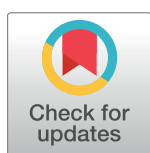**PLOS** | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Deconvolving sequence features that discriminate between overlapping regulatory annotations

Akshay Kakumanu[1], Silvia Velasco[2], Esteban Mazzoni[2], Shaun Mahony[1]*

1 Center for Eukaryotic Gene Regulation, Department of Biochemistry & Molecular Biology, The Pennsylvania State University, University Park, PA, United States of America, **2** Department of Biology, New York University, 100 Washington Square East, New York, NY, United States of America

* mahony@psu.edu

Check for updates

Abstract

Genomic loci with regulatory potential can be annotated with various properties. For example, genomic sites bound by a given transcription factor (TF) can be divided according to whether they are proximal or distal to known promoters. Sites can be further labeled according to the cell types and conditions in which they are active. Given such a collection of labeled sites, it is natural to ask what sequence features are associated with each annotation label. However, discovering such label-specific sequence features is often confounded by overlaps between the labels; e.g. if regulatory sites specific to a given cell type are also more likely to be promoter-proximal, it is difficult to assess whether motifs identified in that set of sites are associated with the cell type or associated with promoters. In order to meet this challenge, we developed SeqUnwinder, a principled approach to deconvolving interpretable discriminative sequence features associated with overlapping annotation labels. We demonstrate the novel analysis abilities of SeqUnwinder using three examples. Firstly, SeqUnwinder is able to unravel sequence features associated with the dynamic binding behavior of TFs during motor neuron programming from features associated with chromatin state in the initial embryonic stem cells. Secondly, we characterize distinct sequence properties of multi-condition and cell-specific TF binding sites after controlling for uneven associations with promoter proximity. Finally, we demonstrate the scalability of SeqUnwinder to discover cell-specific sequence features from over one hundred thousand genomic loci that display DNase I hypersensitivity in one or more ENCODE cell lines.

Author summary

Transcription factor proteins control gene expression by recognizing and interacting with short DNA sequence patterns in regulatory regions on the genome. Current genomics experiments allow us to find regulatory regions associated with a particular biochemical activity over the entire genome; for example, all regions where a particular transcription factor interacts with the genome in a given cell type. Given a collection of regulatory regions, we often aim to discover short DNA sequence patterns that are more common in

the collection than in other regions. Performing such "DNA motif-finding" analysis can give us hints about the patterns that determine gene regulation in the analyzed cell type.

Here we describe a new method for DNA motif-finding called SeqUnwinder. Our approach analyzes collections of regulatory regions where each has been labeled according to various biological properties. For example, the labels could correspond to various cell types in which the regulatory region is active. SeqUnwinder then performs machine-learning analysis to unravel DNA sequence features that are characteristic of each label (e.g. features that distinguish regulatory regions in each cell type from other cell types). SeqUnwinder is the first method to enable analysis of regulatory region collections that contain several overlapping labels.

Introduction

Many regulatory genomics analyses focus on finding DNA sequence features that are characteristic of a biological property. Given a set of sequences that are bound by a particular transcription factor (TF), for example, we typically aim to discover short, degenerate DNA patterns that may represent the DNA binding preferences of the TF itself, the binding preferences of coincident TFs, or general properties of the regions that make them favorable for binding.

The *de novo* DNA motif-finding problem is typically cast in the context of two mutually exclusive sequence sets. Most popular motif-finding methods use unsupervised machine-learning approaches to discover motifs in 'foreground' input sequences that are over-represented with respect to a set of 'background' sequences (e.g. "bound" vs. "unbound", respectively) [1,2]. Several other methods explicitly solve a two-class classification problem, where the goal is to find sequence features that discriminate between two mutually exclusive class labels [3–6].

Current characterizations of regulatory sites move beyond binary labels such as "bound" and "unbound". For example, in a given cell type, each regulatory element could be labeled as bound or unbound by each of several TFs and enriched or depleted for several chromatin states [7–9]. As we add more regulatory class labels, it becomes difficult to define mutually exclusive sets of sequences that are representative of each label. Relatedly, our analyses may become confounded by uneven degrees of overlap between the class labels, leading to incorrect associations between sequence features and regulatory activities. Therefore, a simple recasting of discriminative motif-finding as a multi-class classification problem (where classes are required to be mutually exclusive) is not always appropriate.

As an example, consider the hypothetical scenario presented in Fig 1A. In this example, a given TF's binding sites have been profiled in types A, B, and C. Thus, each TF binding event can be labeled as specific to a cell type or common to all or a subset. Let's assume that after further labeling the sites as being proximal or distal to promoters (Pr and Di, respectively), we find that the TF's binding sites in cell A are more likely to be promoter proximal than sites in other cell types. Promoter regions have sequence features that are distinct from distal regions (e.g. the presence of core promoter elements and distinct GC-content patterns). Therefore, if we search for sequence features that are discriminative of cell A's sites without accounting for the uneven overlaps with other labels, it is likely that some discovered features will actually be generic properties of proximal regions. Such results could in turn affect our conclusions regarding the biological mechanisms of TF binding in cell A. To resolve DNA features
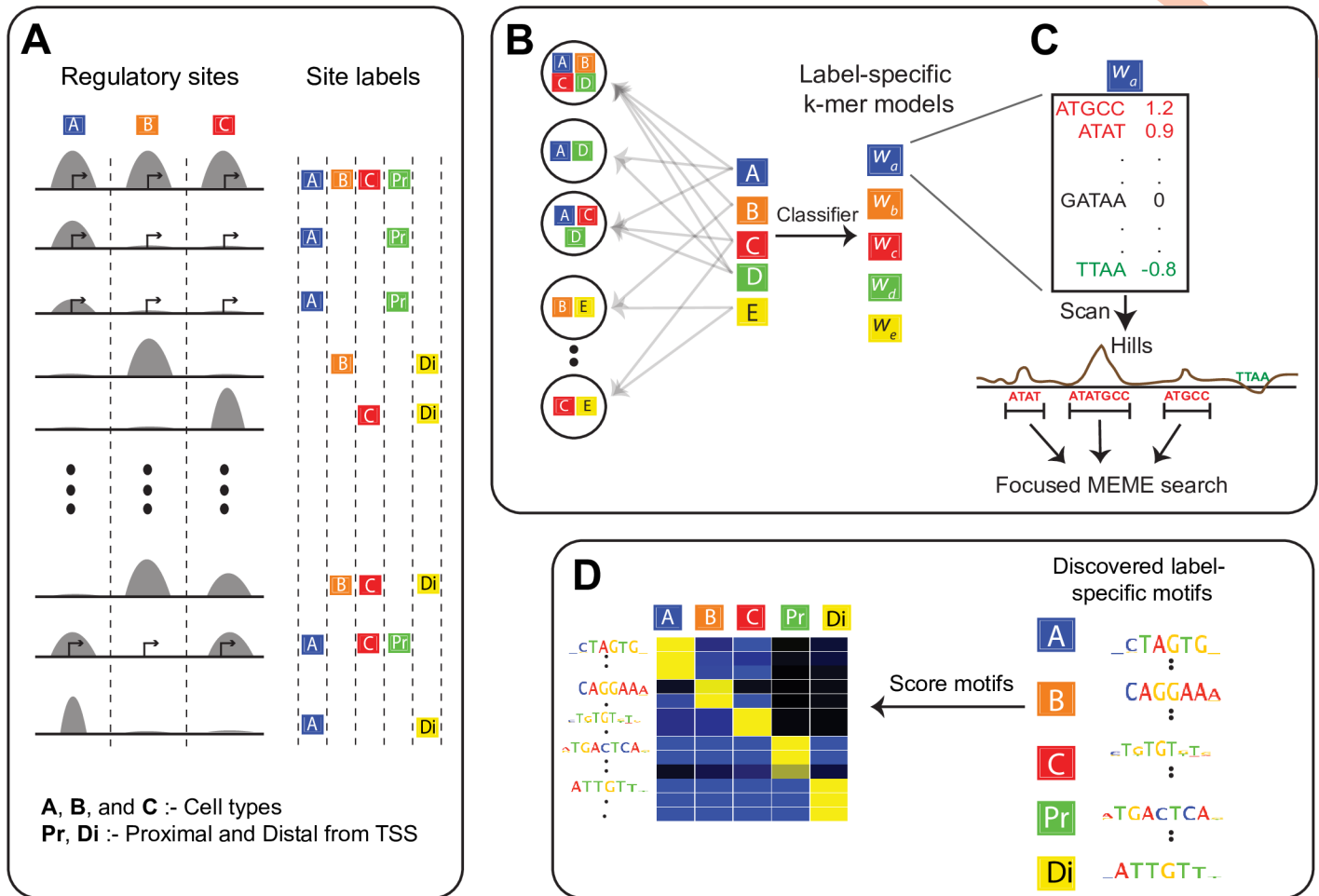
Fig 1. Overview of SeqUnwinder, which takes an input list of annotated genomic sites and identifies label-specific discriminative motifs. (A) Schematic showing a typical input instance for SeqUnwinder: a list of genomic coordinates and corresponding annotation labels. **(B)** The underlying classification framework implemented in SeqUnwinder. Subclasses (combination of annotation labels) are treated as different classes in a multi-class classification framework. The label-specific properties are implicitly modeled using L1-regularization. **(C)** Weighted *k*-mer models are used to identify 10-15bp focus regions called hills. MEME is used to identify motifs at hills. **(D)** *De novo* identified motifs in C) are scored using the weighted *k*-mer model to obtain label-specific scores.

https://doi.org/10.1371/journal.pcbi.1005795.g001

associated with each cell type's label from those associated with confounding labels (e.g. promoter proximity), we need motif-finders that are able to analyze multiple labels in parallel.

Almost all existing discriminative motif-finders assume that the class labels are mutually exclusive, and therefore cannot appropriately handle scenarios such as that outlined in Fig 1A. For example, the multi-class discriminative sequence feature frameworks proposed by Tavazoie and colleagues [3,10,11] are limited to analysis of mutually exclusive classes. A few existing methods do allow a limited analysis of datasets where annotation labels partially overlap, but these approaches were designed for two-class classification problems where the multi-task framework enables modeling of the "common" task in addition to the two classes. For example, Arvey, *et al.* [4] used a multi-task SVM classifier to learn sequence features associated with cell type-specific TF binding across two cell types, along with features shared by TF binding sites in both cell types. The group lasso based logistic regression classifier SeqGL [5] also implements a similar multi-task framework to identify features that are discriminative between two classes and features that are common to both. No existing discriminative feature discovery

method is applicable to multi-label classification scenarios where a set of genomic sequences contains several annotation labels with arbitrary rates of overlap between them.

