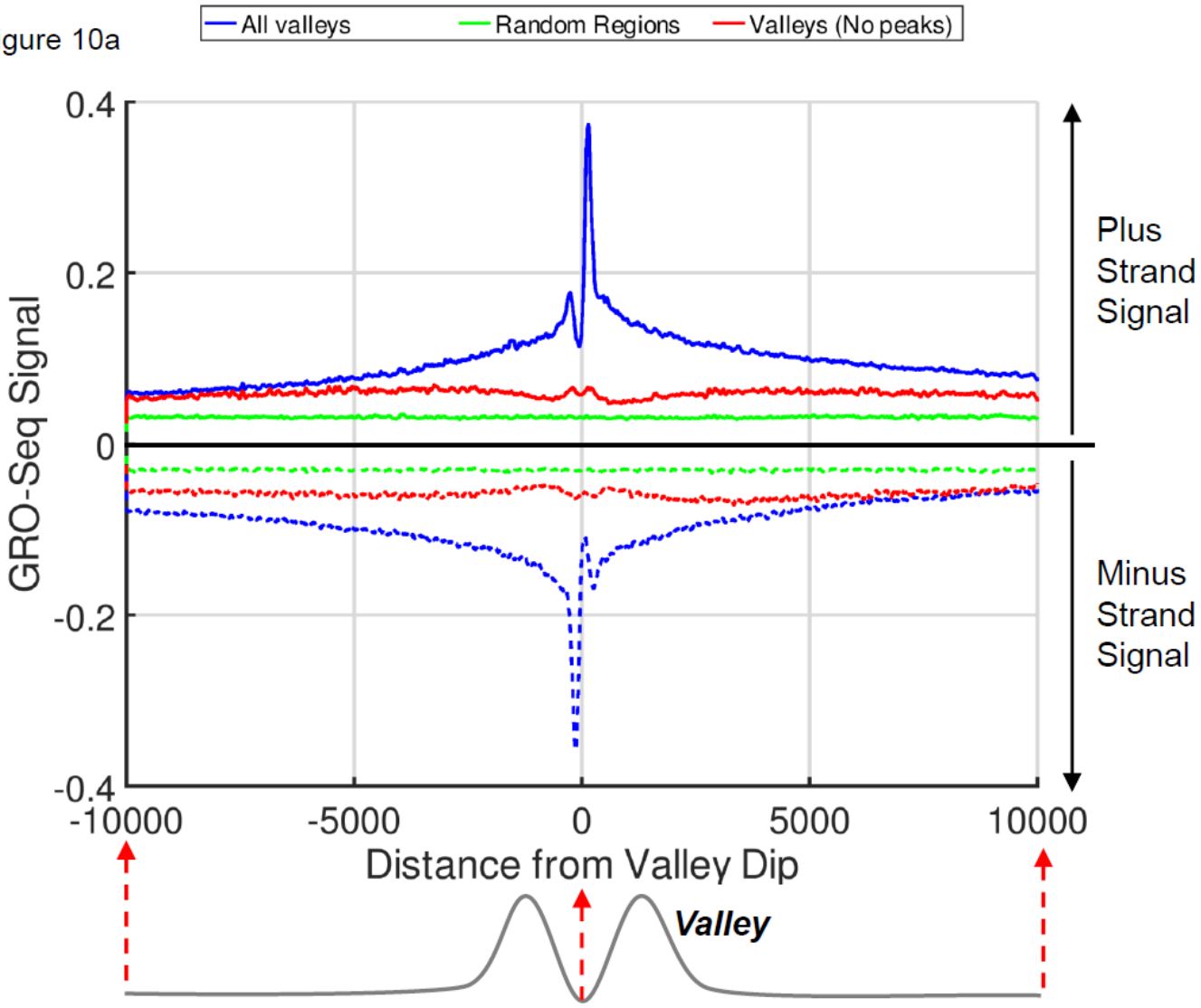


Supplementary Figure 8a, b, c: Impact of p-value computation on accuracy. The scatter plot of left and right summit heights (a) shows a clear correlation between the summit heights. Each dot represents a valley and x and y axis coordinates are the left and right summit heights, respectively. The sensitivity of the detected valleys with different p-value computations is shown in (b). The intersection based binomial p-value combinations provide the most sensitive valley predictions. The fraction of top valleys (with respect to p-value) that overlap with transcription factor (TF) peaks is shown in (c). X-axis shows the number of top peaks and y-axis shows the fraction of valleys that overlap with TF peaks.



