An integrated software system for analyzing ChIP-chip and ChIP-seq data

(Supplementary Information)

Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, Wing H Wong

SUPPLEMENTARY NOTES

A review of ChIP-chip and ChIP-seq data analysis

ChIP-chip and ChIP-seq are powerful technologies to study transcriptional regulation in complex genomes, however, mining information from the huge datasets generated by these high-throughput technologies remains to be a non-trivial task. To analyze a ChIP-chip experiment, one usually starts with data exploration and then goes through normalization, binding region detection, adding gene annotations and finding enriched sequence motifs. This is a multiple step procedure, and the data involved are heterogeneous. In the past few years, a number of tools targeting each individual steps of the ChIP-chip analysis have been developed. For example, microarray blob remover (MBR)[24] has been developed to detect and remove blob-like defects from array images. Quantile normalization[50] which was originally developed for normalizing probe intensities across multiple expression arrays, is also used widely in the tiling array analysis. MAT model[11] was proposed to remove sequence-dependent probe effects in the Affymetrix tiling arrays, and MA2C[22], a model-based normalization approach based on the GC content of probes, was developed for two-color tiling arrays. For detecting binding regions from normalized array data, methods based on moving windows (MAT[11], TileMap[12], TAS[13]), hidden Markov models (HMMTiling[51], TileMap[12], Du et al.[17]), hierarchical mixture models (TileHGMM[15], BAC[52]), as well as regression and kernel deconvolution (MPeak[14], JBD[18], MeDiChI[23]) have been proposed. Tilescope[21] provides a web-based data processing pipeline for analyzing tiling arrays, and Ringo[53] is a R/Bioconductor package for ChIP-chip analysis which allows users to take the advantage of various functions in R. Other methods often used include TAMAL[20], ACME[19], TiMAT (http://sourceforge.net/projects/timat2), Splitter (http://zlab.bu.edu/splitter), and ADM-1 (http://www.cs.technion.ac.il/users/wwwb/cgi-bin/tr-get.cgi/2007/PHD/PHD-2007-05.pdf). For motif discovery, various methods have been developed, among which the most popular ones are MEME[42], Gibbs Motif Sampler[35], and variants of them[54-56]. Methods that try to identify cis-regulatory modules (CisModule[36], Gibbs Module Sampler[57], EMCMODULE[58]) and that incorporates cross-species information into the *de novo* motif discovery (Wasserman et al.[59], PhyloCon[60], PhyME[61], CompareProspector[62], PhyloGibbs[63], Ortholog sampler[64], MultiModule[65], etc.) have also been developed. In addition to these general motif discovery methods, methods specifically targeting at motif analysis on ChIP-chip data are also available, examples including MDSCAN[25] and MotifBooster[26]. To retrieve gene annotations, tools such as Galaxy[43] and CEAS[28] have been made available.

Despite the development of various tools, mining information from the huge tiling array datasets remains to be a non-trivial task. This is due to multiple reasons. First, many tools for the upstream data processing (normalization and peak detection) are designed for a single array platform (e.g., TAS and MAT for Affymetrix, MPeak for NimbleGen), making it awkward to compare data collected from multiple array platforms. A few tools such as Tilescope can support multiple array platforms, but they are incapable of doing ChIP-seq analysis. Therefore it is still difficult to integrate information from the ChIP-chip with information from ChIP-seq which becomes increasingly useful. Second, different upstream and downstream analysis functions are distributed in a dozen of tools, often incompatible with each other. Significant amount of work are required to reformat output of one piece of software before feeding it to the other. The web-service CEAS[28] makes efforts to integrate multiple downstream analysis functions including sequence retrieval, adding gene annotations and motif discovery. However, it performs the analysis and returns the results in a pre-defined manner, and there is limited flexibility for users to customize the analysis procedures to meet their diversified needs. Galaxy allows users to do analyses on genomic intervals in a flexible way, but it does not support various kinds of upstream analyses (e.g. peak detection) and downstream analyses (e.g. motif discovery) that are particularly useful for ChIP data analyses. Third, the ability to visualize the data easily and interactively is a critical requirement for effective analysis. Although tools like IGB (http://www.affymetrix.com/support/developer/tools/download_igb.affx) and SignalMap (http://www.nimblegen.com/products/software/signalmap.html) have been developed for visualizing array signals along chromosomes, the former is mainly designed for Affymetrix tiling arrays, and the latter is a proprietary software provided for NimbleGen users. General-purpose genome browsers such as UCSC[33] and Ensembl[34] are useful tools to visualize ChIP data. However, when thousands of predictions and tens of millions of data points need to be visually examined in a large-scale interactive analysis, these browsers become highly inefficient due to the need of transferring data over the internet. Moreover, many visualization functions particularly useful in the ChIP data analyses are not provided by these tools. For example, to visualize motif information, one needs to go to other websites such as WebLogo[44]. Therefore, in order for a bench biologist to efficiently perform all the upstream and downstream analyses, an integrated tool that can support flexible and seamless analyses of ChIP-chip data is urgently needed.

Analysis of a ChIP-seq experiment begins with aligning reads to the genome and finding read enriched regions. The predicted regions can then be used for downstream analyses including motif discovery and annotation retrieval. Although the downstream analyses can be performed in a similar fashion as the ChIP-chip analysis, development of methods for the upstream analyses of ChIP-seq data is still at its infancy. ELAND (Cox A., unpublished) provides a fast algorithm to align millions of reads to the genome, allowing up to two mismatches and no gaps in the alignment. More recently, new tools such as SOAP[66], RMAP[67], ZOOM[68] and SeqMap[47] have also been developed to align reads generated by massively parallel sequencing to reference genomes. Given the aligned reads, early ChIP-seq studies[4-7] used in-house analysis pipelines to detect binding regions which are often difficult for general bench biologists to use. Recently, a few tools targeting general users have been developed, including GeneTrack[29], QuEST[30] and SISSRs[31]. GeneTrack uses a Gaussian smoothing procedure to produce a continuous curve representing signals across the genome, it then looks for peaks by finding maxima in the curve. It does not provide statistical error rate measurements. QuEST combines a Gaussian kernel with the read directionality information to infer binding sites. It provides a false discovery rate (FDR) estimate which is obtained using information from the negative control sample. In order to compute the FDR, the negative control sample is required to have twice as many reads as the ChIP sample, so that it can be divided into two parts to mimic a random two sample comparison. When no negative control sample is available, QuEST is not able to provide error rate estimates. SISSRs combines the read enrichment with the read directionality to identify binding regions. It uses a Poisson model to estimate FDR when only the ChIP'd sample is available. When the negative control sample is available, the control sample is used to control specificity and sensitivity of the predictions. The control of sensitivity is based on the empirical read distribution in the negative control sample, and the control of specificity is based on empirical p-values computed for fold changes between the ChIP'd and control sample. No FDR is provided in this context. Despite the recent development, our understanding on the basic characteristics of ChIP-seq data is still limited. In particular, there are two types of ChIP-seq experiments: experiments involving negative control samples (two-sample analyses) and experiments that contain only ChIP'd samples (one-sample analyses). Knowledge about their relative merits and limitations is limited due to lack of a direct comparison between the two types of analyses. Moreover, in the one-sample analyses, since no negative control information is available to estimate the noise level, evaluation of statistical significance is challenging. Currently, the published methods either use a Poisson model[5,10,31] or use Monte Carol simulations[8] to construct the null distribution in the one-sample context. Both approaches implicitly assume that the background read occurrence rate is a constant, which is an assumption that has not been carefully examined before. Not only do we know so little about the data, our ability to handle the data is also limited due to the same reasons that caused the bottleneck in the ChIP-chip analyses. These include lack of tools to efficiently visualize tens of millions of reads in the ChIP-seq data (without the need to transfer over the internet), tools to integrate ChIP-seq with ChIP-chip data, tools that seamlessly connect the upstream analyses to downstream analyses, and tools that allow users to flexibly design analysis pipelines to meet the needs of individual studies.

SUPPLEMENTARY METHODS

TileMapv2 – CisGenome's internal ChIP-chip peak caller

CisGenome incorporated a new version of TileMap[12] as the internal ChIP-chip peak caller. Compared to the old version, the new TileMapv2 has incorporated several new features.

First, the old version supports only the two-sample and multiple-sample tests. In the new version, support for one-sample test has been added. The one-sample test is needed for analyzing NimbleGen and Agilent ChIP-chip data. Unlike the Affymetrix technology where the ChIP and control samples are hybridized to separate arrays, the NimbleGen and Agilent technology typically hybridize a ChIP sample and a control sample simultaneously to a single array. Each array will produce a log ratio between the two samples. From statistical point of view, we only have one group of data (i.e. log ratios) instead of two groups of data (i.e. IP vs. control, such as the data in Affymetrix arrays). The one-sample test tries to detect binding regions by evaluating whether the log ratios are significantly bigger than zero.

Second, TileMapv2 has incorporated a new option for computing FDR under the moving average (MA) mode. The original FDR computation was based on an unbalanced mixture subtraction method (UMS), which tends to generate very conservative FDR estimates. The new option makes the assumption that when the distribution of the MA statistics of all probes is plotted in a histogram, the left tail of the distribution represents the noise. To estimate FDR, one first specifies a MA cutoff c to detect peaks. All probes with a MA statistic $\geq c$ are selected to form peaks. The program will then detect negative control peaks by selecting probes with a MA statistic $\leq -c$. Let y_1 denote the total number of peaks, and let y_2 denote the total number of negative control peaks. The FDR at the MA cutoff c is then estimated as y_2/y_1. Indeed the program does not only compute the FDR for the user-specified cutoff c, but also compute the FDR for all peaks, by applying the same approach to the MA statistic associated with each individual peak. Under this new option, choosing a cutoff around FDR≤5-10% usually optimizes the cutoff under the E-O distance criteria defined in ref. 41. This new FDR computation is now set as the default for the TileMap-MA and was used throughout the paper.

Third, a new option has been added to TileMapv2 to allow users to exclude from the analysis the outlier probes listed in the outlier section of the Affymetrix CEL files. With this option, users can now process files generated by Microarray Blob Remover (MBR)[24] which is a software tool to remove certain array artifacts.

Transcription factor binding site mapping

Both consensus sequences and position specific weight matrices can be mapped to genomes, lists of genomic regions, or FASTA sequence files. To map consensus sequences, one can use degenerate patterns and choose allowed number of mismatches. To map a matrix, the matrix is used to scan the genome. At each position, the likelihood ratio (LR) between the motif model and a third order background Markov model is computed. Sites with LR greater than a user-chosen cutoff are reported. Users can choose to compute the background model from the input sequences or use pre-computed background models. In the latter case, the background models are computed using the whole genome sequences and are allowed to vary across the genome.

De novo motif discovery

Gibbs motif sampler[35] is provided for *de novo* motif discovery. CisModule[36] is provided for novel *cis*-regulatory module discovery. Performance of these algorithms was discussed in their original publications as well as in ref. 69-71. New motif discovery algorithms will be added in future if they can improve the performance substantially.

Motif enrichment analysis based on matched genomic control regions

When applied to analyzing ChIP data, current *de novo* motif finders often return multiple motifs. It is not always clear as to which motif corresponds to the key pattern directly recognized by the transcription factor in question, since the reported motif scores (e.g., the MDSCAN[25] score) often rank the real target motif lower than the more abundant but less relevant GC-rich or highly repeated motifs. We have previously shown that this problem can be solved by re-ranking motifs according to their relative enrichment levels[37]. The relative enrichment level of a motif is computed as its occurrence rate in the binding regions divided by its occurrence rate in negative genomic control regions. When the negative control regions are carefully chosen to match the physical distribution of the binding regions, the key motif will usually stand out as the one with the highest relative enrichment level. The use of the matched genomic controls is critical, since the method will not work if the negative control regions are randomly chosen from the genome. Given a genome and a list of binding regions, CisGenome provides a function to generate matched genomic control regions using the method described in ref. 37.

Support for different species

Many CisGenome functions are species-independent (e.g., ChIP-chip/ChIP-seq peak detection and *de novo* motif discovery). The others require information of a particular species (e.g., gene-peak association and sequence retrieval). We build genome databases to support species-dependent analyses. The databases are coded into binary formats to facilitate efficient data access and visualization. Precompiled databases for four commonly used species including human, mouse, Drosophila and Arabidopsis can be downloaded from the CisGenome website and are ready to use. We routinely update these databases to support analyses on new genome assemblies. Databases for other species will be gradually added in future. Meanwhile, users can build their own databases for other species by applying the database construction functions provided by CisGenome to raw data downloaded from the UCSC genome browser.

Developing language and operating systems

CisGenome is developed in ANSI C/C++. The core data analysis functions can be compiled and run on multiple platforms including MS Windows, Linux and MacOS. The current version of GUI and browser can only be used with MS windows.

ChIP-seq read mapping by SeqMap

SeqMap[47] is a fast sequence mapping software. Unlike BLAT, SeqMap indexes the short sequences rather than the genome. Given the maximal numbers of mutations, insertions and deletions allowed, SeqMap splits the short sequences into several parts. By keeping some parts rather than all of them to be fixed, the non-candidates can be eliminated in the very first step. All the candidates that are left will then be collected and a local alignment algorithm will be run on them to finally determine the matched targets. Similar algorithm has been used several times in some paper[72] and software (ELAND by Illumina/Solexa). However, to the best of our knowledge, SeqMap is the first to extend this algorithm for insertion/deletion detection. .

Model fitting and FDR computation in the one-sample ChIP-seq analysis

To fit the Poisson model, for each n = 0, 1, 2, …, we count how many windows have n sequence reads and denote the counts by u_n. We assume that windows with a small number of reads are mainly background. Since $\Pr(n=1)/\Pr(n=0) = \lambda_0$, the ratio u_1/u_o provides an estimate for λ_0. Similarly, u_2/u_1 provides an estimate for $\lambda_0/2$, etc. One can take averages of u_1/u_o, $2u_2/u_1$, … to estimate λ_0.

To fit the negative binomial model, since $r_1=\Pr(n=1)/\Pr(n=0) = \alpha/(\beta+1)$, and $r_2=\Pr(n=2)/\Pr(n=1) = (\alpha+1)/[2(\beta+1)]$, we have $\alpha = r_1/(2r_2-r_1)$, and $\beta = 1/(2r_2-r_1)-1$. We first use u_1/u_o to estimate r_1 and use u_2/u_1 to estimate r_2, then we plug in r_1 and r_2 to the formulas above to estimate α and β.

With λ_0, α and β estimated, we also estimate what percentage of windows are background. This is estimated by taking the ratio between the theoretical $\Pr(n=0)$ and the observed frequency of $n=0$ (One may also use (n=0)+(n=1) instead of (n=0) only).

With the fitted null model available, CisGenome then counts the number of windows that contain n reads for each n = 0, 1, 2, …. The observed number is compared with the number expected by the null model, and the ratio between the two is reported, from which a false discovery rate can be computed for each n. One can then choose an appropriate cutoff accordingly.

FDR computation in the two-sample ChIP-seq analysis

Genome is divided into non-overlapping windows with length w. For each window, the number of reads in the ChIP sample k_{1i}, the number of reads in the negative control sample k_{2i} and the total read number $n_i= k_{1i}+k_{2i}$ are counted. Using windows that contain a small number of reads (i.e., small n_i), the expected sampling ratio between the ChIP and the negative control sample in non-binding regions is estimated as $r_0=\Sigma\, k_{1i} / \Sigma\, k_{2i}$. It should be pointed out that this rate usually is different from the ratio between the total number of reads in the ChIP sample and the total number of reads in the control sample, because in the ChIP sample, a large proportion of reads are sampled from binding regions, whereas in the control sample this is not the case.

Next, we group windows according to n_i. For each group (n = 0, 1, 2, …), the observed distribution of k_{1i} is compared with what is expected by Binomial(n, $p_0=r_0/(1+r_0)$), and a false discovery rate is computed accordingly.

With this information, all windows with $n_i \geq c$ and $k_{1i}-k_{2i}$ large enough to pass the selected FDR cutoff will be selected to form binding regions. Here, the cutoff c for n_i serves as an auxiliary criterion and can be chosen based on the negative binomial model described above.

Post-processing in the ChIP-seq analysis

In most sequencing technologies (e.g., Illumina/Solexa), reads are generated from both ends of ChIP fragments through 5' -> 3' DNA synthesis. Therefore, when one considers reads that are aligned to the forward strand of the genome separately from the reads that are aligned to the reverse complement strand, one would observe two peaks separated by certain offset at each binding location[49]. The forward strand peak is located on the left, and protein-DNA interaction is sitting in between them (**Supplementary Fig. 1, Fig. 2d**).

In the post-processing step, we take advantage of the separation between the forward strand and reverse strand reads to refine binding region boundaries. We first use a w bp sliding window to scan each binding region and count forward strand and reverse strand reads separately. This will produce two smooth curves of read counts. We then identify the modes of the two curves and use their locations to define binding region boundaries (**Fig. 2d**). This *boundary refinement* step can greatly improve the resolution of binding region detection. As an optional *single strand filtering* step, one can further filter binding regions by repeating the exploration and peak detection step separately for forward strand reads and reverse strand reads, and reporting a binding region only if it contains a significant forward strand peak and a significant reverse strand peak simultaneously. Regions that are retained after the boundary refinement and single strand filtering are defined as high quality binding regions.

SUPPLEMENTARY DATA

1. Evaluation of the FDR estimation method in the one-sample ChIP-seq analysis

Analysis of real data

We applied both the negative binomial and Poisson model to three real ChIP samples (NRSF[4], Oct4 (ref. 10) and Nanog[10]) (**Supplementary Table 1**). The estimated FDR based on these two models were then compared to a model-independent reference FDR. In order to obtain the reference FDR for a particular ChIP sample, we applied the one-sample analysis to the corresponding negative control sample with the same number of reads. The reference FDR was computed as the (No. of predictions in the control sample / No. of predictions in the ChIP sample). The reference FDR was independent of any parametric model assumptions. For all three transcription factors, the Poisson model consistently underestimated the reference FDR. In contrast, the negative binomial model provided conservative (in the case of NRSF and Oct4) or reasonable (in the case of Nanog) FDR estimates when the FDR is in the range from 0.1 to 0.5 (**Fig. 2c**). When the reference FDR was small (FDR < 0.1), the negative binomial model was too optimistic. In real applications, a comprehensive prediction list is typically obtained by using FDR cutoffs in the range 0.1 to 0.5. A smaller FDR may serve as a stringent cutoff to pick up high confidence predictions for experimental follow-up. However, for the purpose of selecting candidates for experimental follow-up, the limiting factor is usually the number of candidates we can afford to test rather than the number of statistically significant candidates. In this context, peak ranking is the primary aim and FDR estimation becomes less relevant.

Analysis of simulated spike-in data

As indicated by **Figure 2b**, the empirical distribution of the window read count in negative control samples tends to have a heavier tail than the negative binomial model fitting. As a result, the negative binomial model tends to underestimate the FDR when the window read count is high (**Fig. 2c**). This suggests that not all variations in the background can be explained by the negative binomial model. To understand how much bias this lack-of-fit could introduce to the FDR estimation, we performed a systematic simulation study. 18 simulated spike-in data sets were generated by introducing varying number of peaks with varying enrichment levels into two real negative control samples. The two real negative control samples were collected from the NRSF study[4] and the embryonic stem cell study[10] respectively (**Supplementary Table 1,2**).

To generate a simulated spike-in dataset from a real negative control sample, we first computed n_b = (No. of reads in the real negative control sample / No. of non-overlapping 100bp windows in the genome). We then randomly picked up p locations in the genome to serve as simulated peak centers. For each peak i, we generated n_i=r_i*n_b reads where r_i was a random number drawn from a exponential distribution with mean r. The exponential distribution was chosen because it roughly matched the observed distribution of IP enrichment in real ChIP samples. r was used to control the overall binding strength (or the IP enrichment). In the real NRSF, Oct4 and Nanog ChIP'd samples, this average enrichment was estimated to be 692, 70 and 66 respectively (estimated by averaging the ratios [No. of reads in the 100bp window at the peak center / No. of reads expected in a random 100bp window] across all the peaks). After reads are generated, they are randomly distributed around each peak center, with the distances to the peak center sampled from a normal distribution N(μ=0, σ^2=225). Finally, the computer generated reads were combined with the reads from the real negative control sample to create the spike-in data. By setting peak number *p*=2000, 10000, 50000, and enrichment ratio *r*=20, 100, 500, nine data sets were generated for each of the two negative control samples (**Supplementary Table 2**).

For each simulated dataset, we applied the one-sample analysis to find peaks and estimate the FDR. The estimated FDR was compared with the real FDR. The results (**Supplementary Fig. 3,4**) show that the bias of FDR estimation is correlated with the peak number and the binding strength. When the number of peaks was relatively small (*p*=2000) and the binding signal was weak (*r*=20), the negative binomial model significantly underestimated the real FDR. The bias diminished when the number of real peaks *p* increased or the binding signal *r* became stronger. When *p* and *r* were in the range of what we observed in real data, the estimated FDR was reasonably close to the real FDR.

Also consistent with the simulation results, when we applied the one-sample analysis to the two real negative control samples, a small number of peaks were detected at the 10% FDR level, even though no peaks should be expected from negative controls (**Supplementary Table 3**). The false predictions are caused by the biased estimates of FDR.

Nature of the bias

To understand the nature of unexplained background variations that caused the bias in FDR estimation, we have taken a closer look at the "peaks" detected from the negative control samples. The peaks from the NRSF control sample covered 0.5% of the reads in the sample. Among the reads covered by the peaks, 96.4% were aligned to repeat elements. The peaks from the ES control sample covered 1% of the reads, among which 65.4% were aligned to repeats (**Supplementary Table 3**). As a comparison, peaks detected from the real ChIP samples covered 12-18% of the reads, and among them only 13-28% were aligned to repeats. This suggests that repeat regions contributed a significant portion of background variation that was not explained by the negative binomial model. Artifacts in repeat regions can happen in many possible ways, including but not limited to sequencing errors, polymorphisms, and misalignment caused by errors in reference genome assemblies. How to incorporate these artifacts into the background model is an interesting topic for future research.

Besides the peaks associated with repeats, there are also some peaks detected in non-repeat elements. No firm explanations were established for the enrichment of reads in these regions although open versus closed chromatin structure has been proposed as a potential cause.

The peaks detected from the negative control samples represent artifacts that the current one-sample analyses are not able to control. To remove such biases, two-sample analyses are needed. When we checked the length of the "peaks" detected from the negative control samples, it was found that reads covered by these peaks were clustered within 150bp in 92% of the peaks detected from the NRSF control sample and in 80% of the peaks from the ES control sample (see **Supplementary Fig. 5a-d** for some examples). The remaining peaks (which mainly came from repeat regions) can often be decomposed into small clusters of reads, with each cluster occupying ~100bp (**Supplementary Fig. 5e**). Thus, the span of these artifacts matched well with the window size (w=100-200bp) typically used in the two-sample analyses.

Diagnosis of problematic FDR estimation

Due to the potential bias of the FDR estimation, two-sample experiments are always the preferred design. When one-sample experiments are performed for cost consideration or other reasons, it would be useful to have some guidelines to tell whether the data quality is good enough so that the FDR estimates based on the negative binomial model are not problematic. There are multiple types of information that may indicate low data quality (in terms of FDR estimation).

First, the overall signals can be indicated by the number of reads contributing to the peaks. If the percentage of reads that are covered by the peaks is low, it may indicate that the FDR estimation is problematic. In the simulation study, datasets where the FDR estimation performed well all contained more than 5% of the reads within the peak regions. In contrast, in all datasets where peaks covered only ~1% of the reads, the FDR estimates were problematic (**Supplementary Table 2, Supplementary Fig. 3,4**). When the real ChIP-seq data were analyzed, in the two negative control samples, only ≤1% of the reads were covered by "peaks" detected at the 10% FDR level. In contrast, at the same FDR level, peaks identified from the three ChIP samples covered ≥10% of the ChIP reads (**Supplementary Table 3**). Even at the very stringent FDR level (e.g. FDR=10^{-6}), the detected peaks still covered ≥5% of the reads in the ChIP samples (**Supplementary Fig. 6**).

Second, if the detected peaks are repeat-rich, it may indicate low data quality (**Supplementary Table 3**).

Third, if the binding motif of the transcription factor is known, it can be used as an independent source of information to evaluate data quality. In peaks detected from the real ChIP samples, we often observe significant enrichment of the key motif, and the enrichment level is expected to clearly decrease with the peak rank (**Supplementary Fig. 7**). When no such pattern is observed and motif enrichment level is low, there is indication of problematic data quality.

Based on our current knowledge, it is recommended to always use multiple criteria to evaluate data quality when one-sample experiments are performed. When the predicted peaks cover ≤10% of the reads, or ≥50% reads in the peaks are aligned to repeats, or no expected motif enrichment is observed, adding a negative control sample to the experiment is recommended. CisGenome provides a useful tool for analyzing such experiments, as the wide range of functionalities offered by the software makes these multiple types of composite analyses accessible to the bench biologists.

