UNIVERSITY OF CALIFORNIA SAN DIEGO

Score-based Transcription Factor Motif Enrichment Strategies For Analysis of Transcriptional Regulation Of Immune Response

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

by

Nathaniel Delos Santos

Committee in charge:

Professor Christopher Benner, Chair
Professor Melissa Gymrek, Co-Chair
Professor James Kadonaga
Professor Tsung-Ting Kuo
Professor Wei Wang

2022

The dissertation of Nathaniel Delos Santos is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

- AME: Analysis of Motif Enrichment, software package
- ATAC-seq: Assay for transposase accessible chromatin sequencing, a method for profiling chromatin accessibility
- BED: Browser Extensible Data, file format
- BMDM: Bone-marrow derived macrophage
- ChIP: Chromatin Immunoprecipitation
- ChIPed TF: The transcription factor that is the immunoprecipitation target of a ChIP-seq assay; the ChIP-seq target transcription factor
- CNN: Convolutional Neural Network
- csRNA-seq: capped short RNA-sequencing
- CpG: Cytosine-phosphatase-Guanine
- DPE: Downstream Promoter Element
- ENCODE: Encyclopedia of DNA Elements
- ESC: Embryonic Stem Cell
- FDR: False Discovery Rate
- GC content: Guanine-Cytosine content
- GSEA: Gene Set Enrichment Analysis, software package and method
- GRASP: Genome-Wide Repository of Associations Between SNPs and Phenotypes
- GWAS: Genome-Wide Association Study
- H3K27ac: Histone H3 lysine 27 acetylation
- HOMER: Hypergeometric Optimization of Motif EnRichment, software package
- IFN: Interferon
- INR: Initiator
- IRF: Interferon Regulatory Factor
- ISGF3: Interferon-stimulated gene factor 3
- KLA: Kdo2-lipid A
- LD: Linkage Disequilibrium

- LMPP: Learning Motifs from Positional Priors, software method
- LPS: Lipopolysaccharide
- MACS2: Model-based Analysis of ChIP-Seq, software package
- MEA: Motif Enrichment Analysis
- MEIRLOP: Motif Enrichment in Ranked Lists Of Peaks, software package
- MEME: Multiple EM for Motif Elicitation, software package
- MEPP: Motif Enrichment Positional Profiling, software package
- MMLIS: Modified Murray Lung Injury Score
- MNase-seq: Micrococcal nuclease digestion with deep sequencing
- MSE: Mean Squared Error
- MOODS: Motif Occurrence Detection Suite, software package
- PRO-cap: Precision Run-On and capped RNA-sequencing
- PFM: Position Frequency Matrix
- PWM: Position Weight Matrix
- SNP: Single Nucleotide Polymorphism
- STAT: Signal Transducer and Activator of Transcription
- TF: Transcription factor
- TFEA: Transcription Factor Enrichment Analysis, software package
- TSS: Transcription Start Site
- ZOOPS: Zero or one occurrence per sequence

LIST OF SUPPLEMENTAL FILES

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Christopher Benner for his support as chair of my committee and as my PI. He has provided valuable mentorship, guidance, and opportunities in navigating this field of bioinformatics. His experience and advice on bioinformatics methods development have been invaluable in my work and its accessibility as a resource. His guidance over many manuscript drafts and presentations have been instrumental in sharing my work with confidence.

I would also like to thank Sascha Duttke, an alumnus postdoctoral researcher of the Benner Lab. His support in honing my scientific communication skills, and our mutual exchanges over data and methods, have helped me communicate the value of my work. The opportunities created by his protocols have been instrumental in demonstrating the need for the methods described in this dissertation.

I would also like to thank Carlos Guzman, a graduate student in the neighboring Heinz Lab. He has been central to the lab community prior to the lockdowns, and our frequent exchanges over data science tasks in those days have informed my inclination towards methods development and bioinformatics mentorship. His feedback and guidance over my presentations have been invaluable.

Chapter 1, in full, is a modified reprint of the material as it appears in “MEIRLOP: improving score-based motif enrichment by incorporating sequence bias covariates” in BMC Bioinformatics, 2020. Delos Santos, Nathaniel P.; Texari, Lorane; Benner, Chistopher, Springer Nature, 2020. Modifications have been made to the text to ensure consistency with dissertation formatting. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, has been submitted for publication of the material as it may appear in Nucleic Acids Research Genomics and Bioinformatics, 2022, Delos Santos, Nathaniel P.; Duttke, Sascha; Heinz, Sven; Benner, Christopher, Oxford University Press 2022. Modifications have been made to the text to reflect minor editorial feedback. The dissertation author was the primary investigator and author of this paper.

Chapter 3 includes analysis of data that is derived downstream of analysis from the following preprint:

Lam MTY, Duttke SH, Odish MF, Le HD, Hansen EA, Nguyen CT, Trescott S, Kim R, Deota S, Chang MW, Patel A, Hepokoski M, Alotaibi M, Rolfsen M, Perofsky K, Warden AS, Foley J, Ramirez SI,

Dan JM, Abbott RK, Crotty S, Crotty Alexander LE, Malhotra A, Panda S, Benner CW, Coufal NG. “Profiling Transcription Initiation in Peripheral Leukocytes Reveals Severity-Associated Cis-Regulatory Elements in Critical COVID-19” bioRxiv. 2021.

It has been used in this work with the permission of M.T.Y Lam. The dissertation author was the primary author of this chapter.

VITA

2014 Bachelor of Science, University of California San Diego

2014 - 2016 Staff Research Assistant, Jamieson Lab, University of California San Diego

2016 - 2022 Graduate Student Researcher, University of California San Diego

2022 Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Delos Santos NP, Duttke S, Heinz S, Benner C. MEPP: more transparent motif enrichment by profiling positional correlations. Submitted to NAR Genomics and Bioinformatics 2022

Duttke S, Guzman C, Chang M , Xie J, Delos Santos NP, Carlin AF, Heinz S, Benner C. Position-dependent function of human sequence-specific transcription factors. Submitted to Nature 2022

Delos Santos NP, Texari L, Benner C. MEIRLOP: improving score-based motif enrichment by incorporating sequence bias covariates. BMC Bioinformatics. 2020;21(1):410. PMID: 32938397

Brigidi GS, Hayes MGB, Delos Santos NP, et al. Genomic decoding of neuronal depolarization by stimulus-specific npas4 heterodimers. Cell. 2019;179(2):373-391.e27. PMID: 31585079

Jiang Q, Isquith J, Zipeto MA, Diep RH, Pham J, Delos Santos N, et al. Hyper-Editing of Cell-Cycle Regulatory and Tumor Suppressor RNA Promotes Malignant Progenitor Propagation. Cancer Cell. 2019;35(1):81-94.e7. PMID: 30612940

Lazzari E, Mondala PK, Santos ND, et al. Alu-dependent RNA editing of GLI1 promotes malignant regeneration in multiple myeloma. Nat Commun. 2017;8(1):1922. PMID: 29203771

Crews LA, Balaian L, Delos Santos NP, et al. RNA Splicing Modulation Selectively Impairs Leukemia Stem Cell Maintenance in Secondary Human AML. Cell Stem Cell. 2016;19(5):599-612. PMID: 27570067

Pineda G, Lennon KM, Delos Santos NP, et al. Tracking of Normal and Malignant Progenitor Cell Cycle Transit in a Defined Niche. Scientific Reports. 2016;6:23885. PMID: 27041210

Holm F, Hellqvist E, Mason CN, Ali SA, Delos-Santos NP, et al. Reversion to an embryonic alternative splicing program enhances leukemia stem cell self-renewal. Proc Natl Acad Sci. 2015;112(50):15444-15449. PMID: 26621726

FIELDS OF STUDY

Major Field: Bioengineering

Studies in Biotechnology

Major Field: Bioinformatics

Studies in Biomedical Informatics

Professor Christopher Benner

ABSTRACT OF THE DISSERTATION

Score-based Transcription Factor Motif Enrichment Strategies For Analysis of Transcriptional Regulation Of Immune Response

by

Nathaniel Delos Santos

Doctor of Philosophy in Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

University of California San Diego 2022

Professor Christopher Benner, Chair

Professor Melissa Gymrek, Co-Chair

Transcription factors (TFs) mediate transcriptional responses, allowing cells to respond to changing internal or external stimuli, including infection. The DNA sequences bound by TFs in regulatory regions are called motifs. Researchers employ Motif Enrichment Analysis (MEA) methods to study transcriptional regulation, which analyzes DNA sequences from regulatory regions and determines the

statistical overrepresentation of motifs in those sequences, allowing inference of relevant TFs for that set of regulatory regions.

However, most MEA tools feature oversimplifications of one or more pertinent axes of the data, obscuring potential insights into transcriptional regulation.

For example, most MEA require thresholding of DNA sequences into two sets in order to determine motif enrichment, thus oversimplifying the underlying biological scores that determine those sets, which can prevent biological discovery. We introduce Motif Enrichment In Ranked Lists of Peaks (MEIRLOP), a score-based MEA method that allows researchers to determine the enrichment of motifs within a dataset of scored regulatory region DNA sequences. MEIRLOP uniquely utilizes a logistic regression model that also accounts for lower order levels of sequence bias and other covariates. We demonstrate its utility on multiple ChIP-seq datasets, where it proves more capable (relative to other methods) of finding the enrichment of key transcription factor binding motifs, including the enrichment of binding sites key to immune response.

An overlooked axis in most MEA is the position of motifs relative to anchor features such as transcription start sites (TSS), which can be characterized at high positional resolution using capped short RNA-sequencing (csRNA-seq). We introduce Motif Enrichment Positional Profiling (MEPP), which uses specialized convolutional neural networks to create a profile that characterizes motif enrichment at different motif positions over a dataset of scored sequences. We also introduce Learning Motifs from Positional Priors (LMPP), which uses machine learning to perform the opposite of MEPP: Learning a motif whose positional enrichment resembles a target profile. We use both methods to analyze multiple TSS from csRNA-seq datasets, revealing the positional preferences of transcription factors key to antibacterial and antiviral responses.

Overall, this dissertation presents novel methods by which researchers may analyze transcriptional regulation in and beyond immune response.

Introduction

Regulation of transcription, the process of converting DNA to RNA, is vital to multiple biological processes, including immune response [1]. Key to transcriptional regulation are the actions of transcription factors (TFs), which bind onto specific sequences in regulatory region DNA, called binding motifs, whereupon they recruit other factors or transcriptional machinery, leading to induction or repression of transcriptional activity [2,3]. This process coordinates changes in the transcription of multiple genes across the genome, giving rise to host protective functions, such as the expression of proteins that act to clear the host of pathogens [1].

Given the importance of transcription factors and their binding motifs to transcriptional regulation, multiple bioinformatics methods have been developed to perform Motif Enrichment Analysis (MEA). MEA methods analyze sets of regulatory region DNA sequences for motifs, and quantify the statistical overrepresentation of motifs within those sequences [4,5]. Researchers can then use these motifs to determine important regulatory region sequence features and infer which transcription factors mediate transcriptional regulation under different conditions or comparisons of interest [6]. For example, to determine TFs relevant to immune response, one could analyze the differential activity of regulatory regions between infected and control conditions, then determine the motifs enriched in the set of regulatory regions that appear more active during infection. The TFs that bind to those motifs could then be implicated as potential therapeutic targets. However, by conceiving of MEA in this set-oriented way, we ignore one axis of the data, while oversimplifying another, effectively fitting the data to the limitations of a method.

The previous example demonstrates an issue with many MEA analyses: due to their formulation of motif enrichment, they require researchers to potentially oversimplify one axis of the data. By designating a subset of regulatory regions as being more active in one condition, a fundamental truth about biology is ignored: the activity of regulatory regions (and biological activity in general) typically follows degrees of response rather than discrete categories [7]. By oversimplifying these degrees of

response into discrete sets, researchers risk mischaracterizing motif enrichment along these degrees of response, potentially leading to under-interpretation of their data and missed discoveries.

This issue is the main focus of our first chapter, on Motif Enrichment In Ranked Lists Of Peaks (MEIRLOP), but it consistently carries through into subsequent chapters as well: The methods developed in this dissertation all accept scored sequences from regulatory regions, preventing researchers from oversimplifying the activity of regulatory regions into discrete sets and thus oversimplifying their data. MEIRLOP in particular also pays special attention to covariates that may otherwise confound the process of motif enrichment, and in turn, the inference of relevant transcription factors. We demonstrate the utility of MEIRLOP on multiple ChIP-seq datasets, and find that its attention to degrees to transcriptional response and controlling for covariates allows it to determine the enrichment of key binding motifs for the biological processes under study, including the innate immune response.

While biological activity is an oversimplified axis in many analyses, many MEA methods including MEIRLOP ignore a different axis altogether: Position. In many MEA that ignore position, motif enrichment is characterized as a function of whether or not a motif is present, without accounting for where that motif appears. While ignoring position may suffice for most analysis on lower-resolution assays of regulatory region activity (e.g. ChIP-seq), other assays allow for profiling of transcription start sites (TSS): These assays include CAGE-seq, PRO-cap, and the more recent csRNA-seq [8–10]. Position can be important when analyzing regulatory region sequence anchored on TSS, because the location and presence of motifs can determine where transcription is initiated [11]. Furthermore, motifs can be constrained relative to other motifs, due to the interactions of TFs with each other and with transcriptional machinery [12–14]. There is thus a need for tools that allow the characterization of motif enrichment while accounting for both motif position and degrees of transcriptional response.

Our second chapter on Motif Enrichment Positional Profiling (MEPP), presents a novel MEA method to address this need. Rather than returning a single enrichment score for the enrichment of a motif along a set of scored sequences, as MEIRLOP does, MEPP returns a series of scores comprising an enrichment positionality profile, describing the enrichment of that motif at multiple positions surrounding important anchor features, such as TSS. It also generates a transparent 2D heatmap visualization of where motifs appear throughout both axes of sequence score and motif position. In this

way, MEPP allows researchers to perform both qualitative and quantitative assessments of motif enrichment across both axes. We demonstrate the utility of this method on multiple datasets, including those relevant to pathways for detecting viral infection.

The methods previously described deal with enrichment of motifs whose sequence identity is already known. However, the need to know all the motifs to analyze beforehand can limit the potential for biological discovery. The final chapter, on Learning Motifs with Positional Priors (LMPP), extends MEPP by effectively reversing its operation: Rather than obtaining an enrichment profile from the enrichment of a known motif, LMPP instead uses machine learning to recognize motifs whose enrichment profiles approximate a user-specified positional prior. We demonstrate this method's capabilities in learning motifs from multiple datasets, including an analysis of COVID-19 symptom severity, where we learn motifs that are known or likely to be relevant to SARS-CoV-2 viral infection.

Overall, this dissertation presents a suite of methods intended to allow researchers to make discoveries about TFs and their binding motifs, by more fully utilizing otherwise oversimplified or ignored axes of their data. We have developed, demonstrated, and released these methods with the goal that they may be used by the wider scientific community, to uncover further biological insights about transcriptional regulation both within and beyond immune response.

Chapter 1. MEIRLOP: improving score-based motif enrichment by incorporating sequence bias covariates

1.1 Abstract

1.1.1 Background

Motif enrichment analysis (MEA) identifies over-represented transcription factor binding (TF) motifs in the DNA sequence of regulatory regions, enabling researchers to infer which transcription factors can regulate transcriptional response to a stimulus, or identify sequence features found near a target protein in a ChIP-seq experiment. Score-based MEA determines motifs enriched in regions exhibiting extreme differences in regulatory activity, but existing methods do not control for biases in GC content or dinucleotide composition. This lack of control for sequence bias, such as those often found in CpG islands, can obscure the enrichment of biologically relevant motifs.

1.1.2 Results

We developed Motif Enrichment In Ranked Lists of Peaks (MEIRLOP), a novel MEA method that determines enrichment of TF binding motifs in a list of scored regulatory regions, while controlling for sequence bias. In this study, we compare MEIRLOP against other MEA methods in identifying binding motifs found enriched in differentially active regulatory regions after interferon-beta stimulus, finding that using logistic regression and covariates improves the ability to call enrichment of ISGF3 binding motifs from differential acetylation ChIP-seq data compared to other methods. Our method achieves similar or better performance compared to other methods when quantifying the enrichment of TF binding motifs

from ENCODE TF ChIP-seq datasets. We also demonstrate how MEIRLOP is broadly applicable to the analysis of numerous types of NGS assays and experimental designs.

1.1.3 Conclusions

Our results demonstrate the importance of controlling for sequence bias when accurately identifying enriched DNA sequence motifs using score-based MEA. MEIRLOP is available for download from https://github.com/npdeloss/meirlop under the MIT license.

1.2 Background

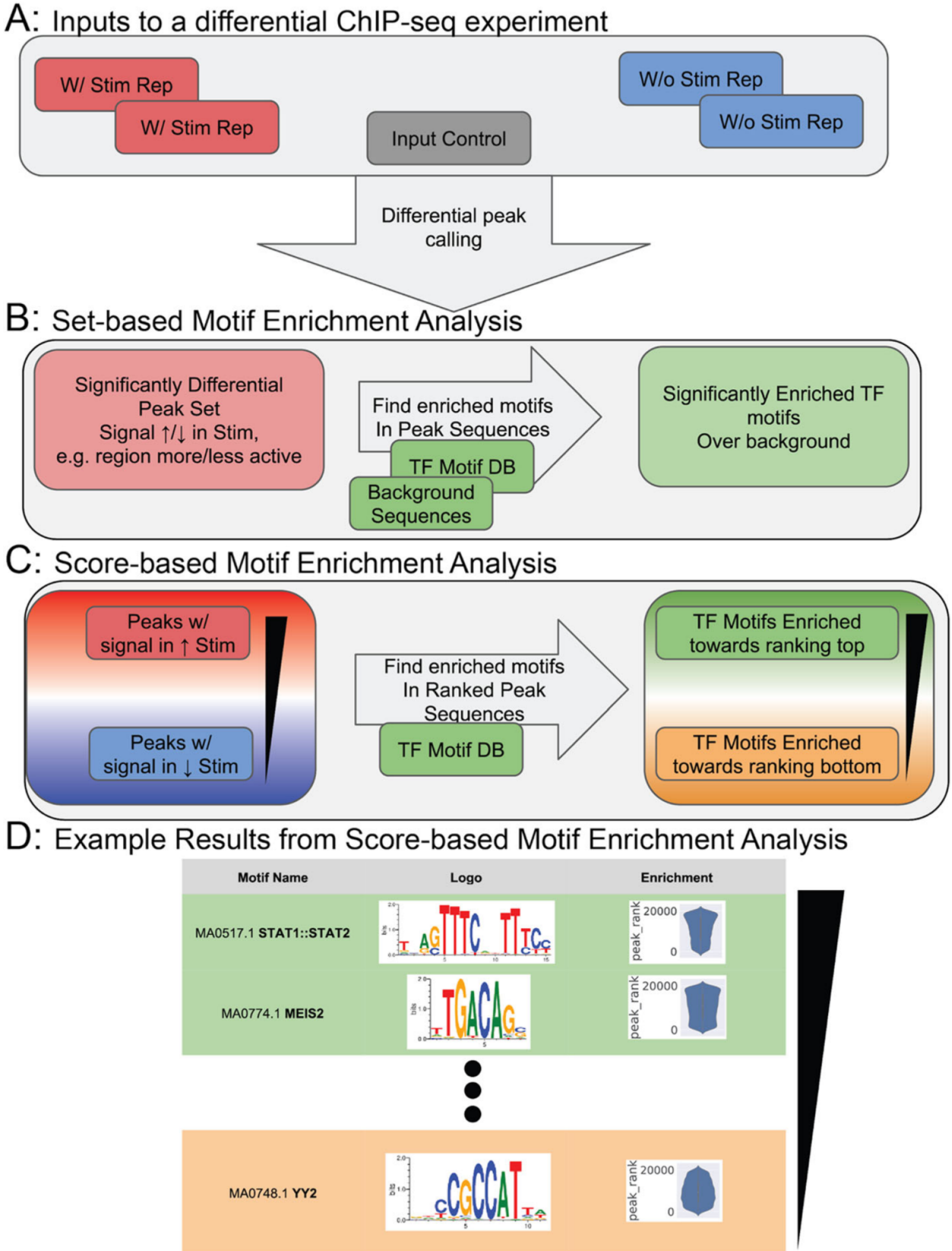
Transcription factors (TFs) mediate transcriptional responses, inducing or repressing transcription of genes by binding to DNA at regulatory regions and recruiting RNA polymerase or accessory factors [2]. Motif enrichment analysis (MEA) of regulatory sequences is a method used to identify over-represented DNA sequence patterns (motifs) in regulatory regions [4,5]. Bioinformatics methods usually represent TF binding motifs as position weight matrices (PWMs), which describe the DNA-binding specificities of transcription factors, and can predict potential binding sites in regulatory regions [3,15]. Researchers use the enrichment of TF binding motifs to infer which TFs may contribute to specific transcriptional responses by binding those motifs in regulatory region sequences [6]. This enables the inference of which TFs mediate a cell's transcriptional response to a condition, such as infection or disease state, marking those TFs as potential therapeutic targets.

When using most MEA methods, regulatory regions are typically scored by their biological response in transcriptomic or epigenomic assays. For example, H3K27ac localization on chromatin serves as a marker for active regulatory regions [16]. MEA on differential H3K27ac ChIP-seq data can reveal which motifs and TFs regulate transcription in regions that change their activity in response to stimulation [17]. Researchers typically filter these regulatory regions by a score threshold and place them

into sets to yield contrasting categories (e.g., regulatory regions with higher activation levels after stimulation vs. those with lower activation levels after stimulation) (Fig. 1.1A,B). Then, motif scanners detect motifs in sequences within those categories, followed by set enrichment tests (e.g. the Fisher exact test) which determine the overrepresentation of motifs in each category. This allows the imputation of motifs and transcription factors that influence transcriptional response in those conditions (Fig. 1.1B). We term this process set-based MEA, for its thresholding of sequences into categorical sets.

Figure 1.1: Motif enrichment analysis can be based on discrete sets of sequences or scored sequences.

(A) Diagram of inputs to a typical differential ChIP-seq binding experiment, consisting of replicates in stimulated and unstimulated conditions, and an input experiment(s).
(B) Diagram of set-based motif enrichment methods. These methods receive significantly differentially regulated subsets of regulatory sequences (e.g. peaks) and return a significantly enriched subset of TF motifs.
(C) Diagram of score-based motif enrichment methods. These receive a list of scored sequences based on differential signal between stimulated and unstimulated conditions, then return TF motifs enriched towards the top or bottom of the scored sequences.
(D) Illustration of the potential output of a score-based enrichment method, featuring motif name, motif logo, and a description of the enrichment of a motif towards the top or bottom of scored sequences


A: Inputs to a differential ChIP-seq experiment
W/ Stim Rep
W/ Stim Rep
Input Control
W/o Stim Rep
W/o Stim Rep
Differential peak calling
B: Set-based Motif Enrichment Analysis
Significantly Differential Peak Set
Signal ↑/↓ in Stim,
e.g. region more/less active
Find enriched motifs In Peak Sequences
TF Motif DB
Background Sequences
Significantly Enriched TF motifs
Over background
C: Score-based Motif Enrichment Analysis
Peaks w/ signal in ↑ Stim
Peaks w/ signal in ↓ Stim
Find enriched motifs In Ranked Peak Sequences
TF Motif DB
TF Motifs Enriched towards ranking top
TF Motifs Enriched towards ranking bottom
D: Example Results from Score-based Motif Enrichment Analysis
Motif Name
Logo
Enrichment
MA0517.1 STAT1::STAT2
MA0774.1 MEIS2
MA0748.1 YY2
peak_rank
20000
0

However, this thresholding ignores how the magnitude of a regulatory region's response can occur along a continuum, and removes relevant information from the score: For example, by collapsing strongly activated and weakly activated regulatory regions into the same set of "activated regulatory regions", set-based MEA obscures potentially important differences between those two categories. In addition, preset thresholds can lead to few regulatory regions being present in one or more sets, reducing confidence in downstream enrichment results. Such sparsity can occur because of technical reasons: the choice of underlying normalization and differential assay analysis influences the number of genomic features that pass thresholds for differential expression/activation [18,19]. When this occurs, a researcher may attempt to revisit the underlying scores, applying different thresholds to achieve a suitable enrichment: However, after testing many thresholds, a significant enrichment can appear purely by chance. This motivates the use of score-based MEA that uses the biological response scores of regulatory regions directly.

Threshold-free motif enrichment asks: Given a list of TF binding motifs and a sorted list of scored/ranked sequences, what motifs are enriched at the top or bottom of that list (Fig. 1.1C, D)? This follows from the fact that most biological data, differential or otherwise, naturally lends itself to ranking [7]. While individual regulatory regions may not fulfill pre-set thresholding criteria for significantly differential signal, motifs may appear in regulatory regions biased towards the top or bottom of a list, showing a correlation with the quantities underlying any potential differential signal thresholds. This concept has precedence in gene expression and ontology analyses: GSEA Prerank identifies significantly enriched gene sets at the top or bottom of a differential ranking of genes [20].

Related work has developed multiple score-based MEA methods for inferring relevant TFs. A handful of implementations share many similarities with set-based MEA, except that they determine a data-driven partitioning threshold along the scored regulatory regions. The chosen threshold maximizes the resulting enrichment from a Fisher exact test or hypergeometric test [21,22]. An alternative method is to perform a rank-sum test, comparing the score ranks of regulatory regions that contain a motif against those that do not. Other strategies involve finding the correlation between the strength of a motif in a

regulatory region's sequence against the score of the regulatory region, under the reasoning that stronger instances of a relevant motif should contribute more to the score. The software package AME (from the MEME suite) implements many of these strategies, including variable threshold Fisher exact tests, rank-sum MEA, and correlation with Pearson coefficients for linear correlation, as well as Spearman coefficients for rank correlation [4].

One complication that can arise when performing MEA is that a nonrandom distribution of nucleotide or dinucleotide frequencies can lead to a bias in sequence content among sets of regulatory region sequences. Thus, MEA methods that do not control for this effect may assign significance to a motif that preferentially recognizes sequences favored by the biased sequence composition. For example, CpG islands are present in promoter regions of genes, so a motif scanner is more likely to observe GC-rich motifs there by chance. Similarly, dinucleotides found in the first two bases of codons may be over-represented in DNA sequence sampled near coding exons, and thus can contribute to spurious enrichment values for motifs over-represented near intron borders [23]. Multiple set-based MEA methods control for this effect: AME and CentriMo from the MEME Suite allow specification of a control sequence file for comparative motif enrichment, and can generate the control/background sequences by shuffling dinucleotides from the target sequence set [4,5]. HOMER can automatically select sequences from a reference genome that resemble the GC content of a target set of motifs, and can also re-weight a sequence's contribution to a contingency table to minimize differences in weighted dinucleotide frequency between sets [24]. Methods such as Pscan/Pscan-ChIP, CLOVER, BiasAway, and GENRE generate or retrieve background sequences to control for sequence properties of a target sequence set [6,23,25–27]. However, such control is absent in current score-based MEA methods that identify motifs through a PWM.

To provide a score-based MEA method more robust to covariate effects, we developed MEIRLOP (Motif Enrichment In Ranked Lists of Peaks) [28]. MEIRLOP uses logistic regression to model the probability of a regulatory region sequence containing a motif as a function of a regulatory region's activity score. To account for sequence bias, our method derives covariates from the dinucleotide frequencies of the regulatory region sequences, and incorporates these into the logistic regression model. The

regression model's coefficient for the score then summarizes whether a motif is more likely to appear in regulatory regions with higher or lower response scores. In addition, we avoid multicollinearity in the logistic regression model by reducing the covariates into a set of principal components summarizing 99% of the variance. This dimensional reduction maintains the stability of the regression model and improves estimation of regression coefficients [29].

Although prior works have used logistic regression in motif analysis, these methods have not used quantitative changes in regulation as the predictor variable. Keles et al. employ logistic regression as another form of set-based MEA: They use the presence or absence of a motif as a predictor of whether a sequence is in a target set [30]. Yao et al. set binding site presence or absence as outcome variables of logistic regression. However, their method uses a predictive variable derived not from sequencing coverage of regulatory regions, but from the overall count of motifs in a region's sequence [31].

MEIRLOP enables enrichment of biologically relevant TF binding motifs where other methods may fixate on nucleotide or dinucleotide sequence bias. On differential H3K27ac data, it achieves superior accuracy over other methods in identifying significant enrichment of motifs known to mediate interferon signaling response. On ENCODE ChIP-seq data, it achieves improved performance relative to other score-based MEA methods implemented in AME, but returns results over twenty times faster. Finally, we demonstrate how MEIRLOP can be applied to a variety of different NGS profiling methods and different activity scores to improve the identification and interpretation of functionally relevant TF motifs.

1.3 Implementation

1.3.1 MEIRLOP motif enrichment procedure

MEIRLOP is based on a logistic regression model for motif enrichment. At minimum, it accepts a list of scored sequences (in the AME scored FASTA format, with sequence headers consisting of a name and score separated by a space), and a motif database (in JASPAR format). MEIRLOP's novel motif enrichment procedure is executed in three parts, described below:

1. Scanning sequences for motifs
2. Principal component reduction of covariates
3. Logistic regression for motif enrichment

1.3.1.1 Scanning sequences for motifs

To detect transcription factor binding motifs in genomic sequence, we use the MOODS motif scanner to scan for sequence matching PWMs from the input motif set, which is provided to MEIRLOP in the JASPAR format [15,32]. In our command lines, we refer to this motif set with the filename: "jaspar.txt".

The MOODS motif scanner internally takes two parameters to determine if a subsequence matches a motif matrix: a pseudocount and a p-value, which default to 0.001. MEIRLOP sets these parameters using arguments '--pcount' and '--pval'.

In this work we use the JASPAR 2018 CORE vertebrate non-redundant motif set, which consists of 579 motifs [15]. We selected this motif set to limit redundancy between different versions of what are essentially the same motif (i.e. avoid many motifs matching the same transcription factor), and to restrict the motifs tested to those relevant for analyzing human data. In addition, this set of motifs is available in formats compatible with AME, facilitating comparison of our method with AME.

1.3.1.2 Principal component reduction of covariates

Although it is possible to directly input sequence-derived covariates into a logistic regression model, when using the k-mer frequencies of sequences as covariates, the multicollinearity of these frequencies (due to e.g. CpG islands) can lead to model instability and inaccurate parameter estimation [29]. To account for this while preserving the ability to control for k-mer frequencies, we adapt the strategy of Aguilera et al.: We reduce multiple k-mer frequency covariates into a lower-dimensional set of principal components [29], converting the multicollinear predictors into a set of linearly uncorrelated predictors explaining 99% of the variance. We use the PCA implementation available in scikit-learn [33]. We refer to a single resulting reduced covariate as x_c.

By default, MEIRLOP controls for dinucleotide frequency covariates, but this behavior can be adjusted with the '--kmer' argument to control for single nucleotide frequencies ('--kmer 1') or no covariates ('--kmer 0'). Additional experimentally derived covariates can be incorporated using the '--covariates' argument.

1.3.1.3 Logistic regression for motif enrichment

Although prior work has previously applied logistic regression to the analysis of transcription factor binding sites [30,31], we present a different logistic regression model that more closely follows the example of linear regression as applied in AME [4], while controlling for the effect of collinear covariates (e.g. short k-mer frequencies).

Let p be the probability of a sequence containing a given motif m, and let x_s be the score assigned to the sequence, e.g. H3K27ac ChIP-seq signal. We then model the log-odds of a sequence with score x_s containing motif m as:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_s x_s + \sum_c^n \beta_c x_c$$

Where x_c refers to one of the n reduced covariates previously described, with β_c being the corresponding coefficient. After maximum likelihood estimation of the coefficients and bias term, β_s can be interpreted as the change in the log-odds of a sequence containing motif m, for a one unit increase in x_s.

To ensure that the maximum likelihood estimation converges, all predictor variables are standardized. The significance of the coefficient β_s is determined using the Wald test, as per the Statsmodels implementation of logistic regression [34]. To control for multiple hypothesis testing across multiple motifs, we applied Benjamini-Hochberg correction to the Wald test *p*-values [35].

This approach to motif enrichment allows controlling for sequence derived covariates such as GC content and k-mer frequencies, allowing the model to compensate for sequence bias similarly to certain set-based methods such as HOMER [24].

MEIRLOP displays logistic regression results using an interactive HTML table powered by Datatables.net, with motif logos generated by Logomaker [36].

1.3.2 Differential analysis of IFN-β stimulation against control

We started with cell culturing and ChIP-seq for data acquisition. This was followed by differential ChIP-seq bioinformatics analysis comparing stimulated samples vs. unstimulated controls. Finally, we ran motif enrichment analysis on the differential ChIP-seq peaks using MEIRLOP and AME in different configurations.

We conducted the data acquisition in two parts, described below:

1. Cell culture and treatment
2. ChIP-seq and Crosslinks

The data analysis and comparison was conducted in three parts, also described below:

1. Differential ChIP-seq analysis (preprocessing)
2. Motif enrichment analysis (running MEIRLOP & AME)
3. Motif enrichment accuracy evaluation

1.3.2.1 Cell culture and treatment

HCT116 CMV-osTIR1 RAD21-mAC cells were obtained from Masato T. Kanemaki [37] and cultured in McCoy's 5A medium supplemented with 10% FBS. Cells were grown in a 37 °C incubator with 5% CO2. Cells were treated with 0.1% final DMSO for 6 h and then treated with either IFN-β (1000 unit/ml) for one hour or not further treated.

1.3.2.2 ChIP-seq and crosslinking

Crosslinking and ChIP-seq was performed as described in Heinz et al., 2018 with few adjustments [38]. Briefly, cells were fixed directly by adding formaldehyde into media to a final concentration of 1% formaldehyde for 10 min at room temperature and quenched with 125 mM Glycine. Cells were then pelleted at 300 g for 5 min at 4 °C, washed twice with cold PBS (with 0.5% BSA), snap frozen in liquid nitrogen and stored at − 80 °C.

ChIP-seq was performed on 500,000 cells as described in Heinz et al., 2018 [38]. H3K27ac antibodies were obtained from Active Motif (cat#:39133). Libraries were single-end sequenced for 84 bp to a depth of 5.5–8.6 reads on an Illumina NextSeq500 instrument.

1.3.2.3 Differential ChIP-seq analysis

After sequencing and adapter trimming with FastP [39], 5.5 M - 8.6 M reads per library were aligned to the GRCh38 reference genome using bowtie2 at overall alignment rates of 98.5–99.5% [40]. For each stimulation (n = 2) and control sample (n = 2), MACS2 was used to call peaks of median length 1kbp relative to a background input sample (of 16 M reads) [41]. DiffBind was used to call differentially acetylated peaks between stimulated and unstimulated conditions [42]: Internally, DiffBind counts reads from each sample within each peak, then uses DESeq2 to calculate differential statistics such as log 2 fold change while accounting for library size and replicate/batch effects.

In order to obtain scores for these peaks and their sequences, we used the log2 fold change for each feature as called by DiffBind as the score. For each peak, the sequence +/− 500 bp of each feature's

center was extracted. The motif enrichment analysis methods we evaluated then received these scores and sequences for input, in the scored FASTA file “deseq2-ifnb.diff-peaks.scored.fa”.

1.3.2.4 Motif enrichment analysis

When evaluating MEIRLOP with dinucleotide frequency covariates, we used the command:

“OMP_NUM_THREADS=25 $(which time) --verbose python -m meirlop --jobs 25 --kmer 2 --pcount 0.01 --pval 0.001 --html --scan --fa deseq2-ifnb.diff-peaks.scored.fa jaspar.txt {meirlop_output_directory}”.

The argument “--kmer 2” sets MEIRLOP to use dinucleotide frequency covariates. The file “deseq2-ifnb.diff-peaks.scored.fa” is available in the “data” directory of the MEIRLOP Github repository.

When evaluating MEIRLOP without covariates, we used the command:

“OMP_NUM_THREADS=25 $(which time) --verbose python -m meirlop --jobs 25 --kmer 0 --pcount 0.01 --pval 0.001 --html --scan --fa deseq2-ifnb.diff-peaks.scored.fa jaspar.txt {meirlop_output_directory}”.

The argument “--kmer 0” sets MEIRLOP to use no k-mer frequency covariates.

MEIRLOP gives positive enrichment values for motifs overrepresented in sequences with higher scores, while AME gives positive enrichment values for motifs overrepresented in sequences with lower scores. Due to this difference in scoring/enrichment conventions we must transform the input sequence scores for AME relative to those used by MEIRLOP: we multiplied scores by -1 prior to input for AME and its implemented methods [4]. AME version 5.0.2 was used unless otherwise specified [4].

1.3.2.5 Motif enrichment accuracy evaluation

To assess motif enrichment results across multiple methods, we adopt the strategy previously used to assess different motif enrichment methods in AME [4]. We reduce motif enrichment scores to the

percentile rank of the motif enrichment score. The formula for this percentile rank accuracy (PRA) metric is reproduced below from eq. 8 of McLeay & Bailey 2010 [4]:

$$PRA = \frac{R_k}{N}$$

Where Rk is the rank of the target motif Mk within a method's enrichment results, and N is the total number of motifs in the database. Since true positive results are expected to have larger positive enrichment scores in these method comparison experiments, Rk is higher for more highly positive scores and reflects a better result for a method.

Exclusion from the results table of any method defaults to a rank score of 0, and ties resolve to the lowest applicable rank. This penalizes methods that do not provide results for the correct motif, or that simply assign many motifs the same high enrichment score. For methods implemented in AME, the rank is higher the earlier the motif appears in AME's result table. For our method, the rank is correlated with the logistic regression coefficient of regulatory region score against motif presence. To obtain percentile ranks, the rank is divided by the number of motifs in the reference motif database.

We use the percentile rank accuracy metric to assess the accuracy of both AME and MEIRLOP, so this calculation is not implemented in MEIRLOP itself, but is instead evaluated after running both AME and MEIRLOP.

1.3.3 Comparison of MEIRLOP and AME on ENCODE ChIP-seq data

We started by selecting ENCODE ChIP-seq experiments and extracting scored FASTA files for input into MEIRLOP and AME. Then, we ran MEIRLOP and AME to analyze motif enrichment in the scored sequences in these FASTA files and evaluated each method's performance using the percentile rank accuracy metric previously described.

