CASSys: an integrated software-system for the interactive analysis of ChIP-seq data

Malik Alawi[1,2], Stefan Kurtz[1] and Michael Beckstette[1,*]

[1]Center for Bioinformatics, University of Hamburg, Bundesstraße 43, D-20146 Hamburg, Germany, http://www.zbh.uni-hamburg.de

[2]Hamburg Center for Experimental Therapy Research (HEXT), University Medical Center Hamburg-Eppendorf, Martinistraße 52, D-20246 Hamburg, Germany, http://www.uke.de

Summary

The mapping of DNA-protein interactions is crucial for a full understanding of transcriptional regulation. Chromatin-immunoprecipitation followed by massively parallel sequencing (ChIP-seq) has become the standard technique for analyzing these interactions on a genome-wide scale. We have developed a software system called *CASSys* (**C**hIP-seq data **A**nalysis **S**oftware **Sys**tem) spanning all steps of ChIP-seq data analysis. It supersedes the laborious application of several single command line tools. *CASSys* provides functionality ranging from quality assessment and -control of short reads, over the mapping of reads against a reference genome (readmapping) and the detection of enriched regions (peakdetection) to various follow-up analyses. The latter are accessible via a state-of-the-art web interface and can be performed interactively by the user. The follow-up analyses allow for flexible user defined association of putative interaction sites with genes, visualization of their genomic context with an integrated genome browser, the detection of putative binding motifs, the identification of over-represented Gene Ontology-terms, pathway analysis and the visualization of interaction networks. The system is client-server based, accessible via a web browser and does not require any software installation on the client side. To demonstrate *CASSys*'s functionality we used the system for the complete data analysis of a publicly available Chip-seq study that investigated the role of the transcription factor estrogen receptor-α in breast cancer cells.

1 Introduction

ChIP-seq, chromatin immunoprecipitation followed by massively parallel sequencing, is a method for profiling DNA-protein interactions on a genome-wide scale. Offering a higher resolution and benefiting from the decreasing costs of second-generation sequencing, ChIP-seq is replacing its array-based predecessor ChIP-chip as the method of choice for identifying transcription factor binding sites and histone modifications. In a typical ChIP-seq experiment, immunoprecipitated DNA fragments, about 200 bp long, are sequenced from both ends with modern short read sequencing technology. As a result, millions of short sequence reads with a length of 25 - 100 bp are obtained. These short reads are the starting point for the subsequent computational data analysis. Figure 1 shows, that the computational analysis of ChIP-seq experiments consists of four sequential steps. The first step is usually the quality assessment and

[*]To whom correspondence should be addressed. Email: beckstette@zbh.uni-hamburg.de

doi:10.2390/biecoll-jib-2011-155

control (QA/QC) of raw reads. Reads satisfying the applied quality criteria are then mapped against a reference genome (readmapping) and from these mappings the locations of read enriched regions are determined to identify sites where the immunoprecipitated protein and DNA interact (peakdetection). These are called *interaction sites* in the following. To understand mechanisms of transcriptional regulation it is not sufficient to know the mere locations of interaction sites. Therefore, subsequently to peakdetection, further steps, like the analysis of protein-protein interactions, are required. These follow-up analyses are diverse and often require the integration of genomic annotation data and other information.

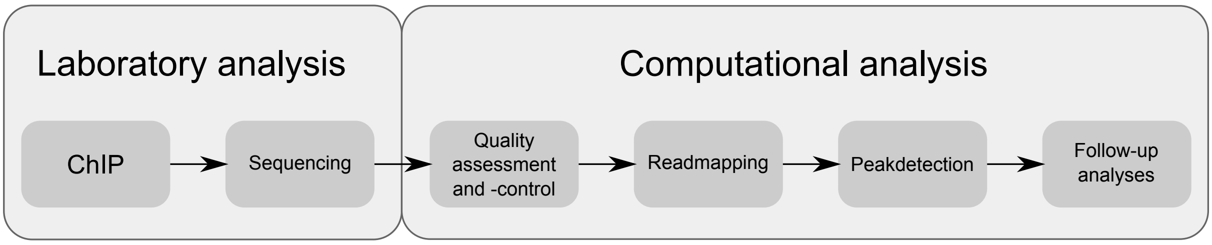


Figure 1: The typical workflow of a ChIP-seq experiment can be divided into a laboratory and a computational part. The laboratory part covers chromatin-immunoprecipitation (ChIP) and massively parallel sequencing. The computational data analysis part consist of four sequentially performed steps. Namely, (1) the quality control and assessment of short reads, (2) their mapping against a reference genome and (3) the derivation of putative interaction sites. In a final step (4) various follow-up analyses to elucidate the biological implications of identified interaction sites can follow.

For readmapping and peakdetection several programs have been developed, for a recent review see [33, 24]. While the readmapping tools are used in several other applications of short read sequencing, the peakdetection tools are specific to ChIP-seq data analysis. Most of these programs are command line tools. Stand-alone software is also available for some follow-up analyses. Especially for the detection of sequence motifs several established programs [5, 21, 23, 25] exist. Standard genome browsers, like the *UCSC Genome Browser* [17], can be used to visualize binding sites within their genomic context. But overall, computational analysis of ChIP-seq data remains a patchwork of manual application of different tools.