2. Examination of array-specific peaks using non-canonical NRSF motifs

The motif used in the NRSF analysis (**Table 1**) was the canonical NRSF motif which contained two half sites separated by 11bp (**Supplementary Fig. 9**). In a previous study, Johnson et al.[4] showed that many binding regions that do not contain the canonical NRSF motif may contain non-canonical NRSF motifs. In the non-canonical motifs, the two half sites are separated by 16~20bp.

In the NRSF analysis, we have shown that only 1.23% of the array-specific peaks contained the canonical NRSF motif. A natural question is whether the array-specific peaks are more likely to contain non-canonical NRSF motifs and therefore do not represent false discoveries. To address this issue, we mapped the non-canonical NRSF motifs to the binding regions (using LR≥500 as the cutoff) and summarized the results in **Supplementary Table 7**. Based on the results, only 0.53% (29/5,517) of the array-specific peaks that do not contain the canonical NRSF motif contained the non-canonical NRSF motifs. As a comparison, 2.47% (176/7,114) of all array peaks, 9.24% (128/1,385) of ChIP-seq specific peaks, and 9.14% (145/1,587) of peaks common to all three analyses contained the non-canonical motifs.

We further mapped the two half sites of the NRSF motif to the genome and asked whether the half sites were enriched in array-specific binding regions (**Supplementary Table 8**). None of the half sites was enriched in array-specific binding regions, but both were enriched in peaks common to the ChIP-chip and ChIP-seq analyses, and enriched in peaks that were detected by ChIP-seq only.

Together, these suggest that the non-canonical NRSF motifs are not enriched in array-specific peaks. It further confirms that the array-specific peaks may represent technical noise.

3. Analysis of Oct4 and Nanog ChIP-seq data

We collected the Oct4 and Nanog ChIP-seq data from ref. 10 (**Supplementary Table 1**). The experiment contains a negative control sample that was used in both the Oct4 and Nanog analyses. We applied both the one-sample and two-sample analyses to these two transcription factors, and the peak detection results are shown in **Supplementary Fig. 10**. For Oct4, the concordance between one-sample and two-sample analysis was 74% before post-processing. After post-processing, the concordance increased to 96%. For Nanog, the concordance between one-sample and two-sample analysis before post-processing was 59%. After post-processing, the concordance increased to 83%. Saturation analysis (**Supplementary Fig. 11**) showed that the Nanog data was close to saturation. Both the Oct4 and NRSF did not show saturation before the post-processing, but the curve corresponding to peaks after post-processing started to level off.

4. Searching TRANSFAC database for matches to the novel motif

We searched TRANSFAC database (Professional 10.5) to find potential matches to the novel motif using three different approaches.

First, we used the PATCH function provided by the TRANSFAC to search for all known binding sites ≥10 bp long that match any part of the GGACTACAATTCCCAGCAA consensus with ≥70% identity. The returned results contained binding sites that are recognized by transcription factors c-Rel, NF-kappaB, p50, RelA-p65, Ncx, STAT3, PEA3, PU.1, STAT5 and STAT6. When the sequence logos of the corresponding binding motifs were examined, no pattern was found to match the novel motif.

Next, we collected all 525 human and mouse motif matrices from the TRANSFAC. We generated forward and reverse complement sequence logos for all of them using CisGenome browser and visually examined them one by one. No match to the novel motif was found.

Finally, in order to make sure that the visual examination did not miss any potential matches, we computed Euclidian distances between the TRANSFAC motifs and the novel motif. Let L1 denote the length of the novel motif. For each TRANSFAC motif with length L2, we slid the TRANSFAC motif along the novel motif and examined all the L1+L2-1 possible alignment windows (e.g., if L2=20, then the alignment between position 1→1 of motif 1 and position 20→20 of motif 2, between position 1→2 of motif 1 and position 19→20 of motif 2, between position 1→3 of motif 1 and position 18→20 of motif 2, etc. were examined). For each alignment window, a Euclidian distance is computed as the square root of $\Sigma_i\Sigma_j(p_{1ij}-p_{2ij})^2$, where p_{kij} is the occurrence frequency of nucleotide j at the window position i for motif k, $j \in \{A,C,G,T\}$, and $i \in [1, \text{window length}]$. We repeated the same procedure on the reverse complement strand of the TRANSFAC motif. Among all the possible alignment windows, the maximal window length is min(L1,L2), and the minimal window length is 1. For each window length, the smallest Euclidian distance is recorded for the TRANSFAC motif. After this computation, each TRANSFAC motif will have several distances recorded, one for each window length. In total, the number of distances recorded for each TRANSFAC motif is min(L1,L2). After all TRANSFAC motifs have been processed, for each possible window length from 6 to 18, we identified the top 5 TRANSFAC motifs that had the smallest distances to the novel motif. We carefully examined the corresponding sequence logos. No matches to the novel motif were found in this analysis. We also tried to replace the Euclidian distance by Kullback-Leibler distance, i.e., replacing $\Sigma_i\Sigma_j(p_{1ij}-p_{2ij})^2$ by $\min\{\Sigma_i\Sigma_j\ p_{1ij}*\log_2(p_{1ij}/p_{2ij}),\ \Sigma_i\Sigma_j\ p_{2ij}*\log_2(p_{2ij}/p_{1ij})\}$, and again no matches were found.

5. Functional context of the novel motif

The novel motif was discovered by analyzing human Sox2 and Nanog ChIP-chip data set on promoter arrays. It can either represent a motif that functions specifically in the embryonic stem cell context, or it can represent a general promoter element not directly related to the stem cell function. To see whether it is related to general stem cell functions, we further analyzed the whole genome ChIP-PET data for Oct4 and Nanog in mouse[73]. *De novo* motif discovery on the mouse Oct4 and Nanog ChIP-PET binding regions did not find this motif. Further examination showed that although the motif was enriched in human Sox2 and Nanog ChIP-chip binding regions identified by promoter arrays, it was not enriched in mouse Oct4 and Nanog binding regions identified by the genome-wide ChIP-PET (**Supplementary Table 10**), suggesting that this motif may not have a direct role in embryonic stem cells but is more likely to be a general promoter element. The strong evidence for the motif being functional though indicates that future investigation of the motif in a more general context is worthwhile.

6. Comparison of ChIP-chip analysis algorithms

CisGenome uses an upgraded version of TileMap, TileMapv2, as the internal ChIP-chip peak detection algorithm (see **Supplementary Methods**). We compared TileMapv2 with a number of other ChIP-chip analysis algorithms using the recently published spike-in data[41]. The benchmark datasets contained spike-in ChIP-chip data generated by different labs and from three different array platforms (Affymetrix, Agilent, NimbleGen).

To analyze the Affymetrix arrays, raw data were quantile normalized, and TileMapv2 was run under the Moving Average (MA) mode. To analyze the NimbleGen and Agilent arrays, raw Cy5 and Cy3 data from all arrays within an experiment were quantile normalized. Log2(Cy5/Cy3) ratio was computed, and TileMapv2 (MA) was then applied to the log ratios.

TileMap-MA method requires users to set a window size W. In the original TileMap paper, W=5 was recommended for analyzing Affymetrix arrays with a 35bp probe spacing. Under this setting, information from 2*5+1=11 probes will be pooled to compute the MA statistics for the center probe. To analyze the spike-in data on Affymetrix arrays, we adjusted the W based on the platform-specific probe spacing. For Affymetrix Encode 2.0R arrays with a 7bp probe spacing, W=25 (=5*35/7) was used. For Affymetrix Encode 1.0R arrays with a 22bp probe spacing, W=8 (=5*35/22) was used. The NimbleGen Encode arrays had a 38bp probe spacing, and W=5 (=5*35/38) was used. Agilent arrays had a 100bp probe spacing, representing a much lower density. If we were to use the same principle, W should be set to 2 (=5*35/100). However, according to our previous experience on analyzing Agilent custom arrays with a 125bp spacing[74], W=2 would not be sufficient to eliminate random noise, and setting W=3-5 would generate more robust results. Therefore, for analyzing the Agilent arrays, W was set to 3.

In TileMapv2, probes with a MA statistic bigger than certain MA cutoff will be picked up to form potential binding regions. For each potential binding region, a FDR will then be computed. Users have the freedom to choose a FDR cutoff after getting the peak predictions. In the analyses here, FDR≤10% was used to define the final peak list in all spike-in data sets, consistent with the analysis of the NRSF ChIP-chip data. The default MA cutoff is MA≥3. In some analyses, these produced fewer than 100 peaks. When this was the case, we relaxed the cutoff to MA≥2.5 in order to obtain approximately 100 peaks, so that the number of true positives (#TP), false negatives (#FN) and false positives (#FP) among the top 100 peaks can be compared with the other algorithms. Other than the principles described above, we did not try to optimize the TileMap parameters.

Following the previously described procedure in Johnson et al.[41], we derived the ROC-like curve for TileMap (**Supplementary Fig. 13**) and computed the area under the ROC curve (AUC) (**Supplementary Fig. 14)**, the E-O distance (i.e., the distance between the TileMap 10% FDR cutoff and the optimal cutoff), as well as the number of true positives, false negatives and false positives among the top 100 peaks (**Supplementary Table 11**). Compared with the other algorithms, TileMap performed as the best or among the best in almost all cases, as indicated by the bigger AUC (**Supplementary Fig. 13,14)**, higher numbers of true positives and lower numbers of false positives in the top 100 predictions (**Supplementary Table 11**). For example, when analyzing the Affymetrix data, TileMap outperformed MAT in four out of the five analyses. In the only case where MAT outperformed TileMap, all algorithms performed poorly, which is an indication of extremely low signal-to-noise ratio. In this case, the sequence based background correction provided by MAT may help improve the analysis by removing part of the systematic variation in the data. Examination of the E-O distance (**Supplementary Table 11**) suggests that the cutoff based on FDR≤10% performed reasonably well to balance the sensitivity and specificity when reporting the final peak lists.

7. Comparison of ChIP-seq analysis algorithms

We compared CisGenome's internal ChIP-seq peak caller with QuEST[30] and two other existing ChIP-seq peak detection algorithms ChIP-seq Peak Finder (CPF)[4] and GeneTrack[29]. Unlike CisGenome, CPF and GeneTrack do not provide statistical estimates of FDR, making it difficult to choose a cutoff. For example, when applied to the NRSF data, GeneTrack produced a list of 1,450,624 predictions. It is unlikely that all of them were true. The QuEST algorithm can provide an estimate of FDR. However, QuEST only produces FDR for two-sample analyses, and it does not support the one-sample analysis in which only the ChIP sample is available. Moreover, to estimate the FDR in the two-sample analysis, QuEST requires an extra negative control sample that has the same number of reads as the original ChIP sample. This extra negative control sample is required in addition to the original negative control sample. In other words, if the original experiment involves one million ChIP reads and one million control reads, then one needs to generate an additional one million control reads in order to be able to compute FDR. In the FDR analysis, the second control sample will serve as a mock-ChIP, and the original control sample will serve as the control. Since the NRSF data had about the same number of ChIP reads and control reads, we were not able to apply the QuEST FDR estimation procedure to the NRSF analysis. Compared to these methods, CisGenome is the only tool that can produce FDR estimates for both one-sample analyses and two-sample analyses. In the two-sample analyses, CisGenome does not pose any special requirement on control read numbers, since the ratio $p_0 = r_0/(1+r_0)$ will be used as the baseline to normalize the data and evaluate signal enrichment, and the ratio can be estimated from the data. When we were revising the paper, a new tool SISSRs[31] has become available. This tool uses a Poisson model to estimate FDR in the one-sample analysis. In the two-sample analysis, the control sample is used to control specificity and sensitivity of the predictions. The control of sensitivity is based on the empirical read distribution in the negative control sample, and the control of specificity is based on empirical p-values computed for fold changes between the ChIP'd and control sample. No FDR is provided in this context by SISSRs.

When the top 1,500 peaks of CisGenome, QuEST, GeneTrack and CPF were compared, predictions made by GeneTrack had a lower probability to cover the NRSF motif but a longer peak length compared to CisGenome predictions (**Supplementary Fig. 15a,b**). Peaks predicted by the CPF had a little higher probability to cover the NRSF motif, but this was because their average peak length was ≥10 times longer than CisGenome predictions. QuEST and CisGenome had about the same performance in terms of NRSF motif coverage, and they produced the highest NRSF motif occurrence rate (i.e., no. of motif sites per kb) in the predicted peaks (**Supplementary Fig. 15c**). We also compared the ChIP-seq analysis results with the ChIP-chip analysis results obtained using MAT and TileMap. All ChIP-seq analyses produced better results than ChIP-chip analyses.

8. More on CisGenome's FDR estimation for ChIP-seq analysis

In this study, we investigated the basic characteristics of the ChIP-seq data and found that the read sampling rate in the background non-binding regions is not a constant. Similar results were obtained when analyzing multiple transcription factors, suggesting that the observation is likely to be a general phenomenon. This basic data property has important implications in estimating the FDR. Previous studies[5,8,10,31] either use a Poisson model or use Monte Carlo simulations to describe what is expected under no binding. In the Monte Carlo simulations, reads are randomly re-distributed to the genome to characterize the expected noise level. Both approaches implicitly assume that the background read sampling rate is a constant, which is not true as suggested by the current study. FDR estimates based on constant rate assumptions therefore are very likely to underestimate the false positive rate. In order to better characterize the underlying variability of the data, we propose to use the negative binomial model to estimate the FDR in the one-sample analysis. **Figure 2b** suggests that even with the negative binomial model, one may still underestimate the tail probabilities in the negative control sample. However, compared to the Poisson model and the other constant rate based methods, FDR based on the negative binomial model represents a more reasonable error rate estimation for excluding background noise. It should be pointed out that our negative binomial method was designed for the one-sample analysis in which the negative control sample is not available. When the negative control sample is not available, it is difficult to perfectly characterize the underlying variability of the noise. To the best of our knowledge, the negative binomial model here represents the best solution that we currently have for handling this situation. On the other hand, as shown in **Supplementary Data 1**, the negative binomial method may seriously underestimate FDR when signals are not strong. Therefore, users still need to take cautious when using the model. When the negative control sample is available, FDR can be estimated without the negative binomial assumption. In this case, CisGenome uses a conditional binomial model to estimate the FDR.

Using negative binomial as the null model may produce substantially different error rate estimates from those based on constant rate assumptions. For example, using a Poisson assumption, Robertson et al.[5] estimated that there were 41,582 STAT1 binding regions in IFN-–stimulated HeLa S3 cells cells at a 0.1% FDR level. At the same FDR level, CisGenome's one-sample analysis only found 18,896 regions. Only at a 10% FDR level, CisGenome identified 48,523 regions.

9. Factors that may cause the observed differences between the NRSF ChIP-chip and ChIP-seq results

Our analysis showed that, compared to the NRSF ChIP-chip results, binding regions detected from the NRSF ChIP-seq data had a higher resolution (i.e. shorter peak length), higher signal-to-noise ratio (i.e. higher probability to cover the NRSF motif), and a more comprehensive genome coverage (i.e. array-specific regions are likely to represent noise, but a significant fraction of the ChIP-seq specific regions still contain NRSF motifs and are likely to be real signals).

There are two potential reasons for the higher resolution observed in the ChIP-seq data. First, the DNA fragment length in the ChIP-chip experiments are around 1kb. The long fragments are required to hybridize to multiple probes in order to generate reliable ChIP-chip signals. On the contrary, in the preparation of the Solexa library, a size selection step was introduced to select DNA fragments ~150-300bp long for sequencing[4]. The smaller pieces of DNA are expected to improve the resolution of binding site identification, besides increasing the colony size uniformity and the effective read number one can obtain. If one were to put the same ChIP sample with ~150-300bp long DNA fragments on tiling arrays, most of the signals would be buried among the noise, due to the small number of probes that can be covered by the ChIP fragments (data not shown). Second, in the ChIP-seq analysis, the offset between the 5' and 3' reads produced a natural source of information to determine a "confidence interval" of binding sites (i.e., the boundary refinement). Such information is not available in the ChIP-chip experiments.

To understand what contributed to the higher signal-to-noise ratio and a relatively more comprehensive coverage of the ChIP-seq analysis, we first analyzed the effects of read numbers. Our current analysis of NRSF involved ~2.2 million ChIP reads and ~2.8 million control reads. We did a simulation study in which 25%, 50% and 75% of the reads were randomly excluded from the analysis. With decreasing read number, the number of binding regions that can be detected by ChIP-seq also decreased (**Supplementary Fig. 11**). Using 25% of the original reads, the one-sample analysis only detected 1,973 binding regions at the 10% FDR level (compared to the 3,312 peaks detected using all reads). Among the 1,339 lost binding regions, 303 can be found by the TileMap ChIP-chip analysis, and 25.08% of them (76/303) contained ≥1 NRSF motif. This suggests that with fewer reads, ChIP-seq will start to lose sensitivity and miss true binding regions. Therefore, the relatively comprehensive coverage of the NRSF motif observed in the current study was at least partly due to the increased read number.

On the other hand, when only a fixed number of top peaks were compared, decreasing the read number did not decrease the percentage of ChIP-seq peaks that cover the NRSF motif, neither did it change the peak length and the motif occurrence rate (**Supplementary Fig. 16**). In other words, the effect of reducing the read number is to miss weak peaks, but it will not introduce additional noise to the peak predictions. As a result, the specificity of the predictions (i.e., # of false positives / [# of false positives + # of true positives]) will not decrease when we fix the number of total predicted peaks. This is likely due to the fact that when reads are generated, the stronger binding regions always have a higher probability to be sampled first. Thus the effect of increasing read number is to find weaker peaks and to increase the comprehensiveness of the prediction, and it will not affect the intrinsic signal-to-noise ratio that the technology can achieve. Importantly, even with the reduced number of reads, the ChIP-seq predictions were still more likely to cover the NRSF motif than the ChIP-chip binding regions. The ChIP-seq predictions made using reduced read number were still shorter than ChIP-chip binding regions and still had a higher NRSF occurrence rate (i.e. # of NRSF motif per kb). The performance of the ChIP-chip results is unlikely due to the specific algorithms used here, since in addition to TileMap, we also applied MAT to make predictions. At the 10% FDR level, MAT generated 7,054 NRSF binding regions (median length = 1161bp). **Supplementary Fig. 16** shows that MAT and TileMap performed similarly in the NRSF analysis, and both produced results worse than the ChIP-seq analyses.

We next asked whether the cross-hybridization in the arrays could potentially cause the lower specificity of the array predictions. In the array probe design, repeats were masked from the genome by the RepeatMasker. For this reason, we use large segmental duplications in the genome to study the potential effect of cross-hybridization. Among the 5,517 array-specific peaks, 547 (10.4%) had ≥50bp overlap with segmental duplications. As a comparison, only 2.9% (46/1,587) of peaks common to the ChIP-chip and ChIP-seq analyses (i.e., the intersection of ChIP-chip, one-sample and two-sample ChIP-seq analyses), and 2.2% (31/1,385) of the ChIP-seq specific peaks (i.e., the intersection of the one-sample and two-sample ChIP-seq analyses) contained ≥50bp overlap with segmental duplications. Therefore, the array-specific peaks were more likely to cover sequences that occur more than once in the genome. This suggests that part of the noise in the array was likely due to cross-hybridization issues.

We then explored whether the array design may affect the comprehensiveness of peak detection results. Among the 1,385 ChIP-seq specific peaks, 153 (11.1%) were not tiled in the arrays. 143 (93.5%) out of the 153 contained repeat elements that occupied more than 50% of the peak length. These regions were likely excluded from the array design due to repeat masking. On the other hand, 88.9% of the ChIP-seq specific peaks were covered by the array design. We did not find a clear reason why they were not detected as peaks by ChIP-chip.

Finally, another major factor that may affect the performance of the two technologies is the sample preparation. Although preparation of the ChIP sample in the two experiments followed the same protocol, a size selection step was introduced in preparing the Solexa library before sequencing. This step is unique to ChIP-seq and was not applied to ChIP-chip. We speculate that it may also affect the signal-to-noise ratio of the final sample to be sequenced. In order to test this, a systematic experimental study that compares each individual sample preparation steps is needed in future. Such a study is already beyond the scope of our current paper which mainly aimed at addressing computational challenges of data analysis.

To summarize, the observed differences between the NRSF ChIP-chip and ChIP-seq results are likely to be contributed by multiple factors, including but not limited to increased depth of sequencing, cross-hybridization in the arrays, array design as well as sample preparation procedures.

10. Potential reasons for the differences between the one-sample and two-sample ChIP-seq analysis results

In order to see what contributed to the observed differences in the one-sample and two-sample NRSF ChIP-seq analysis results, we focused on the high quality binding regions identified by these analyses. Among the high-quality binding regions detected by these two analyses, 69 were specific to the one-sample analysis, and there was no region specific to the two-sample analysis (**Fig. 3b**). Among the 69 one-sample analysis specific regions, 61 (88.4%) contained repeat elements that that occupied more than 50% of the peak length. As a comparison, only 21.3% (88/414) of the high quality regions common to the two analyses contained the same level of repeats. This suggests that the one-sample analysis is more likely to pick up repeat-rich regions. Although we only used uniquely mapped sequence reads in the analysis, there could be a chance to misalign reads from the repeat regions to a wrong location due to SNPs, sequencing errors, or errors in the reference genome assembly. We speculate that this may correlate with the observation here although a further examination is needed in future to fully address this issue. In the two-sample analysis, these repeat-rich regions were eliminated perhaps due to the same bias exists in the negative control sample.

11. Are the observations in this study a general phenomenon?

In the NRSF analyses, ChIP-seq results are more sensitive and more specific than ChIP-chip results. In general, whether this conclusion will continue to hold true for other transcription factors is an important question whose definitive answer must await the availability of similar data for a larger number of different transcription factors.

However, certain aspects of the results in this study might be general. In particular, the fact that the background read sampling rate across the genome is not uniform may hold true for other ChIP-seq studies. In fact, in addition to the NRSF analysis, the initial analysis of a number of other transcription factors showed similar results.

Secondly, the fact that the ChIP-seq can provide a higher resolution (i.e., shorter peak length) than ChIP-chip in the determination of transcription factor binding sites is unlikely to be specific to this study. The high resolution is partly due to the size selection in the ChIP-seq protocol, and partly due to the additional information provided by the read directionalities. Both are applicable to future ChIP-seq studies. On the other hand, the size selection may not be applicable to ChIP-chip where long DNA fragments are needed in order to hybridize to multiple probes to generate reliable signals. Thus the resolution that the current ChIP-chip can achieve is limited intrinsically.

Finally, many discussions in **Supplementary Data 9** may hold true for future studies of other transcription factors, such as the potential effect of increasing read numbers and the potential effect of cross-hybridization issues.

12. How do the observations in the current study relate to previous observations?

Our analysis suggests that the NRSF ChIP-seq analysis performed better than NRSF ChIP-chip. In a previous study, Euskirchen et al.[46] compared the ChIP-chip and ChIP-PET, and they found that the array and sequencing based studies had comparable performance and are complementary to each other. The inconsistency in these two studies is likely due to the differences between the objects that are compared (e.g., ChIP-seq and ChIP-PET are two different technologies with different sample preparation protocols). With the availability of the recently developed massively parallel sequencing platform, our current ChIP-seq study involved 2.2 million uniquely mapped ChIP reads, representing a increased depth of sequencing compared to the ChIP-PET data in ref.

46 which contained ~726k paired end tags representing ~328k distinct ChIP DNA fragments. The ChIP DNA fragments in the ChIP-PET data ranged from 0.1-6k, whereas the DNA fragment length of the ChIP-seq data in our current study is ~150-300bp due to the size selection. Furthermore, in our current analysis, we tried to use the read directionalities to refine peak boundaries, and this information was not explicitly used in ref. 46. The ChIP-chip data in our current study were produced using Affymetrix tiling arrays with 25 bp short oligonucleotides as probes, and the ChIP-chip data in ref. 46 were produced with NimbleGen 50 bp oligonucleotide arrays. This is another potential factor that may cause differences of the results of the two studies.

In another study by Robertson et al.[5], STAT1 ChIP-chip and ChIP-seq were compared. The authors observed that there were much more ChIP-seq peaks at the 0.1% FDR level than ChIP-chip peaks at the 1% FDR level on the same chromosomes. However, this conclusion was based on peaks determined using a Poisson background model, therefore the observation may be partly due to an underestimated ChIP-seq FDR. The study did not compare the one-sample analyses with two-sample analyses, and it did not compare the gain of using read directionalities. Both issues are handled by our current study.

13. Discussion on different types of negative controls

The negative control sample used in the current ChIP-seq study is a noIP control (i.e., crosslink was reversed but the immnunoprecipitation was bypassed). We do not use mock IP in the ChIP-seq experiments, since there is so little DNA when we do a mock IP that the sequence reads are extremely biased to only a few fragments. Sometimes, ChIP-seq experiments can be performed using different cell types, and the cell type that lack the ChIP target may serve as the negative control. For example, in ref. 5, the authors studied IFN-γ-stimulated and unstimulated HeLa S3 cells. The unstimutated cells served as a control for detecting STAT1 targets responsive to the stimulation. However, when this type of control is used, it is typically used to detect differences of the protein-DNA binding between cell types. It is likely that the cell type “lacking” the ChIP targets still contain some base line level binding (e.g., STAT1 binds to a number of targets even in the unstimulated cells). If this is the case and if the purpose of the study is to find all binding regions as well as differential binding, then one can perform the one-sample analyses to identify all binding regions (when the noIP control is not available), and perform the two-sample analyses to identify differential protein-DNA binding.