This comparison was conducted in three parts, two of which are described below:

1. ENCODE ChIP-seq experiment selection and sequence extraction (preprocessing)

2. Motif enrichment analysis on ENCODE ChIP-seq data (running MEIRLOP & AME)
3. Motif enrichment accuracy evaluation (as previously described)

1.3.3.1 ENCODE ChIP-seq experiment selection and sequence extraction

In order to determine the performance of logistic regression with covariates on other ChIP-seq datasets, we used MEIRLOP on scored ChIP-seq peaks from 582 ENCODE ChIP-seq TF binding experiments, summarized in supplementary Table 1.S3 [43,44]. These were assays on human cell lines for which the ChIPed TF could be matched to a TF binding motif in the JASPAR 2018 non-redundant motif database by gene name. These experiments were selected to exclude those with severe (red flag) audit categories. From each experiment, approximately 300K peaks were obtained, with peak scores assigned by SPP. Peaks were obtained prior to filtering for IDR in order to obtain enrichments across a range of peaks including those lacking high affinity binding motifs. Motif enrichment was restricted to sequence within +/− 200 bp of the peak center.

1.3.3.2 Motif enrichment analysis on ENCODE ChIP-seq data

MEIRLOP was run with covariates controlling for sequence dinucleotide frequencies. MEIRLOP was invoked multiple times on data for different ENCODE accession numbers. E.g., for ENCODE accession ID “ENCSR976TBC”, MEIRLOP was run using the command line:

> “OMP_NUM_THREADS=25 $(which time) --verbose python -m meirlop --jobs 25 --html --pcount 0.1 --pval 0.001 --kmer 2 --fa scored_fastas/ENCSR976TBC.fa jaspar.txt meirlop_outputs/ENCSR976TBC”.

We parallelized AME runs using GNU Parallel [45]. As described previously, sequence scores were inverted for use with AME.

1.3.4 Differential csRNA-seq analysis

We started by extracting sequences and scores for differential TSS data previously generated as described in Duttke et al. [10]. We then ran MEIRLOP on the scored TSS sequences.

This analysis was performed in two parts, described below:

1. TSS sequence extraction and scoring (preprocessing)
2. MEIRLOP analysis of TSS

1.3.4.1 TSS sequence extraction and scoring

Differential TSS were found from murine BMDMs as described in Duttke et al. [10]. We performed enrichment for motifs found within +/− 150 bp of the TSS, and scored sequences by the differential log 2 fold change between KLA stimulated and unstimulated control conditions as computed by HOMER getDiffExpression.pl (which wraps DESeq2 to calculate differential statistics while accounting for library size and replicates).

1.3.4.2 MEIRLOP analysis of TSS

MEIRLOP was run with covariates controlling for sequencing dinucleotide frequencies. We ran MEIRLOP with the command line:

> “OMP_NUM_THREADS=30 $(which time) --verbose python -m meirlop --jobs 30 --kmer 2 --pcount 0.001 --pval 0.001 --html --scan --sortabs --bed differential_tss.bed --fi genome.fa jaspar.txt differential_tss.meirlop”.

Using MEIRLOP with input bed files this way requires bedtools version 2.29.0 [46,47]: Later versions (specifically 2.29.2) have incompatible behavior for the “bedtools getfasta” subcommand. The file “differential_tss.bed” is available in the “data” directory of the MEIRLOP Github repository.

1.3.5 Analysis of DHS scored by histone ChIP-seq ratios

We started by retrieving DNase-seq and histone ChIP-seq data from ENCODE, then extracting sequence, coverage, and signal data [43,44]. These were converted into scored sequence FASTA files and a covariate file, where each entry in these files corresponded to one DHS. We then ran MEIRLOP on the scored DHS sequences and their covariates.

This analysis was performed in two parts, described below:

1. DHS sequence extraction
2. MEIRLOP analysis of DHS

1.3.5.1 DHS sequence extraction, scoring, and covariates

DHS for K562 cells were taken from ENCODE DNase-seq experiment accession ENCSR000EOT [43,44]. We used the narrowPeak BED file ENCFF821KDJ for the BED intervals [43,44]. DHS were then resized to +/− 75 bp around the DNase-seq peak center. To score DHS by DNase-coverage scores as a covariate, we converted the ENCODE DNase-seq bam file ENCFF156LGK into a coverage bigwig using deeptools bamCoverage, then ran deeptools multiBigwigSummary to obtain the mean values across the DHS [43,44,48]. To score DHS by histone-ChIP-seq, we downloaded alignments for H3K27ac, H3K27me3, H3K4me3, and H3K4me1 ChIP-seq from ENCODE, corresponding to ENCODE biosamples ENCBS639AAA (isogenic replicate 1) and ENCBS674MPN (isogenic replicate 2) [43,44]. Details on these downloaded files are available in supplementary Table 1.S4. We obtained bigwigs corresponding to log2 ratios of histone ChIP-seq coverage (H3K27ac over H3K27me3, and H3K4me3 over H3Kme1) for each replicate using bamCompare [48]. Then, we ran bigwigCompare to obtain a single bigwig for each log2 ratio summarizing the mean across both replicates. Finally, multiBigwigSummary was used to obtain the mean values for these log2 ratios across genomic ranges +/− 500 bp of the center for each DHS. DHS were assigned scores from the mean histone log2 ratios corresponding to their centers.

1.3.5.2 MEIRLOP analysis of DHS

MEIRLOP was run to analyze these scored DHS sequences with covariates controlling for dinucleotide frequencies and DNase-seq coverage for each DHS. MEIRLOP was invoked multiple times

for each histone ratio. E.g., for sequences scored by log2 ratio of H3K27ac over H3K27me3, while controlling for dinucleotide frequencies and DNase-seq signal as a covariate we used the command line:

> "OMP_NUM_THREADS=45 $(which time) --verbose python -m meirlop --jobs 45 --kmer 2 --covariates dhs.dnase.covariates.tsv --pcount 0.001 --pval 0.001 --html --scan --sortabs --fa dhs.h3k27ac_over_h3k27me3_mean.scored.fa jaspar.txt {h3k27ac_over_h3k27me3_meirlop_output_directory}".

A "walkthrough" Jupyter notebook for this analysis, starting from ENCODE downloads, is available on the MEIRLOP Github repository. To perform the analysis without the custom DNase-seq signal covariate, we ran a similar command but without the "--covariates" argument.

1.4 Results

1.4.1 MEIRLOP uses covariates to accurately call enrichment of relevant TF motifs

To show the utility of our method, we performed a differential motif enrichment analysis of regulatory elements modulated by interferon beta (IFN-β) treatment in HCT116 cells as measured by H3K27ac ChIP-seq. IFN-β treatment stimulates the type I interferon pathway, which ultimately leads to the activation of STAT and IRF family transcription factors to stimulate the expression of genes with antiviral activity [49]. While the type I interferon pathway is responsive in most cell types, HCT116 and other cell lines often exhibit a limited response to IFN-β treatment relative to primary myeloid cells [38], making their analysis more challenging. MACS2 found initial H3K27ac peaksets in each H3K27ac replicate and condition relative to input; then diffBind found 20,087 total H3K27ac peaks of median length 1kbp in HCT116 cells after controlling for coverage in an input experiment, and merging adjacent peaks across IFN-β- and mock-treated (control) conditions. Using diffBind's DESeq2 analysis to find differentially acetylated peaks, we found only 2 peaks with significantly increased H3K27ac signal after stimulation and

accounting for library size and replicates (adjusted p-value < 0.05, log 2 fold-change > 1.0) (Fig. 1.S1). As a result, set-based MEA on this small set of 2 peaks failed to return results for two known IFN-β response relevant motifs.

MEIRLOP uses logistic regression and does not rely on thresholding of regulatory region sequences into sets. We extracted the central 1K bp sequence of each peak and scored the response of each regulatory region by the log2 fold-change of the H3K27ac signal with and without IFN-β stimulation. PCA yielded 11 covariates summarizing 99% of the variance in dinucleotide frequencies across these sequences. When accounting for PCA-based covariates derived from dinucleotide frequency, our method can identify significant enrichment of known relevant motifs in the top 10 results: These are the motifs for IRF9 and the STAT1::STAT2 heterodimer, which together bind to interferon-sensitive response elements to activate pathways in the innate immune response [49]. Upon ranking the enrichment scores of these two relevant binding motifs compared to 577 others present in the motif database, we find that the IRF9 motif achieves a percentile ranking of 99.6 (rank 577 out of 579, just below similar IRF motifs), while the STAT1::STAT2 motif achieves a percentile ranking of 97.4 (rank 564 out of 579) (Fig. 1.2A). Thus, our method can call enrichments of relevant TF binding motifs in differential ChIP-seq analysis, even when the difference between conditions is subtle and difficult to quantify with conventional approaches such as diffBind [50].

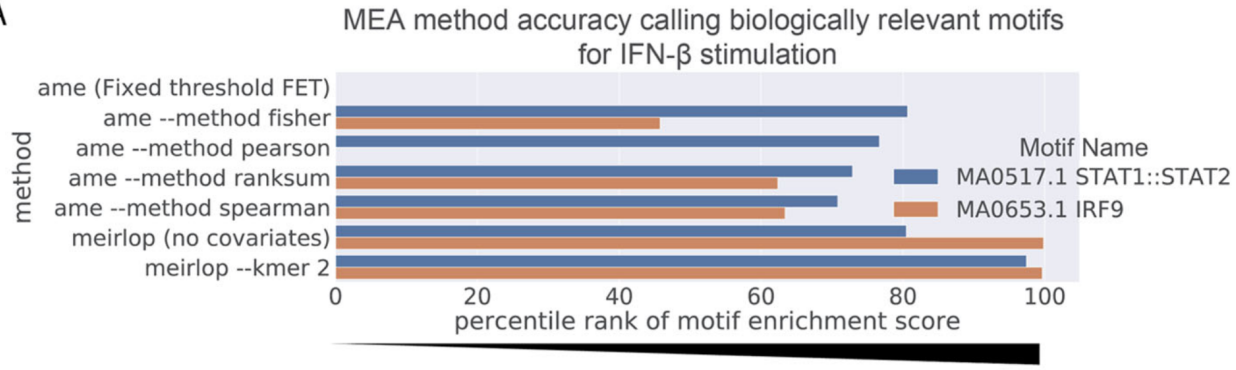
Figure 1.2: Logistic regression with covariates recovers motifs mediating IFN-β stimulation response in differentially acetylated ChIP-seq peaks.

(A) Ranks of the motif enrichment scores for motifs bound by TFs activated by IFN-β across score-based MEA methods. Higher percentile ranks indicate the motif considered more significantly enriched relative to other motifs in the results of each method. Motifs excluded from enrichment reports are assigned a rank of 0.

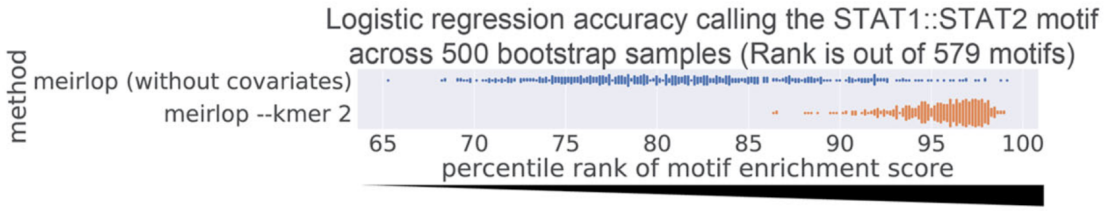
(B) Reproducibility of logistic regression accuracy of MEIRLOP when calling enrichment of the STAT1::STAT2 binding motif, with and without controlling for dinucleotide frequency covariates. Each point represents enrichment results on a bootstrap sample of the same dataset.

(C) Swarm plot depicting the distributions of score ranks for peaks with and without STAT1::STAT2 binding motif instances. Each point represents a scored peak.

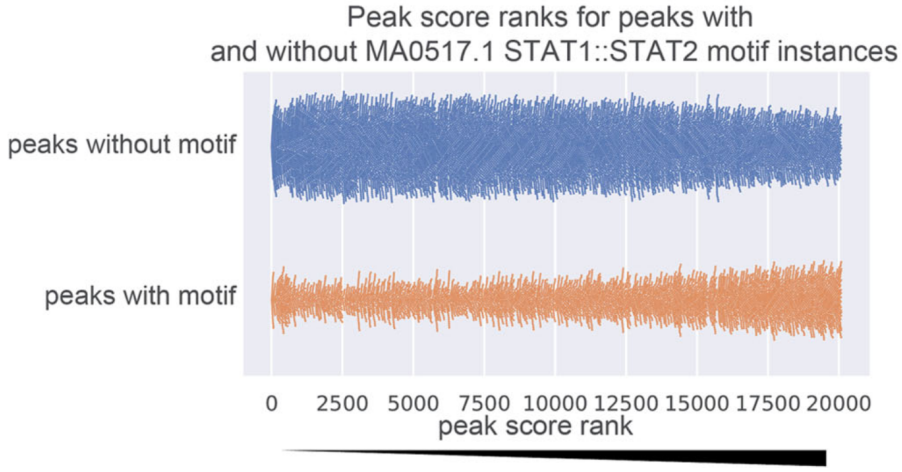
A


MEA method accuracy calling biologically relevant motifs
for IFN-β stimulation
method
ame (Fixed threshold FET)
ame --method fisher
ame --method pearson
ame --method ranksum
ame --method spearman
meirlop (no covariates)
meirlop --kmer 2
Motif Name
MA0517.1 STAT1::STAT2
MA0653.1 IRF9
0
20
40
60
80
100
percentile rank of motif enrichment score

B


Logistic regression accuracy calling the STAT1::STAT2 motif
across 500 bootstrap samples (Rank is out of 579 motifs)
method
meirlop (without covariates)
meirlop --kmer 2
65
70
75
80
85
90
95
100
percentile rank of motif enrichment score

C


Peak score ranks for peaks with
and without MA0517.1 STAT1::STAT2 motif instances
peaks without motif
peaks with motif
0
2500
5000
7500
10000
12500
15000
17500
20000
peak score rank

To establish the effect of sequence bias on the analysis, we repeated the enrichment without controlling for covariates derived from dinucleotide frequencies. While it slightly improved the enrichment for the IRF9 motif (percentile rank 99.8; rank 578 out of 579), the STAT1::STAT2 motif no longer appeared in the top ten enrichment results (percentile rank 80.5; rank 466 out of 579) (Fig. 1.2a, Fig. 1.S2A). Instead, more AT-rich homeobox transcription factors appeared ahead of the STAT1::STAT2 motif, comprising most of the top results (Fig. 1.S2A).

To determine whether the covariate-dependent difference in enrichment of the STAT1::STAT2 motif was because of chance, we created 500 bootstrap samples of 20,087 scored regulatory region sequences by sampling the original scored sequences with replacement, and performed motif enrichment on these with and without covariates. We found that the change in the rank of the STAT1::STAT2 motif enrichment was significant (Wilcoxon p-value 1.33e-83), with an average rank of 553.7 with covariates, and 471.5 without (Fig. 1.2B). When performing a limited enrichment controlling only for GC content, we found that GC content negatively correlated with the probability of a regulatory region sequence containing the motif (logistic regression coefficient -0.90, Wald test p-value < 0.01), consistent with the GC-poor composition of the motifs that appear ahead of the STAT1::STAT2 motif in the enrichment results. Despite the effect of GC composition, STAT1::STAT2 appears more frequently in the sequences of higher ranked peaks than in lower ranked peaks (Fig. 1.2C). Thus, controlling for sequence bias allows our logistic regression-based method to accurately call the enrichment of TF motifs relevant to IFN-β stimulation.

1.4.2 MEIRLOP outperforms other score-based MEA in differential ChIP-seq analysis

To evaluate the performance of our method relative to multiple approaches to score-based MEA, we applied multiple methods implemented in AME to the analysis of the 20,087 regulatory region sequences scored for IFN-β response. We found that most of the top enriched motifs found by these

methods were homeobox transcription factors, consistent with the results of MEIRLOP when not controlling for sequence bias (Fig. 1.2A, 1.S2A,B). IRF9 and STAT1::STAT2 motifs never obtained enrichment scores past percentile rank 86.3 (rank 500 out of 579) (Fig. 1.S2C). Thus, multiple score-based motif enrichment methods that do not control for sequence bias cannot accurately call the enrichment of relevant TFs in our differential ChIP-seq analysis in their top results.

1.4.3 MEIRLOP achieves similar or better accuracy on TF ChIP-seq data

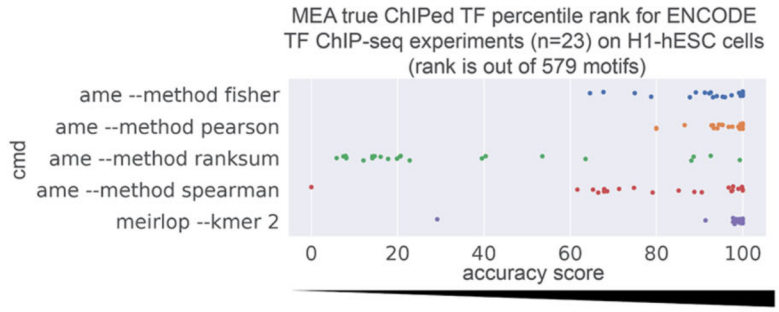
To determine the relative accuracy of our method compared to more commonly used MEA implementations across a wide range of experiments, we analyzed scored genomic regions from a set of 23 ENCODE TF ChIP-seq experiments on H1 human embryonic stem cells (H1-hESCs) [43,44] that analyzed TFs with known motifs in the JASPAR 2018 motif database [15]. From these scored regions, we extracted 200 bp sequences from the center of the peak for motif enrichment testing, yielding approximately 300,000 scored sequences per experiment. To evaluate the accuracy of similarly implemented MEA methods, we compared our method's accuracy against the accuracy of AME's implementations on the same datasets of scored regions. Similar to the differential ChIP-seq analysis, the evaluation metric was the rank of the enrichment score of the known motif for the ChIPed TF.

Our method (meirlop --kmer 2) achieved an average percentile rank of 95.9 (rank 555.2 out of 579) when recovering the motif corresponding to the true ChIPed TF (Fig. 1.3A). Scored MEA using Pearson correlation coefficients achieved an average percentile rank of 95.4 (rank 552.5 out of 579), not significantly different from our method's performance (Wilcoxon p-value = 0.09). Other MEA methods achieved lower average ranks (Fig. 1.3A). Thus, logistic regression can achieve accuracy similar to or better than existing score-based MEA methods on scored sequence data from TF ChIP-seq experiments. However, MEIRLOP is multithreaded and takes less time to achieve similar accuracy compared to other methods, taking an average of 10.8 min per experiment, while the Pearson correlation method in AME took 21.9 h on average.

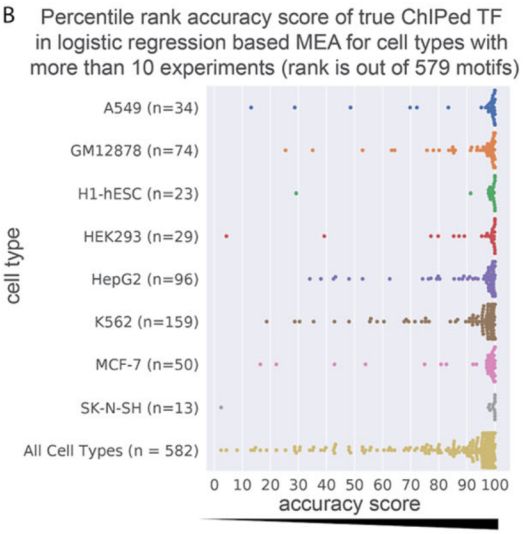
Figure 1.3: Logistic regression with covariates accurately calls motifs for ChIPed TFs in ENCODE experiments.

(A) Plot of the rank for the enrichment score of the ChIPed TF's motif in 23 ENCODE H1-hESC ChIP-seq experiments, across multiple score-based MEA.
(B) Plot of the rank for the enrichment score of the ChIPed TF's motif across multiple ENCODE cell lines. Only cell lines with more than 10 experiments analyzed are represented, as well as a separate category summarizing all cell lines.
(C) Bar plot of Fisher exact test significance for enrichment of ENCODE warning flags among 36 (out of 582 eligible) experiments for which logistic regression with covariates failed to call the true ChIP-ed TF as significant, showing partial characterization of ChIP-seq antibodies as a significantly enriched warning term (Fisher exact test p-value = 6.8e-5, FDR = 1.563e-3)

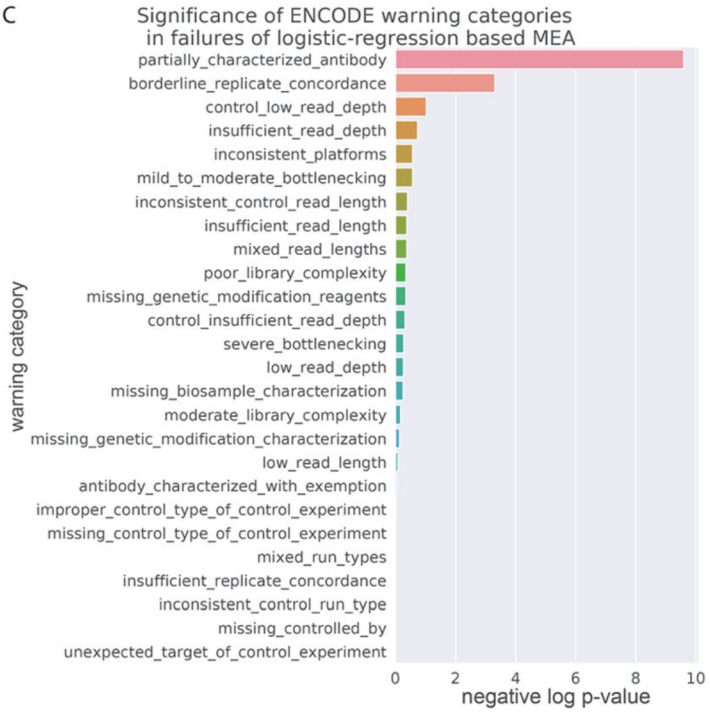
A


MEA true ChIPed TF percentile rank for ENCODE TF ChIP-seq experiments (n=23) on H1-hESC cells (rank is out of 579 motifs)
cmd
ame --method fisher
ame --method pearson
ame --method ranksum
ame --method spearman
meirlop --kmer 2
0
20
40
60
80
100
accuracy score

B


Percentile rank accuracy score of true ChIPed TF in logistic regression based MEA for cell types with more than 10 experiments (rank is out of 579 motifs)
cell type
A549 (n=34)
GM12878 (n=74)
H1-hESC (n=23)
HEK293 (n=29)
HepG2 (n=96)
K562 (n=159)
MCF-7 (n=50)
SK-N-SH (n=13)
All Cell Types (n = 582)
0 10 20 30 40 50 60 70 80 90 100
accuracy score

C


Significance of ENCODE warning categories in failures of logistic-regression based MEA
warning category
partially_characterized_antibody
borderline_replicate_concordance
control_low_read_depth
insufficient_read_depth
inconsistent_platforms
mild_to_moderate_bottlenecking
inconsistent_control_read_length
insufficient_read_length
mixed_read_lengths
poor_library_complexity
missing_genetic_modification_reagents
control_insufficient_read_depth
severe_bottlenecking
low_read_depth
missing_biosample_characterization
moderate_library_complexity
missing_genetic_modification_characterization
low_read_length
antibody_characterized_with_exemption
improper_control_type_of_control_experiment
missing_control_type_of_control_experiment
mixed_run_types
insufficient_replicate_concordance
inconsistent_control_run_type
missing_controlled_by
unexpected_target_of_control_experiment
0
2
4
6
8
10
negative log p-value

To determine the performance of our method on TF ChIP-seq for other cell types, we used MEIRLOP to measure the enrichment of motifs in scored ChIP-seq peaks from 582 ENCODE ChIP-seq TF binding experiments on human cell lines. On average, our method achieved a percentile rank of 93.3 (rank 540.1 out of 579), placing the motif for the ChIPed TF towards the top 90% of enrichment results (Fig. 1.3B).

Our method failed to call significant enrichment of a motif for the true ChIPed TF in 36 of 582 experiments. To diagnose these failure cases, we retrieved ENCODE audit warning flags for each of the 582 experiments. From these warning flags, 26 of the 582 experiments had the “partially characterized antibodies” warning. Of the 36 experiments that our method failed, 8 had this warning flag present and enriched (Fisher exact test p-value = 6.8e-5), with the enrichment remaining significant after multiple testing of 22 other warning flags (adjusted p-value = 1.767e-3) (Fig. 1.3C). Thus, the antibodies used in these ChIP-seq experiments may not have been specifically targeting the proper TFs, leading to a failure to identify the expected known motif.

1.4.4 MEIRLOP identifies motifs of TFs mediating KLA response from csRNA-seq

Score-based MEA is well suited to analyze changes in regulatory states by incorporating the magnitude of transcriptional change into the motif enrichment calculation. To demonstrate our method’s efficacy when analyzing changes in transcription with diverse data types, we applied MEIRLOP to the analysis of capped short (cs)RNA-seq data generated in macrophages activated by TLR4-agonist KLA for 1 h or mock-treated controls. csRNA-seq is an approach that isolates initiating transcripts at both promoters and enhancers to directly assess the transcriptional activity of regulatory elements genome-wide [10]. We analyzed 90,857 transcription start sites (TSS) from csRNA-seq profiling of murine bone-marrow derived macrophages (BMDMs) activated by Kdo2-lipid A (KLA), as previously analyzed in Duttke et al. [10]. We searched for motifs within +/− 150 bp of the TSS, and scored them by their log2 fold change of csRNA-seq signal between KLA stimulation and control. MEIRLOP readily identified enrichment of motifs for the NF-κB and JUN/AP1 TF families in KLA-induced TSS (Fig. 1.4) [10]. These

TFs are known to mediate KLA response [51]. In KLA-repressed TSS, MEIRLOP identified enrichment of motifs for PU.1/SPI1 (ETS), MITF (bHLH), and the MEF2 TF family (Fig. 1.4). These findings are consistent with previous literature indicating lipopolysaccharide (LPS) -induced downregulation of these TFs [52,53].

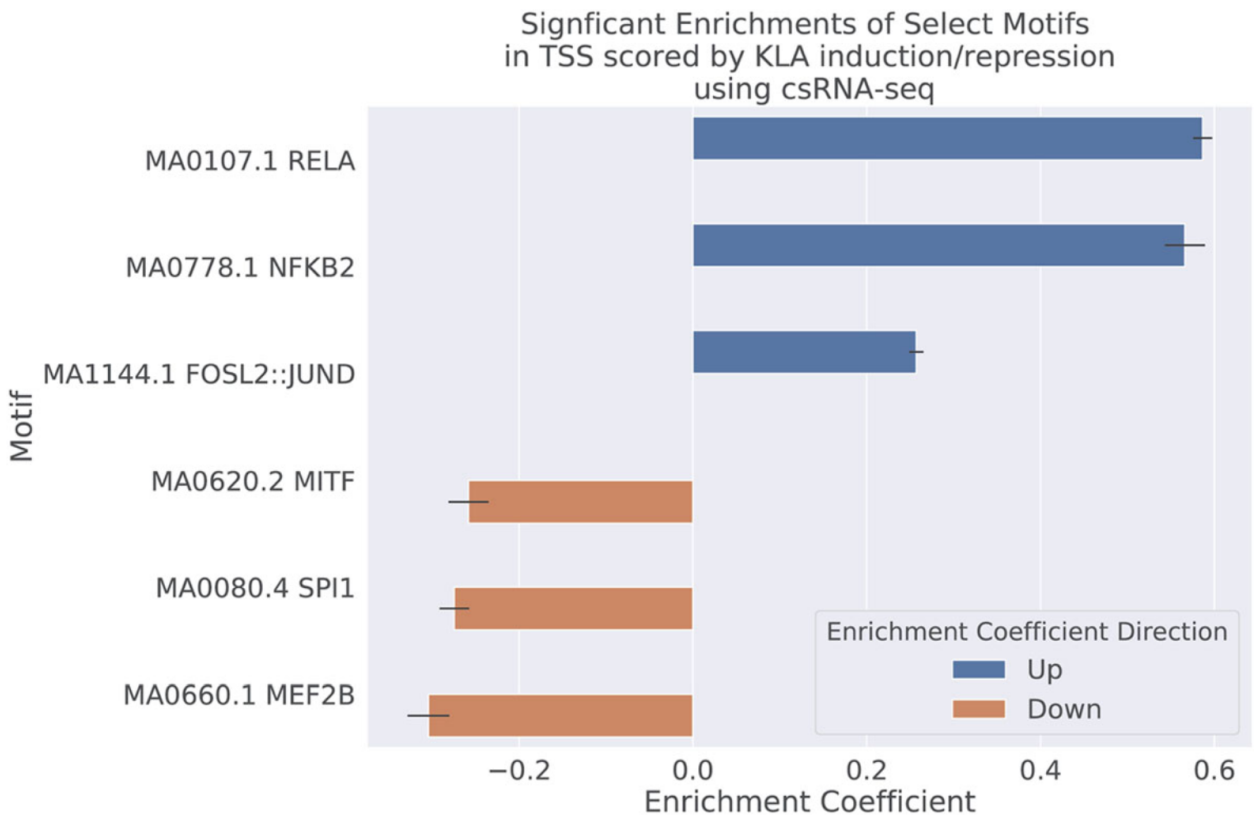


Figure 1.4: Logistic regression with covariates finds motifs for TFs mediating KLA response.

Bar plot of significant logistic regression coefficients for select motifs on csRNA-seq TSS scored by differential nascent transcription signal between KLA stimulated and control conditions

1.4.5 MEIRLOP identifies enriched TF binding motifs in DNase I hypersensitive sites

To further demonstrate the flexibility conferred by scored-based MEA, we used MEIRLOP to assign regulatory functions to TF motifs found enriched in DNase I Hypersensitive sites (DHS). Extensive chromatin profiling from ENCODE provides key information about the regulatory states at each DHS, although the levels of each chromatin modification are often correlated with the DNase hypersensitivity signal, which reflects the relative fraction of cells with open chromatin at that site in the population of cells [43,44]. To identify DHS that are associated with specific epigenetic profiles independent of hypersensitivity levels, we computed composite scores across 235,220 DHS in K562 cells based on the log2 ratios of ChIP-seq coverage for the following scores: H3K27ac (active promoters and enhancers) over H3K27me3 (polycomb, repressive), reflecting active vs. repressed regulatory regions [16]; And H3K4me3 (promoters) over H3K4me1 (enhancers), reflecting DHS with promoter vs. enhancer characteristics [54]. Composite scores were derived from histone ChIP-seq coverage +/− 500 bp from the peak centers. To account for correlation of these scores with DNase hypersensitivity signal, we incorporated DNase-seq coverage +/− 75 bp from the peak centers as a covariate. We searched for motifs +/− 75 bp from the peak centers.

MEIRLOP analysis of the H3K4me3/H3K4me1 composite ratio identified motifs for several TFs with known roles in promoters or enhancers. MEIRLOP found the motifs for ELF1, NRF1, and YY1 in the top 10 significant enrichments for more promoter-like DHS with a greater ratio of H3K4me3 over H3K4me1 (Fig. 1.5a). This finding has precedence in a larger study by Anderson et al., which also found motifs for these TFs significantly enriched in promoters compared to enhancers [55].The transcription factor ZNF143 also acts as a promoter-bound transcriptional activator, suggested to bind next to POL2 [56,57].

1.4.6 MEIRLOP identifies enriched TF binding motifs in DNase I hypersensitive sites

To further demonstrate the flexibility conferred by scored-based MEA, we used MEIRLOP to assign regulatory functions to TF motifs found enriched in DNase I Hypersensitive sites (DHS). Extensive chromatin profiling from ENCODE provides key information about the regulatory states at each DHS, although the levels of each chromatin modification are often correlated with the DNase hypersensitivity signal, which reflects the relative fraction of cells with open chromatin at that site in the population of cells [43,44]. To identify DHS that are associated with specific epigenetic profiles independent of hypersensitivity levels, we computed composite scores across 235,220 DHS in K562 cells based on the log2 ratios of ChIP-seq coverage for the following scores: H3K27ac (active promoters and enhancers) over H3K27me3 (polycomb, repressive), reflecting active vs. repressed regulatory regions [16]; And H3K4me3 (promoters) over H3K4me1 (enhancers), reflecting DHS with promoter vs. enhancer characteristics [54]. Composite scores were derived from histone ChIP-seq coverage +/− 500 bp from the peak centers. To account for correlation of these scores with DNase hypersensitivity signal, we incorporated DNase-seq coverage +/− 75 bp from the peak centers as a covariate. We searched for motifs +/− 75 bp from the peak centers.

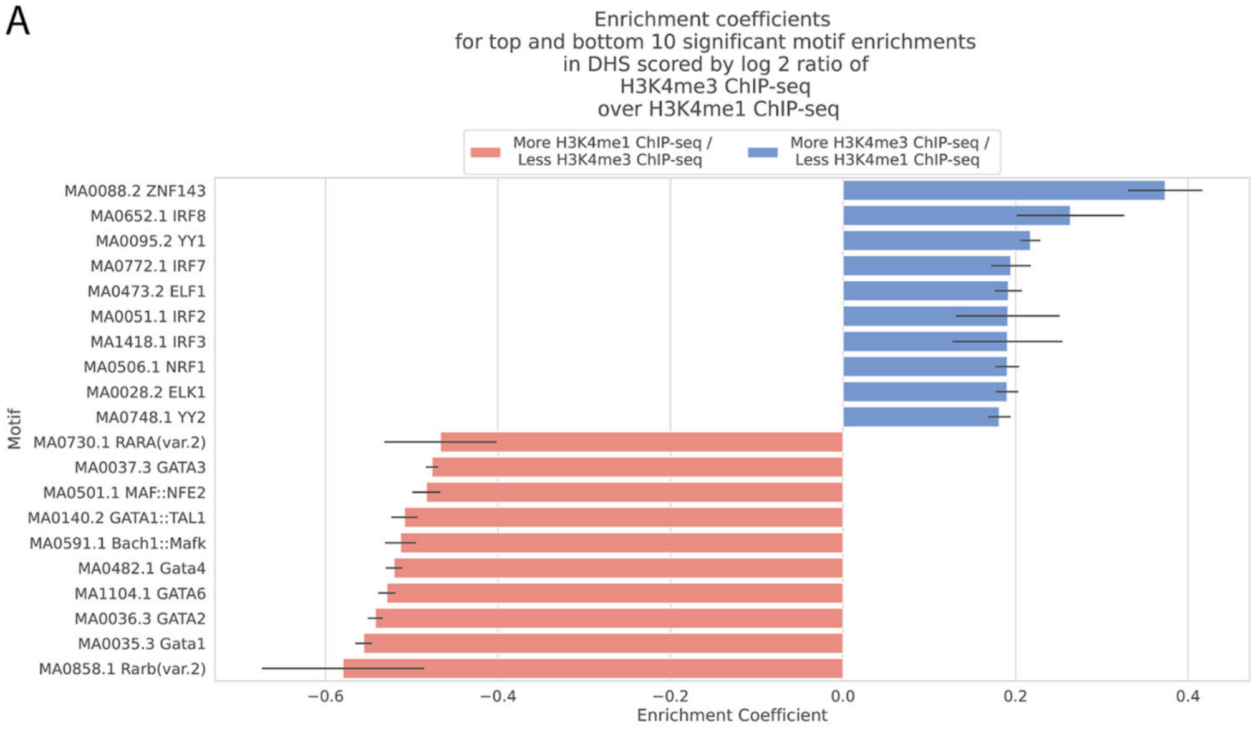
MEIRLOP analysis of the H3K4me3/H3K4me1 composite ratio identified motifs for several TFs with known roles in promoters or enhancers. MEIRLOP found the motifs for ELF1, NRF1, and YY1 in the top 10 significant enrichments for more promoter-like DHS with a greater ratio of H3K4me3 over H3K4me1 (Fig. 1.5A). This finding has precedence in a larger study by Anderson et al., which also found motifs for these TFs significantly enriched in promoters compared to enhancers [55].The transcription factor ZNF143 also acts as a promoter-bound transcriptional activator, suggested to bind next to POL2 [56,57].

Figure 1.5: Logistic regression with covariates finds motifs for TFs associated with the ratios of different histone modifications over DHS.

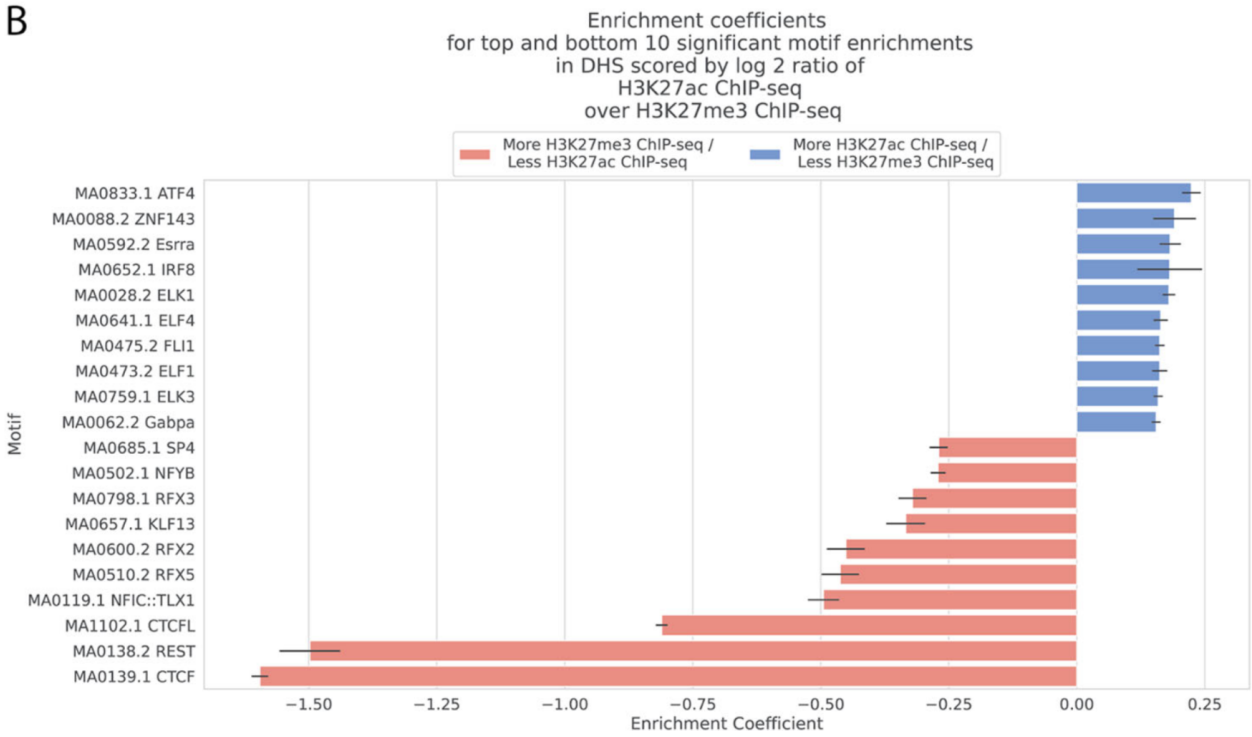
(A) Bar plot of enrichment coefficients for the top and bottom 10 most significantly enriched motifs associated with DHS scored by the $\log_2$ ratio of H3K4me3 over H3K4me1. Error bars represent the standard deviation of the enrichment coefficients.