Some software packages simplify the computational analyses of ChIP-seq data by providing graphical interfaces for peakdetection or integrating follow-up analyses. *W-ChIPeaks* [19] and *ChIP-Seq web server* [1] are examples for peak-detection tools with a web-based user interface. *Sole-Search* [8] additionally provides basic statistics on interaction sites and identifies the genes closest to each site. The most advanced system in this field is *CisGenome* [15], which also integrates a genome browser and provides motif detection capability. The CisGenome GUI can, however, only be used on MS Windows platforms. *MICSA* [9] employs *FindPeaks* [11] for peakdetection and aims at enhancing its results by removing putative interaction sites not bearing certain sequence motifs. Although the tools mentioned above can simplify and improve certain aspects of ChIP-seq analysis, none of them covers all steps of the analysis, see Table 1. With *CASSys* (ChIP-seq Analysis Software System) we present a software system integrating all steps of computational ChIP-seq data analysis.

	W-ChIPeaks	*ChIP-Seq Web Server*	*Sole-Search*	*MICSA*	*CisGenome*	*CASSys*
Quality assessment and control						√
Readmapping						√[1,2]
Peakdetection	√	√	√	√	√[4]	√[3,4]
Motif detection				√[8]	√[5]	√[5,6]
Motif comparison						√[7]
Genome browser					√	√
Over-representation analysis						√
Pathway analysis						√
Interaction analysis						√
Web interface	√	√	√			√
Incorporation of genomic annotations	√		√		√	√

[1] Bowtie [20], [2] BWA [22], [3] MACS [34], [4] FindPeaks [11], [5] MEME [5], [6] Weeder [25], [7] Tomtom [13], [8] flexmodule

Table 1: Most ChIP-seq analysis pipelines focus on providing a graphical user-interface for peakdetection and hardly offer further analysis capabilities following this step. Except for *CASSys*, no system supports readmapping as well as peakdetection, which are absolutely necessary for analysis of ChIP-seq data. In addition, *CASSys* offers an integrated QA/QC processing step and the widest range of possible follow up analyses. Hence, it supports the whole computational data analysis ranging from the processing of raw reads to follow up analyses like motif detection, pathway analysis and interaction network visualization.

2 Methods

CASSys divides the computational analyses into two parts. The first part spans the steps from QA/QC to peakdetection. Since these steps are sequentially executed and—once started—do not require user-interaction, we refer to them as *asynchronous analyses* in the following. Results of the asynchronous part are stored in a local database from where they can be efficiently retrieved for the *interactive analyses* subsequent of the peakdetection step. For persistent storage *CASSys* employs the open source object-relational database system *PostgreSQL*. For an overview of *CASSys*'s system architecture see Figure 2.

2.1 Asynchronous analyses

To determine the location of interaction sites *CASSys* employs the concept of *workflows*. Here a workflow describes the tools and parameters used for QA/QC, readmapping and peakdetection. These sets of parameters are stored in the database. The workflow concept has the following benefits:

- A workflow completely documents the different steps of the asynchronous analysis allowing to fully reproduce them.
- Parts of a workflow can be (re)used in different combinations. If, for example in two workflows, only the parametersets for peakdetection differ, it is not required to repeat the other two steps which came before peakdetection.

The QA/QC step starts with the parsing of short reads in the widespread FastQ-format. A filter removes prefixes, suffixes or whole reads according to user-specified quality and length parameters. The mean quality and base-frequency of each position common to all reads is assessed before and after filtering. The results of this step are summarized, stored in the database and can be displayed and visualized in the systems web interface.

