SOFTWARE TOOL ARTICLE

Expresso: A database and web server for exploring the interaction of transcription factors and their target genes in *Arabidopsis thaliana* using ChIP-Seq peak data [version 1; referees: 2 approved, 1 approved with reservations]

Delasa Aghamirzaie[1], Karthik Raja Velmurugan[1,2], Shuchi Wu[3], Doaa Altarawy[4], Lenwood S. Heath[4], Ruth Grene[5]

[1]Genetics, Bioinformatics, and Computational Biology (GBCB), Virginia Tech, Blacksburg, VA, 24061, USA
[2]Center for Bioinformatics and Genetics and the Primary Care Research Network, Edward Via College of Osteopathic Medicine, Blacksburg, VA, 24060, USA
[3]Department of Horticulture, Virginia Tech, Blacksburg, VA, 24061, USA
[4]Department of Computer Science, Virginia Tech, Blacksburg, VA, 24061, USA
[5]Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech, Blacksburg, VA, 24061, USA

v1 **First published:** 28 Mar 2017, **6**:372 (doi: 10.12688/f1000research.10041.1)
Latest published: 28 Mar 2017, **6**:372 (doi: 10.12688/f1000research.10041.1)

Abstract

Motivation: The increasing availability of chromatin immunoprecipitation sequencing (ChIP-Seq) data enables us to learn more about the action of transcription factors in the regulation of gene expression. Even though *in vivo* transcriptional regulation often involves the concerted action of more than one transcription factor, the format of each individual ChIP-Seq dataset usually represents the action of a single transcription factor. Therefore, a relational database in which available ChIP-Seq datasets are curated is essential.
Results: We present Expresso (database and webserver) as a tool for the collection and integration of available *Arabidopsis* ChIP-Seq peak data, which in turn can be linked to a user's gene expression data. Known target genes of transcription factors were identified by motif analysis of publicly available GEO ChIP-Seq data sets. Expresso currently provides three services: 1) Identification of target genes of a given transcription factor; 2) Identification of transcription factors that regulate a gene of interest; 3) Computation of correlation between the gene expression of transcription factors and their target genes.
Availability: Expresso is freely available at http://bioinformatics.cs.vt.edu/expresso/

This article is included in the GODAN channel.

Open Peer Review

Referee Status: ? ✓ ✓

	Invited Referees		
	1	2	3
version 1 published 28 Mar 2017	? report	✓ report	✓ report

1 **Nicholas J. Provart**, University of Toronto Canada, University of Toronto Canada

2 **Asa Ben-Hur** , Colorado State University USA

3 **Sakiko Okumoto**, Texas A&M University USA

Discuss this article

Comments (0)

Corresponding author: Delasa Aghamirzaie (delasa@vt.edu)

How to cite this article: Aghamirzaie D, Raja Velmurugan K, Wu S *et al.* **Expresso: A database and web server for exploring the interaction of transcription factors and their target genes in *Arabidopsis thaliana* using ChIP-Seq peak data [version 1; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2017, **6**:372 (doi: 10.12688/f1000research.10041.1)

Grant information: This project was supported by National Science Foundation [NSF-MCB-1052145 and NSF-ABI-1062472].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 28 Mar 2017, **6**:372 (doi: 10.12688/f1000research.10041.1)

Introduction

Chromatin immunoprecipitation (ChIP) is a method to investigate DNA-binding sites of DNA-binding proteins, such as transcription factors (TFs) (Valouev *et al.*, 2008). ChIP can provide genome-wide information of *in vivo* protein-DNA interactions (Kaufmann *et al.*, 2010). Therefore, it has become an important tool to assay TF-associated gene regulations (Kaufmann *et al.*, 2010; Park, 2009; Valouev *et al.*, 2008). In a typical ChIP experiment, first the DNA-binding protein of interest is cross-linked to its binding sites. Then the chromatin is sheared, randomly, into short fragments and the protein-DNA complexes are purified by immunoprecipitation using a specific antibody against the DNA-binding protein of interest. Finally, genome-wide profiling of protein binding sites is produced by either genome-tiling arrays (ChIP-ChIP) or next-generation sequencing technologies (ChIP-Seq) (Kaufmann *et al.*, 2010; Valouev *et al.*, 2008). Compared to ChIP-ChIP, ChIP-Seq provides high-resolution data with a better signal-noise ratio. ChIP-seq also requires less initial material and is more cost-effective (Ho *et al.*, 2011; Kaufmann *et al.*, 2010; Valouev *et al.*, 2008). Therefore, ChIP-Seq has displaced ChIP-ChIP rapidly and is currently the most widely used technology for studying the action of transcription factors (Park, 2009; Valouev *et al.*, 2008).

