2019

# Development of Computational Techniques for Identification of Regulatory DNA Motif

Cankun Wang
*South Dakota State University*

# DEVELOPMENT OF COMPUTATIONAL TECHNIQUES FOR IDENTIFICATION OF REGULATORY DNA MOTIF

BY

CANKUN WANG

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Plant Science

South Dakota State University

2019

DEVELOPMENT OF COMPUTATIONAL TECHNIQUES FOR IDENTIFICATION OF REGULATORY DNA MOTIF

CANKUN WANG

This thesis is approved as a creditable and independent investigation by a candidate for the Master of Science degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Anne Fennell, Ph.D.
Thesis Advisor Date

David Wright, Ph.D.
Head, Department of Agronomy, Horticulture and Plant Science Date

Dean, Graduate School Date

# ACKNOWLEDGMENTS

First, I would like to thank my academic advisor Dr. Anne Fennell, who has helped me a lot during my graduate study. Her brilliant insights in bioinformatics have a profound influence on my research, responded to my questions and queries promptly despite her busy schedules. I much appreciate the freedom that Dr. Fennell and Dr. Ma have given me to do my research and develop myself as a researcher in the best possible way. This thesis would not have been possible without her guidance and persistent help. I have learned a lot such as how to learn a new skill, how to give presentations, how to read and write papers, and all the necessary and essential quantities that a graduate student should have.

I want to thank my committee members also as my first-year advisor, Dr. Qin Ma, his insightful comments and useful suggestions in terms of my slides and research plan.

I want to thank my graduate advisory committee member Dr. Trevor Roiger for his time and efforts in helping me obtain my degree.

I would also like to thank all Bioinformatics and Mathematical Biosciences Lab (BMBL) members, Anjun Ma, Juan Xie, Yuzhou Chang, Adam McDermaid, Jinyu Yang, Shaopeng Gu, Zhaoqian Liu, Jing Jiang, Junyi Chen, Weiliang Liu, Zichun Zhang, Minxuan Sun, Jennifer Xu for all types of your support and comments.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ABSTRACT

## DEVELOPMENT OF COMPUTATIONAL TECHNIQUES FOR IDENTIFICATION OF REGULATORY DNA MOTIF

CANKUN WANG

2019

Identifying precise transcription factor binding sites (TFBS) or regulatory DNA motif (motif) plays a fundamental role in researching transcriptional regulatory mechanism in cells and helping construct regulatory networks for biological investigation. Chromatin immunoprecipitation combined with sequencing (ChIP-seq) and lambda exonuclease digestion followed by high-throughput sequencing (ChIP-exo) enables researchers to identify TFBS on a genome-scale with improved resolution. Several algorithms have been developed to perform motif identification, employing widely different methods and often giving divergent results. In addition, these existing methods still suffer from prediction accuracy.

Thesis focuses on the development of improved regulatory DNA motif identification techniques. We designed an integrated framework, WTSA, that can reliably combine the experimental signals from ChIP-exo data in base pair (bp) resolution to predict the statistically significant DNA motifs. The algorithm improves the prediction accuracy and extends the scope of applicability of the existing methods. We have applied the framework to *Escherichia coli k12* genome and evaluated WTSA prediction performance through comparison with seven existing programs. The performance

evaluation indicated that WTSA provides reliable predictive power for regulatory motifs using ChIP-exo data.

An important application of DNA motif identification is to identify transcriptional regulatory mechanisms. The rapid development of single-cell RNA-Sequencing (scRNA-seq) technologies provides an unprecedented opportunity to discover the gene transcriptional regulation at the single-cell level. In the scRNA-seq analyses, a critical step is to identify the cell-type-specific regulons (CTS-Rs), each of which is a group of genes co-regulated by the same transcription regulator in a specific cell type. We developed a web server, IRIS3 (Integrated Cell-type-specific Regulon Inference Server from Single-cell RNA-Seq), to solve this problem by the integration of data pre-processing, cell type prediction, gene module identification, and *cis*-regulatory motif analyses. Compared with other packages, IRIS3 predicts more efficiently and provides more accurate regulon from scRNA-seq data. These CTS-Rs can substantially improve the elucidation of heterogeneous regulatory mechanisms among various cell types and allow reliable constructions of global transcriptional regulation networks encoded in a specific cell type.

Also presented in this thesis is DESSO (DEep Sequence and Shape mOtif (DESSO), using deep neural networks and the binomial distribution model to identify DNA motifs, DESSO outperformed existing tools, including DeepBind, in 690 human ENCODE ChIP-Sequencing datasets. DESSO also further expanded motif identification power by integrating the detection of DNA shape features.

# CHAPTER 1. Introduction

## 1.1 DNA sequence motif

A DNA sequence motif (DNA motif) is defined as a nucleic acid sequence pattern that is short and recurring and has some biological significance such as being DNA binding sites for a regulatory protein, i.e., a transcription factor (TF) [1]. Sequence motifs are short (usually 5 to 20 base-pairs (bp) long), recurring patterns in DNA that are presumed to have a biological function. Often they indicate sequence-specific binding sites (TFBS) for proteins such as nucleases and transcription factors [2], [3], TFBS are frequently located near the transcription start site (TSS) of the gene (i.e. proximal promoter region) or further apart (enhancers, silencers, etc.)[4]–[6] (Figure 1). With the progress of molecular biology, particular types of DNA motifs are recognized: palindromic motifs and spaced dyad (gapped) motifs. The palindromic motif was discovered during the study of restriction endonucleases the late 1960s [7]–[10], known to serve functionalities such as the formation of DNA[11], RNA transcription[12]. From the structure perspective, the palindromic motif is a subsequence that is precisely the same as its reverse complement, and exist in double-stranded and not the single-stranded DNA, for example, the two palindromic sequences recognized by restriction enzyme *Eco* RI. They are usually referred to as "reverse palindromes"[13]:

$$G\,A\,A\,T\,T\,C$$

$$C\,T\,T\,A\,A\,G$$

The spaced dyad motif consists of two smaller conserved sites separated by a spacer (gap) of fixed length but might be slightly variable [14], [15]. In this case, the TF forms a dimer, and each unit that binds to the DNA are conserved but are typically rather small (3–5 bp). Such spaced dyad elements are common to a large class of transcription factors [1].

## 1.2 Representation of DNA motif

A single TF can recognize dozens to hundreds of DNA binding site sequences over a range of binding affinities. Hence, the TF binding specificity cannot be adequately represented using any one DNA sequence. Instead, TF binding specificities are often represented as binding site motifs, which summarize the collection of preferentially bound sequences [16]. The most straightforward model to denote the binding preference of a TF on each position along a motif is the consensus sequence, which is composed of the concatenation of the most frequent nucleotide on each position. Pribnow *et al.* discovered the 'TATAAT box' in 1975, a well-conserved sequence centered around 10 bp upstream of the transcription initiation site of *Escherichia coli* promoters [17]. In such case, we can denote a set of TFBSs with a single oligonucleotide, and the standard ambiguity codes are introduced by International Union of Pure and Applied Chemistry (IUPAC) to indicate possible nucleotides to occur at a given position [18]. For example, $V$ means this binding site position could be recognized as either $G, C$ or $A$; the complete mapping of the IUPAC nucleotide codes are available in Table 1. A case-sensitive extension to the IUPAC codes is proposed to assist in the representation of the rapidly growing space of information in human genetic variation by considering more interrelationships of nucleic acids [19]. However, a set of exactly matched consensus sequences are actually extremely rare, most positions of binding sites do not show a definite preference for a nucleotide. Thus, one of the main restrictions of consensus sequence is that it presents distorted pictures of binding sites [20], for example, a position that is always ‘A’ is treated the same as the position has 70% ‘A’ in an aligned set of DNA sequence motifs.

Although the consensus presents the characteristics of a motif in each position in a simple and clear way, the variations in this motif are absent in this model. A more accurate and most commonly used model is the position weight matrix (PWM) model [21]–[23], it describes the probability of a given nucleotide's occurrence at each position in the DNA binding site [16]. The standard PWM model assumes that each position of the nucleotide contributes independently to the binding. The model first obtains a position frequency matrix (PFM) on each nucleotide position, and a normalized position probability matrix (PPM) is obtained by calculating the relative frequencies of each nucleotide at each position, pseudocounts are usually applied when calculating PPMs in order to avoid matrix entries have a value of 0 [24]. Finally, the PWM is obtained by logarithmic transformation of the PPM divided by the nucleotides’ background probabilities. We calculate the Information Content (IC) [25]–[28] to measure how different a given PWM is from a uniform distribution. It corresponds to the Kullback–Leibler divergence or relative entropy [29], the IC can be calculated as the sum of the expected self-information of every element:

$$I(i) = -\sum_{b} p_{b,i} \log_2 \frac{p_{b,i}}{b_i}$$

Where $b_i$ is the background frequency of base $i$.

With the PWM model, the intuitive visualization method called sequence logo [30] has widely replaced the earlier consensus-based DNA motif representation method, the four possible nucleotides are stacked at each position where the height is scaled with the IC of the base frequencies at that position. The sequence conservation at a particular position in the alignment is defined as the difference between the maximum possible entropy and the entropy of the observed symbol distribution: [30], [31]

$$R_{seq} = R_{max} - R_{obs} = \log_2 N - \left( -\sum_{n=1}^{N} p_n \log_2 p_n \right)$$

Where $n$ is the particular sequence position among all the distinct symbols for the given sequence type $N$, $p_n$ is the observed frequency of symbol $n$. In terms of the DNA motifs with 4 possible letters, the maximum sequence conservation per site is $\log_2 4 = 2$ bit for DNA motifs. Figure 2 shows an example of motif consensus sequences and its logo generated from WebLogo[31].

The PWM model does not to provide a true picture of the sequence specificity, as PWM assumes the base positions of the sequence motif are independent of each other and studies have shown such independent assumption is not true [32], [33], for example, in the binding sites of zinc finger in proteins.

Numerous methods that model dependencies in DNA motifs have been developed, including learning mixture of PWM [34]–[36], HMM-based method [37]–[39], Tree-Based PWM method [40], feature-based method [41], [42], Markov Chain based method [43], Bayesian Markov based method [44]. However, the usefulness of such more complex models has been controversially discussed, most transcription factors PWMs performed as well as more complex models to predict PBM binding strength [32], [45], [46].

### 1.3 ChIP techniques

The rapid development of show chromatin immunoprecipitation (ChIP) technologies [59]–[71] permit the genome-wide identification of protein–DNA interactions and massive yields of data in recent years provide an unprecedented opportunity to discover DNA motif [60], [61]. The most widely used technique is the chromatin immunoprecipitation followed by sequencing of the immuno-precipitated DNA (ChIP-seq), the overview of ChIP-seq protocol is shown in Figure 3, the chromatin is isolated from cells or tissues and fragmented. Antibodies against chromatin-associated proteins are used to enrich for specific chromatin fragments. The DNA is recovered, sequenced and aligned to a reference genome to determine specific protein binding location [62], [63]. Peaks are generated from the alignment results, referring to the site where multiple reads have mapped a pileup, that indicate a higher possibility of identifying a potential DNA motif [64]. An extensive amount of ChIP-Seq data has been generated and is available in the public domain, i.e. ENCODE [65], [66], ChIP-Atlas [67], GTRD [68].

The chromatin immunoprecipitation combined with lambda exonuclease digestion followed by high-throughput sequencing (ChIP-exo) [48] is developed as a variation of ChIP-seq assay to improve sensitivity and positional resolution by up to two orders of magnitude. Compared with ChIP-seq protocol, shown in Figure 4, it uses lambda exonuclease to digest sonicated chromatin to the formaldehyde-induced protein-DNA cross-linking point [69]. By providing near base pair (bp) resolution of protein-DNA interactions, ChIP-exo can identify almost single-nucleotide-resolution binding sites of TFs [70], [71], the binding resolution is significantly better than the ChIP-seq protocol (Figure 5).

## 1.4 Motif identification techniques

Identifying *de novo* DNA motifs has been an essential and challenging task in bioinformatics. The basic computational assumption of motif identification is that they are overrepresented conserved patterns in given sequences, and once identified will show significant conservation compared to background sequences [72], and the researchers understand gene regulatory networks[73], [74]. Over the past decades, numerous DNA motif identification techniques have been developed, Figure 6 displays some of these methods. The methods mainly fall into two categories: word-based methods (i.e., word-based) and profile-based methods[1], [33], [75]. The word-based methods usually use the $(l, d)$-motifs, where where $l$ is the width of a motif and $d$ is the maximum number of mutations between a motif instance and the consensus sequence, such as DREME [76], FMotif [77], RSAT [78], [79], CisFinder [80], SIOMICS[81], Discrover [82], and BoBro [83], additional tools and their brief descriptions are shown in Table 2, For a given motif consensus sequence collection of all possible occurences (with allowed mismatches) could be formalized and significance determined. While used in the early stages of bioinformatics with real case studies, the method is considered too time-consuming for large scale motif identification as it has an exponential complexity [84]–[86]. The word-based methods search the sequences with a fixed length and a tolerance of mutations since several TFs are already known from the accumulations of research studies. The identified candidate motifs can be compared to known motifs databases with tools like TOMTOM [87] and obtain a similarity score. The word-based strategy can identify optimal global solutions but suffers from high false-positive ratio issue and high computational complexity when applied to large biological datasets. Profile-based methods usually take from the motif profile score [88], [89], i.e. IC from PWM, or randomly selected [90], the profile-based methods try to find a collection of DNA segments, giving rise to a motif profile with the highest score among all the combinations of candidates. The profile-based methods have better performance when predicting motifs with complex mutations. However, profile-based methods are limited in detecting multiple motifs when the data size is large.

DNA motif identification techniques have been applied to different types of data. Recently, studies show that ChIP techniques can be effectively integrated into the motif discovery, the provided high-throughput peak signals bound by the TF investigated are the ones showing enrichment over a control sample, expressed as the difference between the number of times each base pair of the genome has appeared in the sequenced IP sample versus the control [64]. The DNA sequence sets obtained from ChIP-enriched peak regions of the input is significantly larger than thr traditional approaches which use a provided set of DNA promoter sequences, typically a few hundred sequences. Some widely used motif identification tools, i.e. MEME and RSAT [91], [92] cannot directly use ChIP peak data. Recently, new tools that are specially designed to handle the large volumes of data generated from high-throughput technologies, i.e. MEME-ChIP [93] and HOMER [94]. MEME-ChIP combines two different motif identification algorithms MEME (Multiple expectation maximization (EM) for Motif Elicitation, an extension of the EM algorithm [95], [96]) and DREME (Discriminative Regular Expression Motif Elicitation), to discover novel DNA sequence motifs. To detect enrichment of previously characterized functional motifs for TF or RBP binding sites in the sequences, MEME uses a Fisher's exact test [93] for calculating the significance of relative enrichment of each motif in two sets of sequences. One set is the set of ChIP-Seq peak regions and the other is either similar data from a different ChIP-Seq experiment or shuffled versions of the first sequences (central motif enrichment analysis or CentriMo [97]). Finally, to ease interpretation of the results, MEME-ChIP applies AME algorithm [93] to group the discovered and enriched motifs by similarity to each other [99]. MEME-ChIP used MAST [100] and AMA [101] algorithms for visualizing motifs as well as for binding strength analysis. HOMER (Hypergeometric Optimization of Motif EnRichment) is a suite of tools for DNA motif discovery, Motif scanning, and next-generation sequencing data analysis. HOMER supports various popular assays like ChIP-seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C. HOMER provides the *de novo* motif discovery algorithm that designed to find DNA motifs in a large scale genomics data with about 10,000 targets sequences. The input of HOMER suite requires a target and background sequences, promoters of genes that are likely to be co-regulated and promoters of genes that are not regulated, respectively, or a standard peak file that allows HOMER to extract target and background sequences from a reference genome and a set of known-motif profile from the motif database. When the target and background sequences are set, HOMER starts scanning the specific length and over-represented motif patterns in the target sequences relative to the background sequences, the enrichment score of each motif is calculated using the cumulative hypergeometric distribution, or cumulative binomial distribution for motif scoring if the user specify a large number of input sequences, which is faster to calculate and gives essentially a same results. The latest HOMER version adds a procedure in the workflow by revisiting the input sequences to remove the oligos that are slightly offset from the original motifs, which makes it more sensitive to co-enriched motifs.

In terms of usage, the command-line version of the program provides full control by the researchers and can be integrated into larger workflows, researchers are interested in developing a web server, users with limited computational knowledge or computational resources can efficiently perform DNA motif analysis and visualization through their web browsers. The implementation of motif identification web server generally falls into two categories either a pipeline implementation that incorporates a suite of existing tools into a web server, or implementing novel algorithms into a web server [102]. Generally, DNA motif identification web server requires users to upload input sequences of DNA, RNA, proteins or custom alphabet datasets, and provide an E-mail address so they can be notified when the submitted job is done. However, as bioinformatics web servers continue to grow, the challenge of service maintenance increases, only 45% of all services published on Nucleic Acids Research Web Server Issues between 2003 and 2009 are now positively confirmed functional [103],  and over 95% of sites were running in the first 2 years, but this rate declined to 84% in the third year and continued to decrease gradually [104].

## 1.5 Single-cell RNA sequencing

For the last decades, RNA sequencing (RNA-seq) has been a popular method to study global gene expression changes using next-generation sequencing (NGS) technique to reveal the presence and quantity of RNA in a biological sample at a given moment [105], [106], providing tens to hundreds of millions of sequence read fragments and information on billions of individual bases. The rapid development of single-cell RNA sequencing (scRNA-seq) technologies has provided massive amounts of data.  Mining the activity of thousands of individual cells has allowed researchers to identify gene transcriptional regulation at the single cell level[107], [108]. This technology can be summarized as follows: Isolating the single-cell from the system of interest, this is one of the significant challenges for performing high-throughput and unbiased single cell experiment [109]; reverse transcription (RT); amplification; library generation and sequencing [110], [111]. After obtaining the scRNA-seq data, a typical single-cell RNA-seq analysis workflow can be processed as pre-processing (quality control, normalization, data correction, feature selection, dimensionality reduction, and visualization) and cell-level, gene-level downstream analysis [112]. Relative to traditional profiling methods that assess bulk cell populations, this research direction holds promising potential for providing an unprecedented opportunity to allow researchers to uncover unexpected biological discoveries [113], including but not limited to, predicting cell types [114], analyzing cell trajectory paths [115], and revealing the heterogenous regulatory mechanism [116] in various cell states. Regulon is a maximal group of co-regulated genes by the same TF or the same set of TFs spread out in a genome [117]. The successful identification of regulons at the single-cell level can substantially improve the elucidation of heterogeneous gene regulation mechanisms across various cell types and allow reliable constructions of global transcription regulation networks encoded in a specific cell type.

The higher resolution of cellular differences detected by single-cell sequencing also raises a host of new questions. Performing successful scRNA-seq experiments requires that the expertise from various disciplines, although the data obtained from scRNA-seq are often structurally identical to those from a bulk expression experiment [118], scRNA-seq data is extremely sparse (There is no expression measured for many genes in most cells). Simply applying traditional RNA-seq analysis methods to scRNA-seq data may not obtain a satisfied result, new methods specially design for scRNA-seq are proposed at an astonishing rate, the number of available tools nearly doubled in one year (2018), which increased from 165 to 337 [119].

## 1.6 Outline

The rest of this thesis is organized as follows: Chapter 2 introduces a novel method, WTSA, for identifying DNA motifs from ChIP-exo data. Chapter 3 introduces a web server, IRIS3, which performs cell-type-specific regulon inference from scRNA-seq. Chapter 4 briefly introduces DESSO, a new method for DNA motif prediction using deep neural networks and the binomial distribution model.