In this work, we present SeqUnwinder, a hierarchical classification framework for characterizing interpretable sequence features associated with overlapping sets of genomic annotation labels. We demonstrate the unique analysis abilities of SeqUnwinder using both synthetic sequence datasets and collections of real TF ChIP-seq and DNase-seq experiments. In each demonstration, SeqUnwinder cleanly associates interpretable sequence features with various cell- or condition-specific annotation labels, while simultaneously removing the effects of confounding signals. SeqUnwinder scales effectively to large collections of genomic loci that have been annotated with several overlapping labels, and is thus designed to deal with the complexity of modern data sets.

Results

SeqUnwinder overview

The intuition behind SeqUnwinder is that sequence features associated with a particular annotation label should be similarly enriched across all subclasses spanned by the label (regardless of how the subclasses have been defined). SeqUnwinder's analysis begins by defining genomic site subclasses based on the combinations of labels annotated at these sites (Fig 1B). The site subclasses are treated as distinct classes for a multi-class logistic regression model that uses k-mer frequencies as predictors. At the same time, k-mer models are also learned for each label by incorporating them in an L1 regularization term (see Methods). In other words, while the k-mer weight parameters for each subclass are learned directly from the data, the weight parameters for the labels are learned exclusively through the regularization constraint. The regularization encourages each label's model to take the form of the features that are consistently enriched across the subclasses spanned by that label (Fig 1B). The trained classifier encapsulates weighted k-mer models specific to each label and each subclass (i.e. combination of labels). The label- or subclass-specific k-mer model is scanned across the original genomic sites to identify focused regions (which we term "hills") that contain discriminative sequence signals (Fig 1C). Finally, to aid interpretability, SeqUnwinder identifies over-represented motifs in the hills and scores them using label- and subclass-specific k-mer models (Fig 1D).

SeqUnwinder is easy to use, taking as input a list of DNA sequences or genomic coordinates that are each annotated with a set of user-defined labels. The labels can come from any source, enabling a high degree of analysis flexibility. SeqUnwinder implements a multi-threaded version of the ADMM [12] framework to train the model and typically runs in less than a few hours for most datasets. Output includes both k-mer models and position-specific scoring matrices and weights associating these motifs with each subclass and label.

SeqUnwinder deconvolves sequence features associated with overlapping labels

To demonstrate the properties of SeqUnwinder, we simulated 9,000 regulatory regions and annotated each of them with labels from two overlapping sets: A, B, C and X, Y (Fig 2A). We assigned a different motif to each label. At 70% of the sequences associated with each label, we inserted appropriate motif instances by sampling from the distributions defined by the position-specific scoring matrices of label assigned motifs (Fig 2A). We used this collection of sequences and label assignments to compare SeqUnwinder with a simple multi-class classification approach (MCC). In MCC training, each label was treated as a distinct class and therefore each regulatory sequence is included multiple times in accordance with its annotated labels.
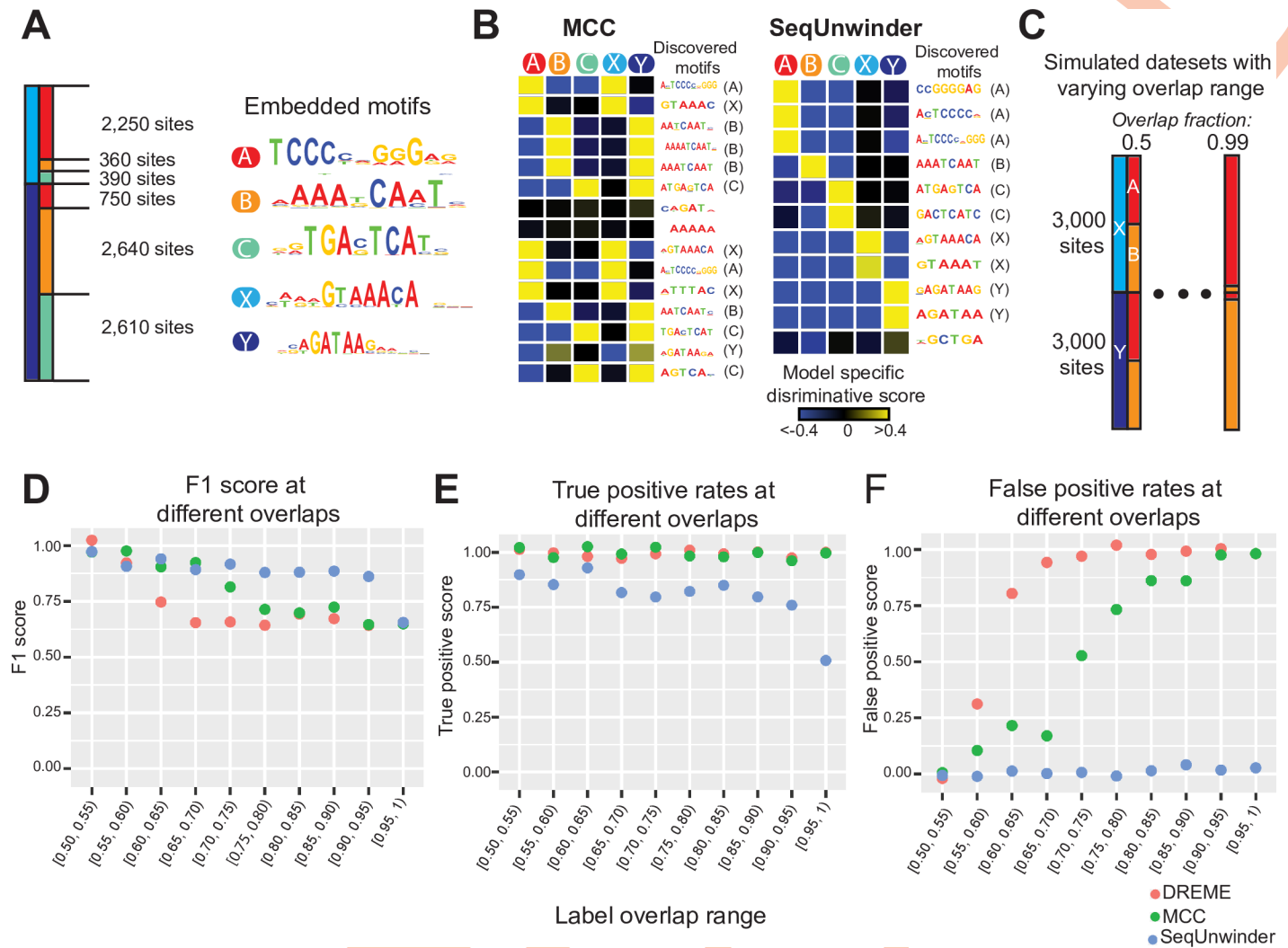
Fig 2. Performance of SeqUnwinder on simulated datasets. (A) 9000 simulated genomic sites with corresponding motif associations. **(B)** Label-specific scores for all *de novo* motifs identified using MCC (left) and SeqUnwinder (right) models on simulated genomic sites in "A". For consistency across figures, we fix the color saturation values to -0.4 and 0.4 **(C)** Schematic showing 100 genomic datasets with 6000 genomic sites and varying degrees of label overlap ranging from 0.5 to 0.99. **(D)** Performance of MCC (multi-class logistic classifier), DREME, and SeqUnwinder on simulated datasets in "C", measured using the F1-score, **(E)** true positive rates, and **(F)** false positive rates.

SeqUnwinder and the MCC model correctly identify motifs similar to all inserted motifs (Fig 2B). However, the MCC approach makes several incorrect motif-label associations, potentially due to high overlap between labels. In contrast, the label-specific scores of the identified motifs in the SeqUnwinder model are not confounded by overlap between annotation labels. For example, even though labels X and A highly overlap, SeqUnwinder correctly assigns each motif to its respective label.

Next, we assessed the performance of SeqUnwinder at different levels of label overlaps. We simulated 100 datasets with 6000 simulated sequences, varying the degree of overlap between two sets of labels ({A, B} and {X, Y}) from 50% to 99% (Fig 2C). We then compared SeqUnwinder with MCC and DREME [1], a popular discriminative motif discovery tool. Since DREME takes only two classes as input: a foreground set and a background set, we ran four different DREME runs for each of the four labels. We calculated the true positive (discovered motif

correctly assigned to a label) and false positive (discovered motif incorrectly assigned to a label) rates based on the true label assignments. We used these measures to calculate the F1 score (harmonic mean of precision and recall) at different overlapping levels (Fig 2D).

Fig 2D demonstrates the range of label overlap rates in which SeqUnwinder outperforms the alternative approaches. When the labels are uncorrelated (i.e. low or random overlap), the sequence features associated with each label do not confound one another and thus all methods perform similarly well in characterizing label-specific motifs. On the other hand, when the labels are highly correlated (i.e. high overlap), it becomes impossible for any method to correctly assign sequence features to the correct labels. SeqUnwinder performs better than the other approaches in the intermediate range of label overlaps, and accurately characterizes label-specific sequence features even when the simulated labels overlap at 90% of sites. More specifically, SeqUnwinder consistently has a false positive rate (incorrectly assigning motifs to labels) of zero at the cost of a modest decrease in true positive rates (recovering all motifs assigned to a label) (Fig 2E and Fig 2F).

SeqUnwinder uncovers co-factor driven TF binding dynamics during iMN programming

To demonstrate its unique abilities in a real analysis problem, we use SeqUnwinder to study TF binding during induced motor neuron (iMN) programming. Ectopic expression of Ngn2, Isl1, and Lhx3 in mouse embryonic stem (ES) cells efficiently converts the resident ES cells into functional spinal motor neurons [13,14]. We recently characterized the dynamics of motor neuron programming by studying TF binding, chromatin dynamics, and gene expression over the course of the 48hr programming process [14]. We found that two of the ectopically expressed TFs, Isl1 & Lhx3, bind together at the vast majority of their targets during the programming process. Using MultiGPS [15], we also found that this cooperative pair of TFs shifted their binding targets during programming, and we used three mutually exclusive labels–early, shared, and late–to annotate Isl1/Lhx3 binding sites according to their observed dynamic occupancy patterns. Early sites were bound by Isl1/Lhx3 only during earlier stages of programming, shared sites were constantly bound over the entire programming process, and late sites were only bound during the final stage of programming.

