nature biotechnology

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

Hongkai Ji[1], Hui Jiang[2], Wenxiu Ma[3], David S Johnson[4,8], Richard M Myers[5] & Wing H Wong[6,7]

We present CisGenome, a software system for analyzing genome-wide chromatin immunoprecipitation (ChIP) data. CisGenome is designed to meet all basic needs of ChIP data analyses, including visualization, data normalization, peak detection, false discovery rate computation, gene-peak association, and sequence and motif analysis. In addition to implementing previously published ChIP–microarray (ChIP-chip) analysis methods, the software contains statistical methods designed specifically for ChIP sequencing (ChIP-seq) data obtained by coupling ChIP with massively parallel sequencing. The modular design of CisGenome enables it to support interactive analyses through a graphic user interface as well as customized batch-mode computation for advanced data mining. A built-in browser allows visualization of array images, signals, gene structure, conservation, and DNA sequence and motif information. We demonstrate the use of these tools by a comparative analysis of ChIP-chip and ChIP-seq data for the transcription factor NRSF/REST, a study of ChIP-seq analysis with or without a negative control sample, and an analysis of a new motif in Nanog- and Sox2-binding regions.

Chromatin immunoprecipitation followed by either genome tiling array analysis (ChIP-chip)[1–3] or massively parallel sequencing (ChIP-seq)[4–10] enables transcriptional regulation to be studied on a genome-wide scale (**Supplementary Fig. 1** online). By systematically identifying protein-DNA interactions of interest, studies using these technologies provide information on *cis*-regulatory circuitry underlying various cellular processes. However, analysis of the massive and heterogeneous datasets from these studies poses several challenges, including effective data visualization, seamless connection of low-level (close to raw data) and high-level (close to answering biological questions) analysis, integration of data from multiple technological platforms, and flexibility to customize the analysis so that specific biological questions can be addressed. Although there are several recently developed programs[11–31] that target some of the individual steps, an integrated tool that can satisfy all basic needs in ChIP data analyses is not yet available (see **Supplementary Notes** online).

We developed a set of methods to meet these needs in ChIP data analyses and implemented them in an integrated software package (**Fig. 1**). CisGenome provides a wide range of functionalities for ChIP data analyses that can be accessed through a menu-driven system in a graphic user interface (GUI). The results are automatically linked to the CisGenome browser, which is designed for data visualization. CisGenome is a standalone system that bench biologists can use to analyze their own data locally on personal computers. At the same time, most CisGenome functionalities can also be accessed in a command-line manner. This modular design allows computational biologists to build large batch jobs for customized analyses on computer servers.

RESULTS

CisGenome basic functionalities are listed below.

Data processing and binding region identification

Finding regions harboring protein-DNA association is the critical first step of ChIP data analyses. CisGenome can detect these regions (or peaks) from raw tiling array probe intensities or mapped sequence reads. For example, using the GUI, one can directly load Affymetrix CEL and BPMAP tiling array data, examine raw array images to detect hybridization artifacts, normalize data across different arrays and then detect binding regions (**Supplementary Fig. 2a–c** online). CisGenome can also take as input the binding regions or peak scores obtained from other preprocessing programs, such as MAT[11] for ChIP-chip and QuEST[30] for ChIP-seq data. CisGenome uses TileMap[12] for internal ChIP-chip peak calling and false discovery rate (FDR) estimation (**Supplementary Methods** online).

Visualization of results

Convenient visualization of raw and processed data provides a useful way to assess data quality and generate scientific hypotheses. In CisGenome, the peak signals, including fold changes and summary statistics, are reported in tables and linked to the CisGenome browser. Using the browser, one can visualize at the probe or read level

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA. [2]Institute for Computational and Mathematical Engineering, Stanford University, Durand Building, 496 Lomita Mall, Stanford, California 94305, USA. [3]Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, California 94305, USA. [4]Department of Genetics, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, California 94305, USA. [5]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, Alabama 35806, USA. [6]Department of Statistics, [7]Department of Health Research and Policy, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, California 94305, USA. [8]Present address: Gene Security Network, Inc., 1442 Cortland Avenue, San Francisco, California 94110, USA. Correspondence should be addressed to W.H.W. (whwong@stanford.edu).

Received 11 June; accepted 3 October; published online 2 November 2008; doi:10.1038/nbt.1505

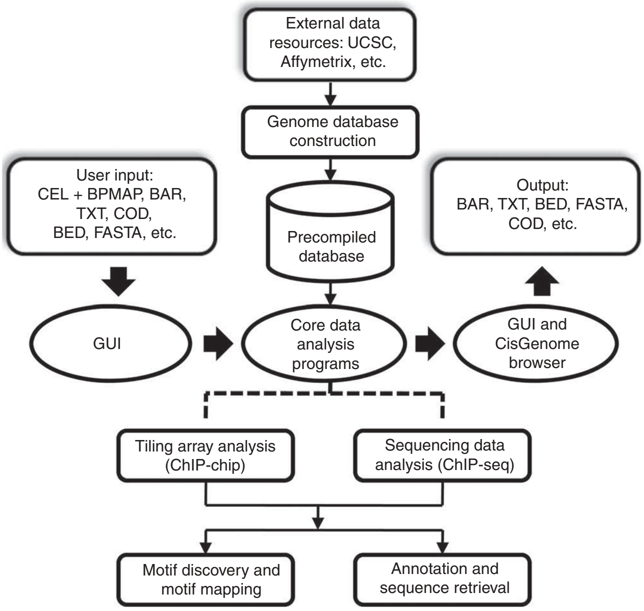

External data resources: UCSC, Affymetrix, etc.
Genome database construction
User input: CEL + BPMAP, BAR, TXT, COD, BED, FASTA, etc.
Precompiled database
Output: BAR, TXT, BED, FASTA, COD, etc.
GUI
Core data analysis programs
GUI and CisGenome browser
Tiling array analysis (ChIP-chip)
Sequencing data analysis (ChIP-seq)
Motif discovery and motif mapping
Annotation and sequence retrieval

Figure 1 The basic framework of CisGenome. CisGenome contains three core components: a GUI, the built-in CisGenome browser and a set of underlying data analysis algorithms. The GUI allows users to load raw data and choose specific analysis functions. Core programs carry out the analysis, and results displayed in the CisGenome browser can be exported in various formats. Precompiled genome databases are required to support analyses involving sequence and gene annotation information. CisGenome contains functions to construct such databases from standard external data resources. Databases for a few commonly used species can be downloaded directly from the CisGenome website.

data together with gene structures, conservation scores and DNA sequences (**Supplementary Fig. 2d**). One can freely zoom in and out, move left and right, search for genes and regions, or add and delete annotation tracks. By clicking a location of interest, one can link to external resources such as NCBI[32], UCSC[33] and Ensembl[34] to obtain more comprehensive information. The CisGenome browser also supports visualization of raw array images and sequence logos of motifs. The memory requirement (~64 M) is minimal. This built-in browser makes it easy and efficient to visualize millions of data points without the need to transfer them to Web services such as the UCSC genome browser. This often becomes inefficient in large-scale analyses.