(B) Bar plot of enrichment coefficients for the top and bottom 10 most significantly enriched motifs associated with DHS scored by the log2 ratio of H3K27ac over H3K27me3. Error bars represent the standard deviation of the enrichment coefficients

A


Enrichment coefficients
for top and bottom 10 significant motif enrichments
in DHS scored by log 2 ratio of
H3K4me3 ChIP-seq
over H3K4me1 ChIP-seq
More H3K4me1 ChIP-seq / Less H3K4me3 ChIP-seq
More H3K4me3 ChIP-seq / Less H3K4me1 ChIP-seq
MA0088.2 ZNF143
MA0652.1 IRF8
MA0095.2 YY1
MA0772.1 IRF7
MA0473.2 ELF1
MA0051.1 IRF2
MA1418.1 IRF3
MA0506.1 NRF1
MA0028.2 ELK1
MA0748.1 YY2
MA0730.1 RARA(var.2)
MA0037.3 GATA3
MA0501.1 MAF::NFE2
MA0140.2 GATA1::TAL1
MA0591.1 Bach1::Mafk
MA0482.1 Gata4
MA1104.1 GATA6
MA0036.3 GATA2
MA0035.3 Gata1
MA0858.1 Rarb(var.2)
Motif
−0.6
−0.4
−0.2
0.0
0.2
0.4
Enrichment Coefficient

B


Enrichment coefficients
for top and bottom 10 significant motif enrichments
in DHS scored by log 2 ratio of
H3K27ac ChIP-seq
over H3K27me3 ChIP-seq
More H3K27me3 ChIP-seq / Less H3K27ac ChIP-seq
More H3K27ac ChIP-seq / Less H3K27me3 ChIP-seq
MA0833.1 ATF4
MA0088.2 ZNF143
MA0592.2 Esrra
MA0652.1 IRF8
MA0028.2 ELK1
MA0641.1 ELF4
MA0475.2 FLI1
MA0473.2 ELF1
MA0759.1 ELK3
MA0062.2 Gabpa
MA0685.1 SP4
MA0502.1 NFYB
MA0798.1 RFX3
MA0657.1 KLF13
MA0600.2 RFX2
MA0510.2 RFX5
MA0119.1 NFIC::TLX1
MA1102.1 CTCFL
MA0138.2 REST
MA0139.1 CTCF
Motif
−1.50
−1.25
−1.00
−0.75
−0.50
−0.25
0.00
0.25
Enrichment Coefficient

In contrast, different sets of motifs were identified in enhancer-like DHS. Rye et al. previously found that NFE2 preferentially mapped to genomic regions marked with H3K4me1 compared to those marked with H3K4me3 [58]. DNA binding activity of GATA1 is correlated with H3K4me1, a histone marker for enhancers [55,59]. However, the enhancer activity of DNA bound by GATA1 is associated with TAL1 co-binding [60,61]. In keeping with these combined roles, MEIRLOP found a motif associated with binding of both TFs significantly enriched in DHS with greater ratios of H3K4me1 over H3K4me3 signal (enrichment coefficient = − 0.51, adjusted p-value < 0.01) (Fig. 1.5A). Significant enrichments for other GATA1-like motifs were also found enriched in DHS with greater H3K4me1 over H3K4me3 signal (Fig. 1.5A).

Analysis of H3K27ac/H3K27me3 identifies the motifs of TFs that may contribute to highly active (high ratio) or repressed (low ratio) regions of chromatin. The high score for the ATF4 motif at active regions is consistent with the evidence that ATF4 recruits histone acetyltransferase [62]. At the other extreme, MEIRLOP found a binding motifs for CTCF and REST significantly enriched in more repressed DHS with higher levels of H3K27me3 over H3K27ac (enrichment coefficients ~ − 1.5, adjusted p-value < 0.01) (Fig. 1.5B). CTCF is an important factor in establishing 3D chromatin structure, while REST is a TF known to silence gene expression through chromatin remodeling [63,64].

KLF13 (also known as BTEB3) represses transcription through interaction with histone deacetylases (HDACs) and competes with Sp1 for DNA-binding [65]. Notably, while MEIRLOP finds enrichment for a KLF13 binding motif associated with transcriptionally silenced chromatin (enrichment coefficient = − 0.33, adjusted p-value <1e-8) (Fig. 1.5B), it requires the additional DNase-seq signal covariate to do so: Running MEIRLOP with only dinucleotide-based covariates does not find a significant enrichment of this motif (enrichment coefficient − 4.8e-2, adjusted p-value = 0.18). This change in enrichment values is consistent with the moderate correlation of DNase-seq coverage and H3K27ac over H3K27me3 ratios across DHS (Spearman correlation coefficient = 0.51).

Overall, MEIRLOP found enrichments for TF motifs associated with key functional processes by analyzing DHS scored as a function of multiple sequencing assays, while also controlling for dinucleotide sequence bias and the quality of the DHS as a function of DNase-seq coverage. These enrichments are consistent with previous findings in the literature and known roles of the TFs binding to these motifs.

1.5 Discussion

By incorporating covariates derived from low-level sequence bias (GC content, dinucleotide frequencies), MEIRLOP enables accurate score-based MEA that achieves accurate enrichment results. These results recapitulate the transcription factors involved in regulating transcriptional response to IFN-β stimulation. When applied to ENCODE TF ChIP-seq data, we find that our method performs just as well or better compared to multiple score-based MEA methods in recovering significant enrichment of the binding motif for the ChIPed TF. Combined with the faster execution of our implementation, this allows the application of our method in situations where researchers must analyze thousands of scored regulatory regions for motifs and TFs that can explain transcriptional regulation in differing conditions. Using a logistic regression model means that our method can also be extended to integrate covariates other than sequence composition, allowing researchers to control for variation in sequence length, batch effects, or systematic coverage biases found in control experiments.

Set-based MEA can remain applicable when the biological signal quantified across conditions for regulatory regions is bimodal, or when the signal can be thresholded into sizable sets of sequences. However, in scenarios where the list of significantly differentially regulated regions is small or the magnitude of regulation is weak, we demonstrated that score-based MEA can be sensitive enough to properly identify differential motif enrichment. But even where other score-based MEA methods may apply, we found that uneven distribution of GC content in the sequences analyzed obscured known biology from being recovered: If a researcher took results from methods that did not account for covariates at face value, then absent other evidence, they might conclude that homeobox transcription factors (with their GC-poor motifs) are more significant to transcriptional regulation of IFN-β response compared to interferon regulatory factors or STAT transcription factors.

We find that our method achieves accuracy similar to existing regression-based MEA methods on non-differential TF ChIP-seq data. TF ChIP-seq data quantifies the binding affinity of TFs to genomic regions, and so directly mirrors the assumption in AME's regression-based methods that stronger binding motifs correlate with stronger binding. But even when applied to data that better reflects AME's regression model, MEIRLOP remains competitive in both accuracy and runtime. Runtime improvements may result from: The use of a faster motif scanning algorithm provided by MOODS [32]; and the use of multithreading. However, the generalizability of the method may stem from its hybridization of set-based and score-based approaches: MEIRLOP treats motif presence as a Boolean variable and regulatory region activity as a continuous variable, rather than treating both as Boolean variables (as per a Fisher exact test) or both as continuous variables (as per regression). Since MEIRLOP takes arbitrary scores for regulatory region activity, it is able to analyze motif enrichments from novel sequencing methods, such as the recently developed csRNA-seq. However, reducing motif presence to a Boolean outcome variable can lead to loss of accuracy where motif matches may be inexact, e.g. where a modified TF does not bind as well to its canonical motif. Here, de novo motif finding methods can recover the true binding motif, then MEIRLOP can determine enrichment of this motif, as performed in Bloodgood et al. [66].

We demonstrated the flexibility of MEIRLOP in situations where the genomic regions quantified were scored using experimental assays that were different from those used to identify them: We found motif enrichments within DHS scored using composite measurements, as log2 ratios of histone ChIP-seq coverage across the DHS. MEIRLOP found multiple enriched motifs consistent with previous findings from the literature and with known roles of TFs corresponding to those motifs. This illustrates that MEIRLOP can determine motif enrichments along diverse scores of regulatory region activity, even where these scores may not have intuitive pre-defined thresholds due to their composite nature or relation to the genomic regions being analyzed. Furthermore, because MEIRLOP offers two-sided hypothesis testing, it enables researchers to investigate motifs enriched towards either extreme of such ratios in a single pass, instead of having to run a motif enrichment analysis tool twice to investigate both extremes. MEIRLOP's ability to incorporate customized covariates also offers researchers further power to detect enrichments

that are otherwise obscured by factors such as the quality of a DHS site, while also accounting for sequence bias. The combined flexibility from score-based motif enrichment and covariate control also serves MEIRLOP well when analyzing data from more recent sequencing strategies, including csRNA-seq. These examples indicate how MEIRLOP can serve researchers when performing motif-based analyses on data from newer techniques, where rule-of-thumb thresholds for set-based motif enrichment may not be fully established, enabling new findings.

1.6 Conclusions

We have demonstrated that MEIRLOP can determine enrichment of sequence motifs where sequence bias or other covariates may confound other methods. To use MEIRLOP, researchers need only score genomic regions across a continuum of biological interest. These scores are flexible, and MEIRLOP's two-sided enrichment can identify motifs enriched towards either extreme of log-ratios assembled from multiple sequencing experiments. Although MEIRLOP is confined to scoring the enrichment of a known motif library, future work will explore the de novo identification of motifs that maximize for enrichment based on MEIRLOP.

1.7 Availability and requirements

- Project name: MEIRLOP.
- Project home page: https://github.com/npdeloss/meirlop
- Operating system(s): Linux.
- Programming language: Python 3.
- Other requirements: Conda for installation and dependency management.
- License: MIT License.
- Any restrictions to use by non-academics: MIT License terms.

1.8 Availability of data and materials

The ENCODE TF ChIP-seq datasets analysed in this study are available in the ENCODE repository, at https://www.encodeproject.org/ (doi: https://doi.org/10.1038/nature11247.). Detailed listings, including experiment accession IDs, download URLs, and the names of labs where the data were generated, are available in supplementary Table 1.S3.

Other ENCODE DNase-seq and histone ChIP-seq datasets in this study are also available in the ENCODE repository, at https://www.encodeproject.org/ (doi: https://doi.org/10.1038/nature11247.). Detailed listings, including experiment accession IDs, download URLs, and the names of labs where the data were generated, are available in supplementary Table 1.S4.

JASPAR 2018 CORE vertebrate non-redundant motif set file used in this study is available from the JASPAR 2018 archive download page, at http://jaspar2018.genereg.net/downloads/ [15].

The differential stimulation H3K27ac datasets used and/or analysed during the current study are available under GEO accession GSE147707.

The code for MEIRLOP is available from its Github repository at https://github.com/npdeloss/meirlop, and can be installed through the Anaconda software distribution platform (for 64-bit Linux systems) with the command: “conda install -c bioconda -c conda-forge -c npdeloss meirlop” [28].

1.9 Publication Acknowledgement Statements

Sascha Duttke and Chiemerie Azubuogu provided significant editorial feedback for this manuscript.

1.9.1 Funding

This work was supported by the NIH (GM134366, PI: Benner), a NLM Training Grant (T15LM011271), and the Katzin Prize Endowed Fund for the work of NPDS. The funding bodies had no direct roles in the design or execution of the study.

1.9.2 Author information

Authors and Affiliations

Department of Biomedical Informatics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0640, USA

Nathaniel P. Delos Santos

Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0640, USA

Lorane Texari & Christopher Benner

1.9.3 Contributions

NPDS designed and implemented MEIRLOP, performed sequence analyses, and was a major contributor in writing the manuscript. LT designed and performed the differential H3K27ac ChIP-seq experiment whose data is used in this study, and contributed the methods described for generation of this dataset. NPDS and CB wrote the manuscript. The author(s) read and approved the final manuscript.
Corresponding author

Correspondence to Christopher Benner.

1.9.4 Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable. This work uses human cell line data from the ENCODE Project [43,44].

Competing interests

The authors declare that they have no competing interests.

1.9.5 Additional information

Publisher’s Note

1.9.6 Rights and permissions

(http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

1.9.7 Changes Made

This paper has been reformatted from its original print to satisfy dissertation formatting requirements and navigatability. Changes include header numbering, figure numbering, and reference formats, as well as font and spacing changes. Supplemental figures and tables have been moved to the Appendix, under section “A.1 Supplemental Material for Chapter 1”. Abbreviations have been moved to the preface pages of this dissertation.

1.10 Acknowledgements

Chapter 1, in full, is a modified reprint of the material as it appears in “MEIRLOP: improving score-based motif enrichment by incorporating sequence bias covariates” in BMC Bioinformatics, 2020. Delos Santos, Nathaniel P.; Texari, Lorane; Benner, Chistopher, Springer Nature, 2020. Modifications have been made to the text to ensure consistency with dissertation formatting. The dissertation author was the primary investigator and author of this paper.

Chapter 2. MEPP: More transparent motif enrichment by profiling positional correlations

2.1 Abstract

Score-based Motif Enrichment Analysis (MEA) is typically applied to regulatory DNA to infer transcription factors (TFs) that may modulate transcription and chromatin state in different conditions. Most MEA methods determine motif enrichment independent of motif position within a sequence, even when those sequences harbor anchor points that motifs and their bound TFs may functionally interact with in a distance-dependent fashion, such as other TF binding motifs, transcription start sites (TSS), sequencing assay cleavage sites, or other biologically meaningful features. We developed Motif Enrichment Positional Profiling (MEPP), a novel MEA method that outputs a positional enrichment profile of a given TF's binding motif relative to key anchor points (e.g. transcription start sites, or other motifs) within the analyzed sequences while accounting for lower-order nucleotide bias. Using transcription initiation and TF binding as test cases, we demonstrate MEPP's utility in determining the sequence positions where motif presence correlates with measures of biological activity, inferring positional dependencies of binding site function. We demonstrate how MEPP can be applied to interpretation and hypothesis generation from experiments that quantify transcription initiation, chromatin structure, or TF binding measurements. MEPP is available for download from https://github.com/npdeloss/mepp.

2.2 Introduction

Transcription factors (TFs) coordinate cellular transcriptional responses to external or changing signals [17]. Motif enrichment analysis (MEA) allows researchers to infer the TFs responsible for altering gene expression or chromatin state in response to internal or external stimuli. MEA achieves this through quantifying the enrichment of TF binding motifs in regulatory element sequences that exhibit a measurable biological response of interest, such as chromatin opening, histone modification, TF binding

or transcription. Methods such as ATAC-seq, ChIP-seq, or csRNA-seq quantify these responses and are widely used to study transcription regulation [10,67,68].

Most MEA methods analyze biological sequences for the simple presence or absence of a motif without regard to the motif's position within the sequence. However, the position of TF binding motifs can play important biological roles [69]. For example, some transcription factors play a role in directing the selection of TSS, or preventing ectopic TSS utilization [11]. The position of TF binding motifs relative to other motifs can also be important for establishing functional regulatory modules and TF co-binding, as reflected in regulatory motif grammars [12–14].

Recent sequencing advances allow the definition of TSS and TF binding sites at base resolution, thus enabling analysis of the functional aspects of motif positioning. PRO-cap or csRNA-seq assays reveal and quantify nascent transcription start sites, providing high-resolution transcription initiation data in both genomic and temporal axes [10,70]. Similarly, methods such as ChIP-exo and ChIP-nexus pinpoint TF binding locations at high resolution [67,71]. Other assays, such as ATAC-seq and MNase-seq, define cleavage sites in open chromatin or at nucleosome boundaries [68,72]. Proper analysis of these high-resolution measurements of biological or functional features can provide a more precise characterization of nearby motif positions and their regulatory functions.

There is a need for methods that visualize and quantify the spatial relationships between TF binding motifs and biologically relevant anchor points such as TSS. CentriMo [73] and TFEA [74] provide bin- and quartile- based approaches to analyzing these relationships, but there remains an unfulfilled need for a method that uses score information directly. To fulfill this need, we have developed Motif Enrichment Positional Profiles (MEPP). MEPP identifies sequence motifs enriched at positions relative to biologically meaningful features, thereby integrating position as an additional layer of information. This provides the user with knowledge about which motif positions are optimal for context-specific binding site functions, in the form of a positional profile. Because this profile correlates relative positions of predicted binding sites with biological function, it can narrow down a search for the most functional binding sites from hundreds of base pairs to a local neighborhood: This refinement directly addresses the concerns of Wasserman and Sandlin's futility theorem, which states that almost all predicted binding sites lack function [75].

Input for MEPP comprises scored genomic sequences of uniform length. Scores for these sequences can come from a biological readout (e.g. transcription level measured by a sequencing assay). To contextualize the position of motifs, the sequences should be centered on a biologically meaningful position, for example the location of TSSs, cleavage sites, other sequence motifs, or other meaningful features. Rather than calculating a singular enrichment score for a motif, like standard MEA methods, MEPP calculates a position-dependent enrichment profile for each motif. In this profile, highly positive values at a position correspond to a stronger positive correlation of the motif presence with the biological score assigned to each sequence. By contrast, more negative values at a position correspond to a stronger negative correlation with the biological score. The resulting profile thus reveals positions of motifs that are most likely to activate or repress the scored biological features. In addition, MEPP visualizes the distribution of the motif across the input dataset as a 2D heatmap of the motif's strength and presence across both motif positions and sequence ranks (based on the assigned biological score). These results help identify not just relevant motifs, but the positional constraints of motifs that delineate context or position-dependent function. The score-based principle of this enrichment method further avoids issues with arbitrary threshold selection, while controlling for sequence bias.

2.3 Materials and Methods

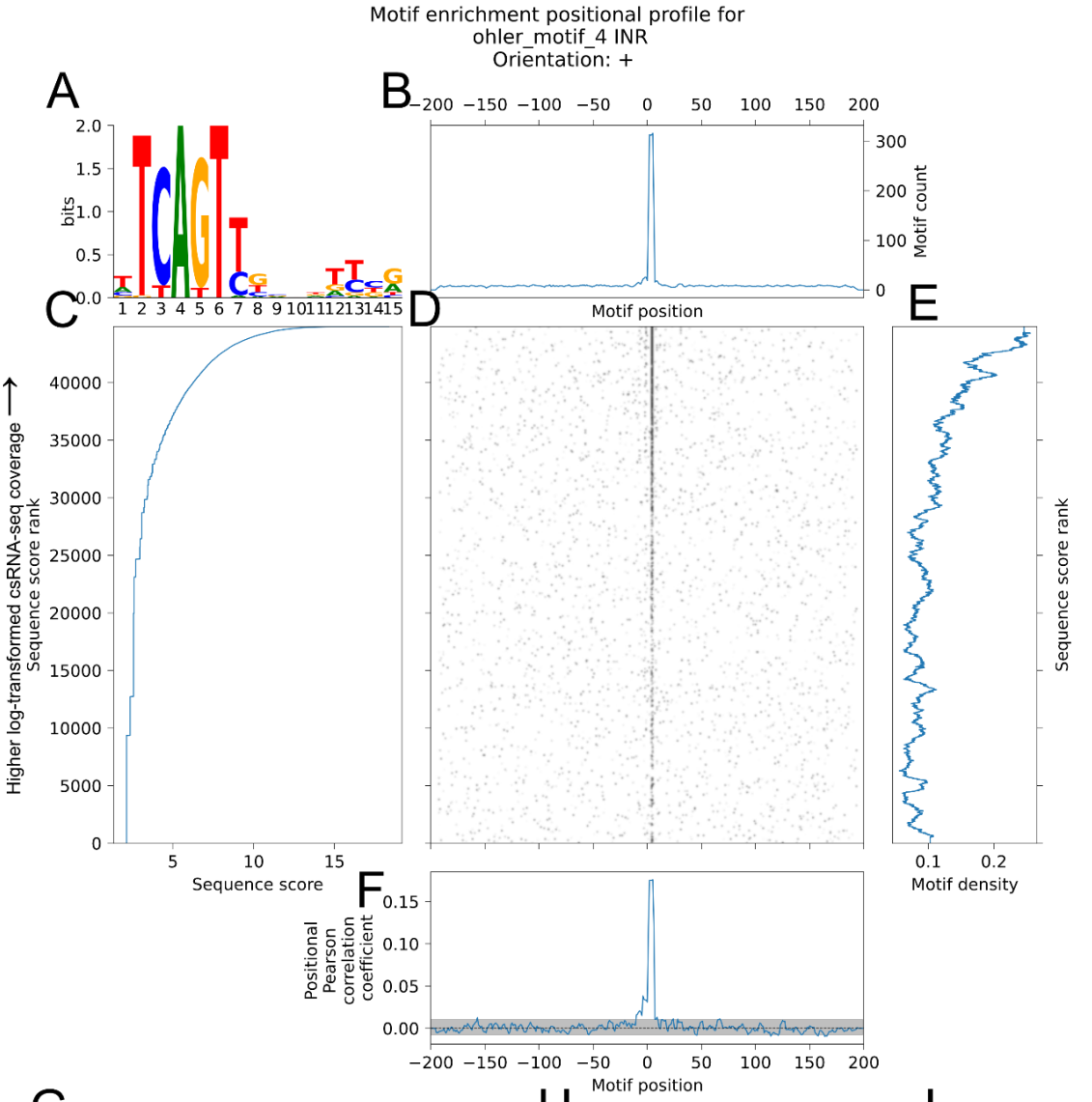
2.3.1 MEPP Implementation

In order to visualize and quantify local enrichment of motifs, we developed and employed Motif Enrichment Positional Profiles (MEPP). The typical execution of MEPP occurs in 5 parts:

1. Input data and pre-processing
2. Motif heatmap generation
3. Positional profile computation
4. Per-motif visualization
5. Motif dataset visualization

Figure 2.1: MEPP visualizes and quantifies core promoter motifs near Drosophila melanogaster transcription start sites.

(A) Motif logo for the Drosophila Initiator motif
(B) Visualization of smoothed motif counts over each position across the 400 bp sequence, centered on the TSS quantified by csRNA-seq.
(C) Line plot relating the sequence score (log-transformed csRNA-seq coverage) to the rank of the sequence score. Sequences are arranged in order of descending score in the dataset.
(D) 2D motif heatmap summarizing motif occurrences across the whole dataset, with the horizontal axis indicating motif position, and the vertical axis indicating the rank of the sequence score. Each black spot represents a motif occurrence, with darker spots for stronger/more motifs in a downsampled neighborhood.
(E) Line plot summarizing smoothed density of motifs across sequences in the dataset, with the vertical axis representing sequence score rank.
(F) Visualization of the partial Pearson correlation values of motif strength/presence with score, quantified at each possible motif position surrounding the TSS, after controlling for sequence GC content. A 95% confidence interval is shaded in gray.
(G) Partial screenshot of MEPP's interactive table output. Motifs are identifiable in MEPP's interactive table by motif ID and sequence logo.
(H) Motif positional profiles are summarized using a sparkline visualization, allowing exploration of profiles at a glance. Pictured are the Initiator motif, TATA Box motif, and DPE motif.
(I) Extremes (minima, maxima) of motif positional profiles are summarized, including the values and where they occur relative to the sequence center.



G H I

motif_id	logo	positional_r_profile_scaled_sparkline	extreme_r	extreme_r_pos
ohler_motif_4 INR			0.175835	5
ohler_motif_9 DPE			0.138242	25
ohler_motif_1 1			0.107824	-3
ohler_motif_3 TATA			0.097780	-28
ohler_motif_2 DRE			0.071272	-51
ohler_motif_10 10			0.068335	20

2.3.1.1 Input data and pre-processing

MEPP accepts input comprising a series of uniform-length scored DNA sequences in the scored FASTA file format, where the sequence score follows after the sequence header, separated by a space. The score for each sequence resembles the score column of a bed-file, with its meaning dependent on the assay in question. For example, when analyzing csRNA-seq, the user may assign the score to TSS usage/csRNA-seq signal, or the $\log_2$ fold change in TSS usage between two experimental conditions (Fig. 2.S1A). To simplify the generation of input data, we include a helper script, '`mepp.get_scored_fasta`' which generates a scored FASTA file from a scored BED file and a reference genome FASTA file.

Degenerate sequences, sequences from repetitive regions, and sequences sampled from overlapping genomic intervals can negatively affect the interpretation of the MEPP results. We describe optional steps to filter out these sequences in the Supplemental Methods.

2.3.1.2 Motif heatmap generation

Position weight matrices (PWMs) represent TF binding motifs as a matrix of nucleotide specificities. The match of a given DNA subsequence to a PWM occurs at variable strengths, quantified as the log-odds score of the match between subsequence and PWM [32,76,77]. PWMs usually accompany a log-odds score threshold above which a subsequence is determined to be a match to the motif PWM [32].

MEPP accepts a list of motifs in JASPAR format [76]. For each motif *j*, MEPP creates a convolutional model function f_j that accepts a one-hot encoded DNA sequence S_i and outputs log-odds match scores to the given motif (Fig. 2.S1E). All sequences S are expected to have the same length. Using functions from the Motif Occurrence Detection Suite (MOODS) [32], we calculate a log-odds score threshold b_j describing the minimum threshold log-likelihood match score for motif *j* under a given nucleotide background, pseudocount, and p-value threshold (Fig. 2.S1E). Thus, given a motif *j* and sequence *i*, MEPP computes a heatmap row vector as:

$$H_{i,j} = h_j(S_i) = pad(max(0, f_j(S_i) - b_j))$$

Where pad is a 0-value padding function that ensures $H_{i,j}$ and S_i have the same length. When considering both orientations of a motif, MEPP will instead compute the row $H_{i,j}$ as:

$$H_{i,j} = max(h_j(S_i), reverse(h_j(revcomp(S_i))))$$

Where *reverse(X)* reverses an array of motif scores, and *revcomp*(S_i) computes the reverse complement of one-hot sequence S_i. When all sequences S are sorted according to score, the matrix of all rows $H_{i,j}$ over sequence indices i is the motif score heatmap H_j for motif j. The central plot generated by MEPP displays the heatmap H_j (Fig. 2.1D) with the horizontal axis corresponding to motif position, and the vertical axis corresponding to each of the input sequences sorted in descending order based on their sequence scores. Motif position is measured from the center of sequences in S to the center of motif j. Rendering of the motif heatmap has additional considerations described in the Supplemental Methods.

To account for inexact motif positioning, we later optionally apply average pooling with a stride of 1 to the rows of the motif score heatmap $H_{i,j}$. The size of the pooling window w can be modified to adjust the resolution of the positional profiles, and is computed as w=1+2m, where m is a user-defined motif-margin. For high-resolution datasets that describe features at single nucleotide resolution, such as TSS found with csRNA-seq, the motif margin used should be small, usually 2 bp, but for lower-resolution datasets where the definition of the anchor may be less precise, such as the position of ChIP-seq peaks, a higher motif margin (e.g. 10 bp) may increase sensitivity. We apply 0-value padding to the data so that convolution and average pooling operations result in tensors matching the dimensions of the original one-hot encoded sequence. For simplicity, we refer to the smoothed form of row $H_{i,j}$ as $H_{smoothed,i,j}$

2.3.1.3 Positional profile computation

To calculate local motif enrichment at each sequence position across the dataset, at each position column X of the motif heatmap $H_{smoothed,j}$, we calculate the partial Pearson correlation $p_{XY \cdot Z}$ of the motif score matrix against the vector of sequence scores Y, while controlling for the vector of sequence-wide GC ratios Z. The resulting vector $P_{XY \cdot Z}$ of positional correlation coefficients describes the enrichment of the motif across all positions. We term a motif's vector $P_{XY \cdot Z}$ the positional profile or positional Pearson correlation for that motif. This enrichment method is comparable to that used by Analysis of Motif

Enrichment (AME) [4], which by default calculates enrichment as the correlation between average motif match scores across each sequence and the sequence scores. For each motif, the positional profile $P_{XY \cdot Z}$ is plotted across the same motif position axis as the central heatmap (Fig. 2.1F).

In order to determine statistical significance, we use permutation testing to calculate positional profiles on multiple null permutations of the data. The permutation test shuffles sequence scores to break the relationship between motif position/presence and score. The resulting null enrichment profiles are used to derive confidence intervals and p-values for the scores in the positional profile. Confidence intervals are shaded in gray beneath the positional profile (Fig. 2.1F).

We also calculate the count of motifs at each position summed up across the datasets. A rolling average with window size *w* smooths these values, which our method plots above the central motif heatmap of the MEPP visualization (Fig. 2.1B).

2.3.1.4 Per-motif visualization

For each motif, MEPP creates a plot with multiple subplots visualizing different aspects of the motif enrichment. These include the central heatmap (Fig. 2.1D), positional profile (Fig. 2.1F), and smoothed motif counts over positions (Fig. 2.1B) previously described, as well as the motif logo generated by Logomaker (Fig. 2.1A) [36]. In addition, the left-hand-side plot displays the relationship between the rank of the sequence score vs. the score (Fig. 2.1C), helping diagnose issues caused by non-normal score distributions that may throw off the correlation metrics. To contextualize the results a user might expect from non-positional score-based MEA, the right-hand-side plot displays the density of motifs as they occur along the dataset, smoothed along the sequence score rank axis for display (Fig. 2.1E).

2.3.1.5 Motif dataset visualization

MEPP also provides an interactive table for navigating to the profiles generated for each motif in a dataset (Fig. 2.1G,H,I). For each motif, MEPP renders the positional profile and its confidence interval in

a sparkline format (Fig. 2.1H), alongside an illustration of the motif matrix itself (Fig. 2.1G). The method identifies the extreme values of the positional profiles (Fig. 2.1I) and records them alongside their confidence interval and associated p-values. To control for false positives, the Benjamini-Yekutieli [78] correction adjusts p-values by correcting across all positional p-values and all motifs; We use the correction implemented by statsmodels [34]. MEPP renders the resulting table in HTML, augmented with interactive sorting and filtering features using the DataTables Javascript library (https://datatables.net/).

To aid in data exploration, MEPP renders a custom HTML output (Fig. 2.3B-E), placing the motif matrices next to a heatmap and dendrogram displaying the positional profiles and their clustering hierarchy; This custom interactive clustermap displays motifs along the vertical axis. To keep output legible, we use interactive CSS to expand rows of the heatmap on mouseover.

MEPP clusters the motifs by their positional profiles using UPGMA hierarchical agglomerative clustering (https://doi.org/10.1007/978-1-4020-6754-9_17806) with a correlation clustering metric. The clustering assignments of each motif profile follow the defaults for scipy's 'dendrogram' function (Fig. 2.3B) [79].

Both the table and clustermap HTML output generated by MEPP allow users to navigate to the individual MEPP plots for each motif through hyperlinks.

2.3.2 Public data used and analyzed

We used MEPP to analyze multiple public datasets. For each dataset, we sample sequences surrounding measured events from the relevant sequencing assay, and score these sequences according to either normalized read count coverage or the $\log_2$ fold change when comparing conditions. We then input the resulting scored sequences to MEPP. For convenience, these analyses are summarized in Table 2.1, while full analysis details are provided in the Supplemental Methods (See Appendix A.2). Access information and lab attribution for public data used from other studies is recorded in Supplemental Table 2.S4.

Table 2.1: Public data used in this study

Discussion section title	Methods/Supplemental section title	Sequencing assay	Measured event	Sequence center	Sequence length (bp)	Comparison	Sequence scores	Score polarity (Meaning of higher scores)	Genome	Dataset accession (s)
MEPP visualizes and quantifies positions of core promoter motifs	Analysis of Drosophila melanogaster TSS	csRNA-seq	Nascent transcription initiation	Nascent transcription start sites	400	Nascent transcription vs. none	log-transformed fragment 5' end coverage	Higher transcription	dm6	GSE203135
MEPP visualizes ChIP-seq peaks	MEPP analysis of GATA1 ChIP-seq	ChIP-seq	GATA1 binding	MACS2-called peak summits	400	ChIP vs. input	log2 Fold Change GATA1 ChIP-seq vs input	Greater GATA1 ChIP-seq signal	hg38	ENCSR000EFT (GATA1 ChIP-seq), ENCSR000EHM (Control)
MEPP visualizes cell-type specific TF binding motif spacing	Analysis of differential chromatin accessibility between cell types	ATAC-seq	Cleavage of accessible chromatin	GATA1 binding motifs	200	HCT116 vs. K562	log2 Fold Change HCT116 vs. K562 ATAC-seq	Greater accessibility HCT116	hg38	ENCSR483RKN (K562 ATAC-seq), ENCSR872WG (HCT116 ATAC-seq)
MEPP identifies helical spacing for motifs associated with cooperative Nanog binding	Analysis of Nanog motif binding in Mus musculus	ChIP-seq	Nanog binding	Nanog binding motif	400	Normalized coverage	RPKM-transformed Nanog ChIP-seq coverage	Higher coverage	mm10	GSE144577
MEPP visualizes differing positional specificities of TF binding assays	MEPP Analysis of Mouse ChIP-nexus and ChIP-seq	ChIP-seq	Nanog binding	MACS2-called peak summits	400	Binding signal vs. background	MACS2 peak signal	Higher ChIP-seq signal	mm10	GSE137193
MEPP visualizes differing positional specificities of TF binding assays	MEPP Analysis of Mouse ChIP-nexus and ChIP-seq	ChIP-nexus	Nanog binding	MACS2-called peak summits	400	Binding signal vs. background	MACS2 peak signal	Higher ChIP-nexus signal	mm10	GSE137193
MEPP visualizes differing positional specificities of TF binding assays	MEPP Analysis of Mouse ChIP-nexus and ChIP-seq	ChIP-nexus	Nanog binding	5' read end locations	400	Normalized coverage	rlog-normalized fragment 5' end coverage	Higher ChIP-nexus 5' end coverage	mm10	GSE137193
MEPP yields concordant profiles for assays of differential LPS response	Differential csRNA-seq analysis	csRNA-seq	Nascent transcription initiation	Nascent transcription start sites	400	KLA-stimulated vs. control	log2 Fold Change KLA vs. control nascent transcription	Higher transcription in KLA stimulation	mm10	GSE135498
MEPP yields concordant profiles for assays of differential LPS response	Differential cleavage site analysis	ATAC-seq	Cleavage of accessible chromatin	Cleavage sites	400	LPS-stimulated vs. control	log2 Fold Change LPS vs. control nascent cleavage	Higher transcription in LPS stimulation	mm10	GSE119693
MEPP yields concordant profiles for assays of differential LPS response	Differential cleavage site analysis	MNase-seq	Cleavage of accessible chromatin	Cleavage sites	400	LPS-stimulated vs. control	log2 Fold Change LPS vs. control nascent cleavage	Higher transcription in LPS stimulation	mm10	GSE119693

2.4 Results

2.4.1 MEPP visualizes and quantifies positions of core promoter motifs

To demonstrate the utility of our method in identifying known positional dependencies for DNA motifs, we analyzed transcription start sites (TSS) in *Drosophila melanogaster* embryo cells. Using capped small (cs)RNA-seq, we identified 44,877 high confidence TSSs (read count >3 while controlling for background input, repetitive DNA content, and overlapping sites, see Supplemental Methods). We extracted sequences covering +/-200 bp from each TSS and scored them by the log-transformed csRNA-seq coverage of their TSS centers, where higher scores correspond to TSS with higher rates of initiation. We then ran MEPP using a library of motifs previously found to be enriched in *Drosophila* promoters [80], focusing on the positioning of the TATA-box, Initiator (Inr), and Downstream Promoter Element (DPE) motifs, which are expected to appear upstream, on, and downstream of TSS, respectively.

MEPP visualizes a motif's occurrences in a scored sequence dataset in one figure comprising multiple plots with aligned axes (Fig. 2.1). The central 2D motif heatmap allows more direct visualization and qualitative evaluation of motif distributions across the dataset: The presence of the Inr motif on the TSS, in the sequence center, is clearly visible and is more well defined for TSS with greater transcriptional activity (Fig. 2.1D). The right-hand plot of motif density over sequence ranks also reflects the association of pronounced Inr motifs with greater TSS activity, and resembles the data as it would appear to a motif enrichment method using the zero-or-one-occurrence per sequence (ZOOPs) model (Fig. 2.1E). However, it is the enrichment positional profile plotted at the bottom that summarizes the positionality of this enrichment (Fig. 2.1F): This plot illustrates that the association of Inr motif strength/presence with TSS strength is most positively correlated at the sequence center, on the TSS itself, which matches expectations. Similarly, the TATA-box motif profile peaks upstream of the TSS at -28 bp, while the DPE motif profile peaks 25 bp downstream of the TSS (Fig. 2.S2A,B). Positions are reported using the distance from the center of the motif to the center of the sequence (which is defined here to be the TSS).

MEPP performs an analysis of multiple motifs for a dataset and summarizes them in an interactive table (Fig. 2.1G,H,I). The top 2 results for the most extreme positional profile values

corresponded to profiles for the Inr and DPR motifs, while 4th result corresponded to the profile for the TATA-box. The results table also describes the location of these extreme values across sequences centered on the TSS (as determined by the position of the largest absolute values in the profile). As expected, maximum correlation of the TATA-box motif with transcriptional initiation occurs upstream at -28 bp relative to the TSS, while the DPE's maximum correlation occurs downstream at +25 bp relative to the TSS. This is consistent with the known positioning characteristics of these core promoter elements [81,82]. Thus, we demonstrate how MEPP's multiple readouts recapitulate ground truths about promoter organization from a high-resolution nascent transcriptional assay.