# CHAPTER 2. WTSA - An integrative framework using ChIP-exo data for accurate prediction of DNA motifs

## 2.1 Introduction

Identifying DNA motif has been a major challenge to unravel the regulation of gene expression mechanism, the recent development of high throughput sequencing technologies have revolutionized our understanding of transcriptional regulation by providing an unprecedented opportunity to interrogate *in vivo* transcription factor binding [120]. ChIP-Seq provided a view of genome-wide interactions between DNA and DNA-associated proteins and employed extensively to discover motifs from overrepresented sequences in ChIP-seq peaks [121]. Although a variety of popular methods have been developed for ChIP-Seq data mining and modeling, both computational and experimental challenges remain for the accurate and exhaustive identification of DNA motifs. The ChIP-seq assay requires a large amount of sample material and output a relatively low resolution (200–500 bp) due to the size of DNA fragments generated by chromatin sonication. Recent studies show the observations that ChIP-seq peak scores fail to differentiate between bound versus unbound genomic sequences [122]–[124], the question of what constitutes the minimal sequence determinants for DNA motifs *in vivo* has become increasingly uncertain.

A Higher resolution mapping of bound genomic sequences has been facilitated by the development of ChIP with lambda exonuclease digestion and sequencing (ChIP-exo), by performing several enzymatic reactions including the lambda exonuclease digestion step while protein-DNA complexes are still on the beads prior to sequencing [125], ChIP-exo has significantly improved the resolution of the ChIP-based technologies.

Although numerous DNA motif finding algorithms and tools have been developed for ChIP assays, current methods rely exclusively on DNA sequences extracted from ChIP-enriched regions, and look for DNA motifs of a specific length from an input parameter; the optimal motif lengths was simply obtained by the strategy of iterating the method multiple times for a vector of fixed length [93], [94], [126]–[133], and filter the result based on the significance values. Thus, the results suffer from both lack of specificity (false predictions) and high computation time. To push the prediction accuracy further, compared with ChIP-seq that ChIP-exo achieves near base pair resolution and a piece of structural information on genome-wide binding proteins [134], we hypothesized such nucleotide level sequencing reads information from ChIP-exo can be integrated as an enhancement in the DNA motif identification process.

We designed a weighted two-stage alignment (WTSA) tool by specifically obtaining each nucleotide level weight score from the enriched ChIP-exo regions and inheriting the two-stage alignment and graph-theory based model from BoBro [83]. WTSA has the following unique features to improve the state-of-the-art performance: (i) developed a binomial distribution scoring model to handle the unquantifiable scoring preference when dealing with by motif length window (ii) integrated a weight matrix extracted from the normalized ChIP-exo reads, to assess the probability of nucleotides to be within a DNA motif (iii) employed a dynamic extension strategy to optimize the motif length to be more closely to the actual DNA motif length.

## 2.2 Methods

The WTSA workflow (Figure 7) can be described as follows:

*Step 1: data pre-processing.* To obtain a nucleotide level weight score from ChIP-exo enriched regions, WTSA requires two input file, reference genome file (FASTA format) and ChIP-exo read alignment file (SAM/BAM), two integrated tools, MACE [135] and BEDTOOLS, are used to perform ChIP-exo peak calling and extract the DNA sequences and weight scores, respectively. We select flanking regions 100 bp centered on each ChIP-exo peak, we define a format with the file extension 'wtsa' as the WTSA input. Similar to the FASTA format, the wtsa format begins with a single-line description, followed by lines of sequence data, while the third line represents the weighted scores extracted from BEDTOOLS.

*Step 2: weighted two-stage alignment.* Read the previous defined WTSA format file, initialize with a normalized ChIP-exo weight scores matrix $M^h$, and two auxiliary matrix $M^1$ and $M^2$ with the duple size of $M^h$ (the even rows represent the reversed complementary sequences), and set all elements equal to 0. For all segment pairs $s_{ij}$and $s_{pq}$ of length $l$ with $k$ position identity in the input sequences and their reversed complementary reversed sequences, we calculate $f$and $f'$ also, store the $f'$to $M^1$:

$$f' = f(s_{ij}, s_{pq}) \times (M^h_{ij} + M^h_{pq})$$

where $f(s,t) = -\lg\left(\sum_k^l B(l,k,p)\right)$, B(.) is binomial distribution and p=0.25. And set $f$ = 0 if $\sum_k^l B(l,k,p) > 0.01$. $f'$ is among the top *t* in this alignment between two sequences or f >3, add 1 or 0.5 (if the two nucleotides before $s_{ij}$ and $s_{pq}$ are identical) to $M^1_{ij}$ $and$ $M^1_{pq}$.

Similarly, we calculate $f$and $f'$ also, store the $f'$to $M^2$:

$$f' = f(s_{ij}, s_{pq}) \times \max_{\substack{j-2\le j'\le j+2 \\ q-2\le q'\le q+2}} (M^1_{ij'} + M^1_{pq'})$$

*Step 3. Graph construction and optimize clique finding*. For each segment pair, if $f'$ is among the top *t* in this alignment between two sequences, we build a graph $G$ with $s_{ij}$ and $s_{pq}$ be vertex, and edge between $s_{ij}$ and $s_{pq}$ with weight $f'$ if and only if:

$$f' = f(s_{ij}, s_{pq}) \times \max_{\substack{j\le j'\le j \\ q\le q'\le q}} (M^2_{ij'} + M^2_{pq'})$$

For an empty set $C$, choose an edge $(u,v)$ with the largest $N_G(u) \cap N_G(v)$ with $N_G(x)$ representing all the vertices incident to vertex $x$; add $u$ and $v$ to the current clique $C$; Repeat the above on the sub-graph induced by $N_G(u) \cap N_G(v)$ until the subgraph is empty; remove the current clique from $G$, and repeat this step on the remaining graph for $w$ times (the default is $w$=10). To obtain an optimized motif length, for a set $C$ of motif candidates of $l$ bp length long, calculate overlapped region frequency from candidate motif, and obtain an optimized length $l'$ from the combined overlapped regions with the maximum frequency, repeat the process of constructing set $C$ with the length $l'$.

*Step 4. Motif evaluation.* We found p(x) is very close to a Poisson distribution, we define a profile matrix $P_C$ of $C$ as

$$P_C = \left(\log \frac{P(i,j)}{q(i)}\right)_{4\times l'}$$

Where $P(i,j)$ is the probability of nucleotide type $i$ appearing at position $j$ in the alignment, and $q(i)$ is the probability of $i$ appearing in the simulated background sequence. Define the match score between a candidate motif and a profile matrix as the sum of corresponding values of the matrix based on the specific nucleotide in each position of the motif. Let $x$ be a random variable denoting the number of sequence segments of length $l'$ from a set of random nucleotide sequences, $p(x)$ is the probability distribution of $x$, we found $p(x)$ is very close to a Poisson distribution, we calculate the $P$-value of a set of candidate motif by summing up $p(x)$ if $x$ is larger than the average match score over all the sequence segments $C$, where

$$p(x) \approx \frac{e^{-\mu}\mu^{x}}{x!}$$

Finally, the motif closures are sorted in the increasing order of their $P$-values, and output top $o$ results, with $o$ being a parameter set by the user with default value 10.

### 2.3 Dataset

We have extracted publicly available SRA datasets for 10 different TFs from experiments performed in *Escherichia coli* (*E. coli)* K-12 by ChIP-exo (Fur, Cra, ArgR, GadE, GadW, GadY, OxyR, UvrY, SoxR and SoxS), 3 of them (Fur, Cra and

ArgR) were used for the evaluation as others have limited annotations on RegulonDB [136]  or TOMTOM motif database and cannot used for evaluation, details of the used datasets are described in Table 3.

To assess the motif-finding performance of our method, we compared the prediction results of WTSA with six de-novo motif finding tools: BoBro[83], Bioprospector [130], MEME-ChIP [93], HOMER [94], rGADEM [137] (Genetic Algorithm guided the formation of spaced Dyads coupled with EM for Motif identification) and ChIPMunk [138]. BoBro is the previous version of WTSA; The Bioprospector is a well-known conventional motif discovery program; MEME-ChIP integrated the classic MEME program, HOMER, rGADEM and ChIPMunk are designed able to handle the large volumes of data generated from these high-throughput technologies. The goal of the proposed algorithm is to precisely identify the TFBS location from 100 bp long DNA sequence at single nucleotide resolution and binding site resolution. That is, for each nucleotide of the input sequence, we aim to determine whether the base-pair categorizes to the binding sites from the RegulonDB and TOMTOM Prodoric database, we have compared the performance of WTSA and other tools in terms of precision, recall and F-score. We use default parameters for each of them.

## 2.4 Result

For each target binding site with overlapping predicted binding sites in an input sequence, we use the following values previously defined as the DNA motif evaluation matrics: nTP (true positive), the number of target binding site positions predicted as binding site positions; nTN (true negative), the number of non-target binding site positions predicted as non-binding site positions; nFP (false positive), the number of non-target binding site positions predicted as binding site positions; nFN (false negative), the number of target binding site positions predicted as non-binding site positions. sTP is the number of known sites overlapped by predicted sites; sFN is the number of known sites not overlapped by predicted sites; sFP is the number of predicted sites not overlapped by known sites; [73], [139]

The sensitivity on nucleotide level nSN and site level sSN are defined as:

$$nSN = \frac{nTP}{nTP+nFN} \quad sSN = \frac{sTP}{sTP+sFN}$$

specificity on nucleotide level nSN and site level sSN are defined as:

$$nSP = \frac{nTN}{nTN+nFP} \quad sSP = \frac{sTN}{sTN+sFP}$$

positive prediction value on site level is defined as:

$$sPPV = \frac{sTP}{sTP + nFP}$$

We also used the F-score or F1-score as the overall accuracy measurement. Compared with geometric or arithmetic mean, it tends to penalize more the imbalance of sensitivity and specificity. The nucleotide and TFBS level F-score are defined as $nFscore$ and $sFscore$, respectively:

$$nFscore = \frac{2 \times nSN \times nPPN}{nSN + nPPV}$$

$$sFscore = \frac{2 \times sSN \times sPPN}{sSN + sPPV}$$

The combined motif logo from all methods, including the reference logo from RegulonDB, is shown in Figure 8. We began by making similarity comparisons between motifs predicted by WTSA from ArgR, Fur, Cra TF ChIP-exo datasets and the experimentally confirmed, strongly validated and weakly validated motifs in the RegulonDB databases. These comparisons were extended to the other seven existing methods described previously. Figure 9A shows that WTSA achieved a stable high motif prediction performance on the TFBS level F-score comparisons, the rGADEM program outperforms on the Cra TF data at the TFBS level F-score, while WTSA has the best positive prediction value on the Cra TF data (Figure 9B). The nucleotide level performance comparison shows WTSA achieved the highest F-scores, sensitivity and positive prediction value on all three datasets, indicating the integration of base pair resolution ChIP-exo data weight scores enhances the ability to accurately predict the actual DNA motif region (Figure 10A and Figure 10B).

To assess the similarity of query motifs against validated motifs, TOMTOM was used to compare the statistical significance (i.e., E-value, and q-value) across JASPAR and Prodoric database for DNA motifs that were predicted by all the methods in comparison. Figure 11 shows WTSA provides stable prediction results on the $-log2$(E-value) and $-log2$(Q-value) metrics, WTSA outperforms on Fur and ArgR TF than all other methods, MEME-ChIP slightly performed better than WTSA on Cra TF (Figure 10).

## 2.5 Summary

The combination of large-scale ChIP-exo data holds a promising potential in the DNA motif identification. However, existing DNA motif identification tools fail to generate satisfactory results from high-resolution ChIP-exo data due to the lack of full consideration of the intrinsic characteristics of ChIP-exo data. Validation using comprehensive data sets showed that WTSA reliably identifies the correct DNA motifs with improved base pair level quality.

# CHAPTER 3. IRIS3 - Integrated Cell-type-specific Regulon Inference Server from Single-cell RNA-Seq

## 3.1 Introduction

One of the major challenges in molecular biology is reverse-engineering the *cis*-regulatory logic that plays a significant role in the control of gene expression, application of DNA motif identification has limitedly applied in parallel with methods used in underlying gene regulatory mechanisms that induce the identity of cell types or physiological states usually uncovered in the scRNA-seq analyses. A critical step in this process is to identify the cell-type-specific regulons (CTS-Rs), defined to denote a group of genes controlled by the same transcription regulator (e.g., TF and long non-coding RNA) in a specific cell type. Intuitively, the component genes of a CTS-R tend to be co-expressed in the specific cell type and share the same conserved *cis*-regulatory motif (DNA motif) of the underlying regulator. The identification of CTS-Rs is non-trivial and essential in characterizing the transcriptomic heterogeneity of cell components in tissues.

For the first time, the SCENIC pipeline identified 151 regulons and, based on which, predicted eight cell types from 3,005 adult mouse brain cells [140]. Specifically, this pipeline identified TF based on co-expression analysis, identified the gene modules significantly enriched with TF-binding motifs as regulons and predicted cell types by clustering of a regulon-cell matrix containing regulon enrichment values in each cell. Based on a modified SCENIC pipeline, a Mouse Cell Network Atlas was built in 2018 to construct a global gene regulatory network of

202 cell-type-specific regulons (CTS-Rs), containing 8,461 genes in 61,637 cells sampled from 98 cell types [141]. More regulon inference methods that are specifically designed for scRNA-seq data have been developed: SCODE [142], PIDC [143], recent studies have shown both bulk RNA-seq and scRNA-seq methods perform poorly on predicting gene regulatory structures from scRNA-seq data [144]. There is still a room for the regulon prediction performance considering the false positive issues in cell type prediction, gene module identification, and motif discovery; The practical usage of SCENIC pipeline requires substantial programming experience in R, even with a detailed tutorial; and the identified CTS-R are not intuitively and comprehensively represented through a web server.

We have developed IRIS3, the Integrated Cell-type-specific Regulon Inference Server from Single-cell RNA-Seq, as the first-of-its-kind web server for CTS-R inference for multiple species, described in the pipeline overview (Figure 12), IRIS3 solve the problem computationally by the integration of data pre-processing, cell type prediction, gene module identification, and DNA motif analyses.

### 3.2 Overview

IRIS3 requires one input file, which is a gene expression matrix (GEMAT) with unique gene IDs (rows) and cell names (columns). Both Gene Symbols [145] (e.g., HSPA9) and Ensembl [146] Gene IDs (e.g., ENSG00000113013) are allowed in the GEMAT file, and their expression values can be raw/normalized reads counts or 10x Genomics feature-barcodes matrices. Optionally, a two-column-cell-label file can be used for evaluating the predicted cell types from SC3/Seurat and CTS-R inference. The first column of this cell label file being the cell names exactly match the columns of the GEMAT file and the second column being experimentally validated cell type labels.

## 3.3 Methods

IRIS3 integrates five widely-used tools: Seurat [147], [148], SC3 [149], QUBIC [150], [151],DMINDA2.0 [152], [153] and MEME [126] for CTS-R inference from the scRNA-seq GEMAT, Additionally, several powerful tools and databases such as Enrichr [154], [155] , Clustergrammer [156], Plotly, Ensembl, and GeneCards [157] are implemented in support of the comprehensive interpretation of the identified CTS-Rs. As shown in Figure 12, six steps are included in the IRIS3 pipeline: (*i*) pre-processing, (*ii*) gene module detection, (*iii*) cell type prediction, (*iv*) CTS-gene-module assignment, (*v*) CTS-R inference and (*vi*) Quantifying CTS-R specificity. More details of the six steps and the outputs of IRIS3 have been listed in below:

*Step I: Pre-processing*. The uploaded GEMAT is first pre-processed for universal low-quality gene and cell filtering by removing genes with zero values in more than 95% cells and cells with zero values in 99% genes. Both filtrations are optional but highly recommended to obtain reliable and robust analytical performances [158]. Data normalization, PCA, t-SNE [159], UMAP [160] and marker genes detection are performed on the filtered GEMAT by Seurat.

*Step II: Gene module detection.* The pre-processed GEMAT in *Step I* is then analyzed by our in-house biclustering tool, QUBIC, for co-expressed gene module detection. Each of the identified biclusters represents a group of co-expressed genes in a specific subset of cells. QUBIC has been proven to be one of the top performing methods in capturing a high proportion of biclusters effectively and efficiently [161], [162], which are enriched with functional biological pathways. All the identified biclusters in this step will be saved in support of the following steps.

*Step III: Cell type prediction.* Based on the GEMAT from *Step I*, if the total cell numbers are less than 5000, the cell types are predicted in SC3 by gene distance calculation, PCA dimension reduction, tSNE-k-means clustering, and consensus clustering, all SC3 parameters are set to default, and the optimal number of clusters (*k*) is estimated based on the Tracy-Widom theory on random matrices; Otherwise, the cell types are predicted using Seurat. The output of this step is a two-column-cell-label file with the same format as described above regarding the optional input cell label file.

*Step IV: CTS gene module assignment.* We consider the component genes of a bicluster respond to the regulatory signal in a specific cell type if the cells in the bicluster are highly consistent with the cells in the cell type. To determine the consistency, a hypergeometric enrichment test is performed using the cell components of identified biclusters from Step II, and the cell types predicted from SC3 from Step III (or the uploaded ground-truth cell types). To infer CTS-gene modules, the probability of having x cells of the same cell type in a bicluster of size n from the dataset with a total of N cells can be computed using the following hypergeometric function:

$$P(X = x|N, p, n) = \binom{pN}{x}\binom{(1-p)N}{n-x}/\binom{N}{n}$$

where $p$ is the percentage of that cell type among all cell types in the data set. The p-value of getting such enriched bicluster is calculated as:

$$p\text{-}value = P(X \geq x) = 1 - P(X - x) = 1 - \sum_{i=0}^{x-1} \frac{\binom{pN}{i}\binom{(1-p)N}{n-i}}{\binom{N}{n}}$$

The p-value of a bicluster corresponding to a specific cell type is Bonferroni-adjusted by multiplying by the total number of cell types, and such bicluster is assigned as a CTS-gene module if its adjusted p-value to that cell type is significant (adj.p<0.05). The functional enrichment analysis between genes in the CTS-R and databases, such as KEGG, GO, ProteomicDB, etc. uses the same hypergeometric test and p-value adjustment shown above but compares genes instead of cells. All the enrichment analysis in IRIS3 are performed using EnrichR.

A bicluster is a so-called CTS-gene-module if its cell components are significantly consistent with a cell type (p-value < 0.05, Bonferroni adjusted). Thus, a CTS-gene module is possibly found in multiple cell types, as long as it is significantly enriched in those cell types. The output of this step is the CTS-gene-modules, if present, of each of the identified cell types, which lays a solid foundation of the CTS-R identification.

*Step V: CTS-R inference.* The genes from all the CTS-gene-modules in one cell type are grouped into a nonredundant list, and their 1,000-bp upstream promoter sequences are extracted based on the DMINDA2.0 web server. These sequences are used for DNA motif prediction using both DMINDA2.0 and MEME, with the default parameters. All the identified motifs in a specific cell type are clustered into subgroups using the motif comparison functionality (BBC on DMINDA2.0 or TOMTOM on MEME). For each of the motif clusters, the corresponding nonredundant gene list is named as a CTS-R. Users with developed gene modules that they are interested in or identified by their preferred module detection methods can also upload these modules to IRIS3 and have them analyzed to identify the "module-specific regulon".

*Step VI: Quantifying CTS-R specificity.* To quantify cell-type specificity of a regulon, Suo *et al.* defined a regulon specificity score (RSS) [141], we modified this procedure by adopting an entropy-based strategy [163] and gene set variation analysis (GSVA) [164] that were previously used for gene expression data analysis. For each CTS-R, we use GSVA to calculate the regulon activity score (RAS). To filter non-significant CTS-Rs, we perform Wilcoxon rank sum test on each regulon, the null hypothesis states that the medians from the two RAS populations (whether the cell of that RAS belongs to the specific cell type) are the same. We use the Benjamini-Hochberg procedure to control the false discovery rate, CTS-Rs with adjusted p-value > 0.05 are removed.