SUPPLEMENTARY REFERENCES

50. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

51. Li, W., Meyer, C.A. & Liu, X.S. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **21**(Suppl. 1), i274-i282 (2005).

52. Gottardo, R., Li, W., Johnson, W.E. & Liu, X.S. A flexible and powerful bayesian hierarchical model for ChIP-Chip experiments. *Biometrics* **64**, 468-478 (2008).

53. Toedling, J. *et al.* Ringo -- an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* **8**, 221 (2007).

54. Lawrence, C.E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208-214 (1993).

55. Roth, F.P., Hughes, J.D., Estep, P.W. & Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939-945 (1998).

56. Liu, X.S., Brutlag, D.L. & Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* 127-138 (2001).

57. Thompson, W., Palumbo, M.J., Wasserman, W.W., Liu, J.S. & Lawrence, C.E. Decoding human regulatory circuits. *Genome Res.* **14**, 1967-1974 (2004).

58. Gupta, M. & Liu, J. S. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* **102**, 7079-7084 (2005).

59. Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W. & Lawrence, C.E. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26**, 225-228 (2000).

60. Wang, T. & Stormo, G.D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**, 2369-2380 (2003).

61. Sinha, S., Blanchette, M. & Tompa, M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5**, 170 (2004).

62. Liu, Y., Liu, X.S., Wei, L., Altman, R.B. & Batzoglou, S. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.* **14**, 451-458 (2004).

63. Siddharthan, R., Siggia, E.D. & van Nimwegen, E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* **1**:e67 (2005).

64. Li, X. & Wong, W. H. Sampling motifs on phylogenetic trees. *Proc. Natl. Acad. Sci. USA* **102**, 9481-9486 (2005).

65. Zhou, Q. & Wong, W.H. Coupling hidden Markov models for the discovery of cis-regulatory modules in multiple species. *Ann. Appl. Stat.* **1**, 36-65 (2007).

66. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-714 (2008).

67. Smith, A.D., Xuan, Z. & Zhang, M.Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, **9**, 128 (2008).

68. Lin, H, Zhang, Z., Zhang, M.Q., Ma, B. & Li, M. ZOOM! Zillions of oligos mapped. *Bioinformatics* doi:10.1093 (2008).

69. Jensen, S.T., Liu, X.S., Zhou, Q. & Liu, J.S. Computational discovery of gene regulatory binding motifs: a Bayesian perspective. *Statist. Sci.* **19**, 188-204 (2004).

70. Tompa, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23**, 137-144 (2005).

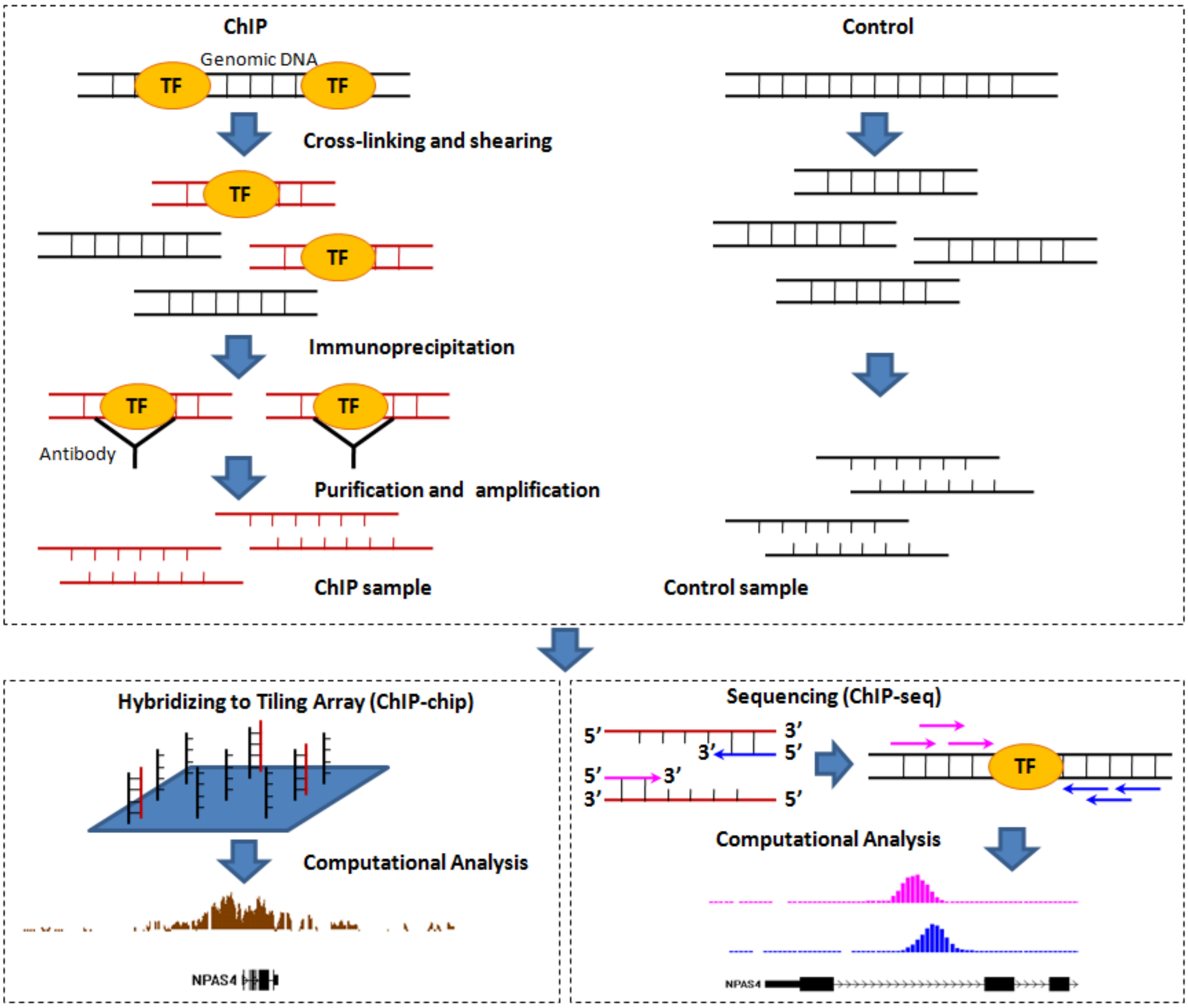
71. Ji, H.K. & Wong, W.H. Computational biology: towards deciphering gene regulatory information in mammalian genomes. *Biometrics*, **62**, 645-663 (2006).

72. Manku, G.S., Jain A. & Sarma, A.D. Detecting near-duplicates for web crawling. *WWW07* 141-150 (2007).

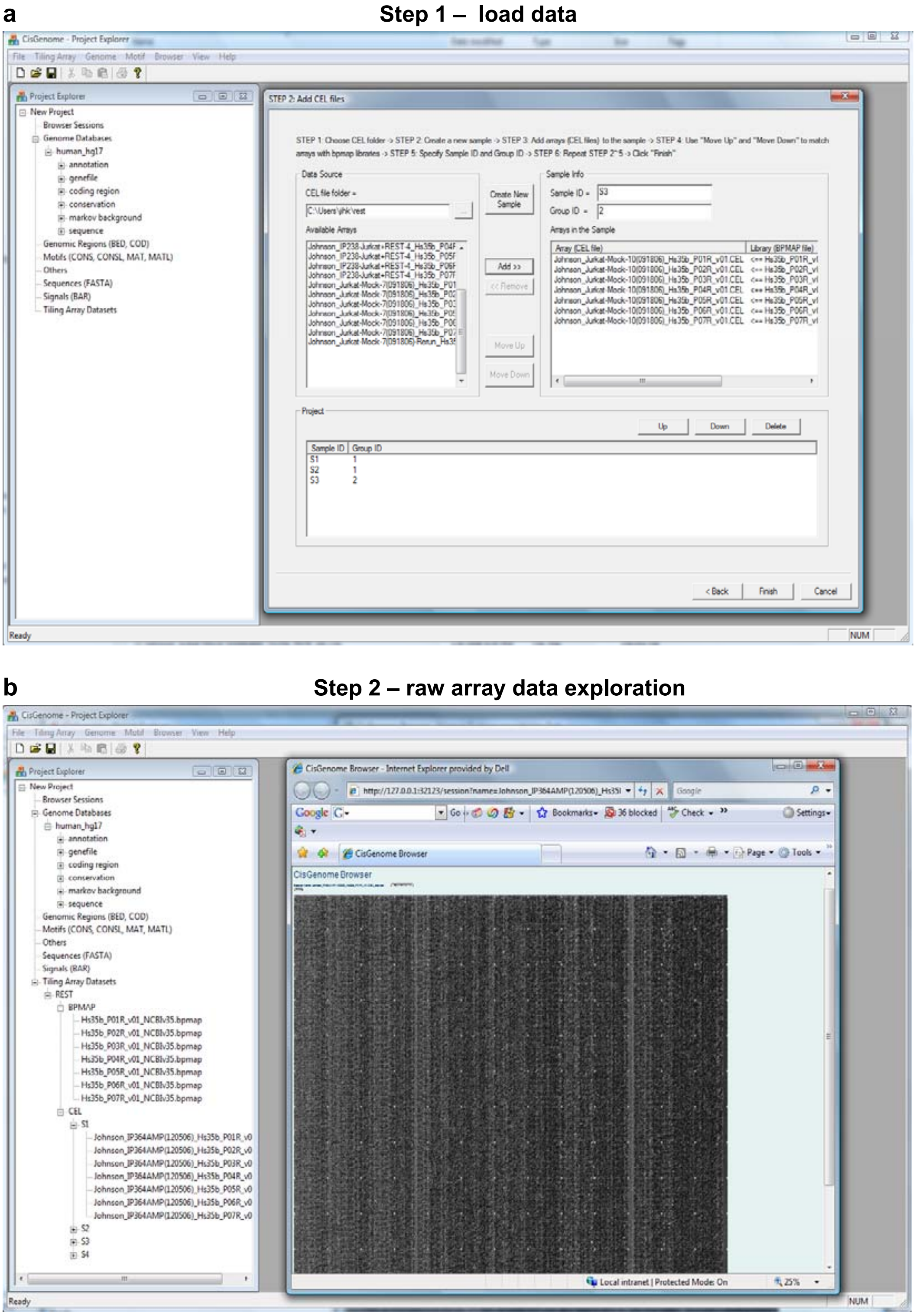
73. Loh, Y.H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431-440 (2006).

74. Vokes, S.A. *et al.* Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning. *Development* **134**, 1977-1989 (2007).

75. Lucas, I. *et al.* High-throughput mapping of origins of replication in human cells. *EMBO Rep.* **8,** 770–777 (2007).



Supplementary Figure 1 ChIP-chip and ChIP-seq. Both technologies start by preparing a ChIP sample enriched in protein bound DNAs. The ChIP sample will either be hybridized to microarrays that contain probes interrogating the whole genome (ChIP-chip), or be sequenced from both ends to generate millions of short reads using ultra high throughput sequencing (ChIP-seq). To eliminate unknown bias that may arise during sample preparation, hybridization or sequencing procedures, people often also include one or more control samples (e.g., Input or mock IP) in the experiments.



STEP 2: Add CEL files

STEP 1: Choose CEL folder -> STEP 2: Create a new sample -> STEP 3: Add arrays (CEL files) to the sample -> STEP 4: Use "Move Up" and "Move Down" to match arrays with bpmap libraries -> STEP 5: Specify Sample ID and Group ID -> STEP 6: Repeat STEP 2~5 -> Click "Finish"

Data Source

CEL file folder = C:\Users\jhk\rest

Available Arrays

```
Johnson_IP238-Jurkat+REST-4_Hs35b_P04F
Johnson_IP238-Jurkat+REST-4_Hs35b_P05F
Johnson_IP238-Jurkat+REST-4_Hs35b_P06F
Johnson_IP238-Jurkat+REST-4_Hs35b_P07F
Johnson_Jurkat-Mock-7(091806)_Hs35b_P01
Johnson_Jurkat-Mock-7(091806)_Hs35b_P02
Johnson_Jurkat-Mock-7(091806)_Hs35b_P03
Johnson_Jurkat-Mock-7(091806)_Hs35b_P05
Johnson_Jurkat-Mock-7(091806)_Hs35b_P06
Johnson_Jurkat-Mock-7(091806)_Hs35b_P07
Johnson_Jurkat-Mock-7(091806)-Rerun_Hs35
```

Sample Info

Sample ID = S3

Group ID = 2

Arrays in the Sample

Array (CEL file)	Library (BPMAP file)
Johnson_Jurkat-Mock-10(091806)_Hs35b_P01R_v01.CEL	<== Hs35b_P01R_vl
Johnson_Jurkat-Mock-10(091806)_Hs35b_P02R_v01.CEL	<== Hs35b_P02R_vl
Johnson_Jurkat-Mock-10(091806)_Hs35b_P03R_v01.CEL	<== Hs35b_P03R_vl
Johnson_Jurkat-Mock-10(091806)_Hs35b_P04R_v01.CEL	<== Hs35b_P04R_vl
Johnson_Jurkat-Mock-10(091806)_Hs35b_P05R_v01.CEL	<== Hs35b_P05R_vl
Johnson_Jurkat-Mock-10(091806)_Hs35b_P06R_v01.CEL	<== Hs35b_P06R_vl
Johnson_Jurkat-Mock-10(091806)_Hs35b_P07R_v01.CEL	<== Hs35b_P07R_vl

Project

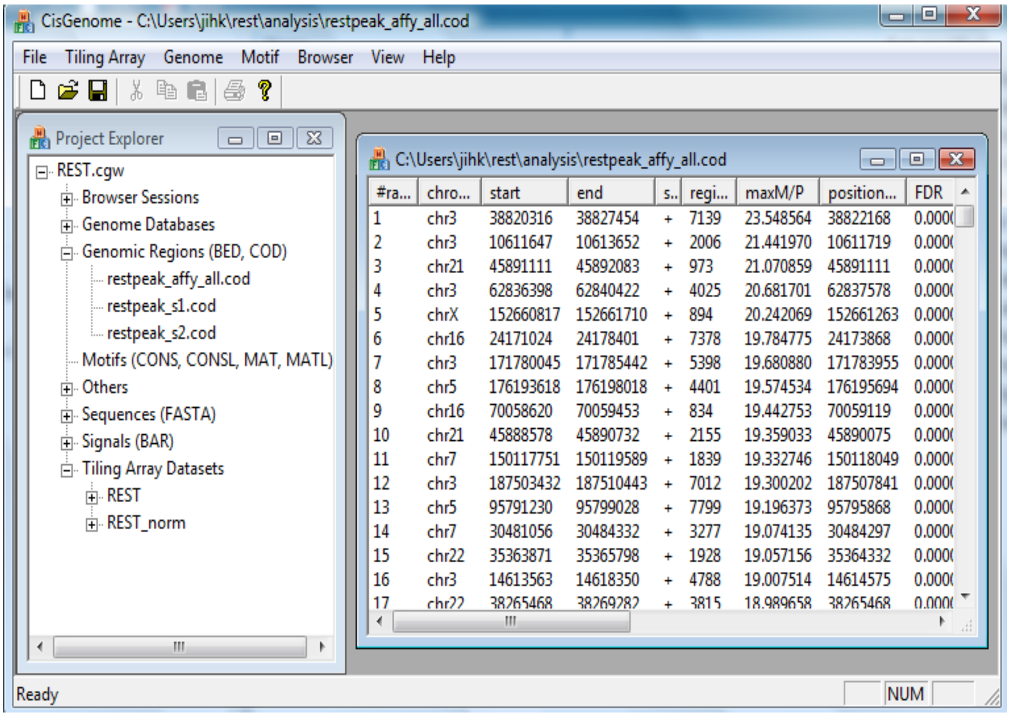
Sample ID	Group ID
S1	1
S2	1
S3	2

Up
Down
Delete
< Back
Finish
Cancel
b
Step 2 – raw array data exploration
CisGenome - Project Explorer
File Tiling Array Genome Motif Browser View Help
Project Explorer
New Project
Browser Sessions
Genome Databases
human_hg17
annotation
genefile
coding region
conservation
markov background
sequence
Genomic Regions (BED, COD)
Motifs (CONS, CONSL, MAT, MATL)
Others
Sequences (FASTA)
Signals (BAR)
Tiling Array Datasets
REST
BPMAP
Hs35b_P01R_v01_NCBIv35.bpmap
Hs35b_P02R_v01_NCBIv35.bpmap
Hs35b_P03R_v01_NCBIv35.bpmap
Hs35b_P04R_v01_NCBIv35.bpmap
Hs35b_P05R_v01_NCBIv35.bpmap
Hs35b_P06R_v01_NCBIv35.bpmap
Hs35b_P07R_v01_NCBIv35.bpmap
CEL
S1
Johnson_IP364AMP(120506)_Hs35b_P01R_v0
Johnson_IP364AMP(120506)_Hs35b_P02R_v0
Johnson_IP364AMP(120506)_Hs35b_P03R_v0
Johnson_IP364AMP(120506)_Hs35b_P04R_v0
Johnson_IP364AMP(120506)_Hs35b_P05R_v0
Johnson_IP364AMP(120506)_Hs35b_P06R_v0
Johnson_IP364AMP(120506)_Hs35b_P07R_v0
S2
S3
S4
CisGenome Browser - Internet Explorer provided by Dell
http://127.0.0.1:32123/session?name=Johnson_IP364AMP(120506)_Hs35l
Google
Go
Bookmarks
36 blocked
Check
Settings
CisGenome Browser
Page
Tools
CisGenome Browser
Local intranet | Protected Mode: On
25%
Ready
NUM

Supplementary Figure 2 ChIP-chip analysis using CisGenome.

c

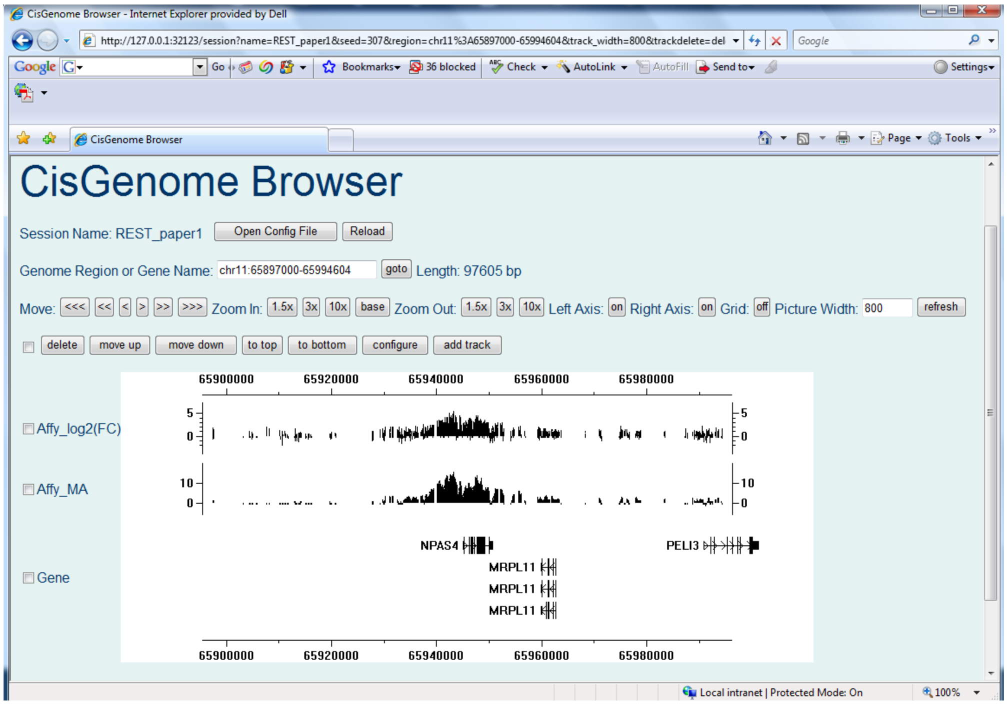
Step 3 – normalization & peak detection



#ra...	chro...	start	end	s..	regi...	maxM/P	position...	FDR
1	chr3	38820316	38827454	+	7139	23.548564	38822168	0.000(
2	chr3	10611647	10613652	+	2006	21.441970	10611719	0.000(
3	chr21	45891111	45892083	+	973	21.070859	45891111	0.000(
4	chr3	62836398	62840422	+	4025	20.681701	62837578	0.000(
5	chrX	152660817	152661710	+	894	20.242069	152661263	0.000(
6	chr16	24171024	24178401	+	7378	19.784775	24173868	0.000(
7	chr3	171780045	171785442	+	5398	19.680880	171783955	0.000(
8	chr5	176193618	176198018	+	4401	19.574534	176195694	0.000(
9	chr16	70058620	70059453	+	834	19.442753	70059119	0.000(
10	chr21	45888578	45890732	+	2155	19.359033	45890075	0.000(
11	chr7	150117751	150119589	+	1839	19.332746	150118049	0.000(
12	chr3	187503432	187510443	+	7012	19.300202	187507841	0.000(
13	chr5	95791230	95799028	+	7799	19.196373	95795868	0.000(
14	chr7	30481056	30484332	+	3277	19.074135	30484297	0.000(
15	chr22	35363871	35365798	+	1928	19.057156	35364332	0.000(
16	chr3	14613563	14618350	+	4788	19.007514	14614575	0.000(
17	chr22	38265468	38269282	+	3815	18.989658	38265468	0.000(

d

Step 4 – peak & gene structure visualization


CisGenome Browser - Internet Explorer provided by Dell
http://127.0.0.1:32123/session?name=REST_paper1&seed=307&region=chr11%3A65897000-65994604&track_width=800&trackdelete=del
Google
Go
Bookmarks
36 blocked
Check
AutoLink
AutoFill
Send to
Settings
CisGenome Browser
Page
Tools
65900000
65920000
65940000
65960000
65980000
Affy_log2(FC)
5
0
Affy_MA
10
0
Gene
NPAS4
PELI3
MRPL11
MRPL11
MRPL11
Local intranet | Protected Mode: On
100%

CisGenome Browser

Session Name: REST_paper1 Open Config File Reload

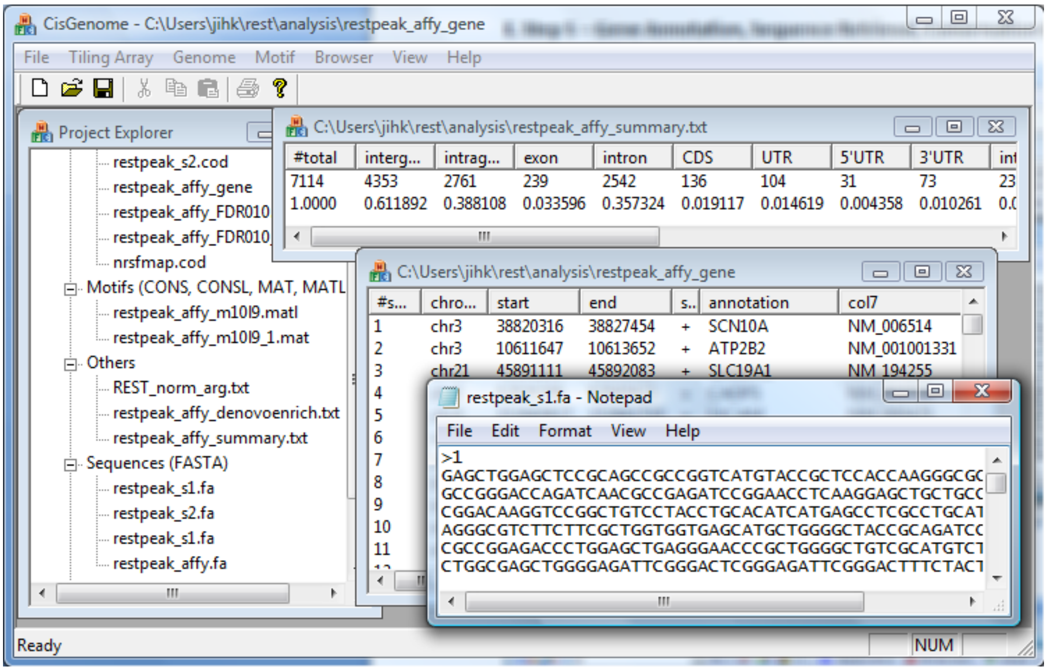
Genome Region or Gene Name: chr11:65897000-65994604 goto Length: 97605 bp

Move: <<< << < > >> >>> Zoom In: 1.5x 3x 10x base Zoom Out: 1.5x 3x 10x Left Axis: on Right Axis: on Grid: off Picture Width: 800 refresh