2.4.2 MEPP visualizes ChIP-seq peaks

To demonstrate our method's ability to visualize known motif content in more general sequencing assays with less exact positioning, we analyzed ChIP-seq peak summits for GATA1 in K562 cells. We used MACS2 on ENCODE GATA1 ChIP-seq alignment files and the corresponding Control ChIP-seq alignment files to extract over 5K non-overlapping sequences sampled +/- 200bp from GATA1 ChIP or Control ChIP summits and scored by the log2 fold change between GATA1 ChIP and control. MEPP analysis on these scored sequences found centrally positioned enrichment of the GATA1 motif correlated with higher coverage in GATA1 ChIP over control, as expected (Fig. 2.2).

Unlike the previous analysis, this enrichment profile reflects positional sequence matches regardless of the orientation of the GATA1 motif. The maximum correlation signal in the positional profile provided by MEPP is less sharp compared to the analysis of core promoter elements relative to single nucleotide-resolution TSS, reflecting the less positionally specific nature of the ChIP-seq assay compared to the csRNA-seq assay. Similar to the previous result on TSS, this demonstrates that MEPP can identify known motif distribution patterns correctly, even when the assay in question has less distinct positional landmarks. This marks its utility in characterizing such assays as a visualization and quality control tool.

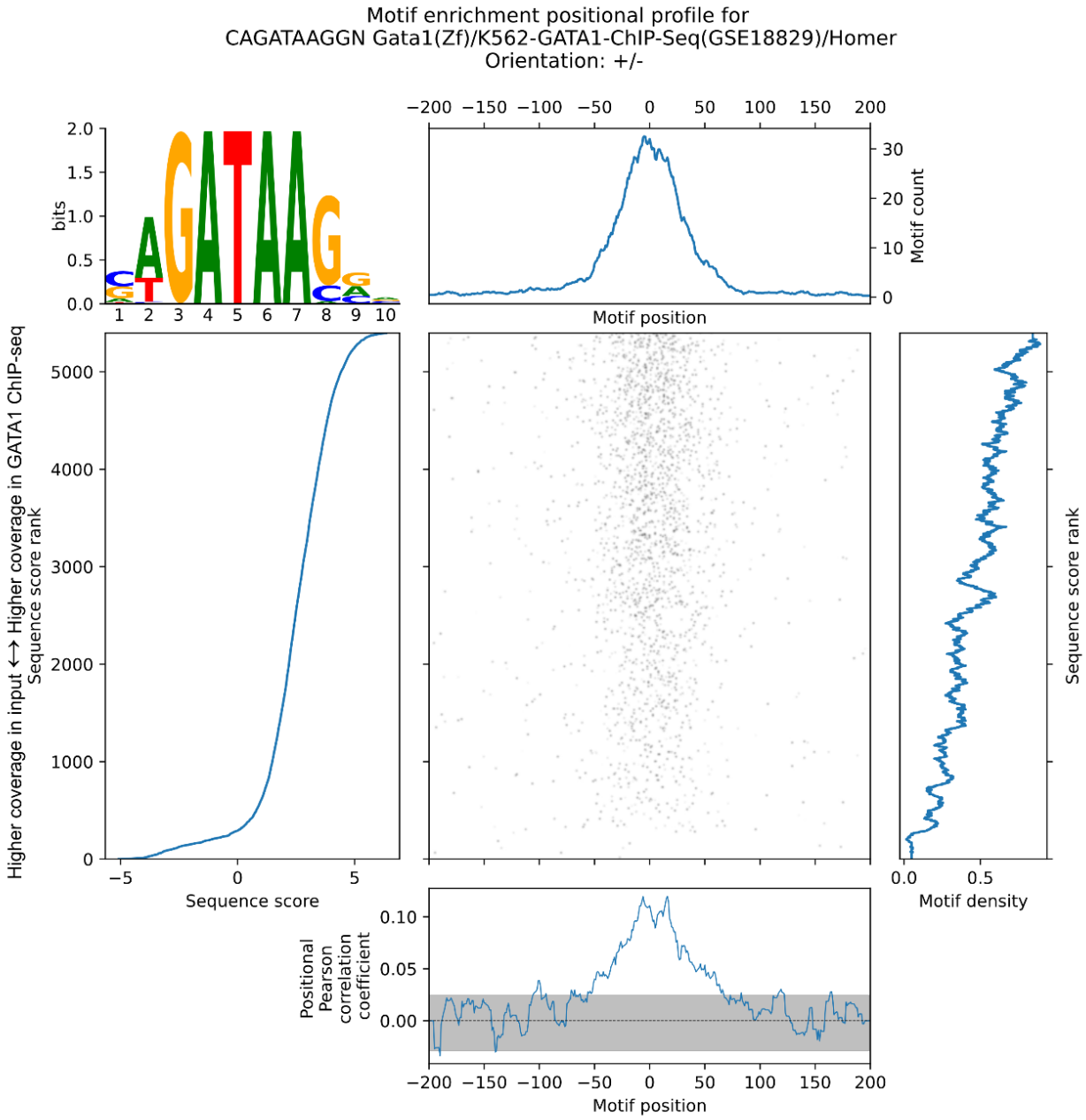


Figure 2.2: MEPP visualizes and quantifies the GATA1 binding motif in GATA1 ChIP-seq binding sites.

MEPP plot for the GATA1 motif, on sequences +/- 200bp of GATA1 ChIP-seq peak centers sampled from the hg38 reference genome, which are scored by differential ChIP-seq coverage (Log2 Fold Change of GATA1 ChIP-seq vs. input Control).

2.4.3 MEPP visualizes cell-type specific TF binding motif spacing

To demonstrate MEPP's ability to identify cell-type specific regulatory grammars, we analyzed the occurrence of motifs surrounding GATA1 binding motif sequences in K562 and HCT116 cells. Instead of using features of NGS profiling to determine analysis anchors (e.g. TSS, ChIP-seq peak summits), here we anchor our analysis on GATA motifs and analyze how the presence of other nearby TF motifs are associated with regulatory element activity. Over 500K GATA1 binding motifs appear in the human genome, but these are not in equally accessible chromatin, especially across cell types. To determine if increased cell-type specific chromatin accessibility associates with a spacing preference between GATA1 and other motifs, we used MEPP to analyze the positions of other binding motifs surrounding GATA1 binding motifs. We extracted genomic sequence +/-100 bp around GATA1 binding motifs, then scored these sequences by the $\log_2$ fold change of chromatin accessibility between HCT116 and K562 cell types; High scores corresponded to higher accessibility in HCT116 than in K562, as measured using ATAC-seq. These scored sequences comprised our input to MEPP for this analysis.

The transcription factor GATA1 plays a key role in hematopoiesis and erythroid gene expression [83]. The heatmap for the SCL motif (also known as TAL1) indicates preferential positioning of this motif around 12 bp upstream of the GATA1 motif, as measured from the center of SCL motif to the 5' end of the GATA motif (Fig. 2.3A). This is consistent with the approximate requirements for binding of a complex assembled by Lmo2 including SCL and GATA1 [83], and is consistent with previous reports characterizing composite GATA:Ebox motifs bound by the factors during erythroid maturation [84]. The enrichment profile generated by MEPP indicates that this positioning of the SCL motif has enrichment surrounding GATA motif sites with greater chromatin accessibility in K562 cells, but not HCT116 cells (Fig. 2.3A), as indicated by the negatively scored valley in the profile at that upstream position. This is consistent with the erythroleukemia origin of K562 cells, where GATA1 and SCL/TAL1 transcription factors play important roles in hematopoietic differentiation of the erythroid lineage. Although the motif heatmap indicates that this profile derives from motifs in a relatively small section of the heatmap (Fig. 2.3A, yellow squares), these sections still reflect motifs present in thousands of the most extremely scored sequences.

Figure 2.3: MEPP visualizes and quantifies the SCL/TAL1 binding motif near GATA1 binding motif locations

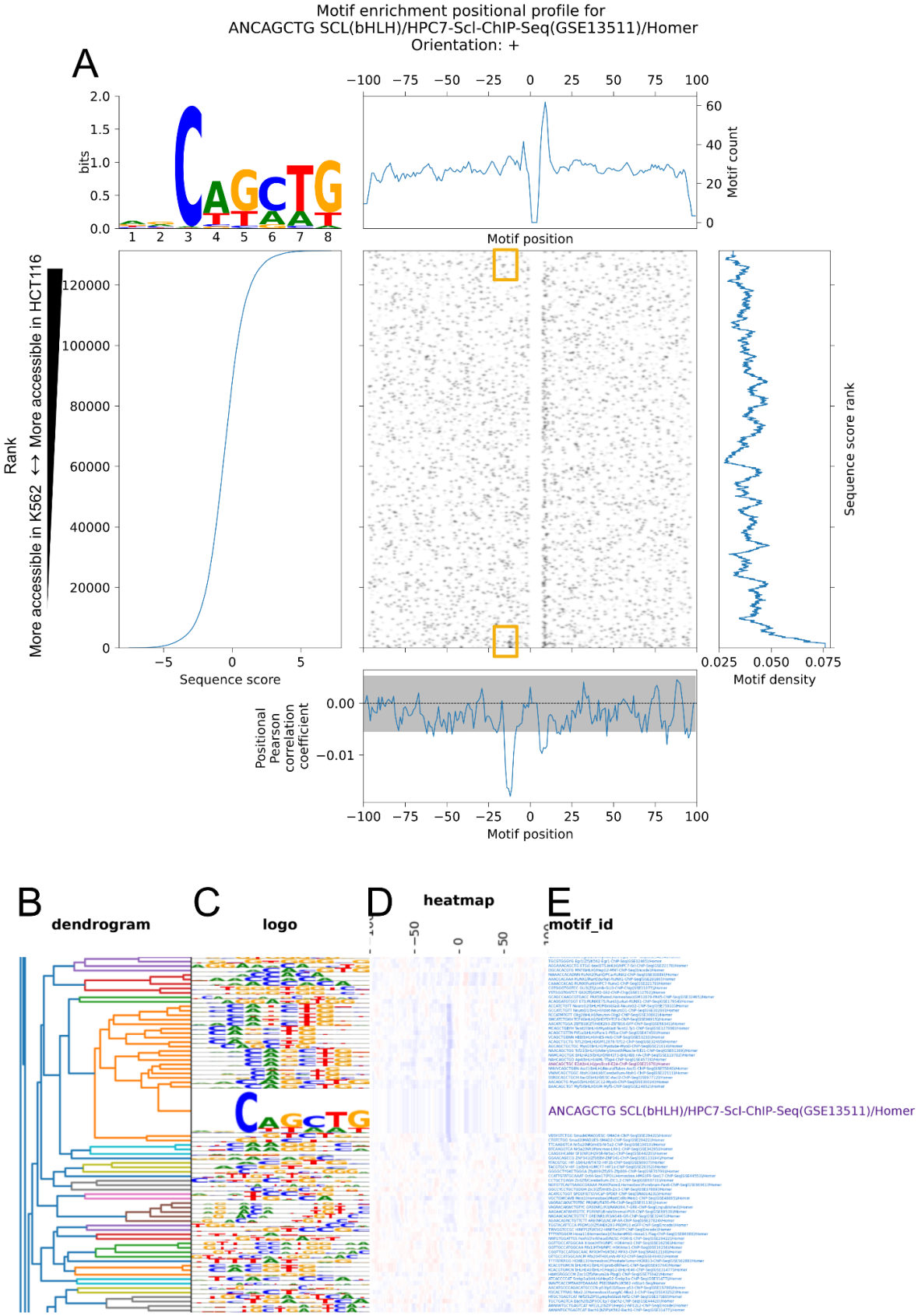
(A) MEPP plot for the SCL/TAL1 binding motif, on sequences +/- 100bp of GATA1 ChIP-seq peak centers sampled from the hg38 reference genome, which are scored by differential chromatin accessibility score (Log2 Fold Change of HCT116 over K562, by ATAC-seq). Yellow boxes indicate extrema of the heatmap where SCL motif presence contributes to the enrichment profile's minima.

(B) Dendrogram illustrating cluster membership of motifs characterized by enrichment positional profiles

(C) Motif logos represented in compacted form alongside enrichment profiles, with full logos visible on mouseover

(D) Heatmap visualizing motif enrichment profiles as rows of color bars, with red, white, and blue coloration signifying positive, zero, and negative correlation with sequence score

(E) Motif names with hyperlinks to full MEPP plots, with enlarged font scaling on mouseover for readability


Motif enrichment positional profile for
ANCAGCTG SCL(bHLH)/HPC7-Scl-ChIP-Seq(GSE13511)/Homer
Orientation: +
A
bits
Motif count
Motif position
Rank
More accessible in K562 ↔ More accessible in HCT116
Sequence score
Sequence score rank
Motif density
Positional Pearson correlation coefficient
Motif position
B
dendrogram
C
logo
D
heatmap
E
motif_id
ANCAGCTG SCL(bHLH)/HPC7-Scl-ChIP-Seq(GSE13511)/Homer

In order to evaluate the behavior of another method for evaluating local motif enrichment, we also analyzed this dataset with CentriMo [73]. Because CentriMo takes contrasting sets of sequences as input, rather than continuously scored sequences, we submitted the top 10% of scored sequences as the "positive" set for enrichment, and the bottom 10% as a contrasting "negative" set. The resulting local enrichment plot (Fig. 2.S3) yields a profile that does not differentiate between a motif being simply prevalent at a position, or more enriched in the "negative" set of sequences. Instead, this profile has two positive peaks, consistent with the peaks in MEPP's plot of motif counts over positions across sequences (Fig. 2.S3). While a second profile is plotted as a dashed line reflecting enrichment in the negative set of sequences, its interpretation relies on the selection of the negative set of sequences, and comparison against the profile for enrichment in the positive set (Fig. 2.S3). This underscores a key difference in MEPP's enrichment profile output from current methods like CentriMo: Rather than only quantifying a motif's positional prevalence in a thresholded selection of a dataset, MEPP, quantifies motif's positional relevance towards a higher or lower scoring sequence, as measured by the local correlation of motif score and sequence score. In addition, CentriMo only accounts for the position of the best match to the motif within a sequence, while MEPP quantifies and visualizes all motif instances within a sequence.

MEPP visualizes and clusters the profiles generated for multiple motifs as an interactive HTML clustermap (Fig. 2.3B-E). This allows users to determine regimes of positional dependencies shared by similar motifs. For example, the SCL motif clusters with similar basic helix-loop-helix binding motifs. Unlike conventional non-interactive heatmaps, our approach to visualization allows users to expand motif logos and text, as well as to click through hyperlinks to the full MEPP profile. This allows for determination of profile similarities across a full motif set at a glance, rendering no singular row permanently unreadable. This combination of novel interactive visualization techniques and positional, score-based motif enrichment is unique to MEPP's approach, and enables users to identify cell-type specific regulatory grammar.

2.4.4 MEPP identifies helical spacing for motifs associated with cooperative Nanog binding

To demonstrate MEPP's ability to identify complex relationships between motifs that have roles in cooperative TF binding, we performed an analysis of Nanog binding in mouse embryonic stem cells (ESC). The Nanog motif is relatively common in the genome, but not all instances of this motif are bound. The Nanog motif instances that are bound often have varying rates of association with Nanog as measured by ChIP-seq. To identify other motifs near Nanog motifs that have positional specificities in their ability to influence Nanog binding, we performed a MEPP analysis of Nanog motif sites across the mm10 reference genome. We scored Nanog motif sites by their Nanog binding activity as quantified by Nanog ChIP-seq in mouse embryonic stem cells (mESCs) [85]. The analysis processed over 3M sequences sampled +/- 200bp of Nanog motif sites, after cluster deduplication and filtering out sequences containing 50% repetitive or degenerate bases as annotated byRepeatMasker.

MEPP analysis showed that motifs bound by pluripotent transcription factors often revealed helical spacing preferences to Nanog motifs bound by Nanog in mESCs. The MEPP plot for enrichment of Sox2 motifs surrounding central Nanog motifs reveals periodicity in the enrichment positional profile with a period of approximately 10 bp (Fig. 2.4). This periodicity is less visible when simply plotting Sox2 motif counts over positions relative to Nanog (Fig. 2.4). Positive peaks in the enrichment positional profile represent a stronger local correlation of Sox2 motif strength/presence with Nanog binding at those periodically spaced positions, suggesting that cooperative binding of Sox2 and Nanog depends on a helical syntax that preserves the relative rotational positions of the factors along the DNA. Other approaches leveraging machine learning models have also found helical binding periodicities between Nanog and Sox2 motifs [86,87]. However, our method does not require the training or interpretation of machine learning models, but yields concordant results. Importantly, due to the cluster deduplication step in the data preprocessing, our results do not reflect repetition of the Nanog motif around itself, ensuring that these findings are not due to e.g. a single Sox2 motif appearing near multiple Nanog motifs that are spaced periodically with each other, as might occur in unannotated repetitive genome sequence. By

combining MEPP with careful preprocessing, we demonstrate the ability to identify properties of motif spacing more complex than single peaks of positional enrichment.

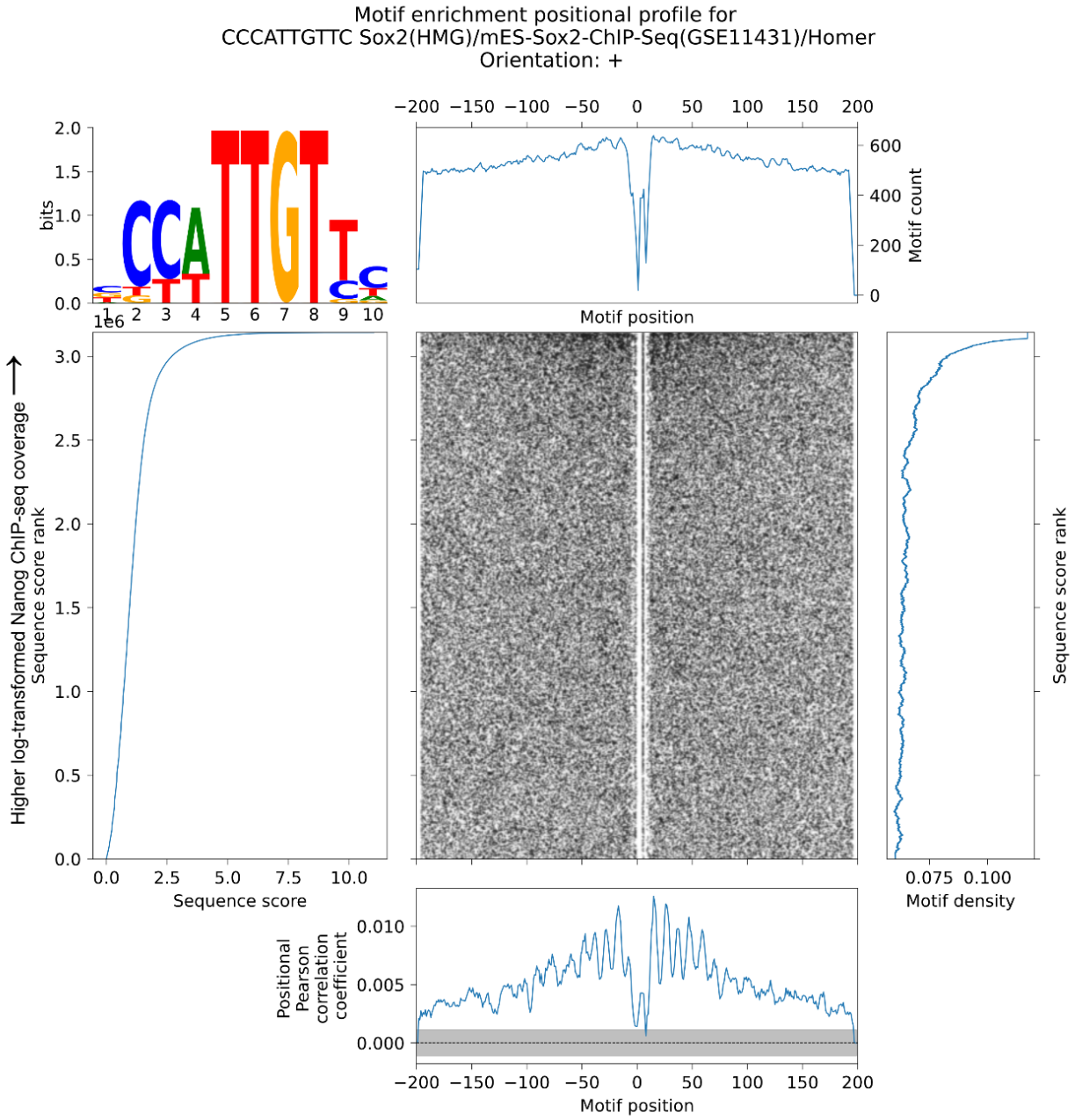


Figure 2.4: MEPP visualizes and quantifies the Sox2 motif near bound Nanog motif sites.

MEPP plot for the Sox2 motif, on sequences +/- 200bp of GATA1 ChIP-seq peak centers sampled from the mm10 reference genome, which are scored by log-transformed Nanog ChIP-seq coverage.

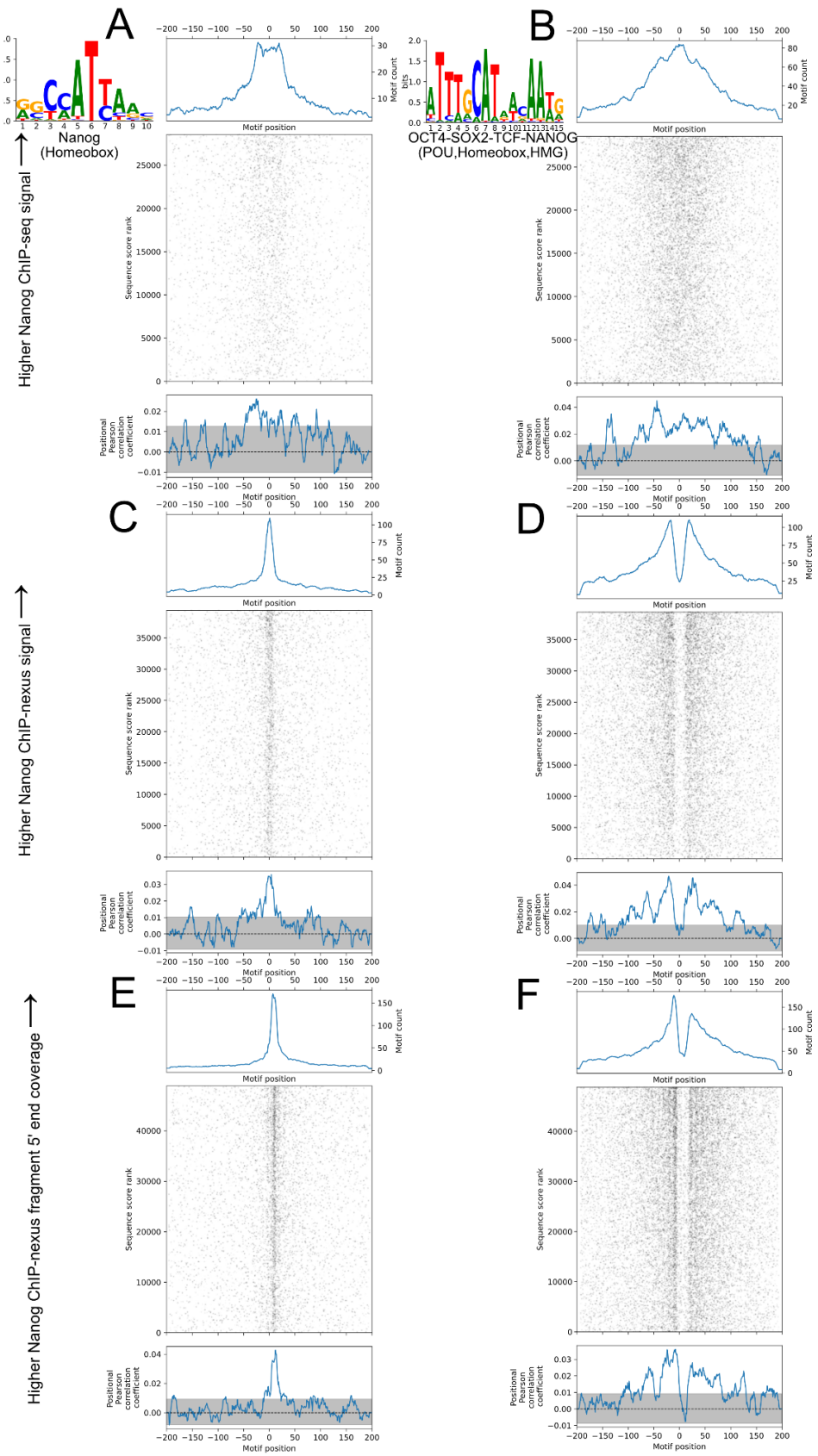
2.4.5 MEPP visualizes differing positional specificities of TF binding assays

To demonstrate the effect of assay type positionality on the positional profiles derived by MEPP, we analyzed ChIP-nexus and ChIP-seq Nanog binding assays in mouse embryonic stem cells, as carried out by Avsec et al. [86]. ChIP-nexus assays use exonucleases to precisely map the locations where crosslinked proteins protect the DNA, suggesting that ChIP-nexus peaks should provide greater precision than ChIP-seq peaks with respect to binding motifs [67]. MEPP analyzed 39K Nanog ChIP-nexus peaks and over 28K Nanog ChIP-seq peaks, using sequence sampled from +/-200bp of each peak summit and scores taken from the signal values in the MACS2 narrowpeak calls. To account for the lack of strand specificity in ChIP-seq, MEPP correlated sequence scores against both forward and reverse orientations of each motif.

As expected, the Nanog motif positional profiles derived from Nanog ChIP-nexus peaks showed greater positional specificity, with the positional profile indicating a positive peak centered directly on the peak summit (Fig. 2.5C). In contrast, the Nanog motif positional profile from the Nanog ChIP-seq experiment indicates a broader, less well-defined central peak that does not rise as far above the 95% confidence interval (Fig. 2.5A). Additionally, positional profiles for the Oct4-Sox2-TCF-Nanog composite motif follow a similar pattern, with the Nanog ChIP-nexus derived profile having enough granularity to resolve two peaks on either side of the peak summit, as opposed to the broader profile reflected from the Nanog ChIP-seq experiment (Fig. 2.5B,D). The motif heatmap visualization offered by MEPP enables researchers to visualize the underlying two-dimensional distribution of motifs surrounding each experiment's peak summits, providing further feedback on the positional properties of each dataset. Thus, MEPP results capably reflect the positional specificities of different sequencing assays, allowing both quantitative and qualitative feedback on sequence features enriched in the surrounding assay peak summits.

Figure 2.5: MEPP differences in positional specificity between ChIP-seq and ChIP-nexus

(A) MEPP plot for Nanog binding motif, on sequences +/- 200bp of Nanog ChIP-seq peak summits sampled from the mm10 reference genome, scored by MACS2 signal value for each peak.
(B) MEPP plot for Oct4-Sox2-TCF-Nanog composite binding motif, on sequences +/- 200bp of Nanog ChIP-seq peak summits sampled from the mm10 reference genome, scored by MACS2 signal value for each peak.
(C) Same, as (A), but for sequences sampled and scored from Nanog ChIP-nexus peak summits.
(D) Same, as (B), but for sequences sampled and scored from Nanog ChIP-nexus peak summits.
(E) Same, as (A), but for sequences sampled and scored from Nanog ChIP-nexus fragment 5' ends found and scored using an alternate HOMER analysis pipeline adapted from use on csRNA-seq.
(F) Same, as (B), but for sequences sampled and scored from Nanog ChIP-nexus fragment 5' ends found and scored using an alternate HOMER analysis pipeline adapted from use on csRNA-seq.


A
B
C
D
E
F
Nanog
(Homeobox)
OCT4-SOX2-TCF-NANOG
(POU,Homeobox,HMG)
Higher Nanog ChIP-seq signal
Higher Nanog ChIP-nexus signal
Higher Nanog ChIP-nexus fragment 5' end coverage
Motif position
Motif count
Sequence score rank
Positional Pearson correlation coefficient
bits

The positional specificity visualized by MEPP is further enhanced using analysis methods that leverage the positional information provided by the ends of the reads, such as those developed for csRNA-seq. To demonstrate, we re-analyzed the ChIP-nexus data that follows the example of csRNA-seq, by identifying and scoring prominent ChIP-nexus 5' protection boundaries from the 5' ends of the Nanog ChIP-nexus reads. This alternative analysis identified 48K potential binding sites, scored by the (DESeq2) rlog-transformed coverage of the Nanog ChIP-nexus 5' read ends [88]. MEPP was used to analyze sequence +/-200 bp of these binding sites for motif enrichment. The resulting profiles for the Nanog motif and the Oct4-Sox2-TCF-Nanog are similar to the previous Nanog ChIP-nexus experiment (Fig. 2.5E,F). However, the enhanced specificity and data density appears as visible vertical striations of motif presence on the central motif heatmaps, which provides a clearer profile peak center in the case of the Nanog motif profile (Fig. 2.5E). Thus, MEPP results reflect specificities from both assay types and analysis approaches, providing both quantitative and qualitative feedback to researchers developing or refining methods for assay or analysis.

2.4.6 MEPP yields concordant profiles for assays of differential LPS response

To demonstrate the applicability of MEPP to multiple types of sequencing experiments, we performed a differential analysis of TSS measured by csRNA-seq and cleavage sites from ATAC-seq and MNase-seq experiments. These experiments compared the state of mouse bone marrow-derived macrophages (BMDMs) before and after 1 hour of LPS stimulation [10,89]. which activates innate immunity pathways by triggering Toll-like receptor 4 (TLR4) signaling. In each experiment, sequences were sampled from +/-200 bp of genomic coordinates taken from the 5' ends of reads: in csRNA-seq, these represent TSS, while in ATAC-seq and MNase-seq, these represent cleavage sites for accessible DNA by the assay's respective enzyme (Fig. 2.6A,B, adapted from Tsompana & Buck 2014) [68,72,90]. In the case of MNase-seq, digested chromatin was further ChIPed for H3K27ac, reflecting transcriptionally active nucleosomes [89]. All TSS/cleavage sites and their associated sequences were scored by log2 fold change comparing pre- and post- stimulation coverage as calculated by DESeq2.

Figure 2.6: MEPP Plots summarize NF-kB motif enrichment across csRNA-/ATAC-/MNase-seq TSS/Cleavage sites

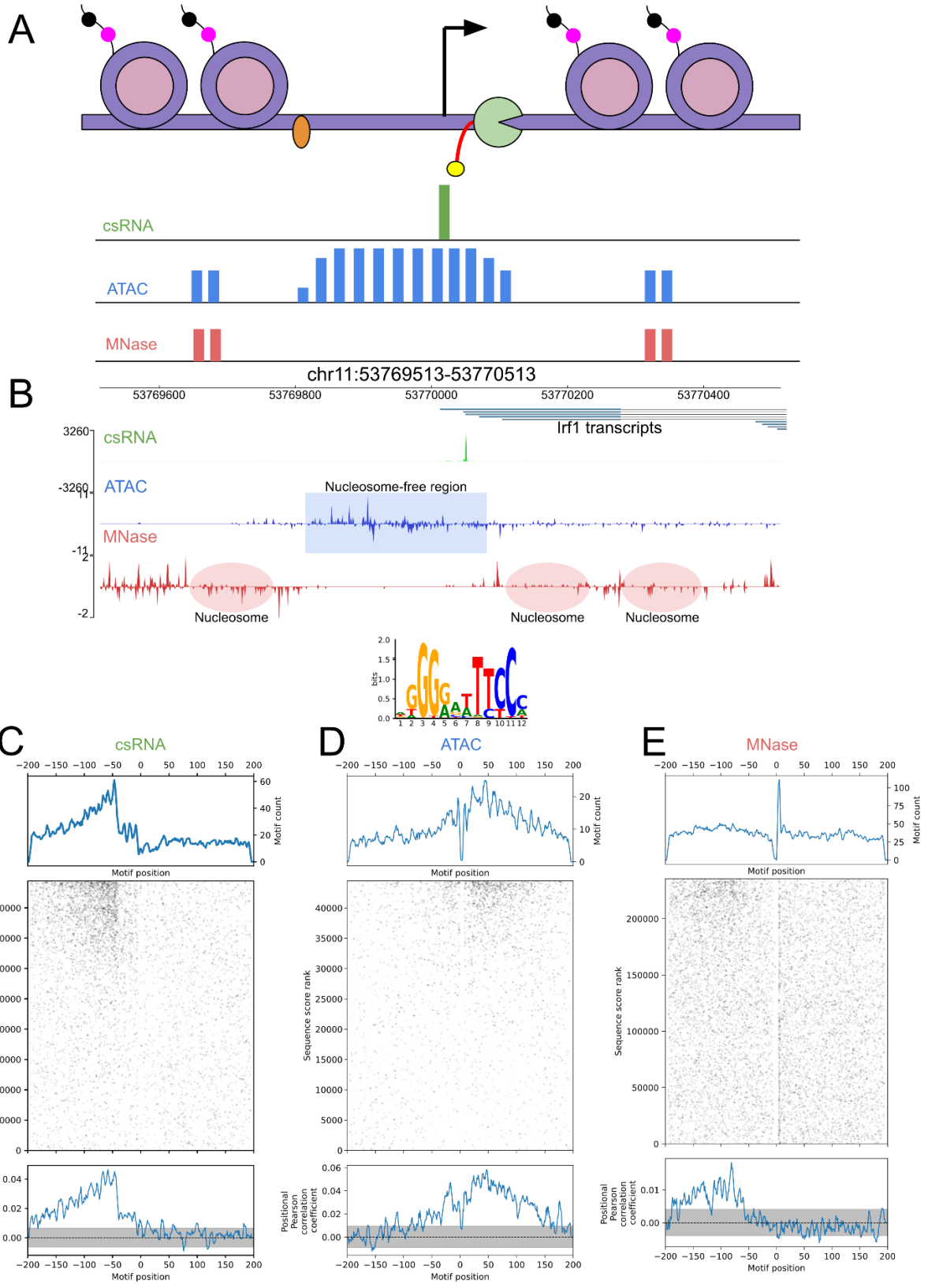
(A) Diagram illustrating read coverage from csRNA-seq (Green), ATAC-seq (Blue), and MNase-seq + H3K27ac ChIP (Red) experiments. Adapted from Tsompana & Buck 2014 [90]. csRNA-seq assays nascent TSS from 5' capped short RNA transcripts, while ATAC-seq and MNase-seq assay open chromatin. MNase-seq from Comoglio et al. includes immunoprecipitation of H3K27ac.

(B) Integrated Genome Browser visualization of coverage from 5' ends of csRNA-seq, ATAC-seq, and MNase-seq reads near the *Irf1* transcription start site in mm10.

(C) MEPP plot for NF-kB binding motif, on sequences centered on csRNA-seq derived TSS, and scored by differential TSS nascent transcription between 1h LPS stimulation vs. 0h control.

(D) Similar as (B), but for sequences centered on ATAC-seq cleavage sites, and scored by differential 5' read coverage between 1h LPS stimulation vs. 0h control.

(E) Similar as (C), but for sequences centered on H3K27ac MNase-seq cleavage sites.


A
csRNA
ATAC
MNase
B
chr11:53769513-53770513
53769600
53769800
53770000
53770200
53770400
Irf1 transcripts
csRNA
ATAC
Nucleosome-free region
MNase
Nucleosome
Nucleosome
Nucleosome
C
csRNA
D
ATAC
E
MNase
Motif count
Motif position
Sequence score rank
Positional Pearson correlation coefficient

The transcription factor NF-kappa B (NF-kB) is known to induce strong changes in transcription in response to activation of TLR4 by LPS [91]. Thus, MEPPs for the NF-kB binding motif all feature concordantly positive peaks. In the csRNA-seq derived MEPP analysis of this motif, there is a clear positional peak 58bp upstream of TSS implying NF-kB binding to this position potently initiates transcription after activation (Fig. 2.6C). Similarly, in the H3K27ac MNase-seq analysis, the MEPP for the NF-kB binding motif exhibits a positive peak at 81bp upstream of the MNase cleavage site (Fig. 2.6E), indicating NF-kB binding likely increases histone acetylation on nucleosomes or repositions acetylated nucleosomes with their edge approximately 80 bp from of the NF-kB motif. Notably, this peak is distinct from the location where the same motif is most prevalent in sequence, just downstream of the cleavage site. Such a distinction underscores the ability of MEPP to distinguish motif relevance to biological signal, as opposed to motif prevalence across a set of sequences agnostic to biological signal.

Unlike the profiles for TSS and nucleosome edges, ATAC-seq derived MEPP analysis of the NF-kB binding motif revealed a strong preference approximately 45 bp downstream of the Tn5 cleavage site, generally placing NF-kB-DNA contacts on the fragments isolated in the ATAC-seq assay. There is also a positive association of NF-kB binding just upstream of the cleavage site, suggesting NF-kB binding may enhance the accessibility of sizable regions surrounding the NF-kB motif. (Fig. 2.6A,B,D) [92]. Similarly, there is positive motif enrichment both up and downstream of the central cleavage site, reflecting ATAC-seq read coverage surrounding a TF binding footprint. However, these profiles are still concordant with increased enrichment of the NF-kB motif in regulatory regions more accessible after LPS stimulation and its role in innate immune response. Thus, while peaks in the NF-kB motif profiles have concordant characteristics, differences in the profiles still reflect meaningful distinctions between the reads selected and sequenced for each assay. Such distinctions would not appear in analyses that report enrichment scores for motifs that do not take motif position into account, highlighting an advantage of MEPP's positional approach to motif enrichment.

