BMC Genomics

ChIPseek, a web-based analysis tool for ChIP data

Ting-Wen Chen[1,2†], Hsin-Pai Li[1,3,4†], Chi-Ching Lee[1,2], Ruei-Chi Gan[1,2,5], Po-Jung Huang[1,2], Timothy H Wu[6], Cheng-Yang Lee[1,2], Yi-Feng Chang[1,6] and Petrus Tang[1,2,3*]

Abstract

Background: Chromatin is a dynamic but highly regulated structure. DNA-binding proteins such as transcription factors, epigenetic and chromatin modifiers are responsible for regulating specific gene expression pattern and may result in different phenotypes. To reveal the identity of the proteins associated with the specific region on DNA, chromatin immunoprecipitation (ChIP) is the most widely used technique. ChIP assay followed by next generation sequencing (ChIP-seq) or microarray (ChIP-chip) is often used to study patterns of protein-binding profiles in different cell types and in cancer samples on a genome-wide scale. However, only a limited number of bioinformatics tools are available for ChIP datasets analysis.

Results: We present ChIPseek, a web-based tool for ChIP data analysis providing summary statistics in graphs and offering several commonly demanded analyses. ChIPseek can provide statistical summary of the dataset including histogram of peak length distribution, histogram of distances to the nearest transcription start site (TSS), and pie chart (or bar chart) of genomic locations for users to have a comprehensive view on the dataset for further analysis. For examining the potential functions of peaks, ChIPseek provides peak annotation, visualization of peak genomic location, motif identification, sequence extraction, and comparison between datasets. Beyond that, ChIPseek also offers users the flexibility to filter peaks and re-analyze the filtered subset of peaks. ChIPseek supports 20 different genome assemblies for 12 model organisms including human, mouse, rat, worm, fly, frog, zebrafish, chicken, yeast, fission yeast, *Arabidopsis*, and rice. We use demo datasets to demonstrate the usage and intuitive user interface of ChIPseek.

Conclusions: ChIPseek provides a user-friendly interface for biologists to analyze large-scale ChIP data without requiring any programing skills. All the results and figures produced by ChIPseek can be downloaded for further analysis. The analysis tools built into ChIPseek, especially the ones for selecting and examine a subset of peaks from ChIP data, provides invaluable helps for exploring the high through-put data from either ChIP-seq or ChIP-chip. ChIPseek is freely available at http://chipseek.cgu.edu.tw.

Keywords: ChIP-seq, ChIP-chip, Analysis tool, Web-services, Peak annotation, Motif identification, Filter tools, Comparison

Background

The chromatin immunoprecipitation (ChIP) assay is a powerful technique to examine the specific interaction between proteins and DNA within living cells [1-3]. The DNA-binding proteins play important roles in regulating many cellular processes including gene expression, replication, recombination, repair, methylation and chromatin remodeling. The ChIP assay provides comprehensive analysis to reveal specific DNA and protein interaction according to: modifications on DNA-binding proteins (phosphorylation, acetylation, methylation, ribosylation and ubiquitination); DNA methylation; spatial and temporal regulation during development; different cell types and physiological conditions. Application of ChIP assay is therefore versatile, specific applications are listed as follows: identifying the transcription factor binding sites [4], locating the sites of histone modifications [5], mapping the hypermethylated DNA region in stem cells [6], identifying the changes in epigenetic modifications according to different developmental stages [7], and identifying the

* Correspondence: petang@mail.cgu.edu.tw
†Equal contributors
[1]Molecular Medicine Research Center, Chang Gung University, Taoyuan, Taiwan
[2]Bioinformatics Center, Chang Gung University, Taoyuan, Taiwan
Full list of author information is available at the end of the article

epigenetic changes in different cell types [8] and in cancers [5,9].

To perform ChIP assay, proteins and DNA are in vivo-crosslinked by formaldehyde, chromatin is sonicated into small fragments, and an antibody is used to immunoprecipitate the protein-DNA complexes [1]. The DNA fragments are further purified and subjected to various analyses, for instance, PCR, cloning, hybridization and sequencing. The ChIP assay not only allows quantitative detection of a given protein on a particular DNA site but also permits the mapping of genome-wide DNA-protein interaction. With the advancement of sequencing technology, ChIP assay has been adapted to use high throughput sequencing (ChIP-seq) for mapping the DNA region or even defining the consensus sequences that are involved in DNA-protein interaction [10-13]. Several peak calling tools have been proposed for ChIP-seq data analysis, such as MACS, PeakSeq and HPeak [14-16]. However, a major challenge comes after the ChIP-seq—processing, analyzing and presenting the ChIP-seq data. Several tools have been developed for ChIP-data analysis. For instance, CisGenome provides analysis and visualization of ChIP-data [17]; Hypergeometric Optimization of Motif EnRichment (HOMER) can find motifs from peak files and annotated peaks [18]; *P*ositioning *data*base and *a*nalysis *t*ool (Podbat) provides analysis and visualization of peaks [19]; Cistrome contains many ChIP data analysis packages [20]; EpiExplorer offers an interactive web interface for visualization of peak distribution [21]; Genomic Association Test (GAT) can test whether multiple sets of peaks significantly overlap with each other [22]; PscanChIP can identify the enriched transcription factor binding sites [23]; and PAVIS focuses on annotation and visualization of peaks [24]. A summary of these tools is shown in Table 1.

As shown in Table 1, most of the existing tools do not have built-in genomic location features for comparison. There are already many large-scale projects that focus on ChIP-seq of transcription factors and histone modifications, such as the ENCODE and modENCODE projects [26,27]. The ENCODE consortia have already completed ChIP-seq for 161 transcription factors in more than 100 cell types using the same guidelines and quality-controls [27]. These databases are highly valuable and very informative, especially in terms of their breadth and quality. Yet most of these tools do not provide the functionality to filter based on peak length or on position relative to a transcription start site (TSS). However, after obtaining a list of peaks from either ChIP-seq or ChIP-chip, researchers want to obtain an overview of all peaks and then pick out the probable or significant peaks for further validation and investigation. The common procedure to obtain the candidate peaks is to filter and then re-analyze the subset of data. To exploit the well-established databases and provide the functionality to filter peaks, we present a web-based analysis tool,

Table 1 Summary of available tools

	CisGenome [17]	BEDTool [25]	HOMER [18]	Podbat [19]	Cistrome [20]	EpiExplorer [21]	GAT [22]	PscanChIP [23]	PAVIS [24]	ChIPseek
Web-based analysis tool				✓	✓	✓		✓	✓	✓
Graphic user interface	✓			✓	✓	✓		✓	✓	✓
Annotation	✓		✓	✓	✓	✓		✓	✓	✓
Plot of position relative to TSS	✓			✓	✓	✓		✓		✓
Plot of peak length distribution						✓				✓
Filter (by peak length)						✓				✓
Filter (by position relative to TSS)										✓
Post-filter analysis										✓
Motif identification	✓		✓		✓			✓		✓
Comparison	✓	✓		✓	✓	✓+	✓			✓
Link to UCSC genome browser	✓	✓			✓	✓		✓		✓
Plot of genomic location	✓			✓	✓				✓	✓
Sequence extraction		✓								✓
Built-in peak databases	✓					✓				✓
Enrichment analysis				✓	✓		✓	✓	✓	

+EpiExplorer provides comparison between an uploaded file and a built-in database.

ChIPseek, which can (1) annotate peak files; (2) filter datasets based on the size of peaks; (3) filter datasets based on the distance to the nearest TSS; (4) identify enriched motifs based on uploaded datasets or filtered datasets; (5) compare two sets of peaks; (6) extract peak sequences and motif identification; (7) compare an uploaded dataset or filtered dataset with built-in datasets; (8) plot peaks on chromosome ideograms, and (9) select peaks based on their closest genes corresponding to the gene ontology (GO) or signaling pathways (KEGG database) [28,29].