□ delete move up move down to top to bottom configure add track

Supplementary Figure 2 ChIP-chip analysis using CisGenome (cont.).

e **Step 5 – gene annotation, sequence retrieval, conservation & location analysis**

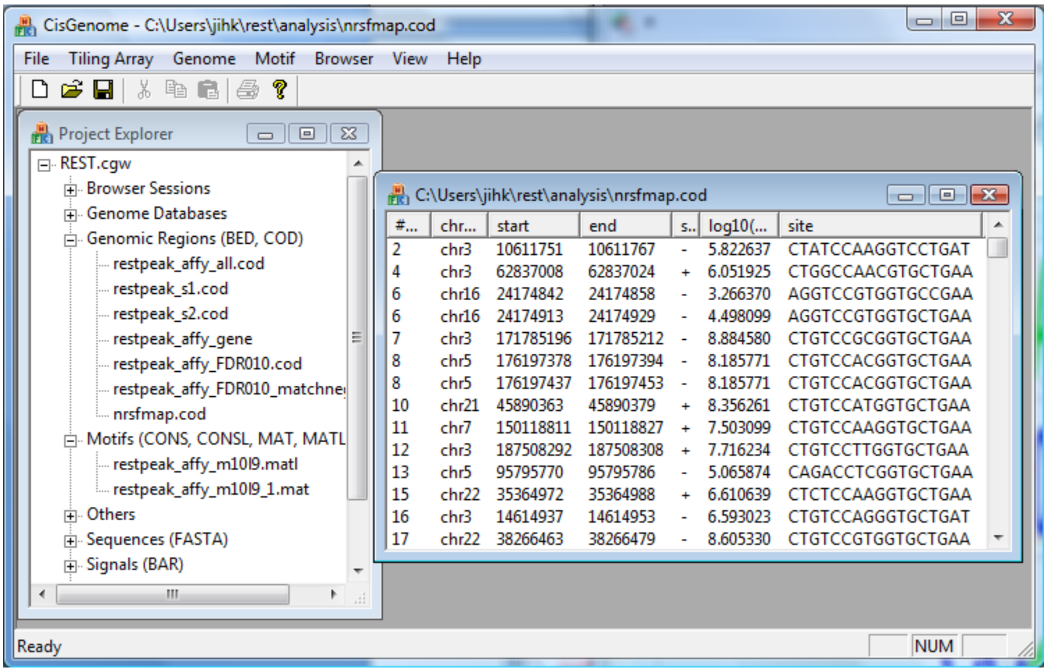

CisGenome - C:\Users\jihk\rest\analysis\restpeak_affy_gene
File Tiling Array Genome Motif Browser View Help
Project Explorer
restpeak_s2.cod
restpeak_affy_gene
restpeak_affy_FDR010
restpeak_affy_FDR010
nrsfmap.cod
Motifs (CONS, CONSL, MAT, MATL
restpeak_affy_m10l9.matl
restpeak_affy_m10l9_1.mat
Others
REST_norm_arg.txt
restpeak_affy_denovoenrich.txt
restpeak_affy_summary.txt
Sequences (FASTA)
restpeak_s1.fa
restpeak_s2.fa
restpeak_s1.fa
restpeak_affy.fa
C:\Users\jihk\rest\analysis\restpeak_affy_summary.txt
C:\Users\jihk\rest\analysis\restpeak_affy_gene
restpeak_s1.fa - Notepad
File Edit Format View Help
Ready
NUM

#total	interg...	intrag...	exon	intron	CDS	UTR	5'UTR	3'UTR	in
7114	4353	2761	239	2542	136	104	31	73	23
1.0000	0.611892	0.388108	0.033596	0.357324	0.019117	0.014619	0.004358	0.010261	0.0

#s...	chro...	start	end	s..	annotation	col7
1	chr3	38820316	38827454	+	SCN10A	NM_006514
2	chr3	10611647	10613652	+	ATP2B2	NM_001001331
3	chr21	45891111	45892083	+	SLC19A1	NM 194255
4						
5						
6						
7						
8						
9						
10						
11						

```
>1
GAGCTGGAGCTCCGCAGCCGCCGGTCATGTACCGCTCCACCAAGGGCGC
GCCGGGACCAGATCAACGCCGAGATCCGGAACCTCAAGGAGCTGCTGCC
CGGACAAGGTCCGGCTGTCCTACCTGCACATCATGAGCCTCGCCTGCAT
AGGGCGTCTTCTTCGCTGGTGGTGAGCATGCTGGGGCTACCGCAGATCC
CGCCGGAGACCCTGGAGCTGAGGGAACCCGCTGGGGCTGTCGCATGTCT
CTGGCGAGCTGGGGAGATTCGGGACTCGGGAGATTCGGGACTTTCTACT
```

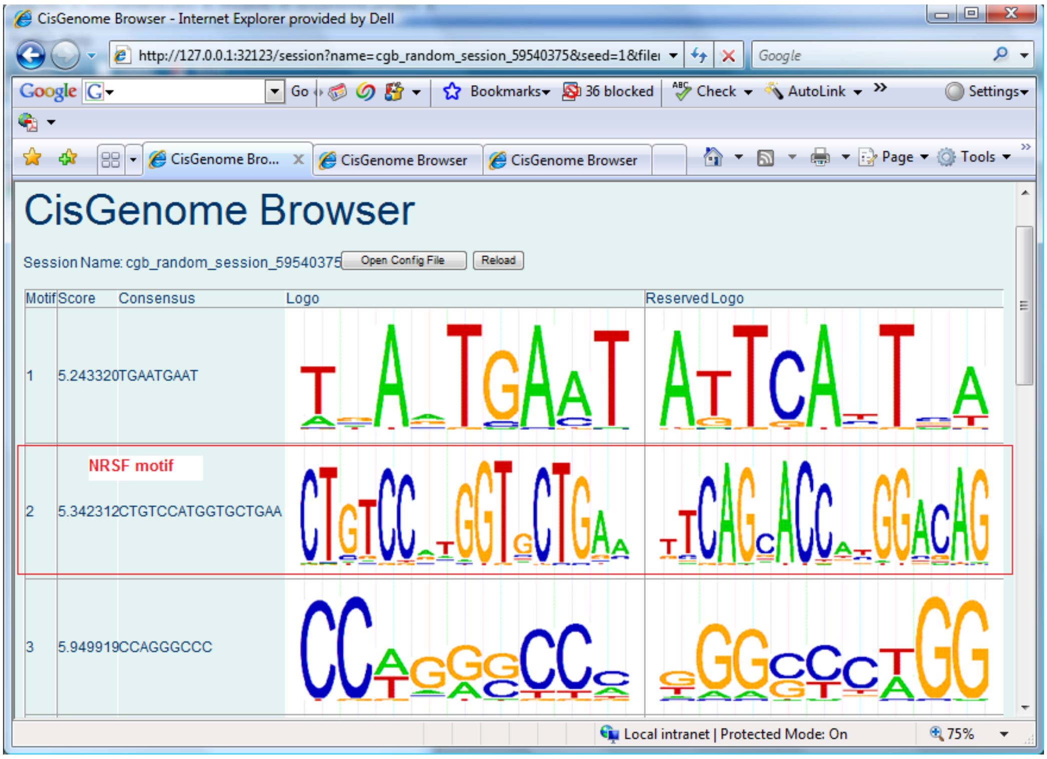
f **Step 6 – known motif mapping**


CisGenome - C:\Users\jihk\rest\analysis\nrsfmap.cod
File Tiling Array Genome Motif Browser View Help
Project Explorer
REST.cgw
Browser Sessions
Genome Databases
Genomic Regions (BED, COD)
restpeak_affy_all.cod
restpeak_s1.cod
restpeak_s2.cod
restpeak_affy_gene
restpeak_affy_FDR010.cod
restpeak_affy_FDR010_matchne
nrsfmap.cod
Motifs (CONS, CONSL, MAT, MATL
restpeak_affy_m10l9.matl
restpeak_affy_m10l9_1.mat
Others
Sequences (FASTA)
Signals (BAR)
C:\Users\jihk\rest\analysis\nrsfmap.cod
Ready
NUM

#...	chr...	start	end	s..	log10(...	site
2	chr3	10611751	10611767	-	5.822637	CTATCCAAGGTCCTGAT
4	chr3	62837008	62837024	+	6.051925	CTGGCCAACGTGCTGAA
6	chr16	24174842	24174858	-	3.266370	AGGTCCGTGGTGCCGAA
6	chr16	24174913	24174929	-	4.498099	AGGTCCGTGGTGCTGAA
7	chr3	171785196	171785212	-	8.884580	CTGTCCGCGGTGCTGAA
8	chr5	176197378	176197394	-	8.185771	CTGTCCACGGTGCTGAA
8	chr5	176197437	176197453	-	8.185771	CTGTCCACGGTGCTGAA
10	chr21	45890363	45890379	+	8.356261	CTGTCCATGGTGCTGAA
11	chr7	150118811	150118827	+	7.503099	CTGTCCAAGGTGCTGAA
12	chr3	187508292	187508308	+	7.716234	CTGTCCTTGGTGCTGAA
13	chr5	95795770	95795786	-	5.065874	CAGACCTCGGTGCTGAA
15	chr22	35364972	35364988	+	6.610639	CTCTCCAAGGTGCTGAA
16	chr3	14614937	14614953	-	6.593023	CTGTCCAGGGTGCTGAT
17	chr22	38266463	38266479	-	8.605330	CTGTCCGTGGTGCTGAA

Supplementary Figure 2 ChIP-chip analysis using CisGenome (cont.).

g **Step 7 – *de novo* motif & module discovery**

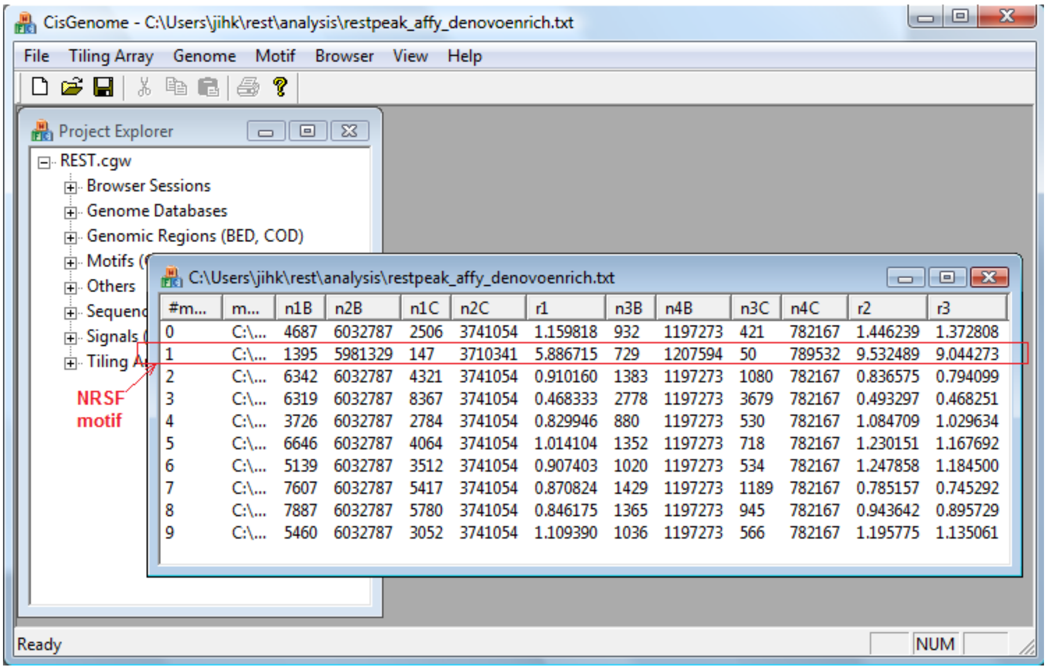

CisGenome Browser - Internet Explorer provided by Dell
http://127.0.0.1:32123/session?name=cgb_random_session_59540375&seed=1&filei
Google
Go
Bookmarks
36 blocked
Check
AutoLink
Settings
CisGenome Bro...
CisGenome Browser
CisGenome Browser
Page
Tools
Open Config File
Reload
NRSF motif
Local intranet | Protected Mode: On
75%

CisGenome Browser

Session Name: cgb_random_session_59540375

Motif	Score	Consensus	Logo	Reserved Logo
1	5.243320	TGAATGAAT		
2	5.342312	CTGTCCATGGTGCTGAA		
3	5.949919	CCAGGGCCC		

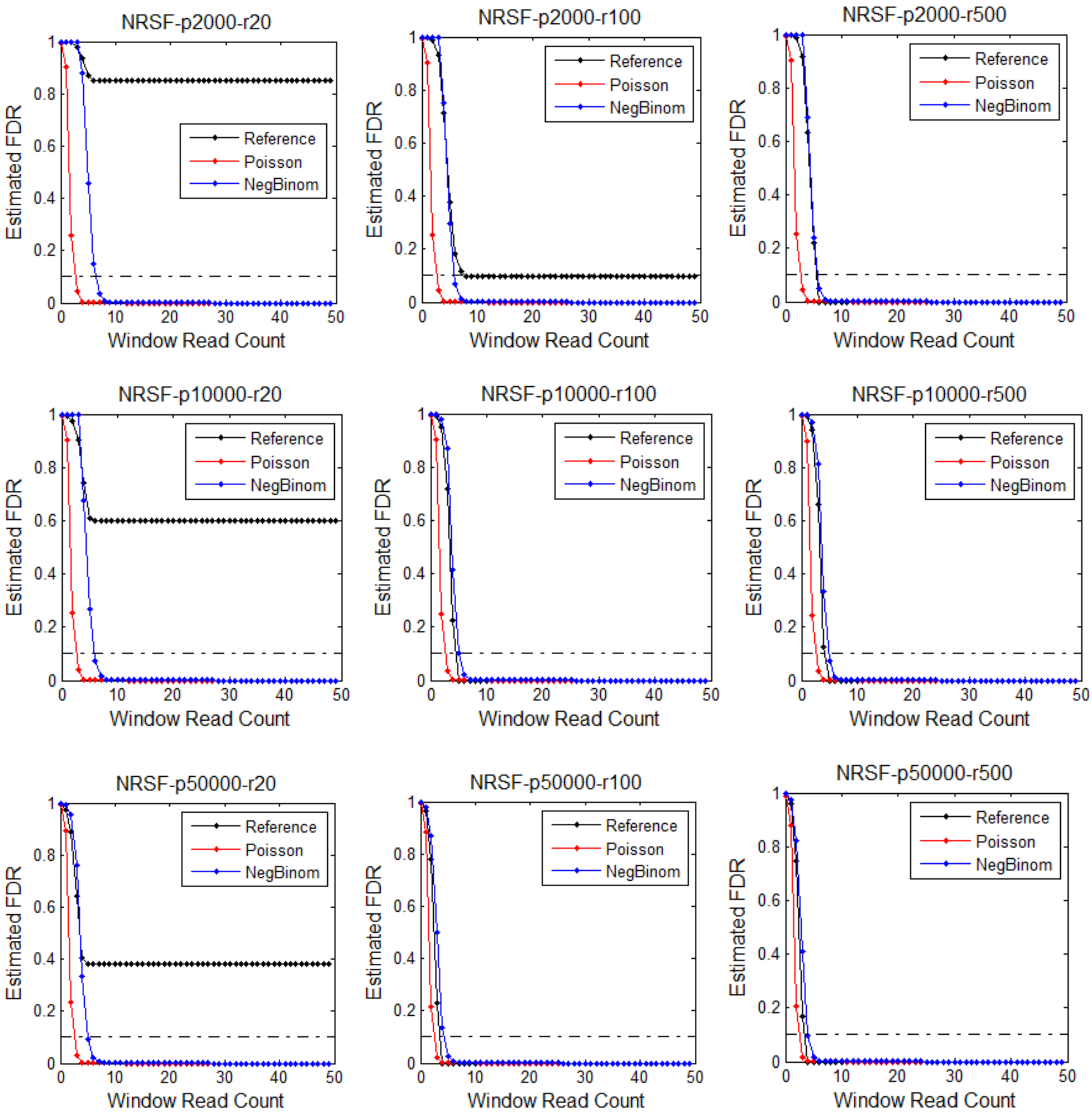
h **Step 8 – matched genomic control selection & motif enrichment analysis**



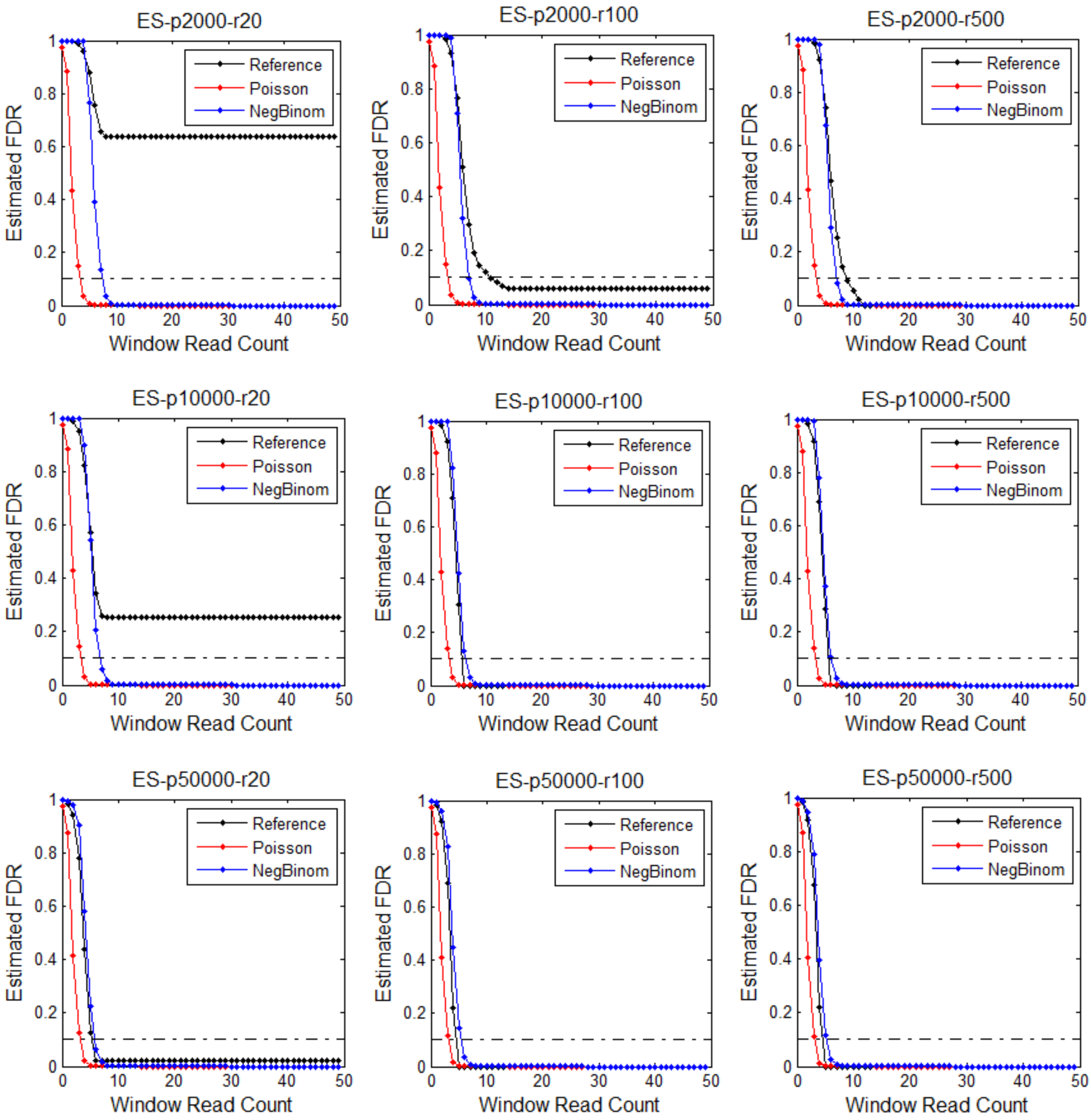
C:\Users\jihk\rest\analysis\restpeak_affy_denovoenrich.txt

#m...	m...	n1B	n2B	n1C	n2C	r1	n3B	n4B	n3C	n4C	r2	r3
0	C:\...	4687	6032787	2506	3741054	1.159818	932	1197273	421	782167	1.446239	1.372808
1	C:\...	1395	5981329	147	3710341	5.886715	729	1207594	50	789532	9.532489	9.044273
2	C:\...	6342	6032787	4321	3741054	0.910160	1383	1197273	1080	782167	0.836575	0.794099
3	C:\...	6319	6032787	8367	3741054	0.468333	2778	1197273	3679	782167	0.493297	0.468251
4	C:\...	3726	6032787	2784	3741054	0.829946	880	1197273	530	782167	1.084709	1.029634
5	C:\...	6646	6032787	4064	3741054	1.014104	1352	1197273	718	782167	1.230151	1.167692
6	C:\...	5139	6032787	3512	3741054	0.907403	1020	1197273	534	782167	1.247858	1.184500
7	C:\...	7607	6032787	5417	3741054	0.870824	1429	1197273	1189	782167	0.785157	0.745292
8	C:\...	7887	6032787	5780	3741054	0.846175	1365	1197273	945	782167	0.943642	0.895729
9	C:\...	5460	6032787	3052	3741054	1.109390	1036	1197273	566	782167	1.195775	1.135061

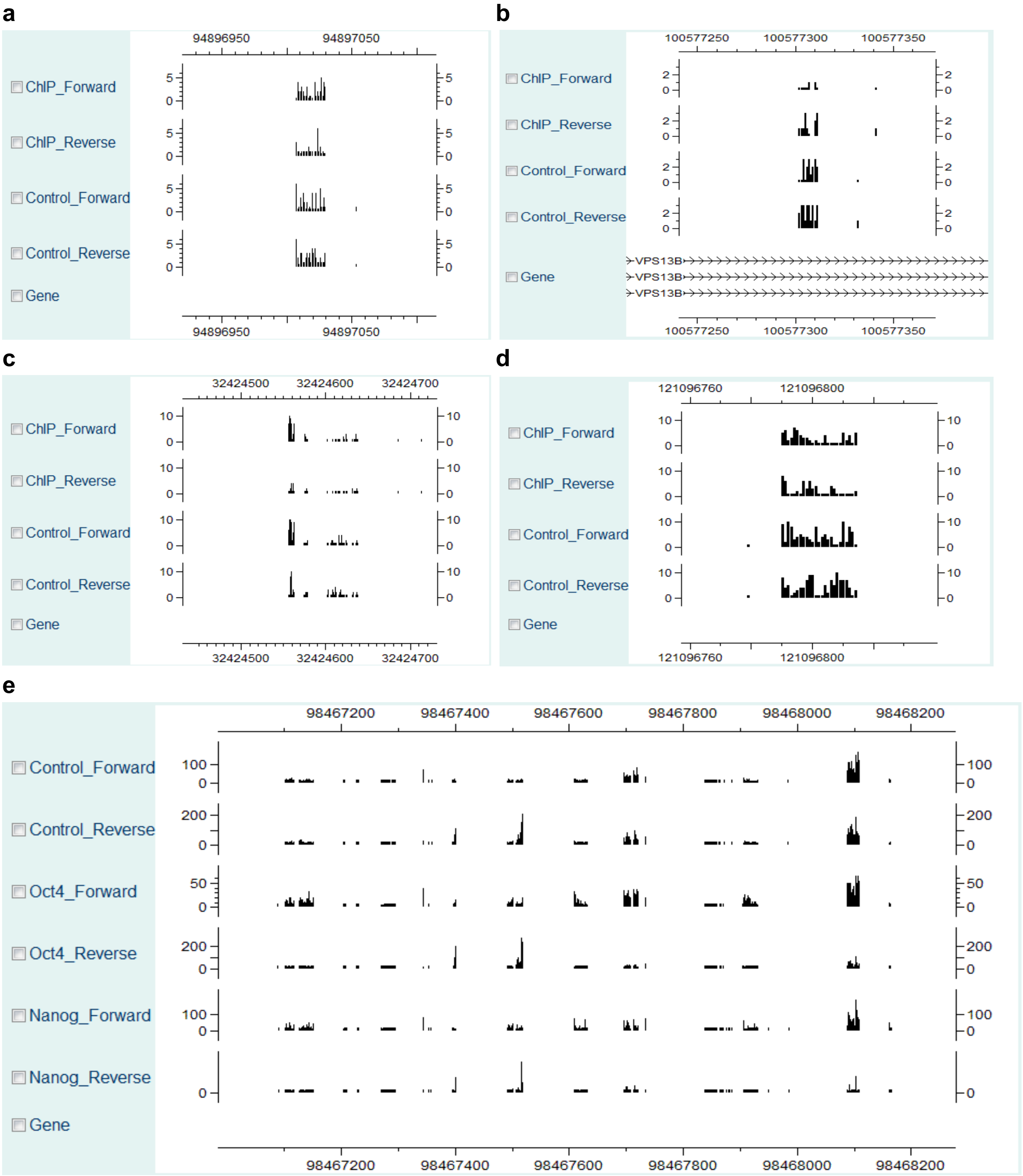
Supplementary Figure 2 ChIP-chip analysis using CisGenome (cont.).