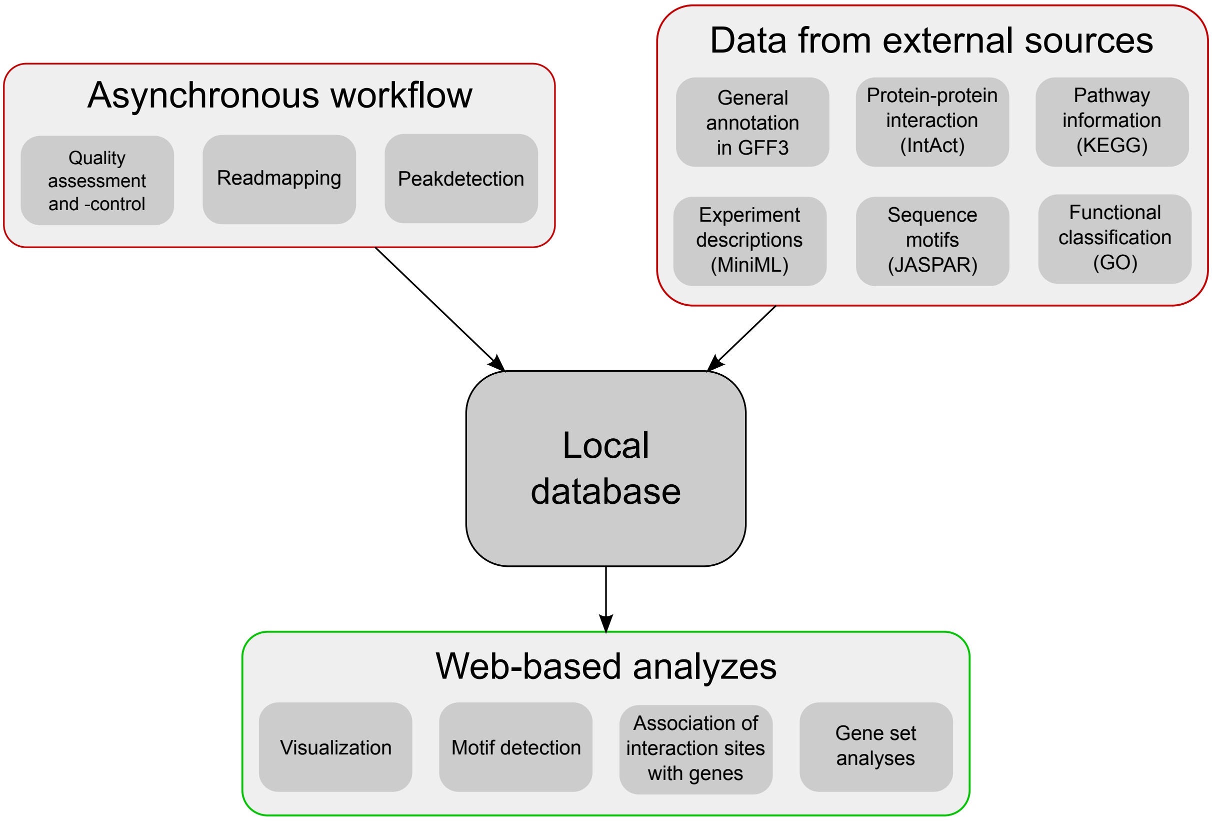

Asynchronous workflow
Quality assessment and -control
Readmapping
Peakdetection
Data from external sources
General annotation in GFF3
Protein-protein interaction (IntAct)
Pathway information (KEGG)
Experiment descriptions (MiniML)
Sequence motifs (JASPAR)
Functional classification (GO)
Local database
Web-based analyzes
Visualization
Motif detection
Association of interaction sites with genes
Gene set analyses

Figure 2: An overview of *CASSys*'s system architecture. The asynchronous part (red) spans the computational analysis steps up to peakdetection and the parsing of data from external sources. Refined data resulting from these processes is persistently stored in a local database, and can be accessed in the web-based analyses of the interactive part (green).

CASSys employs a selection of widely used software tools for readmapping and peakdetection. Moreover it offers the infrastructure to quickly integrate new tools. This allows users to choose from different methods for these central processing steps depending on data quality and experiment design. In particular, different peak-calling algorithms show severe variations in the obtained results on different datasets [31] which may lead to dramatic changes in the drawn biological conclusions [18]. Currently, for readmapping *CASSys* integrates *Bowtie* [20] and *BWA* [22], while for peakdetection the system employs *MACS* [34] and *FindPeaks* [11]. Besides determining interaction sites, the asynchronous part covers the parsing of annotation data. This process is required only once for each studied species. The annotations provide a biological context for interpreting interaction sites. Genomic annotations can be provided in GFF3-format [2]. In addition, *CASSys* parses other annotation from *IntAct* [3] and *KEGG* [16] in the respective formats. Experiment descriptions are imported in *MiniML* (MIAMI Notation In Markup Language) format. This allows to easily import complete experiments with their associated data files from NCBI's Gene Expression Omnibus (*GEO*) [6] which, as of April 2011 contains *MiniML*-files for over 570 studies.

In ChIP-seq data analysis the identification of genes that are co-regulated by the immunoprecipitated protein is of central interest. This requires to associate interaction sites identified in the peakdetection step with annotated genes. To address this, *CASSys* calculates, based on the parsed annotation data and the detected putative interaction sites, the distance between the center of each site and the closest transcription start of a gene occurring upstream or downstream on the forward or the reverse strand. In this way, every site is initially associated with up to four genes. *CASSys* also provides the option to only include functional genes when associating interaction sites and genes. A gene is considered to be functional if and only if it is listed in the *KEGG* database. This allows to ignore pseudogenes and other non-coding elements when determining the set of site/gene associations. As a consequence, up to eight genes can be associated with each interaction site. This number results from the combination of the features upstream/downstream, forward/reverse and functional/not necessarily functional. For each interaction site *CASSys* considers these (up to) eight genes to be the most likely regulated genes and stores them in the local database. Observe that this precalculated set of interaction site/gene associations only constitutes the set of all possible associations from which in the interactive analyses subsets can be generated based on user defined criteria. This precalculation optimizes access times and avoids redundant calculation during the interactive analyses.

2.2 Interactive analyses

The analyses subsequent to the peakdetection step are interactive and accessible via *CASSys*' web interface. This is designed to be used by experimentalists. The parameters of all software tools used in the interactive analyses can be defined in the web interface. As the different analysis steps are very fast and all data is retrieved from the local database, the effects of changing parameters are instantly visible.

CASSys supports the following three types of web-based analyses which are also depicted in Figure 3:

- The derivation and analysis of sets of candidate genes (gene set analyses).
- The detection and interpretation of sequence motifs in interaction sites (motif detection).
- The visualization and comparison of interaction sites in their genomic context (genome browsing).

The analyses rely on information which was stored in the database during the asynchronous part of *CASSys*. As *CASSys* uses a local database to connect the asynchronous and interactive analyses, it does not require access to third party databases at runtime.