ChIPseek takes BED, GFF and text files as input and provides annotation for the peaks. In addition to peak annotation, ChIPseek provides pie charts, bar charts and histograms for peak location distribution, which offer an intuitive way to understand the distribution of all peaks. ChIPseek also generates histograms indicating the frequency distribution of the peak length and the distance to the nearest TSS. All these figures or tables generated by ChIPseek can be downloaded for further analysis. Currently, ChIPseek supports data from *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Xenopus tropicalis*, *Danio rerio*, *Gallus gallus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana* and *Oryza sativa*. Overall, ChIPseek provides both systematic visualization and filter tools for ChIP data analysis through a user-friendly interface. We believe that ChIPseek is a useful and valuable tool for further ChIP and epigenetic studies.

Implementation
ChIPseek is mainly constructed using python and the website is written in mod-python, Google Charts, jQuery, JavaScript and PHP [30-32]. ChIPseek also integrates several well-developed and widely used tools, such as HOMER and BEDTools [18,25], to provide better peak annotation and analysis.

Annotation table
The first result generated by ChIPseek is an annotation table. After uploading a file(s), ChIPseek first checks for file format. If the uploaded file(s) contains the proper location information, then ChIPseek can further utilize annotatePeaks.pl from HOMER [18] to annotate the peaks according to their genomic locations.

Links to the UCSC Genome browser
In the annotation table, ChIPseek provides a link to the UCSC genome browser to allow interactive access to the genomic sequence for each peak. ChIPseek also adds custom tracks of uploaded files on the UCSC website. To add custom tracks, ChIPseek incorporates BEDTools [25] and two tools downloaded from UCSC ftp [33,34]. ChIPseek first uses fetchChromSizes to produce a file of

recorded sizes of chromosomes belonging to the organism of interest. ChIPseek then uses mergeBed and sortBed to ensure that the input peak file is compatible with further conversion. Finally, ChIPseek employs bedGraphToBigWig to convert the peak file, together with chromosome size information, into bigWig file. The input for bedGraphToBigWig must be non-redundant and sorted by position. ChIPseek utilizes shell scripts to prepare a proper input BED file for bedGraphToBigWig. The bigWig file can then be uploaded to UCSC to visualize the peak information. If users upload multiple peak files, then ChIPseek can generate multiple tracks and display them simultaneously in the UCSC genome browser.

Statistics of peaks
ChIPseek generates several distribution figures based on the annotation results. These plots are generated in real time by python scripts, mod_python and JavaScript libraries from Google charts [30].

Filter criteria
In order to provide a comprehensive analysis of the distribution and location of the protein binding sites/peaks relative to TSS, and to classify the biological impact of the targets to a specific group of gene in the same metabolic or signal pathway, users can further subdivide the peaks by using six filter criteria based on (a) Distance to nearest TSS, (b) Peak length, (c) Peak location, (d) Gene list, (e) Gene Ontology (GO), and (f) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway. ChIPseek uses JavaScript in filter criteria (a) Distance to TSS and (b) Peak length to provide slider bars and text boxes for users to specify a range (bp) to filter the peaks according to distance to TSS or peak length. ChIPseek then generates histogram with mod_python and Google chart libraries for the filtered subset of peaks. For filter criteria (e) GO and (f) KEGG, the relationships between GO terms are downloaded from the GO website (http://www.geneontology.org). The functional annotations of KEGG are downloaded from KEGG BRITE database (http://www.genome.jp/kegg/brite.html). KEGG ID conversion API is used to convert the NCBI gene ID and KEGG Orthology (KO) ID. Several in-house scripts are used to integrate all functional hierarchies into tree structures. Fancytree from jQuery Plugin Registry is used to show the relationship between those functional annotation terms [35].

Get genomic sequences
ChIPseek employs fastaFromBed to extract corresponding genomic sequences from the corresponding genome. The output FASTA files are converted into tab-delimited text files with in-house python scripts.

Motif identification

For enriched motif identification, ChIPseek integrates findMotifsGenome.pl, a tool from HOMER [18] which can identify enriched motifs in genomic regions, and use Weblogo and Ghostscript for sequence logo generation [35,36]. The motif identification includes the following steps: extraction of sequences, background selection, GC-content normalization (for background sequences), and finally motif discovery (including novel motifs and known motifs). ChIPseek provides four options (±25 bp, ±100 bp, ±250 bp and ±500 bp) for users to specify the size of the region to search for motifs.

Chromosome ideograms

To achieve a better understanding of overall peak distribution in the genome, ChIPseek provides ideograms of chromosomes indicating the relative positions of the peaks associated with them. The ideograms are implemented in JavaScript, and we adapted the ideogram plotting code Ideogram Viewer from MD Anderson Center to ChIPseek [37]. We also made several modifications: (1) we added options for different colors to mark the peaks on a chromosome; (2) we added a hyperlink to each mark and provide a link to the UCSC genome browser; (3) we changed the content that is displayed when the mouse pointer hovers over an item; and (4) we changed the ends of the chromosome into round ends. The cytoband information used in ideograms is downloaded from UCSC [38].

Intersection of peaks

ChIPseek provides the option to compare two set of peaks. The comparison generates the relationship between peaks from the two sets. To obtain these peak relationships, ChIPseek employs several in-house scripts, as well as intersectBed and coverageBed from BEDdtools [25].

Database of transcription factor binding sites

ChIPseek also provides a binding sites database of transcription factors for human genome assemblies (hg19, hg18 and hg17). This database is generated from the Encyclopedia of DNA Elements (ENCODE) project [26], which contains 161 transcription factors in its database. The latest ENCODE dataset (hg19) is obtained from the UCSC Genome Browser download server. We further applied liftOver [39] to cover the binding site to obtain corresponding coordinates for hg18 and hg17. All of these databases are available in the comparison section.
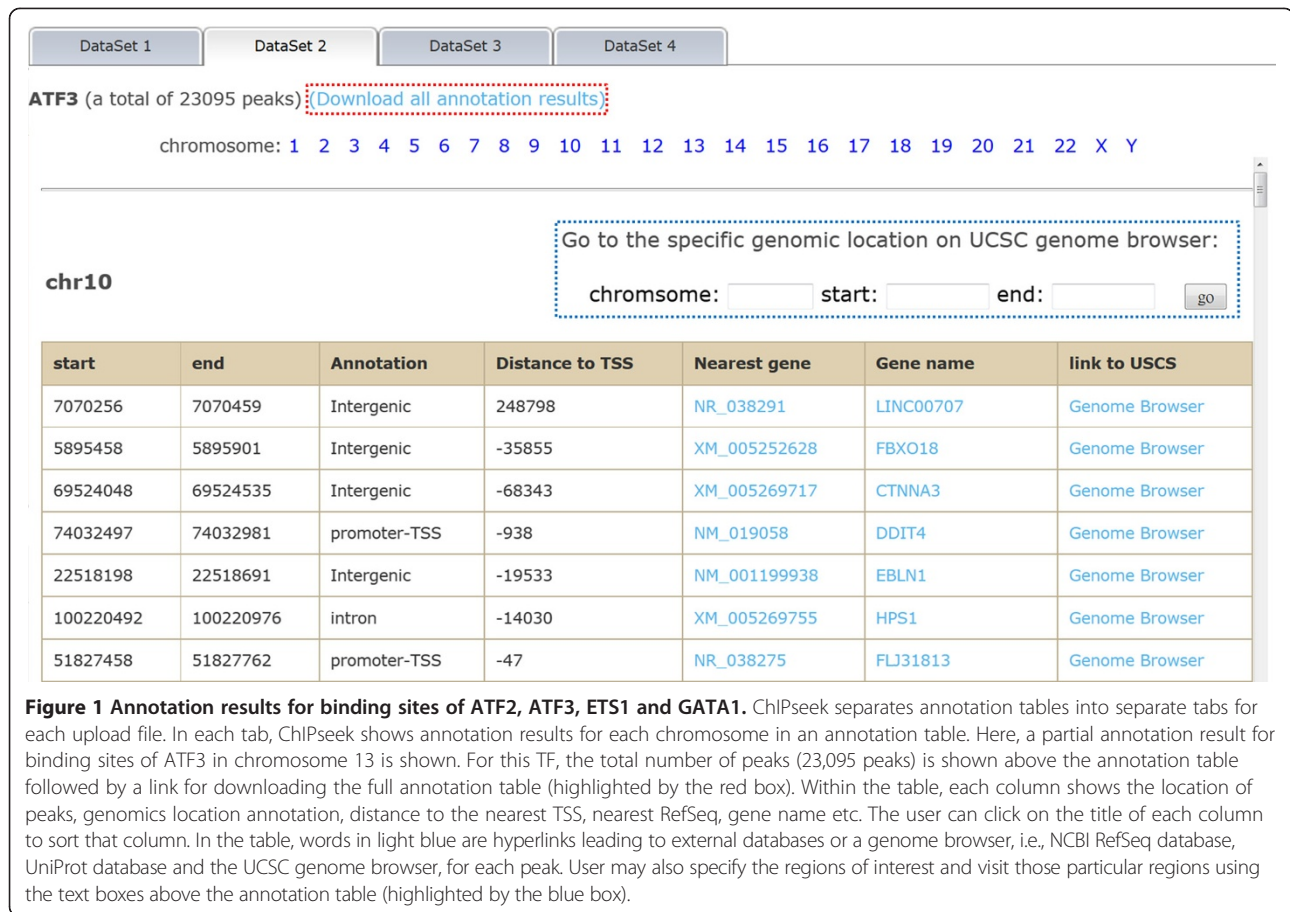
Results and discussion

