Current Plant Biology 3-4 (2015) 48–55

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

CressInt: A user-friendly web resource for genome-scale exploration of gene regulation in *Arabidopsis thaliana*

Xiaoting Chen[a,b,1], Kevin Ernst[a,b,1], Frances Soman[a], Mike Borowczak[a,b,2], Matthew T. Weirauch[b,c,*]

[a] *Department of Electrical Engineering and Computing Systems, College of Engineering and Applied Sciences, University of Cincinnati, Cincinnati, OH 45221, United States*
[b] *Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, OH 45229, United States*
[c] *Division of Biomedical Informatics and Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, OH 45229, United States*

ARTICLE INFO

Article history:
Received 16 April 2015
Received in revised form 25 August 2015
Accepted 9 September 2015

Keywords:
Arabidopsis
Functional genomics
Transcription factors
Gene regulation
Systems biology
Web server
Computational tools

ABSTRACT

The thale cress *Arabidopsis thaliana* is a powerful model organism for studying a wide variety of biological processes. Recent advances in sequencing technology have resulted in a wealth of information describing numerous aspects of *A. thaliana* genome function. However, there is a relative paucity of computational systems for efficiently and effectively using these data to create testable hypotheses. We present *CressInt*, a user-friendly web resource for exploring gene regulatory mechanisms in *A. thaliana* on a genomic scale. The *CressInt* system incorporates a variety of genome-wide data types relevant to gene regulation, including transcription factor (TF) binding site models, ChIP-seq, DNase-seq, eQTLs, and GWAS. We demonstrate the utility of *CressInt* by showing how the system can be used to (1) Identify TFs binding to the promoter of a gene of interest; (2) identify genetic variants that are likely to impact TF binding based on a ChIP-seq dataset; and (3) identify specific TFs whose binding might be impacted by phenotype-associated variants. *CressInt* is freely available at http://cressint.cchmc.org.

1. Introduction

The sequencing of the *Arabidopsis thaliana* genome over 15 years ago [1] enabled a new era of scientific exploration of this versatile model organism. As "next generation" sequencing technologies continue to mature, datasets capable of measuring function on a genome-wide scale continue to become more prevalent. Despite an exponential increase in our ability to generate data probing function on a genome-scale, there remains a lag in our analytical capability to effectively analyze these data to attain new biological insights.

Several useful bioinformatics tools are currently in widespread use in the *Arabidopsis* community (see de Lucas et al. [2], Bassel et al. [3], and Brady and Provart [4] for reviews). However, as more complex and higher resolution data types become available, there is an increasing need for the development of user-friendly computational tools for their analysis. In the past five years alone, *Arabidopsis* data have been released describing genetic variants associated with particular traits [5] or altered gene expression levels [6], open chromatin regions in multiple tissue types and conditions [7], and DNA binding specificities for hundreds of transcription factors (TFs) [8]. Collectively, these data offer new opportunities to probe gene regulation and genome function. However, access to the wide range of analytical capabilities afforded by these data remains largely limited to bioinformaticians.

We present CressInt (thale cress data intersector), a user-friendly, freely accessible web server for integrating and analyzing genome-scale *A. thaliana gene* datasets. Conceptually, CressInt is similar to visually analyzing data in a genome browser such as those

Abbreviations: TF, transcription factor; ChIP-seq, Chromatin immunoprecipitation followed by sequencing; DNase-seq, sequencing of DNase I hypersensitive sites; eQTL, expression quantitative trait locus; GWAS, genome-wide association study; PBM, protein binding microarray.

* Corresponding author at: Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, OH 45229, United States.
E-mail address: Matthew.Weirauch@cchmc.org (M.T. Weirauch).

[1] These authors contributed equally.
[2] Current affiliation: Erebus Labs, Laramie, WY 82073, United States.

http://dx.doi.org/10.1016/j.cpb.2015.09.001

provided by UC Santa Cruz [9] or Ensembl [10], with the key differences that (1) up to thousands of loci of interest can be queried at once; (2) quality-controlled data specific to *A. thaliana* are preloaded into the CressInt system; and (3) results are downloadable in formats easily amenable to further downstream analysis. CressInt combines data from a wide variety of sources, including TF genomic binding regions (from ChIP-seq), TF DNA binding specificities (from Protein Binding Microarrays (PBMs) [11]), chromatin accessibility (DNAse-seq), and genetic variants associated with specific phenotypes (from GWAS) or genotype-dependent gene expression levels (*i.e.*, expression quantitative trait loci or eQTLs). The CressInt system enables a wide range of queries, from simple (*e.g.*, identifying all datasets intersecting genomic regions of interest), to complex (*e.g.*, identifying genetic variants of interest likely to affect the binding of specific TFs). To our knowledge, there is currently no web server capable of performing these operations on *A. thaliana* datasets that are already integrated into the system. This not only enables easy access to these data for non-computational experts, but also saves hours of time that would otherwise be spent identifying, obtaining, quality checking, and re-formatting the various data sets.

CressInt's intuitive graphical user interface is designed to be easy to use for non-bioinformaticians, while maintaining sufficient power and capabilities to enable downstream computational analysis. Using three case studies, we demonstrate the ability of CressInt to effectively use functional genomics data to generate testable hypotheses involving genes or phenotypes of interest. The CressInt web server is freely available at https://cressint.cchmc.org.