In our previous work, we demonstrated that the early Isl1/Lhx3 sites were more accessible in the initial pluripotent cells, and we suggested that some early sites are the result of opportunistic Isl1/Lhx3 binding to ES enhancer regions [14]. However, this raises a question that was not addressed in our earlier work: if we discover sequence features at early sites, how can we tell if those features are specifically associated with Isl1/Lhx3 as opposed to reflecting on coincident properties of ES enhancers?

In order to assess the potential confounding effects of ES regulatory sites, we trained a random forest classifier to further categorize all Isl1/Lhx3 bound sites using two additional labels: "ES-active and "ES-inactive" (see Methods). Annotating Isl1/Lhx3 sites using both sets of labels (Isl1/Lhx3 binding dynamics and ES activity) results in six different subclasses. As can be seen from Fig 3A, early sites have a higher propensity to also be active prior to ectopic TF expression in the starting ES cells. Conversely, the late sites are more likely to be inactive in ES cells.

We next trained SeqUnwinder on the multi-label Isl1/Lhx3 dataset, and compared the results with those of DREME and the simple MCC approach described in the previous section (Fig 3B, S1 Fig, S2 Fig, S1 Table). All methods discover similar sets of motifs. For example, both the SeqUnwinder and MCC approaches find motifs corresponding to the binding preferences of Oct4, Zfp281, Onecut-family TFs, and homeodomain TF motifs corresponding to the
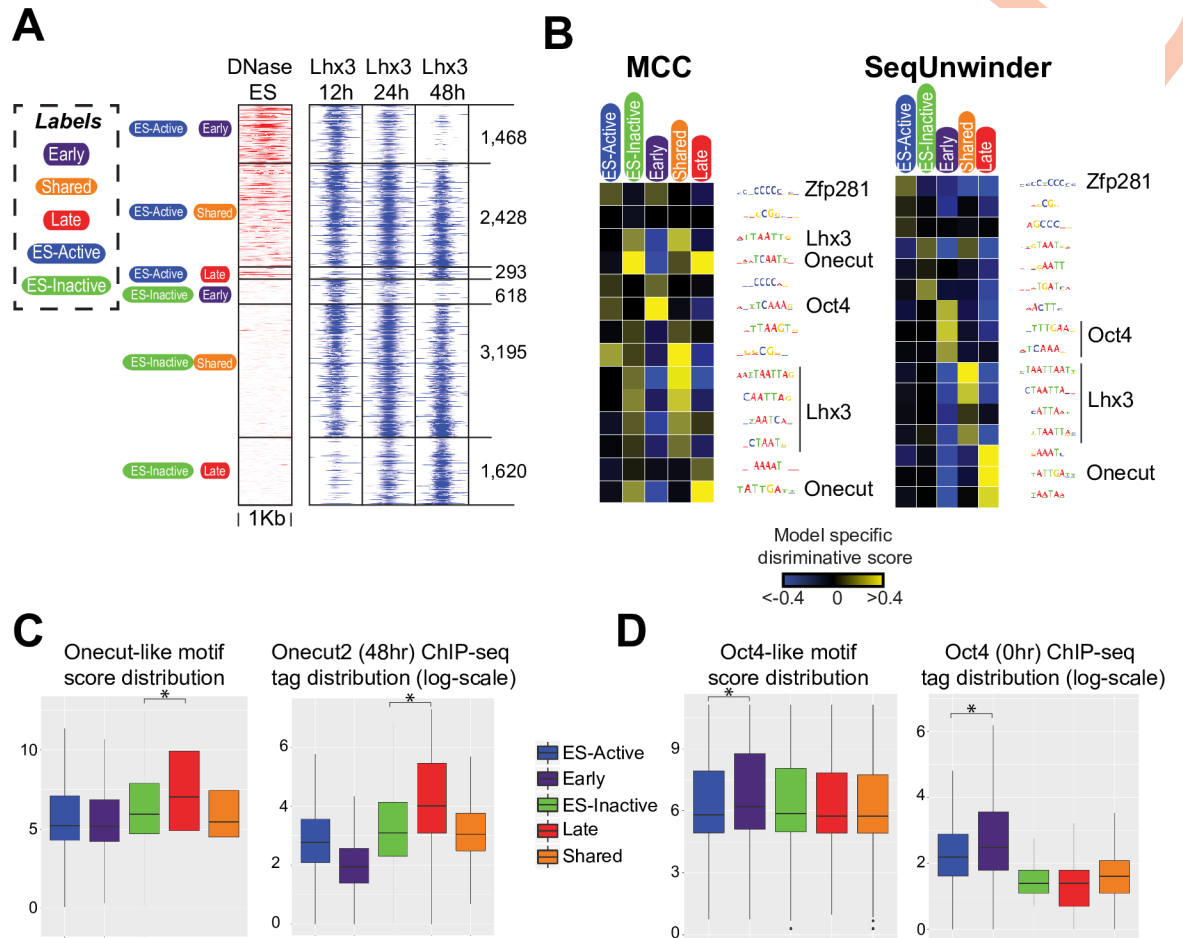
Fig 3. SeqUnwinder analysis of Lhx3 binding classes during iMN programming. (A) Lhx3 binding sites labeled using their dynamic binding behavior and ES chromatin activity statuses. **(B)** Label-specific scores of *de novo* motifs identified at Lhx3 binding sites defined in "A" using MCC (left) and SeqUnwinder (right) models. For consistency across figures, we fix the color saturation values to -0.4 and 0.4. **(C)** Log-odds score distribution of *de novo* discovered Onecut-like motif at "ES-active", "ES-inactive", "Early", "Shared", and "Late" sites (left panel). Distribution of Onecut2 (48hr) ChIP-seq tag counts in log-scale at "ES-active", "ES-inactive", "Early", "Shared", and "Late" sites (right panel). **(D)** Log-odds score distribution of *de novo* discovered Oct4-like motif at "ES-active", "ES-inactive", "Early", "Shared", and "Late" sites (left panel). Distribution of Oct4 (0hr) ChIP-seq tag counts in log-scale at "ES-active", "ES-inactive", "Early", "Shared", and "Late" sites (right panel). Statistical significance calculated using Mann-Whitney-Wilcoxon test (*: P-value < 0.001).

https://doi.org/10.1371/journal.pcbi.1005795.g003

cognate Isl1/Lhx3 binding preference (Fig 3B). However, the different approaches produce different associations between motifs and labels. For example, the MCC approach associates the Oct4 motif with both the "early" and "ES-active" labels, and it associates the Onecut motif with both "late" and "ES-inactive" labels (S1 Fig). DREME similarly makes ambiguous associations (S2 Fig). SeqUnwinder, in contrast, makes much cleaner associations; the Oct4 motif is only associated with the "early" label, and the Onecut motif is only associated with the "late" label, suggesting that these motifs are not merely coincidental features due to the ES activity status of the binding sites.

The SeqUnwinder motif-label associations suggest that Isl1/Lhx3 bind cooperatively with Oct4 and Onecut TFs at subsets of early and late binding sites, respectively. As described in our earlier work, we characterized Onecut2 binding to be highly enriched at late Isl1/Lhx3 sites during iMN programming [14]. We also found that late sites are not bound by Isl1/Lhx3

(and iMN programming does not proceed) in cellular conditions under which Onecut TFs are not expressed [14], supporting a model in which late Isl1/Lhx3 binding is dependent on Onecut TFs. Analysis of motif log-odds scores and Onecut2 ChIP enrichment further support SeqUnwinder's prediction that the Onecut motif is not merely a general feature of ES-inactive sites (Fig 3C).

Conversely, Oct4 is predicted by SeqUnwinder to be a specific feature of "early" binding sites, and not merely an artifact associated with "ES-active" sites. Using ChIP-seq, we profiled the binding of Oct4 immediately before NIL induction. As shown in Fig 3D, Oct4 motif log-odds scores and ChIP-seq tags show a preferential enrichment at early Isl1/Lhx3 sites, in line with SeqUnwinder's prediction.

Interestingly, the motif features that are most highly associated with shared binding sites all correspond to homeodomain TF motifs of the type bound by Isl1/Lhx3 (Fig 3B and S1 Fig). One possible explanation is that there are stronger or more frequent cognate motif instances at sites bound by a given TF across multiple timepoints, or indeed across multiple unrelated cell types. We further assess this hypothesis in the following section.

Our analysis of Isl1/Lhx3 binding during iMN programming thus serves as an example analysis scenario in which SeqUnwinder identifies motif features associated with multiple overlapping labels, leading to testable hypotheses about co-factors that serve mechanistic roles at subsets of binding sites.

Multi-condition TF binding sites are characterized by stronger cognate motif instances

The sequence properties of tissue-specific TF binding sites have been extensively studied [4,5,16]. As might be expected, sites that are bound by a given TF in only one cell type are often enriched for motifs of other TFs expressed in that cell type. Therefore, a given TF's cell-specific binding activity is likely determined by context-specific interactions with other expressed regulators.

Most TFs also display cell-invariant binding activities; each TF typically has a cohort of sites that appear bound in all or most cellular conditions in which that TF is active. Despite the potential regulatory significance of such multi-condition binding sites, little is known about the sequence properties that enable a TF to bind them regardless of cellular conditions. Studies of individual TFs suggest that binding affinity to cognate motif instances may play a role in distinguishing multi-condition binding sites from tissue-specific sites [15,17].

In order to characterize sequence discriminants of multi-condition TF binding sites across a wider range of TFs, we curated multi-condition ChIP-seq experiments from the ENCODE project. We restricted our analysis to the 17 sequence-specific TFs profiled in all 3 primary ENCODE cell-lines (K562, GM12878, and H1-hESC; see Methods) [18]. For each TF, we used MultiGPS analysis to curate sets of tissue-specific sites in each cell type, and a further set of sites that are "shared" across all three cell types (see Methods).