ChIPseek is a web-based tool designed for ChIP data analysis. This software includes the following features: annotation of peaks; motif identification; peak sequence extraction; peak plotting on chromosome ideograms;

generation of histograms of the peak length distribution; generation of histograms of the distance to the nearest TSS; generation of pie charts of the genomic location, and peak filtering by peak length, distance to the nearest TSS, peak location, gene list, GO categories, and KEGG pathways, etc. Currently, ChIPseek mainly focuses on the human genome (hg17, hg18 and hg19) but also provides most of the analytic functions for other model organisms, such as mice (mm8, mm9 and mm10), rats (rn4 and rn5), worms (ce6 and ce10), flies (dm3), frogs (xenTro3, xenTro2), zebrafish (danRer7), chickens (galGal4), yeast (sacCer2 and sacCer3), *Schizosaccharomyces pombe* (ASM294v1), Arabidopsis (tair10) and rice (*Oryza sativa*, msu6). ChIPseek supports three types of input file formats: text file, BED format and GFF format. Users can upload their own data to perform all supported analyses through the web interface. Here, we demonstrate the usage and analysis functions of ChIPseek with two datasets: (1) the DNA binding sites of four transcriptional activators, ATF2, ATF3, ETS1 and GATA1, and (2) binding sites of the transcriptional repressor CTCF.

Demo dataset 1: Transcription factor binding sites of ATF2, ATF3, ETS1 and GATA1

The binding sites of transcriptional factors ATF2, ATF3, ETS1 and GATA1 are generated by the ENCODE project. We downloaded the BED files from the UCSC download server and used them as a demo dataset for ChIPseek. We uploaded four BED files and selected Human/hg19 as our genome assembly version from the multiple file upload page. The input files are checked and then annotated by ChIPseek. The results are shown in the annotation tables (as shown in Figure 1). If users upload multiple files, then the annotation tables for each file are listed in separate tabs. The annotation results are separated according to different chromosomes. The annotation table for each chromosome indicates start, end, location annotation, distance to the nearest TSS, nearest RefSeq, nearest gene name and link to the UCSC genome browser. Users can also link to external databases, NCBI and UniProt [40,41], by clicking on the RefSeq or gene symbol. The total number of peaks is shown on top of each tab. In our demo samples, there are a total of 23095, 26026, 13737 and 20310 peaks for ATF2, ATF3, ETS1 and GATA1, respectively. Due to limited space on the browser, seven columns of peak information are displayed. The detailed information of the peaks can be downloaded by clicking the hyperlink "Download all annotation results" on the top-left corner within each separate tab. The full annotation results include the distance to the nearest TSS, nearest promoter ID, Entrez Gene ID, nearest Uniqene ID, nearest Refseq, nearest Ensembl gene name, gene symbol, gene aliases, gene

| | DataSet 1 | DataSet 2 | DataSet 3 | DataSet 4 |

ATF3 (a total of 23095 peaks) (Download all annotation results)

chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

chr10

Go to the specific genomic location on UCSC genome browser:

chromosome: [   ] start: [   ] end: [   ] [go]

start	end	Annotation	Distance to TSS	Nearest gene	Gene name	link to USCS
7070256	7070459	Intergenic	248798	NR_038291	LINC00707	Genome Browser
5895458	5895901	Intergenic	-35855	XM_005252628	FBXO18	Genome Browser
69524048	69524535	Intergenic	-68343	XM_005269717	CTNNA3	Genome Browser
74032497	74032981	promoter-TSS	-938	NM_019058	DDIT4	Genome Browser
22518198	22518691	Intergenic	-19533	NM_001199938	EBLN1	Genome Browser
100220492	100220976	intron	-14030	XM_005269755	HPS1	Genome Browser
51827458	51827762	promoter-TSS	-47	NR_038275	FLJ31813	Genome Browser