For each filtered CTS-Rs, we use a vector to represent the distribution of RAS in the cell population:

$$P^R = (p_1^R, \cdots, p_n^R)$$

where $n$ is the total number of the cells, and RAS are normalized so that:

$$\sum_{i=1}^{n} p_i^R = 1$$

Then we use a vector to indicate whether a cell belongs to a specific cell type:

$$P^C = (p_1^C, \cdots, p_n^C)$$

where

$$P_i^C = \begin{cases} 1, & cell\ belongs\ to\ the\ specific\ cell\ type \\ 0, & otherwise \end{cases}$$

The vector is also normalized so that:

$$\sum_{i=1}^{n} p_i^C = 1$$

Next, we evaluate the Jensen-Shannon Divergence (JSD), which is a commonly used metric for quantifying the difference between two probability distributions, defined as:

$$JSD(P^R, P^C) = H\left(\frac{P^R + P^C}{2}\right) - \frac{H(P^R) + H(P^C)}{2}$$

where $H(P) = -\sum p_i \log p_i$, represents the Shannon entropy of a probability distribution $P$. The range of JSD values is between 0 and 1, where 0 means identical distribution and 1 means extreme difference. Finally, the RSS is defined by converting JSD to a similarity score:

$$RSS(R, C) = 1 - \sqrt{JSD(P^R, P^C)}$$

The identified CTS-Rs are first compared with the marker genes list generated from Seurat, and ranked by the number of overlapped marker genes in that CTS-R, next ranked by the RSS in the decreasing order.

### 3.4 Regulon analytical interpretation

IRIS3 provides detailed analyses for an individual CTS-R to interpret detailed information for the associated genes and motifs. As shown in Figure 15, six co-regulated genes are included in the CT1S-R1, and each of the Gene Symbol and Ensembl Gene ID was linked to its corresponding profiles on the GeneCards and Ensembl datasets, respectively. A local heatmap can be achieved by clicking the “Show Heatmap” button to display the expression level of the eight genes among cells in Cell Type 1. To better illustrate the expression value of the gene sets, we applied a log-transformed for heatmap interpretation:

$$VN_i = \lg(1 + VE_i) - \sum_{i=1}^{n} \lg(1 + VE_i) / n$$

where $VN_i$ indicates the normalized value for gene $i$, $VE_i$indicates the expression value of gene $i$, and $n$ is the total number of cells.

The gene enrichment analysis for the genes can be performed using the enrichment function integrated into the heatmap as described above, or by clicking the "Send gene list to Enrichr" button to view the complete enrichment results on the Enrichr website. The additional ATAC-Seq validation function can be achieved by clicking on the “Show ATAC-Seq peak enrichment” button for human (74 tissues) and mouse (117 tissues) using peak files downloaded from CistromeDB [165]. The TAD supported supplementary gene function can be achieved by clicking on the “Show TAD covered genes” button. Considering that each CTS-R is inferred from CTS-gene modules that may be assigned to multiple cell types, it is likely to have CTS-Rs found in other cell types holding similar motifs that regulating a similar group of genes. To find such similar CTS-Rs in other cell types, IRIS3 performs the motif comparison between selected CTS-R and all CTS-Rs in other cell types. A user can click on the “Show similar CTS-Rs” button under each CTS-R to achieve the result of this function.

Using all default parameters of IRIS3, 678 CTS-Rs were identified in a total of six predicted cell types from the above example data. To interpret all CTS-Rs in each cell type, we integrated Clustergrammer, a powerful and interactive heatmap visualization tool, for the CTS-cell-gene-regulon heatmap display. Both gene compositions of these CTS-Rs and their expression values across different cell types can be intuitively displayed in such a heatmap (Figure 14A). Due to space limitation, only the top 15 CTS-Rs and their corresponding genes are showcased. The CTS-Rs are ranked in the increasing order of the overlapped marker genes and regulon specificity score as described above, and each specific CTS-R is renamed as CTnS-

Rm, where n represents the ID of a cell type and m represents the rank ID of a CTS-R. The green rectangles under a CTS-R indicate the presence of its component genes. The heatmap shows the log-transformed expression level of each gene across all cells. The representative motif shown on the right panel and the interactive motif logo (the 12-bp consensus sequence) can direct users to a detailed motif mapping result page, including the motif p-value, related genes, binding site occurrences, and motif position weight matrix. Performed by Seurat, for each CTS-R, IRIS3 generates two t-SNE plot, Colored by cell type or level of regulon activity, respectively (Figure 17).

Further motif validations can be carried out by comparing the sequence to the JASPER [166] and HOCOMOCO [167] databases using TOMTOM. For the user's convenience, we directly listed the top five matched TFs in the two databases in the table below the two buttons (Figure 15), more DNA motif details can be accessed from clicking the motif logo, including motif logo, motif length, P-value, number of motif instances, and detailed information of motif instances, motif positions on the promoters (Figure 16). Besides the above interpretations of the identified CTS-Rs, IRIS3 provides visualizations and evaluations of the predicted cell types. IRIS3 provides a detailed tutorial page for users who need more information about the web server (Figure 17).

### 3.5 Web server implementation

IRIS3 runs on a Red Hat Enterprise seven Linux system with 16 core Intel Xeon E5-2650 CPU and 48GB RAM, and each task is assigned to four cores and scalable based on the server load. The front-end builds on top of technologies such as JQuery and Bootstrap, the interactive tables and figures are generated utilizing libraries such as DataTables, Plotly.js, and Clustergrammer. We employed PHP for the back-end server implementation, and the data parser workflow is aggregated using the R programming language. All data are stored and managed using a MySQL database.

## 3.6 Summary

The IRIS3 web server is a highly powerful and easy-to-use platform for CTS-R inference with interactive and informative result interpretations. The identified CTS-Rs can substantially improve the elucidation of heterogeneous gene regulation mechanisms across various cell types and allow reliable constructions of systematic transcription regulation networks encoded in a specific cell type. IRIS3 supports the analysis of multiple species, including but not limited to human and mouse, hence, users can upload integrated expression data formed by dual species, e.g. one matrix containing genes from both human and mouse. However, the time complexity might be a limitation in the practical application and usage of IRIS3, when more than 10 cell types provided. In such a situation, IRIS3 tends to identify a relatively large number of regulons and their visualization and interpretations based on multi-omics data are usually time-consuming.

To facilitate more users in scRNA-seq data analysis, we plan to develop a more integrative CTS-R inference pipeline capable of adapting raw sequencing data and providing more functionalities based on the current IRIS3 framework and an in-house RNA-Seq data analysis Shiny server IRIS-EDA (http://bmbl.sdstate.edu/iris-eda) (29). On the other hand, an integrated computational model is under development to handle dropout issue in scRNA-seq, gene module detection, and cell types prediction using an iterative manner in support of more accurate CTS-R inference. The ultimate goal of IRIS3 is to build up a web database consisting of both cell types and the CTS-Rs and link these predictions and their interpretations with specific tumors or other diseased cells. This will lay a solid foundation to infer the underlying global and local gene regulatory networks and their impact on disease development and treatment. It can be further combined with studies in biomedical research such as therapeutic research, cell trajectory analysis, and cancer treatment.

# CHAPTER 4. DESSO – a new method for DNA motif prediction using deep neural networks and the binomial distribution model

## 4.1 Introduction

The identification of transcription factor binding sites and *cis*-regulatory motifs is a frontier whereupon the rules governing protein-DNA binding are being revealed. Here, we developed a new method (DEep Sequence and Shape mOtif or DESSO) for *cis*-regulatory motif prediction using deep neural networks and the binomial distribution model. DESSO outperformed existing tools, including DeepBind, in predicting motifs in 690 human ENCODE ChIP-Sequencing datasets. Furthermore, the deep-learning framework of DESSO expanded motif discovery beyond the state-of-the-art by allowing the identification of known and new protein-protein-DNA tethering interactions in human TFs. Specifically, 61 putative tethering interactions were identified among the 100 TFs expressed in the K562 cell line. In this work, the power of DESSO was further expanded by integrating the detection of DNA shape features. We found that shape information has strong predictive power for TF-DNA binding and provides new putative shape motif information for human TFs. Thus, DESSO improves in the identification and structural analysis of TF binding sites by integrating the complexities of DNA binding into a deep-learning framework.

## 4.2 Methods

The DESSO framework is composed of (*i*) a CNN model for extracting motif patterns from given ChIP-Seq peaks, and (*ii*) a statistical model based on the binomial distribution for optimizing the identification of motif instances (i.e., TFBSs). This framework can accept both DNA sequences and DNA shape features as input to identify sequence and shape motifs, respectively. DESSO enables the extraction of more complex motif patterns compared to existing motif prediction methods owing to its multi-layer network architecture. We designed a binomial-based model in DESSO to identify all the significant TFBSs under the statistical hypothesis that the number of random sequence segments that contain the motif of interest in the human genome is binomially distributed (Figure 19).

The first layer of the CNN model contains multiple convolutional filters, which were used to identify low-level features from given ChIP-Seq peaks. A subsequent max pooling layer and a fully connected layer were used to extract high-level features based on the output from the convolutional layer. Specifically, the CNN model takes DNA sequences centered on the ChIP-Seq peaks as input query sequences and learns motif patterns using convolutional filters (denoted as motif detectors). Then, a large set of background sequences was selected from the human genome, considering GC content, CpG frequency, and promoter and repeat overlap to eliminate biases created by these features. Both the query and background sequences were then aligned as sequence matrices, where each row represents a distinct sequence. For each optimized motif detector, two motif signal matrices were derived by sliding the detector along the query sequence matrix and background sequence matrix, respectively. Each element of a signal matrix represents the occurrence probability of the corresponding motif detector on a sequence segment in the corresponding sequence matrix. These two motif signal matrices were then used to generate motif candidates by varying a motif instance signal cutoff in a predefined interval. For each value of the motif signal cutoff, the motif instance candidates in the query sequence matrix and background sequence matrix were obtained and then used to calculate a p-value according to the binomial distribution. The optimal motif instances for a motif detector were finally determined as the motif instance candidates in the query sequence matrix that correspond to the minimum p-value.

## 4.3 Results

We began by making similarity comparisons between motifs predicted by DESSO from 690 ENCODE TF ChIP-Seq datasets and experimentally validated motifs in the human JASPAR and TRANSFAC databases using TOMTOM. These comparisons were extended to other five existing methods in this field, i.e., DeepBind, Basset, MEME-ChIP, KMAC, and gkm-SVM. The results showed that DESSO significantly improved the motif prediction performance on 161 TFs in 91 cell lines (Figure 20A), covered by the above ChIP-Seq datasets. Known motifs and undocumented motifs are grouped by whether motif can be matched in TOMTOM databases (Figure 20B).

DESSO also outperform of DNA shape in predicting TF-DNA binding specificity (Figure 21). DESSO was applied to five different inputs, i.e., HelT, MGW, ProT, Roll, and DNA shape combination. (b) The AUC of the five inputs above using single and two convolutional layers based on the 690 ChIP-Seq datasets. The Wilcoxon test p-values between one-layer and two-layer model. (c) The contribution of HelT (32%), MGW (9%), ProT (22%), and Roll (37%) in DNA shape combination in predicting TF-DNA binding specificity. (d) The heat map is a more detailed analysis of diagram (c), indicating the contribution of each DNA shape feature on the 690 datasets, where each column represents a dataset. Those columns were organized by hierarchical clustering based on Pearson correlation and complete linkage. The structural class of ChIP-ed TF in each dataset was showcased at the bottom. (e) A performance comparison between sequence and the combination of sequence and shape (Sequence + DNA Shape) against structural classes in terms of AUC. The red two red boxes indicate the classes with the most significant AUC improvement by combining Sequence and Shape compared to Sequence only.

## 4.4 Conclusion

DESSO improved the state-of-the-art performance of *cis*-regulatory motif prediction and TFBSs identification and showcased the potential of a DL framework for identification and rationalization of results. Results demonstrate that DESSO was able to identify a number of previously unidentified DNA motifs and shape factors that contribute to TF-DNA binding mechanisms and can infer the indirect regulation mechanisms through tethering binding activities and co-factor motifs predictions.

# REFERENCES

[1] M. K. Das and H.-K. Dai, “A survey of DNA motif finding algorithms,” *BMC Bioinformatics*, vol. 8, no. Suppl 7, p. S21, Nov. 2007.

[2] P. D’haeseleer, *What are DNA sequence motifs?*, vol. 24. 2006.

[3] M. Geertz and S. J. Maerkl, “Experimental strategies for studying transcription factor–DNA binding specificities,” *Brief Funct Genomics*, vol. 9, no. 5–6, pp. 362–373, Dec. 2010.

[4] B. Lemon and R. Tjian, “Orchestrated response: a symphony of transcription factors for gene control,” *Genes Dev.*, vol. 14, no. 20, pp. 2551–2569, Oct. 2000.

[5] M. Levine and R. Tjian, “Transcription regulation and animal diversity,” *Nature*, vol. 424, no. 6945, pp. 147–151, Jul. 2003.

[6] G. Pavesi, G. Mauri, and G. Pesole, “In silico representation and discovery of transcription factor binding sites,” *Brief Bioinform*, vol. 5, no. 3, pp. 217–236, Sep. 2004.

[7] K. Danna and D. Nathans, “Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae,” *PNAS*, vol. 68, no. 12, pp. 2913–2917, Dec. 1971.

[8] E. Ford and H. W. Boyer, “Degradation of enteric bacterial deoxyribonucleic acid by the Escherichia coli B restriction endonuclease,” *J. Bacteriol.*, vol. 104, no. 1, pp. 594–595, Oct. 1970.

[9] D. Roulland-Dussoix and H. W. Boyer, “The Escherichia coli B restriction endonuclease,” *Biochim. Biophys. Acta*, vol. 195, no. 1, pp. 219–229, Nov. 1969.

[10] R. Yuan and M. Meselson, “A Specific Complex between a Restriction Endonuclease and Its DNA Substrate,” *PNAS*, vol. 65, no. 2, pp. 357–362, Feb. 1970.

[11] K. Willwand, E. Mumtsidu, G. Kuntz-Simon, and J. Rommelaere, “Initiation of DNA Replication at Palindromic Telomeres Is Mediated by a Duplex-to-Hairpin Transition

Induced by the Minute Virus of Mice Nonstructural Protein NS1," *J. Biol. Chem.*, vol. 273, no. 2, pp. 1165–1174, Sep. 1998.

[12] W.-M. Chu, R. E. Ballard, and C. W. Schmid, "Palindromic sequences preceding the terminator increase polymerase III template activity," *Nucleic Acids Res*, vol. 25, no. 11, pp. 2077–2082, Jun. 1997.

[13] A. Sheari *et al.*, "A tale of two symmetrical tails: Structural and functional characteristics of palindromes in proteins," *BMC Bioinformatics*, vol. 9, no. 1, p. 274, Jun. 2008.

[14] H. Li, V. Rhodius, C. Gross, and E. D. Siggia, "Identification of the binding sites of regulatory proteins in bacterial genomes," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 18, pp. 11772–11777, Sep. 2002.

[15] J. van Helden, Alma. F. Rios, and J. Collado-Vides, "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads," *Nucleic Acids Res*, vol. 28, no. 8, pp. 1808–1818, Apr. 2000.

[16] S. Inukai, K. H. Kock, and M. L. Bulyk, "Transcription factor–DNA binding: beyond binding site motifs," *Curr Opin Genet Dev*, vol. 43, pp. 110–119, Apr. 2017.

[17] D. Pribnow, "Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter.," *Proc Natl Acad Sci U S A*, vol. 72, no. 3, pp. 784–788, Mar. 1975.

[18] A. Cornish-Bowden, "Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.," *Nucleic Acids Res*, vol. 13, no. 9, pp. 3021–3030, May 1985.

[19] A. D. Johnson, "An extended IUPAC nomenclature code for polymorphic nucleic acids," *Bioinformatics*, vol. 26, no. 10, pp. 1386–1389, May 2010.

[20] T. D. Schneider, "Consensus Sequence Zen," *Appl Bioinformatics*, vol. 1, no. 3, pp. 111–119, 2002.

[21] G. D. Stormo, T. D. Schneider, and L. M. Gold, “Characterization of translational initiation sites in E. coli.,” *Nucleic Acids Res*, vol. 10, no. 9, pp. 2971–2996, May 1982.

[22] G. D. Stormo and G. W. Hartzell, “Identifying protein-binding sites from unaligned DNA fragments,” *PNAS*, vol. 86, no. 4, pp. 1183–1187, Feb. 1989.

[23] G. Z. Hertz, G. W. Hartzell, and G. D. Stormo, “Identification of consensus patterns in unaligned DNA sequences known to be functionally related,” *Bioinformatics*, vol. 6, no. 2, pp. 81–92, Apr. 1990.

[24] K. Nishida, M. C. Frith, and K. Nakai, “Pseudocounts for transcription factor binding sites,” *Nucleic Acids Res*, vol. 37, no. 3, pp. 939–944, Feb. 2009.

[25] T. D. Schneider, “Information content of individual genetic sequences,” *J. Theor. Biol.*, vol. 189, no. 4, pp. 427–441, Dec. 1997.

[26] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, “Information content of binding sites on nucleotide sequences,” *Journal of Molecular Biology*, vol. 188, no. 3, pp. 415–431, Apr. 1986.

[27] G. D. Stormo, “Information Content and Free Energy in DNA–Protein Interactions,” *Journal of Theoretical Biology*, vol. 195, no. 1, pp. 135–137, Nov. 1998.

[28] R. Siddharthan, “Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix,” *PLOS ONE*, vol. 5, no. 3, p. e9722, Mar. 2010.

[29] L. B. Heilprin, “Information Theory and Statistics. Solomon Kullback. Wiley, New York; Chapman and Hall, London, 1959. xvii + 395 pp. Illus. $12.50,” *Science*, vol. 131, no. 3404, pp. 917–918, Mar. 1960.

[30] T. D. Schneider and R. M. Stephens, “Sequence logos: a new way to display consensus sequences.,” *Nucleic Acids Res*, vol. 18, no. 20, pp. 6097–6100, Oct. 1990.

[31] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, “WebLogo: a sequence logo generator,” *Genome Res.*, vol. 14, no. 6, pp. 1188–1190, Jun. 2004.

[32] M. L. Bulyk, P. L. F. Johnson, and G. M. Church, “Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors,” *Nucleic Acids Res*, vol. 30, no. 5, pp. 1255–1261, Mar. 2002.

[33] B. Liu, J. Yang, Y. Li, A. McDermaid, and Q. Ma, “An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data,” *Brief. Bioinformatics*, vol. 19, no. 5, pp. 1069–1081, 28 2018.

[34] B. Georgi and A. Schliep, “Context-specific independence mixture modeling for positional weight matrices,” *Bioinformatics*, vol. 22, no. 14, pp. e166-173, Jul. 2006.

[35] S. Hannenhalli and L.-S. Wang, “Enhanced position weight matrices using mixture models,” *Bioinformatics (Oxford, England)*, vol. 21 Suppl 1, pp. i204-12, Jul. 2005.

[36] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan, “Modeling Dependencies in protein-DNA Binding Sites,” in *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, New York, NY, USA, 2003, pp. 28–37.

[37] R. A. Salama and D. J. Stekel, “Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction,” *Nucleic Acids Res.*, vol. 38, no. 12, p. e135, Jul. 2010.

[38] V. D. Marinescu, I. S. Kohane, and A. Riva, “MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes,” *BMC Bioinformatics*, vol. 6, p. 79, Mar. 2005.

[39] P. Mehta, D. J. Schwab, and A. M. Sengupta, “Statistical Mechanics of Transcription-Factor Binding Site Discovery Using Hidden Markov Models,” *J Stat Phys*, vol. 142, no. 6, pp. 1187–1205, Apr. 2011.