In contrast to the biomedical field, the use of ChIP-Seq in plant biology is limited (Kaufmann *et al.*, 2010). For example, the GEO database (https://www.ncbi.nlm.nih.gov/gds) currently contains 8,486 ChIP-Seq human datasets (as of October 2016), but has only 200 *Arabidopsis* datasets. The delay in the use of ChIP-Seq technology in plant research may be due to the specific properties of plant tissue, such as the presence of the cell wall and abundant secondary metabolites that affect the quality of protein-DNA complex extraction (Kaufmann *et al.*, 2010). However, with the improvement of ChIP-Seq protocols and reduction of next-generation sequencing costs, an increasing number of plant scientists are choosing ChIP-Seq to study function of transcription factors in detail.

ChIP datasets currently available for *Arabidopsis* are isolated, fragmentary and they lack a uniform format. Thus a major gap exists between the capabilities of *in vitro* methods, such as ChIP Seq and the goal of understanding the complexities of transcriptional regulation. We report on the curation of the Expresso database to collect and integrate *Arabidopsis* ChIP-Seq data (available as peaks), which in turn can be linked to a user-provided *Arabidopsis* gene expression data. Expresso compiles 20 groups of selected *Arabidopsis* ChIP-Seq peak datasets downloaded from NCBI GEO or supplemental data of the corresponding paper. All collected ChIP-Seq peak datasets were re-analyzed by the Expresso processing pipeline to create a coherent and unified results which bridge the gap among multiple ChIP-Seq studies, and to provide a consensus access to TFs, target genes and DNA-binding motifs. In summary, instead of going though separate ChIP-Seq datasets, Expresso provides a more rapid and integrated method for the systematic study of the action of plant transcription factors.

Methods

The Expresso computational analysis pipeline comprises preprocessing of peak loci reported by at each reference dataset, finding conserved motifs using MEME-suite (Bailey *et al.*, 2009), identifying potential target genes for each transcription factor, and finally storing target genes and motifs linked to TFs into the database. Data-formatting primarily involves the extraction of a peak locus peak, peak summit and DNA sequences in fasta format from the *Arabidopsis thaliana* genome. Of the 50 datasets, almost all were found to be in distinct formats and only 20 had the peak information available either on GEO or at their supplemental material section of their corresponding published manuscript. We restructured the downloaded data into a unique format by extracting a specific set of information including: peak ID, chromosome number, peak start and end positions and genes in 1kbp distance of the peak summit. All the codes for preprocessing of the input data are available at Expresso GitHub page under "preprocessing".

Candidate target gene finding using motif search: Given the chromosome number and peak start and end positions, the corresponding genomic sequence was extracted and trimmed, and then were subject to motif search using MEME-suite tool (http://meme-suite.org/), with following parameters: -nmotifs 20 -minw 5 -maxw 30 -dna. While the distribution of the length of the untrimmed peak sequences of each dataset varied widely, the reported peak summit lengths were usually 200 to 500 bases long upstream and downstream from the middle of the summit (Bailey *et al.*, 2009; Immink *et al.*, 2012; Valouev *et al.*, 2008). For a few datasets, the summit length was not provided in the article, so the largest summit length found, 500 bases, was used. Motif width was set to the length of the reported motif (if any). Otherwise, motif width was set to 5 to 30 bps, and significant motifs (E-value < 0.05) together with the candidate target genes possessing those motifs were uploaded to the database. Hence, a gene should have the following properties to be eligible to be uploaded to the database: i) should be among the target genes provided by a ChIP-Seq experiment, or within 1kbps distance of the peak summit ii) should have a significantly enriched motif in its peak binding site. Moreover, the presence of the motif found by MEME was validated by the reported motif in the reference paper. If the reported motif was not found using the MEME search tool on the peak sequences, the resulting motifs were not uploaded to the database.

Results

Expresso provides a user-friendly environment to facilitate exploring different transcription factors and target genes through motif analysis. ChIP-Seq experiments in Expresso are available under the "Experiments" tab. Expresso currently provides three services for identifying: 1) the target genes of a given transcription factor, 2) the transcription factors that regulate genes of interest and 3) the correlation of gene expression between transcription factors and their target genes.