Figure 1 Annotation results for binding sites of ATF2, ATF3, ETS1 and GATA1. ChIPseek separates annotation tables into separate tabs for each upload file. In each tab, ChIPseek shows annotation results for each chromosome in an annotation table. Here, a partial annotation result for binding sites of ATF3 in chromosome 13 is shown. For this TF, the total number of peaks (23,095 peaks) is shown above the annotation table followed by a link for downloading the full annotation table (highlighted by the red box). Within the table, each column shows the location of peaks, genomics location annotation, distance to the nearest TSS, nearest RefSeq, gene name etc. The user can click on the title of each column to sort that column. In the table, words in light blue are hyperlinks leading to external databases or a genome browser, i.e., NCBI RefSeq database, UniProt database and the UCSC genome browser, for each peak. User may also specify the regions of interest and visit those particular regions using the text boxes above the annotation table (highlighted by the blue box).
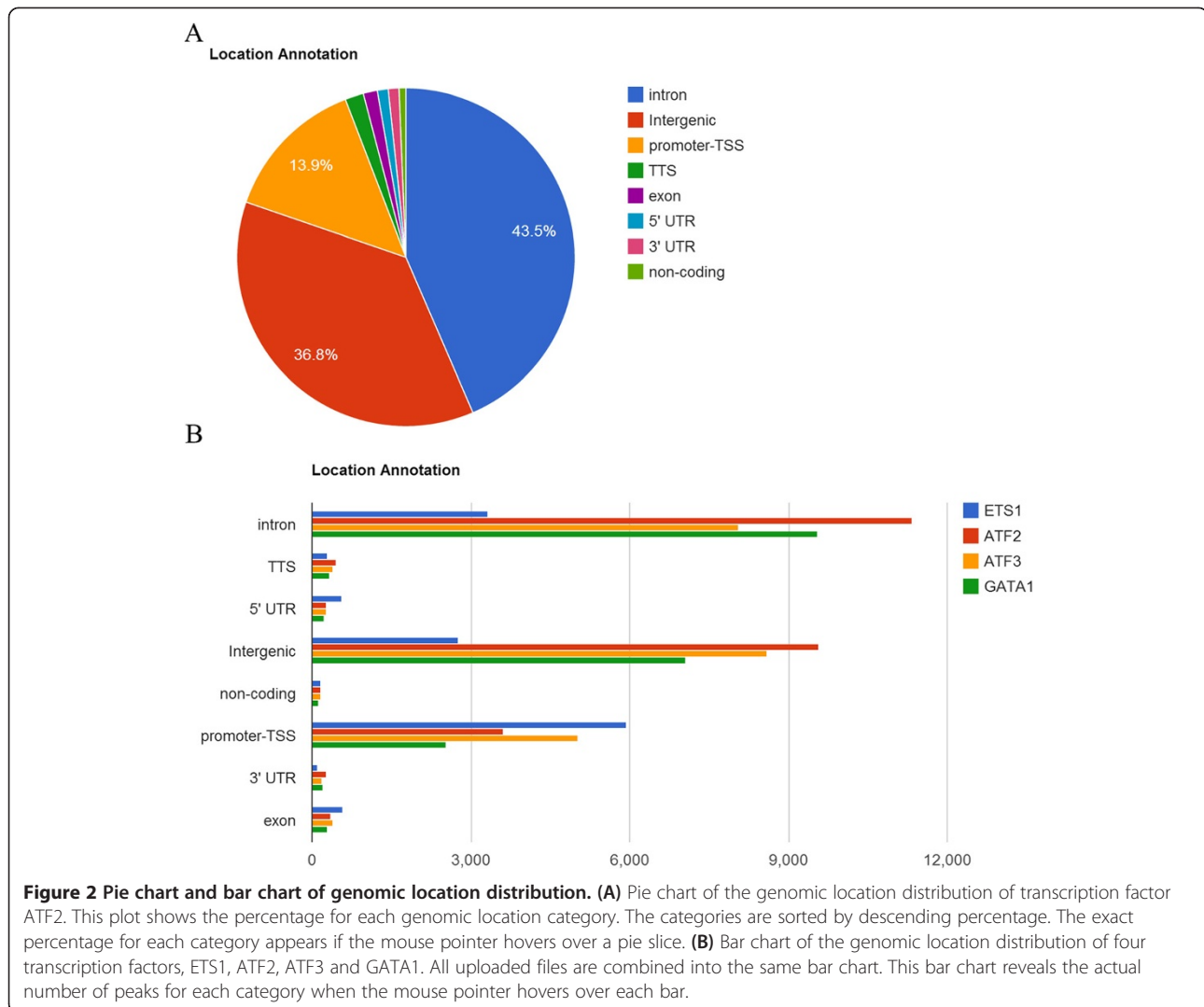
description, genomic location annotation, and hyperlinks to external databases in text format.

ChIPseek provides links to the UCSC genome browser and displays uploaded peak data on the genome browser. UCSC integrates many biological information sources such as SNP, expression profile and regulation elements [34,38], allowing users to access and explore peaks of interest at many levels. In the last column of the annotation table, ChIPseek provides a hyperlink for each peak, which can lead users to the corresponding region for that particular peak. By default, the region extends 1,000 bp upstream and downstream. Users can directly select a peak and explore other biological feature tracks in adjacent regions on the UCSC genome browser. ChIPseek also provides the flexibility for users to explore a specific region of interest with the genome browser. As shown in Figure 1, users can enter a specified region in the text boxes above the annotation table, click on the 'go' button, and be linked to the UCSC genome browser, focusing on that particular region in a new window.

The scores (height) of custom tracks indicated in the UCSC genome browser are either from users' uploaded files or given by ChIPseek. If users uploaded BED files, GFF files or text files with score information, then the score will be shown on the UCSC genome browser. If there is no score information in the uploaded text file, ChIPseek will assign all peaks a score of 1. In that case, all peaks will have the same height on the UCSC genome browser. Moreover, the peaks in a single track on the UCSC genome browser cannot overlap. Therefore, if there are overlapping peaks in a single uploaded file, ChIPseek will merge those overlapping peaks and use the average score as a representation for the final merged peak.

To visualize the peak distribution within the genome, the examination of location distribution is usually the first step in analyzing ChIP data. The location represents where each peak located and the value includes the 3′UTR (untranslated region), 5′UTR, exon, intergenic region, intron, non-coding, promoter-TSS and transcription termination site (TTS). Both pie charts and bar charts for visualizing the overall peak distribution are available in ChIPseek. If users upload multiple files, separate pie charts for each dataset and a bar chart containing all datasets for comparison are presented as shown in Figure 2. Users may choose a suitable chart to present their data. If users are interested in the detailed composition of each category, then the raw data of genomic location is also available in the full annotation table as mentioned previously.

Figure 2 Pie chart and bar chart of genomic location distribution. (A) Pie chart of the genomic location distribution of transcription factor ATF2. This plot shows the percentage for each genomic location category. The categories are sorted by descending percentage. The exact percentage for each category appears if the mouse pointer hovers over a pie slice. **(B)** Bar chart of the genomic location distribution of four transcription factors, ETS1, ATF2, ATF3 and GATA1. All uploaded files are combined into the same bar chart. This bar chart reveals the actual number of peaks for each category when the mouse pointer hovers over each bar.

To display the frequency and distribution of the distance to the nearest TSS or distribution of peak lengths, histograms are used. As shown in Figure 3, the x-axis for the histogram of distance to the nearest TSS is centered at 0 and divided into 100 bins. The x-axis for the histogram of the length distribution starts at 0 and is divided into 50 bins. One of the unique features of ChIP-seek is that it offers six filters allowing users to obtain a subset of peaks for further analysis. These six filters are established based on the following criteria: distance to the nearest TSS, lengths of peaks, genomic location, uploaded gene list, selected GO terms, and KEGG groups. Because the transcription factors are likely to be located near the TSS of regulated genes, we can filter the binding sites of ATF2 and ATF3 based on the "Distance to the nearest TSS". Using this filter criterion, only the peaks located between −200 and +200 bp of the nearest TSS are selected. After this step, a total of 2,847 out of 26,026 ATF2 peaks and 3,999 out of 23,095 ATF3 peaks fit this

criterion. A new histogram of the subset peaks in real time as shown in Figure 3. Two new files containing the two subsets of peaks will be generated by and saved in ChIP-seek, and will be automatically named ATF2_-200-200_TSS and ATF3_-200-200_TSS, respectively. The naming system starts with the original uploaded file name, followed by the range and then "TSS", which shows the filter criteria based on the distance to TSS. These two saved files can be used in the subsequent analysis.