For most examined TFs, the majority of shared binding sites were located in promoter proximal regions (S3 Fig). As outlined in the Introduction, promoter proximal sites are known to have distinct sequence biases, which could confound the discovery of sequence features associated with shared sites. We therefore further labeled each TF's binding sites as being located proximal or distal to annotated TSSs. Introducing the proximal and distal labels should marginalize the proximal bias at shared sites, as the sequence features learned by SeqUnwinder at shared sites must be consistently enriched at both proximal and distal sites. Each examined TF's binding sites is thus categorized into eight subclasses, each of which is composed of combinations of six distinct labels.

We applied SeqUnwinder to each labeled sequence collection in order to characterize label-specific sequence features (see S2 Table for cross-validation classification performance values). We illustrate the process with SeqUnwinder's results for YY1. We started with a total of ~35,000 YY1 binding events called by MultiGPS across the three cell types, categorized into the eight aforementioned subclasses (Fig 4A). SeqUnwinder identifies several *de novo* motifs in this collection (Fig 4B). Interestingly, SeqUnwinder predicts that a motif matching the cognate YY1 motif is strongly associated with the "shared" label. The cell-type specific, proximal and distal labels had low or negative scores for this cognate motif. Note here that a non-positive label-specific score for a motif does not necessarily imply complete absence of instances of that motif. A significant depletion of motif instances at sites annotated by a label compared to other labels can very likely result in non-positive scores. Cell-type specific sites had higher scores for co-factor motifs. For example, H1-hESC specific sites were enriched in instances of a TEAD-like motif, while K562-specific sites and GM12878-specific sites were enriched for a GATA-like motif and an IRF-like motif, respectively. In fact, GATA2 ChIP-seq reads in K562, IRF4 ChIP-seq reads in GM12878, and TEAD4 ChIP-seq reads in H1hESC showed striking enrichment at corresponding cell-specific YY1 binding sites (Fig 4A).
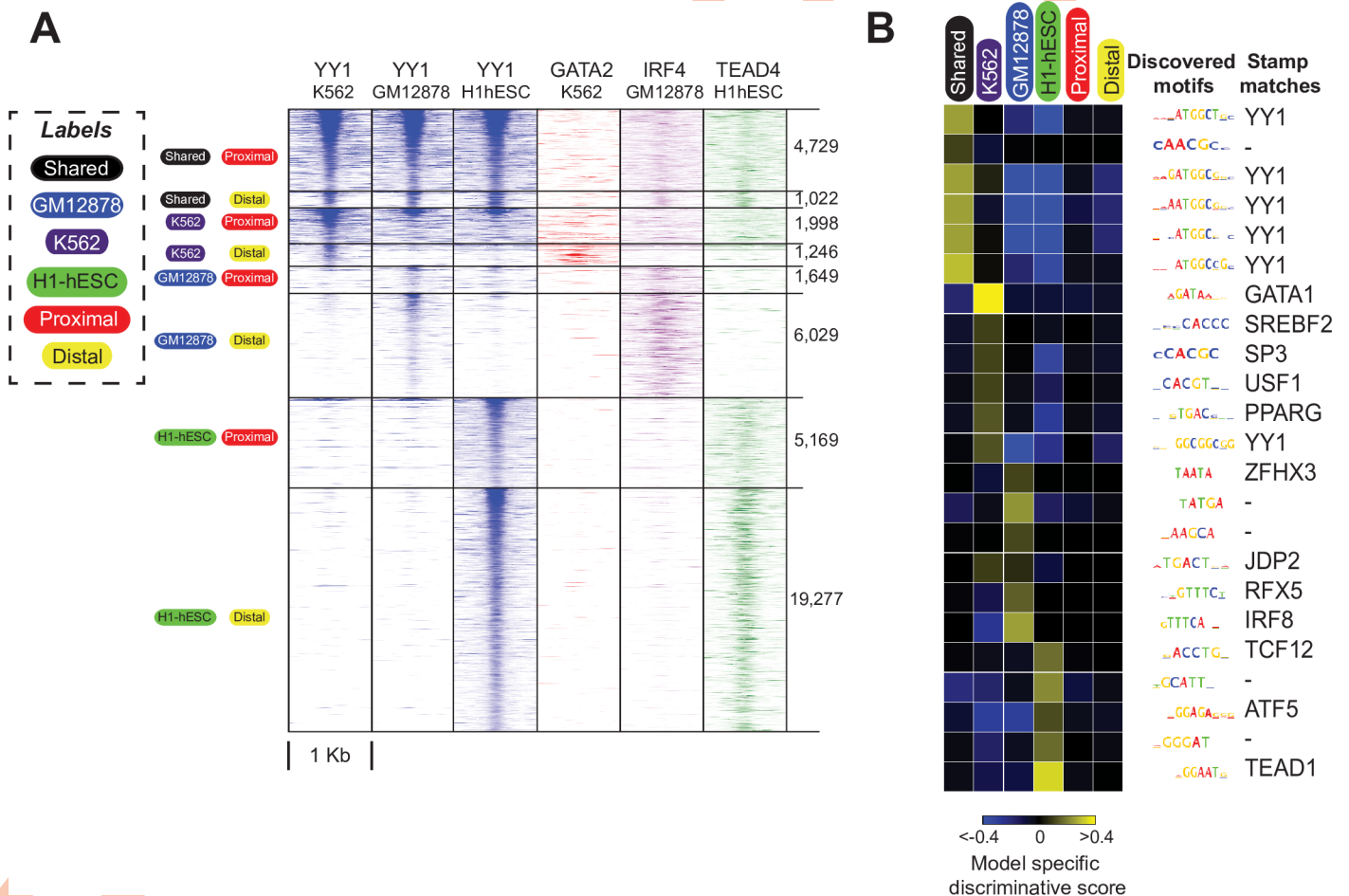

Fig 4. SeqUnwinder analysis of sequence features at multi-condition TF binding sites for ENCODE YY1 datasets. (A) Heatmaps showing the YY1 ChIP-seq reads at curated YY1 binding sites, stratified based on binding across cell-lines and distance from annotated mRNA TSS. The order of subclasses is: Shared and Proximal, Shared and Distal, K562 and Proximal, K562 and Distal, GM12878 and Proximal, GM12878 and Distal, H1-hESC and Proximal, and H1-hESC and Distal. **(B)** *De novo* motifs and corresponding label-specific scores identified using SeqUnwinder at events defined in A). For consistency across figures, we fix the color saturation values to -0.4 and 0.4.

https://doi.org/10.1371/journal.pcbi.1005795.g004

Analogous results were observed for many of the examined factors. For 13 out of the 17 examined factors, SeqUnwinder predicts that the cognate motif is highly associated with the "shared" label (Fig 5A). Despite significant overlaps between shared sites and promoter proximal sites (S3 Fig), the cognate motifs were not found to be predictive of any TF's "proximal"
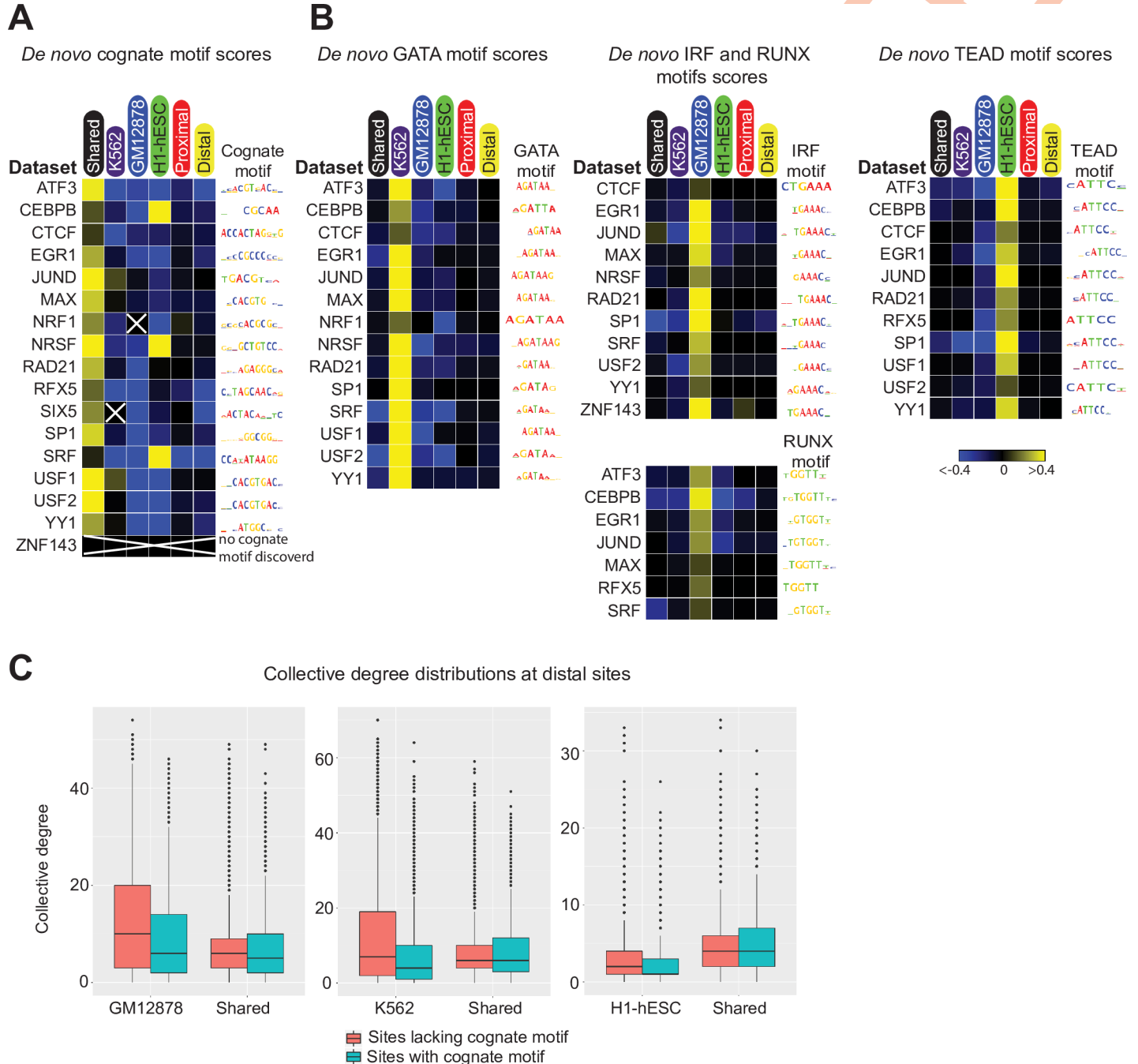

Fig 5. SeqUnwinder analysis of sequence features at multi-condition TF binding sites for 17 ENCODE TFs. (A) Label-specific scores of *de novo* discovered cognate motifs across all 17 ENCODE TF datasets. SeqUnwinder did not discover a cognate motif for ZNF143. GM12878-enriched sites for NRF1 and H1-hESC-specific sites for SRF were excluded because of low number of binding sites. **(B)** Label-specific scores of *de novo* discovered GATA-like, IRF-like, RUNX-like and TEAD-like motifs. **(C)** Collective degree distributions at distal shared and cell-specific sites further stratified based on presence of cognate motif. For consistency across figures, we fix the color saturation values to -0.4 and 0.4.

https://doi.org/10.1371/journal.pcbi.1005795.g005

label (Fig 5A). Furthermore, the cognate motif was not specifically predictive of cell-type-specific labels for the examined TFs, with the exception of H1-hESC-specific sites for CEBPB, NRSF and SRF. An orthogonal analysis of log-odds motif scoring distributions across each TF's labels is consistent with the SeqUnwinder results (S4 Fig). When we ran DREME on the same datasets for comparison, the association of cognate motif to shared sites was less clear. For 9 of the tested factors, DREME results associated the cognate motif with more than one label (S5 Fig).