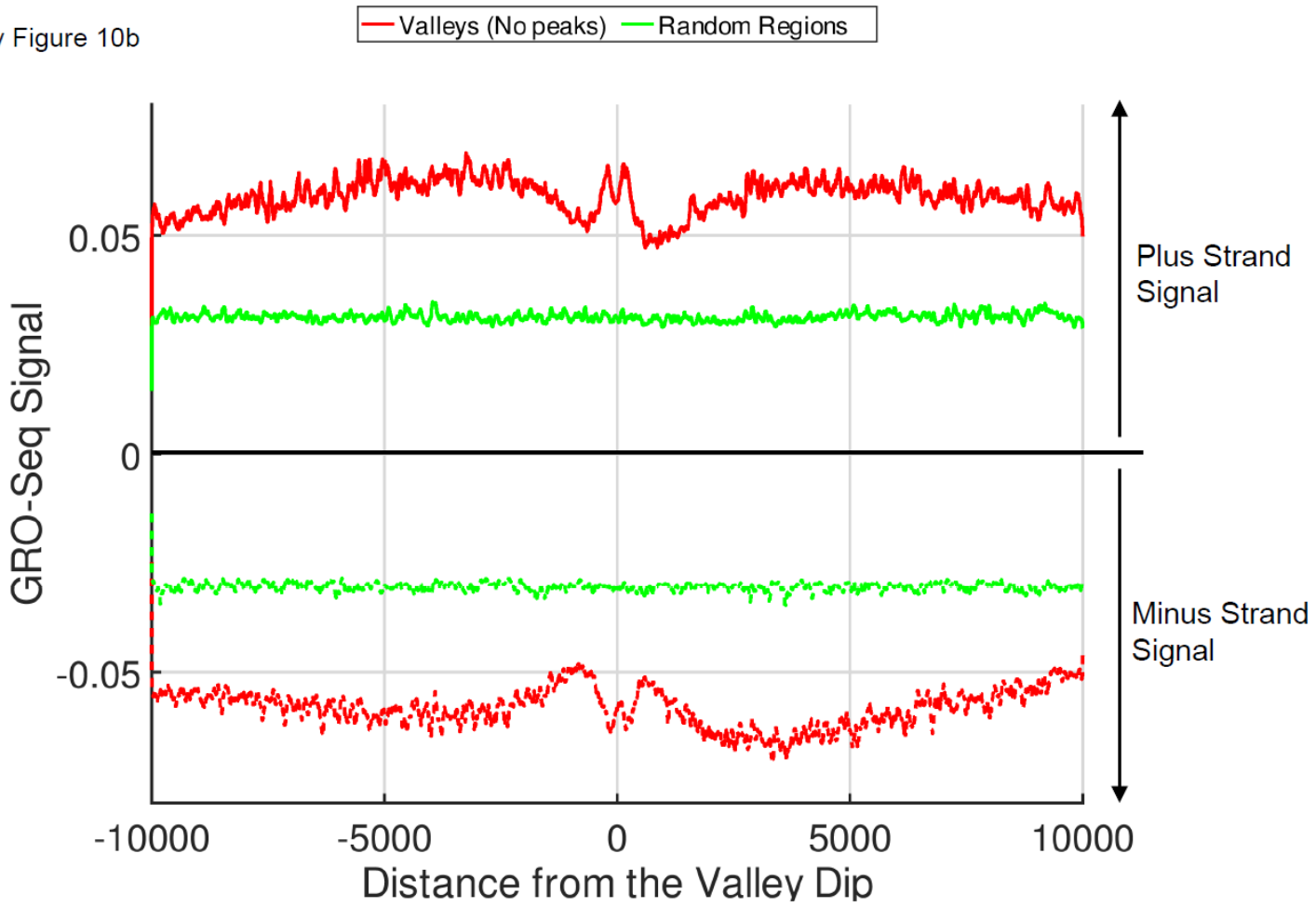
Supplementary Figure 9: Stratification of transcription factors overlapping with H3K4me3 valleys. The number of overlaps between the top 1000 valleys and different transcription factors. X-axis shows the transcription factor and Y-axis shows the number of valleys that overlap with a peak of the corresponding transcription factor. The transcription factors are sorted with respect to decreasing number of valleys overlapping with them.