In addition t.o generating plots for exploring the properties of peaks, ChIPseek also extracts the sequences of the peaks. These sequences can be used for further analysis such as TFsearch, DMEME, CentriMo or other ChIP data analysis tools etc. [42-45]. After clicking the "Peak sequences" on the menu, users can obtain a table of partial extracted sequences. The extracted sequences are presented in two forms: FASTA format and tab-delimited text format. The tab-delimited format contains 5 columns: chromosome, start position, end position,
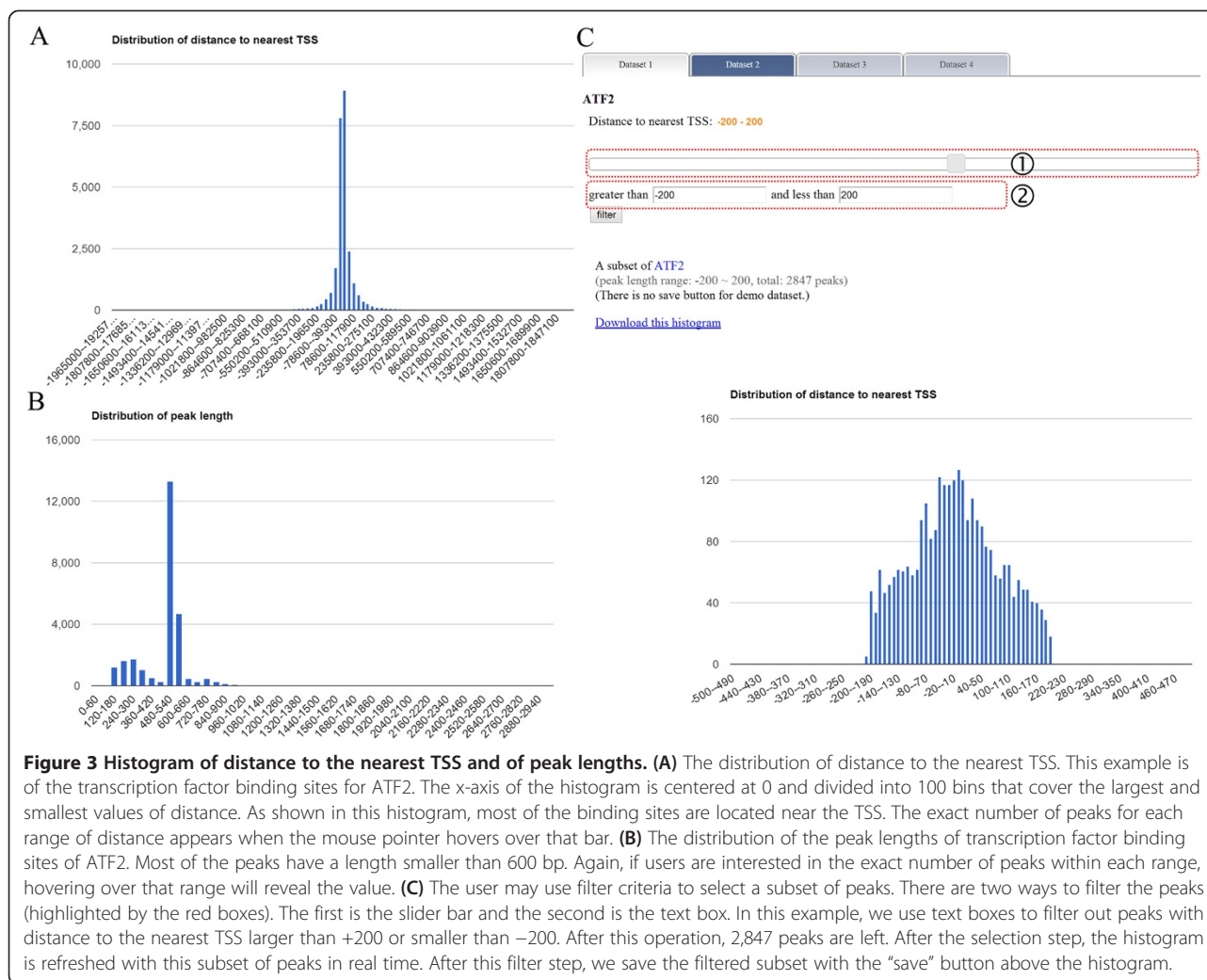
Figure 3 Histogram of distance to the nearest TSS and of peak lengths. (A) The distribution of distance to the nearest TSS. This example is of the transcription factor binding sites for ATF2. The x-axis of the histogram is centered at 0 and divided into 100 bins that cover the largest and smallest values of distance. As shown in this histogram, most of the binding sites are located near the TSS. The exact number of peaks for each range of distance appears when the mouse pointer hovers over that bar. **(B)** The distribution of the peak lengths of transcription factor binding sites of ATF2. Most of the peaks have a length smaller than 600 bp. Again, if users are interested in the exact number of peaks within each range, hovering over that range will reveal the value. **(C)** The user may use filter criteria to select a subset of peaks. There are two ways to filter the peaks (highlighted by the red boxes). The first is the slider bar and the second is the text box. In this example, we use text boxes to filter out peaks with distance to the nearest TSS larger than +200 or smaller than −200. After this operation, 2,847 peaks are left. After the selection step, the histogram is refreshed with this subset of peaks in real time. After this filter step, we save the filtered subset with the "save" button above the histogram.

length of sequences and the DNA sequences. Our tools can also extract sequences for the subsets of uploaded files. As we already saved a subset of peaks (ATF2_-200-200_TSS and ATF3_-200-200_TSS) in the previous step, we can retrieve the sequences for these two saved subsets by clicking on "Get peak sequences" on the menu.

ChIPseek further allows peak comparison between two peak files. As a demonstration, we performed a comparison using the two saved subsets of peaks, ATF2_-200-200_TSS and ATF3_-200-200_TSS. In Figure 4, the results of the comparison are displayed in four separate tabs. The first tab is a Venn diagram that shows the number of unique peaks for each dataset and the number of overlapped peaks between the two datasets. Users should note that the overlapped peak section might have redundant peaks. For example, if one peak from ATF3_-200-200_TSS is overlapped with two peaks in ATF2_-200-200_TSS, then they will be counted twice. Therefore, the total number of unique peaks plus overlapped peaks may not necessarily be equal to the total number of peaks in

the original datasets. The second and third tabs show tables of unique peaks for each dataset. As shown in Figure 5, the final tab contains tables of various chromosomes with the listed overlapped peaks. In addition to directly comparing the users' datasets, ChIPseek also includes a database of binding sites for 161 transcription factors generated by ChIP-seq from the ENCODE project [26]. Users may compare their datasets with the experimentally based transcription factor binding sites. As the ENCODE project is focused on human genome, the comparison of transcription factor binding sites is only available for human datasets.

Demo dataset 2: transcriptional repressor CTCF binding sites

CTCF is known as a DNA methylation-dependent chromatin insulator. Aberrant methylation pattern of the CTCF binding sites of an imprint gene *Igf2/H19* causes the development of Beckwith-Wiedemann Syndrome (BWS) [46-48]. Children with BWS have an over-growth

Figure 4 Comparison between binding sites of ATF2 and ATF3 and Venn diagram. After selecting ATF2_-200-200_TSS and ATF3_-200-200_TSS for comparison, ChIPseek compares peaks from these two datasets. The overall comparison result is shown as a Venn diagram in the first tab. As shown here, a total of 1,359 peaks are ATF2 unique, 2,499 peaks are ATF3 unique and 1,499 peak pairs are overlapped between the biding sites of ATF2 and ATF3. The detailed peak information for unique peaks and overlapped peak pairs can be found in the following three tabs.

syndrome and predispositions to develop pediatric tumors including Wilms tumor [49]. We downloaded the binding sites of CTCF as our demo dataset. The CTCF dataset is uploaded as a BED file. To demonstrate how to filter the dataset by the length of peaks, we used "Peak length" filter criteria to obtain a subset of peaks that have length range from 100 bp to 150 bp. After that, a subset of 67 peaks out of the original 162,209 peaks is selected and saved. ChIPseek automatically named this subset "CTCF_100-150_len". The suffix "len" indicates that the peaks in this subset have been filtered according to the peak length.