We also examined the motifs that SeqUnwinder predicts to be associated with cell-type-specific binding labels. Interestingly, we found IRF and RUNX motifs enriched at GM12878-specific binding sites for 11 and 7 of the 17 examined TFs, respectively. Similarly, the GATA motif was predictive of K562-specific binding for 14 of the 17 examined TFs. A TEAD-like motif was predictive of H1-hESC specific sites for 11 of the 17 TFs (Fig 5B). The observation that cell-type-specific sites are depleted for cognate motif instances but are enriched for motif instances of other lineage-specific regulators is consistent with the "TF collective" model proposed by Junion and colleagues [19]. Under this model, the cooperative binding of large numbers of TFs is driven by the presence of motifs for a subset of lineage-specific factors that drive recruitment of the rest (i.e. the motifs for some TFs need not always be present).

To further support the "TF collective" interpretation of SeqUnwinder's results, we tested the degree to which TSS-distal cell-type-specific sites are bound by numerous other TFs. We first defined a binding site's "collective degree" as the number of distinct TFs with nearby ChIP-seq peaks. To calculate collective degree, we used a total of 158, 102, and 202 ChIP-seq datasets in GM12878, H1-hESC, and K562 cell-types, respectively. From Fig 5C, it is clear that distal K562- and GM12878-specific sites lacking a cognate motif instance have higher collective degrees. Similar findings were previously identified at the "high occupancy of transcription-related factors (HOT)" regions [20].

SeqUnwinder identifies sequence features at shared and cell-specific DHS in six different ENCODE cell-lines

Finally, we aim to demonstrate the utility of SeqUnwinder in identifying sequence features at large numbers of genomic loci annotated with several labels. We first annotated a large collection of DNase I hypersensitive (DHS) sites with six cell-line labels depending on the enrichment of DNase-seq reads (Fig 6A). If we had used analysis methods that rely on mutually exclusive categories, we would need to restrict analysis to ~97,000 sites labeled as either shared or exclusive to one of the six cell types [21]. Indeed, these strict category definitions may introduce sequence composition biases into each category. However, by taking advantage of SeqUnwinder's unique framework to pool information from all subclasses, we can analyze ~140,000 DHS sites that we annotate into 22 subclasses as shared (i.e. enriched in 5 or more cell types) or specific to one or two cell types (Fig 6A, S3 Table).

SeqUnwinder identifies several motifs in this large collection of DHS sites, including those previously associated with specific cell-types [22–24] (Fig 6B). For example, components of the CTCF motif were highly predictive of shared DHS sites. This result is consistent with previous findings suggesting relatively invariant CTCF binding across cellular contexts [25,26]. RUNX, IRF and NF-κB motifs were enriched at GM12878-specific DHS sites. These motifs were also discovered by others at GM12878-specific DHS sites [5,23]. Motifs corresponding to GATA TFs, key regulators of erythroid development [27–29], were enriched at K562-specific DHS sites. SNAI and TEAD motifs were enriched at H1-hESC sites, consistent with previous observations [5]. JUND and FOS motifs were enriched at HeLa-S3-specific DHS sites. Motifs for HNF4A and FOX TFs, known master regulator of hepatocytes [30–33], were enriched at
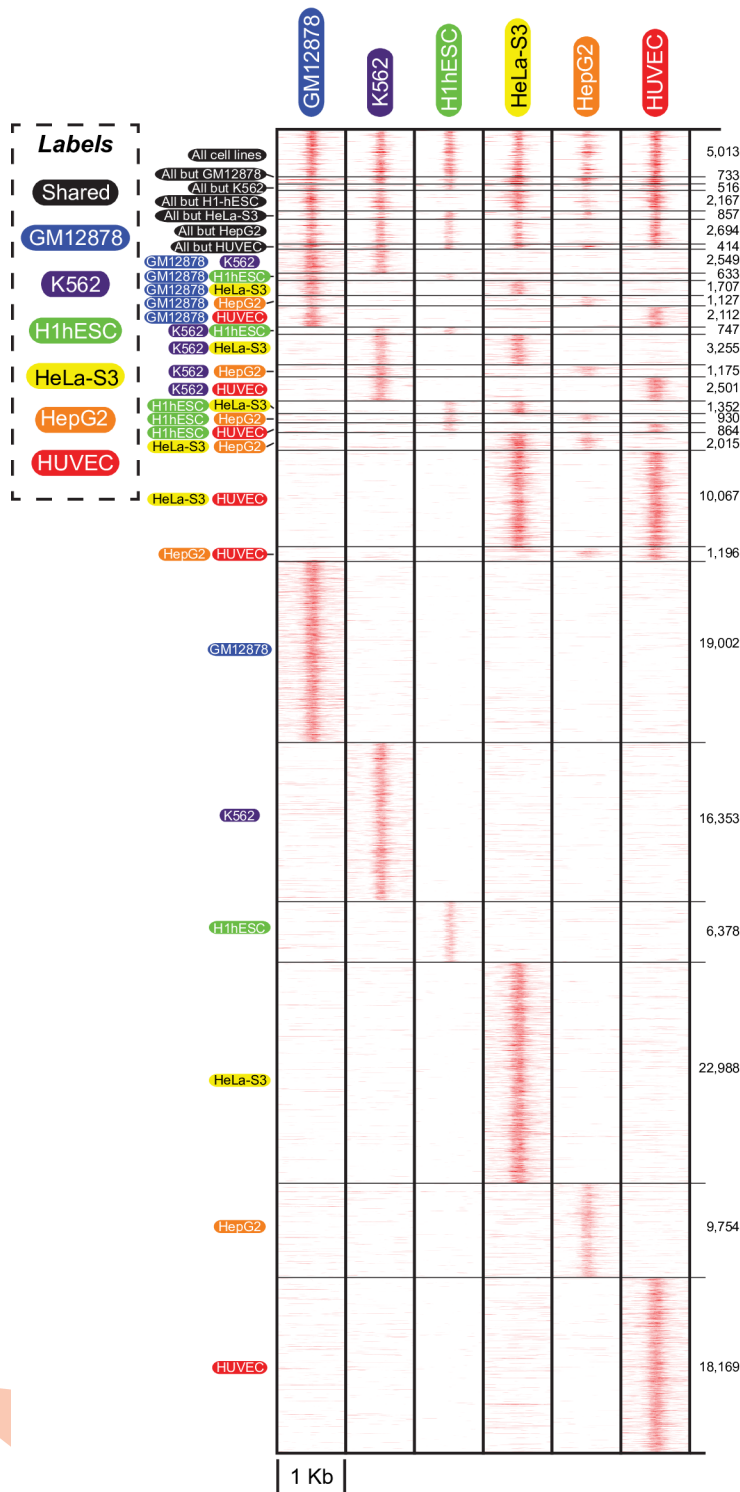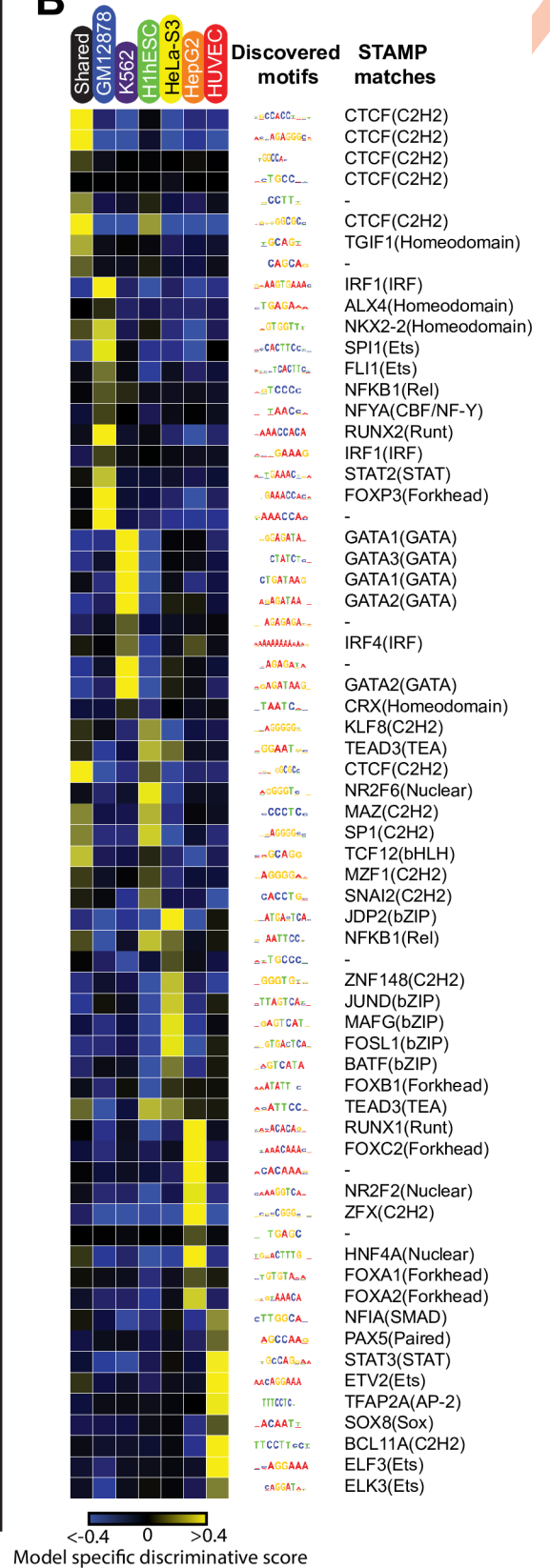
Fig 6. Discriminative sequence feature analysis at DHS sites in 6 different ENCODE cell-lines using SeqUnwinder. (A) ~140K DHSs sites annotated with 6 different cell-line labels used to identify cell-line specific and shared sequence features. **(B)** Label-specific scores of all the *de novo* motifs identified at DHSs sites in "A". For consistency across figures, we fix the color saturation values to -0.4 and 0.4.