2. Materials and methods

2.1. Data and code availability

All source code developed for the web server is available on Bitbucket (https://bitbucket.org/weirauchlab/tf-tools-cressint). All datasets are available from the authors upon request.

2.2. Data collection and quality control

We obtained data from a variety of sources (Table 1). All genome-based datasets are organized by plant tissue type (*e.g.*, seedling, leaf, inflorescence, *etc.*), and stored as UC Santa Cruz BED6 files [9]. DNase-seq data indicating open chromatin regions in *A. thaliana* seedlings exposed to heatshock, darkness, and light were taken from Sullivan et al. [7]. 4,355,790 naturally occurring genetic variants and eQTLs derived from seedlings were obtained from Gan et al. [6]. The eQTL set was filtered to only include SNPs with *P*-values <0.001. GWAS data were obtained from Atwell et al. [5], and genetic variants were included in our set of phenotype-associated variants if they either (1) have associations exceeding genome-wide significance ($P < 2.75 \times 10^{-7}$, which corresponds to the Bonferroni-corrected $P < 0.05$ cutoff used in the original study; 178 SNPs in total) or (2) are among the top 10 most strongly associated variants for each phenotype, regardless of *P*-value (943 SNPs in total). TF binding specificity models were taken from build 1.01 of the CisBP database [8] (http://cisbp.ccbr.utoronto.ca/).

We obtained ChIP-seq data from the gene expression omnibus (GEO) [12]. Beginning with all 26 *A. thaliana* ChIP-seq datasets available in GEO in March of 2015, we used a three-step quality control procedure to ensure that only high-quality datasets are included in the CressInt system. First, we removed any datasets whose peak regions cover >5% of the *A. thaliana* genome, deeming them too non-specific (with the exception of ChIP-seq for histone marks, which mark general regulatory regions and tend to have wider peaks). Next, we removed any datasets where the number of peaks obtained from the GEO dataset did not match the number of peaks reported in the publication associated with the data—this step is necessary because both GEO datasets and methods sections of manuscripts are often insufficiently documented to reproduce the reported peak calls. Finally, we ran all peak sets through the TF DNA binding motif enrichment algorithm used by HOMER [13], and only included datasets where the ChIP'ed TF's motif ranks in the top three of enriched motifs. A total of 16ChIP-seq datasets, taken from 13 different studies, passed our QC process (Table 1).

2.3. Differential binding of transcription factors to genetic variants

We used PBM data describing the DNA binding specificities of 575 *A. thaliana* TFs taken from Weirauch et al. [8], and a similar procedure used in that study and another recent study [14] to identify TFs whose binding might be affected by the alleles of 4,355,790 naturally occurring *A. thaliana* genetic variants [6]. One type of data produced by a PBM experiment is the E-score, which ranges from −0.50 to +0.50, and quantifies the relative preference of the binding of the tested TF to each of the 32,896 possible 8 base sequences [11]. We constructed a matrix containing the PBM 8-mer E-scores for 534 PBM experiments (267 constructs, each assayed on two independent array designs). 466 of these experiments directly assay the DNA binding specificity an *A. thaliana* TF. 68 of them measure a related TF in another organism that has a similar DNA binding domain (DBD) to at least one *A. thaliana* TF (68 experiments). Each PBM experiment was mapped to its "closest" *A. thaliana* TF by either (1) assigning it to the *A. thaliana* TF that was directly measured (trivial); or (2) (for PBMs measuring non-*A. thaliana* TFs) assigning it to the *A. thaliana* TF with the most similar DBD (based on percent amino acid identity in DBD alignments—see Weirauch et al. [8] for details of how thresholds for these inferred binding specificities are established).

We then scored the alleles of each genetic variant using the resulting 8-mer *E*-score matrix. For a given variant, we first determined all 8-mers in the reference genome sequence overlapping each allele—for example, a SNP will overlap eight 8-mers, plus their reverse complements, for each allele. For each PBM experiment, we then identified the highest scoring 8-mer *E*-score attained by any of the reference allele sequences (E_{ref}), and the highest attained by any non-reference allele ($E_{non\text{-}ref}$). We then identified all PBM experiments where only one of E_{ref} and $E_{non\text{-}ref}$ has an *E*-score value exceeding 0.45 (values above this threshold will likely be strongly bound by the given TF [15]). All experiments meeting this criterion were then assigned a final score E_{final}, which is the maximum value of (E_{ref} and $E_{non\text{-}ref}$). Finally, we also calculated the predicted difference in binding strength between the two alleles as $E_{delta} = |E_{ref} - E_{non\text{-}ref}|$. We then created a final ranked list of TFs (sorted by E_{final}) whose binding is likely to be affected by the alleles of a given SNP (*e.g.*, strongly binding to one allele, but not binding to the other).

2.4. Web server implementation

The user interface to the CressInt analysis pipeline is served by a GNU/Linux virtual machine running CentOS 6 and the Apache 2.2 web server. The web front-end is implemented primarily as HTML "templates" rendered through the use of a PHP library (http://twig.sensiolabs.org/), which maintains a separation of concerns between interface and application logic. Client-side JavaScript manages interaction among input form elements in the web front-end, and the form submission is done asynchronously (*via* Ajax), allowing certain types of validation errors such as missing inputs or malformed BED files to be detected and reported without a page reload. Input data for analysis is received and processed by a Perl CGI (Common Gateway Interface) script, which in turn inter-