As mentioned in the previous section, users can use multiple filter criteria to narrow down the number of regions that they are interested in. For example if the user want to narrow down the peaks located at promoter regions due to the fact that the transcription factor binding site near the "promoter-TSS" is consider to have a greater biological impact on the activity of the promoter. By selecting peaks located within "promoter-TSS" under the "Peak location" in the filter criteria, the user can filter a total of 11,328 out of 162,209 peaks. On the other

hand, if users already have lists of genes which they are interested in, they could also upload lists of genes and filter peaks based on these genes. In addition to use a list of genes for filtering, ChIPseek also offers built-in functional annotations downloaded from GO and KEGG [28,29] which allow users to generate gene lists based on their interested biological function or pathways. As the annotation functions from GO and KEGG are both constructed in tree structures, users may select interested functions one by one or directly select a whole branch from the annotation tree. It should be noted that, nodes that are belonged to more than one parent, will be listed multiple times under different parents separately in the annotation tree. By allowing this kind of redundancy, ChIPseek can guarantee that the created gene list includes all possible genes related to these selected function groups. Subsets of filtered peaks can be used in further analysis or filtered again by other filter criteria. Users can also remove selected files too if they found some results are not necessary for further analysis.

To illustrate the distribution of peaks on chromosomes, users may click "Plot peaks on ideogram" on the

Peaks found in both A.) ATF2_-200~200_TSS and B.) ATF3_-200~200_TSS
There are a total of 1499 peak pairs. Download all peaks in text format

chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

chr13

Go to the specific genomic location on UCSC genome browser:

chromsome: [_____] start: [_____] end: [_____] [go]

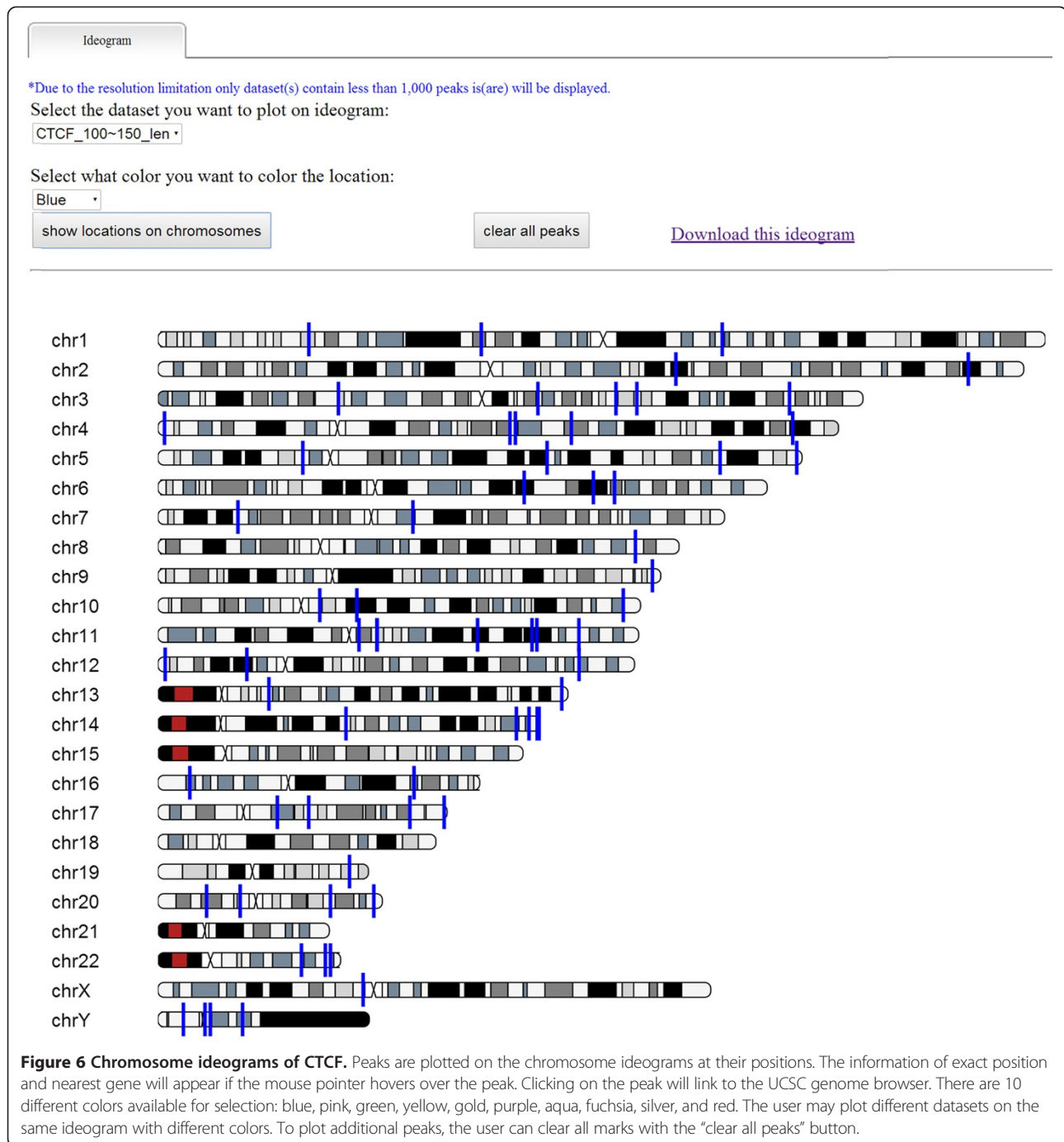| start (A)	end (A)	start (B)	end(B)	relative position		link to USCS
101326997	101327517	101327142	101327378	A B		Genome Browser
103498135	103498276	103497982	103498251	A B		Genome Browser
113951196	113951776	113951142	113951635	A B		Genome Browser
21871903	21872423	21872158	21872472	A B		Genome Browser
24463416	24463936	24463453	24463683	A B		Genome Browser

Figure 5 Overlapped peak pairs. At the top of this page is shown how many peak pairs are found, and a link is provided to download all peak pairs with their annotation. All overlapped peak pairs are separated into different tables according to their location. As shown here, the table lists peak pairs located on chromosome X. The first four columns list the start and end positions of peaks. The fifth column shows the relative positions of the peak pairs. The last column provides links to the UCSC genome browser for that region of interest. Clicking on the title of each column can sort that column.

menu. After that, users may select the dataset and the color bar (with 10 different colors), and click "show locations on chromosomes". Due to the resolution limitation, only datasets having less than 1000 peaks can be displayed in this way. Users may display up to 10 different datasets in different colors on the ideograms. In our demo data, dataset "CTCF_100-150_len" is selected, and the locations of 67 peaks are indicated on the chromosomes by blue bars as shown in Figure 6. The chromosome ideogram provides an intuitive way to visualize the overall genomic distribution of all peaks. Currently, the chromosome ideogram is only available for human assemblies (hg17, hg18 and hg19).

To identify the enriched motif of CTCF biding sites, users may click on "find motif" in the menu. The users need to specify the range of the enriched motifs. ChIPseek utilizes HOMER for motif identification. According to the user manual of HOMER, 50 bp (±25 bp around the centers of peaks) should be enough to identify the motif for a transcription factor. Alternatively, users may choose 200 bp (±100 bp around the center), which should be enough to find both the primary and "co-enriched" motifs for a transcription factor. As for histone marked regions, ranges from 500–1000 bp (±250- ± 500 bp around the center) may be suitable regions to search. To satisfy all ChIP studies focused on transcription factors, histone modifications and CpG islands, ChIPseek provides four

different options: ±25 bp, ±100 bp, ±250 bp and ±500 bp. Users can specify how many nucleotides around the centers of peaks to use as target regions for domain identification according to the users' experimental purpose. It is worth noting that motif identification is time-consuming and the larger the selected regions the longer it will take to identify the consensus sequences. In our case, because CTCF is a DNA binding protein with known consensus sequence CTCCC, we selected ±25 bp around the center of peaks to search for motifs and the results are shown in Figure 7.