2.2.1 Motif detection

If the detected sites interact with the same protein, they likely share common binding motifs. Therefore, *CASSys* allows interaction sites to be screened for common motifs using the established motif detection programs *Weeder* [25] and *MEME* [5]. *Weeder* provides an enumerative approach to motif discovery, while *MEME* is based on a probabilistic model. Results are post-processed and a report with summary statistics including sequence logos of each detected motif is generated. The motifs can automatically be compared to known motifs from *JASPAR* [26] using the *Tomtom* [13] program. Pairwise alignments of high-ranking database motifs and detected motifs are reported along with corresponding E-values.

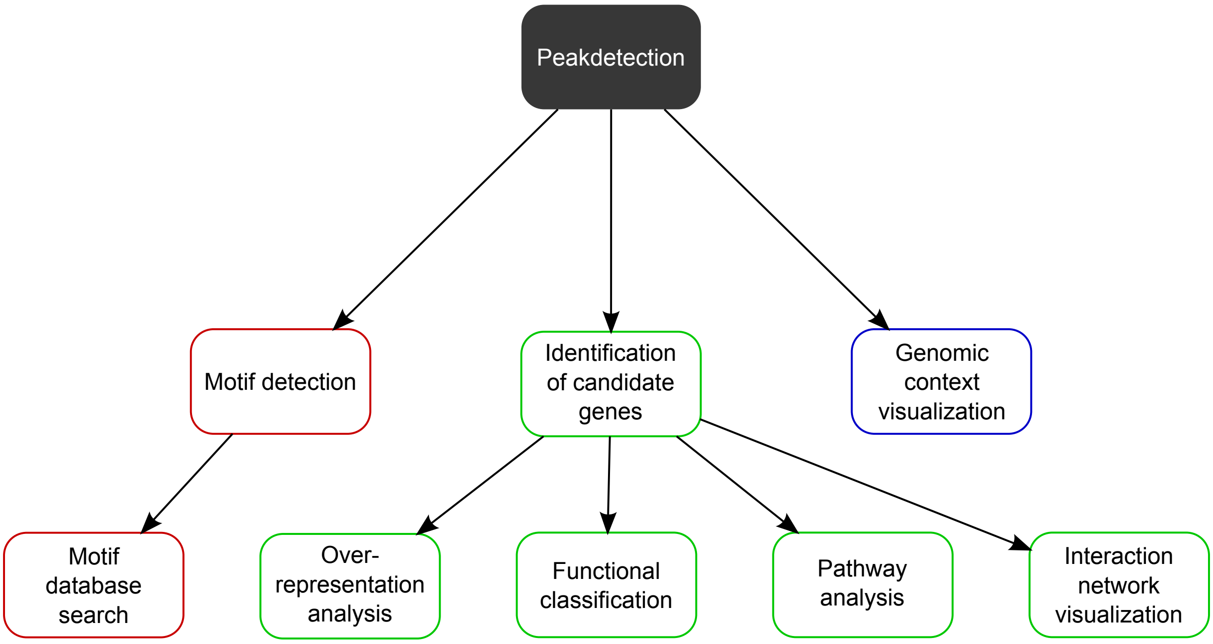

Peakdetection
Motif detection
Identification of candidate genes
Genomic context visualization
Motif database search
Over-representation analysis
Functional classification
Pathway analysis
Interaction network visualization

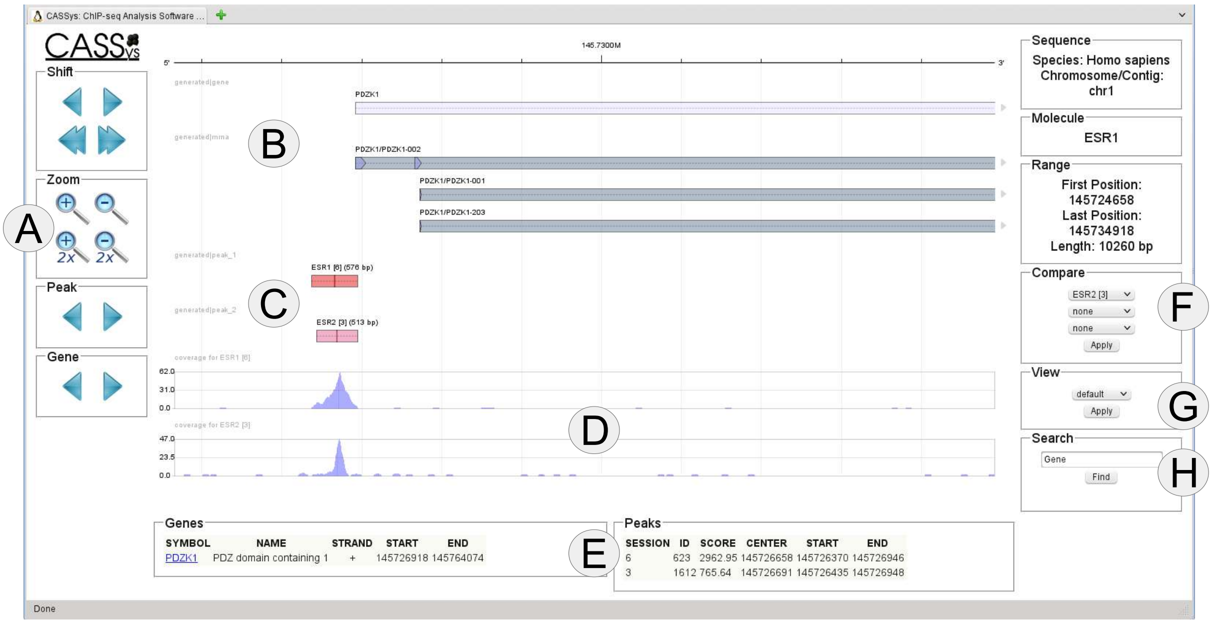
Figure 3: Dataflow in *CASSys* interactive part. The interaction sites detected in the peakdetection step of the asynchronous analysis part are the starting point for three different types of interactive analyses in *CASSys*. The sites are screened for sequence motifs and detected motifs can then be compared to known motifs listed in motif databases (red). A set of candidate genes is constructed and further analyzed (green). The interaction sites can be explored in their genomic context using the *CASSys*' genome browser (blue).