Supplementary Figure 10a

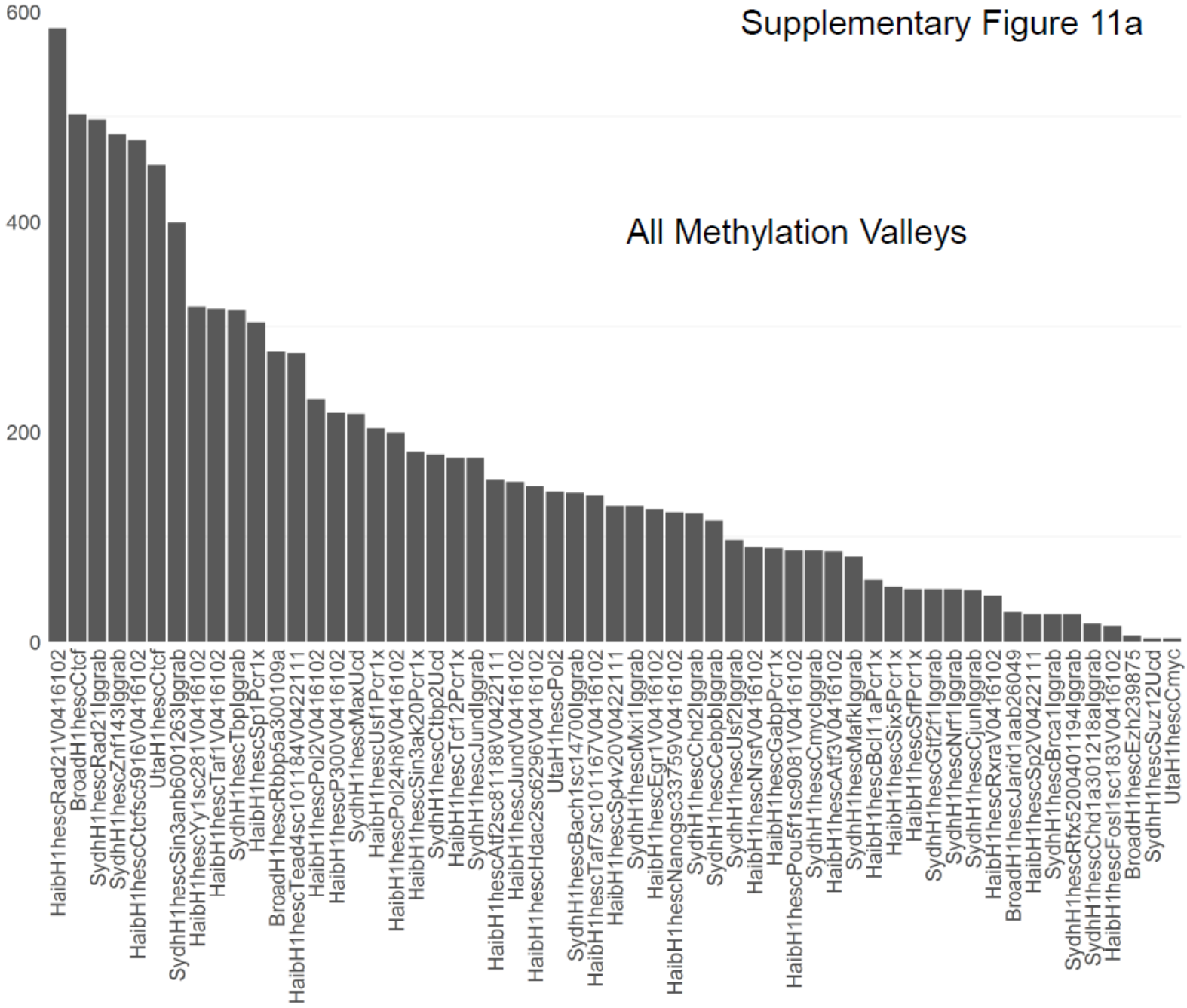


Supplementary Figure 10: Analysis of nascent transcription around the H3K4me3 valleys. a) The aggregation of GRO-Seq signal within 20,000 base pairs of the dips reported by EpiSAFARI. All valleys (Blue), Randomized regions (Green), and valleys that do not overlap with any H3K4me3 peaks (Red) are plotted. Random regions are generated by randomly shifting the valleys within 1 megabase vicinity of the valley's starting position. Plus and minus strand signals are plotted with straight and dashed lines, respectively. The valley below the figure aims to illustrate the valley's positioning within 20 kilobase region. Two humps represent the two summits of the valley. Note that the summit locations in this illustration are not drawn to scale. The red dashed arrows indicate how the dip coordinates align with the x-axis of the aggregation plot.