2.5 Discussion

MEPP correlates the log-odds scores of a motif with biologically relevant measurements as a function of the motif's position to identify spatial relationships in regulatory DNA. In contrast, many MEA methods such as MEIRLOP and HOMER treat motif presence within a sequence under a zero-or-one-occurrence-per-sequence (ZOOPs) model: For enrichment, a motif is either present or absent [24,93]. This ignores how a motif may occur at multiple positions within a sequence and leads to methods that cannot describe positional dependencies of binding site function. Such positional dependencies may hold relevance when an experiment samples sequences from the genome surrounding biologically significant features, such as transcription start sites. By correlating motif presence at multiple positions in the sequences surrounding relevant features, MEPP enables positional profiling of motif enrichment alongside a structured visualization system that illustrates motif prevalence in the dataset. These results recapitulate known relationships such as the positioning of core promoter elements surrounding drosophila melanogaster TSS, and are capable of revealing more complex relationships including periodicities in motif positioning where a scatterplot/heatmap visualization does not provide enough clarity.

When applied to sequences surrounding GATA1 motifs, we find that our method recovers the positional relevance of SCL-to-GATA1 motif spacing to K562 cells, a result supported by the previous characterization of ternary complex formation on a composite GATA:E-box motif [84]. We demonstrate the ability of MEPP to summarize the positional enrichment of all motifs in a dataset, and present them in a novel interactive clustermap format. The clustermap allows the identification of locally co-enriched motifs, such as those with similar consensus sequences, or those that may comprise sub-motifs for binding a larger cis-regulatory mechanism, such as the GATA-SCL motifs for an Lmo2-bridged binding complex. Thus, MEPP's ability to visualize correlated positional relevance of motifs at a glance allows researchers to quickly observe transcriptional regulation mechanisms beyond single motifs, and to better contextualize results for single motifs.

When applied to multiple sequencing assays that present biologically relevant positioning features, such as csRNA-seq, ATAC-seq, or MNAse+ChIP-seq, we find that MEPP yields concordant

profiles whose differences reflect the biochemical specificities of the assays analyzed. Each of these assays produce reads describing biological phenomena such as nucleosome edges or transcription initiation at single nucleotide resolution that MEPP can leverage to investigate the roles that transcription factors play in regulating these phenomena. This can prove invaluable when describing multiple functions of regulatory sequences.

We find that unlike CentriMo or HOMER, which can plot the prevalence of a motif in a dataset of sequences, MEPP plots the positional relevance of motifs along a continuous score. The use of signed enrichment coefficients with a signed score allows researchers to investigate regulatory region sequences that vary between two extremes quantifiable by an assay-based score, such as those exhibiting cell-type- or stimulation- specific expression. While users could run similar analyses by analyzing quantiles or otherwise stratified bins of regulatory region sequences, these still require the user to select thresholds to partition the sequences according to best practices, which are not guaranteed when analyzing novel measures of biological activity. MEPP's motif heatmaps can assist in this task, allowing researchers to visualize motif presence along two dimensions of position and assay-based score, while avoiding overplotting effects. This transparency mitigates the risk of being misled by non-specific local motif prevalence. Similarly, MEPP plots the relationship between assay scores and sequence ranks, avoiding the risk of selecting non-informative thresholds for a score distribution. Thus, when taken together, all elements of a MEPP plot remain powerful in informing decisions for subsequent analyses.

We have demonstrated MEPP as a novel means of quantifying and visualizing the positional relevance of a motif across multiple centered genomic sequences. Similar to our previous work with MEIRLOP, MEPP is usable by scoring genomic regions across a continuum of scores reflecting two extremes of biological interest. Unlike other methods of performing positional motif enrichment, MEPP identifies local motif enrichments towards either extreme, with the sign reflecting a motif’s association with higher or lower scores. MEPP currently functions with a fixed motif library. However, the underlying convolutional network architecture lends itself easily to future work for recognizing and assembling de novo motifs based on correlated positional profiles.

2.6 Publication Acknowledgement Statements

N.P.D.S., S.H. and C.B. oversaw the overall design and execution of the project. The csRNA-seq experiments were performed by S.H.D. The computational analyses were performed by N.P.D.S.. N.P.D.S. and C.B. were primarily responsible for writing the manuscript with input from all authors.

This chapter is a reproduction of a manuscript submitted to and under review at Nucleic Acids Research Genomics and Bioinformatics. Content has been adjusted to satisfy dissertation formatting requirements, including the structure of the acknowledgement and data availability.

2.6.1 Data availability

All raw data generated for this study will be accessible at NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) accession number GSE203135.

2.6.2 Code availability

The code for MEPP is available from its Github repository at https://github.com/npdeloss/mepp, and can be installed through pip, via the command line:

```
pip install git+https://github.com/npdeloss/mepp@main
```

2.6.3 Funding

This work was supported by an NLM Training Grant (T15LM011271) and the Katzin Prize Endowed Fund for the work of N.P.D.S., NIH grants R00GM135515 to S.H.D., R01GM134366, U01DA051972, and U01AI150748 to C.B and R01GM129523 to S.H., who received additional support from NIH grants R01GM134366, P30DK063491, P30DK120515 and U01AI150748. The funding bodies had no direct roles in the design or execution of the study.

2.6.4 Ethics approval and consent to participate

Not applicable.

2.6.5 Consent for publication

Not applicable. This work uses human cell line data from the ENCODE Project [43,44].

2.6.6 Competing interests

The authors declare that they have no competing interests.

2.7 Acknowledgements

Chapter 2, in part, has been submitted for publication of the material as it may appear in Nucleic Acids Research Genomics and Bioinformatics, 2022, Delos Santos, Nathaniel P.; Duttke, Sascha; Heinz, Sven; Benner, Christopher, Oxford University Press 2022. Modifications have been made to the text to reflect minor editorial feedback. The dissertation author was the primary investigator and author of this paper.

Chapter 3. Learning Motifs with Positional Priors

3.1 Abstract

Functional transcription factor binding motifs can be positionally constrained around other features and transcription start sites. By using data-driven positional priors that inform these positional constraints, we can direct the discovery of these positionally constrained motifs, without knowing those motifs beforehand. Here we present a method for Learning Motifs from Positional Priors (LMPP), a novel machine-learning based approach that allows researchers to incorporate data from otherwise disparate experiments and reference frames, in order to direct the discovery of positionally enriched motifs relative to key anchor points (e.g. transcription start sites, binding sites) across a set of sequences scored according to biological signal. By using the enrichment positional profiles of known motifs as ground truths, we demonstrate the ability of LMPP to effectively recover motifs from their positional enrichment towards either extreme of a dataset of scored sequences. We also demonstrate the applicability of our method to discovery of motifs relevant to COVID-19 symptom severity, integrating data from both genome-wide association studies (GWAS) and high resolution nascent transcriptional assays. LMPP is an extension to our previous method, MEPP, and is available for download from https://github.com/npdeloss/mepp.

3.2 Background

Transcription factors (TFs) are key to transcriptional regulation in response to stimuli such as viral infection: By recognizing and binding to sites in regulatory region DNA sequence, they recruit transcriptional machinery across multiple genes, giving rise to feedback loops and host protective functions [94].

The sequence specificities of these TF binding sites are characterized as patterns in DNA, termed motifs. Typical motif enrichment analysis (MEA) enables researchers to compare the binding sites present between two extremes of regulatory regions, quantify the relative enrichment of those binding

sites between those extremes, and thus infer the transcription factors that constitute an explanatory mechanism for differences in regulatory region activity between those extremes [4].

Standard MEA requires that the binding site motifs to be analyzed must be known ahead of time, limiting the potential for discovery. In contrast, de novo motif discovery methods determine new motifs from overrepresented sequence content within regulatory region sequence, without the prior need for a set of known motifs. Interpretation is facilitated by comparing the discovered motifs against known motifs for transcription factors binding sites, but the search space is not confined by prior knowledge of binding site sequence specificities.

However, such de novo MEA methods are in practice confined by prior knowledge of binding site sequence positions, which we term a positional prior: These positional priors are typically broad criteria, such as searching for motifs within e.g. 500 bp upstream of annotated gene transcription start sites (TSS). This is due to many MEA methods generalizing motif presence in a binary manner, only considering motifs as having “zero or one occurrence per sequence” (ZOOPS) [95]. However, such a generalization cannot be accepted lightly: The distance between different binding sites and the TSS can affect TF binding and TSS activity, imposing constraints on the locations of functional binding sites within a TSS-proximal reference frame [13,14]. Motif enrichment positional profiling (MEPP) allows us to visualize the locations of functional binding sites relative to TSS, given a set of differentially scored TSS. However, it requires a set of known motifs as input before characterizing their positional enrichment. In this work, we will remove that requirement, by incorporating positional priors that can be driven by other data of interest, such as a GWAS study on infectious disease.

Infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has varying severity of symptoms across individuals: These symptoms can range from flu-like symptoms and loss of smell, to severe lung injury requiring hospitalization and shortness of breath that persists long after the initial infection. In order to determine the mechanisms for such varied symptoms and severities across individuals, several GWAS studies have been conducted, quantifying associations between genotype and viral response phenotypes. The results of these studies are available from the Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP) [96]. Among these are GWAS results comparing over 8.9 million single nucleotide polymorphisms (SNPs) for association with patients who

developed severe respiratory symptoms from COVID-19 (n=269), and those who contracted COVID-19 but were not hospitalized (n=688) [97].

While these studies utilize genotypes for millions of SNPs, not all of this genetic variation is directly observed: Direct observations for these studies came from genotyping arrays that assayed around 800 thousand markers, including SNPs, and indels [98]. The rest of the genotypes are imputed, taking advantage of how alleles for genetic variants can co-occur in linkage disequilibrium (LD), as observed from reference panels [99]. As a result of this use of linkage disequilibrium, a genotyped SNP may be reported for its indirect association with the true disease-causing SNP [100,101]. Thus, reported lead SNPs corresponding to loci from a GWAS study may correspond to “near misses” correlated with neighboring causal SNPs within regions of high linkage disequilibrium, potentially obscuring insight into mechanisms behind increased disease risk, e.g. disruption of binding sites for key transcription factors for disease response.

Due to issues with linkage disequilibrium and indirect association between SNPs, significant SNP associations from GWAS studies should not be assumed as precise locations of causal variants and disrupted transcription factor binding motifs within a genome-wide reference frame [100]. However, the positional constraints of functional binding in TSS-proximal reference frames present an opportunity: Just as functional binding motifs can be positionally constrained, so too may the SNPs that affect them (as well as nearby indirectly associated SNPs in LD). By marking positions within the TSS-proximal reference frame by a function describing the presence of nearby highly associated SNPs, we can construct a heuristic positional prior. We can then use this heuristic positional prior to search for positionally enriched transcription factor binding sites affected by either highly associated SNPs or a nearby indirectly associated causal SNPs. To ensure that the GWAS and enriched motifs are germane, the TSS should be scored according to differential or correlated expression towards phenotypes analyzed in the GWAS, e.g. severe lung injury vs. no lung injury.

Superficially, the use of positional priors for motif discovery resembles an approach implemented in Multiple EM for Motif Elicitation (MEME), which allows users to specify a “position-specific prior” for each input sequence to direct motif discovery [102]. However, similar to the score-based approaches employed with Motif Enrichment in Ranked Lists of Peaks (MEIRLOP) and MEPP, we have developed a

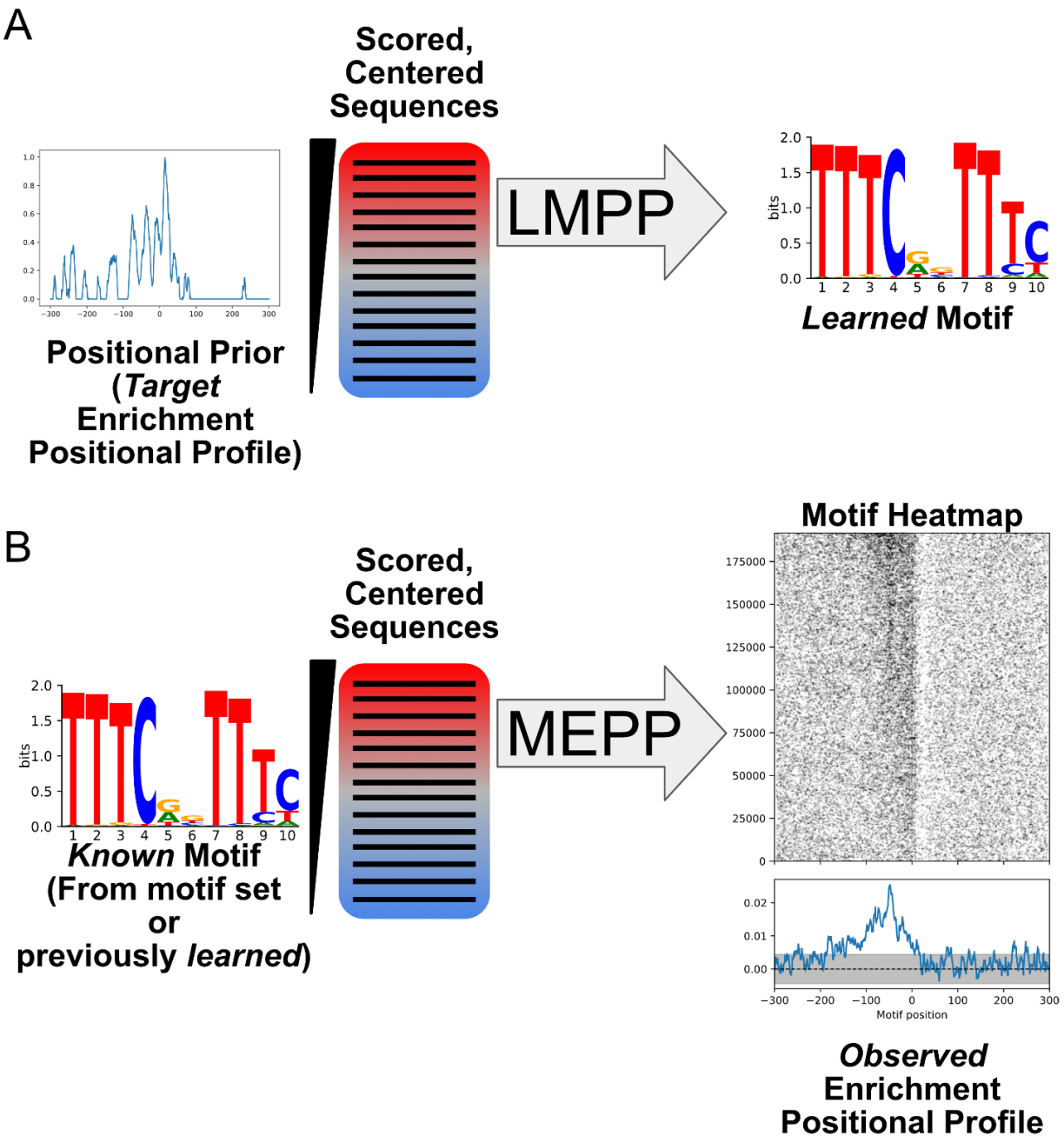
method to utilize positional priors for differential motif discovery across a dataset of scored sequences, rather than relying on set-based motif discovery methods such as Gibbs sampling or MEME's expectation maximization. This avoids previously described issues with set-based motif enrichment methods, such as the arbitrary setting of thresholds, and takes advantage of the fact that most biological data, differential or otherwise, is more naturally analyzed by ranking or scoring [7]. In addition, LMPP's interoperability with MEPP allows transparent visualization of the distribution of learned sequence features across both axes of the data (sequence score and motif position).

This use of positional priors for differential motif discovery across a set of scored sequences is the premise behind our method, which we term Learning Motifs from Positional Priors (LMPP). LMPP utilizes positional priors and specialized machine learning to direct the search for positionally enriched de novo motifs, whose enrichment positional profiles resemble the positional prior. LMPP is an extension to MEPP and its positional priors effectively describe a target enrichment positionality profile for discovered motifs (Fig. 3.1A).

Figure 3.1: LMPP inverts MEPP's motif-to-profile process by learning motifs from positional priors

(A) LMPP accepts a positional prior (i.e. a target enrichment positional profile) and a set of scored and centered sequences, then outputs a learned motif. LMPP's input scored sequence format is identical to MEPP's. Similar to MEPP, sequence anchors may be features such as TSS or binding events.

(B) MEPP accepts known motifs (e.g. from HOMER or JASPAR) and a set of scored and centered sequences, then outputs a motif heatmap and an enrichment positional profile. MEPP accepts JASPAR-formatted motifs, including LMPP's de novo motif output format. LMPP and MEPP are interoperable on the same datasets.


A
Scored, Centered Sequences
LMPP
Positional Prior (*Target* Enrichment Positional Profile)
Learned Motif
B
Scored, Centered Sequences
MEPP
Known Motif (From motif set or previously *learned*)
Motif Heatmap
Motif position
Observed Enrichment Positional Profile

3.3 Methods

3.3.1 LMPP Implementation

In order to learn de novo motifs while incorporating priors from positionally relevant signals, we developed and employed a novel technique, termed Learning Motifs With Positional Priors (LMPP). The typical execution of a de novo motif analysis with LMPP occurs in 6 parts:

1. Input data and pre-processing
2. Scaling positional profile by sequence scores.
3. Convolutional neural network training
4. De novo motif extraction and amplification
5. Positional profile generation and comparison
6. Comparison against known motifs

3.3.1.1 Input data and pre-processing

As a heuristic inversion of MEPP, LMPP uses the same scored sequence data inputs as MEPP, as described previously (Chapter 2, Fig. 3.1A). Briefly, LMPP accepts scored sequence data in the form of a series of uniform-length sequences from a FASTA file format, where the sequence header is annotated with an assigned sequence score. Similar to MEPP, these sequences are expected to be sampled surrounding an interpretable feature, e.g. transcription start sites assayed from csRNA-seq. Degenerate sequences, sequences from repetitive regions, and sequences sampled from overlapping genomic intervals are similarly handled using protocols from MEPP.

Unlike MEPP, LMPP has the additional option of masking out motifs from an existing JASPAR-formatted motif file from the input sequence. This step is optional and allows users to filter out previously discovered motifs in order to recover more unique motifs in subsequent invocations of LMPP. LMPP recovers the positions of these “masking motifs” using MOODS [32], then replaces the bases with the special “blank” character “_”. This blank character serves as an exception to one-hot-encoding of the sequence, and is interpreted as an all zero vector at that position.

In order to compensate for the influence of GC bias on motif finding, we use least squares linear regression to regress out the effect of sequence GC content on the input sequence scores, retaining the residual as the GC-corrected sequence scores for the purposes of later steps. This step is optional and can be turned off.

3.3.1.2 Scaling positional profile by sequence scores

While MEPP accepts known motifs and outputs novel information in the form of positional enrichment profiles, LMPP reverses this process, accepting a target profile *P* to discover de novo motifs that have positional enrichment resembling *P* (Fig. 3.1A,B). Each position in *P* thus corresponds to a motif position in each input sequence S_i, similar to the positional profiles generated by MEPP. To ensure that positional information is incorporated into the training of the convolutional neural networks that LMPP uses to encode motifs, we convert each sequence score y_i into a scaled positional profile Y'_i:

$$Y'_i = P * y_i$$

For best results, values in the positional profile *P* must range from -1 to 1. Similar to the positional profiles in MEPP, positive values target positive enrichment at a given position, while negative values target negative enrichment at a given position. Values closer to zero indicate diminishing target enrichments.

3.3.1.3 Convolutional neural network training

To learn positional motifs, LMPP makes heavy use of a specialized convolutional neural network (CNN). Users specify the motif length *k* and a target number of motifs to learn, *n*. Similar to MEPP, users may also specify a motif margin *m* to account for loose tolerances in motif positioning. The network is then comprised of the following layers (Fig 3.2):

1. Input layer (IL) (one-hot encoded sequence)
2. Dinucleotide convolutional layer (DNCL) (kernel size 2, 32 kernels)
3. Motif convolutional layer (MCL) (kernel size *k*-1, *n* kernels)
4. Dropout layer (dropout rate 0.1)
5. Average pooling layer (window 2*m*+1, stride 1)
6. Zero padding (pad to input layer length)

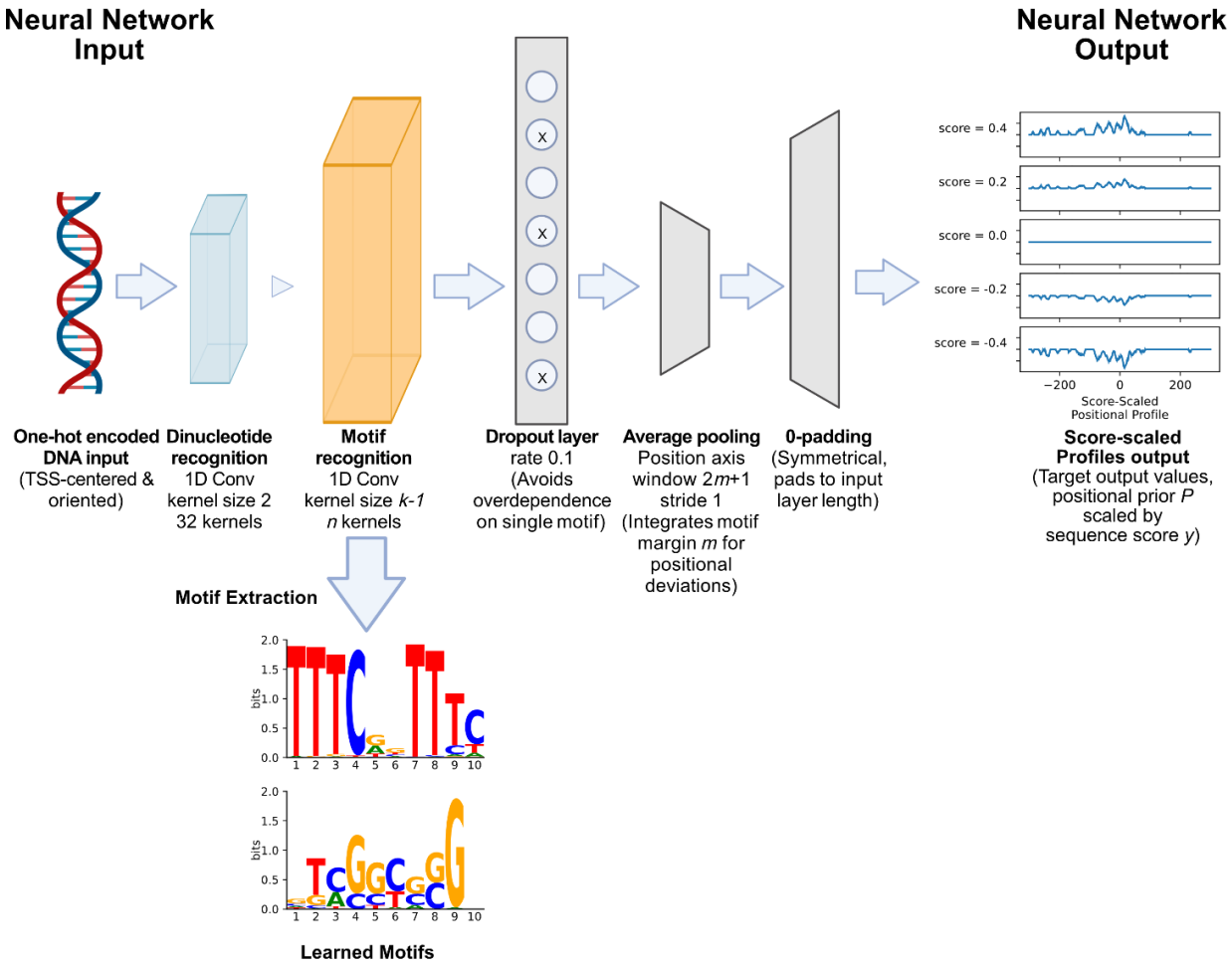


Figure 3.2: LMPP learns motifs using a specialized convolutional neural network

Diagram of LMPP's internal neural network, from right to left, consisting of two convolutional layers followed by auxiliary non-trained layers. Neural network inputs and outputs are indicated. TSS-centering is assumed here, although any centering feature may be used, similar to MEPP. Motifs are extracted from the second convolutional layer, where motifs are recognized.

Data fed into to the input layer consist of one-hot encoded DNA sequence, similar to MEPP or other machine learning tools that operate on DNA sequence. Similar to MEPP, all sequences are expected to share the same length and a common center feature, typically the TSS (Fig. 3.2). Exceptions to the one-hot encoding are: The degenerate nucleotide “N”, represented by a vector of even probabilities of all bases at 0.25; And the special motif-masked symbol ‘_’, which is represented by a vector of all zeroes.

The DNCL allows the network to account for dinucleotide presence and positional interdependence between adjacent bases of motifs. Meanwhile, the MCL is the principal layer where motifs are captured: activation values from this layer reflect motif recognition at valid positions from the input sequence. Both the DNCL and MCL use rectified linear (ReLu) activation functions, which are similar to linear functions, except that values below zero output to zero: This construction allows motifs that below a recognition threshold to be ignored, similar to log-odds detection thresholds for traditional motifs.

To prevent network accuracy from being too dependent on the positional recognition of any singular motif, the dropout layer randomly drops inputs received from the MCL to zero, at a rate of 0.1. The average pooling layer allows the network to tolerate deviations in the position of a motif: the motif margin *m*, may be set higher if applying LMPP to a lower resolution assay such as ChIP-seq. The zero padding layer allows for dimensional compatibility of network outputs and the scaled profiles, and sets the interpretation of motif positions relative to the center of the motif.

For learning motifs from data that lack strandedness information, e.g. ChIP-nexus, LMPP adjusts the network to account for dual motif orientations. Layers 1-3 are evaluated twice: once for the input sequence as-is to obtain verbatim motif recognition values, and once for a reverse complementation of the input sequence. The second output is then reversed, so that positional indices remain consistent, yielding motif recognition values for reverse complemented motifs from the MCL. Finally, a max pooling layer selects the maximum value for each motif at each position in the MCL from either the verbatim motif recognition values, or the reverse complemented motif recognition values. This renders the network internally invariant to reverse complementation.

In order to learn motifs that describe the input data while reflecting enrichment distributions from the positional prior, this network is trained to infer the scaled positional prior Y'_i from one-hot encoded input sequence S_i. The network is trained using an Adam optimizer to minimize the mean squared error (MSE) between its own output and the scaled positional prior Y'_i. By default, LMPP trains for a maximum of 1000 epochs, with an early stopping procedure that stops training when the MSE remains at a stable minimum for a set number of epochs (10 by default).

Because the goal of LMPP is to describe motifs within the input dataset rather than create a robust predictor from it, all input data is used for training, analogous to how all sequences of interest are passed to de novo motif learning algorithms such as HOMER.

3.3.1.4 De novo motif extraction and amplification

In order to recover usable position frequency matrix (PFM) format motifs from the trained CNN, we first evaluate layers 1-3 of the CNN on one-hot encoded sequences, recovering the indices and values of maximum motif recognition. We then extract the maximum activating one-hot encoded subsequences for each motif for each sequence, and scale them by their activation values (thus preventing poor contributions from). This weighted contribution is then summed on a per-motif basis, yielding non-normalized, weighted PFMs. These PFMs are then normalized so that values at each position sum to 1.

To increase Shannon information content (IC) from the normalized motifs, we calculate the IC for a PFM using Logomaker's transform_matrix function, then scale the normalized PFM so that the maximum IC on at least one position of the PFM approaches 2 [36]. This results in less degenerate motifs without modifying the CNN architecture or training.

Due to the dropout layer, some kernels may not return motifs with any information content. These are omitted from the output motifs.

LMPP packages the normalized and amplified motifs in a JASPAR-formatted motif PFM format, for compatibility with other bioinformatics tools including MEPP and MEIRLOP [103]. For ease of use, users may specify a custom prefix for motif IDs.

3.3.1.5 Positional profile generation and comparison

Because LMPP outputs motifs in JASPAR PFM format, and shares a scored sequence input format with MEPP, we can use MEPP to characterize the distribution of de novo motifs within the data. By running MEPP on the same scored sequence inputs as were submitted to LMPP, we can create MEPP heatmaps of where there de novo motifs appear, as well as enrichment positional profiles for these de novo motifs (Fig. 3.1B). This affords transparency as to where and how the discovered motifs appear in the dataset, rather than relying on black box machine learning interpretations.

To quantify the similarity between the target profile and the enrichment positional profiles of de novo motifs, we use the Pearson correlation coefficient of the profile values across all positions.

3.3.1.6 Comparison against known motifs

While LMPP can be executed without a dataset of known TF binding motifs, effective interpretation of de novo motifs should be grounded against known references. To this end, we re-implemented similar motif comparison methods used in versions 3.3 onwards of HOMER: We align de novo and known motifs against each other at multiple offsets and in both relative orientations (forward and reverse-complemented). At each alignment, we evaluate the Pearson correlation coefficient where the motifs overlap, and report the alignment with the greatest coefficient.

3.3.2 Ground truth motif recovery

One of the premises behind LMPP's implementation is that it can reverse the motif-to-enrichment-profile transformation achieved by MEPP (Fig 3.1A,B). To verify this premise, we ran MEPP on scored sequence datasets with known relevant motifs, and created positional priors from the enrichment positional profiles for those motifs, to test if LMPP could reconstruct the known relevant motif from its enrichment profile.

The testing procedure was performed in 4 parts, which were executed separately for each dataset:

1. Ground truth dataset generation
2. Known motif positional profile generation

3. Two-pass known motif recovery
4. Recovered vs. known motif comparison

Each ground truth dataset had its recovered ground truth motifs discovered and evaluated separately.

Table 3.1: Datasets for ground truth dataset generation

Ground Truth Dataset	Motif Margin	Strand Oriented	Motif to Recover	GEO Accession	Notes
fly-tss	2	Yes	INR	GSE203135	Delos Santos et al. 2022 (Reproduced in Chapter 2; Manuscript under review)
			DPE		
			TATA		
siRNA-nrf1-tss	2	Yes	NRF1	GSE199431	Duttke et al. 2022 (Manuscript Under Review)
siRNA-yy1-tss	2	Yes	YY1		
siRNA-nfy-tss	2	Yes	NFY	GSE115110	Oldfield et al. 2019 [11]
nanog-chip-nexus	5	No	Nanog	GSE137193	Avsec et al. 2021 [86]
oct4-chip-nexus	5	No	Oct4		
sox2-chip-nexus	5	No	Sox2		
klf4-chip-nexus	5	No	Klf4		

3.3.2.1 Ground truth dataset generation

We first created a series of scored sequence datasets from csRNA-seq, Start-seq, and ChIP-nexus experiments, for which relevant motifs are known. These datasets are summarized in Table 3.1, including references for their data generation methods.

For ChIP-nexus datasets, the scored sequences were generated as described in the supplemental methods for the previous chapter, MEPP. Briefly, we use data from the MACS2 [104] narrowpeak calls to locate the peak summits and peak signal values, then extract sequence +/- 200 bp of the peak summits. Sequences from peaks were scored by their corresponding peak signal values. A motif margin of 5 is used to account for less positional specificity in ChIP-nexus.

3.3.2.2 Known motif positional profile generation

We then ran MEPP on each dataset of scored sequences, profiling enrichment of known motifs, including the relevant target motifs. For most experiments here we use the same JASPAR-formatted HOMER motif dataset as we used for MEPP. The exception is ground truth analysis on *Drosophila* TSS, for which we use a JASPAR-formatted conversion of core promoter motifs from Ohler et al [80].

For compatibility with LMPP, the enrichment positional profile for each dataset's relevant motifs was converted into a positional prior by scaling the magnitude of the most extreme maximum or minimum to an absolute value of 1. We term the resulting profile a ground truth positional prior.

3.3.2.3 Two-pass known motif recovery

For each scored sequence dataset, we used LMPP to learn motifs while incorporating the previously generated ground truth positional prior. Because these positional priors were specific to a single motif, they recovered multiple similar de novo motifs. To explore more diverse outputs, we performed a second pass to learn more motifs with LMPP, this time masking out the motif whose enrichment positional profile was most correlated with the ground truth positional prior. LMPP Each pass with LMPP was configured to learn a maximum of 10 motifs 10 bp long, over a maximum of 1000 epochs with early stopping.

3.3.2.4 Recovered vs. known motif comparison

Upon completion of both LMPP passes, we then derived MEPP profiles for the de novo motifs, then recovered the de novo motif with the enrichment profile most correlated with the ground truth positional prior. We compared the similarity of this recovered motif against the known ground truth motif, using the motif comparison procedure previously described.

3.3.3 GWAS-informed de novo motif discovery from COVID-19 TSS data

To demonstrate the ability of LMPP to integrate biomedical data, we performed a de novo motif analysis of transcription start sites found from csRNA-seq of COVID-19 patients [105]. In order to focus the motif search on sequence features relevant to COVID-19 induced lung damage, we incorporate a positional profile derived from a COVID-19 GWAS study comparing respiratory symptoms after infection.

3.3.3.1 COVID-19 csRNA-seq analysis

We used csRNA-seq data from whole white blood cells isolated from blood samples drawn from COVID-19 patients, as collected, generated, and described by Lam et al. [105]. Our analysis proceeds after read alignment and tag directory generation for downstream analysis with the HOMER suite. No identifiable sequence data was processed, and de-identified donor codes were used.

Starting from the generated tag directories, we used the HOMER script “getTSSfromReads.pl” to identify individual TSS from all csRNA-seq samples, after controlling for potential false positives using csRNA-seq input control samples, requiring a minimum raw coverage of 7 5’ read ends to call a TSS.

We then quantified the called TSS for all samples, using HOMER’s “annotatePeaks.pl” script, itself a wrapper around DESeq2. We derived both raw counts and quantifications normalized using DESeq2’s variance-stabilizing transformation [88].

To mitigate severe outlier TSS, analysis proceeded only with TSS possessing at least 10 reads in 10 samples. In order to remove the effect of confounding variables, including batch effects, VST-normalized quantifications were further processed using PEER [106,107], accounting for donor codes sequencing date batch effects.

3.3.3.2 Correlation of csRNA-seq TSS with lung health

Using data from Lam et al., csRNA-seq samples were annotated with lung injury severity data for the patient around the date of sampling, quantified as a Modified Murray Lung Injury Score (MMLIS), as previously described by Lam et al. [105]. In order to determine the extent by which transcription from a TSS increased or decreased with COVID-19 induced lung damage, we scored TSS by the Spearman correlation of its expression and MMLIS across all infected samples.

We sampled sequence +/- 300 bp of each correlation-scored TSS from the hg38 human reference genome, then converted these scored TSS sequences into a scored FASTA format compatible with MEPP and LMPP, where sequence headers consist of the TSS ID and the correlation score, separated by a space.

3.3.3.3 Positional prior generation from GWAS SNPs

In order to incorporate positional information on functional sequence features disrupted by genetic variation, we constructed a positional prior using GWAS SNP data from the COVID-19 Host Genetics Initiative, as provided from GRASP [96,97] (Fig. 3.3A). For parity against our correlation-scored TSS, we used GWAS results comparing individuals with very severe COVID-19 respiratory symptoms against individuals who were not hospitalized for COVID infection. This ensured that both the scored TSS and the positional profile characterized features COVID-19 induced lung injury.

To construct a TSS-centric centric positional prior, we calculated the positions of SNPs relative to each scored TSS, accounting for the strand orientation of the TSS (Fig. 3.3B). Thus, SNPs upstream of the TSS consistently had a negative relative position assigned, while SNPs downstream of the TSS consistently had a positive relative position assigned. For each TSS-centric position within +/- 300 bp, we summed a significance-weighted contribution from each SNP's GWAS p-value p, calculated as $-\log_{10}(p)$ if $p <= 0.05$, 0.0 otherwise. We smoothed the profile using a 3rd-degree polynomial Savitzky-Golay filter with a window size of 29. To retain only parts of the profile with high signal, we used a high-pass filter that set below-average profile values to 0. The final GWAS-based positional prior was then scaled so that the maximum of absolute values from the profile was equal to 1, for compatibility with LMPP (Fig. 3.3C).

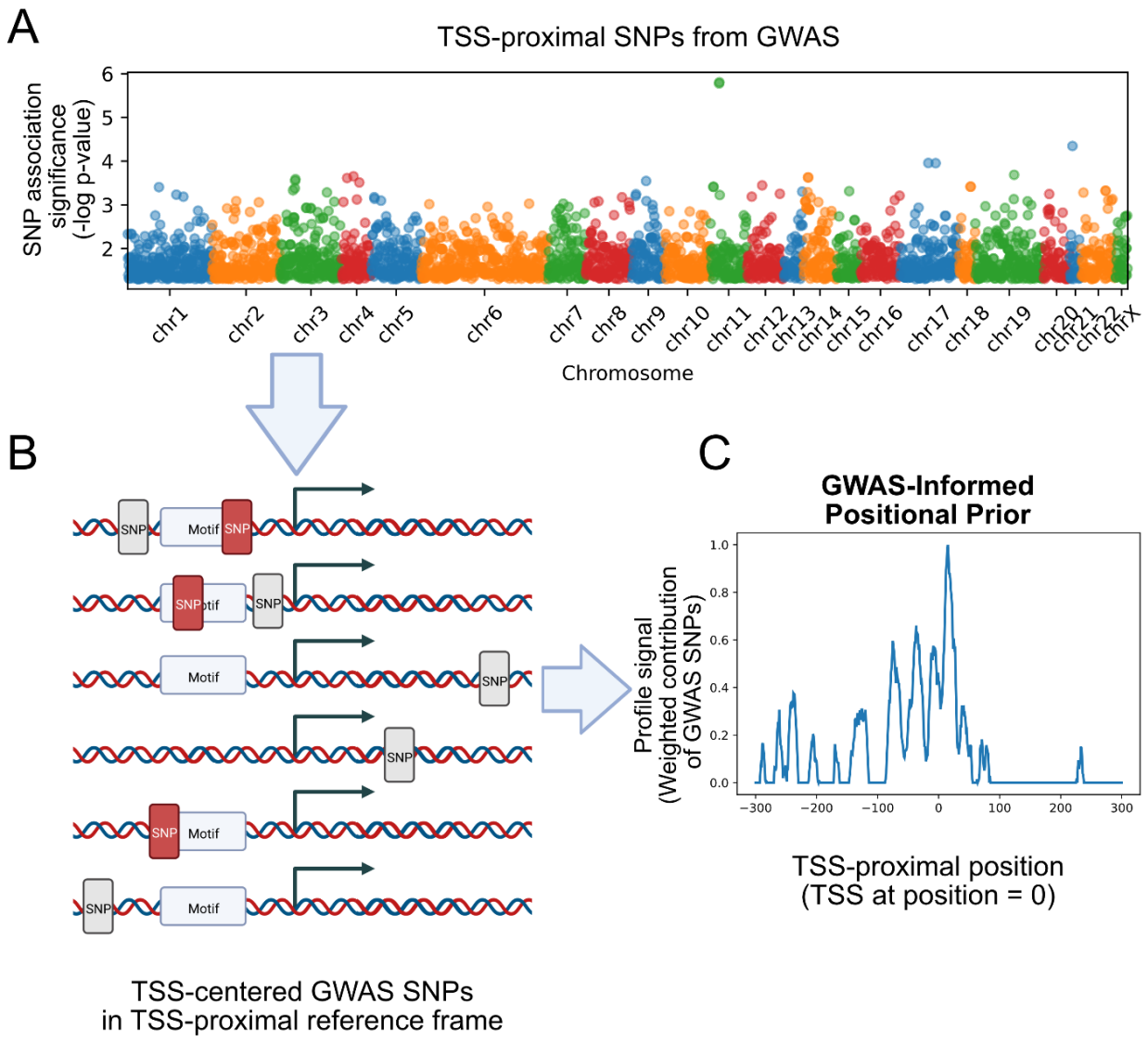


Figure 3.3: Schematic for creation of a positional prior from GWAS SNPs

(A) Manhattan plot of GWAS SNPs found within +/- 300 bp of TSS
(B) Orientation and alignment of TSS-proximal SNPs within TSS-proximal reference frames
(C) A GWAS-informed positional prior created using summed contributions from GWAS SNPs at each TSS-proximal position, weighted by the significance of their association in the GWAS study. The final profile is smoothed using a SavGol filter and normalized to a maximum absolute value of 1.0

3.3.3.4 De novo motif discovery from GWAS positional prior

To search for positionally enriched de novo motifs, we ran LMPP on the dataset of correlation-scored TSS sequence, using the GWAS-based positional prior to optimize for enrichment at TSS-centric locations impacted by SNPs associated with severe respiratory symptoms after COVID-19 infection.