2.2.2 Genome browsing

CASSys implements its own genome browser. The browser is fully integrated into the system and visualizes interaction sites and corresponding fragment coverage data, within their genomic context and augments them with a tabular view containing detailed information on displayed features. Moreover, the genome browser allows the visual comparison of multiple datasets. The level of detail and different color-schemes of annotated genes can be modified by the user. Transcripts, for example, can be visualized as separate items or they can be collapsed to genes. For a screenshot of *CASSys*' integrated genome browser, which makes use of the genome annotation drawing library *AnnotationSketch* [30], see Figure 4.

2.2.3 Derivation and analysis of gene sets

CASSys allows to interactively derive a set of likely co-regulated genes based on user defined criteria from the precalculated set of possible interaction site/gene associations determined in the asynchronous part. These criteria include maximal up- and downstream distances from transcription start sites of genes to the centers of interaction sites, quality values of the interaction sites, and membership of the genes in the *KEGG*-database. Since multiple candidate genes may be associated with a single binding site, the user can also decide whether in such cases all, none or only the gene closest to the center of the interaction site should remain in the set. Summary statistics, including the median quality and length of associated interaction sites as well as the median distance between their center and the transcription start sites of genes, are generated in form of tables. The tables can be exported in CSV-format (comma-separated val-



Sequence
Species: Homo sapiens
Chromosome/Contig: chr1

Molecule
ESR1

Range
First Position: 145724658
Last Position: 145734918
Length: 10260 bp

Genes

SYMBOL	NAME	STRAND	START	END
PDZK1	PDZ domain containing 1	+	145726918	145764074

Peaks

SESSION	ID	SCORE	CENTER	START	END
6	623	2962.95	145726658	145726370	145726946
3	1612	765.64	145726691	145726435	145726948

Figure 4: The *CASSys*'s genome browser displaying two ChIP-seq datasets. Navigation elements provide functionality for shifting, zooming and jumping to certain features (A). Tracks display genomic annotation like genes and transcripts (B). DNA-protein interaction sites of different datasets are color-coded (C). The original coverage of precipitated DNA-fragments is shown in an extra track (D). Displayed genes and interaction sites are also provided in an additional tabular view (E). Multiple datasets can be selected for visual comparison (F). The level of detail and different color-schemes of annotated genes can be chosen (G). A search box allows searching specific genes (H).

ues format) and the sequences of the interaction sites can be downloaded in Fasta-format. This set of candidate genes is the starting point for the following gene set analyses within *CASSys*.

Identification of over-represented *Gene Ontology Terms*

Candidate genes are characterized using the *Gene Ontology Terms* (GO-terms) [4] they are annotated with. A central question is whether or not some GO-term annotations are over-represented in a set of candidate genes. *CASSys* addresses this question by applying a hypergeometric test to detect statistical over-representation and by reporting a Bonferroni-corrected p-value resulting from this test. For this functionality *CASSys* has an interface to *R/Bioconductor* [27, 28] and makes use of existing *Bioconductor* functions.

Pathway analyses

Pathway analyses provide evidence about the role candidate genes play in metabolic and signaling pathways. As a first step, a table of all *KEGG-Pathways* with at least a user-defined minimum of candidate genes is constructed. In a subsequent step, the web-service provided by *KEGG* is used to highlight those genes in the pathways. Additionally the genes are categorized according to the *KEGG-Brite* functional hierarchy.

Protein-protein interaction visualization

It is likely, that the candidate genes are directly regulated by the protein of interest. For studying other important regulatory mechanisms, like feedback-loops and interaction cascades, indirect interactions must also be taken into account. Therefore, the set of candidate genes is enriched with protein-protein interaction data from the *IntAct* database. This allows to construct network graphs containing the immunoprecipitated protein, its target genes and interactions with other proteins. *CASSys* enables users to visualize and interactively explore these networks, for an ex-