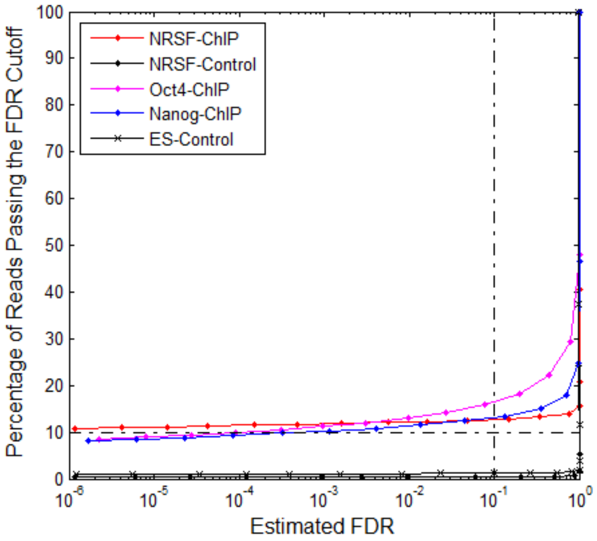
Supplementary Figure 3 Evaluation of the one-sample FDR estimates in the simulated spike-in experiments using NRSF negative control as background. ‘p’: peak number; ‘r’: average IP/control read enrichment ratio.



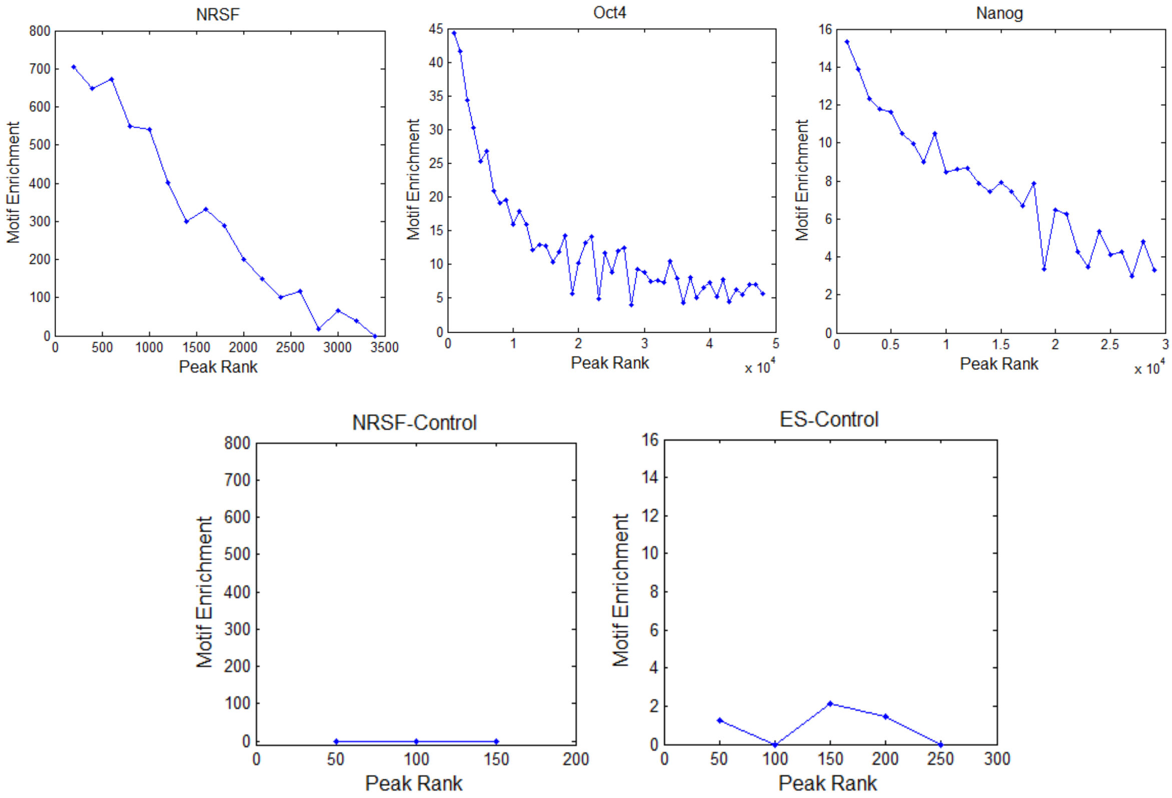
Supplementary Figure 4 Evaluation of the one-sample FDR estimates in the simulated spike-in experiments using ES negative control as background. ‘p’: peak number; ‘r’: average IP/control read enrichment ratio.



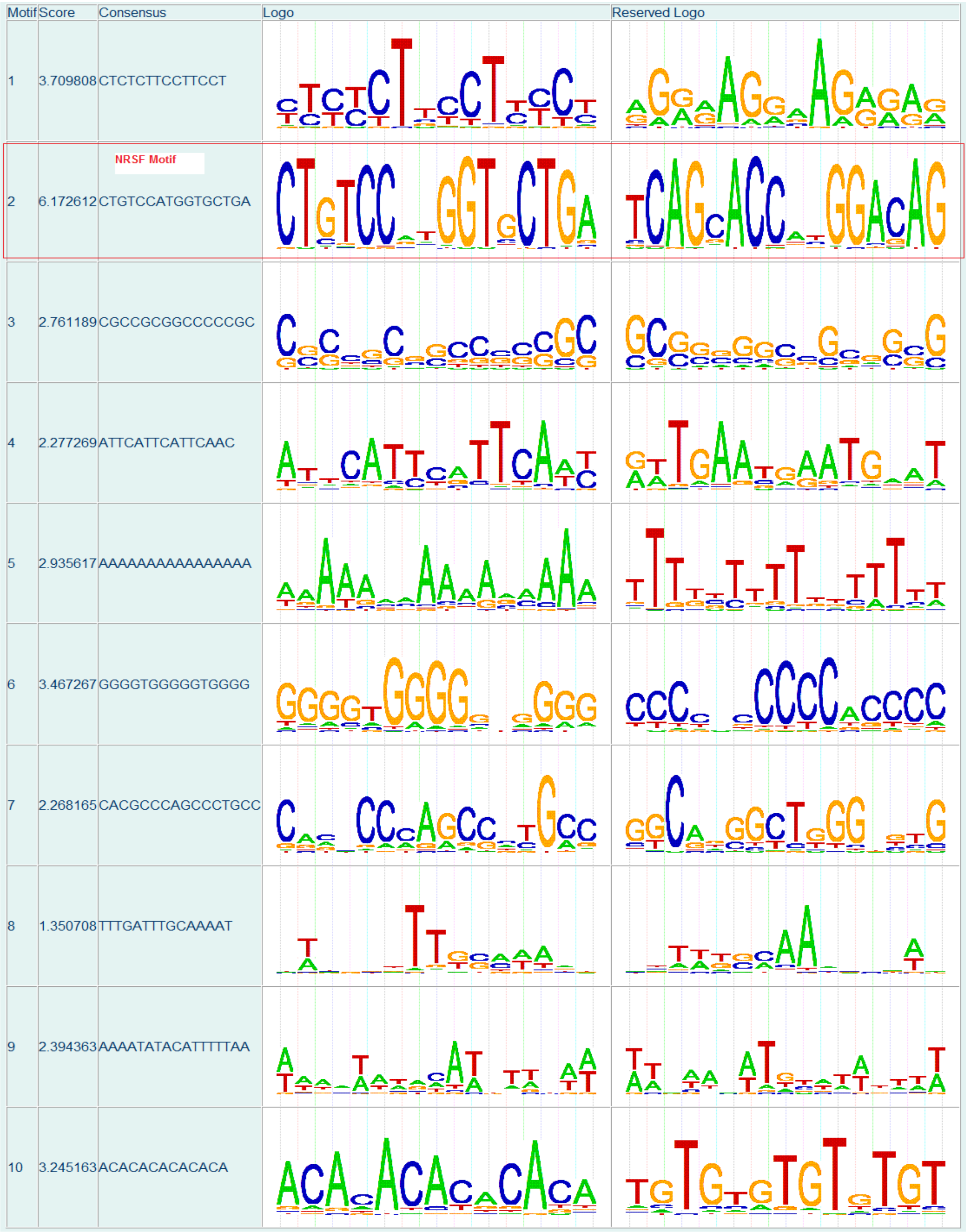
Supplementary Figure 5 Examples of sequencing artifacts detected by the one-sample ChIP-seq analysis. Raw read alignments are shown in the figure. Most artifacts detected in the negative control samples occur within 100-150 bp windows, and they are correlated with artifacts in the corresponding ChIP samples.



Supplementary Figure 6 Percentage of reads covered by peaks at different FDR cutoffs in the one-sample analysis.



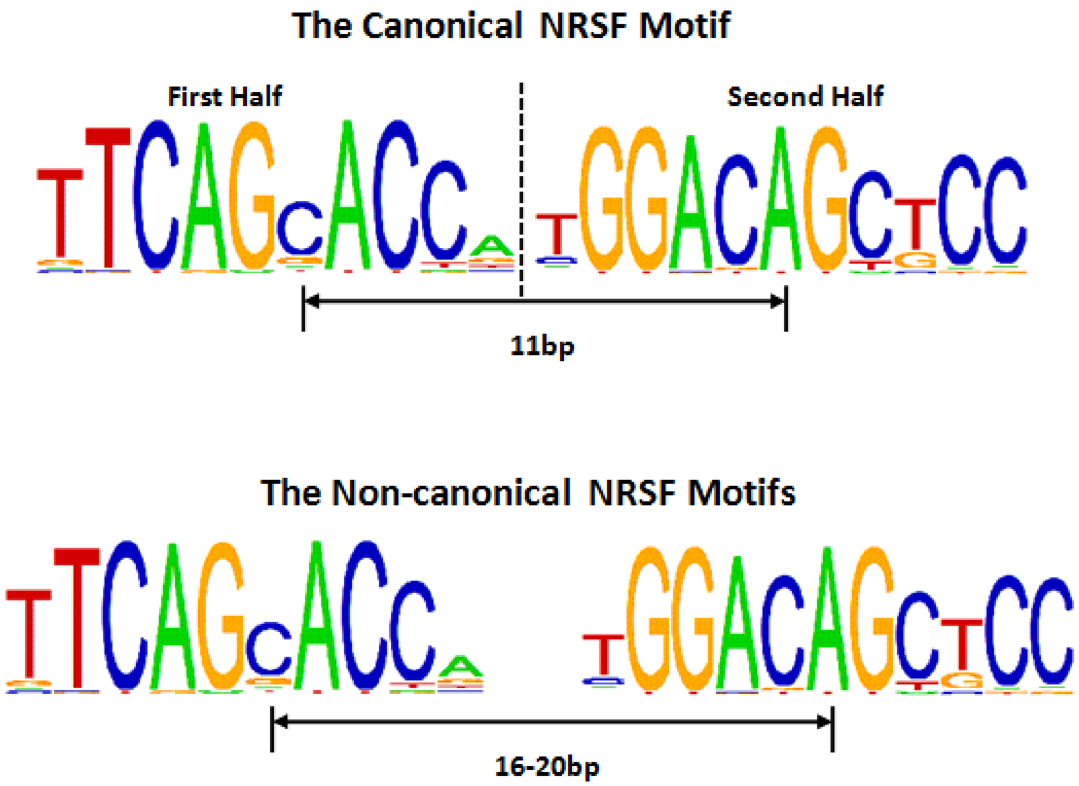
Supplementary Figure 7 Motif enrichment in binding regions predicted by the one-sample analysis. The method to compute enrichment is described in **Supplementary Methods**.



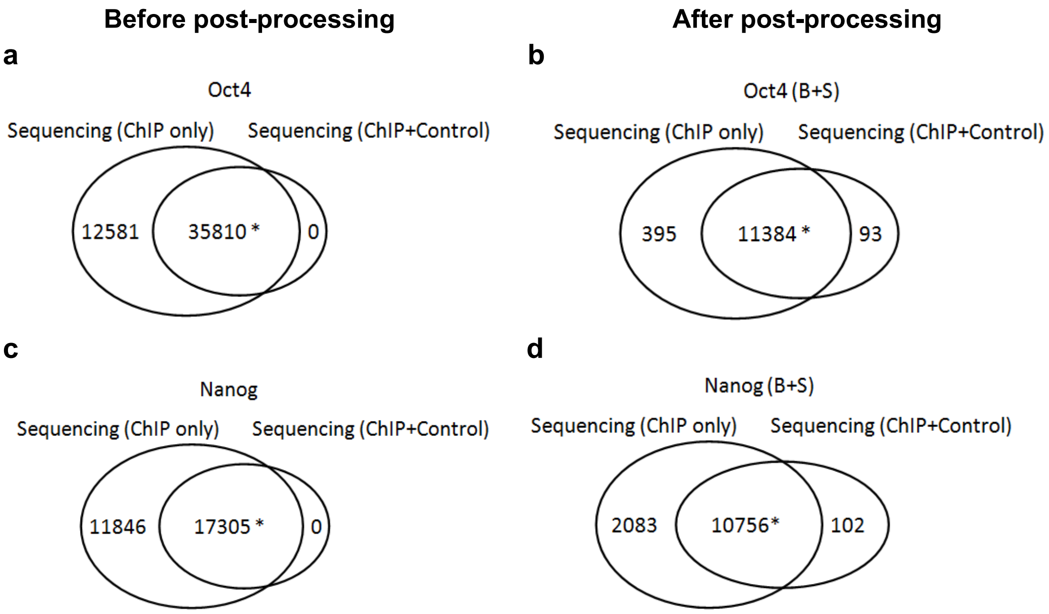
Motif	Score	Consensus	Logo	Reserved Logo
1	3.709808	CTCTCTTCCTTCCT		
2	6.172612	NRSF Motif CTGTCCATGGTGCTGA		
3	2.761189	CGCCGCGGCCCCCGC		
4	2.277269	ATTCATTCATTCAAC		
5	2.935617	AAAAAAAAAAAAAAA		
6	3.467267	GGGGTGGGGGTGGGG		
7	2.268165	CACGCCCAGCCCTGCC		
8	1.350708	TTTGATTTGCAAAAT		
9	2.394363	AAAATATACATTTTTAA		
10	3.245163	ACACACACACACA		

Supplementary Figure 8 *De novo* motif discovery results for NRSF ChIP-seq data.

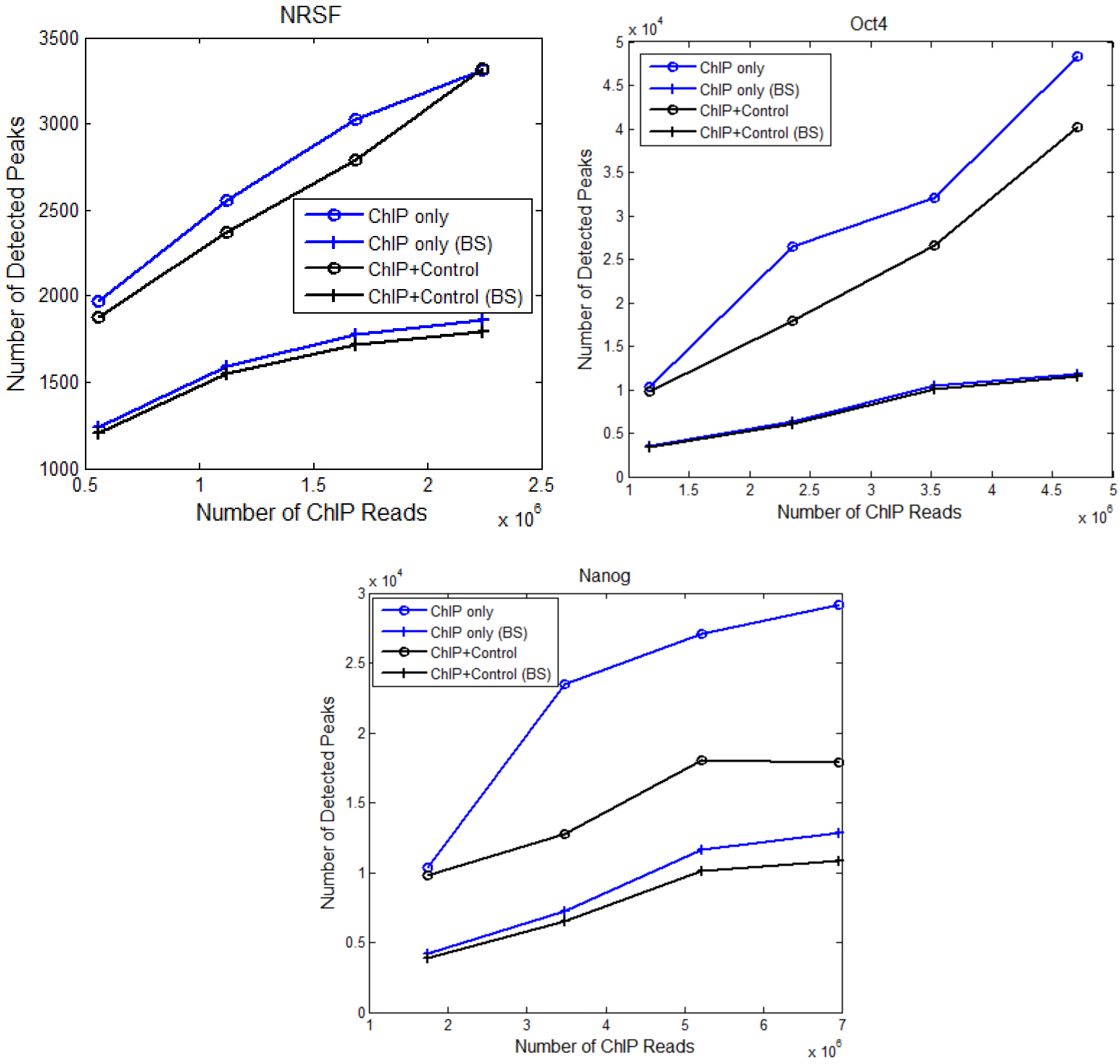
Nature Biotechnology: doi:10.1038/nbt.1505



Supplementary Figure 9 Canonical and non-canonical NRSF motifs.



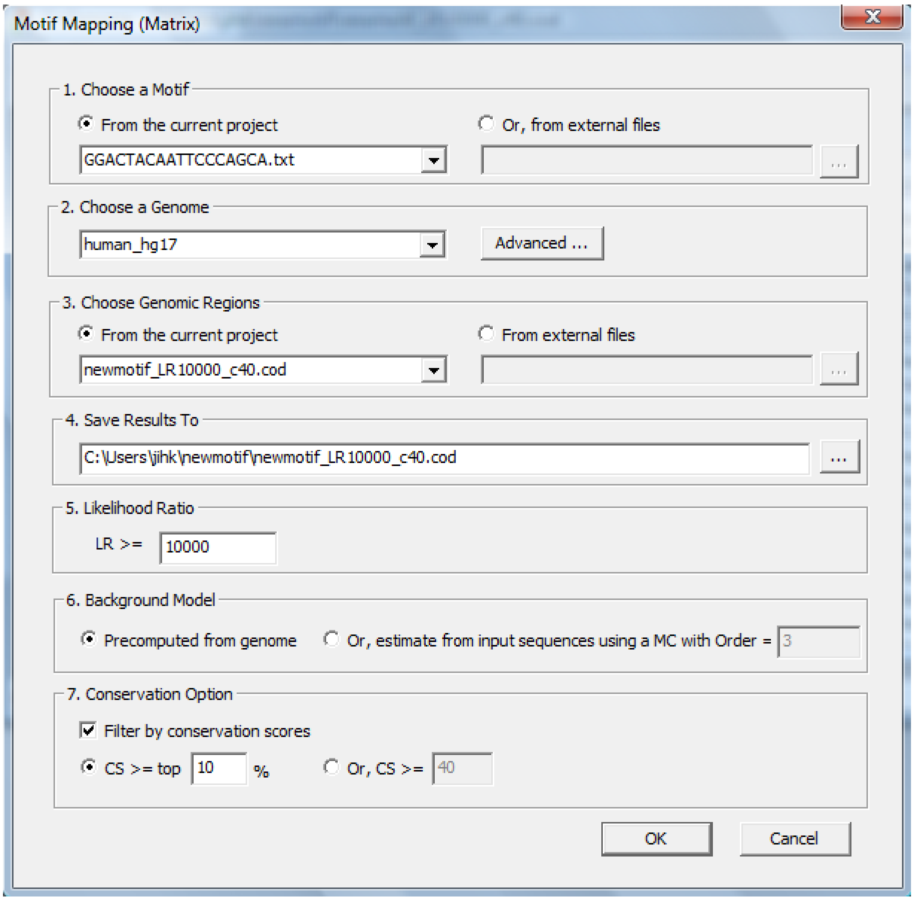
Supplementary Figure 10 Peak detection results for Oct4 and Nanog ChIP-seq data. Left: before post-processing. Right: after post-processing.



Supplementary Figure 11 Number of peaks detected at different sequencing depths. Numbers of peaks before and after post-processing are shown for both the one-sample and two-sample analyses.

a

Map the motif to the genome



Motif Mapping (Matrix)

1. Choose a Motif
   - ● From the current project: GGACTACAATTCCCAGCA.txt
   - ○ Or, from external files

2. Choose a Genome: human_hg17

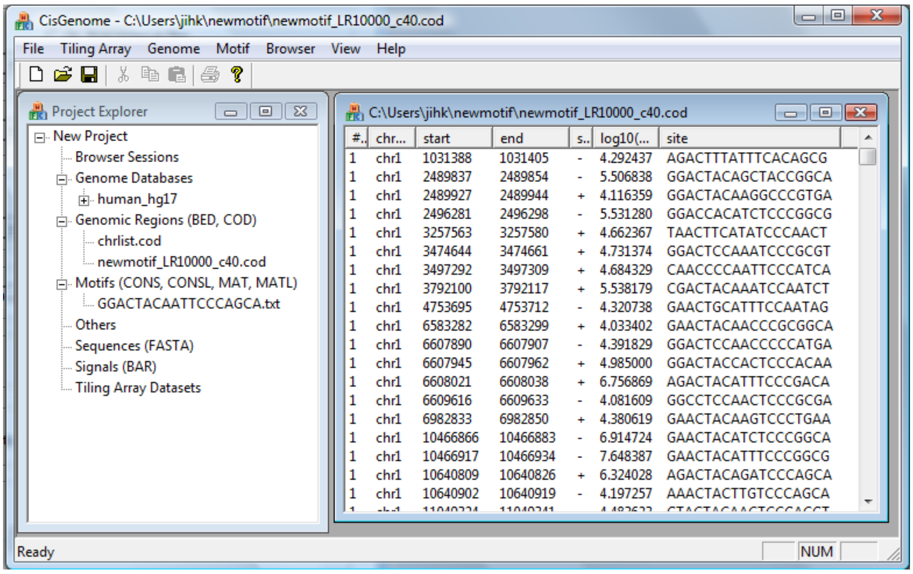
3. Choose Genomic Regions
   - ● From the current project: newmotif_LR10000_c40.cod
   - ○ From external files

4. Save Results To: C:\Users\jihk\newmotif\newmotif_LR10000_c40.cod

5. Likelihood Ratio: LR >= 10000

6. Background Model
   - ● Precomputed from genome
   - ○ Or, estimate from input sequences using a MC with Order = 3

7. Conservation Option
   - ☑ Filter by conservation scores
   - ● CS >= top 10 %
   - ○ Or, CS >= 40



CisGenome - C:\Users\jihk\newmotif\newmotif_LR10000_c40.cod

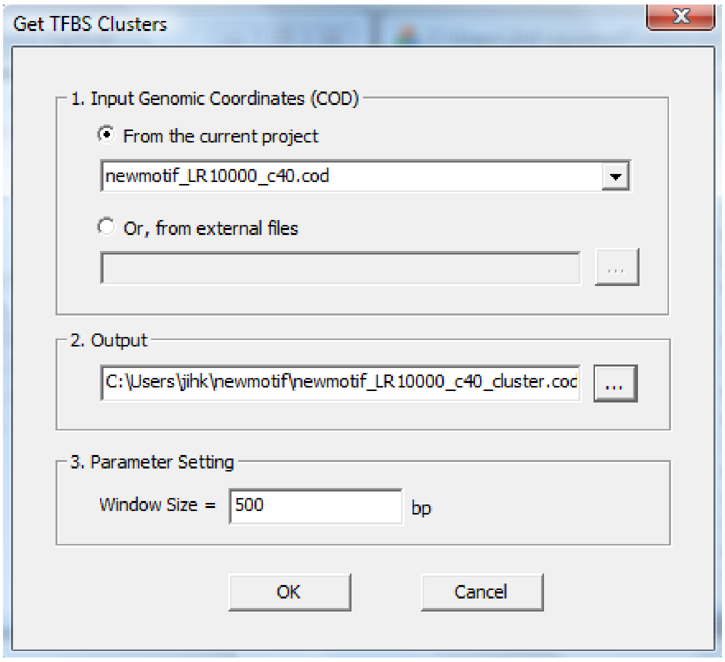
- New Project
  - Browser Sessions
  - Genome Databases
    - human_hg17
  - Genomic Regions (BED, COD)
    - chrlist.cod
    - newmotif_LR10000_c40.cod
  - Motifs (CONS, CONSL, MAT, MATL)
    - GGACTACAATTCCCAGCA.txt
  - Others
  - Sequences (FASTA)
  - Signals (BAR)
  - Tiling Array Datasets