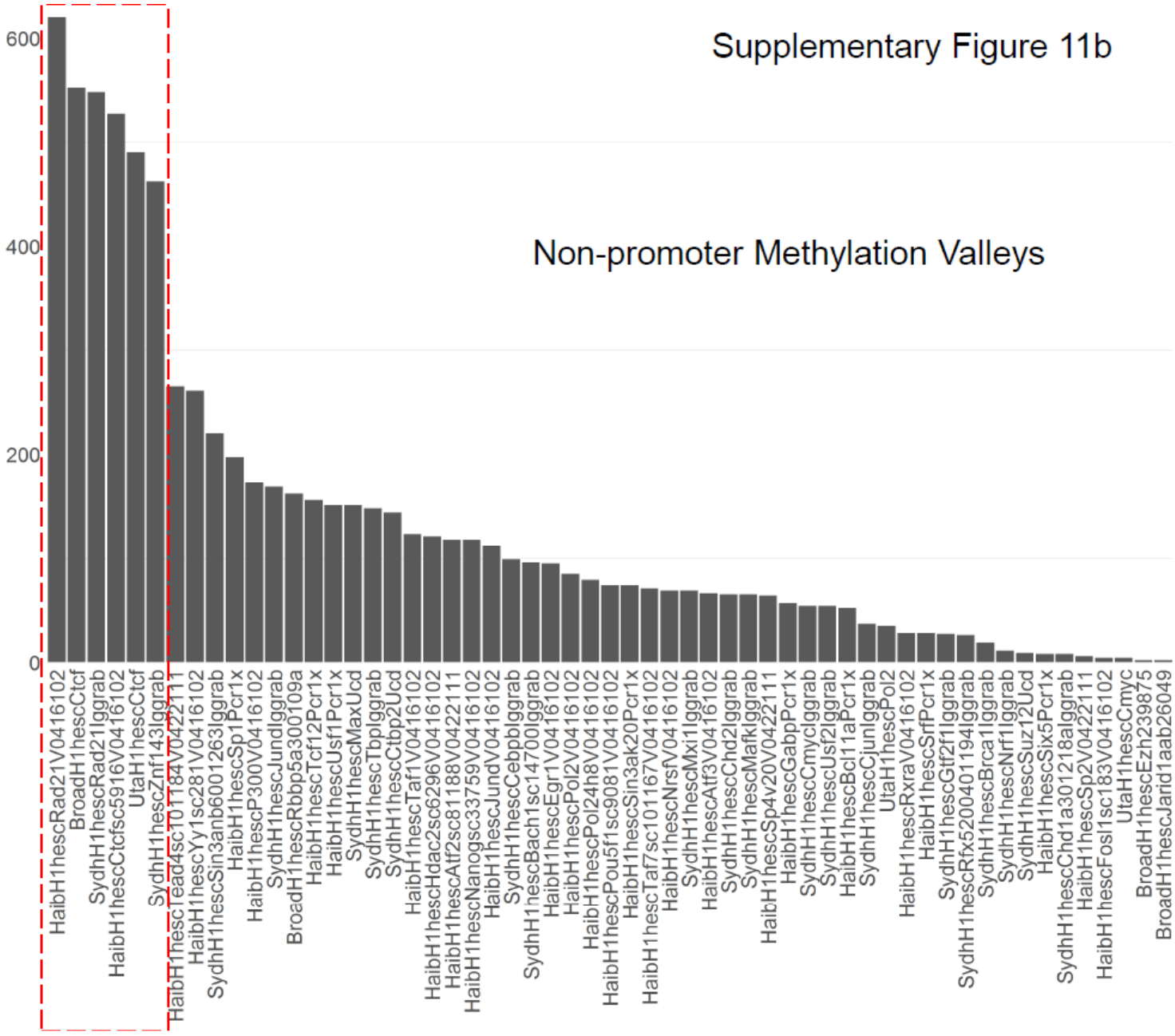
Supplementary Figure 10b



Supplementary Figure 10: Analysis of nascent transcription around the H3K4me3 valleys. b) The aggregation of GRO-Seq signal for valleys that do not overlap with any peaks (Non-peak Valleys) and random regions are plotted to highlight the GRO-Seq signal on non-peak overlapping valleys.

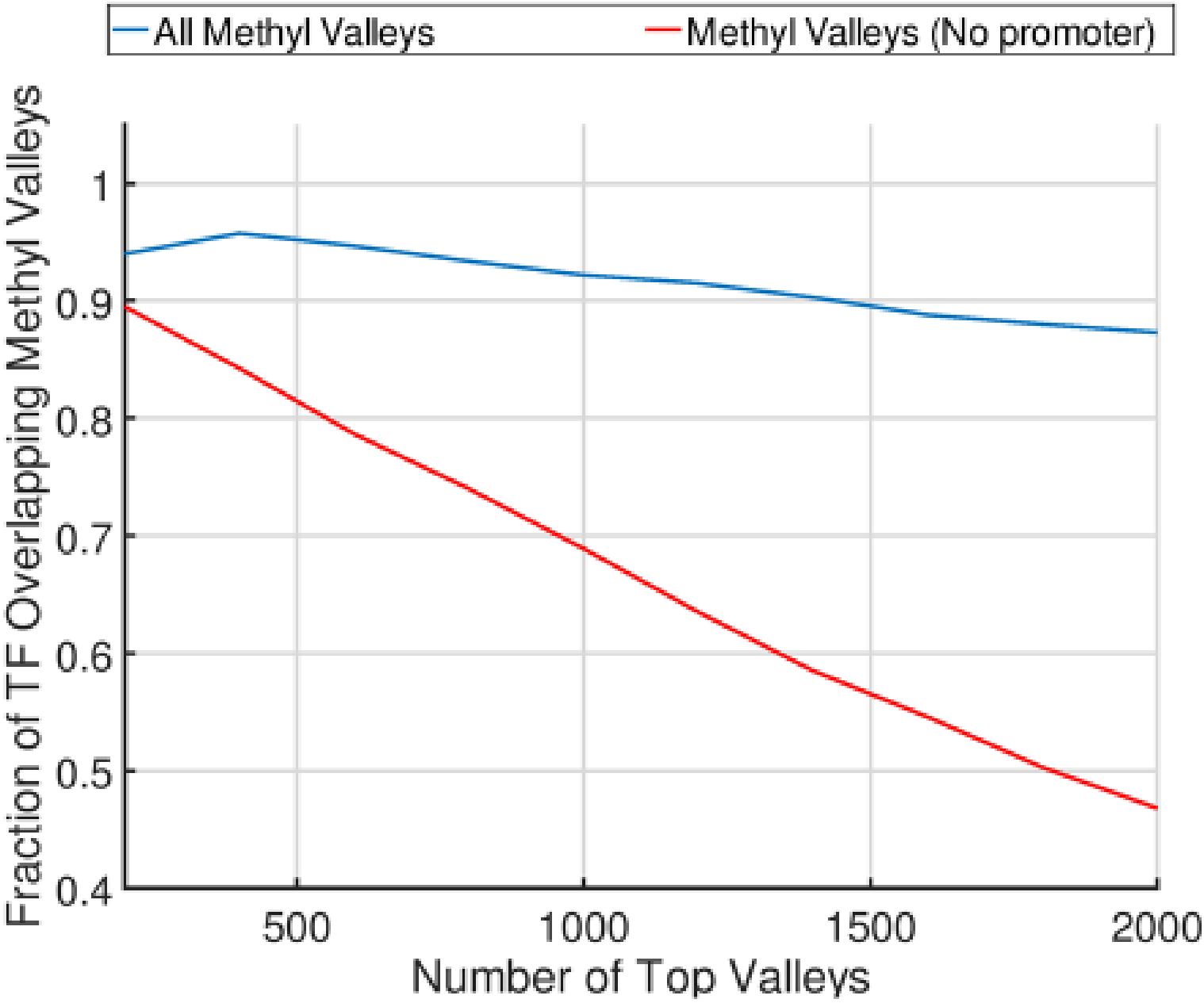


Supplementary Figure 11: Stratification of transcription factor binding around methyl-valleys. a) Fraction of overlap between the top 1000 methyl-valleys and transcription factor peaks. The bars are sorted with respect to decreasing number of overlaps. The transcription factors are shown on x-axis and the number of overlapping methyl-valleys are shown on y-axis.