Users should notice that ChIPseek can identify enriched motifs even if they input random data due to the large scale of input data. A significant p-value may be achieved just by random chance. Because ChIPseek uses HOMER to identify potential motifs, ChIPseek only reports motifs that have p-values smaller than 1e-10, which is the cutoff suggested by the HOMER website. However, there may still be some false positives after this filter. Therefore, it is still very important for the users to determine which patterns may truly be recognized by the binding protein. Users may identify most promising candidates by checking the p-value distribution and setting a customized p-value cutoff. As ChIPseek shows p-values in ascending order, the p-value followed by another dramatically increased p-value may be the proper cutoff. Another clue for candidate motif selection is that
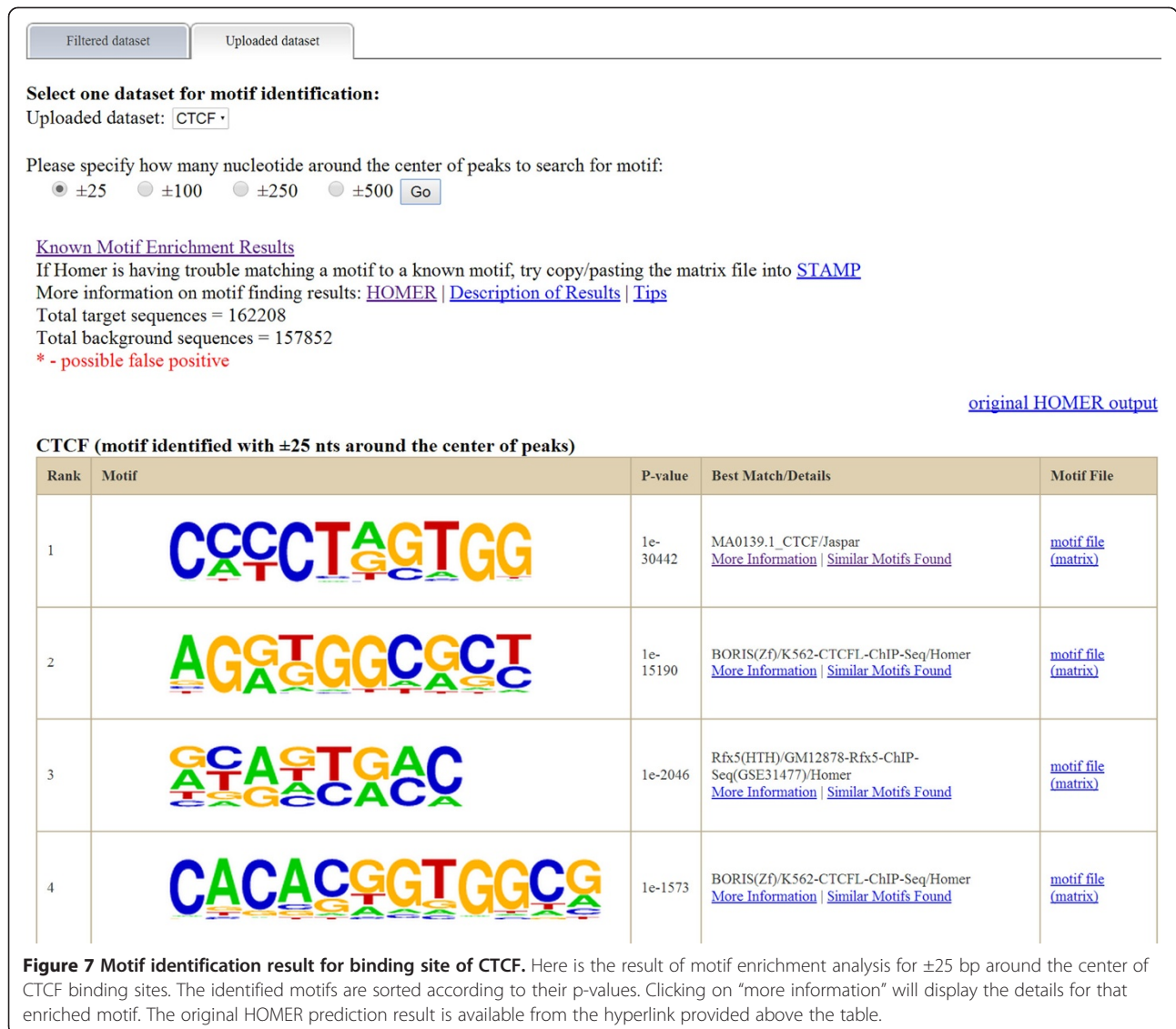
Figure 6 Chromosome ideograms of CTCF. Peaks are plotted on the chromosome ideograms at their positions. The information of exact position and nearest gene will appear if the mouse pointer hovers over the peak. Clicking on the peak will link to the UCSC genome browser. There are 10 different colors available for selection: blue, pink, green, yellow, gold, purple, aqua, fuchsia, silver, and red. The user may plot different datasets on the same ideogram with different colors. To plot additional peaks, the user can clear all marks with the "clear all peaks" button.

different offsets of a true motif may appear in the report several times. For example, in Figure 7, the pattern "GGTGG" shows up several times at the top four sequence logos. Therefore, this GGTGG is likely to be part of a real motif. As the binding consensus sequence of CTCF in the human genome has already been investigated in a previous study [50], we found this pattern to indeed be present in the consensus binding sequences of CTCF. Meanwhile, many motifs have been identified,

and their consensus sequences are well known. ChIPseek also uses HOMER to perform motif enrichment analysis for those known motifs. Users may also explore known motifs by following the hyperlink "Known Motif Enrichment Results" above the table.

Conclusions

Here we present ChIPseek as a web tool for analyzing ChIP-seq and ChIP-chip data. By integrating HOMER

Figure 7 Motif identification result for binding site of CTCF. Here is the result of motif enrichment analysis for ±25 bp around the center of CTCF binding sites. The identified motifs are sorted according to their p-values. Clicking on "more information" will display the details for that enriched motif. The original HOMER prediction result is available from the hyperlink provided above the table.

and BEDTools, ChIPseek is an easy-to-use software package for the first-time user with only basic computer skills who wants to analyze ChIP data. ChIPseek guides the users step-by-step to obtain the peak annotation, locations, sequences, and useful statistics as charts and histograms for visualizing the properties of the peaks. ChIPseek also provides UCSC genome browser links so that the users can investigate peaks further. A unique feature of our tool is that it includes filter tools which allow users to select interested peak subsets based on peak lengths, distance to the nearest TSS, peak location, user uploaded gene lists or genes belong to user selected functions. For human datasets, users may also use ChIPseek to plot peaks on an ideogram, which offers an overview of genomic distribution across different chromosomes. When comparing two datasets, ChIPseek generates a Venn diagram, lists of unique peaks for each dataset and overlapped peak pairs with their relative positions. We also showed that ChIPseek can help users identify enriched binding motifs for transcription factors and DNA binding protein factors. In addition to these convenient analysis tools, ChIPseek also has a built-in database of human transcription factor binding sites available for comparison. In summary, ChIPseek includes many desired functions that are available through an intuitive user interface. ChIPseek is a free, web-based service. The use of ChIPseek requires no knowledge of Linux, no programming, no installation and no login. The ChIP data analysis tools and databases provided by ChIPseek will offer invaluable assistance for biologists whose studies focus on protein-DNA interaction, gene transcription regulation, epigenetics, chromatin organization and cancer biology.

Availability and requirements

ChIPseek is freely available at http://chipseek.cgu.edu.tw and all the source codes of pipelines and parsers are also available upon request. Currently, ChIPseek can support most commonly used browsers such as Chrome, Firefox, Internet Explorer and Safari.

Abbreviations

ChIP: Chromatin immunoprecipitation; TSSs: Transcription start sites; bp: Base pair; UTR: Untranslated region; TTS: Transcription termination site; BWS: Bechwith-Wiedemann Syndrome.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TWC and CYL built the pipeline. CY, THW and CCL designed and constructed the website. TWC and HPL wrote the manuscript together. RCG and PJH maintain the system. PT supervised and revised the manuscript. All authors read and approved the final manuscript.