Statistical summaries

Predicted binding regions need to be linked to gene annotations to provide functional contexts for interpreting the results. Tools for establishing such links automatically are crucial for efficient analysis of thousands of predictions. The CisGenome GUI enables one to associate binding regions with neighboring genes and to study statistical properties of the binding regions in relation to various genome annotation features. For example, one can extract the frequency of regions found in exons, introns and untranslated regions and summarize the conservation level of each individual binding region (**Supplementary Fig. 2e**).

Motif analysis

Many transcription factors recognize specific DNA sequence patterns (that is, motifs). Finding motifs from ChIP data and locating them in the genome will provide clues on how transcriptional regulatory programs are encoded in DNA. CisGenome contains many functions related to sequence and motif analyses. It can be used to retrieve DNA sequences on binding-regions, map transcription factor binding motifs to the genome, and search for novel motifs[35] and *cis*-regulatory modules[36]. A *de novo* motif search may return multiple motifs; CisGenome identifies the functionally relevant ones by comparing the occurrence rates of the motifs in binding regions to those in matching genomic control regions[37] (**Supplementary Fig. 2e–h** and **Supplementary Methods**).

Support for different species

Although CisGenome currently supports only human, mouse, *Drosophila* and *Arabidopsis* for species-dependent analyses (for example, peak-gene association), users can add support for other species (**Supplementary Methods**).

Modular structure

CisGenome has a modular design so that most of its functions can be accessed in command mode and from the GUI. The command mode functions can be conveniently embedded into users' own programs. Interfaces that allow users to link their own programs to the CisGenome browser are provided. Interfaces that allow users to plug their own tools into the CisGenome GUI are being developed.

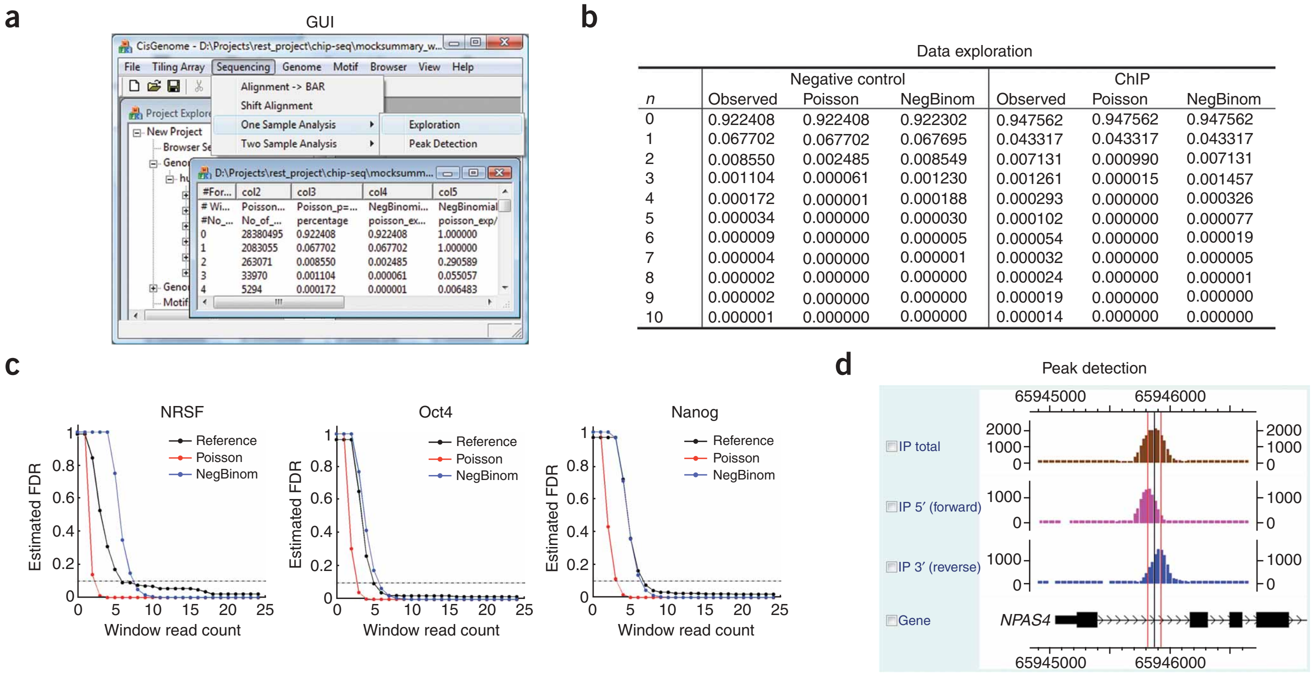
Open source and user support

The program download, frequently asked questions document, file formats, tutorial and user manual can be found at http://biogibbs.stanford.edu/~jihk/CisGenome/index.htm. Developing language and operating systems are discussed in **Supplementary Methods**. We provide source codes to enable customization by users.

Processing of ChIP-seq data

Compared to ChIP-chip, the recently developed ChIP-seq technology is able to provide higher resolution and more comprehensive genome coverage for identification of protein-DNA interactions. The processing of ChIP-seq data, however, is still in its infancy. CisGenome can handle data from two types of designs common in ChIP-seq experiments (see Methods and **Fig. 2**): one-sample analysis, where only a ChIP sample is sequenced[5,9], and two-sample analysis[4,6,8,10], where both a ChIP sample and a negative control sample are sequenced. In one-sample analysis, CisGenome scans the genome with a sliding window and identifies regions with read counts greater than a user-chosen cutoff for *bona fide* binding regions. FDRs are estimated by modeling the read count in nonbinding windows using a negative binomial distribution. In contrast to the constant rate assumed in the widely used Poisson background model, the negative binomial model allows the background rate of occurrence of the reads to vary across the genome and to have a more flexible gamma distribution. For many datasets, the negative binomial model provides a much better fit to the data than does the Poisson model (**Fig. 2b,c**). A systematic evaluation of the method is provided in **Supplementary Data 1**, **Supplementary Figs. 3–7** and **Supplementary Tables 1–3** online.