HepG2-specific DHS sites. Finally, motifs belonging to the ETS class of TFs were enriched at HUVEC-specific DHS sites ([Fig 6B]). ETS factors have been shown to directly convert human fibroblasts to endothelial cells [34]. Interestingly, some of the motifs associated with cell-type specific DHS sites were also found in our analyses of cell-type specific TF binding sites above ([Fig 5B]). For example, IRF, GATA, and TEAD motifs associated with GM12878, K562, and H1-hESC specific DHSs were also predictive of corresponding cell-type specific binding for many of the analyzed TFs.

These results demonstrate that SeqUnwinder scales effectively in characterizing sequence features at thousands of regulatory regions annotated by several different overlapping labels.

Discussion

Classification models have shown great potential in identifying sequence features at defined genomic sites. For example, Lee *et al.* [3] trained an SVM classifier to discriminate putative enhancers from random sequences using an unbiased set of k-mers as predictors. The choice of kernel function is key to the performance of an SVM classifier. Several variants of the basic string kernel (e.g. mismatch kernel [35], di-mismatch kernel [4], wild-card kernel [5,35], and gkm-kernel [36]) have been proposed and have been shown to substantially improve the classifier performance. Several complementary methods using DNA shape features in a classification framework have also provided insight on the role of subtle shape features that distinguish bound from unbound sites [37–39]. More recently, deep learning models have also been harnessed to predict TF binding sites from unbound sites [6].

In this work, we focus not on the form of the training features, but rather on the tangential problem of identifying sequence features that discriminate several annotations applied to a set of genomic locations. Most existing methods have been developed and optimized to identify sequence features that discriminate between mutually exclusive classes (e.g. bound and unbound sites). However, when considering different sets of genomic annotation labels, overlaps between them are very likely and can confound results. To systematically address this, we developed SeqUnwinder.

Using three analysis scenarios based on real ChIP-seq and DNase-seq datasets, we have demonstrated that SeqUnwinder provides a unique ability to deconvolve discriminative sequence features at overlapping sets of labeled sites. Our applications are chosen to demonstrate that SeqUnwinder has the ability to predict the identities of TFs responsible for particular regulatory site properties, while accounting for potential sources of bias.

For example, in our previous characterization of Isl1/Lhx3 binding dynamics during motor neuron programming, we discovered motifs that were enriched at early and late binding site subsets [14]. However, our analyses were potentially confounded by a correlation between TF binding dynamics and the chromatin properties of the sites in the pre-existing ES cells. Therefore, the motifs that we previously assigned to early or late TF binding behaviors could have been merely associated with ES-active and ES-inactive sites, respectively. By implicitly accounting for the effects of overlapping annotation labels, SeqUnwinder can deconvolve sequence features associated with motor neuron programming dynamics and ES chromatin status. Our analyses support an association between Oct4 binding and early Isl1/Lhx3 binding sites, along with our previously confirmed association between Onecut TFs and late Isl1/Lhx3 binding sites [14].

Our analyses of ENCODE ChIP-seq and DNase-seq datasets demonstrate the flexibility and scalability of SeqUnwinder. In analyzing TF binding across multiple cell types, we used SeqUnwinder to account for promoter proximity as a potential confounding feature. Our results add to the growing evidence that multi-condition TF binding sites tend to be distinguished by better quality instances of the primary cognate motif [15,17]. For example, Gertz *et al.*, showed that ER (estrogen receptor) binding sites bound in both ECC1 and T4D7, two human cancer cell lines, had higher affinity instances of EREs (estrogen response elements) compared to cell-specific binding sites. Indeed, even the "shared" binding sites for Isl1/Lhx3 in our first demonstration are characterized by stronger instances of the Isl1/Lhx3 cognate binding motifs (Fig 3B). These results suggest that many TFs have a set of binding sites that are bound across a broad range of cellular contexts, and which are characterized by better quality cognate motif instances. Furthermore, our results support a model in which cell-type-specific sites lacking cognate motif instances are bound in a "TF collective" fashion [19].

Interestingly, SeqUnwinder discovers consistent motif features to be predictive of cell-specific binding sites across several examined TF ChIP-seq collections. For example, SeqUnwinder discovers GATA, IRF and TEAD motifs at K562-, GM12878- and H1hESC-specific TF binding sites, respectively. These same motifs are also discovered by SeqUnwinder to be predictive of appropriate cell-specific DNase I hypersensitivity in a large collection of DHS sites across 6 different cell types. SeqUnwinder's characterization of cell-specific motif features in collections of DNase-seq datasets may therefore serve as a source of predictive features for efforts that aim to predict cell-specific TF binding from accessibility experimental data alone [39–41].

There remain several areas of possible future improvement within SeqUnwinder's hierarchical multi-label classification approach to discriminative motif-finding. SeqUnwinder does not currently model any relationships or correlations between class labels. For example, we might expect similar cell types to have similar cell-specific motif features within their regulatory regions. Incorporating graphs defined by label similarities [42,43] may thus be productive in the context of analyses across cell lineages or developmental time-series. SeqUnwinder may also be easily extended to incorporate different kinds of sequence kernels and DNA shape features [35,36,44].

In summary, SeqUnwinder provides a flexible framework for analyzing sequence features in collections of related regulatory genomic experiments, and uniquely enables the principled discovery of sequence motifs associated with multiple overlapping annotation labels.

Methods

SeqUnwinder model

The core of SeqUnwinder is a multi-class logistic regression classifier trained on subclasses of genomic sites. The training features for the classifier are based on k-mer frequencies in a fixed window around input loci. The value or range of k is user-definable in the SeqUnwinder software, but all analyses in this work use models based on all 4-mers and 5-mers. When counting frequencies, we map each k-mer to the same entry as its reverse complement. To account for differences in the ranges of k-mer frequencies, we standardize the feature vectors such that each k-mer has a zero mean and unit variance across the entire training dataset.

The parameters of SeqUnwinder are k-mer weights for each subclass (combination of annotation labels). In addition, SeqUnwinder also models the label-specific k-mer weights by incorporating them in the L1 regularization term. Briefly, label-specific k-mer weights are encouraged to be similar to k-mer weights in all subclasses the label spans by regularizing on the differences of k-mer weights. Note that our approach is conceptually similar to hierarchical

classification approaches such as that described by [45], although we use L1 regularization as opposed to L2.

The overall objective function of SeqUnwinder is: -

$$-\sum_{i=1}^{M}\sum_{n\in T}b_i y_{in}\log\left[\frac{\exp(w_n x_i)}{\sum_{n\in T}\exp(w_n x_i)}\right] + \lambda\sum_{n\in T}\sum_{p\in\Pi(n)}\|w_n - w_p\|_1 \qquad (1)$$

In the above equation; M is the total number of genomic loci in all subclasses, T is the set of all subclasses, b_i is the weight given to the genomic site i, w_n is the k-mer weight vector for sub-class n, x_i is a vector of k-mer counts for the genomic site i, y_{in} is a binary indicator variable denoting the subclass of genomic site i, λ is the regularization co-efficient, $\Pi(n)$ is the set of all labels spanning the subclass n, and w_p is the k-mer weight vector for label p. Values for b_i are chosen to account for class imbalances. Hence, the value of b_i for a genomic site i belonging to class n is defined as $|n_{max}|/|n|$, where $|n|$ denotes the number of genomic sites in subclass n and $|n_{max}|$ denotes the number of genomic sites in the subclass with maximum sites.

Training the SeqUnwinder model

The w_n and w_p update steps separate out and are iteratively updated until convergence. The w_p update step has a simple closed form solution given by the equation:

$$w_p^k = median(c_p^k); \quad where \quad c_p^k = \{w_j^k \mid j \in C(p)\}$$

Where w_p^k is the kth term of the label-p weight vector. c_p^k is a set of the kth terms of the weight vectors of all the subclasses the label p spans.

The w_n update step is: -

$$w_n = \underset{w_n}{\operatorname{argmin}}\left[-\sum_{i=1}^{M}\sum_{n\in T}b_i y_{in}\log\left(\frac{\exp(w_n x_i)}{\sum_{n\in T}\exp(w_n x_i)}\right) + \lambda\sum_{n\in T}\sum_{p\in\Pi(n)}\|w_n - w_p\|_1\right]$$

The above equation is solved using the scaled alternating direction method of multipliers (ADMM) framework [12]. Briefly, the ADMM framework splits the above problem into 2 smaller sub-problems, which are much easier to solve. ADMM introduces an additional variable z_{np} initialized as follows

$$z_{np} = w_n - w_p;$$

w_n and z_{np} are iteratively estimated until convergence of the ADMM algorithm.

Sub-problem 1:

$$w_n^{t+1} = \underset{w_n}{\operatorname{argmin}}\left[-\sum_{i=1}^{M}\sum_{n\in T}b_i y_{in}\log\left(\frac{\exp(w_n x_i)}{\sum_{n\in T}\exp(w_n x_i)}\right) + \frac{\rho}{2}\sum_{n\in T}\sum_{p\in\Pi(n)}\|w_n - z_{np}^t - w_p + u_{np}^t\|_2^2\right]$$

Where u_{np} is the scaled dual variable. The above sub-problem is solved using the LBFGS (limited-memory Broyden Fletcher Goldfarb Shanno) algorithm [46].