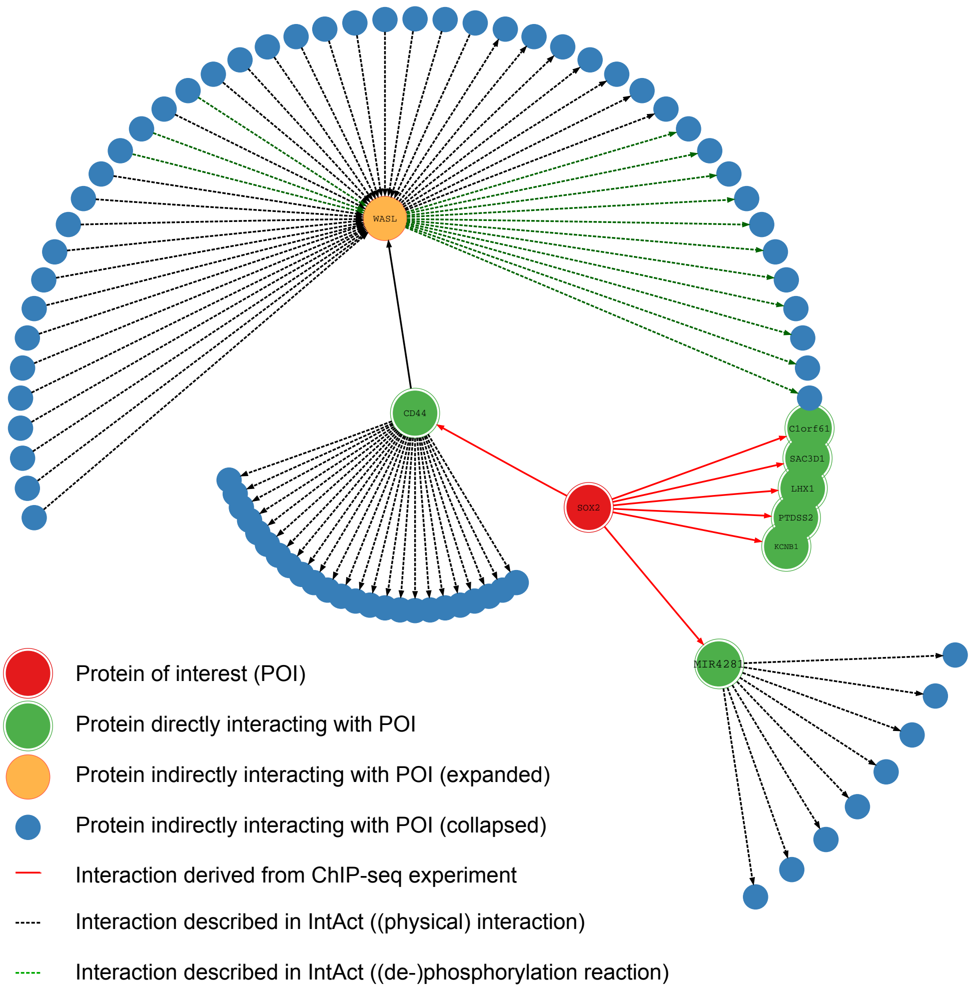

WASL
CD44
SOX2
C1orf61
SAC3D1
LHX1
PTDSS2
KCNB1
MIR4281
Protein of interest (POI)
Protein directly interacting with POI
Protein indirectly interacting with POI (expanded)
Protein indirectly interacting with POI (collapsed)
Interaction derived from ChIP-seq experiment
Interaction described in IntAct ((physical) interaction)
Interaction described in IntAct ((de-)phosphorylation reaction)

Figure 5: Interactive display of protein-protein interaction within *CASSys*. Subgraphs can be expanded or collapsed by the user by clicking on the corresponding nodes. The resulting network graph can be exported in scalable vector graphics format (SVG).

ample see Figure 5. Edges corresponding to different types of interaction are color-coded. The user can expand and collapse subgraphs by clicking on the nodes representing the interacting proteins.

Although *CASSys* was designed to serve the needs associated with ChIP-seq experiments, several of its features have a wide range of possible applications. Especially analyses based on gene sets may also prove useful in other contexts, like mRNA-expression analysis. Therefore, the web-interface of *CASSys* allows to directly submit an arbitrary set of gene identifiers. This set can be analyzed in the same way as a set of candidate genes from a ChIP-seq experiment.

3 Results

To demonstrate *CASSys*' analysis capability and functionality, we used the system to re-evaluate several ChIP-seq datasets from *GEO*. As an example, we briefly describe an analysis of estrogen receptor-α (ER-α) binding sites in breast cancer cells [14]. The analyzed dataset is publicly available in *GEO* (accession GSE19013). It consists of $3,624,955$ short reads resulting from Illumina Solexa sequencing of DNA immunoprecipitated against ER-α. *CASSys* accomplished all asynchronous analyses in ≈ 3 hours, requiring only few manual interactions. During QC/QA, 11% of the reads were removed, because their average Phred quality score was below a threshold of 10. From the remaining reads, 86% could be mapped to the reference genome (human genome, version 19) using *bowtie* (v0.12.7). 32% of the mapped reads were discarded because they could be mapped to more then ten genomic locations. Peakdetection was performed with *MACS* (v1.3.7.1). This delivered $12,131$ putative interaction sites for ER-α. All 643 sites inside a 3 kb region upstream of the transcription start site of a functional gene (as annotated in gencode v6) were selected. Of these sites, 620 sites were unambiguously associated with a gene. The interaction sites have a median length of 214 bases. The distances of the center of each interaction site to the next transcription start site have a median of 786 bases. As some genes are associated with multiple interaction sites, the 620 sites correspond to 570 genes. GO-term analysis revealed, that the terms *protein binding* and *multicellular organismal development* are significantly over-represented within the set of 570 genes. *MEME* (v4.4.0) and *Weeder* (v1.1) both found very similar highly significant sequence motifs within the sequences of the interaction sites. A search in the *JASPAR*-database for the highest scoring sequence motifs revealed that this motif clearly corresponds to a ER-α binding site motif [7], see Figure 6.