In two-sample analysis, where a negative control sample is also available, CisGenome uses a conditional binomial model to identify regions in which the ChIP reads are significantly enriched relative to the control reads. Windows passing a user-specified FDR cutoff are used to generate predicted binding regions. Both one- and two-sample analyses use the directionality of reads to refine peak boundaries and filter out low-quality predictions. These are provided as two post-processing options—boundary refinement and single-strand filtering (**Fig. 2d**).



#For...	col2	col3	col4	col5
# Wi...	Poisson...	Poisson_p=...	NegBinomi...	NegBinomial
#No_...	No_of_...	percentage	poisson_ex...	poisson_exp/
0	28380495	0.922408	0.922408	1.000000
1	2083055	0.067702	0.067702	1.000000
2	263071	0.008550	0.002485	0.290589
3	33970	0.001104	0.000061	0.055057
4	5294	0.000172	0.000001	0.006483

n	Negative control			ChIP		
	Observed	Poisson	NegBinom	Observed	Poisson	NegBinom
0	0.922408	0.922408	0.922302	0.947562	0.947562	0.947562
1	0.067702	0.067702	0.067695	0.043317	0.043317	0.043317
2	0.008550	0.002485	0.008549	0.007131	0.000990	0.007131
3	0.001104	0.000061	0.001230	0.001261	0.000015	0.001457
4	0.000172	0.000001	0.000188	0.000293	0.000000	0.000326
5	0.000034	0.000000	0.000030	0.000102	0.000000	0.000077
6	0.000009	0.000000	0.000005	0.000054	0.000000	0.000019
7	0.000004	0.000000	0.000001	0.000032	0.000000	0.000005
8	0.000002	0.000000	0.000000	0.000024	0.000000	0.000001
9	0.000002	0.000000	0.000000	0.000019	0.000000	0.000000
10	0.000001	0.000000	0.000000	0.000014	0.000000	0.000000

Figure 2 ChIP-seq data processing. (**a**) The GUI can be used to explore and analyze ChIP-seq data. (**b**) In data exploration, parametric models are fitted to describe the distribution of read count *n* in background windows. Both negative control samples and the lower end of ChIP samples can be fitted well by the negative binomial (NegBinom) model, whereas the Poisson model generally cannot provide satisfactory fitting. Fitting to the NRSF data are shown as an example. (**c**) In one-sample analyses of NRSF[4], Oct4 (ref. 10) and Nanog[10] data, FDR estimates based on the negative binomial and Poisson models were compared to model-independent reference FDRs. The reference FDRs were obtained by incorporating information from negative control samples and were defined as number of predictions in the control sample divided by number of predictions in the corresponding ChIP sample with equal amount of reads. (**d**) Peak detection results can be visualized using the CisGenome browser. 5′ Reads that are aligned to the forward strand of the genome (pink) and 3′ reads aligned to the reverse complement strand of the genome (blue) are usually shifted away from each other and form two separate peaks resulting from the nature of sequencing[49] (**Supplementary Fig. 1**). CisGenome uses the modes (red vertical lines) of the 5′ and 3′ peaks to refine the boundaries of binding regions (boundary refinement) and reports the center (black vertical line) as well. CisGenome can also filter out low-quality binding regions if 5′ and 3′ peaks do not show up as a pair (single-strand filtering).

Comparative analysis of NRSF ChIP-chip and ChIP-seq data

To demonstrate the basic functions provided by CisGenome, we analyzed whole-genome ChIP-chip and ChIP-seq datasets generated for the neuron-restrictive silencer factor (NRSF, also known as REST)[38,39] in Jurkat cells (see Methods). NRSF is a zinc finger repressor that negatively regulates many neuronal genes in stem and progenitor cells and nonneuronal cell types. Following the steps shown in **Supplementary Figure 2**, we identified 7,114 binding regions at a 10% FDR level (median length, 616 bp) from the ChIP-chip data. The NRSF motif was successfully discovered by *de novo* motif discovery and had the highest enrichment level among all the discovered motifs.

We then applied both one- and two-sample analyses to the corresponding ChIP-seq data. One-sample analysis identified 3,312 NRSF binding regions before postprocessing (FDR $\leq$ 10%; median length, 269 bp), from which the NRSF motif was recovered by *de novo* motif discovery (**Supplementary Fig. 8** and **Supplementary Table 4** online). Motif mapping (**Table 1**) showed that among the initial 3,312 peaks, 1,277 contained $\geq$1 NRSF motif. Boundary refinement greatly

Table 1 Summary of NRSF ChIP-chip and ChIP-seq binding regions

				Region length percentile (bp)				
Data and analysis method	Peaks	Peaks with NRSF motif	Motifs per kb	10	25	50	75	90
Affymetrix-TileMap	7,114	1,001 (14.1%)	0.15	211	323	616	1,274	2,311
Seq-S1w100	3,312	1,277 (38.6%)	1.26	122	173	269	444	598
Seq-S1w100 (B)	3,312	1,223 (36.9%)	5.54	29	30	60	82	113
Seq-S1w100 (B + S)	1,861	1,051 (56.5%)	6.98	41	59	73	90	122
Seq-S2w100	3,317	1,280 (38.6%)	1.28	116	161	261	445	604
Seq-S2w100 (B)	3,317	1,211 (35.5%)	5.53	29	30	59	85	119
Seq-S2w100 (B + S)	1,794	1,041 (58.0%)	7.31	40	57	73	94	125

S1w100, one-sample analysis for ChIP-seq data (window length $w = 100$ bp); S2w100, two-sample analysis for ChIP-seq data (window length $w = 100$ bp); B, applying boundary refinement; S, applying single-strand filtering. The choice of window size $w = 100$ bp represents a tradeoff between sensitivity and specificity (see Methods). Methods for motif mapping are described in **Supplementary Methods**. A likelihood ratio $\geq$500 was used as the cutoff to define NRSF motif sites. To facilitate a fair comparison between different datasets, the TRANSFAC[40] NRSF motif M00256 was used in the motif mapping. Using the NRSF motif recovered from *de novo* motif discovery did not change the results qualitatively (data not shown).