We ran LMPP in two passes. For each pass, we trained the CNN to learn a maximum of 10 positionally enriched motifs 10 bp in length, after masking out motifs learned in previous passes from the sequences. To characterize the positional enrichments of these learned motifs, we ran MEPP using the same scored TSS sequence dataset. We also compared the learned motifs against known motifs from HOMER.

3.3.4 Characterization of positional prior effects

In order to confirm that the choice of positional prior affects the motifs LMPP learns from data, we created six stereotypical synthetic positional priors, each describing broad positional preferences. These profiles are described in Table 3.2, and are generated as linearly interpolated line segments between the tabulated positions and heights. We then ran LMPP using each synthetic positional prior in two passes to learn a maximum of 10 positionally enriched motifs 10 bp in length. In the second pass, LMPP learned motifs after masking out motifs learned from the first pass from the sequences.

To characterize the motifs learned by MEPP using each synthetic positional prior, we compared the learned motifs against known motifs from HOMER. For each synthetic positional prior, we filtered for motif matches exceeding a correlation of 0.8, then counted the number of matched unique known motifs from each HOMER-annotated motif family. We then normalized these counts by the number of motifs in each motif family. The use of broad motif families prevents subtle nuances in learned motifs from being overinterpreted, accounting for the similarities between motifs within a family.

Table 3.2: Numerical description of synthetic positional priors

Profile Name	Start (TSS-relative) (bp)	End (TSS-relative) (bp)	Locations (TSS-relative) (bp)	Profile Heights
linear_up_down_wide_600bp	-300	300	-300,-100,0,1,100,300	0,0,-1,1,0,0
linear_up_down_narrow_600bp	-300	300	-300,-25,0,1,25,300	0,0,-1,1,0,0
square_up_down_600bp	-300	300	-300,-100,0,1,100,300	-1,-1,-1,1,1,1
linear_down_up_wide_600bp	-300	300	-300,-100,0,1,100,300	0,0,1,-1,0,0
linear_down_up_narrow_600bp	-300	300	-300,-25,0,1,25,300	0,0,1,-1,0,0
square_down_up_600bp	-300	300	-300,-100,0,1,100,300	1,1,1,-1,-1,-1

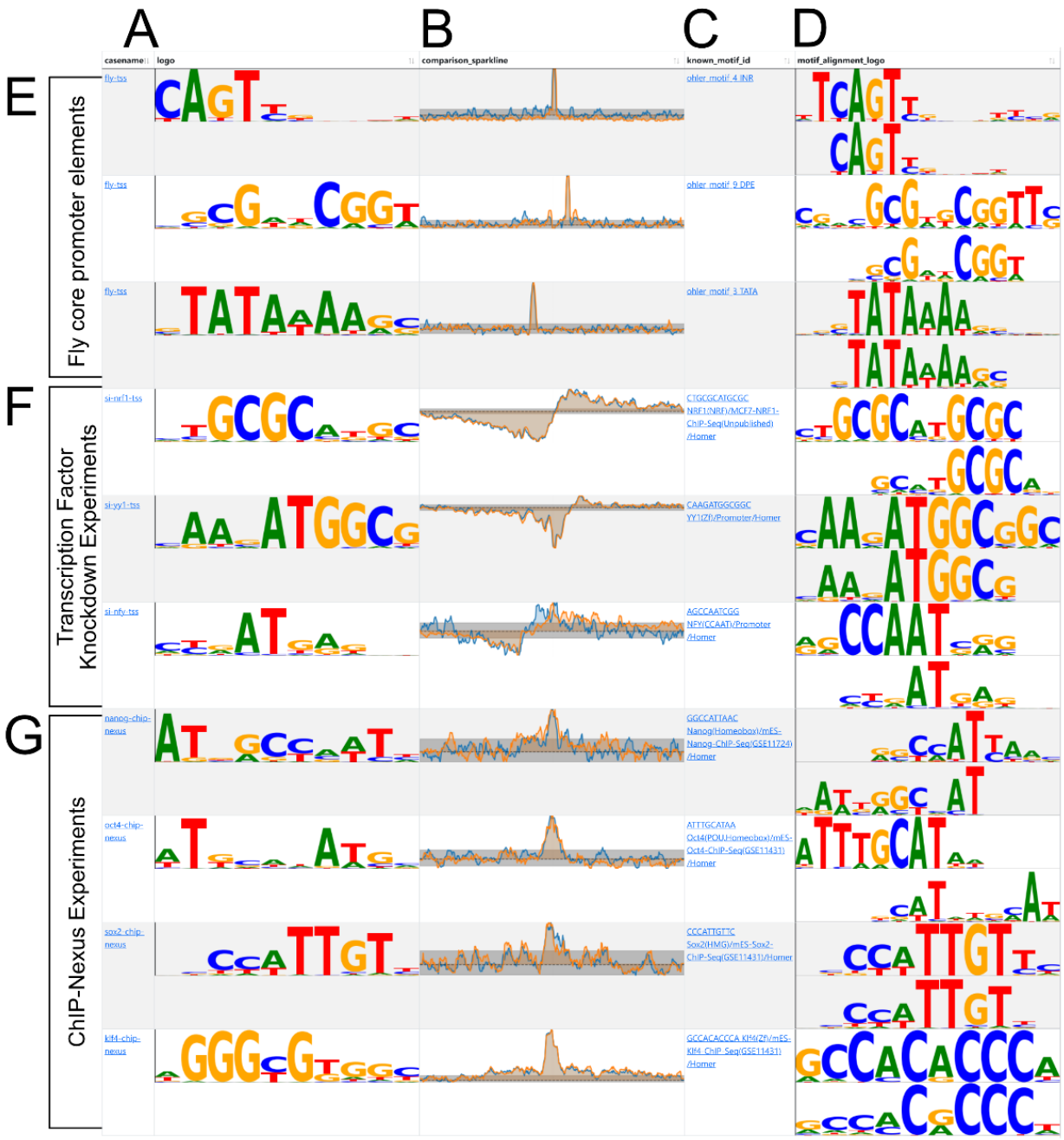
3.4 Results

3.4.1 LMPP recovers motifs from positional enrichment profiles

In order to demonstrate that LMPP learns motifs that have positional enrichment corresponding to the input positional prior, we tested its ability to recover known motifs from positional priors derived from their enrichment positional profiles. Overall, we tested 3 known *Drosophila* core promoter motifs from *Drosophila* TSS, 3 binding motifs for knocked down TFs from differential transcription initiation experiments, and 4 binding motifs for ChIPped TFs from ChIP-nexus experiments.

Figure 3.4: LMPP can recover known motifs from their enrichment positionality profiles

(A) The first two columns display the ground truth case name and the logo of the de novo motif recovered by LMPP.
(B) Double sparkline plot comparing the positional prior (in orange) and the enrichment positionality profile (in blue, generated by MEPP) of the de novo motif recovered by LMPP. For legibility, both profiles have been scaled to the same minimum and maximum, and the area between each profile and the line y=0 has been shaded. In addition, the 95% confidence interval of the enrichment positionality profile has been shaded in gray.
(C) The ID of the known ground truth motif LMPP was tasked to recover. This enrichment positionality profile of this motif (as generated by MEPP) was scaled to provide the positional prior input to LMPP.
(D) Dual motif alignment logos of the known ground truth motif (top), and the de novo motif recovered by LMPP (bottom).
(E) Table rows corresponding to results for recovering the motifs of *Drosophila* core promoter elements from their enrichment positional profiles surrounding *Drosophila* TSS. The TSS were scored by their transcription signal as measured by csRNA-seq.
(F) Table rows corresponding to results for recovering the binding motifs of knocked-down transcription factors from their enrichment positional profiles surrounding. The TSS were scored by their differential transcription between knockdown and control conditions.
(G) Table rows corresponding to results for recovering the binding motifs of ChIPed transcription factors from their enrichment positional profiles. The TSS were scored by their ChIP-nexus peak signal scores.


A
B
C
D
casename
logo
comparison_sparkline
known_motif_id
motif_alignment_logo
E
Fly core promoter elements
fly-tss
ohler_motif_4_INR
fly-tss
ohler_motif_9_DPE
fly-tss
ohler_motif_3_TATA
F
Transcription Factor Knockdown Experiments
si-nrf1-tss
CTGCGCATGCGC NRF1(NRF)/MCF7-NRF1-ChIP-Seq(Unpublished) /Homer
si-yy1-tss
CAAGATGGCGGC YY1(Zf)/Promoter/Homer
si-nfy-tss
AGCCAATCGG NFY(CCAAT)/Promoter /Homer
G
ChIP-Nexus Experiments
nanog-chip-nexus
GGCCATTAAC Nanog(Homeobox)/mES-Nanog-ChIP-Seq(GSE11724) /Homer
oct4-chip-nexus
ATTTGCATAA Oct4(POU,Homeobox)/mES-Oct4-ChIP-Seq(GSE11431) /Homer
sox2-chip-nexus
CCCATTGTTC Sox2(HMG)/mES-Sox2-ChIP-Seq(GSE11431)/Homer
klf4-chip-nexus
GCCACACCCA Klf4(Zf)/mES-Klf4-ChIP-Seq(GSE11431) /Homer

We used LMPP to recover the INR, DPE, and TATA-box motifs using their enrichment profiles as positional priors across over 40K *Drosophila* TSS (Fig. 3.4E). LMPP analyzed 44.9K TSS from *Drosophila* embryos, learning a maximum of 20 motifs (2 passes of 10 motifs) of length 10 bp, and returning the motif with an enrichment profile most correlated with the positional prior. Overall, the recovered motifs bear a strong resemblance ($r > 0.94$) to the ground truth motifs via correlation of their motif matrices (Fig. 3.4A,D). In addition, the positional profiles of these recovered motifs correlate with the positional priors ($r > 0.80$) (Fig. 3.3B).

We repeated the procedure to recover binding motifs for NRF1, YY1, and NFYA, from differential transcription initiation experiments characterizing the siRNA knockdown of their respective transcription factors (Fig. 3.4F). We used the enrichment profiles of each TF's binding motif across the TSS as a positional prior for LMPP. For NRF1, and YY1, these were data from csRNA-seq assays, while for NFYA, this was a Start-seq assay [11]. Similar to the previous results on *Drosophila* TSS, we find that the recovered NRF1 and YY1 motifs match their ground truth counterparts ($r > 0.97$) (Fig. 3.4D). However, recovery of the NFYA motif from its enrichment profile only returned a limited resemblance ($r = 0.77$). Similarly, enrichment profiles for the recovered NRF1 and YY1 motifs are highly correlated with the positional priors ($r > 0.97$), the enrichment profile for the recovered NFYA motif does not fully reflect that of the original known motif ($r = 0.64$) (Fig. 3.4B).

The previous ground truth recovery tasks demonstrate motif recovery from TSS profiling data, and included strand information to orient sequences and motif distributions. In order to test the capability of LMPP to recover motifs from data that does not include strand information, we used LMPP to recover the binding motifs of ChIPed transcription factors from ChIP-nexus data (Fig. 3.4G). In this mode, the LMPP convolutional neural network was configured to recognize motifs in a strand-invariant manner. Recovery of Nanog, Sox2, and Klf4 binding motifs from their enrichment profiles across ChIP-nexus summits was successful, with high correlation between known and recovered motifs ($r > 0.91$) (Fig. 3.4D). Oddly, while the recovered Nanog binding motif matched the known motif ($r = 0.95$), its enrichment profile did not match that of the known motif strongly ($r = 0.59$) (Fig. 3.4B). The Oct4 binding motif, on the other

hand, had a more qualified resemblance to the known motif, in both the motif matrix (r = 0.87), and the enrichment positional profile (r = 0.85).

These results demonstrate the LMPP is capable of recovering motifs from their positional enrichment profiles in almost all cases, indicating that it reverses the operation previously described for MEPP. Thus, we validate LMPP from ground truths as a form of de-novo positional motif enrichment, indicating its capabilities in recovering positionally constrained motifs from user-defined data.

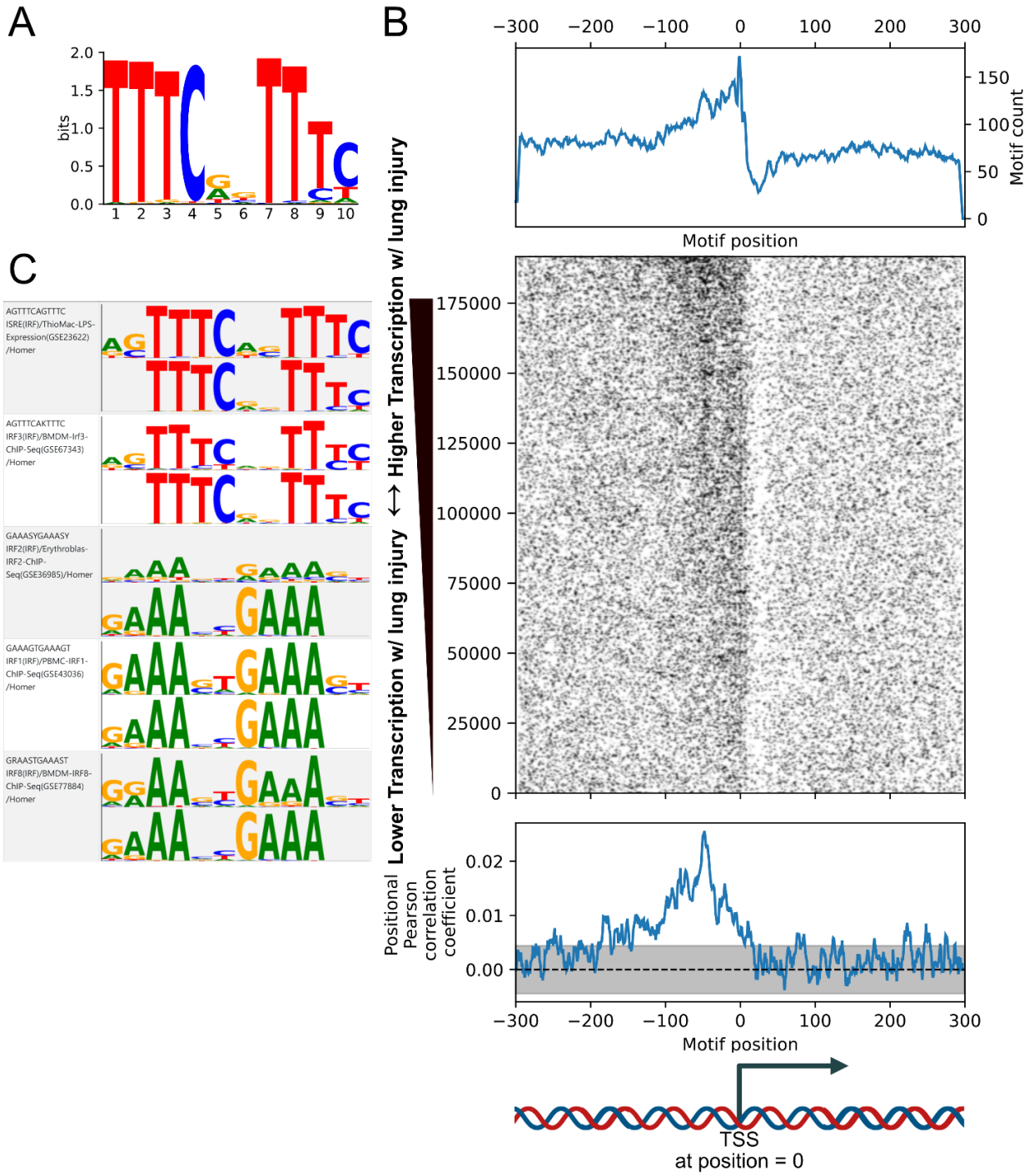
3.4.2 LMPP recovers motifs relevant to COVID-19 infection

To identify motifs whose modification by natural variation could impact COVID-19 respiratory symptoms, we proceeded to use LMPP to determine motifs positionally enriched around both TSS and relevant GWAS SNPs. Our analysis was performed on sequences extracted +/- 300 bp from over 175K TSS, which we scored by the correlation of their expression against patient lung injury (as measured by a modified Murray score): Higher scores indicated TSS that correlated with increased lung injury, while lower scores indicate TSS correlated with decreased lung injury. In order to direct the motif search towards the TSS-relative positions of relevant GWAS SNPs, we derived a positional prior from the contribution of 3274 SNPs found +/- 300 bp of over 175K TSS. We weighted the contribution of these SNPs to the positional prior by the significance ($-\log_{10}$ P-value) of their association against whether subjects experienced severe COVID-19 respiratory symptoms, or did not require hospitalization for COVID infection.

LMPP analyzed both the set of scored TSS sequences and the GWAS-derived positional prior. LMPP's neural network was trained to recognize at least 20 motifs (10 motifs each pass, for 2 passes) that maximized enrichment following the positional prior, directing motif enrichment around more significant GWAS SNPs.

Figure 3.5: LMPP discovers the ISRE motif using a GWAS SNP-based positional prior

(A) Motif logo for the de novo ISRE motif discovered by LMPP.
(B) Partial MEPP plot of the ISRE- motif discovered with LMPP. The motif count over position, motif heatmap, and enrichment positional profile are shown in that order. X-axis coordinates are relative to the TSS. TSS are scored by increasing vs. decreasing transcription correlated with lung injury.
(C) Dual motif alignment logos displaying similarity between the de novo motif and other binding motifs in the IRF family, including the ISRE.

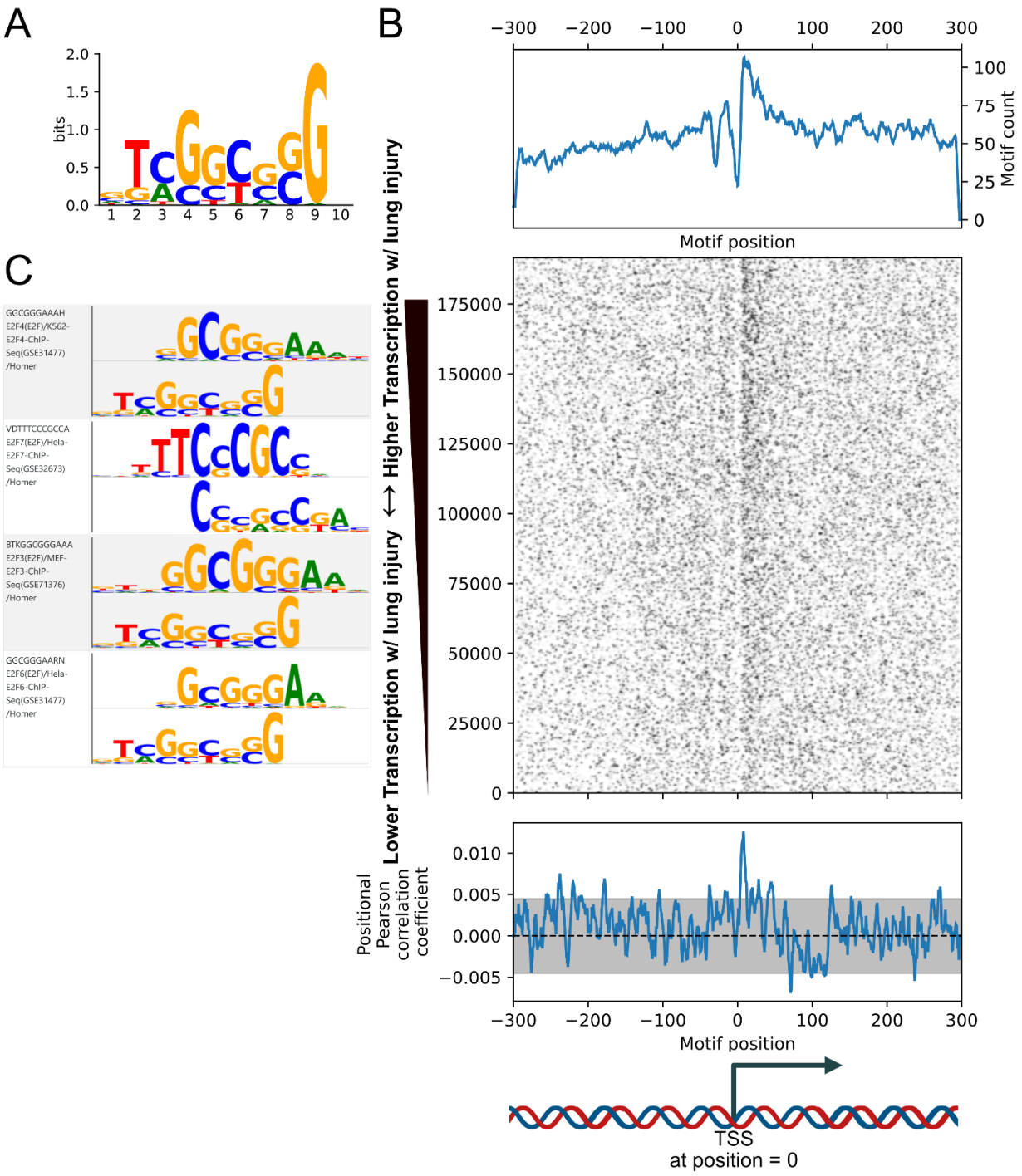

A
bits
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10
B
−300 −200 −100 0 100 200 300
150
100
50
0
Motif count
Motif position
Lower Transcription w/ lung injury ↔ Higher Transcription w/ lung injury
175000
150000
125000
100000
75000
50000
25000
0
Positional Pearson correlation coefficient
0.02
0.01
0.00
−300 −200 −100 0 100 200 300
Motif position
TSS
at position = 0
C
AGTTTCAGTTTC
ISRE(IRF)/ThioMac-LPS-Expression(GSE23622)/Homer
AGTTTCAKTTTC
IRF3(IRF)/BMDM-Irf3-ChIP-Seq(GSE67343)/Homer
GAAASYGAAASY
IRF2(IRF)/Erythroblas-IRF2-ChIP-Seq(GSE36985)/Homer
GAAAGTGAAAGT
IRF1(IRF)/PBMC-IRF1-ChIP-Seq(GSE43036)/Homer
GRAASTGAAAST
IRF8(IRF)/BMDM-IRF8-ChIP-Seq(GSE77884)/Homer

The most positionally enriched (adjusted p-value < 0.001) motif recovered by LMPP strongly matched (r = 0.98) that of the interferon stimulus response element (ISRE) (Fig. 3.5A,C). As expected from the positional prior, this motif is positively enriched upstream of the TSS (Fig. 3.5B), in a direction associated with higher transcription with increased COVID-19 induced lung injury. The ISRE is a known binding site for multiple factors, such as IRF3 and ISGF3 [108]. IRF3 is known to bind to ISRE in response to TLR3 activation upon recognition of viral double-stranded RNA. The SARS-CoV-2 spike protein interacts with IRF3, inhibiting this response [109]. Meanwhile, ISGF3 is activated by type I interferon signaling, and its binding to ISRE gives rise to multiple antiviral responses [110]. Among these responses is the transcriptional upregulation of *MxA*, which encodes a protein known to sequester viral components [94,111]. Thus, we find that LMPP's neural network has identified a motif known to play important roles in antiviral response, which if disrupted by natural variation, can lead to differential health outcomes.

Another positionally enriched (adjusted p-value < 0.001) motif identified by LMPP, which does not resemble binding to IRF-family transcription factors, instead resembles binding motifs for the E2F family, mainly E2F4 (r = 0.78) (Fig. 3.6A,C). The positive enrichment of this motif, which occurs slightly downstream of the TSS, is consistent with increased transcription COVID-19 lung injury increases (Fig 3.6B). E2F cell-cycle related targets were also previously found to be an enriched set of genes when comparing gene expression between COVID-19 patients with and without cancer [112]. The discovery of E2F motifs from this data is corroborated by the analysis of Lam et al., which also found enrichment of E2F motifs in regulatory regions associated with increased lung injury [105]. This corroborated enrichment could be explained by transcriptional regulation of the cell cycle to affect immune cell populations in response to infection [105]. However, the resemblance of the de novo motif is limited, potentially as a result of attempting to match multiple E2F family motifs at once: While the correlation of this motif is highest against the E2F4 motif, the information content more closely resembles that of the E2F6 motif. This reflects a potential weakness of our method when identifying smaller numbers of motifs in a single pass: A single motif may be diluted when generalizing to multiple subsequences within a family of relevant motifs. However, this still reflects LMPP's ability to learn motifs that are likely to be relevant to viral activity and immune response.

Figure 3.6: LMPP discovers an E2F binding motif using a GWAS SNP-based positional prior

(A) Motif logo for the de novo E2F-like motif discovered by LMPP.
(B) Partial MEPP plot of the E2F-like motif discovered with LMPP. The motif count over position, motif heatmap, and enrichment positional profile are shown in that order. X-axis coordinates are relative to the TSS. TSS are scored by increasing vs. decreasing transcription correlated with lung injury.
(C) Dual motif alignment logos displaying similarity between the de novo motif and other binding motifs in the E2F family, including the E2F4 and E2F6.


A
bits
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10
B
−300 −200 −100 0 100 200 300
100
75
50
25
0
Motif count
Motif position
Lower Transcription w/ lung injury ↔ Higher Transcription w/ lung injury
175000
150000
125000
100000
75000
50000
25000
0
0.010
0.005
0.000
−0.005
Positional Pearson correlation coefficient
−300 −200 −100 0 100 200 300
Motif position
TSS
at position = 0
C
GGCGGGAAAH E2F4(E2F)/K562-E2F4-ChIP-Seq(GSE31477)/Homer
VDTTTCCGCCA E2F7(E2F)/Hela-E2F7-ChIP-Seq(GSE32673)/Homer
BTKGGCGGGAAA E2F3(E2F)/MEF-E2F3-ChIP-Seq(GSE71376)/Homer
GGCGGGAARN E2F6(E2F)/Hela-E2F6-ChIP-Seq(GSE31477)/Homer

To verify that SNPs associated with differences in COVID-19 induced respiratory symptoms landed either on or near these motifs (as either causal SNPs or potentially indirectly associated SNPs), we proceeded to scan for these motifs in sequence samples +/- 10 bp of the SNP locations. For the ISRE-like motif, 14 SNPs (with nominal p-values < 0.05) were identified having this motif within +/- 10 bp, with 10 direct hits. For the E2F-like motif, there were 21 SNPs (with nominal p-values < 0.05) identified with the motif within +/- 10 bp, with 18 direct hits. Overall, these results indicate that LMPP is able to identify motifs likely to be relevant to SARS-CoV-2 infection, given the construction and use of a positional prior to direct the learning process in a data-driven manner. Variants on and around these motifs could present targets for specialized genotype arrays for predicting COVID-19 symptom severity and risk factors, enabling fine mapping studies to be informed by otherwise disparate genotyping and transcriptomic data.

3.4.3 Positional priors direct motif discovery

Due to the deliberately limited nature of LMPP's neural network architecture, we expect that the positional prior is an important input to discover relevant families of binding motifs, and that the choice of positional prior affects the motifs discovered. To verify this expectation, we created a set of six stereotypical positional priors, with pairs of these positional priors being inverted versions of each other. One pair corresponded to positional priors approximating Heaviside step functions, while the other two pairs described rising zig-zag patterns of varying width. We invoked LMPP in a two-pass process to discover motifs from the same COVID-19 TSS dataset analyzed previously, but using each stereotypical positional prior rather than only using the GWAS-based one. To avoid over-interpreting minute changes in learned motifs, we counted the number of similar ($r >= 0.8$) motifs learned from each motif family, normalized by the total number of motifs in each family. We used the HOMER known motif dataset as the basis for known motifs, and to annotate the relationship between motifs and motif families.

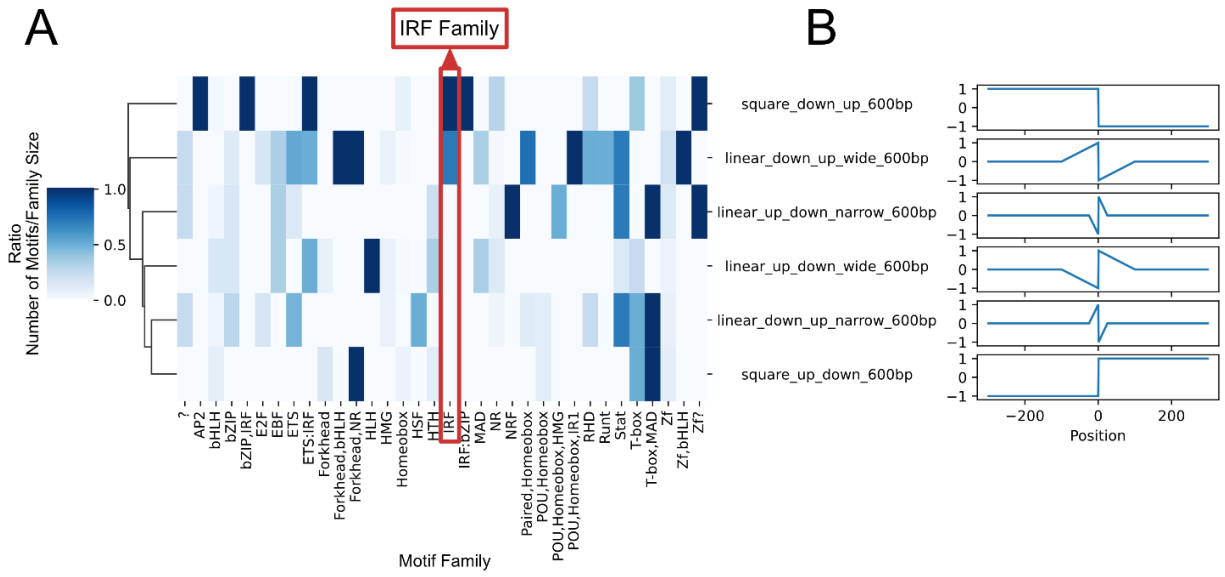


Figure 3.7: LMPP discovers different motif families when given different positional priors

(A) A clustermap of the amount of motifs from each motif family learned, when LMPP is given different stereotypical positional priors. The number of motifs learned from each family is normalized by the number of motifs in that family. The IRF motif family is highlighted in red boxes. Motifs were learned from a dataset of TSS scored by their correlation against increased lung injury.

(B) Line plots of each stereotypical positional prior. Note that motifs from the IRF family were learned by LMPP only for positional priors featuring a strong upstream positive enrichment target.

We find that each positional prior yields a different amount of motifs learned from each family. In particular, only two of the stereotypical positional priors used allowed LMPP to learn motifs resembling those of the IRF family (Fig 3.7A), despite the importance of IRFs to viral response and the phenotype analyzed in this dataset [113]. However, we note that the two positional priors that direct for broad positive enrichment upstream of the TSS do recover IRF family motifs, similar to the GWAS-based positional prior (Fig 3.7B). Yet, there was little correlation between motif families learned from each positional prior (maximum Spearman r = 0.50).

Overall, we find that by substituting different positional priors and recovering different families of motifs, LMPP does utilize these positional priors to direct the process of learning motifs. This implies that selection of data-driven positional priors would allow researchers to use LMPP to perform novel queries for de novo motifs within their data.

3.5 Discussion

By incorporating a positional prior, LMPP enables researchers to use other data within TSS-proximal reference frames to determine positionally enriched de novo motifs. LMPP effectively reverses the enrichment positional profiling operation performed by MEPP: While MEPP takes a known motif and creates a profile describing its enrichment in a dataset of scored sequences across different positions relative to e.g. TSS; LMPP will take a positional prior and search for motifs in a dataset of scored sequences, maximizing motif recognition in the locations and correlations specified by the prior. When applied to ground truth datasets to recover a known motif from its enrichment profile, we find that LMPP can recover the known motif in most cases, demonstrating its inversion of the operation performed by MEPP. Our results demonstrate this motif recovery operation on a suite of ground truths, including the positions of *Drosophila* core promoter elements, differential transcriptional assays of TF knockdown, ChIP-nexus assays for TF binding.

LMPP departs from other methods of de novo motif discovery in a manner similar to MEPP and MEIRLOP, by using biological sequence scores directly, rather than requiring that researchers simplify biological signal into boolean sets to satisfy the requirements of Gibbs sampling or expectation maximization models. In addition, while LMPP outputs motifs in JASPAR format for compatibility with

other downstream tools, LMPP's neural network incorporates dinucleotide interdependencies between adjacent nucleotides in its internal motif model, reflecting the need to control for dinucleotide sequence content as previously demonstrated by MEIRLOP. In addition, while de novo motif finding tools such as HOMER search for motifs enriched in a fixed direction without regard for their position, LMPP allows researchers to drive the discovery of new motifs using other data that can describe their positional enrichment.

We demonstrate this potential for discovery by translating GWAS SNPs into a TSS-proximal reference frame. This created a positional prior that directs motif discovery towards TSS-proximal loci associated with differences in the severity of COVID-19 induced respiratory symptoms. We find that by using this positional prior, LMPP learned motifs with known roles in antiviral response and replication. We also found that these motifs were found in proximity to GWAS SNPs with nominal associations to COVID-19 respiratory symptom severity. These results indicate LMPP as a promising means of identifying binding motifs for causal inference of GWAS SNPs. The motifs discovered by LMPP using GWAS-informed positional priors could be used to guide the creation of specialized genotyping arrays, to facilitate fine mapping studies and the creation of genetic risk scores.

We further demonstrate that the ability of LMPP to recover relevant motifs is a function of the constructed positional prior. In doing so, we find that not all positional priors can recover the IRF family of motifs relevant to the scored TSS from COVID-19 patients: This indicates the specificity of those motifs in our results to the component of our positional prior that specifies increased positive enrichment of motifs upstream and near the TSS. Thus, researchers can expect data-driven construction of the positional priors to guide LMPP to discover relevant motifs, allowing the integration of otherwise disparate datasets. In doing so, researchers may better direct the identification of disease-relevant transcription factors and regulatory networks by leveraging multiple positional priors derived from previous studies, e.g. related GWAS or ChIP-seq experiments: This would better contextualize the discovered motifs and their positional constraints, allowing for better interpretation of binding motif grammars and their associated networks of transcriptional regulation for diseases under study.

LMPP's limited convolutional neural network architecture is intended to encode a more transparent representation of the sequence features it learns. This property is enhanced by MEPP's

heatmap visualization of the distribution of motifs within a dataset of regulatory region sequences, allowing researchers to directly visualize where discovered motifs occur. This presents the opportunity for future work to employ these methods to simplify and visualize the rules of regulatory region motif grammar as learned by more involved machine learning methods, such as generative adversarial networks for synthetic promoter design [114]. This would allow such generative systems to represent less opaque “black boxes” to researchers, presenting the roles of motif grammar in a manner more amenable to hypothesis generation.

3.6 Conclusion

Taken together, we find that LMPP is a novel application of machine learning that allows researchers to find de novo motifs by incorporating sequence and positional data from otherwise disparate datasets. Similar to our previous work with MEIRLOP and MEPP, LMPP retains the ability to analyze sequences from genomic regions that have been scored across a continuum of two extremes of biological interest. Unlike many other de novo motif enrichment methods, LMPP does not search for motifs that are globally enriched towards either extreme, but instead uses the positional prior to specify targeted local motif enrichments. While MEPP previously relied on a fixed motif library, LMPP removes this restriction, taking advantage of a similar yet distinct convolutional neural network architecture to learn positionally enriched motifs. As such, we have made it available as an extension to MEPP.

3.7 Acknowledgements

Figures 3.2, 3.3, 3.5, and 3.6 were created with BioRender.com.

Chapter 3 includes analysis of data that is derived downstream of analysis from the following preprint:

Lam MTY, Duttke SH, Odish MF, Le HD, Hansen EA, Nguyen CT, Trescott S, Kim R, Deota S, Chang MW, Patel A, Hepokoski M, Alotaibi M, Rolfsen M, Perofsky K, Warden AS, Foley J, Ramirez SI, Dan JM, Abbott RK, Crotty S, Crotty Alexander LE, Malhotra A, Panda S, Benner CW, Coufal NG. “Profiling Transcription Initiation in Peripheral Leukocytes Reveals Severity-Associated Cis-Regulatory Elements in Critical COVID-19” bioRxiv. 2021.

It has been used in this work with the permission of M.T.Y Lam. The dissertation author was the primary author of this chapter.