[40] Y. Bi, H. Kim, R. Gupta, and R. V. Davuluri, “Tree-Based Position Weight Matrix Approach to Model Transcription Factor Binding Site Profiles,” *PLOS ONE*, vol. 6, no. 9, p. e24210, Sep. 2011.

[41] F. Chin and H. C. M. Leung, “DNA Motif Representation with Nucleotide Dependency,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 1, pp. 110–119, Jan. 2008.

[42] E. Sharon, S. Lubliner, and E. Segal, “A Feature-Based Approach to Modeling Protein–DNA Interactions,” *PLOS Computational Biology*, vol. 4, no. 8, p. e1000154, Aug. 2008.

[43] C. Wang, J. Xie, and B. A. Craig, “Context dependent models for discovery of transcription factor binding sites,” *Statistical Methodology*, vol. 3, no. 1, pp. 55–68, Jan. 2006.

[44] “Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences | Nucleic Acids Research | Oxford Academic.” [Online]. Available: https://academic.oup.com/nar/article/44/13/6055/2457621. [Accessed: 09-Jul-2019].

[45] P. V. Benos, M. L. Bulyk, and G. D. Stormo, “Additivity in protein–DNA interactions: how good an approximation is it?,” *Nucleic Acids Res*, vol. 30, no. 20, pp. 4442–4451, Oct. 2002.

[46] T.-K. Man and G. D. Stormo, “Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay,” *Nucleic Acids Res*, vol. 29, no. 12, pp. 2471–2478, Jun. 2001.

[47] “ChIP-nexus enables improved detection of in vivo transcription factor binding footprints | Nature Biotechnology.” [Online]. Available: https://www.nature.com/articles/nbt.3121. [Accessed: 09-Jul-2019].

[48] H. S. Rhee and B. F. Pugh, “ChIP-exo: A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy,” *Curr Protoc Mol Biol*, vol. 0 21, Oct. 2012.

[49] P. Collas and J. A. Dahl, “Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation,” *Front. Biosci.*, vol. 13, pp. 929–943, Jan. 2008.

[50] P. J. Park, “ChIP-seq: advantages and challenges of a maturing technology,” *Nat. Rev. Genet.*, vol. 10, no. 10, pp. 669–680, Oct. 2009.

[51] L. Song and G. E. Crawford, “DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells,” *Cold Spring Harb Protoc*, vol. 2010, no. 2, p. pdb.prot5384, Feb. 2010.

[52] A. M. Tsankov *et al.*, “Transcription factor binding dynamics during human ES cell differentiation,” *Nature*, vol. 518, no. 7539, pp. 344–349, Feb. 2015.

[53] F. Wu, B. G. Olson, and J. Yao, “DamID-seq: Genome-wide Mapping of Protein-DNA Interactions by High Throughput Sequencing of Adenine-methylated DNA Fragments,” *J Vis Exp*, no. 107, p. e53620, Jan. 2016.

[54] M. Maragkakis, P. Alexiou, T. Nakaya, and Z. Mourelatos, “CLIPSeqTools--a novel bioinformatics CLIP-seq analysis suite,” *RNA*, vol. 22, no. 1, pp. 1–9, Jan. 2016.

[55] M. Hafner *et al.*, “PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins,” *J Vis Exp*, no. 41, Jul. 2010.

[56] N. T. Ingolia, “Ribosome profiling: new views of translation, from single codons to genome scale,” *Nat. Rev. Genet.*, vol. 15, no. 3, pp. 205–213, 2014.

[57] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb, “FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin,” *Genome Res.*, vol. 17, no. 6, pp. 877–885, Jun. 2007.

[58] R. Nutiu *et al.*, “Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument,” *Nat. Biotechnol.*, vol. 29, no. 7, pp. 659–664, Jun. 2011.

[59] O. Aparicio, J. V. Geisberg, and K. Struhl, “Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo,” *Curr Protoc Cell Biol*, vol. Chapter 17, p. Unit 17.7, Sep. 2004.

[60] “Genome-Wide Mapping of in Vivo Protein-DNA Interactions | Science.” [Online]. Available: https://science.sciencemag.org/content/316/5830/1497. [Accessed: 09-Jul-2019].

[61] G. Robertson *et al.*, “Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing,” *Nature Methods*, vol. 4, no. 8, pp. 651–657, Aug. 2007.

[62] T. Suganuma and J. L. Workman, “Histone modification as a reflection of metabolism,” *Cell Cycle*, vol. 15, no. 4, pp. 481–482, 2016.

[63] H. Qu and X. Fang, “A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project,” *Genomics Proteomics Bioinformatics*, vol. 11, no. 3, pp. 135–141, Jun. 2013.

[64] S. Pepke, B. Wold, and A. Mortazavi, “Computation for ChIP-seq and RNA-seq studies,” *Nature Methods*, vol. 6, no. 11s, pp. S22–S32, Oct. 2009.

[65] C. A. Davis *et al.*, “The Encyclopedia of DNA elements (ENCODE): data portal update,” *Nucleic Acids Res*, vol. 46, no. Database issue, pp. D794–D801, Jan. 2018.

[66] “An Integrated Encyclopedia of DNA Elements in the Human Genome,” *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[67] “ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data | EMBO reports.” [Online]. Available: https://www.embopress.org/lookup/doi/10.15252/embr.201846255. [Accessed: 11-Jul-2019].

[68] “GTRD: a database on gene transcription regulation—2019 update | Nucleic Acids Research | Oxford Academic.” [Online]. Available: https://academic.oup.com/nar/article/47/D1/D100/5184717. [Accessed: 11-Jul-2019].

[69] H. S. Rhee and B. F. Pugh, “Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution,” *Cell*, vol. 147, no. 6, pp. 1408–1419, Dec. 2011.

[70] A. A. Perreault and B. J. Venters, “The ChIP-exo Method: Identifying Protein-DNA Interactions with Near Base Pair Precision,” *J Vis Exp*, no. 118, 23 2016.

[71] J. Gutin *et al.*, “Fine-Resolution Mapping of TF Binding and Chromatin Interactions,” *Cell Rep*, vol. 22, no. 10, pp. 2797–2807, Mar. 2018.

[72] L. A. McCue, W. Thompson, C. S. Carmack, and C. E. Lawrence, “Factors influencing the identification of transcription factor binding sites by cross-species comparison,” *Genome Res.*, vol. 12, no. 10, pp. 1523–1532, Oct. 2002.

[73] M. Tompa *et al.*, “Assessing computational tools for the discovery of transcription factor binding sites,” *Nature Biotechnology*, vol. 23, no. 1, pp. 137–144, Jan. 2005.

[74] G. D. Stormo and Y. Zhao, “Determining the specificity of protein–DNA interactions,” *Nature Reviews Genetics*, vol. 11, no. 11, pp. 751–760, Nov. 2010.

[75] F. Zambelli, G. Pesole, and G. Pavesi, “Motif discovery and transcription factor binding sites before and after the next-generation sequencing era,” *Brief Bioinform*, vol. 14, no. 2, pp. 225–237, Mar. 2013.

[76] T. L. Bailey, “DREME: motif discovery in transcription factor ChIP-seq data,” *Bioinformatics*, vol. 27, no. 12, pp. 1653–1659, Jun. 2011.

[77] “A New Exhaustive Method and Strategy for Finding Motifs in ChIP-Enriched Regions.” [Online]. Available:

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0086044. [Accessed: 09-Jul-2019].

[78] M. Thomas-Chollier, C. Herrmann, M. Defrance, O. Sand, D. Thieffry, and J. van Helden, "RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets," *Nucleic Acids Res.*, vol. 40, no. 4, p. e31, Feb. 2012.

[79] N. T. T. Nguyen *et al.*, "RSAT 2018: regulatory sequence analysis tools 20th anniversary," *Nucleic Acids Res*, vol. 46, no. W1, pp. W209–W214, Jul. 2018.

[80] A. A. Sharov and M. Ko, "Exhaustive search for over-represented DNA sequence motifs with cisfinder," *DNA research : an international journal for rapid publication of reports on genes and genomes*, vol. 16, no. 5, pp. 261–273, Oct. 2009.

[81] J. Ding, H. Hu, and X. Li, "SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data," *Nucleic Acids Res*, vol. 42, no. 5, p. e35, Mar. 2014.

[82] J. Maaskola and N. Rajewsky, "Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models," *Nucleic Acids Res*, vol. 42, no. 21, pp. 12995–13011, Dec. 2014.

[83] G. Li, B. Liu, Q. Ma, and Y. Xu, "A new framework for identifying cis-regulatory motifs in prokaryotes," *Nucleic Acids Research*, vol. 39, no. 7, pp. e42–e42, Apr. 2011.

[84] J. R. Sadler, M. S. Waterman, and T. F. Smith, "Regulatory pattern identification in nucleic acid sequences," *Nucleic Acids Res.*, vol. 11, no. 7, pp. 2221–2231, Apr. 1983.

[85] M. S. Waterman, R. Arratia, and D. J. Galas, "Pattern recognition in several sequences: consensus and alignment," *Bull. Math. Biol.*, vol. 46, no. 4, pp. 515–527, 1984.

[86] D. J. Galas, M. Eggert, and M. S. Waterman, "Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from Escherichia coli," *J. Mol. Biol.*, vol. 186, no. 1, pp. 117–128, Nov. 1985.

[87] "Quantifying similarity between motifs | Genome Biology | Full Text." [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-2-r24. [Accessed: 10-Jul-2019].

[88] Z. Zhou *et al.*, "Mutation-profile-based methods for understanding selection forces in cancer somatic mutations: a comparative analysis," *Oncotarget*, vol. 8, no. 35, pp. 58835–58846, Aug. 2017.

[89] N. E. Wheeler, L. Barquist, R. A. Kingsley, and P. P. Gardner, "A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes," *Bioinformatics*, vol. 32, no. 23, pp. 3566–3574, 01 2016.

[90] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," *Proc Int Conf Intell Syst Mol Biol*, vol. 2, pp. 28–36, 1994.

[91] J. van Helden, B. André, and J. Collado-Vides, "A web site for the computational analysis of yeast regulatory sequences," *Yeast*, vol. 16, no. 2, pp. 177–187, 2000.

[92] J. van Helden, B. André, and J. Collado-Vides, "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies," *J. Mol. Biol.*, vol. 281, no. 5, pp. 827–842, Sep. 1998.

[93] P. Machanick and T. L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets," *Bioinformatics*, vol. 27, no. 12, pp. 1696–1697, Jun. 2011.

[94] S. Heinz *et al.*, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities," *Mol. Cell*, vol. 38, no. 4, pp. 576–589, May 2010.

[95] R. Sundberg, "Maximum likelihood theory for incomplete data from an exponential family," *Scand. J. Statist.*, vol. 1, no. 2, pp. 49–58, 1974.

[96] B. R. Ceppellini, M. Siniscalco, and C. a. B. Smith, “The Estimation of Gene Frequencies in a Random-Mating Population,” *Annals of Human Genetics*, vol. 20, no. 2, pp. 97–115, 1955.

[97] T. L. Bailey and P. Machanick, “Inferring direct DNA binding from ChIP-seq,” *Nucleic Acids Res*, vol. 40, no. 17, p. e128, Sep. 2012.

[98] “Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data | BMC Bioinformatics | Full Text.” [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-165. [Accessed: 10-Jul-2019].

[99] W. Ma, W. S. Noble, and T. L. Bailey, “Motif-based analysis of large nucleotide data sets using MEME-ChIP,” *Nat Protoc*, vol. 9, no. 6, pp. 1428–1450, 2014.

[100] T. Bailey and M. Gribskov, “Combining evidence using p-values: Application to sequence homology searches,” *Bioinformatics (Oxford, England)*, vol. 14, pp. 48–54, Feb. 1998.

[101] F. A. Buske, M. Bodén, D. C. Bauer, and T. L. Bailey, “Assigning roles to DNA regulatory motifs using comparative genomics,” *Bioinformatics*, vol. 26, no. 7, pp. 860–866, Apr. 2010.

[102] N. T. L. Tran and C.-H. Huang, “A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data,” *Biol Direct*, vol. 9, p. 4, Feb. 2014.

[103] S. J. Schultheiss, M.-C. Münch, G. D. Andreeva, and G. Rätsch, “Persistence and Availability of Web Services in Computational Biology,” *PLoS One*, vol. 6, no. 9, Sep. 2011.

[104] Á. Ősz, L. S. Pongor, D. Szirmai, and B. Győrffy, “A snapshot of 3649 Web-based services published between 1994 and 2017 shows a decrease in availability after 2 years,” *Brief Bioinform*, vol. 20, no. 3, pp. 1004–1010, May 2019.

[105] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[106] Y. Chu and D. R. Corey, “RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation,” *Nucleic Acid Ther*, vol. 22, no. 4, pp. 271–274, Aug. 2012.

[107] M. Vera, J. Biswas, A. Senecal, R. H. Singer, and H. Y. Park, “Single-Cell and Single-Molecule Analysis of Gene Expression Regulation,” *Annu. Rev. Genet.*, vol. 50, pp. 267–291, Nov. 2016.

[108] A. Zeisel *et al.*, “Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq,” *Science*, vol. 347, no. 6226, pp. 1138–1142, Mar. 2015.

[109] “Single-cell RNA-sequencing: The future of genome biology is now: RNA Biology: Vol 14, No 5.” [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/15476286.2016.1201618. [Accessed: 10-Jul-2019].

[110] A. M. Klein *et al.*, “Droplet barcoding for single cell transcriptomics applied to embryonic stem cells,” *Cell*, vol. 161, no. 5, pp. 1187–1201, May 2015.

[111] E. Z. Macosko *et al.*, “Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets,” *Cell*, vol. 161, no. 5, pp. 1202–1214, May 2015.

[112] M. D. Luecken and F. J. Theis, “Current best practices in single-cell RNA-seq analysis: a tutorial,” *Molecular Systems Biology*, vol. 15, no. 6, p. e8746, Jun. 2019.

[113] B. Hwang, J. H. Lee, and D. Bang, “Single-cell RNA sequencing technologies and bioinformatics pipelines,” *Experimental & Molecular Medicine*, vol. 50, no. 8, p. 96, Aug. 2018.

[114] “Defining cell types and states with single-cell genomics.” [Online]. Available: https://genome.cshlp.org/content/25/10/1491?top=1. [Accessed: 10-Jul-2019].

[115] C. A. Herring *et al.*, “Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut,” *Cell Syst*, vol. 6, no. 1, pp. 37-51.e9, Jan. 2018.

[116] A. Tanay and A. Regev, “Scaling single-cell genomics from phenomenology to mechanism,” *Nature*, vol. 541, no. 7637, pp. 331–338, 18 2017.

[117] W. K. Maas, “STUDIES ON THE MECHANISM OF REPRESSION OF ARGININE BIOSYNTHESIS IN ESCHERICHIA COLI. II. DOMINANCE OF REPRESSIBILITY IN DIPLOIDS,” *J. Mol. Biol.*, vol. 8, pp. 365–370, Mar. 1964.

[118] R. Bacher and C. Kendziorski, “Design and computational analysis of single-cell RNA-sequencing experiments,” *Genome Biology*, vol. 17, no. 1, p. 63, Apr. 2016.

[119] L. Zappia, B. Phipson, and A. Oshlack, “Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database,” *PLOS Computational Biology*, vol. 14, no. 6, p. e1006245, Jun. 2018.

[120] D. T. Odom, “Identification of Transcription Factor-DNA Interactions In Vivo,” *Subcell. Biochem.*, vol. 52, pp. 175–191, 2011.

[121] D. M. Cohen, H.-W. Lim, K.-J. Won, and D. J. Steger, “Shared nucleotide flanks confer transcriptional competency to bZip core motifs,” *Nucleic Acids Res*, vol. 46, no. 16, pp. 8371–8384, Sep. 2018.

[122] “Role of the chromatin landscape and sequence in determining cell type-specific genomic glucocorticoid receptor binding and gene regulation. - PubMed - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27903902. [Accessed: 11-Jul-2019].

[123] R. Gordân *et al.*, “Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape,” *Cell Rep*, vol. 3, no. 4, pp. 1093–1104, Apr. 2013.

[124] M. J. Rossi, W. K. M. Lai, and B. F. Pugh, “Genome-wide determinants of sequence-specific DNA binding of general regulatory factors,” *Genome Res*, vol. 28, no. 4, pp. 497–508, Apr. 2018.

[125] H. S. Rhee and B. F. Pugh, “ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy,” *Curr Protoc Mol Biol*, vol. Chapter 21, p. Unit 21.24, Oct. 2012.

[126] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, “The MEME Suite,” *Nucleic Acids Res*, vol. 43, no. Web Server issue, pp. W39–W49, Jul. 2015.

[127] A. Kiesel, C. Roth, W. Ge, M. Wess, M. Meier, and J. Söding, “The BaMM web server for de-novo motif discovery and regulatory sequence analysis,” *Nucleic Acids Res*, vol. 46, no. W1, pp. W215–W220, Jul. 2018.

[128] M. Pertea, S. M. Mount, and S. L. Salzberg, “A computational survey of candidate exonic splicing enhancer motifs in the model plant Arabidopsis thaliana,” *BMC Bioinformatics*, vol. 8, no. 1, p. 159, May 2007.

[129] X. S. Liu, D. L. Brutlag, and J. S. Liu, “An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments,” *Nat. Biotechnol.*, vol. 20, no. 8, pp. 835–839, Aug. 2002.

[130] X. Liu, D. L. Brutlag, and J. S. Liu, “BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes,” *Pac Symp Biocomput*, pp. 127–138, 2001.

[131] Y. Liu, X. S. Liu, L. Wei, R. B. Altman, and S. Batzoglou, “Eukaryotic regulatory element conservation analysis and identification using comparative genomics,” *Genome Res.*, vol. 14, no. 3, pp. 451–458, Mar. 2004.

[132] L. T. Dang *et al.*, “TrawlerWeb: an online de novo motif discovery tool for next-generation sequencing datasets,” *BMC Genomics*, vol. 19, no. 1, p. 238, Apr. 2018.

[133] J. Grau, S. Posch, I. Grosse, and J. Keilwagen, “A general approach for discriminative de novo motif discovery from high-throughput data,” *Nucleic Acids Res*, vol. 41, no. 21, p. e197, Nov. 2013.

[134] M. J. Rossi, W. K. M. Lai, and B. F. Pugh, “Simplified ChIP-exo assays,” *Nat Commun*, vol. 9, no. 1, p. 2842, 20 2018.

[135] L. Wang *et al.*, “MACE: model based analysis of ChIP-exo,” *Nucleic Acids Res.*, vol. 42, no. 20, p. e156, Nov. 2014.

[136] H. Salgado *et al.*, “Using RegulonDB, the Escherichia coli K-12 gene regulatory transcriptional network database,” *Curr Protoc Bioinformatics*, vol. 61, no. 1, pp. 1.32.1-1.32.30, Mar. 2018.

[137] “An Integrated Pipeline for the Genome-Wide Analysis of Transcription Factor Binding Sites from ChIP-Seq.” [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0016432. [Accessed: 05-Aug-2019].

[138] V. G. Levitsky *et al.*, “Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data,” *BMC Genomics*, vol. 15, p. 80, Jan. 2014.

[139] J. Hu, B. Li, and D. Kihara, “Limitations and potentials of current motif discovery algorithms,” *Nucleic Acids Res*, vol. 33, no. 15, pp. 4899–4913, 2005.

[140] “SCENIC: Single-cell regulatory network inference and clustering.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5937676/. [Accessed: 10-Jul-2019].

[141] S. Suo, Q. Zhu, A. Saadatpour, L. Fei, G. Guo, and G.-C. Yuan, “Revealing the Critical Regulators of Cell Identity in the Mouse Cell Atlas,” *Cell Reports*, vol. 25, no. 6, pp. 1436-1445.e3, Nov. 2018.