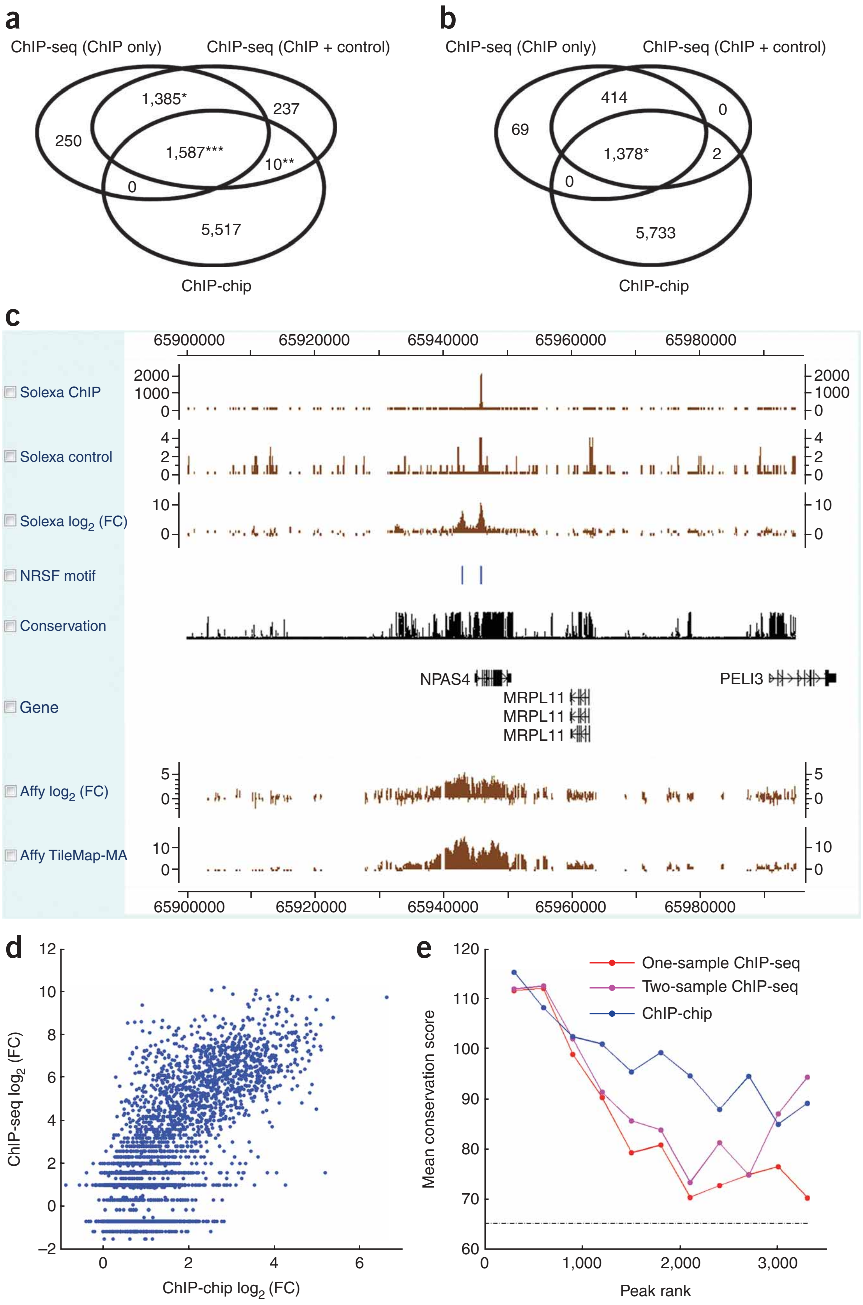


Figure 3 Comparisons between NRSF ChIP-seq and ChIP-chip data. (**a**) Overlap among ChIP-chip and ChIP-seq binding regions before applying boundary refinement and single-strand filtering. *Because a peak from one dataset can overlap multiple peaks from another dataset, the intersection involved 1,385 one-sample and 1,387 two-sample ChIP-seq peaks; **10 ChIP-chip peaks and 22 two-sample ChIP-seq peaks; ***1,587 ChIP-chip peaks and 1,677 one-sample and 1,671 two-sample ChIP-seq peaks. (**b**) Overlap among ChIP-chip and ChIP-seq binding regions after applying postprocessing to ChIP-seq data. *1,378 ChIP-seq and 1,379 ChIP-chip peaks overlapped. (**c**) Visual comparison of ChIP-seq and ChIP-chip signals in CisGenome browser. FC, fold change; MA, moving average. (**d**) Using CisGenome, the NRSF motif was mapped to the human genome, and log$_2$ fold changes for IP over control were extracted for the motif sites from both ChIP-chip and ChIP-seq. Comparison of these site-level signals revealed a strong correlation between ChIP-chip and ChIP-seq ($\rho = 0.73$). The CisGenome functions used here can be applied to construct genome-wide tissue-specific activity maps of transcription factor binding motifs in future studies. (**e**) Conservation levels of ChIP-chip and ChIP-seq binding regions were higher than the corresponding conservation level of randomly chosen nonrepeat genomic regions (dotted line). The ranked binding regions were grouped into tiers of 300. Mean phastCons[48] conservation score was computed for each tier (see Methods). Data shown characterize the conservation at the binding region level rather than motif site level. Results were obtained before postprocessing. Applying postprocessing to ChIP-seq produced similar results (data not shown).

reduced the median length of these 3,312 regions (from 269 to 60 bp), with only a slight decrease in the number of NRSF site–containing regions (from 1,277 to 1,223). The further step of single-strand filtering reduced the number of regions from 3,312 to 1,861 but retained most (1,051 of 1,223) of the NRSF site–containing regions. Even before postprocessing, there were substantially more NRSF sites in ChIP-seq regions (1.26/kb) than in ChIP-chip regions (0.15/kb). The rate was further increased after each step in the postprocessing (to 5.54/kb after boundary refinement and 6.98/kb after single-strand filtering). This increase of signal-to-noise ratio could potentially improve the possibility of *de novo* discovery of weak unknown motifs. Predictions with a higher resolution can also provide more focused targets for future experimental studies, such as those seeking the minimal *cis*-regulatory elements sufficient and necessary to drive target gene expression.

Using both the ChIP and negative control samples, two-sample analysis identified 3,317 initial binding regions (FDR ≤ 10%; median length, 261 bp). Postprocessing reduced the median region length to ~65 bp and the number of regions to 1,794 (**Table 1**). After postprocessing, there was a 96% overlap between the peaks detected in one-sample analysis and those detected in two-sample analysis (**Fig. 3a,b**).