Identifying candidate target genes for a transcription factor (see "Transcription Factors" on the Expresso website: http://bioinformatics.cs.vt.edu/expresso/?q=node/3): Users can select a transcription factor from the list of available transcription factors to view potential target genes. Since target genes for each transcription factor have been compiled from the peaks and motifs data, users can change the cut-off for the motif E-value. The default E-value is set to 0.05. A short functional description (along with a link to TAIR10) and the GEO id for the reference ChIP-Seq experiment is provided for each potential target gene. For example, searching for target genes of TOC1 transcription factor results in 298 genes that have at least one significantly enriched motifs at least one peak located close to their transcription start site.

Identifying potential transcription factors regulating a target gene (see "Genes" on the Expresso website: http://bioinformatics.cs.vt.edu/expresso/?q=node/4): Users can enter a gene or multiple genes and Expresso finds all the transcription factors that might regulate that gene together with the binding motif for that TF. For example, SGP2 (AT3G21700) gene is potentially transcriptionally regulated by PIF3 and KAN1.

Exploring gene expression data: Users can upload gene expression data and Expresso finds genes and transcription factor pairs present in Expresso database and performs Pearson correlation analysis on their corresponding expression data. Upon submission of the gene expression, a task id is assigned to this job. Users need to keep the task id to retrieve the results or check the status of their job. If they provide an email address, they will be notified when the results get ready. To demonstrate the application of correlation analysis on finding potential TF-target gene pairs, a RNA-Seq dataset (Segaran, 2007) has been added to Expresso as a demo (see "Gene Expression" on the Expresso website: http://bioinformatics.cs.vt.edu/expresso/?q=node/5). 100 genes (including some transcription factors) were selected randomly from this dataset, which has expression values for genes from different *Arabidopsis* tissues: leafs, seeds, roots and flowers. 54 genes were found to be target genes of transcription factors in Expresso. 33% of the uploaded genes were found to be targets genes of multiple transcription factors. The correlation of gene expression between a transcription factor and its target genes can be used for inferring their relationship. For example, three out of four target genes of PIF3 show high correlation with the expression of PIF3, although one gene was found to have a negative correlation (R=-0.92). The fact that their expression patterns are correlated with PIF3, suggests that PIF3 plays a dominant role in regulating these three target genes. However, AT3G21700 was found to have a low correlation with PIF3, which suggests that there might be other transcription factors that challenge PIF3 in the regulation of AT3G21700.

Conclusions

ChIP-Seq is a powerful technology that aides in the study of the action of transcription factors, predicting a given transcription factor's target genes and corresponding conserved binding motifs (Ho *et al.*, 2011; Kaufmann *et al.*, 2010; Park, 2009; Valouev *et al.*, 2008). The Expresso database is curated to integrate several available ChIP-Seq datasets. Expresso provides an easy access to 1) potential targets of a given transcription factor and their possible binding sites; 2) candidate transcription factors regulating several genes of interest; 3) correlation analysis of TF and target gene pair using the user's input gene expression data. Taken together, Expresso facilitates an easy access to several ChIP-Seq experiments, making the study of the transcriptional regulation in the cells easier in the context of interaction among several transcription factors.

Software and data availability

Expresso is freely available online: http://bioinformatics.cs.vt.edu/expresso/

Source code available at: https://github.com/doaa-altarawy/Expresso/tree/2.0.0

Archived source code as at time of publication: doi, 10.5281/zenodo.399501 (Altarawy, 2017).

License: MIT

All datasets were publicly available and were downloaded from GEO DataSets. The list of ChIP-Seq datasets available in Expresso is available at 'Experiments' section on Expresso. The list of transcription factors and target genes can be downloaded in the text format.

Author contributions

D. Aghamirzaie contributed to the data analysis, web development, and biological validation of the results. KRV contributed to the data analysis section. SV was involved in the biological validation of the results. D. Altarawy was involved the web development and maintenance of the Expresso website. RG and LH conceived the study. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

This project was supported by National Science Foundation [NSF-MCB-1052145 and NSF-ABI-1062472].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors are grateful to the GBCB (Genetics, Bioinformatics, and Computational Biology) program, especially Dr. David Bevan, for providing the opportunity to work on this project.