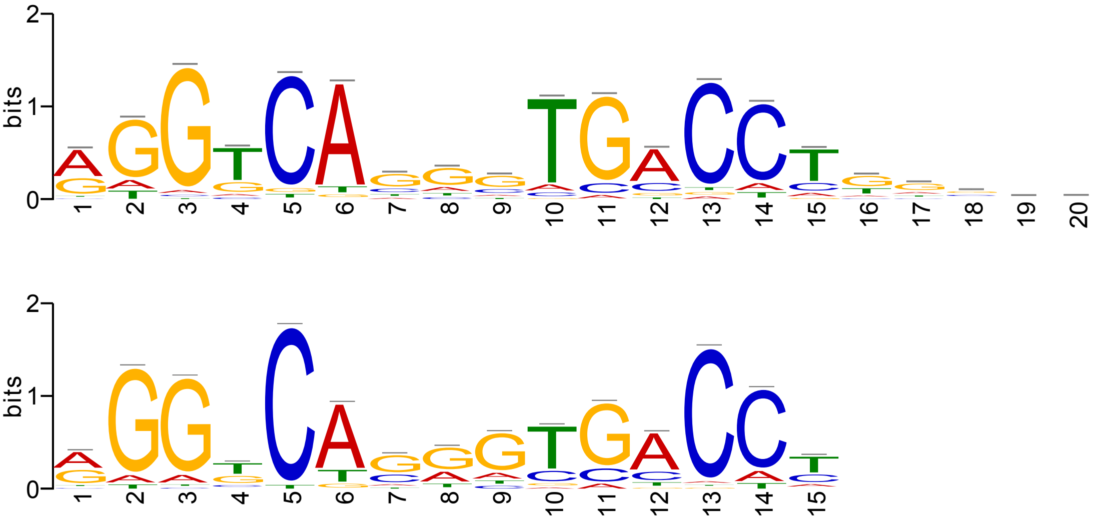


Figure 6: Top: ER binding site motif identified with *CASSys*. Bottom: The motif of ER-α from the JASPAR-database (accession number MA0112.2). Input for motif detection were the 200 most significant interaction sites extracted as described in the main text. All of these sites contained the shown motif.

The twelve genes with the statistically most significant binding site predictions are listed in Table 2. These include *TFF1* (trefoil factor 1), *GREB1* (growth regulation by estrogen in breast cancer 1), *PDZK1* (PDZ domain containing 1) and *OXT* (oxytocin, prepropeptide), all of which are known to be regulated by the estrogen receptor [10, 12, 29].

Gene	Sequence motif	Known interaction with ER	Overlapping ER-β site	associated with GO:0005515	associated with GO:0007275
TFF1	**AGGTCACCGTGGCCA**	√	√	√	
PDZK1	**AGGACTGGGTGACCT**	√	√	√	
ADAMTSL5	**AGGGCGGGGTGACCT**		√	√	
VASN	**GGGCCAGGGCAACCC**		√	√	
GREB1	**GGAGCTGTGTGACCT**	√	√		
ADORA1	**AGGTTAGGGTGACCT**		√	√	√
ANXA9	**GGACCACAGAGACCT**		√	√	
CCDC88C	**GGCCCAGGGCGACCT**			√	
MANEAL	**GGGTCAAACTGTCAA**		√		
OLFML3	**GGGTCACAGTGACCT**		√		√
OXT	**CGGTCAGGCTGACCT**	√	√	√	√
CASP7	**GGGTCAGGGTGAACT**		√	√	√

Table 2: The twelve genes (from a set of 570 candidate genes) with the statistically most significant binding site predictions. The interaction sites of all twelve genes contain the ER-α sequence motif. This is shown in column 2 with the less conserved part of the motif underlined. Four of the listed genes are known to be regulated by the estrogen receptor (column 3). The locations of the interaction sites corresponding to the twelve genes were compared with those derived from the ER-β dataset GSE21770 [32] (workflow as described in the main text). The interaction site upstream of CCDC88C is the only site not overlapping in both datasets (column 4). GO-term analysis reveals that the terms *protein binding* and *multicellular organismal development* are both significantly over-represented in the set of 570 candidate genes. Nine of the twelve listed genes are associated with the GO-term *protein binding* (GO:0005515, column 5) and four with the GO-term *multicellular organismal development* (GO:0007275, column 6).

4 Discussion

The analysis of data from ChIP-seq experiments is a complex process involving the use of several different software tools and information resources. Therefore, in practice such analysis tasks often turn into a patchwork of manual application and script-based glueing of these tools. The output of one program often has to be reformatted before it can serve as an input for another program. Consequently experimentalists without bioinformatic expertise are often unable to efficiently conduct ChIP-seq data analyses and even for bioinformaticians the process is still laborious. For these reasons we developed *CASSys*, an integrated, user-friendly software system, spanning all steps of ChIP-seq data analysis. The software is easy to use, and offers an unprecedented range of functionality, allowing for example, extensive parametrization of each analysis step.