Comparisons between array and sequencing technologies showed that peak signals produced by the two platforms had a clear correlation (**Fig. 3c,d**). However, peaks called in the tiling array analysis were generally longer than the corresponding ChIP-seq peaks, and the array peaks were less likely to contain the NRSF motif (**Table 1**). In all studies, binding regions were more likely to be located near promoters (**Supplementary Table 5** online). They were substantially more conserved than randomly selected genomic regions (**Fig. 3e**), and they were able to cover 10–13% of all NRSF motif sites in the genome (**Supplementary Table 6** online). Notably, 5,517 (78%) of 7,114 array peaks did not overlap with any ChIP-seq peak (**Fig. 3a**). We carried out motif analyses to investigate whether these regions represent noise in the tiling array technology or signals missed by ChIP-seq. *De novo* motif discovery was not able to recover the NRSF motif from the array-specific peaks, and only 68 (1.23%) of 5,517 array-specific peaks

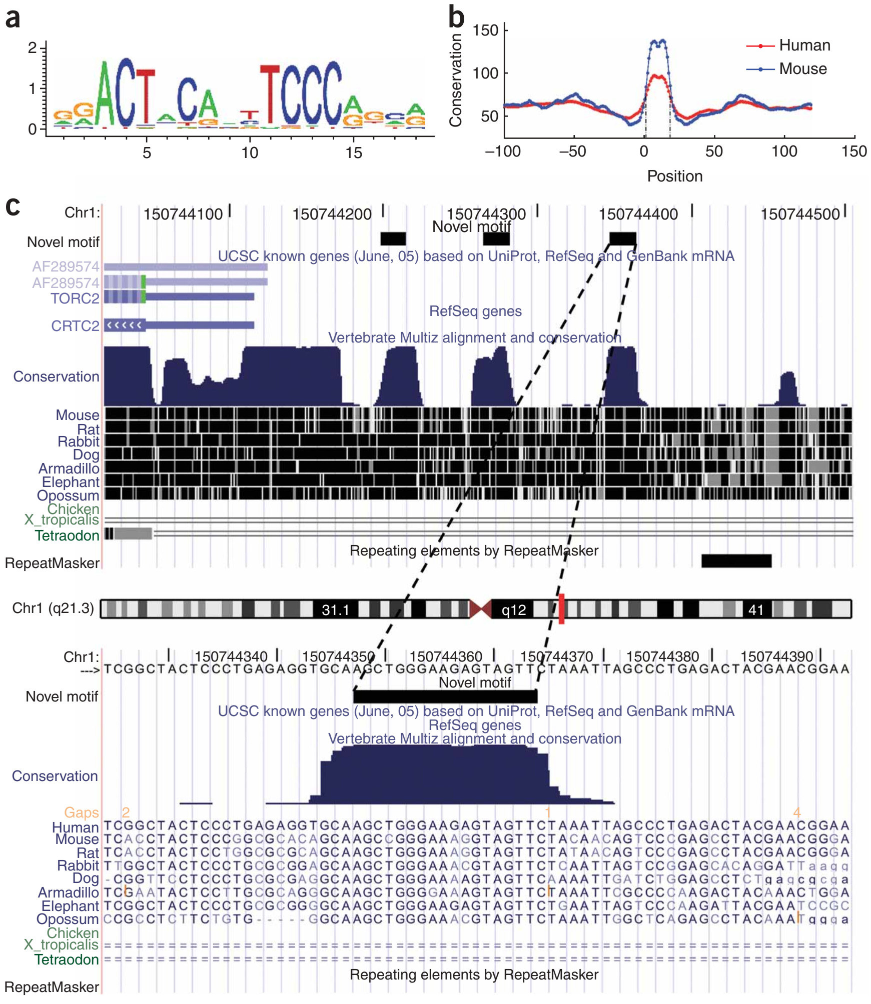


Figure 4 Analysis of a novel motif in Sox2 and Nanog binding regions. (**a**) Sequence logo of the motif visualized using the CisGenome browser. (**b**) Mean phastCons scores for the motif and flanking positions were extracted using CisGenome (**Supplementary Fig. 12d**). The score drops sharply at the motif boundaries, which are indicated by two dotted vertical lines. (**c**) A typical example of clustered motif sites. Sites are indicated by the black blocks in the novel motif track. They coincide well with conserved genomic elements. The UCSC genome browser was used to show that CisGenome allows users to link to external Web resources (**Supplementary Fig. 12c**).

contained ≥1 NRSF motif. For comparison, 1,001 (14.1%) of all 7,114 array peaks, 290 (20.9%) of the 1,385 peaks common to the ChIP-seq analyses but not found by arrays, and 933 (58.8%) of the 1,587 peaks common to all three analyses contained the motif. As analyses using noncanonical NRSF motifs yielded similar results (**Supplementary Data 2**, **Supplementary Fig. 9** and **Supplementary Tables 7** and **8** online), the array-specific peaks in this example are unlikely to represent true signals.

Merits and limitations of one-sample ChIP-seq analyses

One-sample design has been used in many ChIP-seq experiments[5,9]. It allows more biological contexts to be analyzed within a fixed sequencing budget. To study the merits and limitations of this design, we analyzed ChIP-seq data for two additional transcription factors, Oct4 and Nanog, which are crucial regulators for self-renewal and pluripotency of embryonic stem cells[10]. Again, there was good agreement between one-sample and two-sample analyses after postprocessing, with 96% concordance in the case of Oct4 and 83% in the case of Nanog (**Supplementary Data 3** and **Supplementary Figs. 10** and **11** online). These examples suggest that under certain conditions, a one-sample experiment can provide a cost-effective alternative to the two-sample experiment, albeit perhaps at the expense of some specificity.

To gain a better understanding of limitations of one-sample analysis, we applied it to negative control samples. Although no peaks were expected, a small number of peaks were reported at the 10% FDR level (**Supplementary Table 3**). This was caused by the residual background variation that the negative binomial model was not able to explain (Poisson model performed even worse; **Fig. 2b**). Systematic evaluation using simulated spike-in data showed that, although the one-sample analysis can provide reasonable FDR estimates when the overall binding signal is strong, the method may underestimate the real FDR significantly when the overall binding in the sample is weak (**Supplementary Data 1**). Fortunately, poor peak reliability and problematic FDR estimation can often be diagnosed through several criteria, such as highly repeat-rich predictions, predictions covering a low percentage of reads, and lack of motif enrichment (**Supplementary Data 1**). We recommend using two-sample experiments whenever it is affordable or when little is known about the transcription factor. When cost constraints necessitate one-sample analyses, a negative binomial rather than Poisson background model should be used to exclude background noise, and prediction quality should be evaluated using multiple criteria as described above. CisGenome is designed to support these types of analyses.