Author details

[1]Molecular Medicine Research Center, Chang Gung University, Taoyuan, Taiwan. [2]Bioinformatics Center, Chang Gung University, Taoyuan, Taiwan. [3]Graduate Institute of Biomedical Sciences, Chang Gung University, Taoyuan, Taiwan. [4]Department of Microbiology and Immunology, Medical School of Chang Gung University, Taoyuan, Taiwan. [5]Department of Biological Science and Technology, National Chiao Tung University, HsinChu, Taiwan. [6]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan.

References

1. Solomon MJ, Larsen PL, Varshavsky A: **Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene.** *Cell* 1988, **53**(6):937–947.
2. Kuo MH, Allis CD: **In vivo cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment.** *Methods* 1999, **19**(3):425–433.
3. Weinmann AS, Farnham PJ: **Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation.** *Methods* 2002, **26**(1):37–47.
4. Weinmann AS, Yan PS, Oberley MJ, Huang TH-M, Farnham PJ: **Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis.** *Genes Dev* 2002, **16**(2):235–244.
5. Lee SH, Kim J, Kim WH, Lee YM: **Hypoxic silencing of tumor suppressor RUNX3 by histone modification in gastric cancer cells.** *Oncogene* 2009, **28**(2):184–194.
6. Hitchler MJ, Rice JC: **Genome-wide epigenetic analysis of human pluripotent stem cells by ChIP and ChIP-Seq.** *Methods Mol Biol* 2011, **767**:253–267.
7. Rechtsteiner A, Ercan S, Takasaki T, Phippen TM, Egelhofer TA, Wang W, Kimura H, Lieb JD, Strome S: **The histone H3K36 methyltransferase MES-4 acts epigenetically to transmit the memory of germline gene expression to progeny.** *PLoS Genet* 2010, **6**(9):e1001091.
8. Levy A, Eyal M, Hershkovits G, Salmon-Divon M, Klutstein M, Katcoff DJ: **Yeast linker histone Hho1p is required for efficient RNA polymerase I processivity and transcriptional silencing at the ribosomal DNA.** *Proc Natl Acad Sci U S A* 2008, **105**(33):11703–11708.
9. Paradowska A, Fenic I, Konrad L, Sturm K, Wagenlehner F, Weidner W, Steger K: **Aberrant epigenetic modifications in the CTCF binding domain of the IGF2/H19 gene in prostate cancer compared with benign prostate hyperplasia.** *Int J Oncol* 2009, **35**(1):87–96.
10. Schmid CD, Bucher P: **ChIP-Seq data reveal nucleosome architecture of human promoters.** *Cell* 2007, **131**(5):831–832. author reply 832–833.
11. 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* 2008, **36**(16):5221–5231.
12. Kharchenko PV, Tolstorukov MY, Park PJ: **Design and analysis of ChIP-seq experiments for DNA-binding proteins.** *Nat Biotechnol* 2008, **26**(12):1351–1359.
13. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A: **Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.** *Nat Methods* 2008, **5**(9):829–834.
14. Qin ZS, Yu J, Shen J, Maher CA, Hu M, Kalyana-Sundaram S, Yu J, Chinnaiyan AM: **HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data.** *BMC Bioinformatics* 2010, **11**:369.
15. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.** *Nat Biotechnol* 2009, **27**(1):66–75.
16. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9**(9):R137.
17. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: **An integrated software system for analyzing ChIP-chip and ChIP-seq data.** *Nat Biotechnol* 2008, **26**(11):1293–1300.
18. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Mol Cell* 2010, **38**(4):576–589.
19. Sadeghi L, Bonilla C, Stralfors A, Ekwall K, Svensson JP: **Podbat: a novel genomic tool reveals Swr1-independent H2A.Z incorporation at gene coding sequences through epigenetic meta-analysis.** *Plos Comput Biol* 2011, **7**(8):e1002163.
20. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y, Pape UJ, Poidinger M, Chen Y, Yeung K, Brown M, Turpaz Y, Liu XS: **Cistrome: an integrative platform for transcriptional regulation studies.** *Genome Biol* 2011, **12**(8):R83.
21. Halachev K, Bast H, Albrecht F, Lengauer T, Bock C: **EpiExplorer: live exploration and global analysis of large epigenomic datasets.** *Genome Biol* 2012, **13**(10):R96.
22. Heger A, Webber C, Goodson M, Ponting CP, Lunter G: **GAT: a simulation framework for testing the association of genomic intervals.** *Bioinformatics* 2013, **29**(16):2046–2048.
23. Zambelli F, Pesole G, Pavesi G: **PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments.** *Nucleic Acids Res* 2013, **41**(W1):W535–W543.
24. Huang W, Loganantharaj R, Schroeder B, Fargo D, Li L: **PAVIS: a tool for Peak Annotation and Visualization.** *Bioinformatics* 2013, **29**(23):3097–3099.
25. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**(6):841–842.
26. Consortium EP: **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS Biol* 2011, **9**(4):e1001046.
27. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, *et al*: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.** *Genome Res* 2012, **22**(9):1813–1831.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25–29.
29. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36**(Database issue):D480–D484.
30. **Google Charts.** [https://developers.google.com/chart/]

31. mod-python. [http://modpython.org/]
32. jQuery. [http://jquery.org/]
33. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: **BigWig and BigBed: enabling browsing of large distributed datasets.** *Bioinformatics* 2010, **26**(17):2204–2207.
34. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ: **Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser.** *Bioinformatics* 2014, **30**(7):1003–1005.
35. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14**(6):1188–1190.
36. Ghostscript. [http://pages.cs.wisc.edu/~ghost/]
37. Ideogram Viewer. [http://bioinformatics.mdanderson.org/main/Ideogram Viewer:Overview]
38. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ: **The UCSC Genome Browser database: 2014 update.** *Nucleic Acids Res* 2014, **42**(1):D764–D770.
39. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2006.** *Nucleic Acids Res* 2006, **34**(Database issue):D590–D598.
40. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Res* 2009, **37**(Database issue):D32–D36.
41. UniProt C: **Update on activities at the Universal Protein Resource (UniProt) in 2013.** *Nucleic Acids Res* 2013, **41**(Database issue):D43–D47.
42. Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA: **Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL.** *Nucleic Acids Res* 1998, **26**(1):362–367.
43. TFSEARCH: Searching Transcription Factor Binding Sites. [http://www.cbrc.jp/research/db/TFSEARCH.html]
44. Bailey TL: **DREME: motif discovery in transcription factor ChIP-seq data.** *Bioinformatics* 2011, **27**(12):1653–1659.
45. Bailey TL, Machanick P: **Inferring direct DNA binding from ChIP-seq.** *Nucleic Acids Res* 2012, **40**(17):e128.
46. Szabo PE, Tang SH, Silva FJ, Tsark WM, Mann JR: **Role of CTCF binding sites in the Igf2/H19 imprinting control region.** *Mol Cell Biol* 2004, **24**(11):4791–4800.
47. Kanduri C, Pant V, Loukinov D, Pugacheva E, Qi CF, Wolffe A, Ohlsson R, Lobanenkov VV: **Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive.** *Curr Biol* 2000, **10**(14):853–856.
48. Weksberg R, Smith AC, Squire J, Sadowski P: **Beckwith-Wiedemann syndrome demonstrates a role for epigenetic control of normal development.** *Hum Mol Genet* 2003, **12**(Spec No 1):R61–R68.
49. DeBaun MR, Tucker MA: **Risk of cancer during the first four years of life in children from The Beckwith-Wiedemann Syndrome Registry.** *J Pediatr* 1998, **132**(3 Pt 1):398–400.
50. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B: **Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.** *Cell* 2007, **128**(6):1231–1245.