[142] H. Matsumoto *et al.*, “SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation,” *Bioinformatics*, vol. 33, no. 15, pp. 2314–2321, Aug. 2017.

[143] T. E. Chan, M. P. H. Stumpf, and A. C. Babtie, “Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures,” *Cell Syst*, vol. 5, no. 3, pp. 251-267.e3, 27 2017.

[144] S. Chen and J. C. Mar, “Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data,” *BMC Bioinformatics*, vol. 19, no. 1, p. 232, Jun. 2018.

[145] B. Braschi *et al.*, “Genenames.org: the HGNC and VGNC resources in 2019,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D786–D792, Jan. 2019.

[146] F. Cunningham *et al.*, “Ensembl 2019,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D745–D751, Jan. 2019.

[147] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, “Integrating single-cell transcriptomic data across different conditions, technologies, and species,” *Nature Biotechnology*, vol. 36, no. 5, pp. 411–420, May 2018.

[148] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data,” *Cell*, vol. 177, no. 7, pp. 1888-1902.e21, Jun. 2019.

[149] V. Y. Kiselev *et al.*, “SC3: consensus clustering of single-cell RNA-seq data,” *Nat. Methods*, vol. 14, no. 5, pp. 483–486, May 2017.

[150] G. Li, Q. Ma, H. Tang, A. H. Paterson, and Y. Xu, "QUBIC: a qualitative biclustering algorithm for analyses of gene expression data," *Nucleic Acids Res.*, vol. 37, no. 15, p. e101, Aug. 2009.

[151] Y. Zhang, J. Xie, J. Yang, A. Fennell, C. Zhang, and Q. Ma, "QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data," *Bioinformatics*, vol. 33, no. 3, pp. 450–452, 01 2017.

[152] Q. Ma *et al.*, "DMINDA: an integrated web server for DNA motif identification and analyses," *Nucleic Acids Res*, vol. 42, no. Web Server issue, pp. W12–W19, Jul. 2014.

[153] J. Yang, X. Chen, A. McDermaid, and Q. Ma, "DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses," *Bioinformatics*, vol. 33, no. 16, pp. 2586–2588, Aug. 2017.

[154] E. Y. Chen *et al.*, "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, p. 128, Apr. 2013.

[155] M. V. Kuleshov *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Res*, vol. 44, no. Web Server issue, pp. W90–W97, Jul. 2016.

[156] N. F. Fernandez *et al.*, "Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data," *Scientific Data*, vol. 4, p. 170151, Oct. 2017.

[157] G. Stelzer *et al.*, "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses," *Curr Protoc Bioinformatics*, vol. 54, pp. 1.30.1-1.30.33, 20 2016.

[158] C. Soneson and M. D. Robinson, "Bias, robustness and scalability in single-cell differential expression analysis," *Nat. Methods*, vol. 15, no. 4, pp. 255–261, 2018.

[159] L. van der Maaten and G. Hinton, *Visualizing Data using t-SNE*. 2008.

[160] E. Becht *et al.*, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnology*, vol. 37, no. 1, pp. 38–44, Jan. 2019.

[161] K. Eren, M. Deveci, O. Küçüktunç, and Ü. V. Çatalyürek, “A comparative analysis of biclustering algorithms for gene expression data,” *Brief. Bioinformatics*, vol. 14, no. 3, pp. 279–292, May 2013.

[162] W. Saelens, R. Cannoodt, and Y. Saeys, “A comprehensive evaluation of module detection methods for gene expression data,” *Nature Communications*, vol. 9, no. 1, p. 1090, Mar. 2018.

[163] M. N. Cabili *et al.*, “Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses,” *Genes Dev.*, vol. 25, no. 18, pp. 1915–1927, Sep. 2011.

[164] S. Hänzelmann, R. Castelo, and J. Guinney, “GSVA: gene set variation analysis for microarray and RNA-Seq data,” *BMC Bioinformatics*, vol. 14, no. 1, p. 7, Jan. 2013.

[165] R. Zheng *et al.*, “Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D729–D735, Jan. 2019.

[166] A. Khan *et al.*, “JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework,” *Nucleic Acids Res*, vol. 46, no. D1, pp. D260–D266, Jan. 2018.

[167] I. V. Kulakovskiy *et al.*, “HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis,” *Nucleic Acids Res*, vol. 46, no. Database issue, pp. D252–D259, Jan. 2018.

APPENDIX 1: Curriculum Vitae

# Cankun Wang

Department of Agronomy, Horticulture & Plant Science

South Dakota State University

Cankun.Wang@sdstate.edu | 201-618-9390

## Education

| | |
|---|---|
| 01/2018 – Present | M.S. in Plant Science (3.59 GPA) |
| | South Dakota State University, Brookings, SD, USA |
| 09/2013 – 06/2017 | B.S. in Software Engineering (3.21 GPA) |
| | Beijing Jiaotong University, Beijing, China |

## Employment

### Graduate Research Assistant

01/2018 – Present

Department of Plant Science, South Dakota State University, USA

- Research in the identification of DNA transcription factors motif
- Development of Web server based on cell-type-specific regulon inference from Single-cell RNA-Seq

### Research Assistant

09/2017 – 12/2017

Department of Mathematics, Shandong University, China

- Establishes of a test for motif finding efficiency between our algorithm and other popular tools and explored the optimization as well as the feasibility of further iterations.

- Research in correlations between DNA mapped sequencing data and statistical analysis

**Data Analyst Intern**

02/2014 – 07/2014

Hexin technology, Beijing, China

- Development of the software on automatic generating students' wrong answers collections from the collecting of handing-writing test paper
- Monitor and modify the training datasets on the natural language processing algorithm based on deep learning

## Publications

1. Xia, Ye, Seth DeBolt, Qin Ma, Adam McDermaid, **Cankun Wang**, Nicole Shapiro, Tanja Woyke, and Nikos C. Kyrpides. "Improved Draft Genome Sequence of Bacillus Sp. Strain YF23, Which Has Plant Growth-Promoting Activity." Edited by David Rasko. Microbiology Resource Announcements 8, no. 15 (April 11, 2019). https://doi.org/10.1128/MRA.00099-19.
2. Xia, Ye, Seth DeBolt, Qin Ma, Adam McDermaid, **Cankun Wang**, Nicole Shapiro, Tanja Woyke, and Nikos C. Kyrpides. "Improved Draft Genome Sequence of Pseudomonas Poae A2-S9, a Strain with Plant Growth-Promoting Activity." Edited by Irene L. G. Newton. Microbiology Resource Announcements 8, no. 15 (April 11, 2019). https://doi.org/10.1128/MRA.00275-19.
3. Monier, Brandon, Adam McDermaid, **Cankun Wang**, Jing Zhao, Allison Miller, Anne Fennell, and Qin Ma. "IRIS-EDA: An Integrated RNA-Seq Interpretation System for Gene Expression Data Analysis." PLOS Computational Biology 15, no. 2 (February 14, 2019): e1006792. https://doi.org/10.1371/journal.pcbi.1006792.
4. Wang, Yan, Sen Yang, Jing Zhao, Wei Du, Yanchun Liang, **Cankun Wang**, Fengfeng Zhou, Yuan Tian, and Qin Ma. "Using Machine Learning to Measure Relatedness Between Genes: A Multi-Features Model." Scientific Reports 9, no. 1 (December 2018). https://doi.org/10.1038/s41598-019-40780-7.
5. Han, Siyu, Yanchun Liang, Qin Ma, Yangyi Xu, Yu Zhang, Wei Du, **Cankun Wang,** and Ying Li. "LncFinder: An Integrated Platform for Long Non-Coding RNA Identification Utilizing Sequence

Intrinsic Composition, Structural Information and Physicochemical Property." Briefings in Bioinformatics. Accessed November 24, 2018. https://doi.org/10.1093/bib/bby065.

6. McDermaid, Adam, Xin Chen, Yiran Zhang, **Cankun Wang**, Shaopeng Gu, Juan Xie, and Qin Ma. "A New Machine Learning-Based Framework for Mapping Uncertainty Analysis in RNA-Seq Read Alignment and Gene Expression Estimation." Frontiers in Genetics 9 (2018). https://doi.org/10.3389/fgene.2018.00313.

## Presentations

1. Development of Regulatory Motif Identification program. Apr 23, 2019, BioSNTR Plant Science Research Day. Brookings, SD. (Oral Presentation)
2. Combining Computational Methods and Experimental Data for Motif Prediction. Apr 26, 2018, BioSNTR Plant Science Research Day. Brookings, SD. (Poster Presentation)

## Skills

- Next-generation sequence data analyses
- R, C, Python programming
- Web development
- Linux server maintenance
- Database management
- Mathematical & Statistical modelling

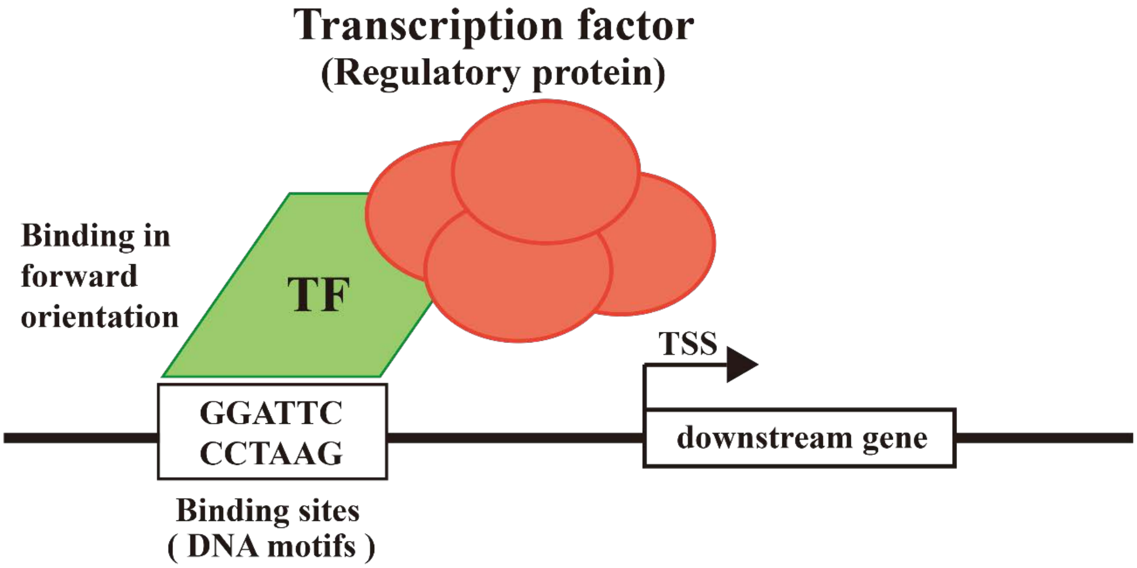
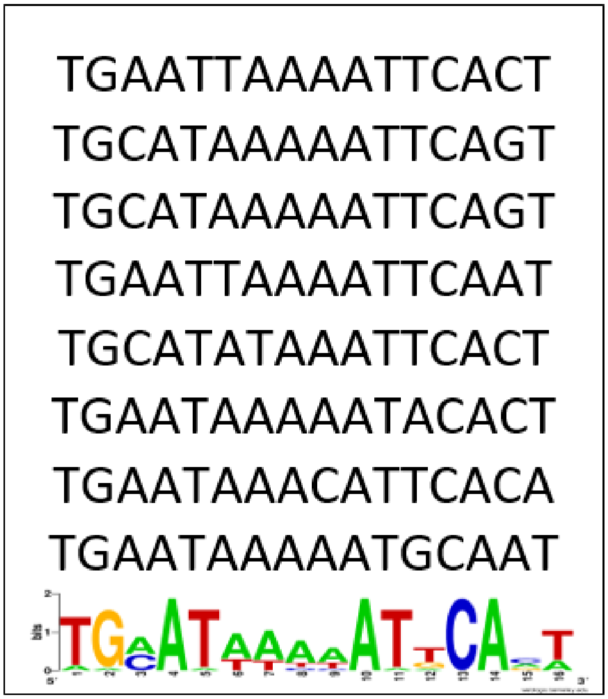


Figure 1. DNA transcription initiation.



TGAATTAAAATTCACT
TGCATAAAAATTCAGT
TGCATAAAAATTCAGT
TGAATTAAAATTCAAT
TGCATATAAATTCACT
TGAATAAAAATACACT
TGAATAAACATTCACA
TGAATAAAAATGCAAT