Analysis of a novel motif in Sox2 and Nanog binding regions

The basic functionalities of CisGenome can be used in combination to address many different biological questions. For example, *de novo* discovery from peak regions may yield new sequence motifs. Bench biologists can use the motif mapping and statistical summary functions to systematically evaluate the functional implications of these motifs. As an example, we studied a novel motif discovered from a Sox2 and Nanog ChIP-chip dataset on human promoter arrays[2]. This motif (**Fig. 4a**), identified by *de novo* motif discovery along with the *bona fide* Oct4 and Sox2 motifs[37], is highly sequence specific but does not correspond to any known motif stored in TRANSFAC[40] (**Supplementary Data 4** online). To address its function, we applied CisGenome to determine whether the motif sites are phylogenetically conserved, whether they function in clusters and whether their locations are associated with structural features of genes (see **Supplementary Fig. 12** online).

Mapping the motif to the human genome yielded a total of 17,740 motif sites, of which 4,543 (25.6%) were phylogenetically conserved. In comparison, only 16.3% of the nonrepeat base pairs in the genome had the same level of conservation (see **Supplementary Table 9** online).

Table 2 Physical distribution of the new motif in human and mouse genomes

	Within 1kb upstream of TSS	Within 1 kb downstream of TES	Intragene	Intergene	Total sites
Human (hg17 assembly)					
All sites	1,920 (10.8%)	179 (1.0%)	7,168 (40.4%)	8,788 (49.5%)	17,740
Clustered sites	835 (49.9%)	37 (2.2%)	599 (35.8%)	336 (20.1%)	1,674
Clustered conserved sites	420 (59.6%)	18 (2.6%)	232 (32.9%)	104 (14.8%)	705
Mouse (mm7 assembly)					
All sites	1,530 (8.5%)	234 (1.3%)	6,532 (36.4%)	9,866 (55.0%)	17,940
Clustered sites	591 (46.7%)	46 (3.6%)	384 (30.4%)	318 (25.1%)	1265
Clustered conserved sites	303 (62.4%)	12 (2.5%)	118 (24.3%)	81 (16.7%)	486

TSS, transcription start site; TES, transcription end site. Number of motif sites and corresponding percentage among total sites are shown for each category.

When motif sites that were physically clustered together were collected, they were more than twice more conserved than nonclustered sites. Among the 1,674 sites that were separated from another site by ≤500 bp, 934 (55.8%) were phylogenetically conserved (versus 4,543 (25.6%) of the 17,740 general sites conserved; **Supplementary Table 9**).

There were 705 clustered conserved motif sites (defined as two conserved sites separated by ≤500 bp). Visual examination showed that, for the majority of these sites, only sequences within the sites were conserved, and the conservation dropped sharply at the site boundaries (**Fig. 4b,c**). Moreover, the most conserved positions coincided well with the most informative positions in the motif. Plotting the mean conservation scores for the flanking positions of the motif clearly verified the observation (**Fig. 4b**).

A summary of physical distributions of the motif sites revealed a strong correlation between the clustered sites and promoters (**Table 2**). Whereas only 1,920 (10.8%) of all 17,740 sites were located within 1 kb upstream of a transcription start site, 835 (49.9%) of the 1,674 clustered sites were within this region. This level increased to 420 (59.6%) of the 705 clustered conserved sites.

Repeating the same analyses on the mouse genome produced essentially the same results (**Table 2**, **Fig. 4** and **Supplementary Table 9**). The motif is thus highly likely to be a functional promoter element. Our findings suggest that future investigation of the motif is worthwhile, although the context of the motif's function awaits further exploration (**Supplementary Data 5** and **Supplementary Table 10** online).

DISCUSSION

Compared to commonly used algorithms such as MAT[11], TAS[13] and Tilescope[21], CisGenome's internal ChIP-chip peak caller provides competitive or higher sensitivity and specificity when applied to the recently published benchmark spike-in datasets[41] for ChIP-chip analysis (**Supplementary Data 6**, **Supplementary Figs. 13** and **14** and **Supplementary Table 11** online). The existing tools for ChIP-seq analysis, GeneTrack[29] and CPF[4], do not provide statistical estimates of FDR. QuEST[30] provides FDR estimates only when the negative control sample is available and when the control has twice as many reads as the ChIP sample. SISSRs[31] estimates FDR in the one-sample analysis based on a Poisson model. Compared to these tools, CisGenome provides not only high sensitivity and specificity, but also better methods for estimating FDR (**Supplementary Data 7** and **8** and **Supplementary Fig. 15** online). In the one-sample analysis, the negative binomial model provides a better model of background. In the two-sample analysis, the conditional binomial model does not impose special requirements on the number of negative control reads.

As summarized in **Supplementary Table 12** online, most peak detection tools do not support both ChIP-chip and ChIP-seq analyses and do not support high-level analyses such as motif discovery and peak-gene association. Traditionally, one requires other tools, such as MEME[42] and MDSCAN[25] (for motif discovery) and Galaxy[43] (for linking peaks to gene annotations). IGB can visualize Affymetrix tiling array data, and SignalMap is a proprietary tool for visualizing NimbleGen data. Both are platform specific and do not handle ChIP-seq data. Genome browsers at UCSC and Ensembl are useful for general purposes but are not optimized for handling ChIP data analyses. They do not provide certain functions that are particularly useful for ChIP data analyses, such as visualization of array images and motif logos, which are currently processed by independent tools such as WebLogo[44]. Furthermore, the need to constantly transfer data over the Internet makes large-scale interactive data analyses inefficient. Thus, the tools required to integrate different types of ChIP data and conduct various upstream and downstream analyses are currently distributed across at least a dozen programs. A considerable effort is required to reformat output of one piece of software before feeding it to the other. Although Web services such as CEAS[28] try to integrate multiple analysis functions, they usually only carry out analyses in a predefined manner, and there is limited flexibility to customize the analysis to answer the questions of most interest to the user (for example, analysis of the novel motif described above). In this context, the development of CisGenome has filled an urgent need for a single user-friendly environment with all the basic functionalities for ChIP-chip and ChIP-seq analyses. We believe the availability of CisGenome will significantly enhance the ability of experimental biologists to extract information from their ChIP datasets and from data provided by large-scale efforts such as the ENCODE[45] project.