C:\Users\jihk\newmotif\newmotif_LR10000_c40.cod

#..	chr...	start	end	s..	log10(...	site
1	chr1	1031388	1031405	-	4.292437	AGACTTTATTTCACAGCG
1	chr1	2489837	2489854	-	5.506838	GGACTACAGCTACCGGCA
1	chr1	2489927	2489944	+	4.116359	GGACTACAAGGCCCGTGA
1	chr1	2496281	2496298	-	5.531280	GGACCACATCTCCCGGCG
1	chr1	3257563	3257580	+	4.662367	TAACTTCATATCCCAACT
1	chr1	3474644	3474661	+	4.731374	GGACTCCAAATCCCGCGT
1	chr1	3497292	3497309	+	4.684329	CAACCCCAATTCCCATCA
1	chr1	3792100	3792117	+	5.538179	CGACTACAAATCCAATCT
1	chr1	4753695	4753712	-	4.320738	GAACTGCATTTCCAATAG
1	chr1	6583282	6583299	+	4.033402	GAACTACAACCCGCGGCA
1	chr1	6607890	6607907	-	4.391829	GGACTCCAACCCCCATGA
1	chr1	6607945	6607962	+	4.985000	GGACTACCACTCCCACAA
1	chr1	6608021	6608038	+	6.756869	AGACTACATTTCCCGACA
1	chr1	6609616	6609633	-	4.081609	GGCCTCCAACTCCCGCGA
1	chr1	6982833	6982850	+	4.380619	GAACTACAAGTCCCTGAA
1	chr1	10466866	10466883	-	6.914724	GAACTACATCTCCCGGCA
1	chr1	10466917	10466934	-	7.648387	GAACTACATTTCCCGGCG
1	chr1	10640809	10640826	+	6.324028	AGACTACAGATCCCAGCA
1	chr1	10640902	10640919	-	4.197257	AAACTACTTGTCCCAGCA

Supplementary Figure 12 Analysis of the novel motif using CisGenome.

b **Retrieve clustered motif sites that are separated by ≤ 500 bp**



Get TFBS Clusters

1. Input Genomic Coordinates (COD)

● From the current project

newmotif_LR10000_c40.cod

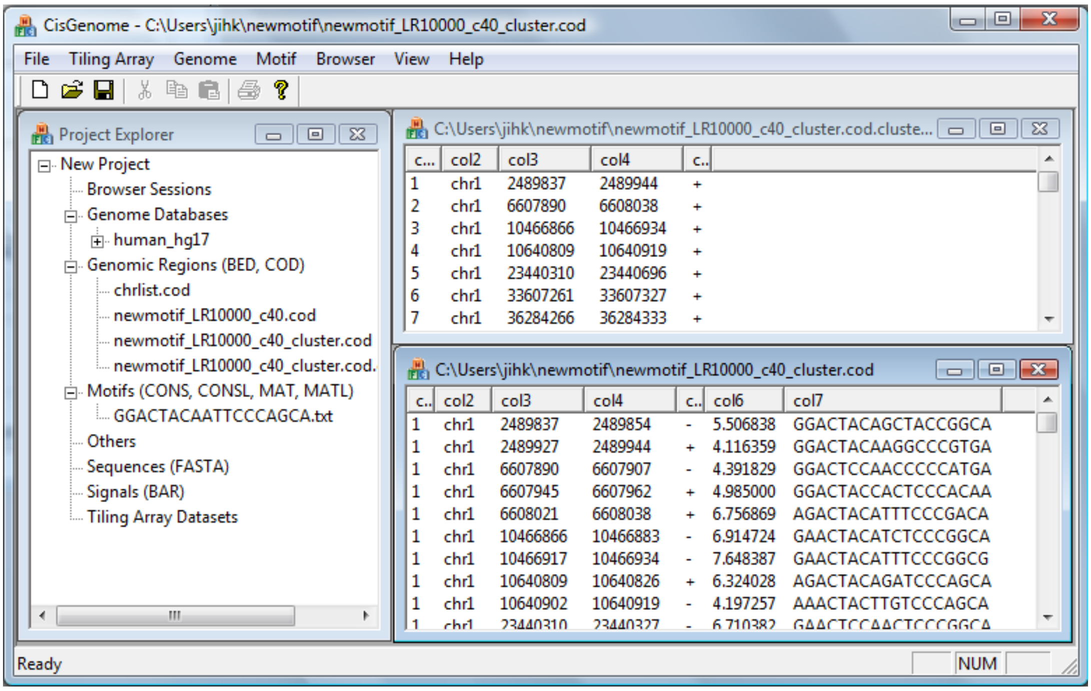
○ Or, from external files

2. Output

C:\Users\jihk\newmotif\newmotif_LR10000_c40_cluster.coc

3. Parameter Setting

Window Size = 500 bp



CisGenome - C:\Users\jihk\newmotif\newmotif_LR10000_c40_cluster.cod

C:\Users\jihk\newmotif\newmotif_LR10000_c40_cluster.cod.cluste...

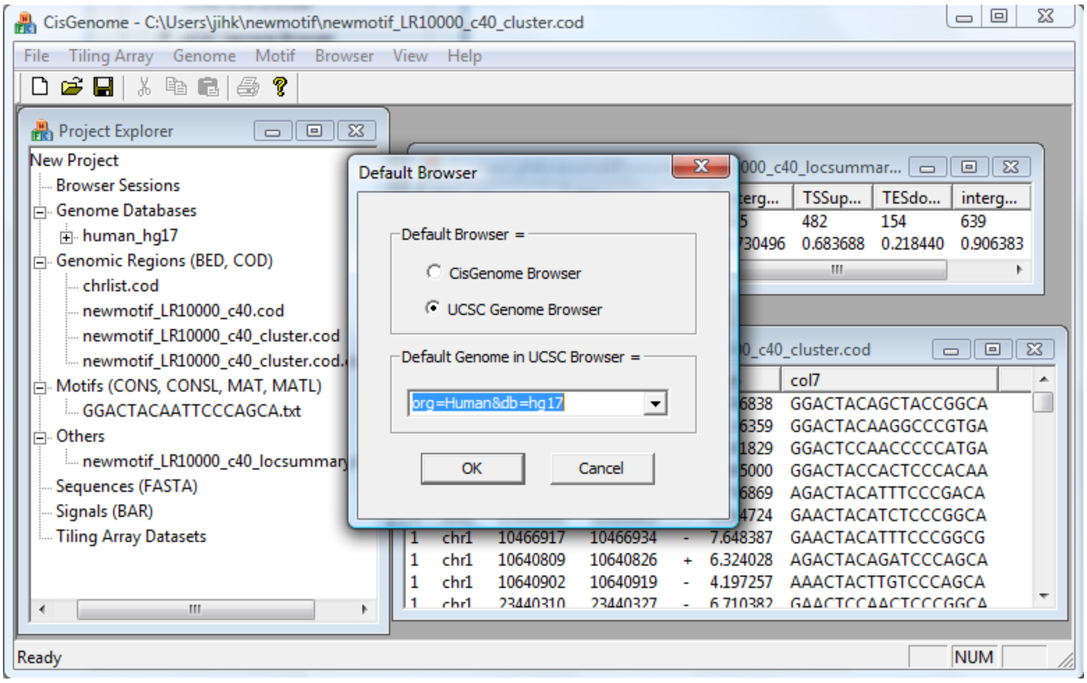
c...	col2	col3	col4	c..
1	chr1	2489837	2489944	+
2	chr1	6607890	6608038	+
3	chr1	10466866	10466934	+
4	chr1	10640809	10640919	+
5	chr1	23440310	23440696	+
6	chr1	33607261	33607327	+
7	chr1	36284266	36284333	+

C:\Users\jihk\newmotif\newmotif_LR10000_c40_cluster.cod

c..	col2	col3	col4	c..	col6	col7
1	chr1	2489837	2489854	-	5.506838	GGACTACAGCTACCGGCA
1	chr1	2489927	2489944	+	4.116359	GGACTACAAGGCCCGTGA
1	chr1	6607890	6607907	-	4.391829	GGACTCCAACCCCCATGA
1	chr1	6607945	6607962	+	4.985000	GGACTACCACTCCCACAA
1	chr1	6608021	6608038	+	6.756869	AGACTACATTTCCCGACA
1	chr1	10466866	10466883	-	6.914724	GAACTACATCTCCCGGCA
1	chr1	10466917	10466934	-	7.648387	GAACTACATTTCCCGGCG
1	chr1	10640809	10640826	+	6.324028	AGACTACAGATCCCAGCA
1	chr1	10640902	10640919	-	4.197257	AAACTACTTGTCCCAGCA
1	chr1	23440310	23440327	-	6.710382	GAACTCCAACTCCCGGCA

Supplementary Figure 12 Analysis of the novel motif using CisGenome (cont.).

c **Choose a browser for visual inspection**



CisGenome - C:\Users\jihk\newmotif\newmotif_LR10000_c40_cluster.cod

New Project
- Browser Sessions
- Genome Databases
  - human_hg17
- Genomic Regions (BED, COD)
  - chrlist.cod
  - newmotif_LR10000_c40.cod
  - newmotif_LR10000_c40_cluster.cod
  - newmotif_LR10000_c40_cluster.cod.
- Motifs (CONS, CONSL, MAT, MATL)
  - GGACTACAATTCCCAGCA.txt
- Others
  - newmotif_LR10000_c40_locsummary
- Sequences (FASTA)
- Signals (BAR)
- Tiling Array Datasets

Default Browser

Default Browser =
○ CisGenome Browser
● UCSC Genome Browser

Default Genome in UCSC Browser = org=Human&db=hg17

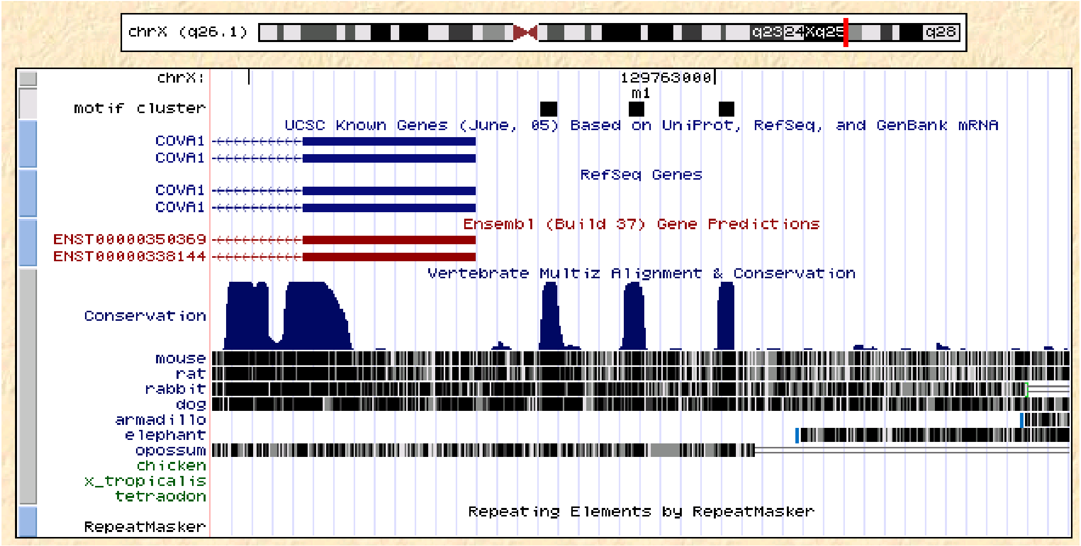
000_c40_locsummar...

```
erg...   TSSup...  TESdo...  interg...
5        482       154       639
30496    0.683688  0.218440  0.906383
```

0_c40_cluster.cod

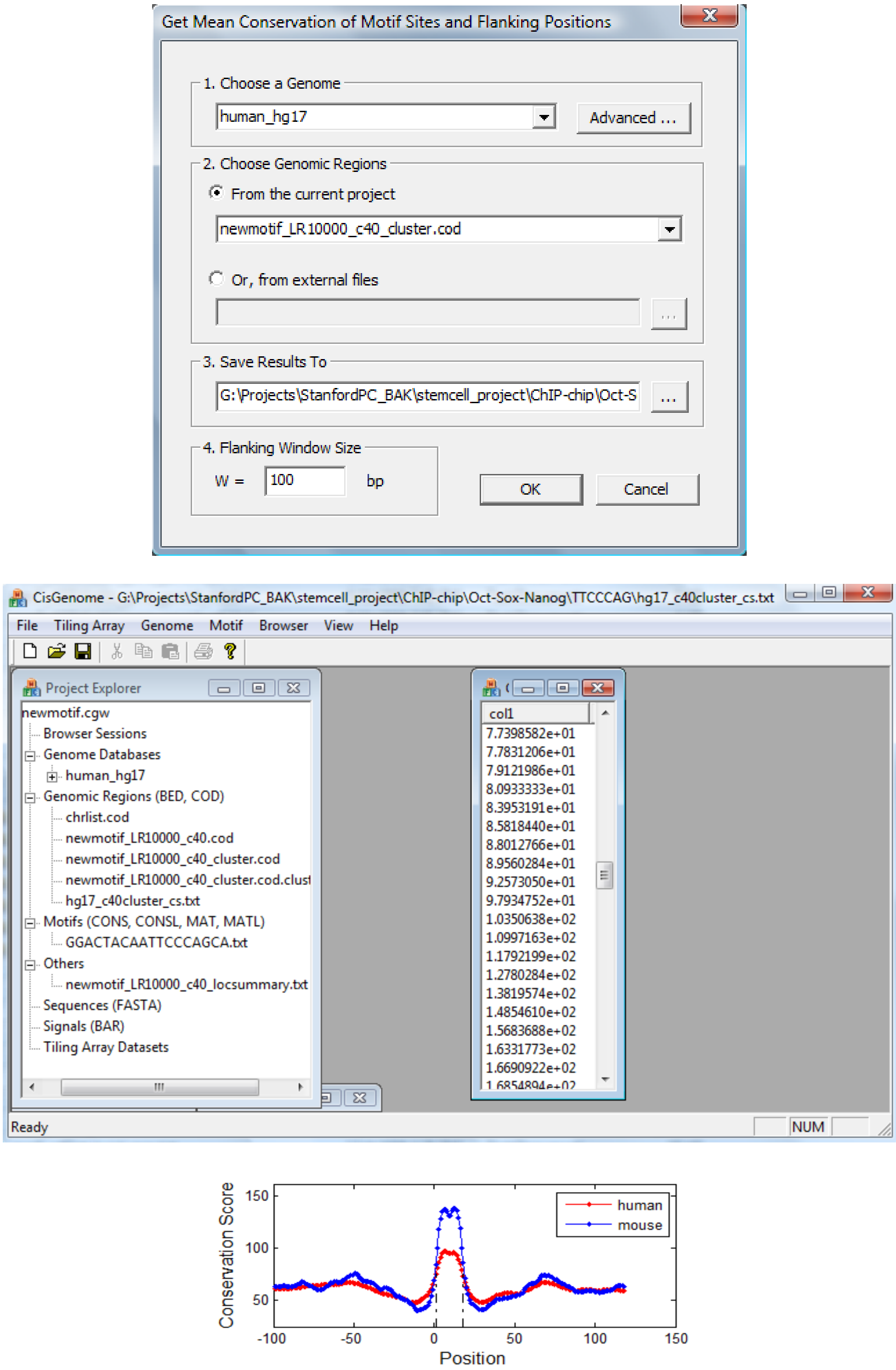
```
                                            col7
                                      6838  GGACTACAGCTACCGGCA
                                      6359  GGACTACAAGGCCCGTGA
                                      1829  GGACTCCAACCCCCATGA
                                      5000  GGACTACCACTCCCACAA
                                      6869  AGACTACATTTCCCGACA
                                      4724  GAACTACATCTCCCGGCA
1  chr1  10466917  10466934  -  7.648387  GAACTACATTTCCCGGCG
1  chr1  10640809  10640826  +  6.324028  AGACTACAGATCCCAGCA
1  chr1  10640902  10640919  -  4.197257  AAACTACTTGTCCCAGCA
1  chr1  23440310  23440327  -  6.710382  GAACTCCAACTCCCGGCA
```

Ready NUM



Supplementary Figure 12 Analysis of the novel motif using CisGenome (cont.).

d **Get conservation scores for flanking positions of the motif**



Get Mean Conservation of Motif Sites and Flanking Positions

1. Choose a Genome

human_hg17 ▾ Advanced ...

2. Choose Genomic Regions

● From the current project

newmotif_LR10000_c40_cluster.cod ▾

○ Or, from external files

...

3. Save Results To

G:\Projects\StanfordPC_BAK\stemcell_project\ChIP-chip\Oct-S ...

4. Flanking Window Size

W = 100 bp

OK Cancel

CisGenome - G:\Projects\StanfordPC_BAK\stemcell_project\ChIP-chip\Oct-Sox-Nanog\TTCCCAG\hg17_c40cluster_cs.txt

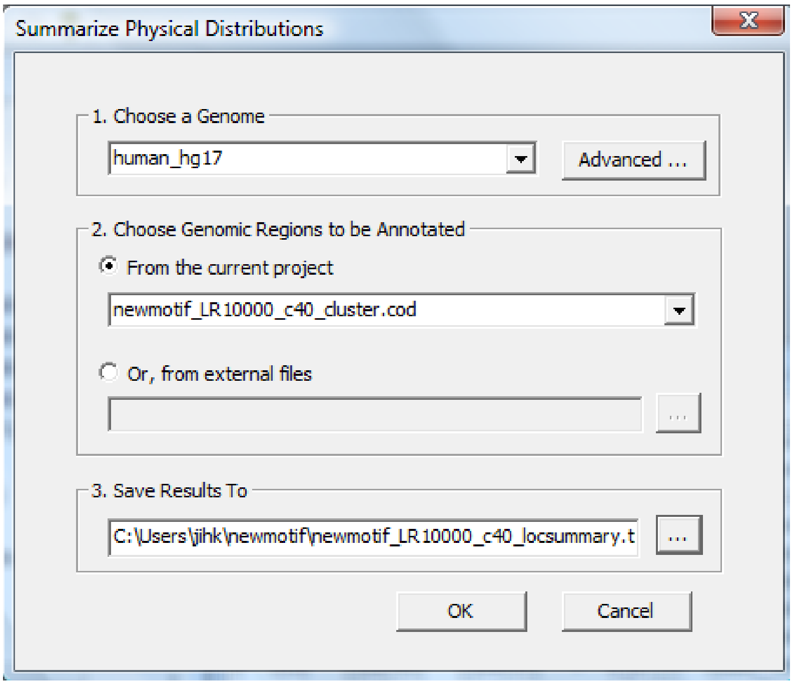
```
newmotif.cgw
  Browser Sessions
  Genome Databases
    human_hg17
  Genomic Regions (BED, COD)
    chrlist.cod
    newmotif_LR10000_c40.cod
    newmotif_LR10000_c40_cluster.cod
    newmotif_LR10000_c40_cluster.cod.clust
    hg17_c40cluster_cs.txt
  Motifs (CONS, CONSL, MAT, MATL)
    GGACTACAATTCCCAGCA.txt
  Others
    newmotif_LR10000_c40_locsummary.txt
  Sequences (FASTA)
  Signals (BAR)
  Tiling Array Datasets
```

col1
7.7398582e+01
7.7831206e+01
7.9121986e+01
8.0933333e+01
8.3953191e+01
8.5818440e+01
8.8012766e+01
8.9560284e+01
9.2573050e+01
9.7934752e+01
1.0350638e+02
1.0997163e+02
1.1792199e+02
1.2780284e+02
1.3819574e+02
1.4854610e+02
1.5683688e+02
1.6331773e+02
1.6690922e+02
1.6854894e+02

Supplementary Figure 12 Analysis of the novel motif using CisGenome (cont.).

e

Summarize distribution of motif sites in relative to gene structures



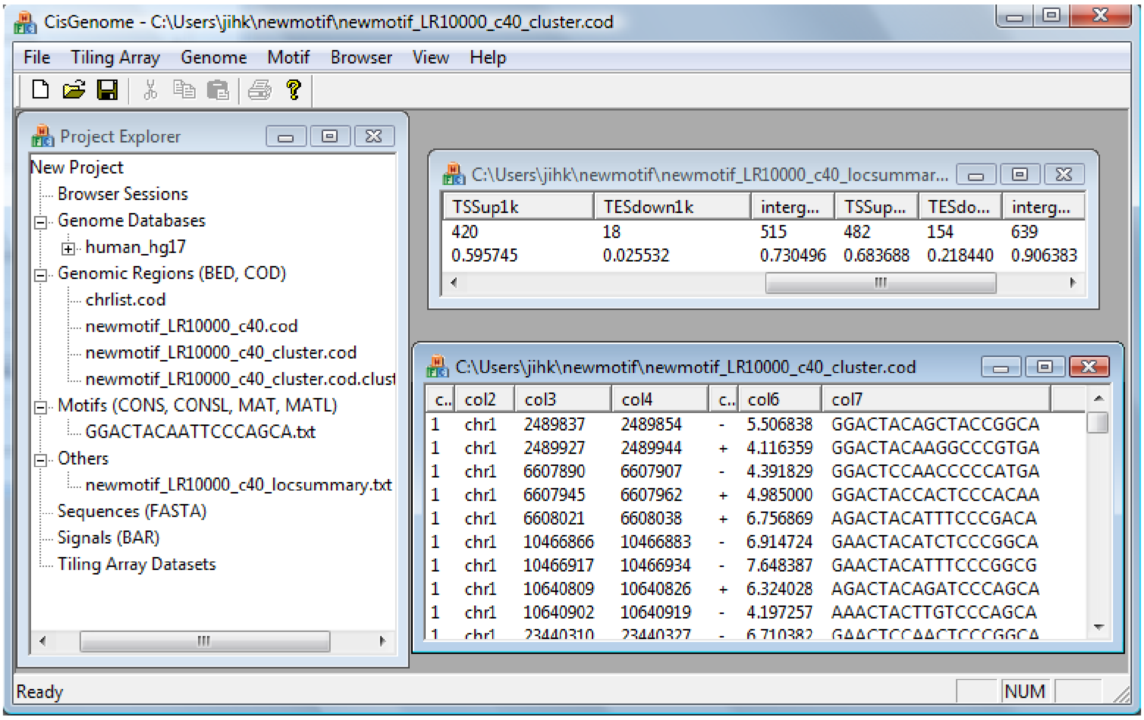
1. Choose a Genome: human_hg17

2. Choose Genomic Regions to be Annotated

● From the current project: newmotif_LR10000_c40_cluster.cod

○ Or, from external files:

3. Save Results To: C:\Users\jihk\newmotif\newmotif_LR10000_c40_locsummary.t


CisGenome - C:\Users\jihk\newmotif\newmotif_LR10000_c40_cluster.cod
File
Tiling Array
Genome
Motif
Browser
View
Help
Project Explorer
Ready
NUM

New Project
- Browser Sessions
- Genome Databases
  - human_hg17
- Genomic Regions (BED, COD)
  - chrlist.cod
  - newmotif_LR10000_c40.cod
  - newmotif_LR10000_c40_cluster.cod
  - newmotif_LR10000_c40_cluster.cod.clust
- Motifs (CONS, CONSL, MAT, MATL)
  - GGACTACAATTCCCAGCA.txt
- Others
  - newmotif_LR10000_c40_locsummary.txt
- Sequences (FASTA)
- Signals (BAR)
- Tiling Array Datasets

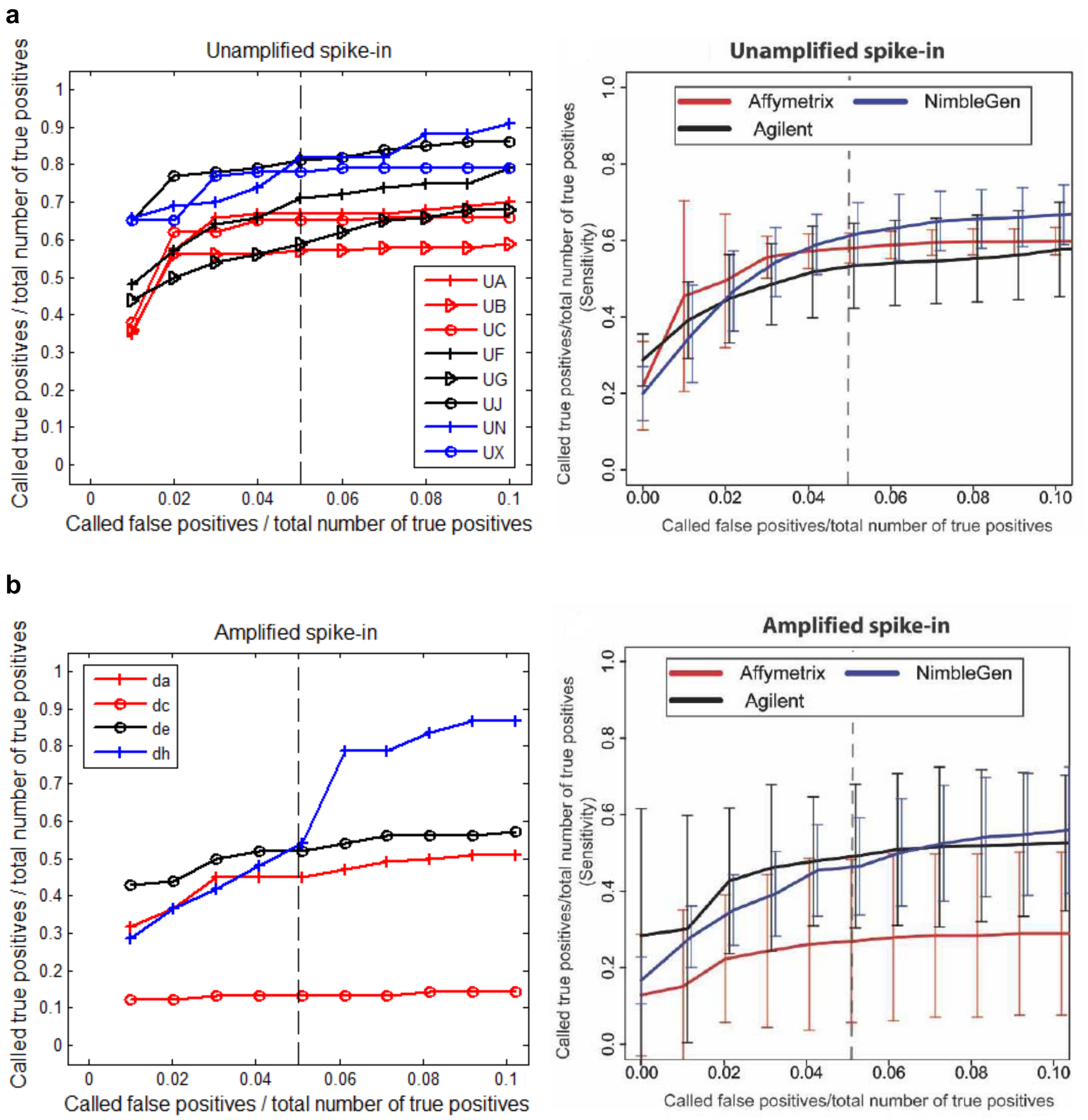
C:\Users\jihk\newmotif\newmotif_LR10000_c40_locsummar...