Figure 2. Example of DNA sequence motif

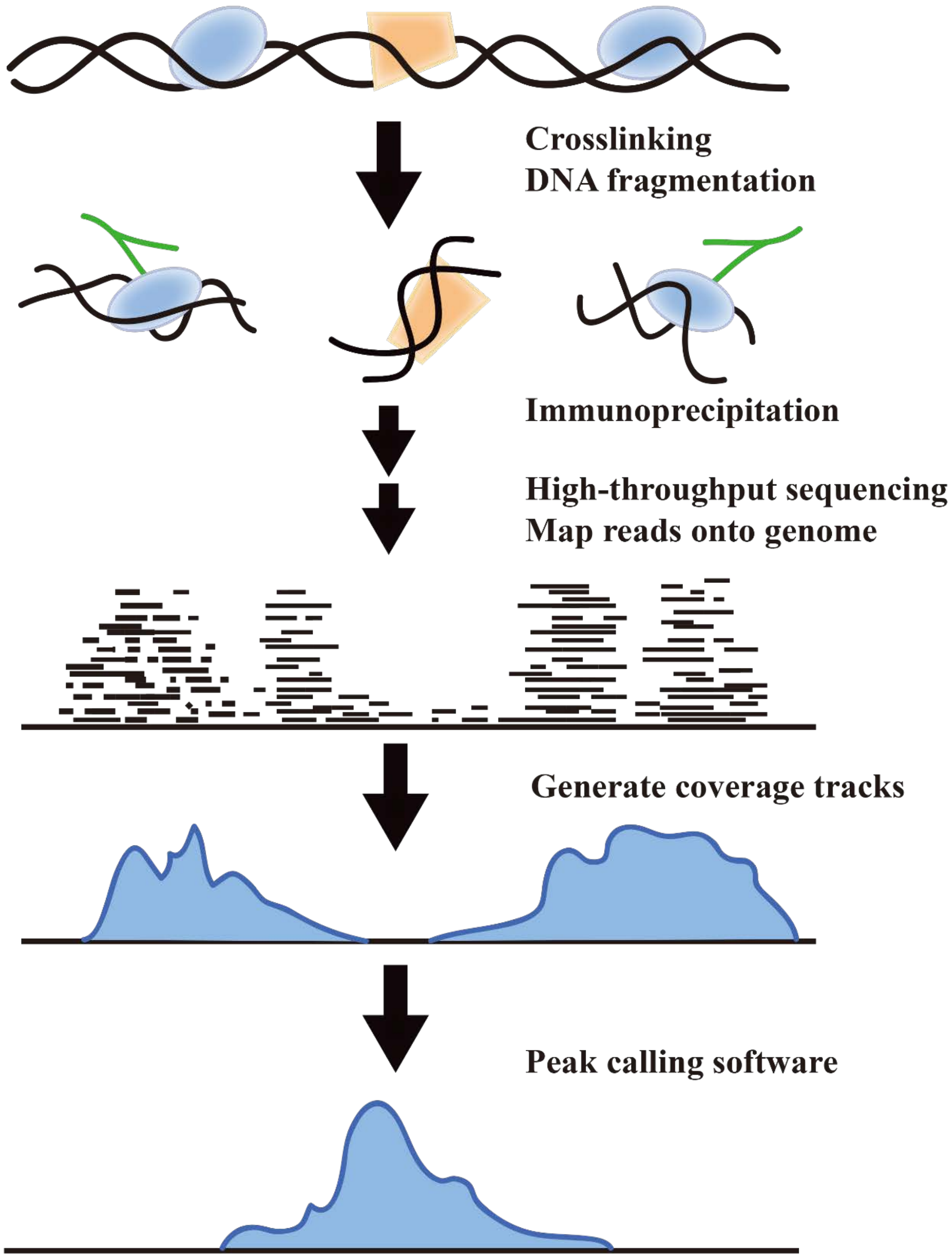


Figure 3. ChIP-seq protocol overview

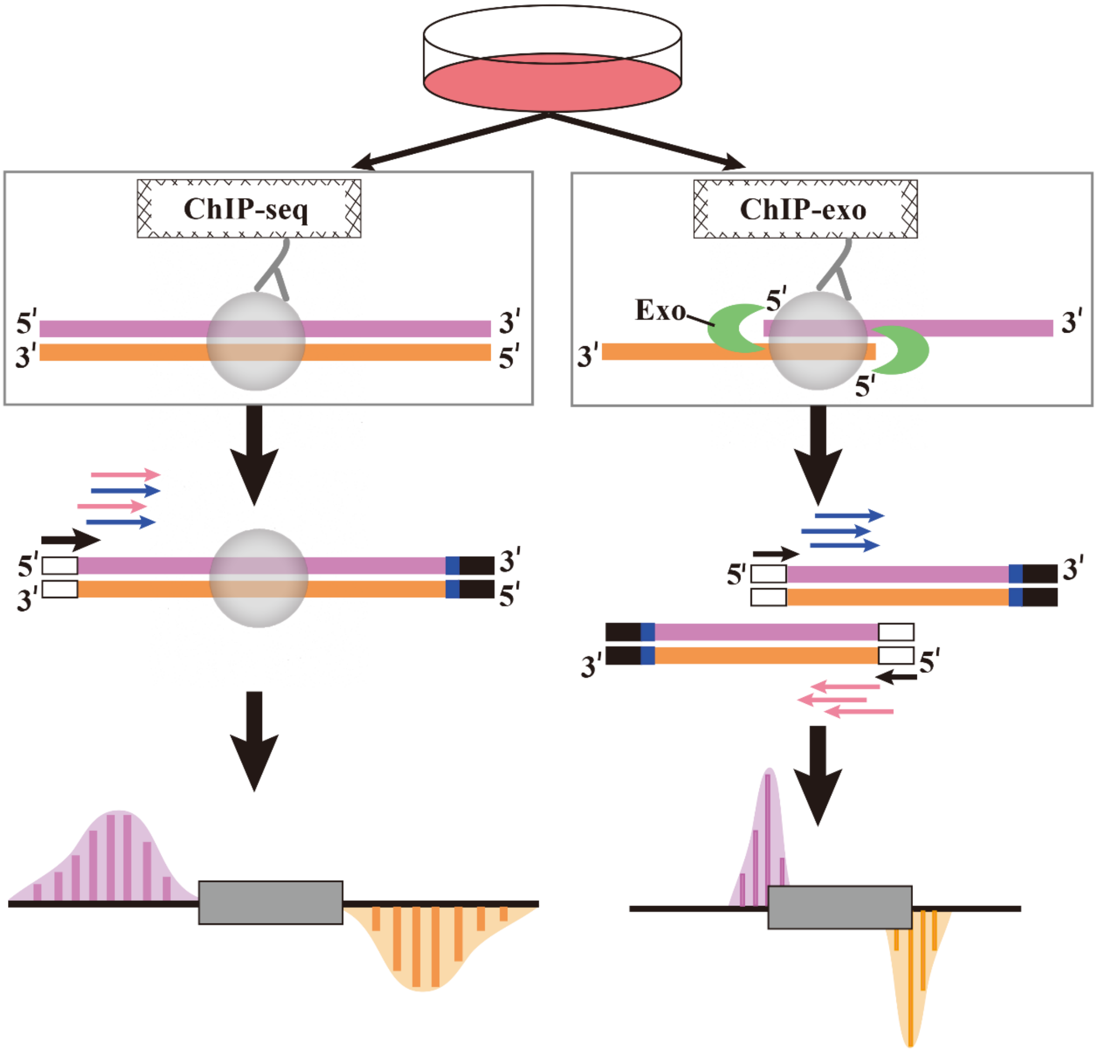


Figure 4. The difference of TF harvested result from ChIP-seq and ChIP-exo

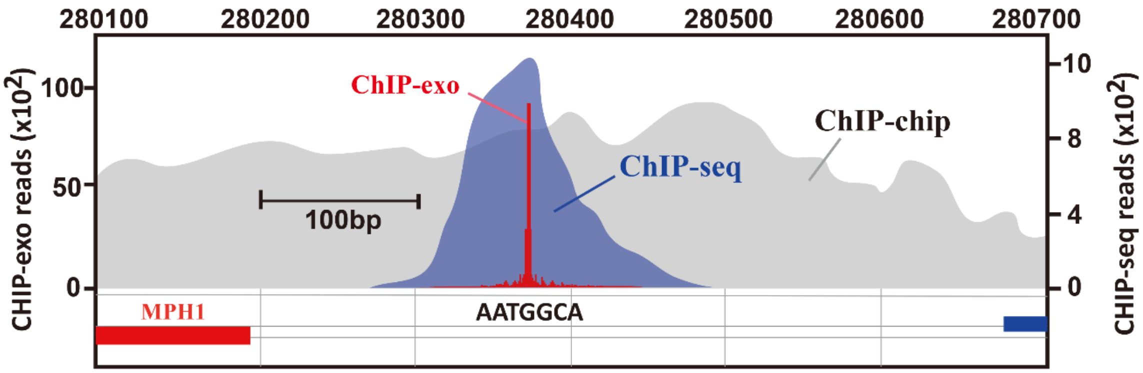


Figure 5. Resolution comparison between ChIP-seq and ChIP-exo

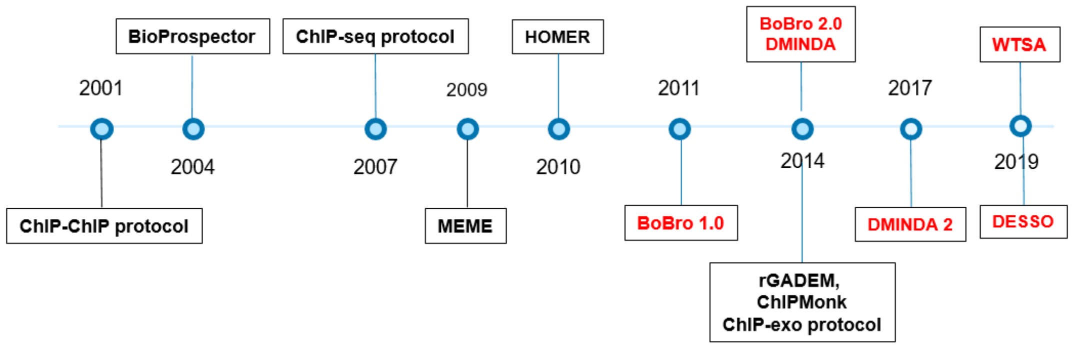


Figure 6. Summary of existing DNA motif identification tools and techniques

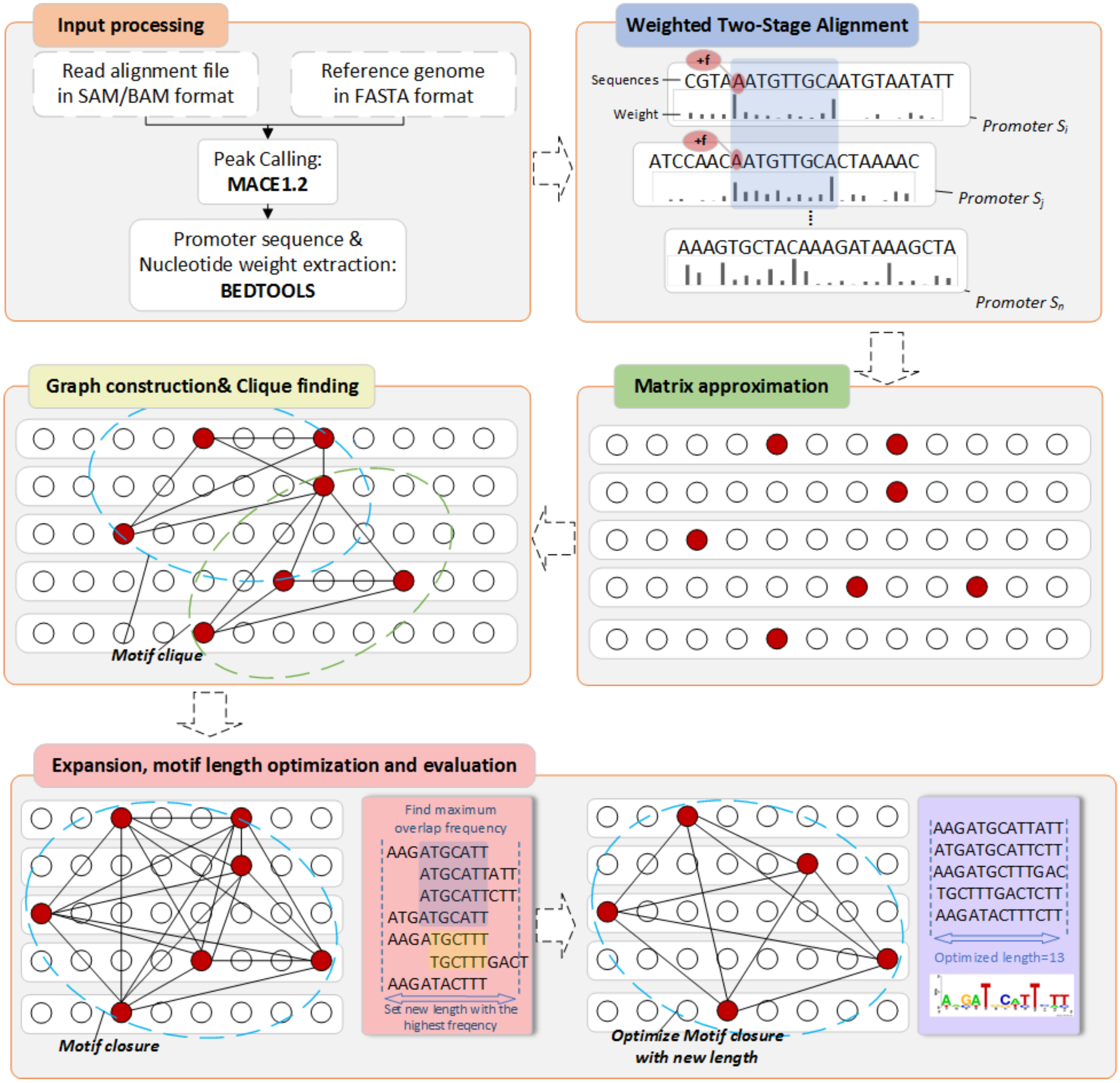


Figure 7. WTSA workflow. This workflow consists of five steps: data pre-processing, weighted two-stage alignment, Matrix approximation, graph construction and clique finding, motif expansion, optimization and evaluation.

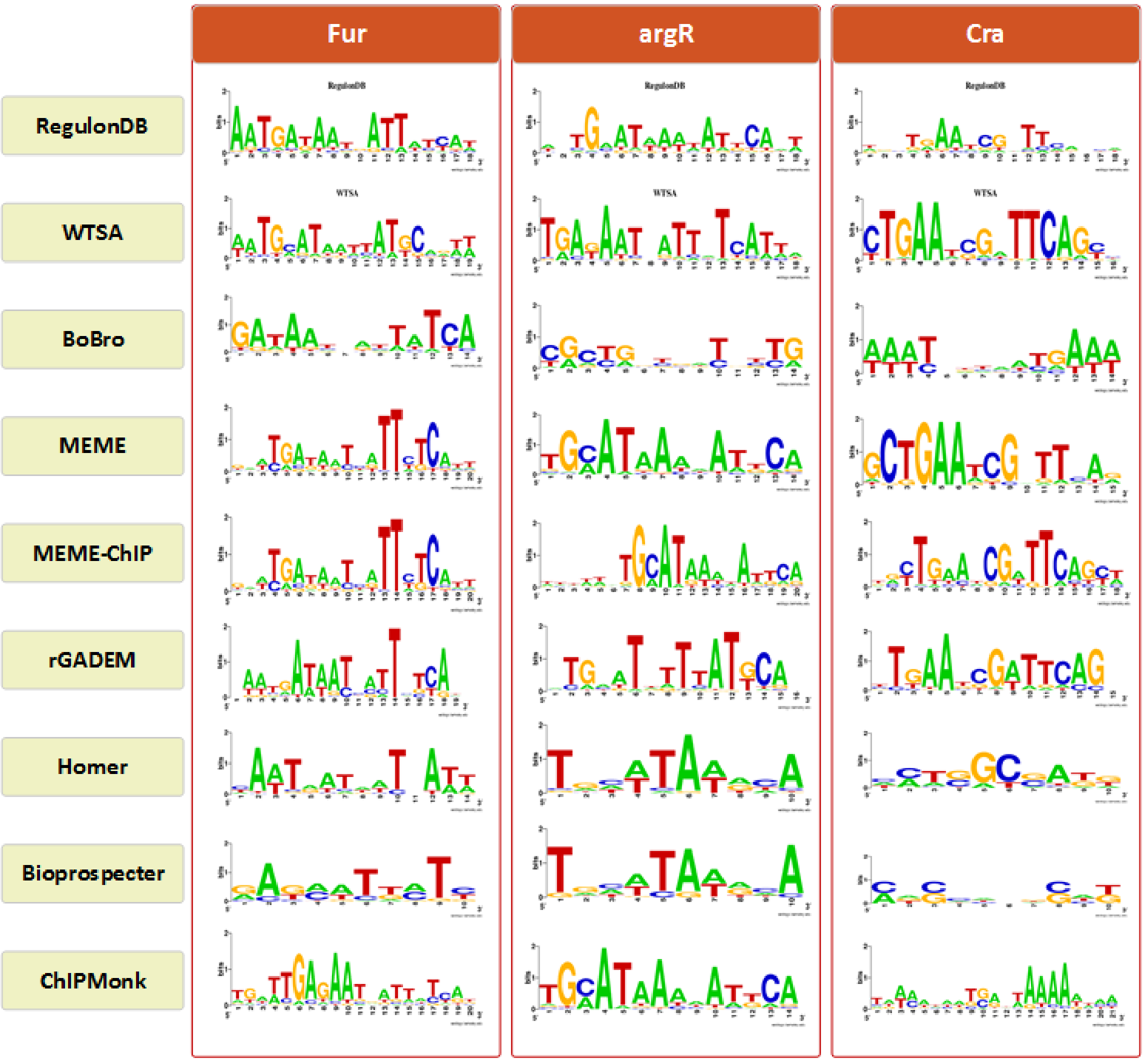


Figure 8. Fur, ArgR, Cra TF motif logo identified by WTSA and other tools, RegulonDB shows the reference ArgR TF motif logo.

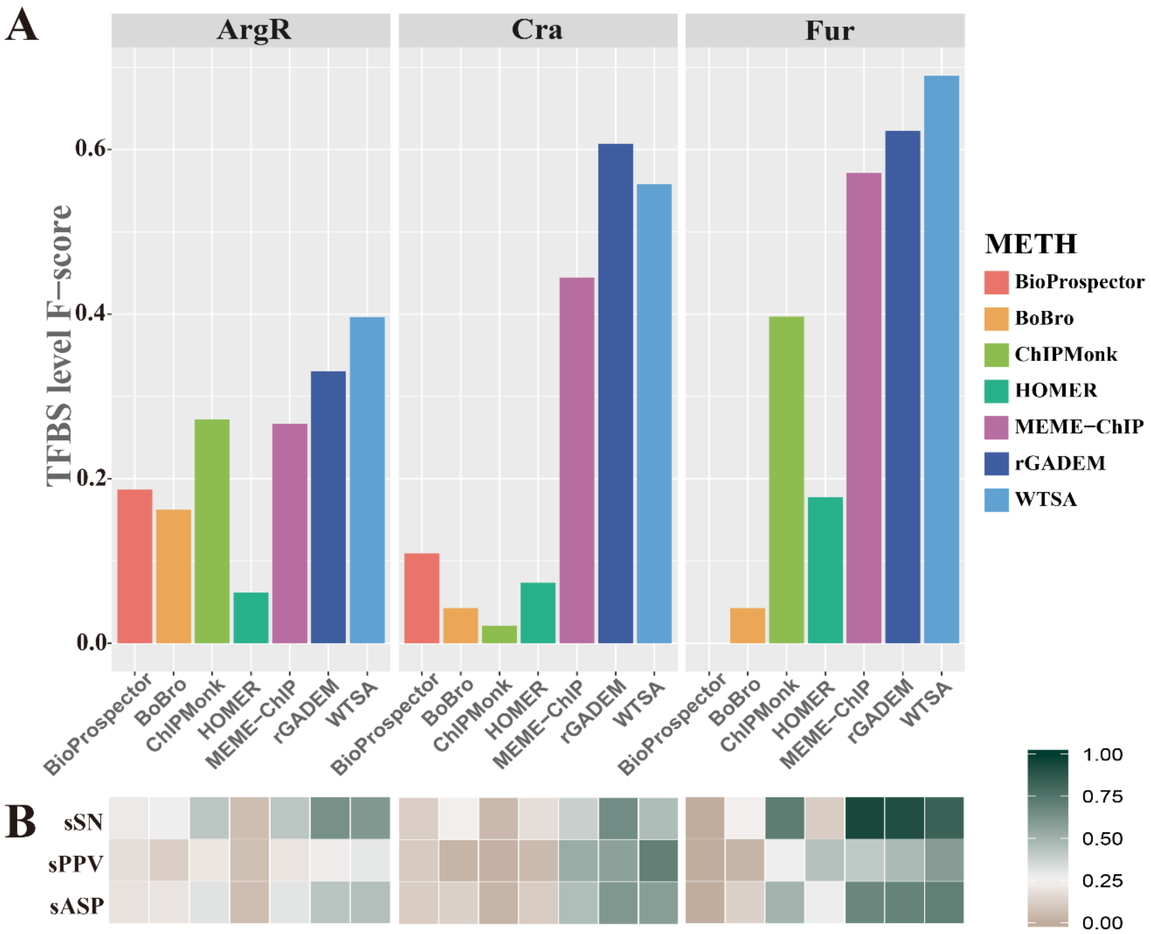


Figure 9. Performance comparison of motif prediction results on TFBS level.

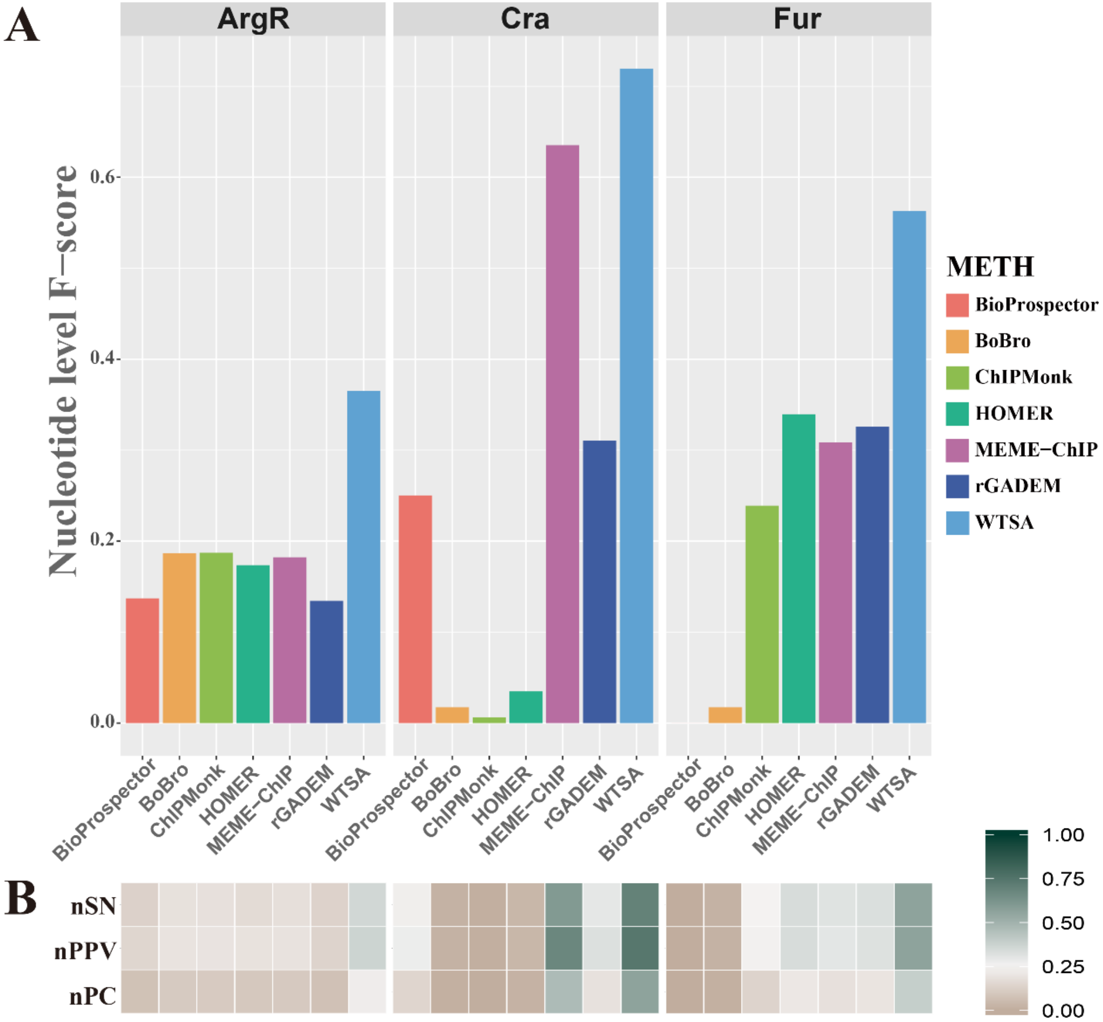


Figure 10. Performance comparison of motif prediction results on nucleotide level.

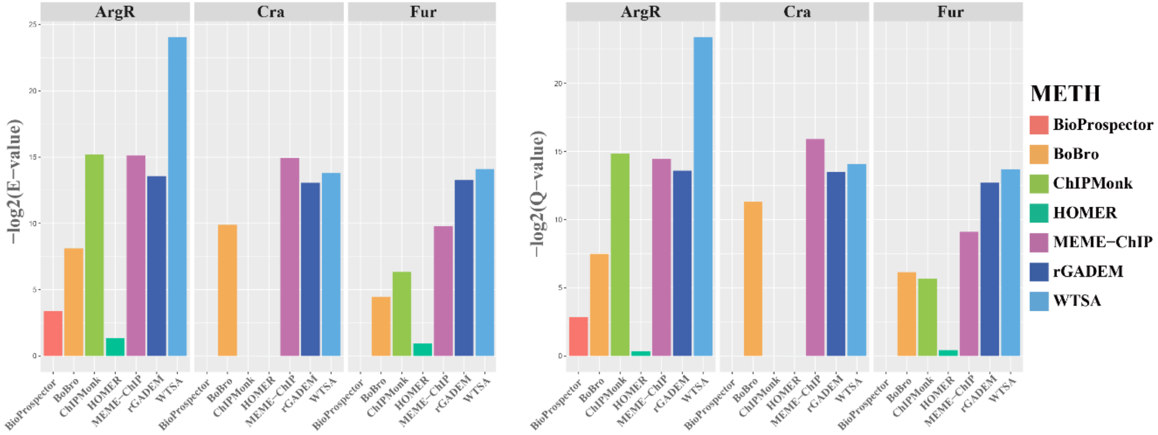


Figure 11. Performance of WTSA, MEME-ChIP, rGADEM, HOMER, Bioprospector, ChIP-monk on TOMTOM profile level comparison.