In the interests of space, we only included in the main text the analyses that directly relate to our demonstration of CisGenome. Many issues not covered are nevertheless important, including likely reasons for the observed differences between the NRSF ChIP-chip and ChIP-seq data, whether these differences represent a general phenomenon, their relationship with previous comparisons of array and sequencing technologies[5,46], and different types of negative controls. Further analyses and discussions of these topics are provided in **Supplementary Data 9–13** and **Supplementary Figure 16** online.

METHODS

Datasets. Data used in this study are summarized in **Supplementary Table 1**. The NRSF ChIP-chip data (GEO accession no. GSE8489) were obtained by analyzing the bound DNA fragments in Jurkat cells with Affymetrix Human Tiling 2.0R arrays. Two independent ChIP samples and two mock immunoprecipitation samples were profiled. The NRSF ChIP-seq data were collected from a previous study[4]. In that study, DNA fragments bound by NRSF in Jurkat cells were sequenced with the next-generation sequencer made by Illumina/Solexa. These experiments involved sequencing a ChIP sample and a negative control sample generated from reverse–cross-linked genomic DNA that had not undergone immunoprecipitation. The Oct4 and Nanog ChIP-seq data were collected from ref. 10.

Outline of ChIP-seq data analysis. Most sequencing platforms will output mapped sequence reads up to a specified number of mismatches and will allow elimination of reads that map to multiple locations. CisGenome can accept the mapped reads as input. CisGenome also accepts mapping output from SeqMap[47], a program that allows mapping of sequence reads in more customized ways, such as accounting for insertions and deletions (see **Supplementary Methods**).

For FDR computation from a ChIP sample only, the genome is divided into nonoverlapping windows with length w (typically 100 bp). The number of reads (n_i) within each window i is counted. It is assumed that in nonbinding regions, $n_i|\lambda_i \sim$ Poisson (λ_i), and $\lambda_I \sim$ gamma (α,β). This implies that the background read occurrence rate varies across the genome, and marginally $n_i \sim$ negative binomial (α,β). To estimate α and β, a truncated negative binomial distribution is fitted to the number of windows with a small number of reads (two or fewer). This estimated null distribution is used to compute the FDR for each level of read counts. In the widely used Poisson model, λ_i is assumed to be a constant λ_0 across the genome, rather than a random variable. To estimate λ_0, a truncated Poisson is fitted using the windows with one or fewer reads. The FDR computation and model fitting details are provided in **Supplementary Methods**. The fitting method assumes that most windows with small read counts represent noise. The assumption usually holds true with sufficient depth of sequencing. For studies in which signals cover a large fraction of the genome (for example, histone modifications) but the sequencing coverage is not deep enough, the true targets may be covered by only one or two reads in a short window. When this is the case, our model-fitting approach may be applicable after increasing the window size, or may not be applicable, depending on how long a typical peak extends.

In a specific location, the counts of the reads from the ChIP sample are subjected to biases that may arise during sample preparation, amplification or sequencing procedures. To correct for these biases, sequence reads can be generated from negative control samples in the same experiments. **Supplementary Figures 5** and **17** and **Supplementary Table 13** online show that the read sampling rates from the ChIP and control samples at the same genomic loci are correlated. Therefore, false signals caused by unknown systematic bias can be eliminated by excluding regions if both the ChIP and negative control samples show strong signals but the former is not significantly stronger than the latter. When reads are also available from a negative control sample, the genome is divided into nonoverlapping windows with length w. For each window i, the number of reads in the ChIP sample (k_{1i}), the number of reads in the control sample (k_{2i}) and the total read number ($n_i = k_{1i} + k_{2i}$) are counted. When there is no IP enrichment in the window, the conditional distribution of the count in the ChIP sample (k_{1i}) given the total count (n_i) is assumed to follow a binomial (n_i, p_0) distribution. p_0 is estimated based on windows with small total counts and used to estimate the FDR associated with each level of n_i and k_{1i}/n_i (see **Supplementary Methods**).

For binding region detection, the genome is scanned with a sliding window of width w to detect all windows with FDR smaller than a user-chosen cutoff. Detected windows that overlap with each other are merged into one region. If a region contains more than one overlapping window, the minimal FDR among the overlapping windows is taken as the FDR of the region. In the two-sample analysis, for each sliding window i, a fold enrichment $(y_i + 1)$ / $(r_0{*}z_i + 1)$ is computed where y_i is the number of ChIP reads in the window, z_i is the number of control reads in the window and $r_0 = p_0$ / $(1 - p_0)$. To avoid dividing by 0, 1 is added to both the numerator and denominator. The biggest fold change among all the overlapping windows within a binding region is recorded as the fold change of the region.

For peak localization and filtering, CisGenome uses the counts of 5′ and 3′ reads within each candidate binding region to further pinpoint the location of transcription factor binding sites within the region (**Fig. 2d**) and to filter out regions enriched for reads of only one direction, based on the assumption that these are unlikely to represent real binding events. Regions that are retained after the boundary refinement and single-strand filtering are defined as high-quality binding regions (see **Supplementary Methods**).

To ensure adjustment for DNA fragment length, CisGenome uses a two-pass algorithm for peak detection. High-quality peaks detected in the first pass will be used to estimate the DNA fragment length, which is computed as the median distance between the modes of the coupling 5′ and 3′ peaks. In the second pass, the reads are shifted toward the center of the ChIP fragments by half of the estimated fragment length, and FDR computation and peak detection will be run again on the shifted reads to get the final predictions.

The default choice of window size w (100 bp) represents a tradeoff between sensitivity and specificity based on the analysis of the NRSF data (**Supplementary Tables 14** and **15** online). With a smaller w, one can get sharper boundaries of binding regions. However, more noise will be introduced, and fewer regions containing the NRSF motif will pass the significance cutoff (FDR $\leq$ 10%). A bigger w may dilute the signals, resulting in a lower resolution of binding region call and a lower percentage of regions that contain the NRSF motif. In future transcription factor studies, one can fine-tune the choice of window size w in a similar fashion by using either the known transcription factor binding motifs or motifs recovered from the *de novo* motif discovery.