Conclusion

We have presented a series of three novel bioinformatics methods for motif enrichment analysis (MEA), and demonstrated their utility on datasets relevant to immune response. These methods address oversimplifications or omissions of key axes of interest to MEA: MEIRLOP expands on score-based MEA, allowing researchers to use biologically derived scores to characterize regulatory regions, rather than oversimplify those scores into threshold-based sets. MEPP transparently visualizes and profiles the enrichment of motifs within regulatory regions in two axes: biologically derived scores, and position relative to relevant anchor points such as TSS. Finally, LMPP inverts the paradigm presented by MEPP, by allowing users to discover de novo motifs whose positional enrichment follows a positional profile that can be informed by otherwise disparate datasets.

Although we demonstrate these methods on datasets that involve immune response pathways, these tools remain broadly applicable to studying other aspects of human health and biology. MEIRLOP itself has seen use in other published research efforts both before [66] and after [115] its own standalone publication. This is owed to the flexibility and verisimilitude afforded by the fact that users need not threshold biologically relevant scores into discretely thresholded sets in order to determine motif enrichment, preserving the underlying truth of degrees of biological response.

Although currently in review, MEPP itself is already being used as part of a larger ongoing effort to characterize how multiple transcription factors can mediate varying levels of transcriptional activation or repression depending on their distance to transcription start sites, giving rise to a spatial grammar of transcriptional regulation (Duttke et al. 2022, manuscript in review). In that same work it is used as both a data visualization and hypothesis generation tool: In particular, its profiling of the relationship between motif distance and TSS activity has been corroborated using a massively parallel reporter assay that captures transcription start site activity across synthetic promoters, called “TSS-MPRA” (Duttke et al. 2022, manuscript in review). We expect that when combined with LMPP to discover de novo motifs that approximate a target relationship between motif distance and TSS activity, the potential for hypothesis generation and discovery should only increase.

MEPP and LMPP both internally leverage convolutional neural networks, but in a much more constrained manner than larger deep learning models such as DeepBind or DeepSEA [116,117]. This use of machine learning components is not incidental: Combined with their visualization of motifs within a datasets, MEPP and LMPP's architectures derive from the goal of combining unbiased data mining methods with less opaque interfaces to help researchers understand the sequence features and patterns mined. Looking forward, we aim to leverage the transparency gained through MEPP and LMPP as a foundation to better understand and visualize the rules of regulatory region sequence grammar learned by more complex methods, such as generative adversarial networks for synthetic promoter generation [114].

Taken as a whole, we have crafted the methods presented in this work to empower researchers in making further discoveries, by making use of otherwise oversimplified or neglected axes of motif enrichment in their data, while maintaining the transparency required to drive interpretation of those motifs and their roles in transcriptional regulation. With MEIRLOP, we used biological scores directly, avoiding assumptions that compress these scores into thresholded sets. In the process, we have created a flexible MEA method used in multiple studies. With MEPP, we profiled the effect of position on motif enrichment and transcription, taking advantage of new transcription initiation profiling protocols that allow better use of TSS position. In the process, we no longer oversimplify MEA to ignore positional effects. And finally, with LMPP, we use both scored sequences and positional data to inform the discovery of de novo motifs, reducing reliance on complete motif datasets, and allowing researchers to incorporate information from otherwise disparate data.

Appendix

A.1 Supplemental Material for Chapter 1

Below we have reproduced supplemental figures and tables for Chapter 1, on MEIRLOP. The exception is supplemental Table 1.S3, which is 600 rows long and so cannot be directly included.

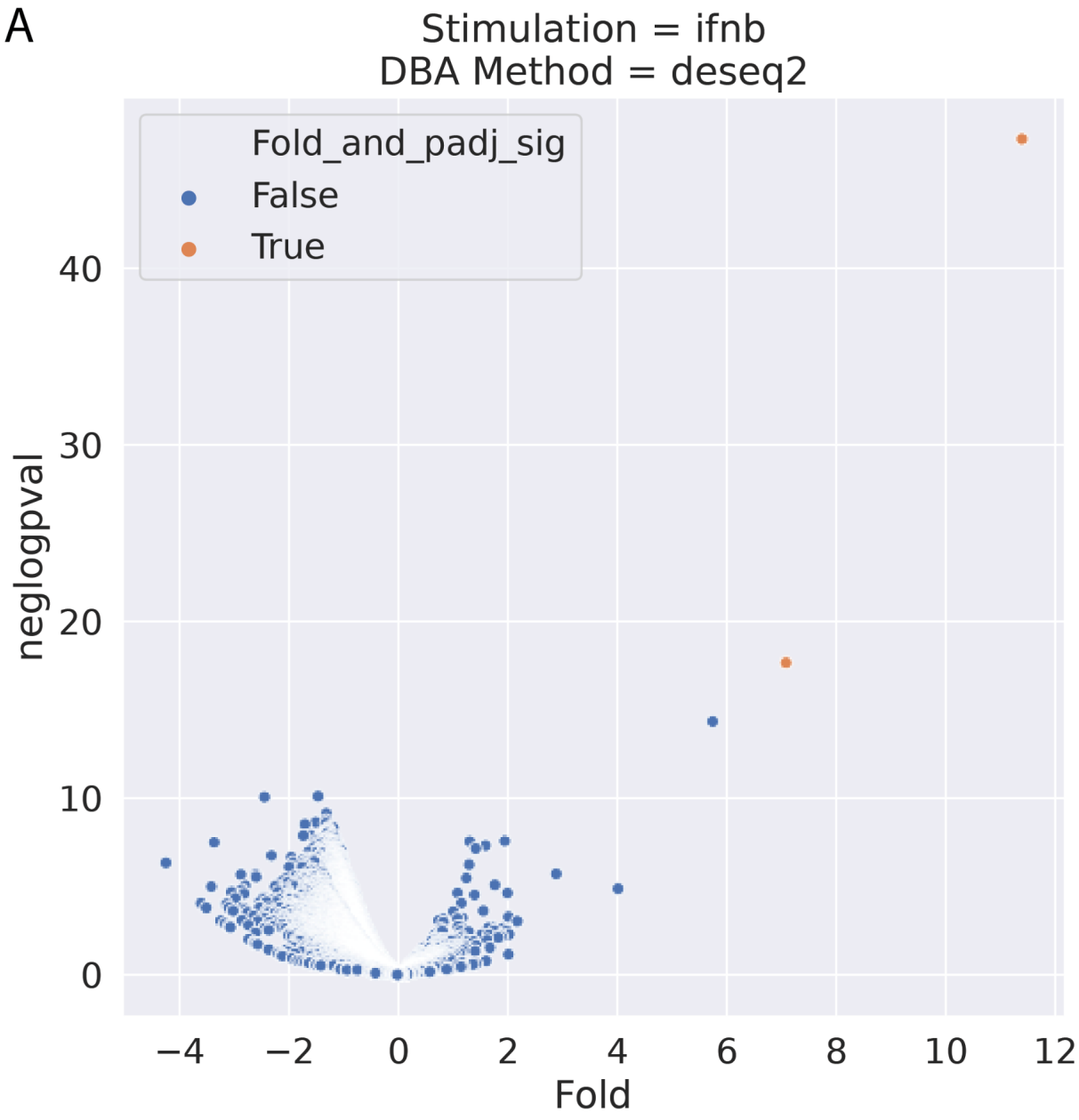


Figure 1.S1: Differential ChIP-seq of HCT116 cells before and after stimulation yields very few significantly differential peaks.

(A) Volcano plot depicting Log2 fold change (Fold) and significance (negative of log p-value, neglogpval) of 20,087 peaks found using MACS2 and DiffBind for HCT116 cells with (n = 2) and without (n = 2) IFN-β stimulation. Peaks matching significantly differential criteria (FDR < 0.05, log 2 fold-change > 1.0) are highlighted in orange (n = 2).

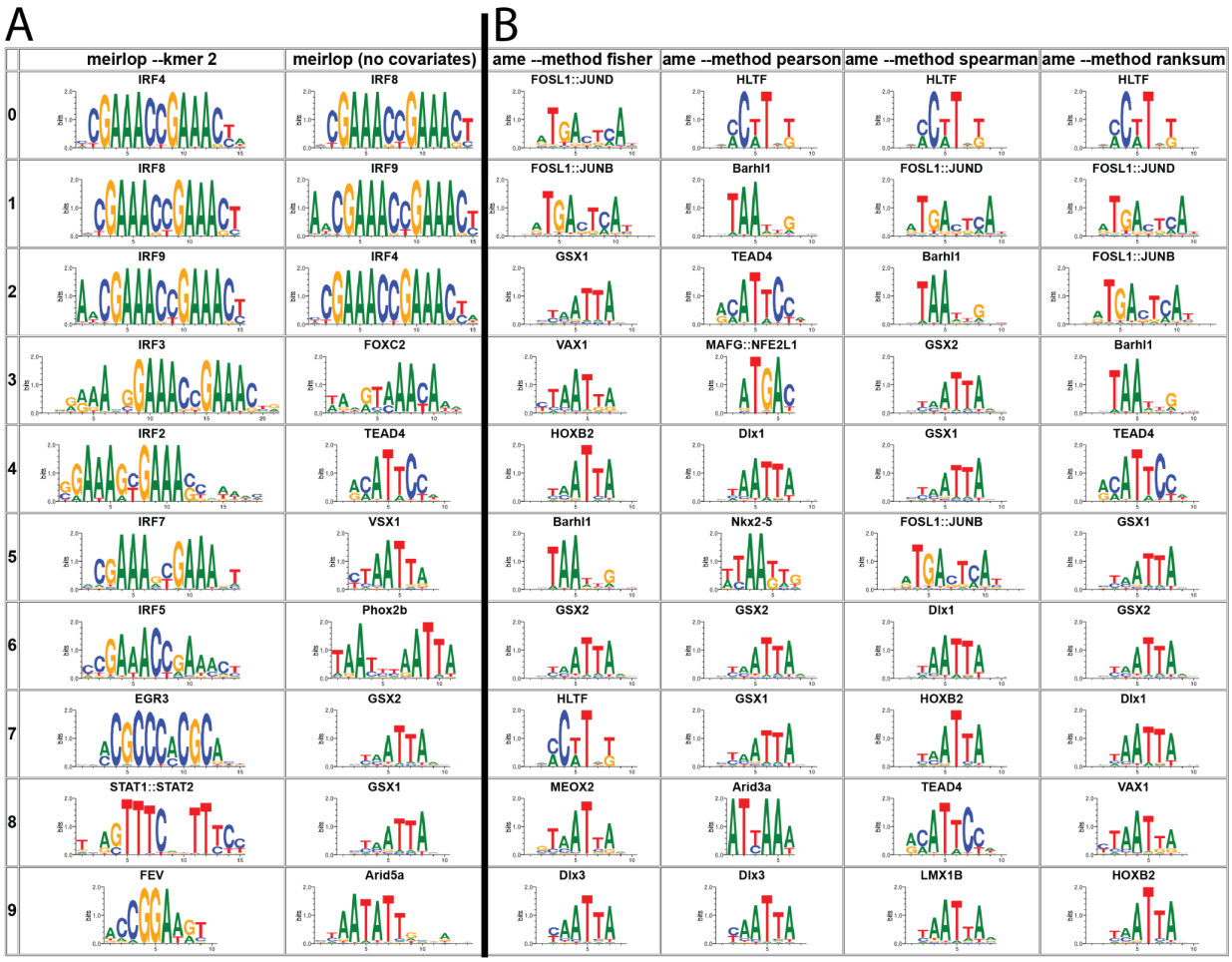


Figure 1.S2: Logistic regression with covariates finds enrichment of IRF9 and STAT1::STAT2 binding motifs ahead of AT-rich homeobox binding motifs.

(A) Top 10 significant enrichment results from our method, with and without covariates. Motifs are ordered as they appear in the HTML enrichment report output.

(B) Top 10 significant enrichment results from other score-based MEA methods. Motifs are ordered as they appear in the HTML enrichment report output.

Table 1.S3: Table of ENCODE ChIP-seq datasets used.

Contains detailed listings of ENCODE datasets used in results section “MEIRLOP achieves similar or better accuracy on TF ChIP-seq data”, including experiment accession IDs, download URLs, and the names of labs where the data were generated. Because this table is over 600 rows long, it cannot be directly included in this document in a feasible manner. In lieu of direct inclusion, a hyperlink (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7493370/bin/12859_2020_3739_MOESM3_ESM.csv) is included to the table in CSV format, as hosted by PubMed Central [93].

Table 1.S4: Table of ENCODE DNase-seq and histone ChIP-seq experiments used.

accession_id	source_url	dest_filename	lab_name
ENCSR000EOT	https://www.encodeproject.org/files/ENCFF156LGK/@@download/ENCFF156LGK.bam	dnase.bam	John Stamatoyannopoulos@UW
ENCSR000EOT	https://www.encodeproject.org/files/ENCFF821KDJ/@@download/ENCFF821KDJ.bed.gz	dnase.bed.gz	
ENCSR000AKU	https://www.encodeproject.org/files/ENCFF633WWH/@@download/ENCFF633WWH.bam	h3k4me3_rep1.bam	Bradley Bernstein@Broad
ENCSR000AKU	https://www.encodeproject.org/files/ENCFF777LZD/@@download/ENCFF777LZD.bam	h3k4me3_rep2.bam	
ENCSR000AKS	https://www.encodeproject.org/files/ENCFF778EZR/@@download/ENCFF778EZR.bam	h3k4me1_rep1.bam	
ENCSR000AKS	https://www.encodeproject.org/files/ENCFF580LGK/@@download/ENCFF580LGK.bam	h3k4me1_rep2.bam	
ENCSR000AKP	https://www.encodeproject.org/files/ENCFF301TVL/@@download/ENCFF301TVL.bam	h3k27ac_rep1.bam	
ENCSR000AKP	https://www.encodeproject.org/files/ENCFF879BWC/@@download/ENCFF879BWC.bam	h3k27ac_rep2.bam	
ENCSR000AKQ	https://www.encodeproject.org/files/ENCFF190OWE/@@download/ENCFF190OWE.bam	h3k27me3_rep1.bam	
ENCSR000AKQ	https://www.encodeproject.org/files/ENCFF692KQZ/@@download/ENCFF692KQZ.bam	h3k27me3_rep2.bam	

A.2 Supplemental Material for Chapter 2

A.2.1 MEPP Supplementary Methods

A.2.1.1 Filtering of degenerate and repetitive sequences

To filter out degenerate sequences, we remove sequences which possess more than a user-selectable threshold percentage of degenerate sequence content (not strictly A, C, G, or T). The resulting thresholded sequences are then one-hot encoded for use with convolutional models later in the analysis.

When used with soft-masked genomes that use lowercase bases to annotate repeats, MEPP considers these lowercase bases equivalent to the degenerate nucleotide code “N”. Users may bypass this behavior by processing the input file with awk’s `toupper` function.

A.2.1.2 Cluster deduplication to filter genomically overlapping sequence

If sequence from multiple overlapping genomic intervals is input into MEPP, this can lead to artificially inflated motif periodicities, as the same genomic subsequence that matches a motif can be repeated multiple times at shifted positions in the dataset. An example of this occurs when analyzing sequence sampled from +/- 200bp of TSS, when multiple TSS appear clustered less than 200bp from each other. To remedy this, we use `bedtools cluster` [118] to label clusters of overlapping genomic intervals, then only select the most informative interval from the cluster; this is typically the interval with the most coverage from the assay, e.g. the strongest of multiple neighboring TSS. We then extract sequences and scores for this “cluster deduplicated” set of intervals, and input these scored sequences into MEPP. Cluster deduplication may be undesirable when studying effects that rely on a feature’s overlapping nature, e.g. spacing of a motif with neighboring instances of itself. In these scenarios, users must avoid over-interpretation of artificial periodicities stemming from overlapping features.

A.2.1.3 Motif heatmap downsampling

Naive downsampling of the motif heatmap to compress the vertical axis would render motif instances invisible: Pixels containing positive matches for a motif would become averaged with surrounding pixels containing no matches for a motif. To counter this, MEPP uses a local maximum function to downsample the original dimensions of the motif heatmap (N sequences x M positions) to MEPP's display resolution.

A.2.1.4 Analysis of Drosophila melanogaster TSS

To identify TF binding motifs with position-dependent functions associated with transcription initiation, we performed csRNA-seq to map initiating transcripts in Drosophila melanogaster embryos. We started with cell culturing and csRNA-seq for data acquisition. This was followed by a csRNA-seq bioinformatics analysis comparing TSS with more vs. less nascent transcription. Finally, we ran motif enrichment positional profiling on these TSS using MEPP.

We conducted the data acquisition in two parts, described below:

1. Fly cell culture and treatment
2. Library prep and csRNA-seq on fly cells

We conducted the data analysis in three parts, also described below:

1. Core promoter element motif library compilation
2. TSS quantification from csRNA-seq
3. MEPP analysis of fly core promoter elements

A.2.1.4.1 Fly cell culture and treatment

Canton S wild-type Drosophila melanogaster flies were grown at 25˚C at 70-80% humidity in population cages. The embryos were collected on molasses-agar plates covered with yeast.

A.2.1.4.2 Library prep and csRNA-seq on fly cells

CsRNA-seq was performed on embryonic (0-12h) Drosophila melanogaster cells as previously described [10]. Small RNAs of ~20-60 nt were size selected form 2-5 µg of total RNA by denaturing gel electrophoresis. A 10% input sample was taken aside and the remainder enriched for 5'Caped RNAs with

3’-OH. Monophosphorylated RNAs were selectively degraded by Terminator 5´-Phosphate-Dependent Exonuclease (Lucigen). Subsequent 5’dephosporylation by CIP (NEB) followed by decapping with RppH (NEB) augments Cap-specific 5’adapter ligation by T4 RNA ligase 1 (NEB). The 3’ adapter was ligated using truncated T4 RNA ligase 2 (NEB) without prior 3’ repair to select against degraded RNA fragments.

A.2.1.4.3 Core promoter element motif library compilation

We transformed the Drosophila core promoter motifs from Ohler et al. from log-odds space to probability space [80], then exported the motifs in JASPAR motif format.

A.2.1.4.4 TSS quantification from csRNA-seq

First, 3’ adapters were cut using the the following command:

```
homerTools trim -3 AGATCGGAAGAGCACACGTCT -mis 2 -minMatchLength 4 -min 20
{reads.fastq.gz}
```

Where {reads.fastq.gz} stands in for the actual read file.

Then the trimmed reads were aligned to the dm6 reference genome using the `STAR` aligner, with splice junctions informed by the corresponding Ensemble Genes GTF file available from UCSC [119,120]. The read alignments were converted into tag directories using HOMER. A HOMER script, `getTSSfromReads.pl`, was then used to identify individual TSS from the assay from the 5’ ends of the aligned reads, and score them by their coverage (using “-min 3” to accept only TSS with a minimum count of 3 reads).

The resulting TSS file was converted to a scored bed file using awk, and coverage scores were transformed using the function log2(x+1), where x was the original score. Cluster deduplication was performed for clusters of TSS within 200 bp of each other and on the same strand, selecting the most highly covered TSS within a cluster. Sequence +/- 200 bp of each TSS was extracted from the dm6 reference genome, and scored by csRNA-seq coverage. We proceeded with these scored sequences for downstream analysis with MEPP.

A.2.1.4.5 MEPP analysis of fly core promoter elements

We ran MEPP on the resulting dataset using 200 permutations for profile permutation testing, and filtered out input sequences that were over 50% degenerate. The effective command line invocation was:

```
mepp --fa {scored_sequences.fa} --motifs {motifs.txt} --out
{output_filepath} --orientations +,- --perms 200 --batch 200 --dgt 50 --jobs 20
--gjobs 10 --nogpu --dpi 100 &> {logfile}
```

A.2.1.5 MEPP analysis of GATA1 ChIP-seq

We started by retrieving alignment files for a GATA1 ChIP-seq experiment from K562 cells (ENCODE experiment ENCSR000EFT), as well as corresponding input experiments from ENCODE (ENCODE experiment ENCSR000EHM) [43,44].

In order to obtain differential scoring information, we first derived coverage bigwig files from the alignment files using deeptools bamCoverage defaults [48]. We then used wiggletools to compute two wig files [121]: one containing the sum of coverage across all replicates, and one containing the log2 fold change of the mean coverage (plus one) between GATA1 ChIP-seq and input control. We calculated the log2 fold change such that higher scores indicated greater sequencing coverage in the GATA1 ChIP-seq experiments. Conversions between wig and bigWig formats were performed using bigWigToWig and wigToBigWig [122].

We then used `macs2 callpeak` to call peak summits on each bam file [104]. We concatenated all unique peak summits, finally using bedtools slop to expand +/-100bp from the summits. Using the bedops utilities wig2bed and bedmap [123], we annotated these summits with scores from the previously generated log2 fold change bigWig files. We ran the bedmap command in `--wmean` mode, so that each interval (surrounding a summit) acquired a score equivalent to the mean of overlapping bigWig intervals, weighted by the percentage of overlap. A similar operation assigned a sum of coverage score to each summit from the previously generated bigWig files.

We then performed cluster deduplication of the motif-centered intervals, keeping only the highest-covered interval among clusters of overlapping intervals, and additionally ensuring a minimum total coverage of 5 reads at each interval. Finally, we extracted scored sequences (sampled from the

hg38 reference genome) from the deduplicated intervals, and submitted these to MEPP. The effective command line invocation was:

```
mepp --fa {scored_sequences.fa} --motifs {motifs.txt} --out {output_filepath} --orientations +/- --margin 5 --perms 100 --batch 200 --dgt 50 --jobs 15 --gjobs 15 --nogpu --dpi 100 &> {logfile}
```

A.2.1.6 Analysis of differential chromatin accessibility between cell types

We started by retrieving alignment files for ATAC-seq experiments on K562 and HCT116 cells from ENCODE [43,44]. Specifically, we used replicate alignment files from experiment ID ENCSR483RKN on K562 cells, and experiment ID ENCSR872WG on HCT116 cells [43,44].

In order to obtain differential scoring information, we first derived coverage bigwig files from the alignment files, then calculated a log2 fold change bigWig file comparing coverage between HCT116 and K562 cells (using the same methods described above for GATA1 ChIP-seq). We calculated the log2 fold change such that higher scores indicated greater chromatin accessibility in HCT116 cells than in K562 cells. Similarly, we calculated the summed coverage across all samples as a bigWig file.

We then scanned the hg38 human genome for instances of the GATA1 binding motif (as stored in HOMER's motif library under `gata.motif`) using the HOMER `scanMotifGenomeWide.pl` command [24]. The effective command line was:

```
scanMotifGenomeWide.pl homer_motifs/gata.motif hg38.fa -bed -5p > gata.scans.bed
```

This yielded a BED file of genomic intervals +/-100bp centered on the 5' end of GATA1 binding motifs, and standed according to motif orientation. We annotated these genomic intervals with scores from the log2 fold change bigWig files, as described for the GATA1 ChIP-seq analysis above.. We also repeated this process to annotate the genomic intervals with their overall summed coverage (to support cluster deduplication).

Then as in previous analyses, we performed cluster deduplication of the motif-centered intervals, keeping only the highest-covered interval among clusters of overlapping intervals and additionally ensuring a minimum total coverage of 5 reads at each interval. Finally, we extracted scored sequences

(sampled from the hg38 reference genome) from the deduplicated intervals, and submitted these to MEPP. The effective command line invocation was:

```
mepp --fa {scored_sequences.fa} --motifs {motifs.txt} --out {output_filepath} --orientations +,- --perms 100 --batch 200 --dgt 50 --jobs 20 --gjobs 10 --nogpu --dpi 100 &> {logfile}
```

A.2.1.7 Analysis of Nanog motif binding in Mus musculus

First, we scanned the mm10 mouse reference genome for matches to the Nanog binding motif. Using the `scanMotifGenomeWide.pl` script we obtained a motif scan bed file of intervals +/- 100bp of the motif 5' end. We reproduce the effective command line below:

```
scanMotifGenomeWide.pl nanog.motif mm10.masked.fa -bed -5p > mm10.nanog_motif_scans.bed
```

As a source of Nanog binding signal, we used the normalized Nanog ChIP-seq read signal from GSM4291126, available through GEO Series GSE144577 [85]. The signal came from a GEO download in the bigWig format, under the file name "GSM4291126_WT_Nanog.bam.bw". To map this Nanog binding signal data to our motif scans, we first converted the bigWig file to bed file with a combination of `bigWigtoWig` and `wig2bed` [122,123]. We then used the bedmap to assign coverage scores from the converted bigWig file to the motif scan bed file, using the 'wmean' operation. After filtering out motif scans with zero coverage and performing cluster deduplication, we recovered the sequences (sampled from the mm10 reference genome) within +/- 200bp of the motif 5' ends. To determine enrichment of motifs in HOMER's vertebrate TF motif collection, we submitted these scored sequences to MEPP for downstream analysis.

A.2.1.8 MEPP Analysis of Mouse ChIP-nexus and ChIP-seq

We started by retrieving MACS2-generated narrowpeak files from GEO accession GSE137193, corresponding to peaks called from ChIP-nexus and ChIP-seq binding assays of transcription factors Nanog, Oct4, and Sox2 on mouse embryonic stem cells [86,104]. For each interval in each narrowpeak file, we relocated both start and end positions to the peak summit (using the peak value in column 10), set

the score to the signal value (column 7), and the strand to positive, outputting the resulting scored peak summits into a bed file. We then extracted scored sequences +/-200bp of each scored peak summit, and analyzed these scored sequences using MEPP. To account for the less positionally specific nature of ChIP-seq, we used a wider motif margin of 5. To account for the lack of strand specificity in the MACS2 ChIP-seq peak calls, we correlate against the higher motif match score of either motif orientation.

We further demonstrate results from an alternative analysis method for ChIP-seq. Following analysis steps as detailed in GEO accession GSM4072778, we started from the reads for the Nanog ChIP-nexus and patchcap experiments, beginning with barcode trimming via nimnexus trim, including alignment, until aligned read deduplication using nimnexus dedup. We converted the resulting alignments into tag directories. The HOMER getTSSFromReads.pl script was used (in a similar manner to its use in csRNA-seq) To enumerate stranded 1bp binding sites from the 5' ends of each tag in these tag directories we used the HOMER getTSSFromReads.pl script, ensuring at least 4 reads were counted per site (-minRaw 4), while controlling for 5' ends from the patchcap experiment. We then used the HOMER annotatePeaks.pl command to count 5' ends of strand-matched tags from both tag directories at each enumerated site, as well as to normalize these counts using DESeq2's rlog method. The resulting 1bp binding sites were scored by the rlog-normalized coverage of tag 5' ends from the Nanog ChIP-nexus experiment, expanded by 200bp in each direction then cluster deduplicated (favoring sites with higher coverage). Finally, we converted the scored and expanded binding sites into scored sequences (sampled from the mm10 reference genome), which we analyzed using MEPP, carrying over parameters from the previous ChIP-nexus analysis.

A.2.1.9 Differential csRNA-seq analysis

csRNA-seq data from mouse bone marrow derived macrophage cells treated with KLA was downloaded from GSE135498 [10]. Subsequently, we ran TSS identification and differential TSS analysis between the KLA-stimulated and control samples.

The data analysis was conducted in three parts, described below:

1. TSS quantification from csRNA-seq on mouse BMDM cells
2. Differential TSS analysis and scoring

3. MEPP analysis of differential TSS

A.2.1.9.1 TSS quantification from csRNA-seq on mouse BMDM cells

TSS quantification proceeded as described above for fly cells, with the exception that alignment was to mm10 references, and with a minimum read count of 5 (-minRaw 5). Using HOMER's `annotatePeaks.pl` command, we then quantified 5' end read counts. We reproduce the effective command line below:

```
annotatePeaks.pl tss.txt genomes/mm10/mm10.fa -strand + -fragLength 1
-raw -d mm10_csrna_raw_kla_0min_replicate_1/
mm10_csrna_raw_kla_0min_replicate_2/ mm10_csrna_raw_kla_50min_replicate_1/
mm10_csrna_raw_kla_50min_replicate_2/ > tss.counts.txt
```

A.2.1.9.2 Differential TSS analysis and scoring

We performed differential csRNA-seq analysis on the previously generated TSS read count quantifications using HOMER's `getDiffExpression.pl` script, (a wrapper around DESeq2) [88]. The comparison found log2 fold change values comparing TSS transcriptional signal in KLA-stimulated vs. control, after controlling for replicate batch effects.

A.2.1.9.3 MEPP analysis of differential TSS

We scored TSS by their log2 fold change values, then performed cluster deduplication, selecting the TSS from each cluster with the highest read coverage. We then extracted sequence +/- 200bp around the TSS, and input the resulting scored sequences into MEPP, to analyze enrichment of motifs in HOMER's vertebrate TF motif collection.

A.2.1.10 Differential cleavage site analysis

We started by retrieving ATAC-seq and H3K27ac MNase-seq data for BMDM cells treated to 1 hour of LPS stimulation vs. control, from the work of Comoglio et al., available under GEO Series GSE119693 [89]. We then trimmed adapters from reads using Trim Galore, then aligned the reads to the

mm10 reference genome using `bowtie2` [40,124]. After converting the read alignments to HOMER tag directories, the rest of the analysis proceeded as described above for differential csRNA-seq analysis.

Figure 2.S1: MEPP uses pre-weighted convolutional kernels to derive motif heatmaps and positional correlation profiles

(A) Typical differential sequencing experiment with replicates, yielding peaks scored by a function of coverage (e.g. normalized or log 2 fold change coverage).
(B) Scored sequences are extracted from scored peaks, centered on some feature of interest, e.g. TSS. Motif locations marked in orange.
(C) Local motif enrichments yield positional correlations of the motif with sequence score.
(D) Input to MEPP consists of a set of scored, one-hot encoded DNA sequences
(E) A known motif is input to form the weights of a convolutional kernel, which evaluates the log odds matching score of a one-hot encoded sequence to the motif. The detection threshold for the motif inverted to form the kernel's bias. The resulting activation value is fed through a ReLU function. The motif's GC% is also evaluated.
(F) The entire set of scored sequences is convolved, forming rows of a motif heatmap. Rows are sorted in descending order from the score of the corresponding input sequence.
(G) Local correlations of motif match scores and sequence scores are calculated for each motif position, controlling for GC% skew using a partial Pearson correlation.

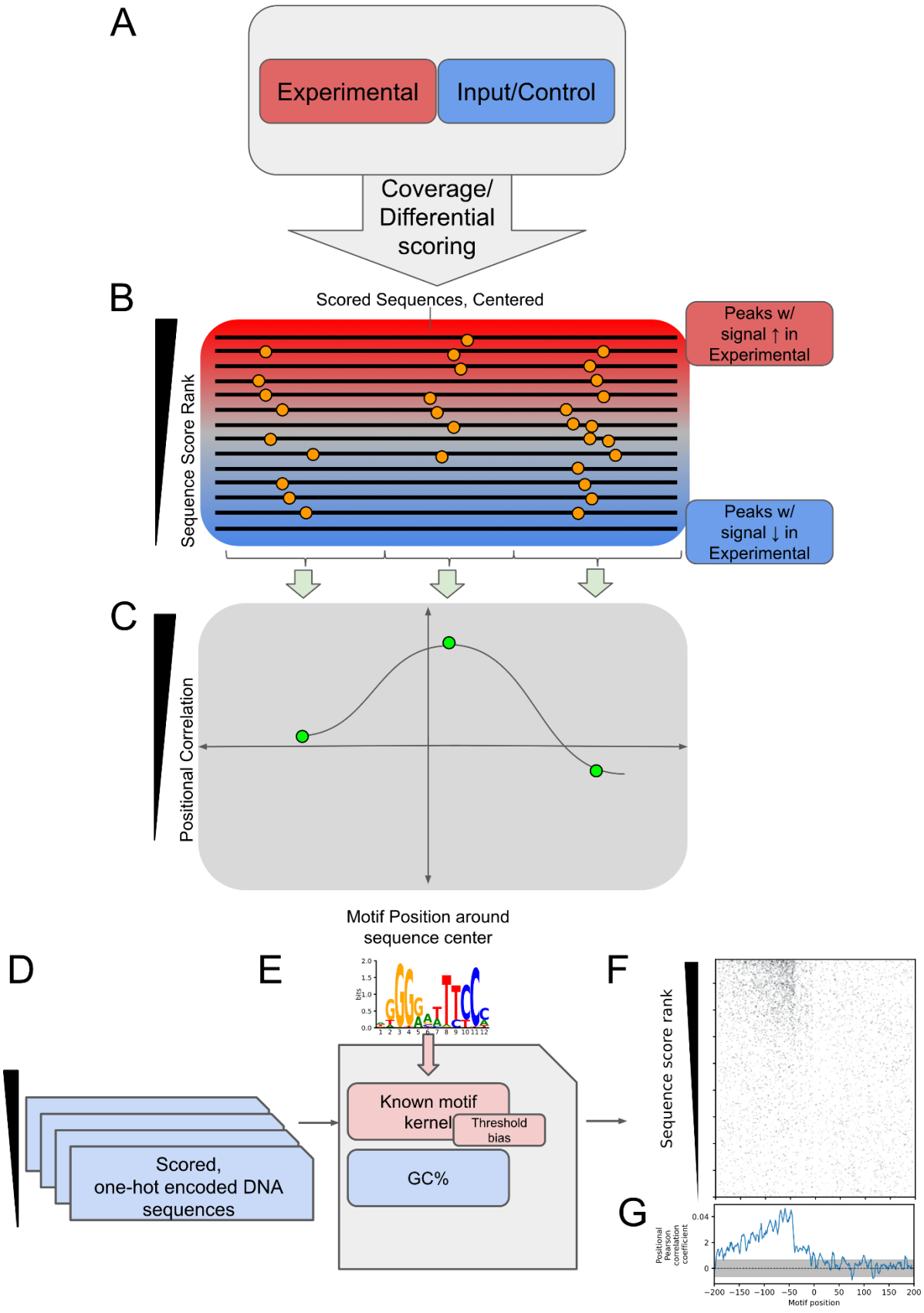
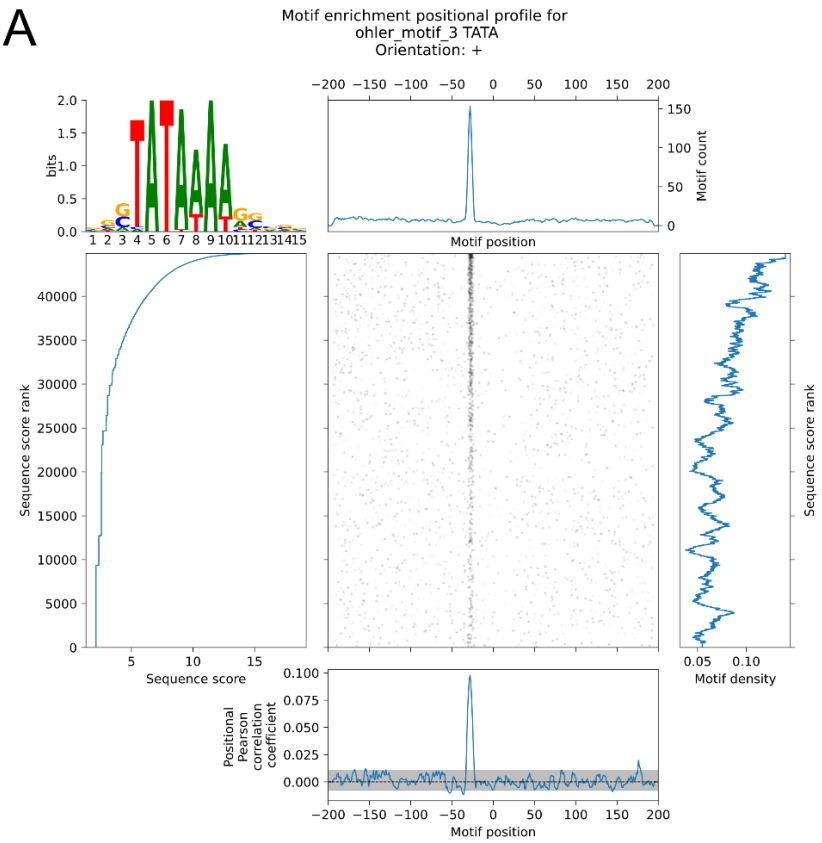

A
Experimental
Input/Control
Coverage/
Differential
scoring
B
Scored Sequences, Centered
Sequence Score Rank
Peaks w/
signal ↑ in
Experimental
Peaks w/
signal ↓ in
Experimental
C
Positional Correlation
Motif Position around
sequence center
D
Scored,
one-hot encoded DNA
sequences
E
Known motif
kernel
Threshold
bias
GC%
F
Sequence score rank
G
Positional
Pearson
correlation
coefficient
Motif position

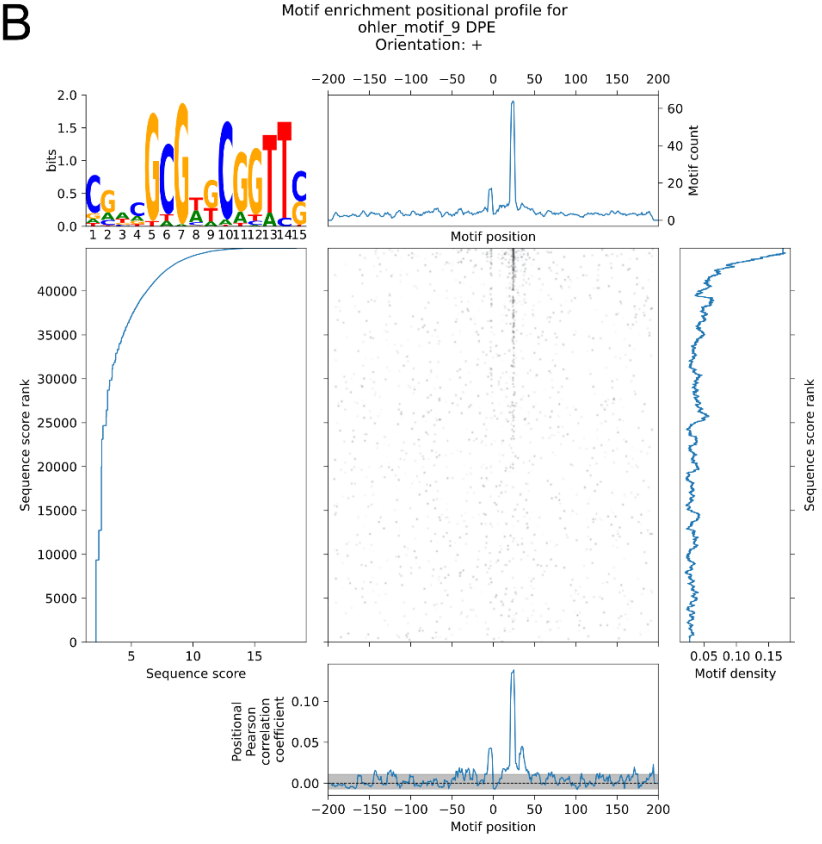
Figure 2.S2: MEPP visualizes and quantifies the TATA-box and DPR motifs near Drosophila melanogaster transcription start sites.