TSSup1k	TESdown1k	interg...	TSSup...	TESdo...	interg...
420	18	515	482	154	639
0.595745	0.025532	0.730496	0.683688	0.218440	0.906383

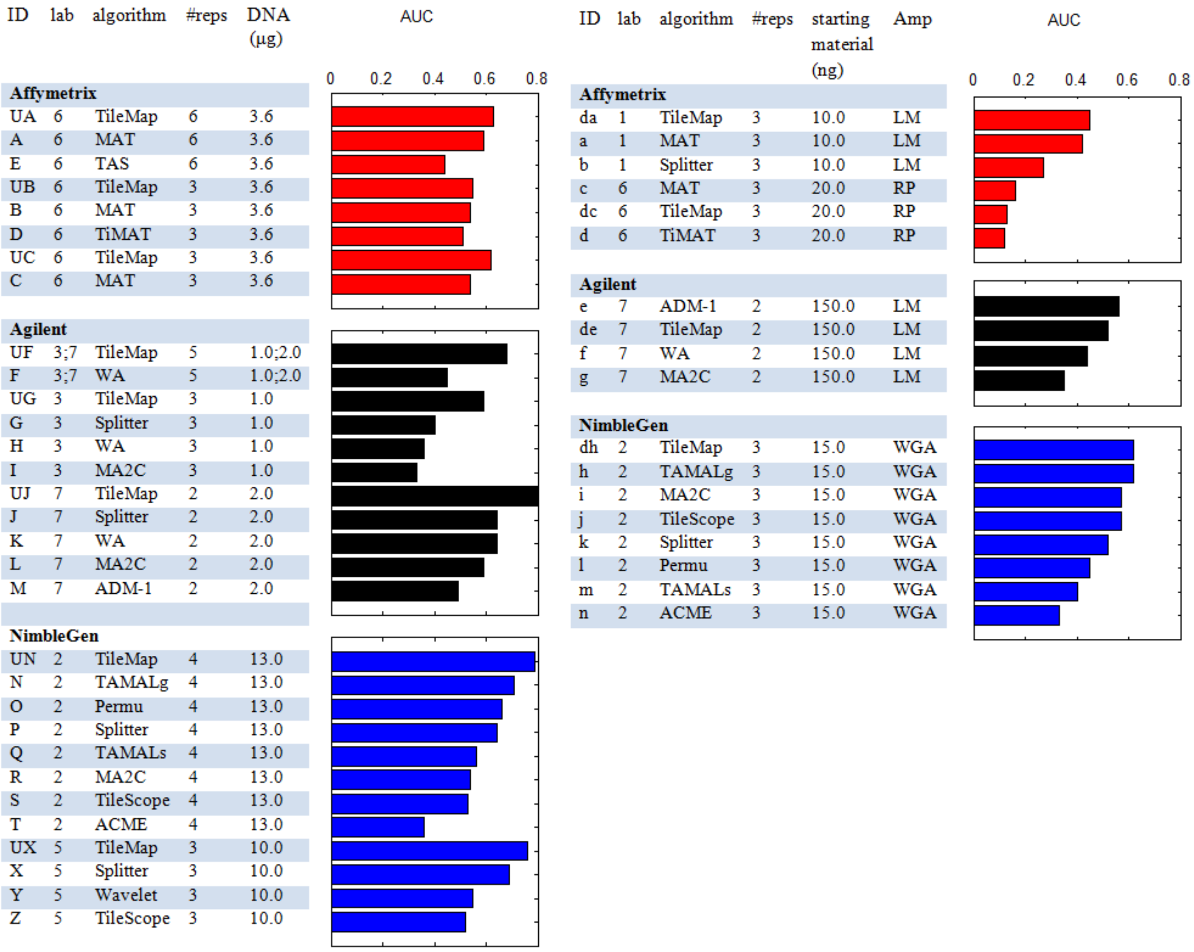
C:\Users\jihk\newmotif\newmotif_LR10000_c40_cluster.cod

c..	col2	col3	col4	c..	col6	col7
1	chr1	2489837	2489854	-	5.506838	GGACTACAGCTACCGGCA
1	chr1	2489927	2489944	+	4.116359	GGACTACAAGGCCCGTGA
1	chr1	6607890	6607907	-	4.391829	GGACTCCAACCCCCATGA
1	chr1	6607945	6607962	+	4.985000	GGACTACCACTCCCACAA
1	chr1	6608021	6608038	+	6.756869	AGACTACATTTCCCGACA
1	chr1	10466866	10466883	-	6.914724	GAACTACATCTCCCGGCA
1	chr1	10466917	10466934	-	7.648387	GAACTACATTTCCCGGCG
1	chr1	10640809	10640826	+	6.324028	AGACTACAGATCCCAGCA
1	chr1	10640902	10640919	-	4.197257	AAACTACTTGTCCCAGCA
1	chr1	23440310	23440327	-	6.710382	GAACTCCAACTCCCGGCA

Supplementary Figure 12 Analysis of the novel motif using CisGenome (cont.).



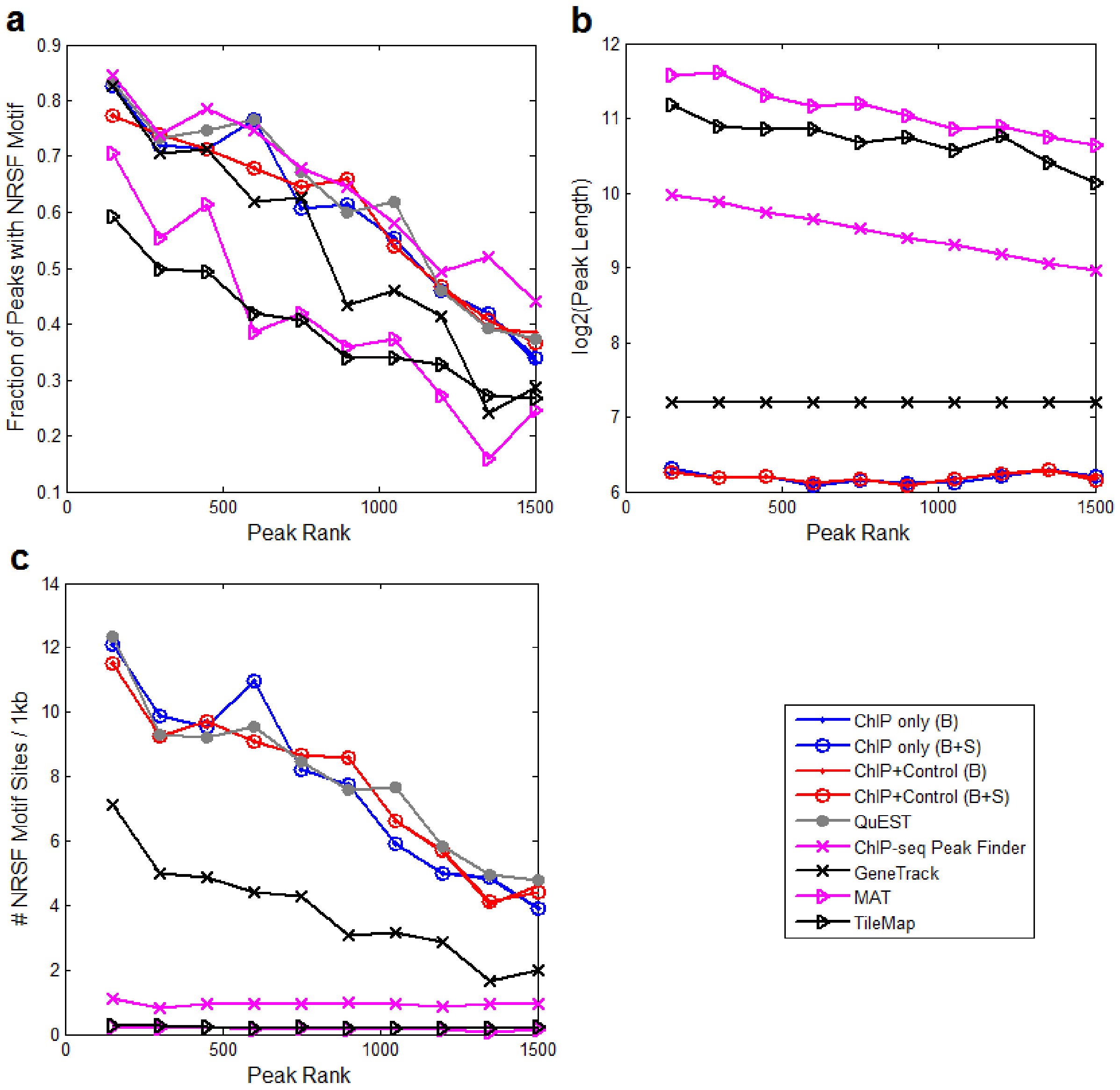
Supplementary Figure 13 ROC-like curve for TileMap. ROC-like curve for unamplified spike-in data (a) and amplified spike-in data (b). The plots in the left column correspond to the TileMap results, and the plots in the right column are the original Fig 2c,d from ref. 41 which correspond to the average performance of all other algorithms. The analysis codes (e.g., UA, UB, da, dc) are defined in Supplementary Figure 14. The dashed vertical line represents the point at which the number of false-positive predictions is equal to 5% of the total number of true-positive spike-ins. For the plots on the right, error bars represent the two-sided 95% confidence interval of the average sensitivity at each false-positive ratio (X-axis).



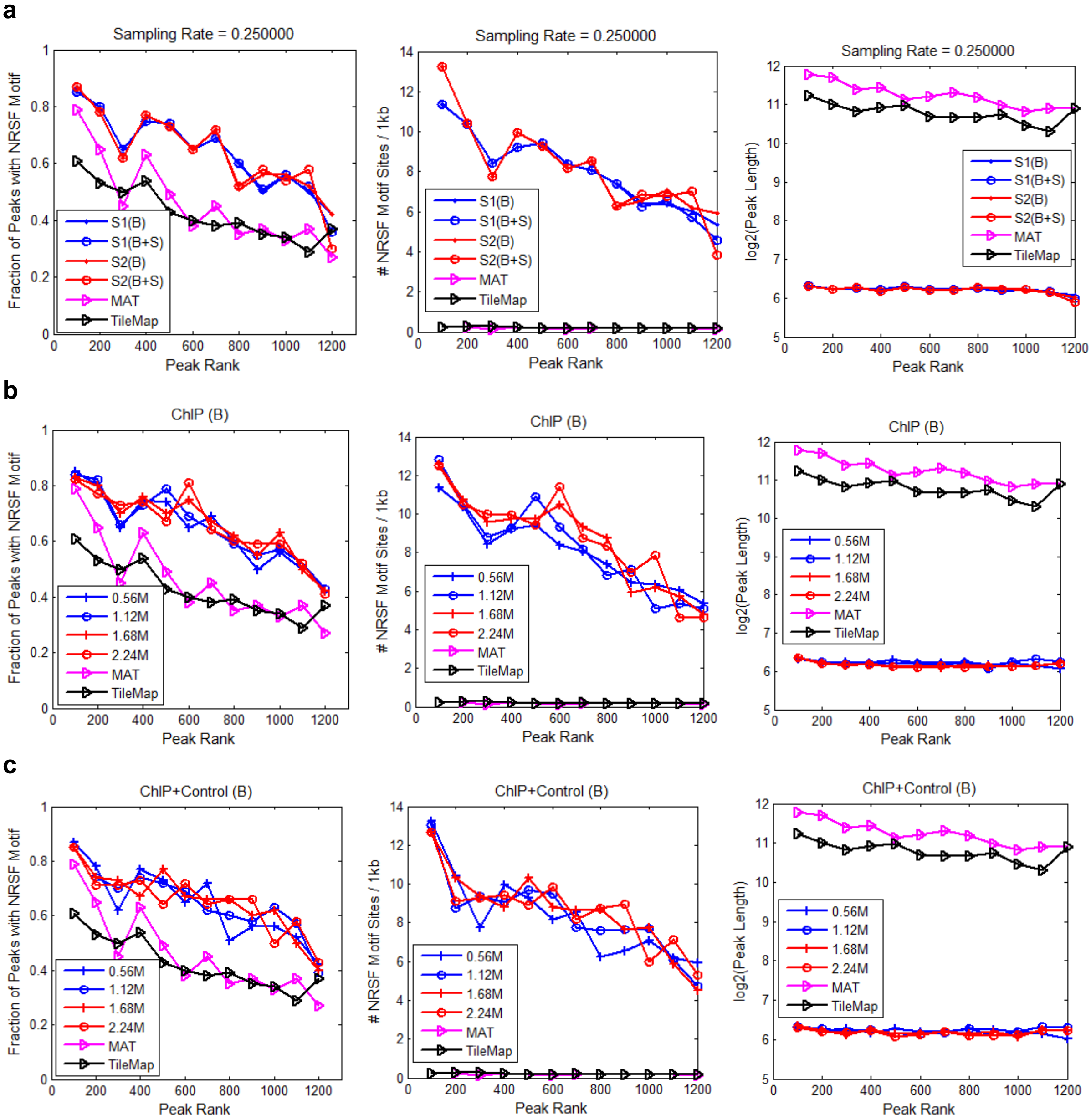
ID	lab	algorithm	#reps	DNA (µg)
Affymetrix				
UA	6	TileMap	6	3.6
A	6	MAT	6	3.6
E	6	TAS	6	3.6
UB	6	TileMap	3	3.6
B	6	MAT	3	3.6
D	6	TiMAT	3	3.6
UC	6	TileMap	3	3.6
C	6	MAT	3	3.6
Agilent				
UF	3;7	TileMap	5	1.0;2.0
F	3;7	WA	5	1.0;2.0
UG	3	TileMap	3	1.0
G	3	Splitter	3	1.0
H	3	WA	3	1.0
I	3	MA2C	3	1.0
UJ	7	TileMap	2	2.0
J	7	Splitter	2	2.0
K	7	WA	2	2.0
L	7	MA2C	2	2.0
M	7	ADM-1	2	2.0
NimbleGen				
UN	2	TileMap	4	13.0
N	2	TAMALg	4	13.0
O	2	Permu	4	13.0
P	2	Splitter	4	13.0
Q	2	TAMALs	4	13.0
R	2	MA2C	4	13.0
S	2	TileScope	4	13.0
T	2	ACME	4	13.0
UX	5	TileMap	3	10.0
X	5	Splitter	3	10.0
Y	5	Wavelet	3	10.0
Z	5	TileScope	3	10.0

ID	lab	algorithm	#reps	starting material (ng)	Amp
Affymetrix					
da	1	TileMap	3	10.0	LM
a	1	MAT	3	10.0	LM
b	1	Splitter	3	10.0	LM
c	6	MAT	3	20.0	RP
dc	6	TileMap	3	20.0	RP
d	6	TiMAT	3	20.0	RP
Agilent					
e	7	ADM-1	2	150.0	LM
de	7	TileMap	2	150.0	LM
f	7	WA	2	150.0	LM
g	7	MA2C	2	150.0	LM
NimbleGen					
dh	2	TileMap	3	15.0	WGA
h	2	TAMALg	3	15.0	WGA
i	2	MA2C	3	15.0	WGA
j	2	TileScope	3	15.0	WGA
k	2	Splitter	3	15.0	WGA
l	2	Permu	3	15.0	WGA
m	2	TAMALs	3	15.0	WGA
n	2	ACME	3	15.0	WGA

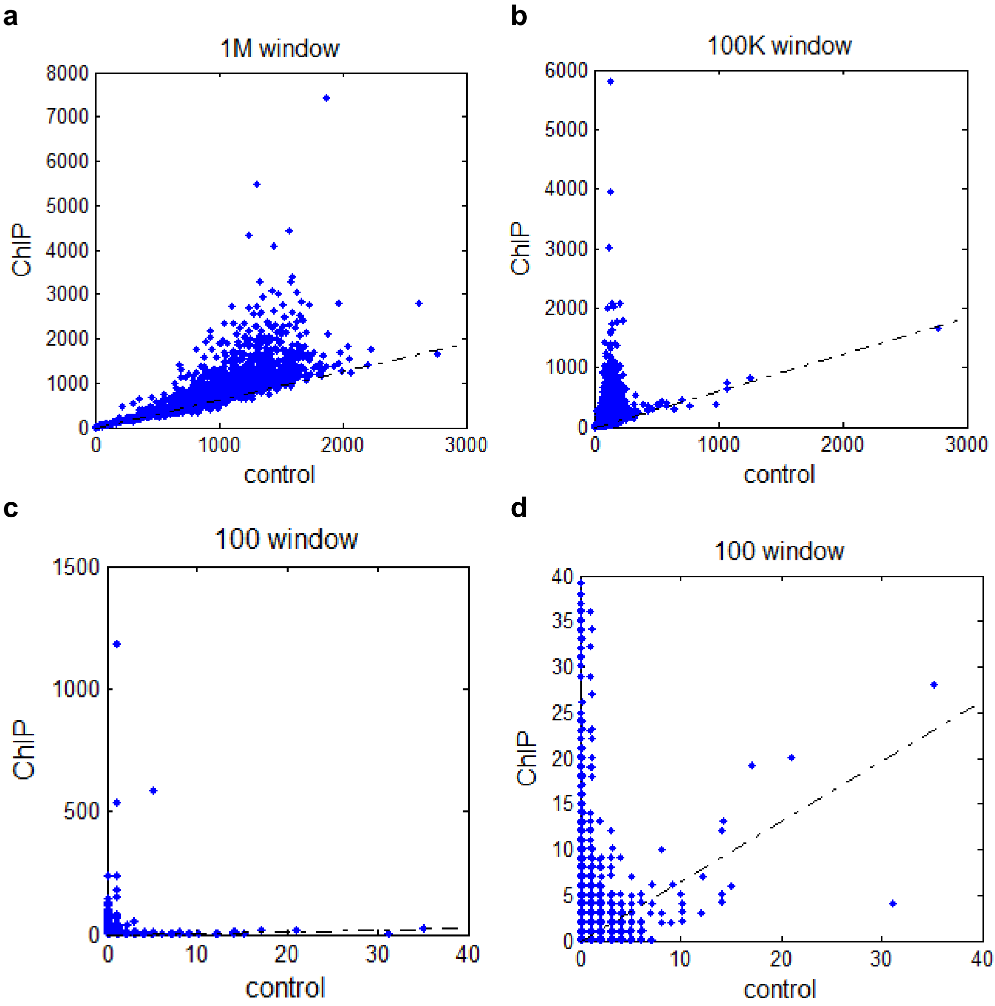
Supplementary Figure 14 Areas under the ROC curve (AUC) of different ChIP-chip peak detection algorithms. The bigger the AUC, the better an algorithm performs. References for the tools are TileMap[12], MAT[11], TAS[13], TiMAT (http://sourceforge.net/projects/timat2), Splitter (http://zlab.bu.edu/splitter), WA (see ref. 41 for a description), MA2C[22], ADM-1 (http://www.cs.technion.ac.il/users/wwwb/cgi-bin/tr-get.cgi/2007/PHD/PHD-2007-05.pdf), TAMAL[20], Permu[75], TileScope[21], ACME[19], Wavelet (A. Karpikov and M. Gerstein, unpubl.). TileMap analysis results are indexed by two letter codes UA, UB, UC, UF, UG, UJ, UN, UX, da, dc, de and dh. Analysis results of the other algorithms are indexed by one letter codes consistent with the original codes in ref. 41.



Supplementary Figure 15 Performance of ChIP-seq peak detection algorithms. CisGenome was run under four modes (i.e., ChIP only with boundary refinement, ChIP only with boundary refinement and single strand filtering, ChIP+control with boundary refinement, and ChIP+control with boundary refinement and single strand filtering). (a) Percentage of peaks that contain ≥1 NRSF motif site. The top 1,500 predictions of each algorithm were grouped into ten tiers. The percentage was computed for each tier. (b) Log2(Peak Length) of the predictions made by different algorithms. (c) NRSF motif occurrence rates in the predicted peaks. QuEST only reports a single coordinate for each peak and does not provide boundary estimates. Thus the peak length for QuEST predictions cannot be computed and was not compared here. For determining NRSF motif coverage and motif occurrence rate, the single QuEST coordinate for each peak has been extended 40bp towards both ends. 40bp represents a slightly longer half fragment length than the half peak length of the CisGenome peaks.



Supplementary Figure 16 Effects of read number on motif coverage and peak length. (a) CisGenome one-sample and two-sample ChIP-seq analyses on 25% of the original reads were compared with TileMap and MAT ChIP-chip analyses. From the left to right, the three plots show the percentage of peaks that contain the NRSF motif, the motif occurrence rate and the log2(peak length). Peaks were ranked and grouped into tiers of size 100. Each tier was analyzed separately. (b) Comparison of ChIP-chip peak detection with CisGenome one-sample ChIP-seq analysis (with boundary refinement but without single strand filtering) when 25%, 50%, 75% and 100% of the original reads were analyzed. The original ChIP read number is 2.24M. (c) Comparison of ChIP-chip peak detection with CisGenome two-sample ChIP-seq analysis (with boundary refinement but without single strand filtering) when 25%, 50%, 75% and 100% of the original reads were analyzed. The original ChIP (control) read number is 2.24M (2.78M).



Supplementary Figure 17 Correlation of read sampling rates between the ChIP and control samples. Genome has been divided into 1Mb (a), 100kb (b) and 100bp (c,d) non-overlapping windows. For each window, the number of reads in the NRSF ChIP sample and the number of reads in the control sample were counted and plotted against each other. Each dot represents a window. (d) is a zoom-in version of (c). The dotted line represents the expected ChIP/control read ratio r_0 in the background regions. r_0 is dependent on the total number of ChIP reads and the total number of control reads. It was derived according to Supplementary Methods. In each plot, there are two types of windows, i.e. windows that contain real binding regions (ChIP reads in these windows are significantly enriched) and windows that do not contain real binding regions (i.e., background windows). For background windows, there was a clear correlation between the ChIP read number and the control read number. In fact, in (c) and (d), when windows with ≥10 ChIP reads and at the same time ≤5 control reads were excluded (these are windows in the upper left part of (c) and (d) which mainly represent real binding signals), the correlation coefficient between the ChIP and control reads for the remaining windows (which mainly represent background) was 0.10 (99% confidence interval = [0.098,0.102] based on Fisher's z transformation), which is significantly greater than 0.