Analysis of phylogenetic conservation. To characterize the conservation level of the binding regions, CisGenome allows users to first choose a t such that x percent of the whole genome has a phastCons[48] score $\geq t$. For each peak, positions with phastCons score $\geq t$ are picked up, and the average phastCons score for these positions is computed to serve as the peak's conservation level. If a peak has no position with phastCons score $\geq t$, its conservation level is 0. A high cutoff t (or a small x) will help users focus on the most conserved part of each binding region. To generate **Figure 3e**, the default value $x = 10$ was used. To generate **Figure 3e**, we used $x = 10$ as the default value. The ranked binding regions were grouped into tiers of 300. Peak conservation levels within a tier were averaged. In CisGenome, phastCons score is transformed linearly from [0, 1] to [0, 255] so that each computer byte can store the score for a single genomic position.

Accession numbers. NRSF ChIP-chip (GEO: GSE8489); NRSF ChIP-seq (GEO: GSE13047); Oct4 and Nanog ChIP-seq (GEO: GSE 11724).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS
We thank W. Li for assistance with analyzing the ChIP-chip spike-in data. This research was supported by National Institutes of Health grant HG003903 (to W.H.W.) and the National Human Genome Research Institute's ENCODE project (to R.M.M.). H. Ji is partially supported by the Johns Hopkins Bloomberg School of Public Health Richard L. Gelb Cancer Research Fund.

AUTHOR CONTRIBUTIONS
H. Ji conceived the study, developed the CisGenome GUI and data analysis algorithms, carried out data analyses and drafted the manuscript. H. Jiang developed the CisGenome browser. W.M. participated in algorithm development and carried out data analyses. D.S.J. and R.M.M. generated NRSF ChIP-chip data. W.H.W. conceived the study and drafted the manuscript. All authors read and revised the manuscript.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
2. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
3. Carroll, J.S. *et al.* Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* **38**, 1289–1297 (2006).
4. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
5. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).

6. Mikkelsen, T.S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
7. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
8. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
9. Wederell, E.D. *et al.* Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.* **36**, 4549–4564 (2008).
10. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521–533 (2008).
11. Johnson, W.E. *et al.* Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci. USA* **103**, 12457–12462 (2006).
12. Ji, H. & Wong, W.H. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* **21**, 3629–3636 (2005).
13. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).
14. Zheng, M., Barrera, L.O., Ren, B. & Wu, Y.N. ChIP-chip: data, model, and analysis. *Biometrics* **63**, 787–796 (2007).
15. Keles, S. Mixture modeling for genome-wide localization of transcription factors. *Biometrics* **63**, 10–21 (2007).
16. Ghosh, S., Hirsch, H.A., Sekinger, E., Struhl, K. & Gingeras, T.R. Rank-statistics based enrichment-site prediction algorithm developed for chromatin immunoprecipitation on chip experiments. *BMC Bioinformatics* **7**, 434 (2006).
17. Du, J. *et al.* A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics* **22**, 3016–3024 (2006).
18. Qi, Y. *et al.* High-resolution computational models of genome binding events. *Nat. Biotechnol.* **24**, 963–970 (2006).
19. Scacheri, P.C., Crawford, G.E. & Davis, S. Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen arrays. *Methods Enzymol.* **411**, 270–282 (2006).
20. Bieda, M., Xu, X., Singer, M.A., Green, R. & Farnham, P.J. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.* **16**, 595–605 (2006).
21. Zhang, Z.D. *et al.* Tilescope: online analysis pipeline for high-density tiling microarray data. *Genome Biol.* **8**, R81 (2007).
22. Song, J.S. *et al.* Model-based analysis of two-color arrays (MA2C). *Genome Biol.* **8**, R178 (2007).
23. Reiss, D.J., Facciotti, M.T. & Baliga, N.S. Model-based deconvolution of genome-wide DNA binding. *Bioinformatics* **24**, 396–403 (2008).
24. Song, J.S. *et al.* Microarray blob-defect removal improves array analysis. *Bioinformatics* **23**, 966–971 (2007).
25. Liu, X.S., Brutlag, D.L. & Liu, J.S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* **20**, 835–839 (2002).
26. Hong, P. *et al.* A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics* **21**, 2636–2643 (2005).
27. Shim, H. & Keles, S. Integrating quantitative information from ChIP-chip experiments into motif finding. *Biostatistics* **9**, 51–65 (2008).
28. Ji, X., Li, W., Song, J., Wei, L. & Liu, X.S. CEAS: cis-regulatory element annotation system. *Nucleic Acids Res.* **34**, W551–554 (2006).
29. Albert, I., Wachi, S., Jiang, C. & Pugh, B.F. GeneTrack–a genomic data processing and visualization framework. *Bioinformatics* **24**, 1305–1306 (2008).
30. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5**, 829–834 (2008).
31. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* **36**, 5221–5231 (2008).
32. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13–D21 (2008).
33. Karolchik, D. *et al.* The UCSC genome browser database: 2008 update. *Nucleic Acids Res.* **36**, D773–D779 (2008).
34. Flicek, P. *et al.* Ensembl 2008. *Nucleic Acids Res.* **36**, D707–D714 (2008).
35. Liu, J.S., Neuwald, A.F. & Lawrence, C.E. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.* **90**, 1156–1170 (1995).
36. Zhou, Q. & Wong, W.H. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA* **101**, 12114–12119 (2004).
37. Ji, H., Vokes, S.A. & Wong, W.H. A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors. *Nucleic Acids Res.* **34**, e146 (2006).
38. Chen, Z.F., Paquette, A.J. & Anderson, D.J. NRSF/REST is required *in vivo* for repression of multiple neuronal target genes during embryogenesis. *Nat. Genet.* **20**, 136–142 (1998).
39. Chong, J.A. *et al.* REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80**, 949–957 (1995).
40. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
41. Johnson, D.S. *et al.* Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res.* **18**, 393–403 (2008).
42. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28–36. AAAI Press, Menlo Park, California, USA, (1994).
43. Giardine, B. *et al.* Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451–1455 (2005).
44. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
45. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
46. Euskirchen, G.M. *et al.* Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res.* **17**, 898–909 (2007).
47. Jiang, H. & Wong, W.H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**, 2395–2396 (2008).
48. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
49. Schmid, C.D. & Bucher, P. ChIP-Seq data reveal nucleosome architecture of human promoters. *Cell* **131**, 831–832 (2007).