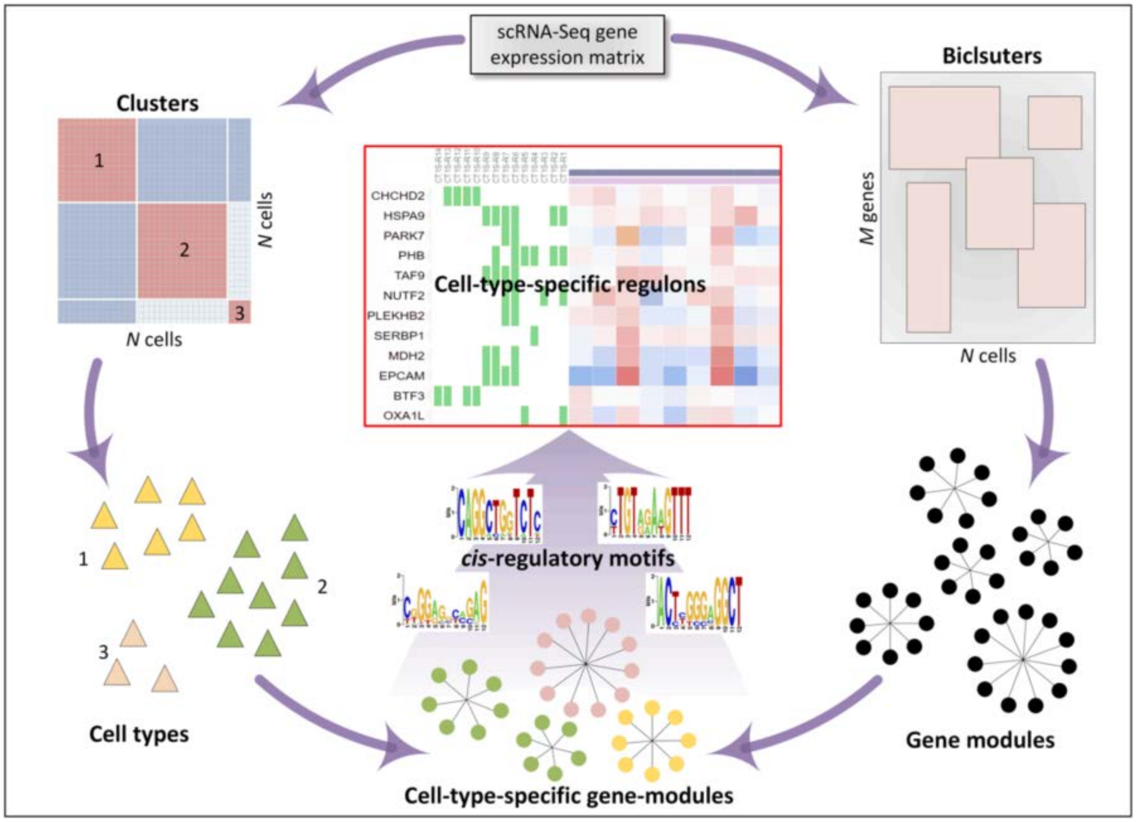


Figure 12. IRIS3 overview.

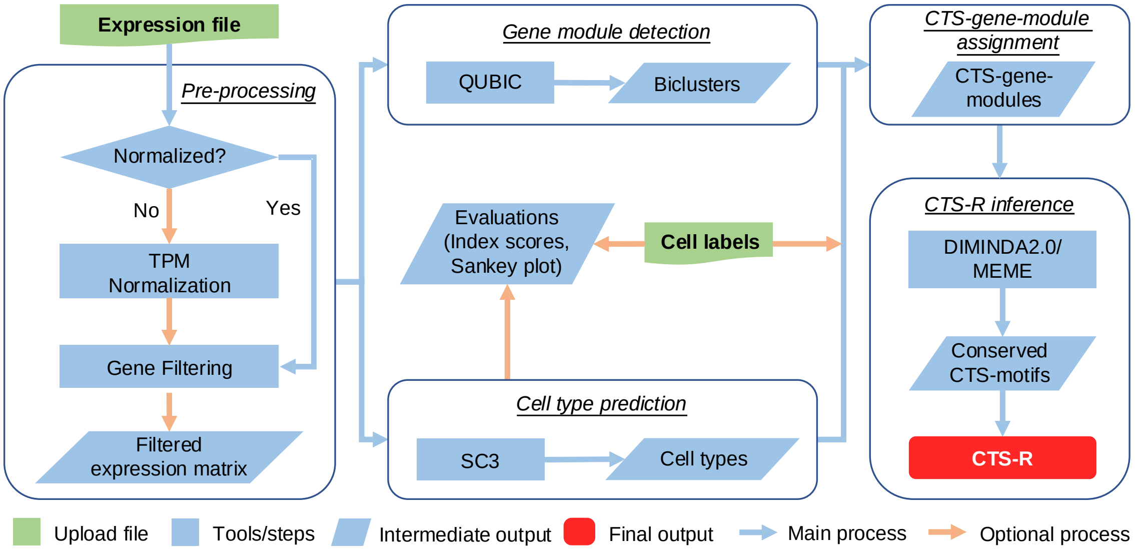


Figure 13. IRIS3 workflow.

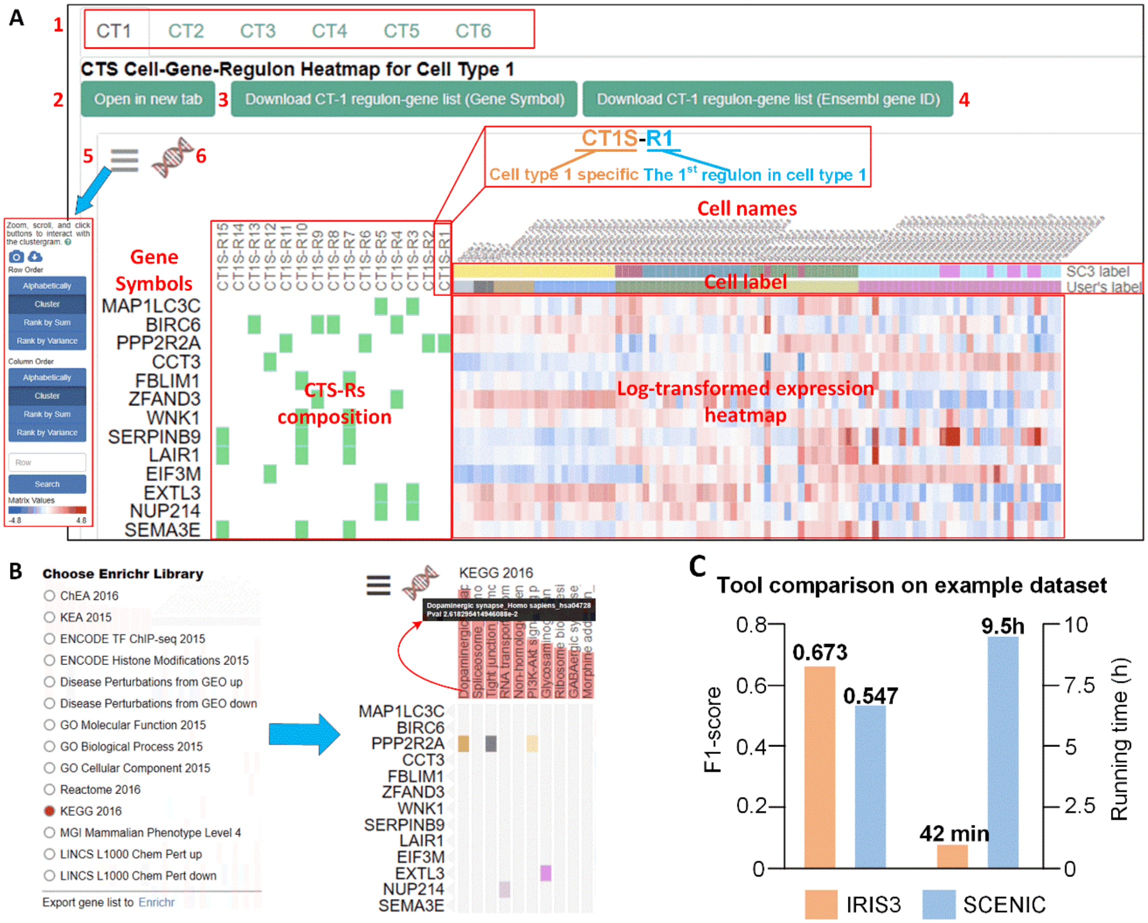
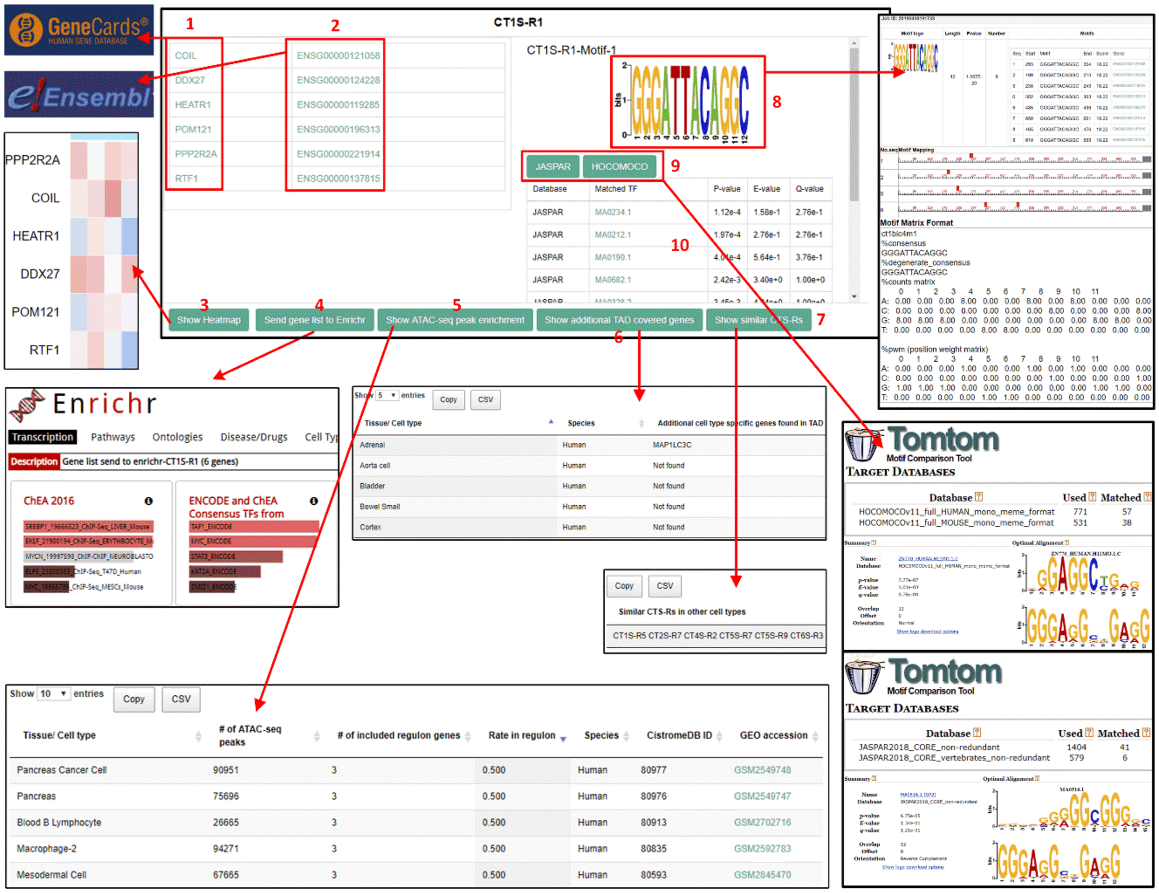


Figure 14. IRIS3 System-level CTS-R inference and performance comparison.



**CT1S-R1**

| | |
|---|---|
| COIL | ENSG00000121058 |
| DDX27 | ENSG00000124228 |
| HEATR1 | ENSG00000119285 |
| POM121 | ENSG00000196313 |
| PPP2R2A | ENSG00000221914 |
| RTF1 | ENSG00000137815 |

CT1S-R1-Motif-1

| Database | Matched TF | P-value | E-value | Q-value |
|---|---|---|---|---|
| JASPAR | MA0234.1 | 1.12e-4 | 1.58e-1 | 2.76e-1 |
| JASPAR | MA0212.1 | 1.97e-4 | 2.76e-1 | 2.76e-1 |
| JASPAR | MA0190.1 | 4.01e-4 | 5.64e-1 | 3.76e-1 |
| JASPAR | MA0682.1 | 2.42e-3 | 3.40e+0 | 1.00e+0 |
| JASPAR | MA0328.2 | 3.45e-3 | 4.84e+0 | 1.00e+0 |

Job ID: 20190408191738

| Motif logo | Length | Pvalue | Number | Motifs |
|---|---|---|---|---|
| GGGATTACAGGC | 12 | 1.067E-29 | 8 | |

| Seq | Start | Motif | End | Score | Gene |
|---|---|---|---|---|---|
| 1 | 293 | GGGATTACAGGC | 304 | 18.22 | [illegible] |
| 2 | 199 | GGGATTACAGGC | 210 | 18.22 | [illegible] |
| 5 | 238 | GGGATTACAGGC | 249 | 18.22 | [illegible] |
| 6 | 352 | GGGATTACAGGC | 363 | 18.22 | [illegible] |
| 6 | 486 | GGGATTACAGGC | 498 | 18.22 | [illegible] |
| 7 | 650 | GGGATTACAGGC | 661 | 18.22 | [illegible] |
| 8 | 466 | GGGATTACAGGC | 476 | 18.22 | [illegible] |
| 8 | 610 | GGGATTACAGGC | [illegible] | 18.22 | [illegible] |

No.seq Motif Mapping

**Motif Matrix Format**

```
ct1bic4m1
%consensus
GGGATTACAGGC
%degenerate_consensus
GGGATTACAGGC
%counts matrix
     0     1     2     3     4     5     6     7     8     9     10    11
A:  0.00  0.00  0.00  8.00  0.00  0.00  8.00  0.00  8.00  0.00  0.00  0.00
C:  0.00  0.00  0.00  0.00  0.00  0.00  0.00  8.00  0.00  0.00  0.00  8.00
G:  8.00  8.00  8.00  0.00  0.00  0.00  0.00  0.00  0.00  8.00  8.00  0.00
T:  0.00  0.00  0.00  0.00  8.00  8.00  0.00  0.00  0.00  0.00  0.00  0.00

%pwm (position weight matrix)
     0     1     2     3     4     5     6     7     8     9     10    11
A:  0.00  0.00  0.00  1.00  0.00  0.00  1.00  0.00  1.00  0.00  0.00  0.00
C:  0.00  0.00  0.00  0.00  0.00  0.00  0.00  1.00  0.00  0.00  0.00  1.00
G:  1.00  1.00  1.00  0.00  0.00  0.00  0.00  0.00  0.00  1.00  1.00  0.00
T:  0.00  0.00  0.00  0.00  1.00  1.00  0.00  0.00  0.00  0.00  0.00  0.00
```

Description Gene list send to enrichr-CT1S-R1 (6 genes)

Show 5 entries

| Tissue/ Cell type | Species | Additional cell type specific genes found in TAD |
|---|---|---|
| Adrenal | Human | MAP1LC3C |
| Aorta cell | Human | Not found |
| Bladder | Human | Not found |
| Bowel Small | Human | Not found |
| Cortex | Human | Not found |

| Similar CTS-Rs in other cell types |
|---|
| CT1S-R5 CT2S-R7 CT4S-R2 CT5S-R7 CT5S-R9 CT6S-R3 |

Show 10 entries

| Tissue/ Cell type | # of ATAC-seq peaks | # of included regulon genes | Rate in regulon | Species | CistromeDB ID | GEO accession |
|---|---|---|---|---|---|---|
| Pancreas Cancer Cell | 90951 | 3 | 0.500 | Human | 80977 | GSM2549748 |
| Pancreas | 75696 | 3 | 0.500 | Human | 80976 | GSM2549747 |
| Blood B Lymphocyte | 26665 | 3 | 0.500 | Human | 80913 | GSM2702716 |
| Macrophage-2 | 94271 | 3 | 0.500 | Human | 80835 | GSM2592783 |
| Mesodermal Cell | 67665 | 3 | 0.500 | Human | 80593 | GSM2845470 |

TARGET DATABASES

| Database | Used | Matched |
|---|---|---|
| HOCOMOCOv11_full_HUMAN_mono_meme_format | 771 | 57 |
| HOCOMOCOv11_full_MOUSE_mono_meme_format | 531 | 38 |

Summary / Optimal Alignment

| | |
|---|---|
| Name | ZN770_HUMAN.H11MO.1.C |
| Database | HOCOMOCOv11_full_HUMAN_mono_meme_format |
| p-value | 7.77e-07 |
| E-value | 1.01e-03 |
| q-value | 9.78e-04 |
| Overlap | 11 |
| Offset | 0 |
| Orientation | Normal |

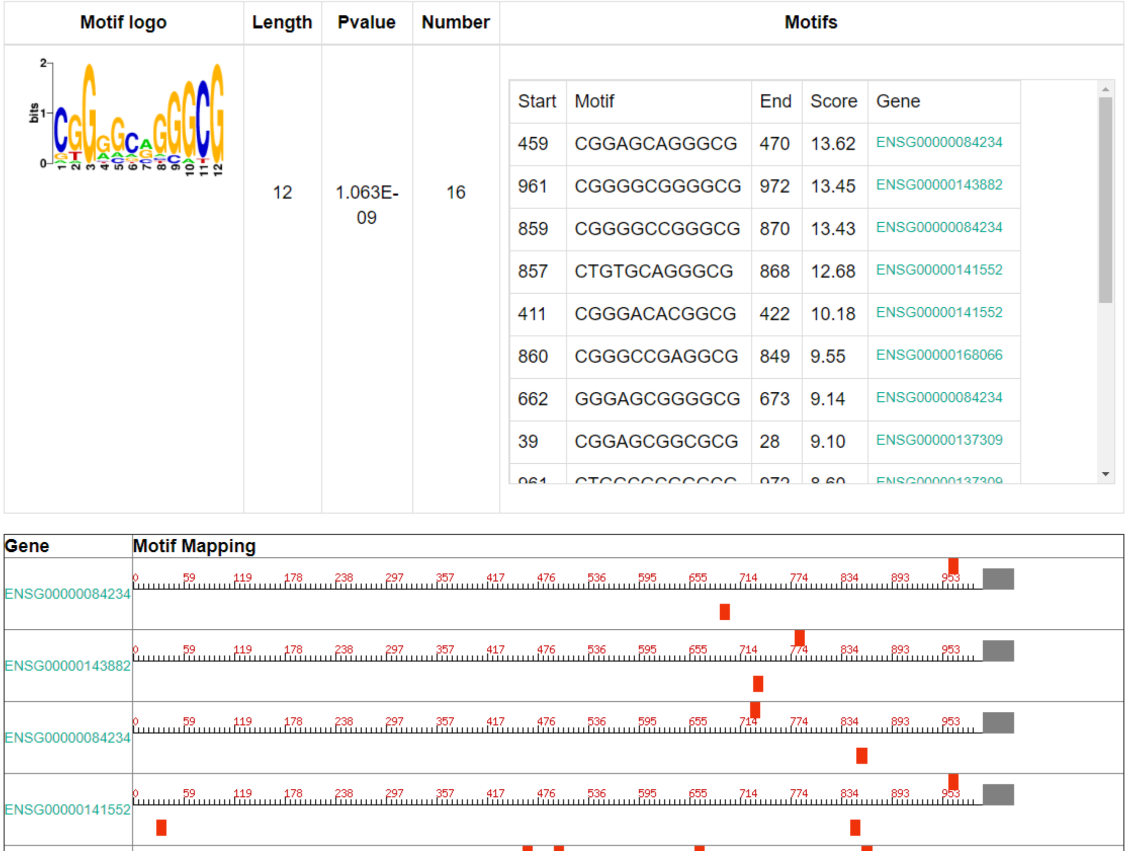
TARGET DATABASES

| Database | Used | Matched |
|---|---|---|
| JASPAR2018_CORE_non-redundant | 1404 | 41 |
| JASPAR2018_CORE_vertebrates_non-redundant | 579 | 6 |

Summary / Optimal Alignment

| | |
|---|---|
| Name | MA0516.1 (SP2) |
| Database | JASPAR2018_CORE_non-redundant |
| p-value | 6.75e-05 |
| E-value | 1.34e-01 |
| q-value | 1.21e-01 |
| Overlap | 12 |
| Offset | 0 |
| Orientation | Reverse Complement |

Figure 15. Example of CT1S-R1 interpretation.



| Motif logo | Length | Pvalue | Number | Motifs |
|---|---|---|---|---|
|  | 12 | 1.063E-09 | 16 |  |

| Start | Motif | End | Score | Gene |
|---|---|---|---|---|
| 459 | CGGAGCAGGGCG | 470 | 13.62 | ENSG00000084234 |
| 961 | CGGGGCGGGGCG | 972 | 13.45 | ENSG00000143882 |
| 859 | CGGGGCCGGGCG | 870 | 13.43 | ENSG00000084234 |
| 857 | CTGTGCAGGGCG | 868 | 12.68 | ENSG00000141552 |
| 411 | CGGGACACGGCG | 422 | 10.18 | ENSG00000141552 |
| 860 | CGGGCCGAGGCG | 849 | 9.55 | ENSG00000168066 |
| 662 | GGGAGCGGGGCG | 673 | 9.14 | ENSG00000084234 |
| 39 | CGGAGCGGCGCG | 28 | 9.10 | ENSG00000137309 |
| 961 | [illegible] | 972 | 8.60 | ENSG00000137309 |

| Gene | Motif Mapping |
|---|---|
| ENSG00000084234 |  |
| ENSG00000143882 |  |
| ENSG00000084234 |  |
| ENSG00000141552 |  |

Figure 16. Result page of identified CTS-R motif details, including motif logo, motif length, P-value, number of motif instances, and detailed information of motif instances, motif positions on the promoters.