References

Altarawy D: **doaa-altarawy/Expresso: Expresso Ver 2.0 [Data set].** *Zenodo.* 2017.
Data Source

Bailey TL, Boden M, Buske FA, *et al.*: **MEME SUITE: tools for motif discovery and searching.** *Nucleic Acids Res.* 2009; **37**(Web Server issue): W202–W208.
PubMed Abstract | Publisher Full Text | Free Full Text

Ho JW, Bishop E, Karchenko PV, *et al.*: **ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis.** *BMC Genomics.* 2011; **12**(1): 134.
PubMed Abstract | Publisher Full Text | Free Full Text

Immink RG, Posé D, Ferrario S, *et al.*: **Characterization of SOC1's central role in flowering by the identification of its upstream and downstream regulators.** *Plant Physiol.* 2012; **160**(1): 433–449.
PubMed Abstract | Publisher Full Text | Free Full Text

Kaufmann K, Muiño JM, Østerås M, *et al.*: **Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP).** *Nat Protoc.* 2010; **5**(3): 457–472.
PubMed Abstract | Publisher Full Text

Park PJ: **ChIP-seq: advantages and challenges of a maturing technology.** *Nat Rev Genet.* 2009; **10**(10): 669–680.
PubMed Abstract | Publisher Full Text | Free Full Text

Segaran T: **Programming collective intelligence: building smart web 2.0 applications.** *O'Reilly Media, Inc.* 2007.
Reference Source

Valouev A, Johnson DS, Sundquist A, *et al.*: **Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.** *Nat Methods.* 2008; **5**(9): 829–834.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ? ✓ ✓

Version 1

Referee Report 02 May 2017

doi:10.5256/f1000research.10818.r22402

✓ **Sakiko Okumoto**
Department of Soil and Crop Sciences, Texas A&M University, College Station, TX, USA

In this manuscript, the authors created a web-based interface which hosts ChIP-seq data from 20 published experiments, and allows the users to 1) access the compiled lists of TF targets that met the criteria set by the authors, 2) identify the transcription factor(s) that regulate his/her gene of interest, and 3) perform co-expression analyses with user-provided RNAseq data.

Although there are other web-based services that allow at least part of what is described above, the authors argue that this is the first platform that provides a uniform format for 20 genes. I would like to agree with the authors about the value of such a format.

Of the above three functions, the first two are fairly straightforward and seems to function as expected. However, I feel that the description found in the manuscript about the co-expression analysis did not contain enough information.

The authors provide a list of 100 genes with their FPKM values in 4 different tissues as a demo. The manuscript describes the set as “100 genes (including some transcription factors) were selected randomly from this dataset”. When comparing this list of 100 with the list of TFs in this database however, I see that 16 out of 20 TFs in the database are included in the list of 100 genes. This seems more than “some” to me- please describe specifically. When I run the demo, none of the 4 that are not in the list of 100 are found to be co-expressed with any of the genes. I think that would make sense because I don’t see how one can deduce co-expression between a given gene and a TF it the TF is not expressed in the data set. If this is indeed a requirement, I would like to see that stated in the manuscript.

Also, typically how many tissues/times would need to be in the dataset? (When I remove one of the columns in the demo set I don’t get any hits, probably due to less statistical power of the data set.) It would be beneficial for the readers to know the approximate number experiments needed for a correlation analysis.

In general, I would really appreciate if the authors could explain how co-expression analysis works – does it first perform co-expression analyses within the genes in the uploaded dataset, identify the TFs in the data base, then select the ones that have the consensus motif? A lay-friendly flow chart would be much appreciated.

Also, it would be nice if the algorithm identified the motif and the distance from the ATG in the co-expressed genes.

Minor points include:

- On the “Experiment” tab- for each TF, would you please include the link to the original publication? One can trace back using GEO NCBI, but it would be easier for the users if the publication is included as an additional column.
- To demonstrate the application of correlation analysis on finding potential TF-target gene pairs, a RNA-Seq dataset (Segaran, 2007) has been added to Expresso as a demo” I am fairly certain that the reference provided here is wrong.
- Methods “a peak locus peak” a peak locus?

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 18 April 2017

doi:10.5256/f1000research.10818.r21320

✓ **Asa Ben-Hur**
Department of Computer Science, Colorado State University, Fort Collins, CO, USA

The authors have created a useful resource that provides unified access to a large number of ChIP-seq experiments in arabidopsis. The database has useful functionality that would be useful for exploring TF binding. Each function of the database has example data that allows users to try it out easily, and the pipeline is available through github.

The following should be addressed:

1. The major issue with the paper is that there is an existing similar resource called ChIPBase (see citations below[1,2]). The authors should cite it and compare their database with it, as it's not obvious what Expresso is adding to what it provides.

2. A figure that summarizes your data analysis pipeline would be beneficial (I saw such a figure on the Expresso website).

3. In the section on motif finding: did you focus on promoter regions for the peaks, and if so, how were those defined? The motifs generated by MEME were compared against those in the corresponding papers, and no motif was added if MEME did not detect a motif. How often did that happen? MEME occasionally misses motifs, and other tools could have possibly found those motifs.