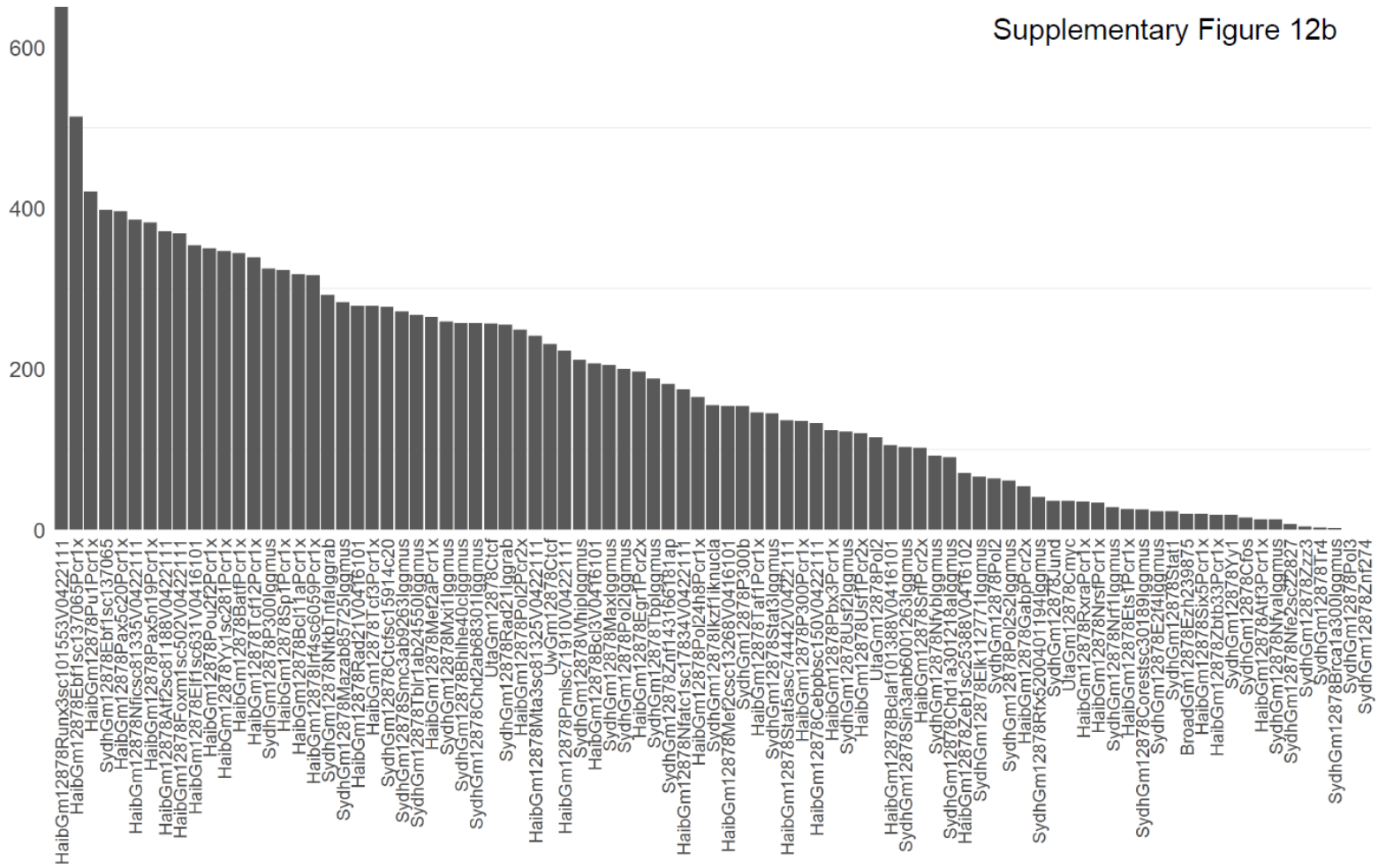


Supplementary Figure 11: Stratification of transcription factor binding around methyl-valleys. b) Overlap of top 1000 non-promoter associated methyl-valleys with transcription factors. The top transcription factors associated with chromatin structure are highlighted with a dashed red rectangle.