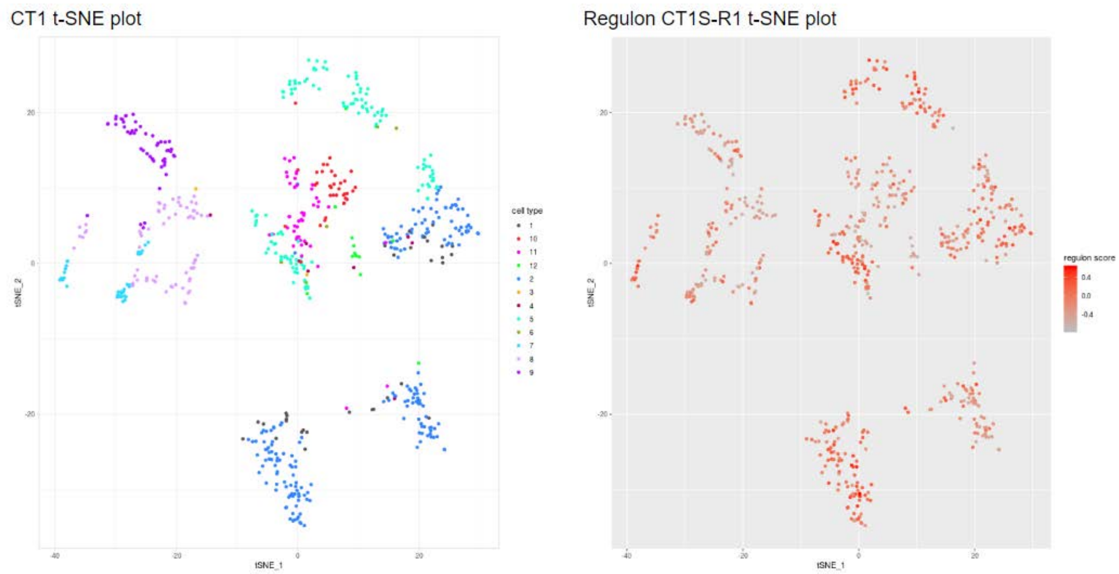
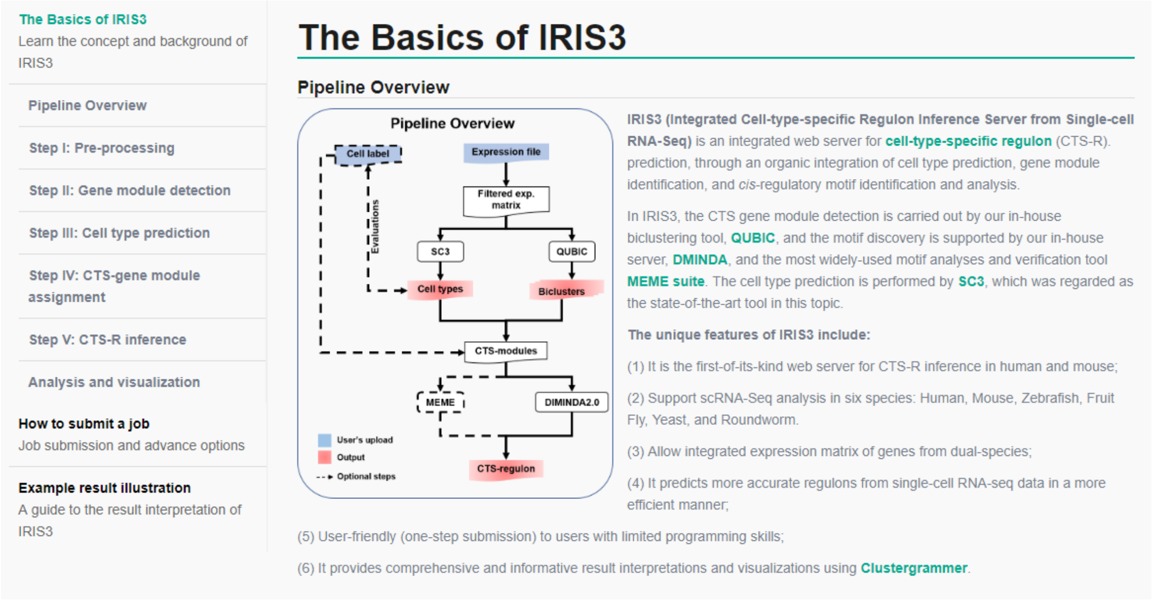


Figure 17. Example t-SNE plot. Colored by cell type or level of regulon activity, respectively.



# The Basics of IRIS3

## Pipeline Overview

**IRIS3 (Integrated Cell-type-specific Regulon Inference Server from Single-cell RNA-Seq)** is an integrated web server for **cell-type-specific regulon** (CTS-R). prediction, through an organic integration of cell type prediction, gene module identification, and *cis*-regulatory motif identification and analysis.

In IRIS3, the CTS gene module detection is carried out by our in-house biclustering tool, **QUBIC**, and the motif discovery is supported by our in-house server, **DMINDA**, and the most widely-used motif analyses and verification tool **MEME suite**. The cell type prediction is performed by **SC3**, which was regarded as the state-of-the-art tool in this topic.

**The unique features of IRIS3 include:**

(1) It is the first-of-its-kind web server for CTS-R inference in human and mouse;

(2) Support scRNA-Seq analysis in six species: Human, Mouse, Zebrafish, Fruit Fly, Yeast, and Roundworm.

(3) Allow integrated expression matrix of genes from dual-species;

(4) It predicts more accurate regulons from single-cell RNA-seq data in a more efficient manner;

(5) User-friendly (one-step submission) to users with limited programming skills;

(6) It provides comprehensive and informative result interpretations and visualizations using **Clustergrammer**.

Figure 18. Result page of IRIS3 tutorials

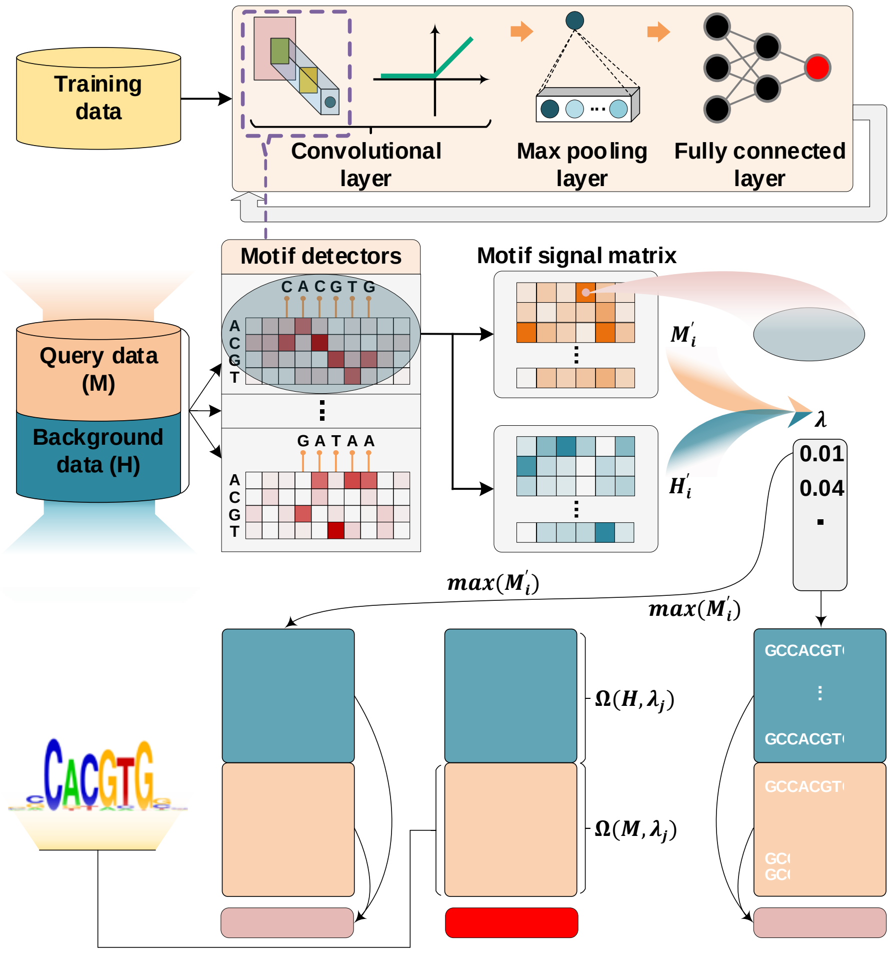


Figure 19. Schematic overview of DESSO framework

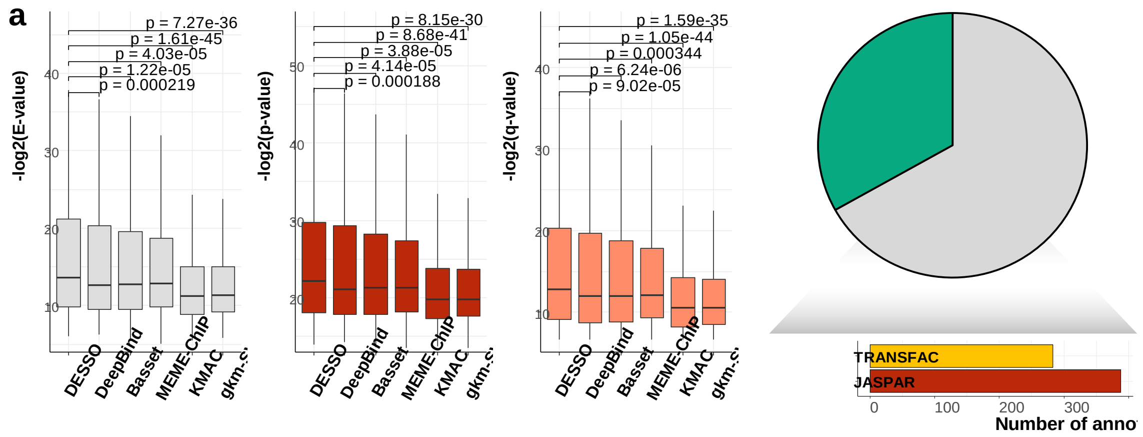


Figure 20. Performance comparison of sequence motif identification accuracy.

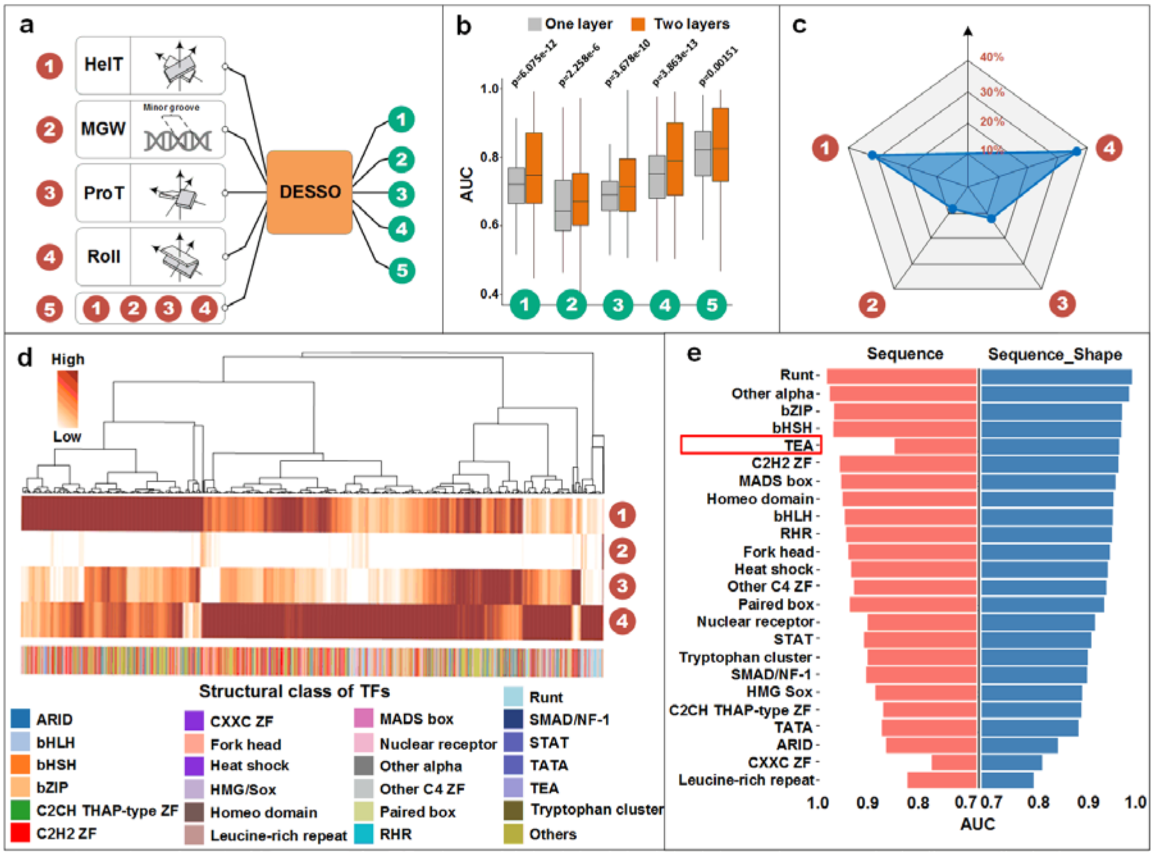


Figure 21. The performance of DNA shape in predicting TF-DNA binding specificity.

Table 1. IUPAC nucleotide code

| Symbol | Mnemonic | Translation |
|---|---|---|
| A | | A (adenine) |
| C | | C (cytosine) |
| G | | G (guanine) |
| T | | T (thymine) |
| U | | U (uracil) |
| R | puRine | A or G (purines) |
| Y | pYrimidine | C or T/U (pyrimidines) |
| M | aMino group | A or C |
| K | Keto group | G or T/U |
| S | Strong interaction | C or G |
| W | Weak interaction | A or T/U |
| H | not G | A, C or T/U |
| B | not A | C, G or T/U |
| V | not T/U | A, C or G |
| D | not C | A, G or T/U |
| N | aNy | A, C, G or T/U |

Table 2. Summary of DNA motif identification tools, citations were collected via Google Scholar as of July 2019.

| Tool | Platform | Citations | Published year | Approach | PMID |
|---|---|---|---|---|---|
| Bioprospector | Command-line tool | 979 | 2001 | Zero to third-order Markov background models | 11262934 |
| MEME | Web server/ command-line tool | 1783 | 2006 | Probabilistic method with expectation-maximization | 16845028 |
| DREME | Web server/ command-line tool | 634 | 2011 | Discriminative Regular Expression Motif Elicitation on ChIP-seq data | 21543442 |
| BoBro | Command-line tool | 27 | 2011 | Two stage alignment and graph based motif finding | 23846744 |
| rGADEM | R package | 43 | 2011 | Genetic Algorithm guided formation of spaced Dyads coupled with EM for Motif identification | 21358819 |
| MEME-ChIP | Web server/ command-line tool | 799 | 2011 | Integrated existing tools: MEME (Multiple EM) and DREME(Discriminative Regular Expression Motif Elicitation algorithm) | 20513432 |
| HOMER | Command-line tool | 3904 | 2010 | Hypergeometric Optimization of Motif EnRichment | 20513432 |
| RSAT peak-motifs | Web server, command-line tool | 178 | 2011 | Implemented RSAT oligo-analysis, RSAT dyad-analysis, RSAT local-word analysis,MEME, ChlPMunk, | 22156162 |
| BammMotif | Web server | 2 | 2018 | Bayesian Markov Models | 29846656 |
| ChIPMonk | Command-line tool | 129, 49, 25 | 2010, 2013, 2014 | Gapless multiple local alignment (GMLA) using the Discrete Information Content (with the Kullback term) | 20736340, 23427986, 24472686 |
| DRAF | Web server | 4 | 2018 | Human database based machine learning model | 29617876 |
| DMINDA2 | Web server | 18 | 2017 | Integrated BoBro | 28419194 |
| CisFinder | Web server | 116 | 2009 | Estimating position frequency matrices (PFMs) directly from n-mer word counts | 19740934 |
| DiNAMO | Command-line tool | 2 | 2018 | An exhaustive and efficient algorithm for IUPAC motif discovery | 29890948 |
| SIOMICS | Command-line tool | 19 | 2014 | Systematic Identification of Motifs In Chip-Seq data | 24322294 |
| Fmotif | Web server | 22 | 2014 | Suffix Tree | 24475069 |
| DeepBind | TF database/ command-line tool | 891 | 2015 | Deep learning based using subjective motifs signals | 26213851 |
| BEEML-PBM | TF database/ command-line tool | 144 | 2011 | Position and effects estimation and modelling weighted regression | 21654662 |

Table 3. Summary of the dataset used for WTSA evaluation

| TF | GEO accession ID | Publish Date | Data ID | #of Bases | #of Identified Sites | Description |
|---|---|---|---|---|---|---|
| Fur | GSE54901 | 2014.9 | GSM1326335 | 193.4M | 556 | From Fur TF with Fe |
| Cra | GSE65643 | 2018.4 | GSM1602341 | 88.6M | 387 | From cra-8myc TF tagged strain_glucose |
| ArgR | GSE60546 | 2015.3 | GSM1482120 | 768.7M | 462 | From ArgR (+arg) rep1 and rep2 |