Supplementary Table 1. Real datasets used in the analyses

Name	Type	Species	Read number (million)	Source
NRSF-Tiling	ChIP-chip	Human	NA	GEO: GSE8489
NRSF-ChIP	ChIP-seq	Human	2.24	Johnson et al.[4]
NRSF-control	ChIP-seq	Human	2.78	Johnson et al. [4]
Oct4-ChIP	ChIP-seq	Mouse	4.71	Marson et al. [10]
Nanog-ChIP	ChIP-seq	Mouse	8.72	Marson et al. [10]
ES-control	ChIP-seq	Mouse	6.96	Marson et al. [10]

Note: Reads were aligned to hg17 and mm8 respectively. Only uniquely mapped reads were counted.

Supplementary Table 2. Summary of simulated datasets for evaluating ChIP-seq FDR estimation

	Enrichment ratio *r*		
Peak number *p*	20	100	500
(NRSF-control)			
2000	0.21% (2.78)	1.26% (2.81)	5.97% (2.95)
10000	1.10% (2.81)	5.89% (2.95)	24.03% (3.66)
50000	5.30% (2.93)	23.94% (3.65)	61.35% (7.19)
(ES-control)			
2000	0.28% (6.98)	1.45% (7.07)	7.00% (7.49)
10000	1.42% (7.07)	6.98% (7.49)	27.63% (9.62)
50000	6.72% (7.47)	27.24% (9.57)	65.37% (20.11)

Note: 18 datasets were generated. For each dataset, the percentage of reads covered by the peaks detected at the 10% FDR level is shown. The total number of reads (i.e., negative control reads + simulated reads, in the unit of million) in the datasets are shown in the brackets.

Supplementary Table 3. Summary of one-sample ChIP-seq peak detection results in different datasets

Data	No. of predicted peaks (FDR≤0.1, after post-processing)	No. of non-repeat peaks[1]	Percentage of peaks that are repeats[2]	% of reads covered by peaks	% of covered reads that are repeats
NRSF-ChIP	1861	1604	13.8%	12%	13.2%
Oct4-ChIP	11780	8487	27.9%	18%	26.5%
Nanog-ChIP	12840	9594	25.3%	13%	27.4%
NRSF-control	149	16	89.3%	0.5%	96.4%
ES-control	244	110	54.9%	1%	65.4%

Note:

1. Non-repeat peaks are peaks in which <50% of base pairs overlap with repeats.
2. Repeat peaks are peaks in which ≥50% of base pairs overlap with repeats

Supplementary Table 4. Relative enrichment levels for motifs discovered in NRSF ChIP-seq data

	LR≥500			LR≥500, CS≥top10%		
Motif	n_{1B}/n_{2B}	n_{1C}/n_{2C}	r_1	n_{3B}/n_{4B}	n_{3C}/n_{4C}	r_2/r_3
1	1041/827813	3644/3408619	1.18	183/195985	731/763198	0.97/1.03
2 NRSF	1612/822967	129/3401641	51.65	870/196462	40/765944	84.80/89.90
3	7250/825389	27818/3405132	1.08	3815/196197	13650/764539	1.09/1.15
4	291/825389	1482/3405132	0.81	48/196197	249/764539	0.75/0.80
5	696/822967	4358/3401641	0.66	96/196462	682/765944	0.55/0.58
6	2242/825389	8426/3405132	1.10	819/196197	2930/764539	1.09/1.15
7	1655/822967	5607/3401641	1.22	741/196462	2383/765944	1.21/1.29
8	370/825389	1695/3405132	0.90	95/196197	382/764539	0.97/1.03
9	428/820540	3555/3398152	0.50	66/196612	585/767392	0.44/0.47
10	604/830238	2715/3412109	0.91	126/195756	549/761891	0.89/0.94

Note: Motif ID in column 1 corresponds to the motif ID displayed in Supplementary Figure 8. LR=likelihood ratio between the motif model and a 3rd order background Markov model. n_{1B} = # of motif sites in binding regions; n_{2B} = total length of non-repeat base pairs in binding regions; n_{1C} = # of motif sites in matched genomic control regions; n_{2C} = total length of non-repeat base pairs in matched genomic control regions. $r_1 = (n_{1B}/n_{2B})/(n_{1C}/n_{2C})$ is the relative enrichment level of the motif. n_{3k} (k = B or C) is the number of phylogenetically conserved motif sites in binding or control regions. n_{4k} is the total length of phylogenetically conserved non-repeat base pairs in binding or control regions. $r_2 = (n_{3B}/n_{4B})/(n_{3C}/n_{4C})$. $r_3=(n_{3B}/n_{2B})/(n_{3C}/n_{2C})$. "Phylogenetically conserved" means that the corresponding phastCons score is within top 10% of the genome. Rationale for using r_1, r_2 and r_3 to characterize relative enrichment levels is discussed in ref. 37.

Supplementary Table 5. A CisGenome summary of locations of NRSF binding regions

Data	Total (%)	Inter-genic	Intra-genic	Exon	Intron	CDS	UTR	5'UTR	3'UTR	TSS up1k	TES down1k
Seq_S1	100	62.32	37.68	7.31	30.68	3.11	4.23	3.56	0.66	9.15	1.48
Seq_S2	100	59.78	40.22	8.62	31.96	3.59	5.10	4.31	0.78	10.85	1.66
Seq_S1(B+S)	100	61.36	38.64	5.80	33.10	2.69	3.12	2.31	0.81	5.05	1.45
Seq_S2(B+S)	100	60.09	39.91	5.85	34.28	2.84	3.01	2.17	0.84	5.30	1.51
Array	100	61.45	38.55	3.34	35.49	1.90	1.45	0.43	1.02	2.12	1.16
Random	100	66.03	33.98	1.78	32.29	1.03	0.78	0.13	0.65	0.68	0.64

Supplementary Table 6. NRSF motif coverage in the human genome by different datasets

	# of motif sites (LR≥500)	# of motif sites covered by peak
Tiling array	10333	1083 (10.48%)
Seq_S1	10333	1351 (13.07%)
Seq_S2	10333	1354 (13.10%)
Seq_S1 (B+S)	10333	1095 (10.60%)
Seq_S2 (B+S)	10333	1085 (10.50%)

Supplementary Table 7. Coverage of non-canonical NRSF motifs by different datasets

Dataset	No. of peaks	Peak with the canonical NRSF motif[1]	Peak with non-canonical NRSF motifs[2]	Percentage of peaks w/o the canonical motif that contain non-canonical motifs[3]
Affymetrix	7114	1001 (14.1%)	176 (2.47%)	2.9% (176/6113)
S1w100	3312	1277 (38.6%)	293 (8.85%)	14.4% (293/2035)
S1w100 (B)	3312	1223 (36.9%)	282 (8.51%)	13.5% (282/2089)
S1w100 (B+S)	1861	1051 (56.5%)	208 (11.2%)	25.7% (208/810)
S2w100	3317	1280 (38.6%)	294 (8.86%)	14.4% (294/2037)
S2w100 (B)	3317	1211 (35.5%)	281 (8.47%)	13.3% (281/2106)
S2w100 (B+S)	1794	1041 (58.0%)	208 (11.6%)	27.6% (208/753)
All three[4]	1587	933 (58.8%)	145 (9.14%)	22.2% (145/654)
Affy only[5]	5517	68 (1.23%)	29 (0.53%)	0.53% (29/5449)
S1&S2 only[6]	1385	290 (20.9%)	128 (9.24%)	11.7% (128/1095)

Note:

[1] No. of peaks with the canonical NRSF motif (percentage of peaks that contain the canonical motif);

[2] No. of peaks without the canonical NRSF motif but containing the non-canonical NRSF motifs (percentage of peaks that do not contain the canonical NRSF motif but contain the non-canonical motifs);

[3] Percentage = (No. of peaks without the canonical NRSF motif but containing the non-canonical NRSF motifs / No. of peaks without the canonical NRSF motif).

[4] Peaks detected by all three analyses (i.e., the intersection among the ChIP-chip, one-sample and two-sample ChIP-seq analyses). Here, ChIP-seq peaks before applying the boundary refinement and single strand filtering were used.

[5] Peaks detected only in ChIP-chip.

[6] Peaks detected in both the one-sample and two-sample ChIP-seq analyses but not in ChIP-chip.

Supplementary Table 8. Enrichment of NRSF half motifs in different datasets

Motif	LR≥500			LR≥500, CS≥top10% of the genome		
	n_{1B}/n_{2B}	n_{1C}/n_{2C}	r_1	n_{3B}/n_{4B}	n_{3C}/n_{4C}	r_2/r_3
Affy only						
NRSF1[1]	545/3178732	574/3737215	1.12	120/564563	137/751258	1.17/1.03
NRSF2[2]	330/3173766	401/3733385	0.97	99/564295	132/751923	1.00/0.88
All three						
NRSF1	1317/2839694	574/3737215	3.02	642/583150	137/751258	6.04/6.17
NRSF2	1059/2838216	401/3733385	3.47	505/584054	132/751923	4.93/5.03
S1&S2 only						
NRSF1	361/227312	574/3737215	10.34	180/60818	137/751258	16.23/21.60
NRSF2	472/226235	401/3733385	19.42	197/60790	132/751923	18.46/24.63

Note:

1. NRSF1 = The first half of the NRSF motif (Supplementary Fig. 9);
2. NRSF2 = The second half of the NRSF motif (Supplementary Fig. 9).

LR=likelihood ratio between the motif model and a 3rd order background Markov model. n_{1B} = # of motif sites in binding regions; n_{2B} = total length of non-repeat base pairs in binding regions; n_{1C} = # of motif sites in matched genomic control regions; n_{2C} = total length of non-repeat base pairs in matched genomic control regions. $r_1 = (n_{1B}/n_{2B})/(n_{1C}/n_{2C})$ is the relative enrichment level of the motif. n_{3k} (k = B or C) is the number of phylogenetically conserved motif sites in binding or control regions. n_{4k} is the total length of phylogenetically conserved non-repeat base pairs in binding or control regions. $r_2 = (n_{3B}/n_{4B})/(n_{3C}/n_{4C})$. $r_3=(n_{3B}/n_{2B})/(n_{3C}/n_{2C})$. “Phylogenetically conserved” means that the corresponding phastCons score is within top 10% of the genome. Rationale for using r_1, r_2 and r_3 to characterize relative enrichment levels is discussed in ref. 37.

Supplementary Table 9. Basic summary statistics of the novel motif

Summary	Human (hg17)	Mouse (mm7)
Conserved non-repeat bp in genome / total non-repeat bp in genome	239652139/1466729425=16.3%	184300142/1457016361=12.7%
Conserved sites / total sites	4543/17740 = 25.6%	3235/17940 = 18.0%
Clustered sites that are conserved / clustered sites	934/ 1674 = 55.8%	647/ 1265 = 51.2%

Note: "Conserved" means that the corresponding phastCons score is within top 10% of the genome. Two motif sites are defined to be "clustered" if they are separated by ≤ 500 bp. In CisGenome one can change the cutoff to define conservation and clustering.

Supplementary Table 10. Enrichment of the novel motif in different datasets

	LR≥500			LR≥500, CS≥top10% of the genome		
Dataset	n_{1B}/n_{2B}	n_{1C}/n_{2C}	r_1	n_{3B}/n_{4B}	n_{3C}/n_{4C}	r_2/r_3
Sox2-human-promoter array	152/307344	1738/6089803	1.73	73/102869	653/1557900	1.69/2.22
Nanog-human-promoter array	174/484667	1296/6456504	1.79	80/157096	515/1640948	1.62/2.07
Oct4- mouse-genome-wide ChIP-PET	141/1073118	759/6477817	1.12	42/220674	180/1118222	1.18/1.41
Nanog-mouse-genome-wide ChIP-PET	90/737726	625/5774175	1.13	27/143248	157/967126	1.16/1.35

Note: see Supplementary Table 4 for meanings of each column.

Supplementary Table 11. Performance of ChIP-chip peak detection algorithms

Sample[1]	Analysis[2]	AUC[3]	E-O distance[4]	#Top sites	#TP[5]	#FN[5]	#FP[5]
UnAmp	Affymetrix_Struhl_6_TileMap	0.63	0	92	71	29	21
UnAmp	Affymetrix_Struhl_6_MAT	0.59	0	100	66	34	34
UnAmp	Affymetrix_Struhl_6_TAS	0.44	-15	93	64	36	29
UnAmp	Affymetrix_Struhl_Gingeras_3_TileMap	0.55	7	100	62	38	38
UnAmp	Affymetrix_Struhl_Gingeras_3_MAT	0.54	-7	100	62	38	38
UnAmp	Affymetrix_Struhl_Gingeras_3_TiMAT	0.51	-15	98	60	40	38
UnAmp	Affymetrix_Struhl_DFCI_3_TileMap	0.62	-9	100	67	33	33
UnAmp	Affymetrix_Struhl_DFCI_3_MAT	0.54	-3	100	62	38	38
UnAmp	Agilent_Myers_WI_5_TileMap	0.68	26	100	81	19	19
UnAmp	Agilent_Myers_WI_5_WA	0.45	-15	100	52	48	48
UnAmp	Agilent_Myers_3_TileMap	0.59	24	100	68	32	32
UnAmp	Agilent_Myers_3_Splitter	0.40	42	98	52	48	46
UnAmp	Agilent_Myers_3_WA	0.36	-10	100	40	60	60
UnAmp	Agilent_Myers_3_MA2C	0.33	44	92	43	57	49
UnAmp	Agilent_WI_2_TileMap	0.80	4	99	86	14	13
UnAmp	Agilent_WI_2_Splitter	0.64	23	100	77	23	23
UnAmp	Agilent_WI_2_WA	0.64	6	100	79	21	21
UnAmp	Agilent_WI_2_MA2C	0.59	27	100	75	25	25
UnAmp	Agilent_WI_2_ADM-1	0.49	-57	86	73	27	13
UnAmp	NimbleGen_Green_4_TileMap	0.79	8	100	90	10	10
UnAmp	NimbleGen_Green_4_TAMALPAISgenerous	0.71	13	100	83	17	17
UnAmp	NimbleGen_Green_4_Permutation	0.66	3	88	77	23	11
UnAmp	NimbleGen_Green_4_Splitter	0.64	4	97	86	14	11
UnAmp	NimbleGen_Green_4_TAMALPAISstrict	0.56	1	61	57	43	4
UnAmp	NimbleGen_Green_4_MA2C	0.54	21	100	83	17	17
UnAmp	NimbleGen_Green_4_TileScope	0.53	10	100	86	14	14
UnAmp	NimbleGen_Green_4_ACME	0.36	53	100	78	22	22
UnAmp	NimbleGen_Snyder_3_TileMap	0.76	0	100	79	21	21
UnAmp	NimbleGen_Snyder_3_Splitter	0.69	19	100	80	20	20
UnAmp	NimbleGen_Snyder_3_Wavelet	0.55	0	66	62	38	4
UnAmp	NimbleGen_Snyder_3_TileScope	0.52	-8	89	77	23	12
Amp	Affymetrix_Brown_LM_3_TileMap	0.45	3	62	51	47	11
Amp	Affymetrix_Brown_LM_3_MAT	0.42	-10	61	46	52	15
Amp	Affymetrix_Brown_LM_3_Splitter	0.27	0	44	35	63	9
Amp	Affymetrix_Struhl_RP_3_MAT	0.16	0	98	29	69	69
Amp	Affymetrix_Struhl_RP_3_TileMap	0.13	20	98	26	72	72
Amp	Affymetrix_Struhl_RP_3_TiMAT	0.12	37	98	20	78	78
Amp	Agilent_WI_LM_2_ADM-1	0.56	-59	66	59	39	7
Amp	Agilent_WI_LM_2_TileMap	0.52	-18	98	65	33	33
Amp	Agilent_WI_LM_2_WA	0.44	11	98	61	37	37
Amp	Agilent_WI_LM_2_MA2C	0.35	34	98	50	48	48
Amp	NimbleGen_Farnham_WGA_3_TAMALPAISgenerous	0.62	-1	98	76	22	22
Amp	NimbleGen_Farnham_WGA_3_TileMap	0.62	18	98	85	13	13
Amp	NimbleGen_Farnham_WGA_3_MA2C	0.57	9	98	81	17	17
Amp	NimbleGen_Farnham_WGA_3_TileScope	0.57	-5	95	82	16	13
Amp	NimbleGen_Farnham_WGA_3_Splitter	0.52	8	98	87	11	11
Amp	NimbleGen_Farnham_WGA_3_Permu	0.45	1	73	65	33	8
Amp	NimbleGen_Farnham_WGA_3_TAMALPAISstrict	0.4	0	44	41	57	3
Amp	NimbleGen_Farnham_WGA_3_ACME	0.33	63	98	74	24	24

Note:
TileMap results were obtained by applying CisGenome to the spike-in data. Results for other algorithms were provided by ref. 41.

1. Sample: “UmAmp” means undiluted spike-in sample; “Amp” means diluted spike-in sample.
2. Analysis: undiluted data sets are labeled by [Array platform]_[Lab generating the data]_[Number of replicates]_[Algorithm] ; diluted data sets are labeled by [Array platform]_[Lab generating the data]_[Amplification protocol]_[Number of replicates]_[Algorithm].
3. AUC = Area under the ROC-like curve. A bigger AUC represents a better overall performance of the algorithm.
4. E-O distance = Distance between the chosen cutoff and the optimal cutoff. A negative distance represents a conservative cutoff, and a positive distance represents a loose cutoff.
5. #TP, #FN, #FP = Number of true positives, false negatives and false positives for the top sites.

Supplementary Table 12. A comparison of representative software tools for ChIP data analyses

	ChIP-chip peak detection	Evaluation of statistical significance for ChIP-chip peaks	ChIP-seq peak detection	Evaluation of FDR for one-sample ChIP-seq	Evaluation of FDR for two-sample ChIP-seq	Peak-gene association	Statistical summary of location/conservation	Large-scale genomic sequence manipulation	De novo motif discovery
CisGenome	+	+	+	+	+	+	+	+	+
TAS[13]	+	+							
MAT[11]	+	+							
Tilescope[21]	+	+				+			
CPF[4]			+						
GeneTrack[29]	+		+						
QuEST[30]			+		+				
SISSRs[31]			+	+					
MEME[42]									+
MDScan[25]									+
CEAS[28]						+	+	+	+
Galaxy[43]						+	+	+	
SignalMap*	+								
IGB**									
UCSC[33]								+	
Ensembl[34]								+	
WebLogo[44]									

	Mapping motif to user-specified genomic regions	Motif enrichment analysis based on matched controls	Genomic region and signal visualization	Motif visualization	GUI	Stand-alone & run locally	Web-based	Allow customization for addressing different questions
CisGenome	+	+	+	+	+	+		+
TAS					+	+		
MAT						+		
Tilescope					+		+	
CPF						+		
GeneTrack			+			+		
QuEST						+		
SISSRs						+		
MEME					+	+	+	
MDScan					+	+	+	
CEAS				+	+		+	
Galaxy			+		+		+	+
SignalMap			+		+	+		
IGB			+		+	+		
UCSC			+		+		+	+
Ensembl			+		+		+	+
WebLogo				+	+		+	

Notes:

* developed by NimbleGen; ** developed by Affymetrix.

According to the original publication, GeneTrack can be used to handle ChIP-chip data. However, similar to its ChIP-seq analysis function, the software does not provide error rate estimates for the ChIP-chip analysis. Moreover, in the original publication, no example was provided to illustrate this ability, and no rigorous tests and systematic evaluation have been presented for the ChIP-chip analysis function. Therefore it remains unclear how its ChIP-chip analysis function performs compared to the other existing ChIP-chip analysis algorithms.

“Stand-alone and run locally” means that the major analysis and visualization functions provided by the tool are self-contained and can be used without the need to transfer data over the internet during the analysis procedure. For example, Galaxy can be installed and run locally, but it uses the UCSC genome browser to display the genomic data which requires transferring the data over the internet, therefore it is not a fully stand-alone software tool in the context of ChIP-chip/ChIP-seq data analysis/visualization.

The comparisons show that CisGenome covers a broad spectrum of functionalities. Only representative software tools were listed here. For example, there are many other ChIP-chip peak detection methods that are compared in **Supplementary Data 6** but not listed here. Similar to TAS, MAT and Tilescope, they typically only handle ChIP-chip data and do not support ChIP-seq analysis as well as downstream sequence/annotation/motif analyses. Also, for *de novo* motif analyses, there are dozens of other tools reviewed and compared in ref. 69-71. In general, they have the same limitations as MEME and MDSCAN that are listed here.

Supplementary Table 13. Correlation of NRSF ChIP and control read number in 100bp windows

		ChIP read		Percentage of windows with ≥1 ChIP read
		=0	≥1	
Control read	=0	2.70 M	0.14 M	4.8%
	≥1	0.22 M	0.02 M	10.1%
Percentage of windows with ≥1 control read		7.4%	14.9%	

Note: number of windows in each category is shown in the unit of million. Chi-square test for correlation yields p-value<1e-10. For windows with 0 control read, 4.8% contain ≥1 ChIP read. For windows with ≥1 control read, 10.1% contain ≥1 ChIP read. Thus, windows that are more likely to contain control reads are also more likely to contain reads in the ChIP sample. This is an analysis complementary to Supplementary Fig. 17c,d. When window size is small, the estimate of read occurrence rate in a window is unstable, and most genomic windows contain no read. Therefore, instead of comparing the read occurrence rate directly (as in Supplementary Fig. 17c,d), the table here compares whether a window that contains control reads are more likely to contain ChIP reads. Together with Supplementary Fig. 17c,d and Supplementary Fig. 5, the results suggest that the background sampling rate of the control sample and the background sampling rate of the ChIP sample at the same loci are correlated at the resolution (w=100-200bp) usually used in the two-sample analyses.

Supplementary Table 14. Length distribution of NRSF ChIP-seq binding regions detected using different window size W

Analysis criteria	No. of NRSF	Percentiles of region length (bp)				
	motif/1kb	10	25	50	75	90
S1w25	3.1606	30	40	56	137	232
S1w50	1.9598	60	86	149	286	448
S1w100	1.2615	122	173	269	444	598
S1w200	0.6217	231	298	403	595	793
S1w25 (B)	6.8840	29	29	31	50	80
S1w50 (B)	6.6882	29	30	42	71	96
S1w100 (B)	5.5388	29	30	60	82	113
S1W200 (B)	2.4235	29	36	96	146	183
S1w25 (B+S)	10.9443	30	40	60	82	103
S1w50 (B+S)	8.6670	30	48	66	87	112
S1w100 (B+S)	6.9799	41	59	73	90	122
S1w200 (B+S)	3.8031	80	111	138	161	184
S2w25	3.3049	29	40	63	144	236
S2w50	2.0469	59	85	152	294	450
S2w100	1.2770	116	161	261	445	604
S2w200	0.7065	227	293	423	605	794
S2w25 (B)	7.4139	29	29	33	51	81
S2w50 (B)	7.2134	29	30	43	70	95
S2w100 (B)	5.5268	29	30	59	85	119
S2w200 (B)	2.4832	29	36	101	156	215
S2w25 (B+S)	11.3630	30	41	60	81	101
S2w50 (B+S)	9.3410	31	48	65	86	109
S2w100 (B+S)	7.3109	40	57	73	94	125
S2w200 (B+S)	3.8733	59	100	137	166	199

Note: S1 = one-sample analysis; S2 = two-sample analysis; B = boundary refinement; S = single strand filtering; w100 means window size w = 100 bp.

Supplementary Table 15. Motif coverage of NRSF ChIP-seq binding regions detected using different window size W

Sample	W	Cutoff[1]	Initial regions[2]	Refine boundary (B)	Boudary+Strand (B+S)
S1	25	7	3581 (1105, 30.9%)	3581 (1067, 29.8%)	1177 (804, 68.3%)
S1	50	7	3240 (1212, 37.4%)	3240 (1163, 35.9%)	1564 (956, 61.1%)
S1	100	8	3312 (1277, 38.6%)	3312 (1223, 36.9%)	1861 (1051, 56.5%)
S1	200	8	4961 (1385, 27.9%)	4961 (1294, 26.1%)	2003 (1092, 54.5%)
S2	25	7	3310 (1105, 33.4%)	3310 (1071, 32.4%)	1157 (804, 69.5%)
S2	50	7	3046 (1212, 39.8%)	3046 (1162, 38.2%)	1507 (954, 63.3%)
S2	100	8	3317 (1280, 38.6%)	3317 (1211, 35.5%)	1794 (1041, 58.0%)
S2	200	9	4264 (1351, 31.7%)	4264 (1202, 28.2%)	1940 (1028, 53.0%)

Note: 1. Cutoff n is the minimal number of reads required to declare a window to be significant. It was chosen to control $\mathrm{FDR} \leq 10\%$. 2. For each analysis the number of binding regions x_1, the number of regions that contain ≥ 1 NRSF motif x_2, and the percentage $y=x_2/x_1$ are reported in the format x_1 (x_2, y).