(A) MEPP plot for TATA-box motif (from the HOMER motif collection), on sequences +/- 200bp of *D. melanogaster* TSS quantified by csRNA-seq, scored by log-transformed csRNA-seq coverage.
(B) Same as (A) but for the DPR motif (from the HOMER motif collection)

A


Motif enrichment positional profile for
ohler_motif_3 TATA
Orientation: +
bits
Motif count
Motif position
Sequence score rank
Sequence score
Motif density
Positional Pearson correlation coefficient

B


Motif enrichment positional profile for
ohler_motif_9 DPE
Orientation: +
bits
Motif count
Motif position
Sequence score rank
Sequence score
Motif density
Positional Pearson correlation coefficient

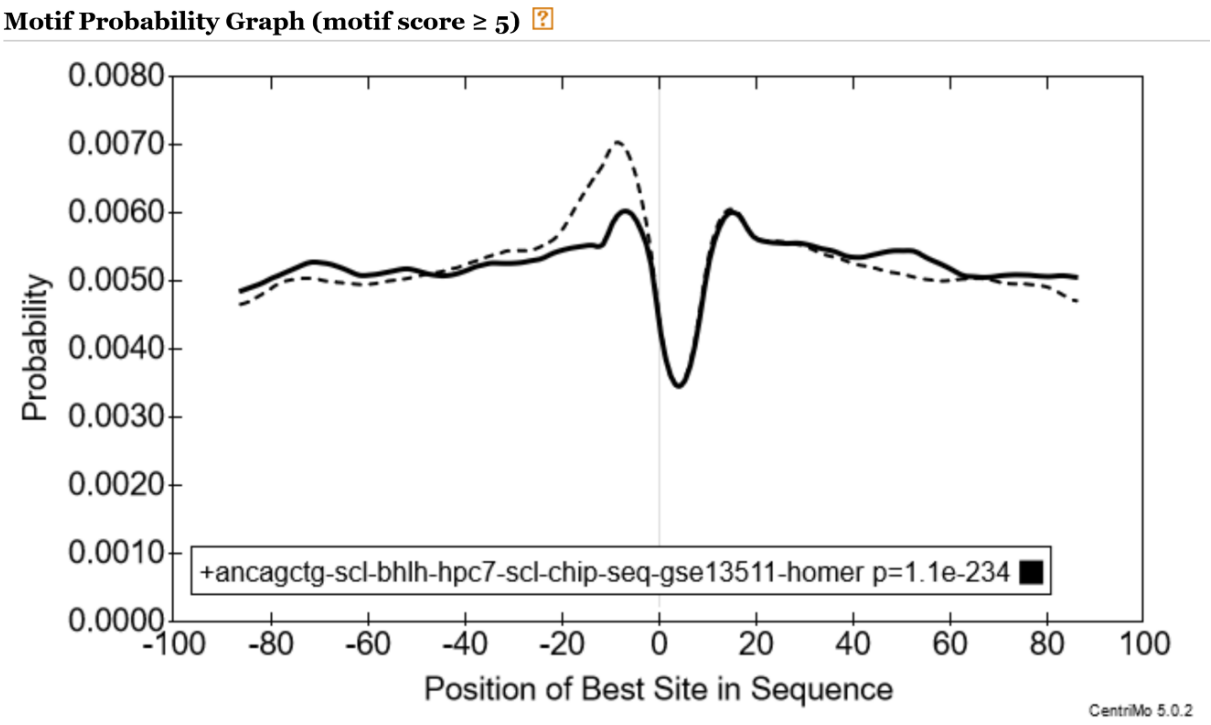


Figure 2.S3: CentriMo visualizes motif prevalence over sequence positions

Local enrichment plot for SCL binding motif locations surrounding GATA1 binding motifs as generated by CentriMo, using default visualization smoothing parameters. Enrichment was performed contrasting the top 10% of scored sequences with the bottom 10% of scored sequences.

Table 2.S4: Public Data accessions and attributions used in this study

Dataset accession (s)	Sample/File accessions	Sample Title	Producing Lab
ENCSR000EFT	ENCFF844WTT.bam	GATA1 ChIP-seq rep 1	Michael Snyder, Stanford
ENCSR000EFT	ENCFF729MTL.bam	GATA1 ChIP-seq rep 2	
ENCSR000EHM	ENCFF769RAH.bam	Control ChIP-seq rep 1	
ENCSR000EHM	ENCFF147YPF.bam	Control ChIP-seq rep 2	
ENCSR483RKN	ENCFF512VEZ.bam	K562 ATAC-seq rep 1	
ENCSR483RKN	ENCFF987XOV.bam	K562 ATAC-seq rep 2	
ENCSR872WG	ENCFF724QHH.bam	HCT116 ATAC-seq rep 1	
ENCSR872WG	ENCFF927YUB.bam	HCT116 ATAC-seq rep 2	
GSE144577	GSM4291126_WT_Nanog.bam.bw	WT_Nanog_ChIP	Xiang Sun, Sun Yat-sen University
GSE137193	GSM4087827_mesc_Nanog_chipseq.idr-optimal-set.narrowPeak.gz	Nanog-ChIP-seq	Julia Zeitlinger, Stowers Institute for Medical Research
GSE137193	GSM4072778_mesc_nanog_nexus.idr-optimal-set.narrowPeak.gz	Nanog-ChIP-nexus	
GSE137193	SRX6827272	Nanog-ChIP-nexus	
GSE135498	GSM4012734	BMDM notx csRNA-seq r1	Christopher Benner, UCSD
GSE135498	GSM4012735	BMDM notx csRNA-seq r2	
GSE135498	GSM4012736	BMDM notx csRNAinput r1	
GSE135498	GSM4012737	BMDM notx csRNAinput r2	
GSE135498	GSM4012738	BMDM KLA1h csRNA-seq r1	
GSE135498	GSM4012739	BMDM KLA1h csRNA-seq r2	
GSE135498	GSM4012740	BMDM KLA1h csRNAinput r1	
GSE135498	GSM4012741	BMDM KLA1h csRNAinput r2	
GSE119693	GSM3380863	ATAC_LPS-1h_rep1	Gioacchino Natoli, Humanitas University (Hunimed)
GSE119693	GSM3380864	ATAC_LPS-1h_rep2	
GSE119693	GSM3380865	ATAC_LPS-1h_rep3	
GSE119693	GSM3380875	ATAC_UT_rep1	
GSE119693	GSM3380876	ATAC_UT_rep2	
GSE119693	GSM3380877	ATAC_UT_rep3	
GSE119693	GSM3380878	H3K27ac_LPS-1h_rep1	
GSE119693	GSM3380879	H3K27ac_LPS-1h_rep2	
GSE119693	GSM3380886	H3K27ac_UT_rep1	
GSE119693	GSM3380887	H3K27ac_UT_rep2	

References

1. Zhong B, Tien P, Shu HB. Innate immune responses: crosstalk of signaling and regulation of gene transcription. Virology [Internet]. 2006 Aug 15;352(1):14–21. Available from: http://dx.doi.org/10.1016/j.virol.2006.04.029

2. Frietze S, Farnham PJ. Transcription factor effector domains. Subcell Biochem [Internet]. 2011;52:261–77. Available from: http://dx.doi.org/10.1007/978-90-481-9069-0_12

3. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. The Human Transcription Factors. Cell [Internet]. 2018 Feb 8;172(4):650–65. Available from: http://dx.doi.org/10.1016/j.cell.2018.01.029

4. McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics [Internet]. 2010 Apr 1;11:165. Available from: http://dx.doi.org/10.1186/1471-2105-11-165

5. Bailey TL, Machanick P. Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res [Internet]. 2012 Sep 1;40(17):e128. Available from: http://dx.doi.org/10.1093/nar/gks433

6. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z. Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res [Internet]. 2004 Feb 26;32(4):1372–81. Available from: http://dx.doi.org/10.1093/nar/gkh299

7. Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol [Internet]. 2007 Mar 23;3(3):e39. Available from: http://dx.doi.org/10.1371/journal.pcbi.0030039

8. de Hoon M, Hayashizaki Y. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. Biotechniques [Internet]. 2008 Apr;44(5):627–8, 630, 632. Available from: http://dx.doi.org/10.2144/000112802

9. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, Waters CT, Munson K, Core LJ, Lis JT. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc [Internet]. 2016 Aug;11(8):1455–76. Available from: http://dx.doi.org/10.1038/nprot.2016.086

10. Duttke SH, Chang MW, Heinz S, Benner C. Identification and dynamic quantification of regulatory elements using total RNA. Genome Res [Internet]. 2019 Nov;29(11):1836–46. Available from: http://dx.doi.org/10.1101/gr.253492.119

11. Oldfield AJ, Henriques T, Kumar D, Burkholder AB, Cinghu S, Paulet D, Bennett BD, Yang P, Scruggs BS, Lavender CA, Rivals E, Adelman K, Jothi R. NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. Nat Commun [Internet]. 2019 Jul 11 [cited 2021 Dec 22];10(1). Available from: https://pubmed.ncbi.nlm.nih.gov/31296853/

12. Acevedo-Luna N, Mariño-Ramírez L, Halbert A, Hansen U, Landsman D, Spouge JL. Most of the tight positional conservation of transcription factor binding sites near the transcription start site reflects their co-localization within regulatory modules. BMC Bioinformatics [Internet]. 2016 Nov 21 [cited 2021 Dec 22];17(1). Available from: https://pubmed.ncbi.nlm.nih.gov/27871221/

13. Tabach Y, Brosh R, Buganim Y, Reiner A, Zuk O, Yitzhaky A, Koudritsky M, Rotter V, Domany E. Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. PLoS One [Internet]. 2007 Aug 29 [cited 2021 Dec 22];2(8). Available from: https://pubmed.ncbi.nlm.nih.gov/17726537/

14. Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc Natl Acad Sci U S A [Internet]. 2016 Jun 7 [cited 2021 Dec 22];113(23). Available from: https://pubmed.ncbi.nlm.nih.gov/27155014/

15. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Chèneby J, Kulkarni SR, Tan G, Baranasic D, Arenillas DJ, Sandelin A, Vandepoele K, Lenhard B, Ballester B, Wasserman WW, Parcy F, Mathelier A. JASPAR 2018: update of the open-access

database of transcription factor binding profiles and its web framework. Nucleic Acids Res [Internet]. 2018 Jan 4;46(D1):D260–6. Available from: http://dx.doi.org/10.1093/nar/gkx1126

16. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A [Internet]. 2010 Dec 14;107(50):21931–6. Available from: http://dx.doi.org/10.1073/pnas.1016071107

17. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol [Internet]. 2015 Mar;16(3):144–54. Available from: http://dx.doi.org/10.1038/nrm3949

18. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. Brief Bioinform [Internet]. 2016 Nov;17(6):953–66. Available from: http://dx.doi.org/10.1093/bib/bbv110

19. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform [Internet]. 2015 Jan;16(1):59–70. Available from: http://dx.doi.org/10.1093/bib/bbt086

20. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A [Internet]. 2005 Oct 25;102(43):15545–50. Available from: http://dx.doi.org/10.1073/pnas.0506580102

21. Leibovich L, Paz I, Yakhini Z, Mandel-Gutfreund Y. DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. Nucleic Acids Res [Internet]. 2013 Jul;41(Web Server issue):W174–9. Available from: http://dx.doi.org/10.1093/nar/gkt407

22. Roider HG, Manke T, O'Keeffe S, Vingron M, Haas SA. PASTAA: identifying transcription factors associated with sets of co-regulated genes. Bioinformatics [Internet]. 2009 Feb 15;25(4):435–42. Available from: http://dx.doi.org/10.1093/bioinformatics/btn627

23. Worsley Hunt R, Mathelier A, Del Peso L, Wasserman WW. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. BMC Genomics [Internet]. 2014 Jun 13;15:472. Available from: http://dx.doi.org/10.1186/1471-2164-15-472

24. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell [Internet]. 2010 May 28;38(4):576–89. Available from: http://dx.doi.org/10.1016/j.molcel.2010.05.004

25. Zambelli F, Pesole G, Pavesi G. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic Acids Res [Internet]. 2009 Jul;37(Web Server issue):W247–52. Available from: http://dx.doi.org/10.1093/nar/gkp464

26. Zambelli F, Pesole G, Pavesi G. PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. Nucleic Acids Res [Internet]. 2013 Jul;41(Web Server issue):W535–43. Available from: http://dx.doi.org/10.1093/nar/gkt448

27. Mariani L, Weinand K, Vedenko A, Barrera LA, Bulyk ML. Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds. Cell Syst [Internet]. 2017 Sep 27;5(3):187–201.e7. Available from: http://dx.doi.org/10.1016/j.cels.2017.06.015

28. Delos Santos N. MEIRLOP: Motif Enrichment In Ranked Lists Of Peaks [Internet]. Github. 2018 [cited 2020 Mar 24]. Available from: https://github.com/npdeloss/meirlop

29. Aguilera AM, Escabias M, Valderrama MJ. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comput Stat Data Anal [Internet]. 2006 Apr;50(8):1905–24. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167947305000630

30. Keles S, van der Laan MJ, Vulpe C. Regulatory motif finding by logic regression. Bioinformatics [Internet]. 2004 Nov 1;20(16):2799–811. Available from:

http://dx.doi.org/10.1093/bioinformatics/bth333

31. Yao Z, Macquarrie KL, Fong AP, Tapscott SJ, Ruzzo WL, Gentleman RC. Discriminative motif analysis of high-throughput dataset. Bioinformatics [Internet]. 2014 Mar 15;30(6):775–83. Available from: http://dx.doi.org/10.1093/bioinformatics/btt615

32. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics [Internet]. 2009 Dec 1;25(23):3181–2. Available from: http://dx.doi.org/10.1093/bioinformatics/btp554

33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perr ot M, Duchesnay É. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research [Internet]. 2011 Oct;12:2825–30. Available from: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

34. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. In: Proceedings of the 9th Python in Science Conference. 2010.

35. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing [Internet]. Vol. 57, Journal of the Royal Statistical Society: Series B (Methodological). 1995. p. 289–300. Available from: http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x

36. Tareen A, Kinney JB. Logomaker: beautiful sequence logos in Python. Bioinformatics [Internet]. 2020 Apr 1;36(7):2272–4. Available from: http://dx.doi.org/10.1093/bioinformatics/btz921

37. Natsume T, Kiyomitsu T, Saga Y, Kanemaki MT. Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. Cell Rep [Internet]. 2016 Apr 5;15(1):210–8. Available from: http://dx.doi.org/10.1016/j.celrep.2016.03.001

38. Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, Rialdi A, White KM, Albrecht RA, Pache L, Marazzi I, García-Sastre A, Shaw ML, Benner C. Transcription Elongation Can Affect Genome 3D Structure. Cell [Internet]. 2018 Sep 6;174(6):1522–36.e22. Available from:

http://dx.doi.org/10.1016/j.cell.2018.07.047

39. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics [Internet]. 2018 Sep 1;34(17):i884–90. Available from: http://dx.doi.org/10.1093/bioinformatics/bty560

40. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods [Internet]. 2012 Mar 4;9(4):357–9. Available from: http://dx.doi.org/10.1038/nmeth.1923

41. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol [Internet]. 2008 Sep 17;9(9):R137. Available from: http://dx.doi.org/10.1186/gb-2008-9-9-r137

42. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR, Ali S, Chin SF, Palmieri C, Caldas C, Carroll JS. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature [Internet]. 2012 Jan 4;481(7381):389–93. Available from: http://dx.doi.org/10.1038/nature10730

43. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature [Internet]. 2012 Sep 6;489(7414):57–74. Available from: http://dx.doi.org/10.1038/nature11247

44. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Strattan JS, Jolanki O, Tanaka FY, Cherry JM. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res [Internet]. 2018 Jan 4;46(D1):D794–801. Available from: http://dx.doi.org/10.1093/nar/gkx1081

45. Tange O. GNU Parallel 2018. 2018 Apr 27 [cited 2020 Mar 23]; Available from: https://zenodo.org/record/1146014

46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features [Internet]. Vol. 26, Bioinformatics. 2010. p. 841–2. Available from: http://dx.doi.org/10.1093/bioinformatics/btq033

47. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics [Internet]. 2011 Dec 15;27(24):3423–4. Available from: http://dx.doi.org/10.1093/bioinformatics/btr539

48. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res [Internet]. 2014 Jul;42(Web Server issue):W187–91. Available from: http://dx.doi.org/10.1093/nar/gku365

49. Paul A, Tang TH, Ng SK. Interferon Regulatory Factor 9 Structure and Regulation. Front Immunol [Internet]. 2018 Aug 10;9:1831. Available from: http://dx.doi.org/10.3389/fimmu.2018.01831

50. McComb S, Cessford E, Alturki NA, Joseph J, Shutinoski B, Startek JB, Gamero AM, Mossman KL, Sad S. Type-I interferon signaling through ISGF3 complex is required for sustained Rip3 activation and necroptosis in macrophages. Proc Natl Acad Sci U S A [Internet]. 2014 Aug 5;111(31):E3206–13. Available from: http://dx.doi.org/10.1073/pnas.1407068111

51. Fujioka S, Niu J, Schmidt C, Sclabas GM, Peng B, Uwagawa T, Li Z, Evans DB, Abbruzzese JL, Chiao PJ. NF-kappaB and AP-1 connection: mechanism of NF-kappaB-dependent regulation of AP-1 activity. Mol Cell Biol [Internet]. 2004 Sep;24(17):7806–19. Available from: http://dx.doi.org/10.1128/MCB.24.17.7806-7819.2004

52. Ishii J, Kitazawa R, Mori K, McHugh KP, Morii E, Kondo T, Kitazawa S. Lipopolysaccharide suppresses RANK gene expression in macrophages by down-regulating PU.1 and MITF. J Cell Biochem [Internet]. 2008 Oct 15;105(3):896–904. Available from: http://dx.doi.org/10.1002/jcb.21886

53. Baillie JK, Arner E, Daub C, De Hoon M, Itoh M, Kawaji H, Lassmann T, Carninci P, Forrest ARR, Hayashizaki Y, FANTOM Consortium, Faulkner GJ, Wells CA, Rehli M, Pavli P, Summers KM, Hume DA. Analysis of the human monocyte-derived macrophage transcriptome and response to lipopolysaccharide provides new insights into genetic aetiology of inflammatory bowel disease. PLoS Genet [Internet]. 2017 Mar;13(3):e1006641. Available from: http://dx.doi.org/10.1371/journal.pgen.1006641

54. Sharifi-Zarchi A, Gerovska D, Adachi K, Totonchi M, Pezeshk H, Taft RJ, Schöler HR, Chitsaz H, Sadeghi M, Baharvand H, Araúzo-Bravo MJ. DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. BMC Genomics [Internet]. 2017 Dec 12;18(1):964. Available from: http://dx.doi.org/10.1186/s12864-017-4353-7

55. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Müller F, Forrest ARR, Carninci P, Rehli M, Sandelin A. An atlas of active enhancers across human cell types and tissues. Nature [Internet]. 2014 Mar 27;507(7493):455–61. Available from: http://dx.doi.org/10.1038/nature12787

56. Myslinski E, Gérard MA, Krol A, Carbon P. A genome scale location analysis of human Staf/ZNF143-binding sites suggests a widespread role for human Staf/ZNF143 in mammalian promoters. J Biol Chem [Internet]. 2006 Dec 29;281(52):39953–62. Available from: http://dx.doi.org/10.1074/jbc.M608507200

57. Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B, Scacheri PC, Haibe-Kains B, Lupien M. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. Nat Commun [Internet]. 2015 Feb 3;2:6186. Available from: http://dx.doi.org/10.1038/ncomms7186

58. Rye M, Sætrom P, Håndstad T, Drabløs F. Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements. BMC Biol [Internet]. 2011 Nov 24;9:80. Available from: http://dx.doi.org/10.1186/1741-7007-9-80

59. Cheng Y, Wu W, Kumar SA, Yu D, Deng W, Tripic T, King DC, Chen KB, Zhang Y, Drautz D, Giardine B, Schuster SC, Miller W, Chiaromonte F, Zhang Y, Blobel GA, Weiss MJ, Hardison RC. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone

modifications, and mRNA expression. Genome Res [Internet]. 2009 Dec;19(12):2172–84. Available from: http://dx.doi.org/10.1101/gr.098921.109

60. Tripic T, Deng W, Cheng Y, Zhang Y, Vakoc CR, Gregory GD, Hardison RC, Blobel GA. SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. Blood [Internet]. 2009 Mar 5;113(10):2191–201. Available from: http://dx.doi.org/10.1182/blood-2008-07-169417

61. Wu W, Morrissey CS, Keller CA, Mishra T, Pimkin M, Blobel GA, Weiss MJ, Hardison RC. Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis. Genome Res [Internet]. 2014 Dec;24(12):1945–62. Available from: http://dx.doi.org/10.1101/gr.164830.113

62. Shan J, Fu L, Balasubramanian MN, Anthony T, Kilberg MS. ATF4-dependent regulation of the JMJD3 gene during amino acid deprivation can be rescued in Atf4-deficient cells by inhibition of deacetylation. J Biol Chem [Internet]. 2012 Oct 19;287(43):36393–403. Available from: http://dx.doi.org/10.1074/jbc.M112.399600

63. Noh KM, Hwang JY, Follenzi A, Athanasiadou R, Miyawaki T, Greally JM, Bennett MVL, Zukin RS. Repressor element-1 silencing transcription factor (REST)-dependent epigenetic remodeling is critical to ischemia-induced neuronal death. Proc Natl Acad Sci U S A [Internet]. 2012 Apr 17;109(16):E962–71. Available from: http://dx.doi.org/10.1073/pnas.1121568109

64. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell [Internet]. 2014 Dec 18;159(7):1665–80. Available from: http://dx.doi.org/10.1016/j.cell.2014.11.021

65. Kaczynski J, Zhang JS, Ellenrieder V, Conley A, Duenes T, Kester H, van Der Burg B, Urrutia R. The Sp1-like protein BTEB3 inhibits transcription via the basic transcription element box by interacting with mSin3A and HDAC-1 co-repressors and competing with Sp1. J Biol Chem [Internet]. 2001 Sep 28;276(39):36749–56. Available from: http://dx.doi.org/10.1074/jbc.M105831200

66. Brigidi GS, Hayes MGB, Delos Santos NP, Hartzell AL, Texari L, Lin PA, Bartlett A, Ecker JR, Benner C, Heinz S, Bloodgood BL. Genomic Decoding of Neuronal Depolarization by Stimulus-Specific NPAS4 Heterodimers. Cell [Internet]. 2019 Oct 3;179(2):373–91.e27. Available from: http://dx.doi.org/10.1016/j.cell.2019.09.004

67. He Q, Johnston J, Zeitlinger J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat Biotechnol [Internet]. 2015 Apr [cited 2021 Dec 22];33(4). Available from: https://pubmed.ncbi.nlm.nih.gov/25751057/

68. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol [Internet]. 2015 Jan 5;109:21.29.1–21.29.9. Available from: http://dx.doi.org/10.1002/0471142727.mb2129s109

69. Westholm JO, Xu F, Ronne H, Komorowski J. Genome-scale study of the importance of binding site context for transcription factor binding and gene regulation. BMC Bioinformatics [Internet]. 2008 Nov 17 [cited 2021 Dec 22];9. Available from: https://pubmed.ncbi.nlm.nih.gov/19014636/

70. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet [Internet]. 2014 Dec;46(12):1311–20. Available from: http://dx.doi.org/10.1038/ng.3142

71. Rhee HS, Pugh BF. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. Curr Protoc Mol Biol [Internet]. 2012 Oct;Chapter 21:Unit 21.24. Available from: http://dx.doi.org/10.1002/0471142727.mb2124s100

72. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. Dynamic regulation of nucleosome positioning in the human genome. Cell [Internet]. 2008 Mar 7;132(5):887–98. Available from: http://dx.doi.org/10.1016/j.cell.2008.02.022

73. Lesluyes T, Johnson J, Machanick P, Bailey TL. Differential motif enrichment analysis of paired ChIP-seq experiments. BMC Genomics [Internet]. 2014 Sep 2;15:752. Available from: http://dx.doi.org/10.1186/1471-2164-15-752

74. Rubin JD, Stanley JT, Sigauke RF, Levandowski CB, Maas ZL, Westfall J, Taatjes DJ, Dowell RD. Transcription factor enrichment analysis (TFEA) quantifies the activity of multiple transcription factors from a single experiment. Communications biology [Internet]. 2021 Jun 2 [cited 2021 Dec 22];4(1). Available from: https://pubmed.ncbi.nlm.nih.gov/34079046/

75. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet [Internet]. 2004 Apr;5(4):276–87. Available from: http://dx.doi.org/10.1038/nrg1315

76. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, Fornes O, Leung TY, Aguirre A, Hammal F, Schmelter D, Baranasic D, Ballester B, Sandelin A, Lenhard B, Vandepoele K, Wasserman WW, Parcy F, Mathelier A. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res [Internet]. 2021 Nov 30; Available from: http://dx.doi.org/10.1093/nar/gkab1113

77. Pizzi C, Rastas P, Ukkonen E. Finding significant matches of position weight matrices in linear time. IEEE/ACM Trans Comput Biol Bioinform [Internet]. 2011 Jan;8(1):69–79. Available from: http://dx.doi.org/10.1109/TCBB.2009.35

78. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. aos [Internet]. 2001 Aug [cited 2021 Dec 22];29(4):1165–88. Available from: https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-4/The-control-of-the-false-discovery-rate-in-multiple-testing/10.1214/aos/1013699998.short

79. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods [Internet]. 2020 Mar;17(3):261–72. Available from: http://dx.doi.org/10.1038/s41592-019-0686-2

80. Ohler U, Liao GC, Niemann H, Rubin GM. Computational analysis of core promoters in the Drosophila genome. Genome Biol [Internet]. 2002 Dec 20;3(12):RESEARCH0087. Available from: http://dx.doi.org/10.1186/gb-2002-3-12-research0087

81. Wang Y, Stumph WE. RNA polymerase II/III transcription specificity determined by TATA box orientation. Proc Natl Acad Sci U S A [Internet]. 1995 Sep 12;92(19):8606–10. Available from: http://dx.doi.org/10.1073/pnas.92.19.8606

82. Butler JEF, Kadonaga JT. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev [Internet]. 2002 Oct 15;16(20):2583–92. Available from: http://dx.doi.org/10.1101/gad.1026202

83. Wadman IA, Osada H, Grütz GG, Agulnick AD, Westphal H, Forster A, Rabbitts TH. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. EMBO J [Internet]. 1997 Jun 2;16(11):3145–57. Available from: http://dx.doi.org/10.1093/emboj/16.11.3145

84. Han GC, Vinayachandran V, Bataille AR, Park B, Chan-Salis KY, Keller CA, Long M, Mahony S, Hardison RC, Pugh BF. Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution. Mol Cell Biol [Internet]. 2016 Jan 1;36(1):157–72. Available from: http://dx.doi.org/10.1128/MCB.00806-15

85. Sun X, Ren Z, Cun Y, Zhao C, Huang X, Zhou J, Hu R, Su X, Ji L, Li P, Mak KLK, Gao F, Yang Y, Xu H, Ding J, Cao N, Li S, Zhang W, Lan P, Sun H, Wang J, Yuan P. Hippo-YAP signaling controls lineage differentiation of mouse embryonic stem cells through modulating the formation of super-enhancers. Nucleic Acids Res [Internet]. 2020 Jul 27;48(13):7182–96. Available from: http://dx.doi.org/10.1093/nar/gkaa482

86. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, Zeitlinger J. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet [Internet]. 2021 Mar;53(3):354–66. Available from: http://dx.doi.org/10.1038/s41588-021-00782-6

87. Miraldi ER, Chen X, Weirauch MT. Deciphering cis-regulatory grammar with deep learning. Nat Genet [Internet]. 2021 Mar;53(3):266–8. Available from: http://dx.doi.org/10.1038/s41588-021-00814-1

88. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol [Internet]. 2014;15(12):550. Available from: http://dx.doi.org/10.1186/s13059-014-0550-8

89. Comoglio F, Simonatto M, Polletti S, Liu X, Smale ST, Barozzi I, Natoli G. Dissection of acute stimulus-inducible nucleosome remodeling in mammalian cells. Genes Dev [Internet]. 2019 Sep 1;33(17-18):1159–74. Available from: http://dx.doi.org/10.1101/gad.326348.119

90. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. Epigenetics Chromatin [Internet]. 2014 Nov 20;7(1):33. Available from: http://dx.doi.org/10.1186/1756-8935-7-33

91. Platanitis E, Decker T. Regulatory Networks Involving STATs, IRFs, and NFκB in Inflammation. Front Immunol [Internet]. 2018 Nov 13;9:2542. Available from: http://dx.doi.org/10.3389/fimmu.2018.02542

92. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biol [Internet]. 2019 Feb 26;20(1):45. Available from: http://dx.doi.org/10.1186/s13059-019-1642-2

93. Delos Santos NP, Texari L, Benner C. MEIRLOP: improving score-based motif enrichment by incorporating sequence bias covariates. BMC Bioinformatics [Internet]. 2020 Sep 16;21(1):410. Available from: http://dx.doi.org/10.1186/s12859-020-03739-4

94. Sadler AJ, Williams BRG. Interferon-inducible antiviral effectors. Nat Rev Immunol [Internet]. 2008 Jul;8(7):559–68. Available from: http://dx.doi.org/10.1038/nri2314

95. Hashim FA, Mabrouk MS, Al-Atabany W. Review of Different Sequence Motif Finding Algorithms. Avicenna J Med Biotechnol [Internet]. 2019 Apr;11(2):130–48. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31057715

96. Thibord F, Chan MV, Chen MH, Johnson AD. A year of COVID-19 GWAS results from the GRASP portal reveals potential genetic risk factors. HGG advances [Internet]. 2022 Apr 14 [cited 2022 Aug 1];3(2). Available from: https://pubmed.ncbi.nlm.nih.gov/35224516/

97. Mapping the human genetic architecture of COVID-19. Nature [Internet]. 2021 Dec [cited 2022 Aug 1];600(7889). Available from: https://pubmed.ncbi.nlm.nih.gov/34237774/

98. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. Nature [Internet]. 2018 Oct;562(7726):203–9. Available from: http://dx.doi.org/10.1038/s41586-018-0579-z

99. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet [Internet]. 2010 Jul;11(7):499–511. Available from: http://dx.doi.org/10.1038/nrg2796

100. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol [Internet]. 2012 Dec 27;8(12):e1002822. Available from: http://dx.doi.org/10.1371/journal.pcbi.1002822

101. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet [Internet]. 2005 Feb;6(2):95–108. Available from: http://dx.doi.org/10.1038/nrg1521

102. Bailey TL, Bodén M, Whitington T, Machanick P. The value of position-specific priors in motif discovery using MEME. BMC Bioinformatics [Internet]. 2010 Apr 9;11:179. Available from: http://dx.doi.org/10.1186/1471-2105-11-179

103. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res [Internet]. 2020 Jan 8;48(D1):D87–92. Available from: http://dx.doi.org/10.1093/nar/gkz1001

104. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc [Internet]. 2012 Sep;7(9):1728–40. Available from: http://dx.doi.org/10.1038/nprot.2012.101

105. Lam MTY, Duttke SH, Odish MF, Le HD, Hansen EA, Nguyen CT, Trescott S, Kim R, Deota S, Chang MW, Patel A, Hepokoski M, Alotaibi M, Rolfsen M, Perofsky K, Warden AS, Foley J, Ramirez SI, Dan JM, Abbott RK, Crotty S, Crotty Alexander LE, Malhotra A, Panda S, Benner CW, Coufal NG. Profiling Transcription Initiation in Peripheral Leukocytes Reveals Severity-Associated Cis-Regulatory Elements in Critical COVID-19 [Internet]. bioRxiv. 2021 [cited 2022 Jul 22]. p. 2021.08.24.457187. Available from: https://www.biorxiv.org/content/10.1101/2021.08.24.457187v1.abstract

106. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol [Internet]. 2010 May 6;6(5):e1000770. Available from: http://dx.doi.org/10.1371/journal.pcbi.1000770

107. Parts L, Stegle O, Winn J, Durbin R. Joint genetic analysis of gene expression data with inferred cellular phenotypes. PLoS Genet [Internet]. 2011 Jan 20;7(1):e1001276. Available from: http://dx.doi.org/10.1371/journal.pgen.1001276

108. Andrilenas KK, Ramlall V, Kurland J, Leung B, Harbaugh AG, Siggers T. DNA-binding landscape of IRF3, IRF5 and IRF7 dimers: implications for dimer-specific gene regulation. Nucleic Acids Res [Internet]. 2018 Mar 16;46(5):2509–20. Available from: http://dx.doi.org/10.1093/nar/gky002

109. Freitas RS, Crum TF, Parvatiyar K. SARS-CoV-2 Spike Antagonizes Innate Antiviral Immunity by Targeting Interferon Regulatory Factor 3. Front Cell Infect Microbiol [Internet]. 2021;11:789462. Available from: http://dx.doi.org/10.3389/fcimb.2021.789462

110. Honda K, Taniguchi T. IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. Nat Rev Immunol [Internet]. 2006 Sep;6(9):644–58. Available from: http://dx.doi.org/10.1038/nri1900

111. Kochs G, Janzen C, Hohenberg H, Haller O. Antivirally active MxA protein sequesters La Crosse virus nucleocapsid protein into perinuclear complexes. Proc Natl Acad Sci U S A [Internet]. 2002 Mar

5;99(5):3153–8. Available from: http://dx.doi.org/10.1073/pnas.052430399

112. Sacconi A, De Vitis C, de Latouliere L, di Martino S, De Nicola F, Goeman F, Mottini C, Paolini F, D'Ascanio M, Ricci A, Tafuri A, Marchetti P, Di Napoli A, De Biase L, Negro A, Napoli C, Anibaldi P, Salvati V, Duffy D, Terrier B, Fanciulli M, Capalbo C, Sciacchitano S, Blandino G, Piaggio G, Mancini R, Ciliberto G. Multi-omic approach identifies a transcriptional network coupling innate immune response to proliferation in the blood of COVID-19 cancer patients. Cell Death Dis [Internet]. 2021 Oct 29;12(11):1019. Available from: http://dx.doi.org/10.1038/s41419-021-04299-y

113. Chiang HS, Liu HM. The Molecular Basis of Viral Inhibition of IRF- and STAT-Dependent Immune Responses. Front Immunol [Internet]. 2018;9:3086. Available from: http://dx.doi.org/10.3389/fimmu.2018.03086

114. Wang Y, Wang H, Wei L, Li S, Liu L, Wang X. Synthetic promoter design in Escherichia coli based on a deep generative network. Nucleic Acids Res [Internet]. 2020 Jul 9;48(12):6403–12. Available from: http://dx.doi.org/10.1093/nar/gkaa325

115. Duttke SH, Montilla-Perez P, Chang MW, Li H, Chen H, Carrette LLG, de Guglielmo G, George O, Palmer AA, Benner C, Telese F. Glucocorticoid Receptor-Regulated Enhancers Play a Central Role in the Gene Regulatory Networks Underlying Drug Addiction. Front Neurosci [Internet]. 2022 May 16;16:858427. Available from: http://dx.doi.org/10.3389/fnins.2022.858427

116. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol [Internet]. 2015 Aug;33(8):831–8. Available from: http://dx.doi.org/10.1038/nbt.3300

117. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods [Internet]. 2015 Oct;12(10):931–4. Available from: http://dx.doi.org/10.1038/nmeth.3547

118. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis [Internet]. Vol. 47, Current Protocols in Bioinformatics. 2014. Available from:

http://dx.doi.org/10.1002/0471250953.bi1112s47

119. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics [Internet]. 2013 Jan 1;29(1):15–21. Available from: http://dx.doi.org/10.1093/bioinformatics/bts635

120. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, El Houdaigui B, Fatima R, Gall A, Garcia Giron C, Grego T, Guijarro-Clarke C, Haggerty L, Hemrom A, Hourlier T, Izuogu OG, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Gonzalez Martinez J, Marugán JC, Maurel T, McMahon AC, Mohanan S, Moore B, Muffato M, Oheh DN, Paraschas D, Parker A, Parton A, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Steed E, Szpak M, Szuba M, Taylor K, Thormann A, Threadgold G, Walts B, Winterbottom A, Chakiachvili M, Chaubal A, De Silva N, Flint B, Frankish A, Hunt SE, IIsley GR, Langridge N, Loveland JE, Martin FJ, Mudge JM, Morales J, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Cunningham F, Yates AD, Zerbino DR, Flicek P. Ensembl 2021. Nucleic Acids Res [Internet]. 2021 Jan 8;49(D1):D884–91. Available from: http://dx.doi.org/10.1093/nar/gkaa942

121. Zerbino DR, Johnson N, Juettemann T, Wilder SP, Flicek P. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. Bioinformatics [Internet]. 2014 Apr 1;30(7):1008–9. Available from: http://dx.doi.org/10.1093/bioinformatics/btt737

122. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics [Internet]. 2010 Sep 1;26(17):2204–7. Available from: http://dx.doi.org/10.1093/bioinformatics/btq351

123. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, Sandstrom R, Humbert R, Stamatoyannopoulos JA. BEDOPS: high-performance genomic feature operations. Bioinformatics [Internet]. 2012 Jul 15;28(14):1919–20. Available from: http://dx.doi.org/10.1093/bioinformatics/bts277

124. Krueger F, James F, Ewels P, Afyounian E, Schuster-Boeckler B. FelixKrueger/TrimGalore: v0.6.7

- DOI via Zenodo [Internet]. Zenodo; 2021. Available from: https://zenodo.org/record/5127899