Sub-problem 2:

$$z_{np}^{t+1} = \underset{z_{np}}{\operatorname{argmin}}\left[\lambda\|z_{np}\|_1 + \frac{\rho}{2}\|w_n^{t+1} - z_{np} - w_p + u_{np}^t\|_2^2\right]$$

The solution to the above equation is given by the shrinkage function defined as follows: -

$$z_{np}^{t+1} = \delta_{\frac{2\lambda}{\rho}}\left(w_n^{t+1} + z_{np}^t - w_p + u_{np}^t\right)$$

$$\delta_k(a) = \begin{cases} a - k, & if\ a > k \\ 0, & if\ \|a\| \le k \\ a + k, & if\ a < -k \end{cases}$$

The update step for the scaled dual variable u_{np} is: -

$$u_{np}^{t+1} = u_{np}^t + w_n^{t+1} - z_{np}^{t+1} - w_p$$

w_n^t, z_{np}^t, and u_{np}^t are iteratively estimated until convergence. The stopping criteria for the ADMM algorithm is:

$$\|\rho(z_{np} - z_{np}^{old})\|^2 < \epsilon^{abs} * K + \epsilon^{rel} * \|\rho * u_{np}\|^2$$

and

$$\|w_n - z_{np} - w_p\|^2 < \epsilon^{abs} * K + \epsilon^{rel} * \max(\|w_n\|^2, \|z_{np}\|^2, \|w_p\|^2)$$

Where ϵ^{abs} and ϵ^{rel} are the absolute and relative tolerance, respectively. Of note, to speed up the implementation of SeqUnwinder, a distributed version of ADMM was implemented. Intuitively, the w_n^{t+1} update step is distributed across multiple threads by spitting the M training examples into smaller subsets. The z_{np}^{t+1} and the u_{np}^{t+1} update steps act as pooling steps where the estimates of different threads are averaged. To further speed up convergence, a relaxed version of ADMM was implemented as described in [12]. In the relaxed version, w_n^{t+1} is replaced by $\alpha w_n^{t+1} + (1 - \alpha)z_{np}^t$ for the z_{np}^{t+1} and u_{np}^{t+1} update steps, where α is the over-relaxation parameter and is set to 1.9 as suggested in [12].

Converting weighted k-mer models into interpretable sequence features

While SeqUnwinder models label-specific sequence features using high-dimensional k-mer weight vectors, it is often desirable to visualize these sequence features in terms of a collection of interpretable position-specific scoring matrices. To do so, we use a combination of k-mer model scanning and local motif-finding in an approach similar to that used by SeqGL for producing interpretable motifs [5]. Specifically, we first scan the k-mer models learned during the training process across fixed-sized sequence windows around the input genomic loci to identify local high-scoring regions called "hills". Label-specific hills are focused regions ranging from 10 to 15bp and are composed of high scoring k-mers. We use a threshold of 0.1 to define hills. Next, we cluster the hills using K-means clustering with Euclidean distance metric and k-mer counts as features. To speed-up implementation, we restrict the unbiased k-mer features to only those k-mers that are present in at least 5% of the hills. We use silhouette index [47] to choose the appropriate value for K. Briefly, we test a range of K values from 2 to 6. For each value of K, we calculate the silhouette index on 30 bootstrap samples. The value of K with highest median silhouette index is chosen as the best value for K. Finally, any clusters with membership size (i.e. numbers of clustered hills) less than 10% of the largest cluster's membership size are merged with their next closest cluster.

MEME [48] is used to identify motifs in different clusters resulting in label-specific discriminative motifs. Each k-mer model further scores MEME-identified motifs as follows:

$$Score_{w_p}(motif_x) = \sum_{j \in motif_x} w_p^j$$

where $j \in motif_x$ is the set of all k-mers that belong to motif "$motif_x$". Note that the heatmaps in each figure which display these label-specific discriminative scores have been generated with a shared color scheme; i.e., the maximum shade of yellow is defined to correspond to a model-specific score of +0.4, while the maximum shade of blue is set to a score of -0.4. In each figure, individual motifs sometimes have scores outside of these bounds, but we chose to use a shared color scheme for consistency of interpretation across figures.

In our experience, the above "hill-finding" method provides a convenient way to convert high-dimensional k-mer models into interpretable position-specific scoring matrices, and is less error-prone than alternative k-mer clustering or assembly approaches. One advantage of the "hill-finding" approach is that it implicitly takes into account positional relationships between high-scoring k-mers on the genome–short stretches that contain multiple high-scoring k-mers will form larger "hills". Focused motif searches in the hills thus can find motifs that are longer than the longest k-mers in the underlying SeqUnwinder model.

Generation of synthetic datasets

To test SeqUnwinder in simulated settings, we generated various synthetic datasets. The sizes of simulated datasets (6,000–9,000 sequences) were chosen to roughly reflect the number of peaks in a typical ChIP-seq dataset. First, we generated 150bp sequences by sampling from a 2nd-order Markov model of the human genome. Our use of a 2nd-order Markov model is motivated by a desire to capture typical di- and tri-nucleotide compositional biases of vertebrate genomes (e.g. CG dinucleotide depletion and poly-A tracts). The exact choice of order of the background Markov model (i.e. 2nd-order versus a higher order) is arbitrary, but should not be expected to affect the relative performances of the methods in correctly associating embedded motifs with correct labels.

Next, we randomly assigned labels to the simulated sequences at different frequencies. The overlap between the labels at the sequences was varied from 0.5 to 0.99. Arbitrarily chosen TF binding motifs were assigned to labels. Each motif instance was sampled from the probability density function defined by the PWM of the motif. Sampled motif instances were inserted at labeled sites at a frequency of 0.7.

Processing iMN programming data-sets

Defining early, shared and late binding labels. MultiGPS was used to call Isl1/Lhx3 binding sites at 12 and 48hrs (datasets were obtained from GSE80321). A q-value cutoff <0.001 was used to call binding sites. All sites with significantly greater Isl1/Lhx3 ChIP enrichment at 12h compared to 48h (q-value cutoff of <0.01) were labeled as "early". Isl1/Lhx3 binding sites called in both 12 and 48h datasets with a further filter of not being differentially bound (q-value cutoff of <0.01), were assigned as "shared" sites. Finally, all sites with significantly greater Isl1/Lhx3 ChIP enrichment at 48h compared to 12h (q-value cutoff of <0.01) were labeled as "late".

Defining active and inactive mES annotation labels. A random forest classifier (see below for implementation details) was trained to classify every Isl1/Lhx3 binding site as either being in accessible/active or inaccessible/unmarked mouse ES chromatin. The classifier was trained using 95 mouse ES ChIP-seq datasets with windowed read-enrichment as predictors.

A union list of 1million 500bp regions comprising the enriched domains (see below) of DNaseI, H3K4me2, H3K4me1, H3K27ac, and H3K4me3 was used as the positive set for training the classifier. An equal number of unmarked 500bp regions were randomly selected and used as the negative set for training the classifier. Every binding site that was predicted to be in accessible/active ES chromatin with a probability of greater than 0.6 was placed in the "ES-active" class, while the remaining sites were placed in the "ES-inactive" class.

Enriched domains for DNaseI, H3K4me2, H3K4me1, H3K27ac, and H3K4me3 were identified using the DomainFinder module in SeqCode (https://github.com/seqcode/seqcode-core/blob/master/src/org/seqcode/projects/seed/DomainFinder.java). Contiguous 50bp genomic bins with significantly higher read enrichment compared to an input experiment were identified (binomial test, p-value < 0.01). Further, contiguous blocks within 200bp were stitched together to call enriched domains.

Weka's implementation of Random Forests was used to train the classifier (https://github.com/seqcode/seqcode-core/blob/master/src/org/seqcode/ml/classification/BaggedRandomForest.java). Briefly, the forest contained 10,000 trees. Each tree was trained with 10 randomly sampled features on 1% bootstrapped samples of the entire dataset.

Processing ENCODE datasets

TF ChIP-seq datasets: We analyzed 17 TF ChIP-seq ENCODE datasets in three primary cell-lines (GM12878, K562, and H1-hESC). The binding profiles for the factors were profiled using MultiGPS [15]. All called binding events for TFs were required to have significant enrichment over corresponding input samples (q-value <0.01) as assessed using MultiGPS' internal binomial test. For a site to be labeled as "shared", the binding site was required to be called in all the 3 cell-lines. Further, binding sites showing significantly differential binding in any of the possible 3 pair-wise comparisons were removed from the shared set. Binding sites labeled as cell-type specific sites were required to have significantly higher ChIP enrichment compared to other cell-lines. All TF binding sites within 5Kbp of a known TSS (defined using UCSC hg19 gene annotations) were labeled as "promoter proximal", while all sites that were more than 5Kbp from known TSSs were labeled as "distal".

DNase-seq datasets: We analyzed the DHS sites at 6 different tier 1 and 2 ENCODE cell-lines (GM12878, K562, H1-hESC, HeLa-S3, HepG2, HUVEC). The DHS sites were called using in-house scripts. Briefly, contiguous 50bp genomic bins with significantly higher read enrichment compared to an input experiment were identified (binomial test, p-value < 0.01). Further, contiguous blocks within 200bp were stitched together to call enriched domains. A 150bp window around the maximum point of read density at enriched domains was considered as the DHS.

Annotation of de novo identified motifs

All *de novo* motifs identified using SeqUnwinder were annotated using the cis-bp database. Briefly, *de novo* motifs were matched against the cis-bp database using STAMP [49]. The best matching hit with a p-value of less than 10e-6 was used to name the *de novo* identified motifs.

Availability and reproducibility

SeqUnwinder is freely available under the MIT open source license from: https://github.com/seqcode/sequnwinder. Complete output files produced by the SeqUnwinder runs described in this work, along with scripts and data for reproducing all analysis figures, are available from: https://github.com/ikaka89/sequnwinderPaper.