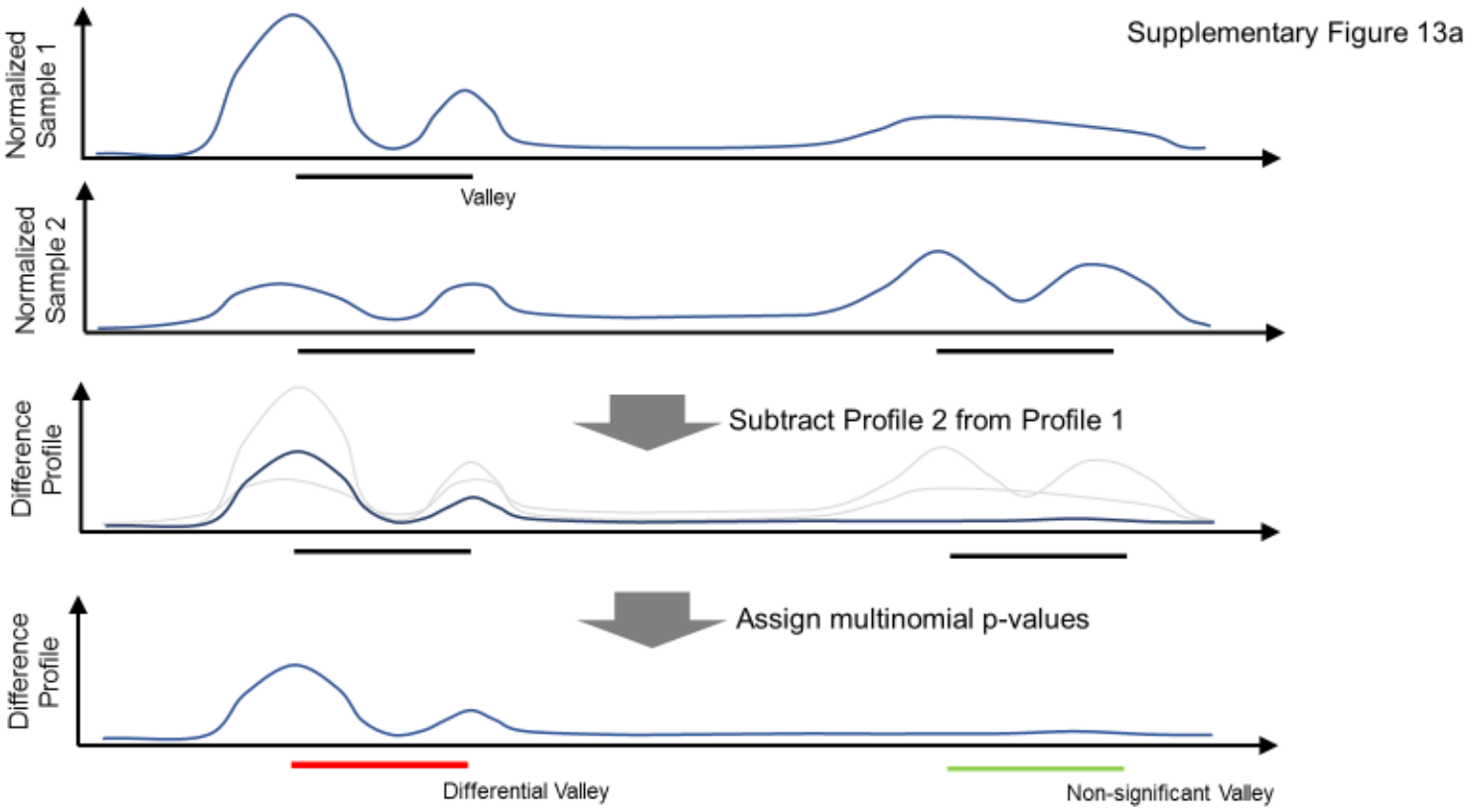
Supplementary Figure 12a



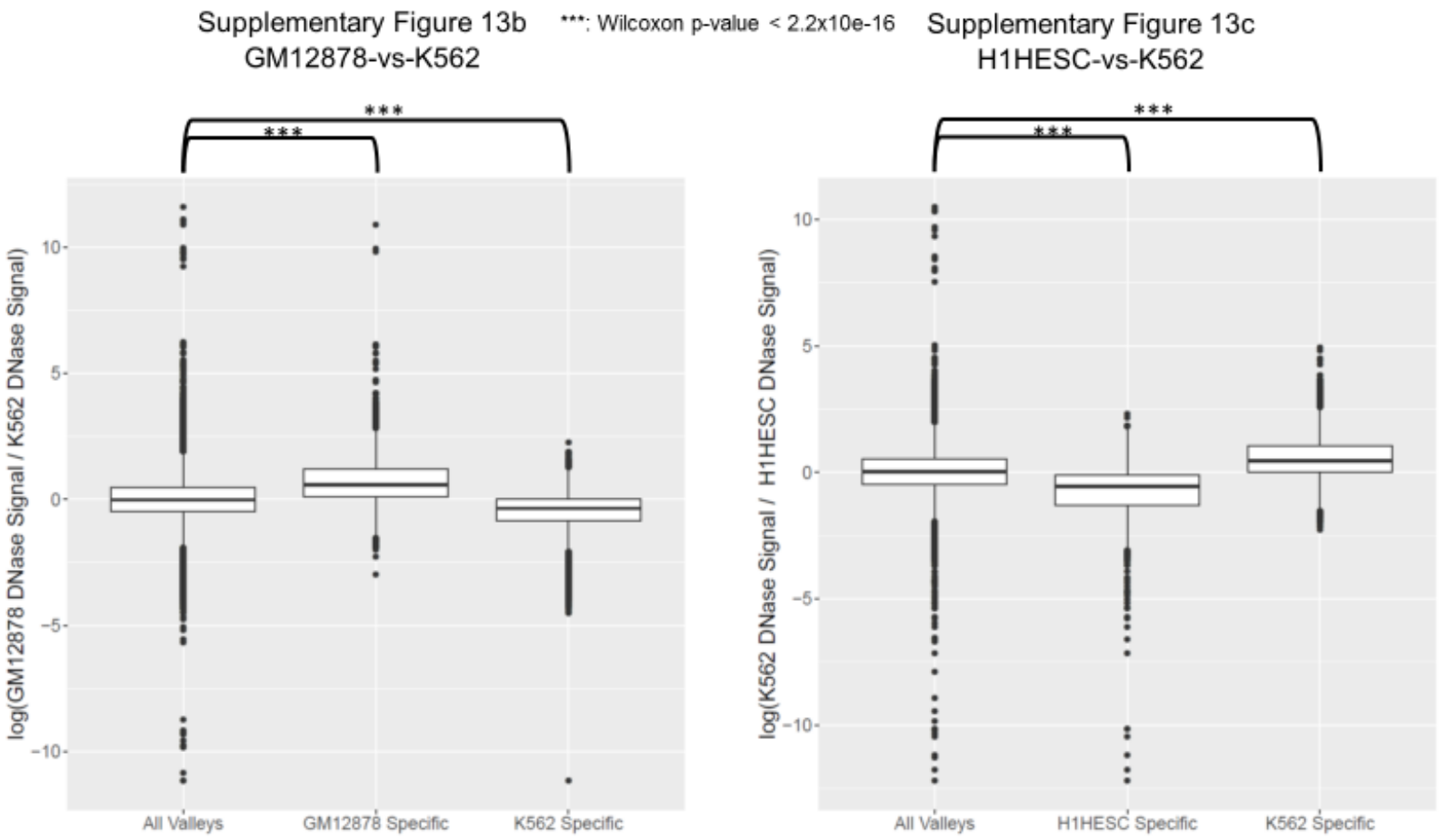
Supplementary Figure 12: a) The fraction of the top 2000 methyl-valleys in GM12878 that overlap with a transcription factor peak.



Supplementary Figure 12: b) The frequency of transcription factors that overlap within top GM12878 methyl-valleys.

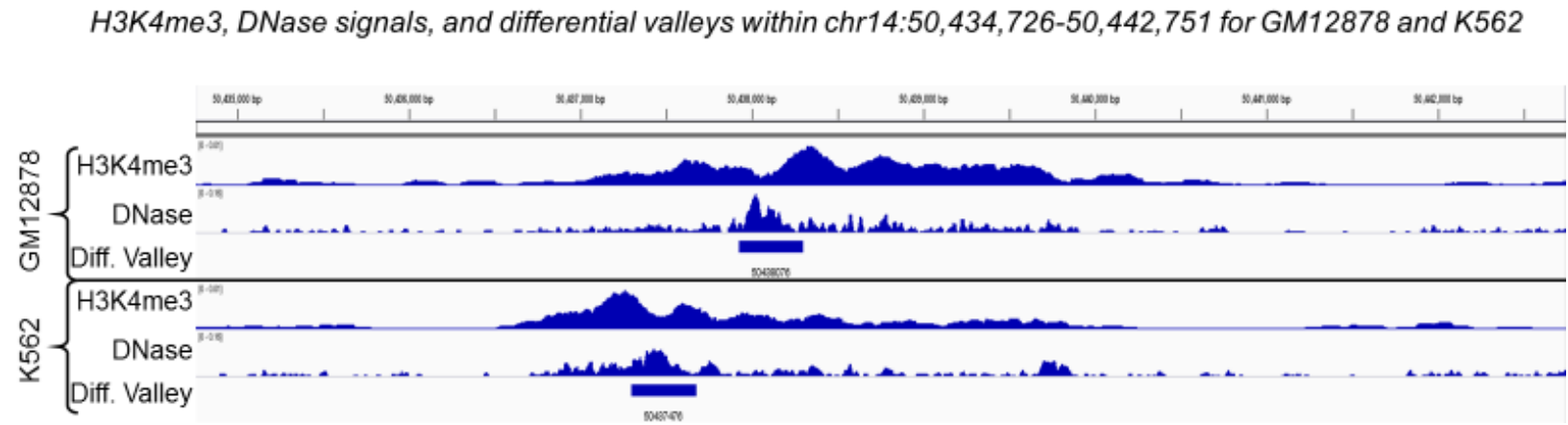


Supplementary Figure 13: a) Detection of differential valleys from two samples. In the example above, the detection of differential valleys in Sample 1 is illustrated. The difference profile is computed by subtracting sample 2's profile from sample 1's profile. Note that the differences that would yield a negative value are set to 0, as shown on the right valley. The difference profile is used to assign differential valley p-value (by multinomial p-value estimation) to each valley in the total set of valleys detected in both samples. In the example above, left valley in sample 1 shows differential activity. However, the right valley detected in sample 2 does not show differential activity in sample 1, as expected. The same comparison is repeated by switching the order of profiles to identify the differential activity of valleys in sample 2.