For the results of ChIP-seq data analysis and hence the biological conclusions drawn from the data, readmapping and peakdetection are notably important, because all follow-up analyses rely on the results obtained in these processing steps. Since excellent programs exist for readmapping, peakdetection and motif detection, *CASSys* employs third party software for these steps. Moreover, it implements a general concept for integrating new tools. This guarantees, that the success of *CASSys* is not bound to a single softwaretool and that, at any time, the best available tools can be used.

5 Availability

The web based part of the *CASSys* system for carrying out all interactive analyses is available at `http://www.zbh.uni-hamburg.de/cassys`.

References

[1] ChIP-Seq web server. *http://ccg.vital-it.ch/chipseq/*.

[2] Generic Feature Format version 3. *http://www.sequenceontology.org/resources/gff3.html*.

[3] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, 38(Database issue):D525–D531, Jan 2010.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, May 2000.

[5] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, 37(Web Server issue):W202–W208, Jul 2009.

[6] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Sobolevashort. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res*, 39(Database issue):D1005–D1010, Jan 2011.

[7] J. M. Berg, J. L. Tymoczko, and L. Stryer. *Stryer Biochemie*. Spektrum Akademischer Verlag, Heidelberg, 2007.

[8] K. R. Blahnik, L. Dou, H. O'Geen, T. McPhillips, X. Xu, A. R. Cao, S. Iyengar, C. M. Nicolet, B. Ludäscher, I. Korf, and P. J. Farnham. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res*, 38(3):e13, Jan 2010.

[9] V. Boeva, D. Surdez, N. Guillon, F. Tirode, A. P. Fejes, O. Delattre, and E. Barillot. De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res*, 38(11):e126, Jun 2010.

[10] J. S. Carroll, X. S. Liu, A. S. Brodsky, W. Li, C. A. Meyer, A. J. Szary, J. Eeckhoute, W. Shao, E. V. Hestermann, T. R. Geistlinger, E. A. Fox, P. A. Silver, and M. Brown. Chromosome-Wide Mapping of Estrogen Receptor Binding Reveals Long-Range Regulation Requiring the Forkhead Protein FoxA1. *Cell*, 122(1):33–43, Jul 2005.

[11] A. P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. J. M. Jones. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, Aug 2008.

[12] M. G. Ghosh, D. A. Thompson, and R. J. Weigel. PDZK1 and GREB1 are estrogen-regulated genes expressed in hormone-responsive breast cancer. *Cancer Res*, 60(22):6367–6375, Nov 2000.

[13] S. Gupta, J. Stamatoyannopoulos, T. Bailey, and W. Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, 2007.

[14] M. Hu, J. Yu, J. M. G. Taylor, A. M. Chinnaiyan, and Z. S. Qin. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res*, 38(7):2154–2167, Apr 2010.

[15] H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*, 26(11):1293–1300, Nov 2008.

[16] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 38(Database issue):D355–D360, Jan 2010.

[17] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res*, 12(6):996–1006, Jun 2002.

[18] T. D. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, and L. L. Elo. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, (10), Dec 2009.

[19] X. Lan, R. Bonneville, J. Apostolos, W. Wu, and V. X. Jin. W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. *Bioinformatics*, 27(3):428–430, Feb 2011.

[20] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[21] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, Oct 1993.

[22] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[23] L. Li. GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J Comput Biol*, 16(2):317–329, Feb 2009.

[24] N. Palmieri and C. Schlötterer. Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. *PLoS One*, 4(7):e6323, 2009.

[25] G. Pavesi, P. Mereghetti, F. Zambelli, M. Stefani, G. Mauri, and G. Pesole. MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res*, 34(Web Server issue):W566–W570, Jul 2006.

[26] E. Portales-Casamar, S. Thongjuea, A. T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W. W. Wasserman, and A. Sandelin. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 38(Database issue):D105–D110, Jan 2010.

[27] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2011.

[28] M. Reimers and V. J. Carey. Bioconductor: an open source framework for bioinformatics and computational biology. *Methods Enzymol*, 411:119–134, 2006.

[29] P. J. Shughrue, T. L. Dellovade, and I. Merchenthaler. Estrogen modulates oxytocin gene expression in regions of the rat supraoptic and paraventricular nuclei that contain estrogen receptor-beta. *Prog Brain Res*, 139:15–29, 2002.

[30] S. Steinbiss, G. Gremme, C. Schärfer, M. Mader, and S. Kurtz. AnnotationSketch: a genome annotation drawing library. *Bioinformatics*, 25(4):533–534, Feb 2009.

[31] A. M. Szalkowski and C. D. Schmid. Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmark efforts. *Briefings in Bioinformatics*, pages 1–8, Nov 2010.

[32] O. I. Vivar, X. Zhao, E. F. Saunier, C. Griffin, O. S. Mayba, M. Tagliaferri, I. Cohen, T. P. Speed, and D. C. Leitman. Estrogen receptor beta binds to and regulates three distinct classes of target genes. *J Biol Chem*, 285(29):22059–22066, Jul 2010.

[33] E. G. Wilbanks and M. T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, 5(7):e11471, 2010.

[34] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9):R137, 2008.