4. In the expression section, please suggest how the user should measure expression to provide good results. For that matter, please provide information on how expression of the TFs is quantified.

Minor comments:

There are some grammar issues that need to be fixed - see below.

1. In the introduction you write that "ChIP can provide genome-wide information...". That is true when performed as ChIP-ChIP or ChIP-seq.

2. TF-associated gene regulations --> regulation

3. "ChIP-Seq in plant biology is limited": I think you meant that it hasn't been as widely used as in mammalian systems.

4. I did not buy your explanation of the delay in adoption of ChIP-seq in plant research. Plant research tends to be a few steps behind, and furthermore, many more people study human than arabidopsis.

5. "All the codes for preprocessing" --> all the code for preprocessing

6. "results in 298 genes that have at least one significantly enriched motifs at least one peak located close to their transcription start site." something unclear here - "one significantly enriched motifs at least one peak" - should there be an "or" or "and" enriched motif AND at least one peak? And the word motif should be singular, and refer to a motif hit/occurrence.

7. targets genes --> target genes

8. "can be downloaded in the text format" --> in text format

References

1. Zhou KR, Liu S, Sun WJ, Zheng LL, Zhou H, Yang JH, Qu LH: ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data.*Nucleic Acids Res*. 2017; **45** (D1): D43-D50 PubMed Abstract | Publisher Full Text

2. Yang JH, Li JH, Jiang S, Zhou H, Qu LH: ChIPBase: a database for decoding the transcriptional

regulation of long non-coding RNA and microRNA genes from ChIP-Seq data.*Nucleic Acids Res*. 2013; **41** (Database issue): D177-87 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Bioinformatics, machine learning, analysis of high throughput sequencing data

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 18 April 2017

doi:10.5256/f1000research.10818.r21317

? **Nicholas J. Provart**[1,2]

[1] Department of Cell and Systems Biology, University of Toronto, Toronto, ON., Canada

[2] Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON., Canada

In principle the Expresso database will be useful to plant researchers. I would like to see a couple of things: what about mention of other databases like AGRIS at OSU and Cistome/ePlant at the BAR? Do these capture the kinds of interactions the authors are describing? Another is HRGRN (http://plantgrn.noble.org/hrgrn/). What are the disadvantages/limitations of these vis a vis Expresso?

Another thing that would a “nice-to-have” would be to include the Ecker Lab’s recent, extensive DAP-seq data set, which the authors (https://www.ncbi.nlm.nih.gov/pubmed/27203113) show to be quite concordant with existing Chip-Seq data. These data are more extensive than the fairly limited number of Chip-seq data sets that Aghamirzaie et al. have collated.

I tried out the software, which worked as promised. The functionality was somewhat basic. It would be quite easy to use table.js or similar on the “Genes” search output page, or on the “Gene Expression” output page to be able to sort the table of favourite genes with their targets as a user expects to be able to do, or to sort by Pearson correlation. It might be nice to let users know how to download the "Genes" search results and load them into Cytoscape in a tutorial section. I was unable to download a file of TFs binding to my favourite genes (URL http://bioinformatics.cs.vt.edu/expresso/Expresso_Codes/getResFile_Genes.php) – the page returned an error of "Unable to select database".

"Run Demo" did not work on the "Gene Expression" page, or at least I thought it didn't until I realized I had to scroll down to see the results, which appeared...but off the bottom of my screen…a little Javascript autoscroll to that section would be helpful after the calculation has finished.

Typos/grammar
In general: it's ChIP-chip, not ChIP-ChIP (the first ChIP is for Chromatin Immuno-Precipitation, the second "chip" refers to microarray)

Be consistent: either ChIP-seq or ChIP-Seq (we see ChIP-seq, ChIP-Seq, and ChIP Seq in the paper)

Candidate target gene finding section: "the corresponding genomic sequences were extracted and trimmed" ("sequences" should be plural as multiple genomic sequences are analyzed, no?)

Candidate target gene finding section: "Otherwise, motif width was set to be between 5 to 30 bp" ("…to 5 to 30" is awkward)

Bottom of page 3: ("along with a link to TAIR10") – TAIR10 refers to the 10th genome build. I'd say rather "along with a link to TAIR". It might be nice to add a link to the Araport record for a given gene too.

Top of page 4: "…motifs in at least one peak located close to…" (missing "in"?)

Midway down page 4: "…the results are complete." (instead of "get ready")

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Cyberinfrastructure, plant bioinformatics, data visualization

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.