Supplementary Figure 13b, c) Normalized DNase signal fold change (FC) distribution on cell line specific valleys in differential H3K4me3 valley analysis between GM12878 vs K562 and H1HESC vs K562 cell lines. b) Leftmost barplot shows the distribution of log(GM12878 DNase signal / K562 DNase signal) on all the valleys detected in GM12878 and K562 cell lines. Middle plot shows the distribution of FC for GM12878 specific valleys. Right plot shows the distribution on K562 specific cell line. Note that the distribution DNase FC on GM12878 specific valleys is are skewed towards positive values and towards negative values for K562 cell line. Comparison of DNase FC distributions on cell line specific valleys are significantly different from the FC distribution on all valleys (Wilcoxon rank sum test p-value < 2.2x10e-16 for all comparisons). c) DNase FC distribution on differential H3K4me3 valleys detected from comparison of K562 and H1HESC cell line.

Supplementary Figure 13d



Supplementary Figure 13d: Example of a region with two differential valleys identified by comparing GM12878 and K562 cell lines. Top 3 rows show the H3K4me3, DNase signals, and the GM12878 specific valley. Bottom 3 rows show the H3K4me3, DNase signals, and the K562 specific valley. Although both cell lines have high H3K4me3 signal at this locus, the structure of the H3K4me3 signal (in terms of valleys) is quite different. The two differential valleys in GM12878 and K562 shows differential DNase signal in corresponding cell lines.

REFERENCES

1. Sethi,A., Gu,M., Gumusgoz,E., Chan,L., Yan,K.-K., Rozowsky,J.S., Barozzi,I., Afzal,V., Akiyama,J., Plajzer-Frick,I., *et al.* (2018) A cross-organism framework for supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *bioRxiv*, 10.1101/385237.

2. Xie,W., Schultz,M.D., Lister,R., Hou,Z., Rajagopal,N., Ray,P., Whitaker,J.W., Tian,S., Hawkins,R.D., Leung,D., *et al.* (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, **153**, 1134–1148.

3. Jeong,M., Sun,D., Luo,M., Huang,Y., Challen,G.A., Rodriguez,B., Zhang,X., Chavez,L., Wang,H., Hannah,R., *et al.* (2014) Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat. Genet.*, **46**, 17–23.

4. Jeong,M., Huang,X., Zhang,X., Su,J., Shamim,M.S., Bochkov,I.D., Reyes,J., Jung,H., Heikamp,E., Aiden,A.P., *et al.* (2017) A Cell type-specific Class of Chromatin Loops Anchored at Large DNA Methylation Nadirs. *bioRxiv*, 10.1101/212928.

5. Dorschner,M.O., Weaver,M., Gartler,S.M., Thomas,S., Sandstrom,R., Hansen,R.S., Stamatoyannopoulos,J.A., Thurman,R.E. and Canfield,T.K. (2009) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci.*, **107**, 139–144.

6. Zhou,W., Sherwood,B. and Ji,H. (2017) Computational prediction of the global functional genomic landscape: Applications, methods, and challenges. *Hum. Hered.*, **81**, 88–105.

7. Sherwood,R.I., Hashimoto,T., O'Donnell,C.W., Lewis,S., Barkal,A.A., Van Hoff,J.P., Karun,V., Jaakkola,T. and Gifford,D.K. (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.*, **32**, 171–178.

8. Mieczkowski,J., Cook,A., Bowman,S.K., Mueller,B., Alver,B.H., Kundu,S., Deaton,A.M., Urban,J.A., Larschan,E., Park,P.J., *et al.* (2016) MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.*, **7**.

9. Pajoro,A., Muiño,J.M., Angenent,G.C. and Kaufmann,K. (2018) Profiling nucleosome occupancy by MNase-seq: Experimental protocol and computational analysis. In *Methods in Molecular Biology*.Vol. 1675, pp. 167–181.

10. Jiang,C. and Pugh,B.F. (2009) Nucleosome positioning and gene regulation: Advances through genomics. *Nat. Rev. Genet.*, **10**, 161–172.

11. Voong,L.N., Xi,L., Sebeson,A.C., Xiong,B., Wang,J.P. and Wang,X. (2016) Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell*, **167**, 1555-1570.e15.

12. Li,W., Xu,S., Zhao,G. and Goh,L.P. (2005) Adaptive knot placement in B-spline curve approximation. In *CAD Computer Aided Design*.Vol. 37, pp. 791–797.

13. Kasowski,M., Kyriazopoulou-Panagiotopoulou,S., Grubert,F., Zaugg,J.B., Kundaje,A., Liu,Y., Boyle,A.P., Zhang,Q.C., Zakharia,F., Spacek,D. V, *et al.* (2013) Extensive variation in chromatin states across humans. *Sci. (New York, NY)*, **342**, 750–752.

14. Jung,Y.L., Luquette,L.J., Ho,J.W.K., Ferrari,F., Tolstorukov,M., Minoda,A., Issner,R., Epstein,C.B., Karpen,G.H., Kuroda,M.I., *et al.* (2014) Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res.*, **42**.

15. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

16. Harmanci,A., Rozowsky,J. and Gerstein,M. (2014) MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol.*, **15**, 474.