Supporting information

S1 Fig. Classification scores associating each *de novo* discovered motif with each Isl1/Lhx3 binding event label. SeqUnwinder and MCC model scores of *de novo* discovered Zfp281-like, Oct4-like, Lhx3-like and Onecut-like motifs in various Isl1/Lhx3 site categories. Positive scores denote that the motif is positively discriminative of the given label (relative to the other labels), and vice versa for negative scores.
(TIF)

S2 Fig. Motifs discovered by DREME for each Isl1/Lhx3 binding event label. This figure summarizes whether DREME finds Zfp281-like, Oct4-like, Lhx3-like and Onecut-like motifs in analyses that aim to discriminate a given label's sites against sites from all other labels. Displayed motifs match the relevant TF's motif in the cis-bp database according to STAMP analysis (values listed under the motifs correspond to STAMP E-values).
(TIF)

S3 Fig. Distributions of distances between ENCODE TF binding events and annotated TSS. The distributions show the distances of TF binding events from annotated mRNA TSS for all 17 examined ENCODE TFs, stratified based on "shared" (black) or "cell line-specific" labels (purple = K562, blue = GM12878, green = H1-hESC). The X-axis represents the distance in base pairs plotted according to a log-scale (natural logarithm).
(TIF)

S4 Fig. Distributions of cognate motif instance scores at labeled ENCODE TF binding events. Log-odds score distributions summarizing highest-scoring instances of SeqUnwinder's *de novo* discovered cognate motifs at "shared", "K562", "GM12878" and "H1-hESC" labeled binding events.
(TIF)

S5 Fig. Cognate motifs discovered by DREME at labeled ENCODE TF binding events. This figure summarizes whether DREME finds cognate motifs in analyses that aim to discriminate a given label's sites against sites from all other labels. Displayed motifs match the relevant cognate motifs in the cis-bp database according to STAMP analysis (values listed under the motifs correspond to STAMP E-values).
(TIF)

S1 Table. Performance of SeqUnwinder in classifying each Isl1/Lhx3 binding site subclass. Area under receiver operating characteristic curve (auROC) values describing the classification performance of SeqUnwinder for each Isl1/Lhx3 subclass. Classification performance is determined using 3-fold cross-validation.
(DOCX)

S2 Table. Performance of SeqUnwinder in classifying subclasses of binding sites for seventeen ENCODE TFs. Area under receiver operating characteristic curve (auROC) values describing the classification performance of SeqUnwinder for each subclass of binding sites. Classification performance is determined using 3-fold cross-validation.
(DOCX)

S3 Table. Performance of SeqUnwinder model in classifying subclasses of DHS sites in 6 different ENCODE cell-lines. Area under receiver operating characteristic curve (auROC) values describing the classification performance of SeqUnwinder for each subclass of DHS sites. Classification performance is determined using 3-fold cross-validation.
(DOCX)

Acknowledgments

Author Contributions

References

1. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011; 27: 1653–1659. https://doi.org/10.1093/bioinformatics/btr261 PMID: 21543442

2. Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci. 1995; 4: 1618–1632. https://doi.org/10.1002/pro.5560040820 PMID: 8520488

3. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 2011; 21: 2167–2180. https://doi.org/10.1101/gr.121905.111 PMID: 21875935

4. Arvey A, Agius P, Noble WS, Leslie C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome Res. 2012; 22: 1723–1734. https://doi.org/10.1101/gr.127712.111 PMID: 22955984

5. Setty M, Leslie CS. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. PLoS Comput Biol. 2015; 11: e1004271. https://doi.org/10.1371/journal.pcbi.1004271 PMID: 26016777

6. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015; 33: 831–838. https://doi.org/10.1038/nbt.3300 PMID: 26213851

7. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012; 9: 215–216. https://doi.org/10.1038/nmeth.1906 PMID: 22373907

8. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012; 9: 473–476. https://doi.org/10.1038/nmeth.1937 PMID: 22426492

9. Zhang Y, An L, Yue F, Hardison RC. Jointly characterizing epigenetic dynamics across multiple human cell types. Nucleic Acids Res. 2016; 44: 6721–6731. https://doi.org/10.1093/nar/gkw278 PMID: 27095202

10. Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell. 2004; 117: 185–198. PMID: 15084257

11. Elemento O, Slonim N, Tavazoie S. A universal framework for regulatory element discovery across all genomes and data types. Mol Cell. 2007; 28: 337–350. https://doi.org/10.1016/j.molcel.2007.09.027 PMID: 17964271

12. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Found Trends Mach Learn. 2011; 3: 1–122. https://doi.org/10.1561/2200000016

13. Mazzoni EO, Mahony S, Closser M, Morrison CA, Nedelec S, Williams DJ, et al. Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. Nat Neurosci. 2013; 16: 1219–1227. https://doi.org/10.1038/nn.3467 PMID: 23872598

14. Velasco S, Ibrahim MM, Kakumanu A, Garipler G, Aydin B, Al-Sayegh MA, et al. A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells. Cell Stem Cell. 2017; 20: 205–217.e8. https://doi.org/10.1016/j.stem.2016.11.006 PMID: 27939218

15. Mahony S, Edwards MD, Mazzoni EO, Sherwood RI, Kakumanu A, Morrison CA, et al. An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding. PLoS Comput Biol. 2014; 10: e1003501. https://doi.org/10.1371/journal.pcbi.1003501 PMID: 24675637

16. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38: 576–589. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432

17. Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, et al. Distinct properties of cell-type-specific and shared transcription factor binding sites. Mol Cell. 2013; 52: 25–36. https://doi.org/10.1016/j.molcel.2013.08.037 PMID: 24076218

18. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489: 57–74. https://doi.org/10.1038/nature11247 PMID: 22955616

19. Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, et al. A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell. 2012; 148: 473–486. https://doi.org/10.1016/j.cell.2012.01.030 PMID: 22304916

20. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. 2012; 13: R48. https://doi.org/10.1186/gb-2012-13-9-r48 PMID: 22950945

21. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012; 488: 116–120. https://doi.org/10.1038/nature11243 PMID: 22763441

22. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012; 22: 1798–1812. https://doi.org/10.1101/gr.139105.112 PMID: 22955990

23. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2014; 42: 2976–2987. https://doi.org/10.1093/nar/gkt1249 PMID: 24335146

24. Lu R, Mucaki EJ, Rogan PK. Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. Nucleic Acids Res. 2017; 45: e27–e27. https://doi.org/10.1093/nar/gkw1036 PMID: 27899659

25. Cuddapah S, Jothi R, Schones DE, Roh T-Y, Cui K, Zhao K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. 2009; 19: 24–32. https://doi.org/10.1101/gr.082800.108 PMID: 19056695

26. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 2007; 128: 1231–1245. https://doi.org/10.1016/j.cell.2006.12.048 PMID: 17382889

27. Pevny L, Lin CS, D'Agati V, Simon MC, Orkin SH, Costantini F. Development of hematopoietic cells lacking transcription factor GATA-1. Development. 1995; 121: 163–172. PMID: 7867497

28. Welch JJ, Watts JA, Vakoc CR, Yao Y, Wang H, Hardison RC, et al. Global regulation of erythroid gene expression by transcription factor GATA-1. Blood. 2004; 104: 3136–3147. https://doi.org/10.1182/blood-2004-04-1603 PMID: 15297311

29. Han GC, Vinayachandran V, Bataille AR, Park B, Chan-Salis KY, Keller CA, et al. Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution. Mol Cell Biol. 2016; 36: 157–172. https://doi.org/10.1128/MCB.00806-15 PMID: 26503782

30. Duncan SA, Navas MA, Dufort D, Rossant J, Stoffel M. Regulation of a transcription factor network required for differentiation and metabolism. Science. 1998; 281: 692–695. PMID: 9685261

31. Friedman JR, Kaestner KH. The Foxa family of transcription factors in development and metabolism. Cell Mol Life Sci. 2006; 63: 2317–2328. https://doi.org/10.1007/s00018-006-6095-6 PMID: 16909212

32. DeLaForest A, Nagaoka M, Si-Tayeb K, Noto FK, Konopka G, Battle MA, et al. HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells. Development. 2011; 138: 4143–4153. https://doi.org/10.1242/dev.062547 PMID: 21852396

33. Alder O, Cullum R, Lee S, Kan AC, Wei W, Yi Y, et al. Hippo Signaling Influences HNF4A and FOXA2 Enhancer Switching during Hepatocyte Differentiation. Cell Rep. 2014; 9: 261–271. https://doi.org/10.1016/j.celrep.2014.08.046 PMID: 25263553

34. Morita R, Suzuki M, Kasahara H, Shimizu N, Shichita T, Sekiya T, et al. ETS transcription factor ETV2 directly converts human fibroblasts into functional endothelial cells. Proc Natl Acad Sci. 2015; 112: 160–165. https://doi.org/10.1073/pnas.1413234112 PMID: 25540418

35. Leslie C, Kuang R. Fast String Kernels Using Inexact Matching for Protein Sequences. J Mach Learn Res. 2004; 5: 1435–1455.

36. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput Biol. 2014; 10: e1003711. https://doi.org/10.1371/journal.pcbi.1003711 PMID: 25033408

37. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, et al. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci U S A. 2015; 112: 4654–4659. https://doi.org/10.1073/pnas.1422023112 PMID: 25775564

38. Chiu T-P, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics. 2016; 32: 1211–1213. https://doi.org/10.1093/bioinformatics/btv735 PMID: 26668005

39. Mathelier A, Xin B, Chiu T-P, Yang L, Rohs R, Wasserman WW. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. Cell Syst. 2016; 3: 278–286.e4. https://doi.org/10.1016/j.cels.2016.07.001 PMID: 27546793

40. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011; 21: 447–455. https://doi.org/10.1101/gr.112623.110 PMID: 21106904

41. Kähärä J, Lähdesmäki H. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. Bioinformatics. 2015; 31: 2852–2859. https://doi.org/10.1093/bioinformatics/btv294 PMID: 25957350

42. Kang F, Jin R, Sukthankar R. Correlated Label Propagation with Application to Multi-label Learning. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Volume 2. IEEE; 2006. pp. 1719–1726. https://doi.org/10.1109/CVPR.2006.90

43. Sohn K-A, Kim S. Joint Estimation of Structured Sparsity and Output Structure in Multiple-Output Regression via Inverse-Covariance Regularization. Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics. 2012. pp. 1081–1089. Available: http://proceedings.mlr.press/v22/sohn12.html

44. Ma W, Yang L, Rohs R, Noble WS. DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. Bioinformatics. 2017; https://doi.org/10.1093/bioinformatics/btx336 PMID: 28541376

45. Gopal S, Yang Y, Niculescu-mizil A. Regularization Framework for Large Scale Hierarchical Classification. Proc ACM SIGKDD. 2013. pp. 257–265.

46. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program. 1989; 45: 503–528. https://doi.org/10.1007/BF01589116

47. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987; 20: 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

48. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994; 2: 28–36. PMID: 7584402

49. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 2007; 35: W253–W258. https://doi.org/10.1093/nar/gkm272 PMID: 17478497