Characterizing regulatory sequence features that discriminate between overlapping annotation labels

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

[1]Center for Eukaryotic Gene Regulation, Department of Biochemistry & Molecular Biology, Penn State University, Pennsylvania, USA

[2]Department of Biology, New York University, 100 Washington Square East, New York, NY 10003, USA

* To whom correspondence should be addressed: mahony@psu.edu

Running Title: SeqUnwinder

Keywords: motif discovery, discriminative features, transcription factor binding sites, DNaseI hypersensitive sites

Abstract

Genomic loci with regulatory potential can be identified and annotated with various labels. For example, sites may be annotated as being bound or unbound by a transcription factor (TF) under particular cellular conditions, or as being proximal or distal to known transcription start sites. Given such a collection of labeled genomic 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 uneven overlaps between annotation labels. 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, we show SeqUnwinder's ability 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 demonstrate that multi-condition TF binding sites are typically characterized by better quality instances of the TF's cognate binding motifs. 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.

Availability: https://github.com/seqcode/sequnwinder

Introduction

Regulatory genomics analyses often focus on finding sequence features associated with genomic sites that share some property or annotation "label". Such problems are typically phrased in terms of a two-class classification. For example, we may wish to find sequence features that discriminate between sites bound by a particular transcription factor (TF) and unbound sites, or between sites that are associated with gene activation and repression. With the increased availability of genome-wide epigenomic datasets such as those generated by the ENCODE and ROADMAP projects (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium *et al*, 2015), it is now possible to provide a more detailed annotation of regulatory sites beyond binary labels such as "bound" and "unbound". For example, a TF's binding sites maybe sub-categorized according to which cell types or conditions it is bound in, or according to whether those sites display coincident ChIP-enrichment of other proteins or histone modifications. Genome segmentation methods (Ernst & Kellis, 2012; Hoffman *et al*, 2012; Zhang *et al*, 2016) provide an automated annotation of promoters, enhancers, and various other chromatin states that can also be overlaid on a TF's binding sites. As regulatory region annotations become more complex, there is a growing need for computational methods that can find sequence features specific to each of several regulatory region subtypes.

Several classification methods have been used to characterize discriminative sequence features in two-class scenarios (Bailey, 2011; Alipanahi *et al*, 2015), including support vector machines (SVM) with various *k*-mer based sequence kernels (Arvey *et al*, 2012; Ghandi *et al*, 2014; Lee *et al*, 2011) and convolutional

neural networks (Alipanahi *et al*, 2015). Some current methods also allow a limited analysis of datasets where annotation labels partially overlap. For example, (Arvey *et al*, 2012) used a multi-task SVM classifier to learn cell-type specific and shared binding preferences of TF binding sites in two cell-types. SeqGL (Setty & Leslie, 2015), a group lasso based logistic regression classifier, also implements a similar multi-task framework. However, these approaches were designed for essentially two-class classification problems where the multi-task framework enables modeling of the "common" task in addition to the two classes.

No existing discriminative feature discovery methods expand beyond two-class problems to scenarios where a set of genomic sites contains multiple annotation labels with arbitrary rates of overlap between them. In cases where all annotation labels are mutually exclusive, the problem of determining label-specific sequence features could be straightforwardly cast as a multi-class classification problem. In more general and realistic scenarios, however, we may wish to find sequence features associated with several types of annotation labels, each of which may overlap and possibly confound the others.

To gain insight into the problems faced by analysis methods in the general multi-label problem, consider the hypothetical scenario presented in Figure 1a, where a given TF's binding sites have been labeled as being bound in cell types A, B, or C. The sites are further characterized as being proximal or distal to TSSs (Pr and Di, respectively), where the latter labels unevenly overlap the cell type labels (e.g. let's suppose that cell type A's sites are more likely to be promoter proximal than sites in other cell types). In such a scenario, simple multi-class classification frameworks

will be limited to either ignoring certain labels or characterizing sequence features associated with each observed subclass (i.e. a particular combination of labels). In either case, it may not be possible to unambiguously assign discovered sequence features to specific annotation labels; continuing the Figure 1a example, features discovered to be enriched in cell type A sites may actually be due to the distinct properties of promoter proximal sites. We therefore require a structured classification framework that can deconvolve sequence features associated with overlapping annotation labels.

In this work, we present SeqUnwinder, a novel classification framework for characterizing interpretable sequence features associated with overlapping sets of genomic annotation labels. SeqUnwinder begins by defining genomic site subclasses based on the combinations of labels annotated at these sites (Figure 1b). The site subclasses are treated as distinct classes for a multiclass logistic regression model that uses *k*-mer frequencies in a fixed window around sites as predictors. However, SeqUnwinder also models each individual label's specific features by incorporating them in an L1 regularization term (see Methods). Regularization encourages consistent features to be shared across subclasses that are spanned by a label, thus implicitly enabling label-specific features to be learned (Figure 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 (Figure 1c). Finally, to aid

interpretability, SeqUnwinder identifies over-represented motifs in the hills and scores them using label- and subclass-specific *k*-mer models (Figure 1d).

We demonstrate the unique abilities of SeqUnwinder using both synthetic sequence datasets and collections of real TF ChIP-seq and DNase-seq experiments. In the real datasets, we begin with a motivating example that analyzes transcription factor binding during the programming of embryonic stem (ES) cells into induced motor neurons (iMNs) (Mazzoni *et al*, 2013; Velasco *et al*, 2016). In this example, we categorize the TF binding sites according to dynamic binding behaviors observed during the programming process. These binding site categories are further (unevenly) split into subclasses according to whether they are in an accessible/active chromatin state in the initial ES cells. We demonstrate that SeqUnwinder can deconvolve sequence features associated with binding dynamics from those associated with initial chromatin state, thereby providing testable hypotheses about the binding mechanisms driving each annotation label.

In two further examples using real epigenomic datasets, we characterize sequence features associated with genomic locations that display regulatory properties across multiple cell types. Sites that are bound by a particular TF in multiple cell types (i.e. "shared" or multi-condition sites) are often strongly biased towards being located in gene promoter regions, in contrast to cell-specific binding sites, which are typically distally located. After controlling for such biases by incorporating labels that annotate proximal and distal sites, SeqUnwinder discovers that shared TF binding sites are characterized by stronger instances of the cognate binding motif than cell-specific sites. In our final example, we demonstrate that

SeqUnwinder scales very well to analyses of over one hundred thousand sites annotated with dozens of label combinations. To show this, we characterize the sequence features at shared and cell-type specific DNase I hypersensitive sites in six different ENCODE cell lines. Interestingly, we find that motifs enriched in cell-type specific DNase I hypersensitive sites are also highly enriched at cell-type specific TF binding sites for a majority of the examined TFs.

Results

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 (Figure 2a). We assigned a motif to each label and inserted sequences by sampling from the distributions defined by the position-specific scoring matrices of label assigned motifs (Figure 2a). When run on this collection of sequences and given knowledge of the label assignments, SeqUnwinder correctly identifies motifs similar to all inserted motifs (Figure 2b). SeqUnwinder also correctly assigns each motif to its respective annotation label with high weight. Further, the label-specific scores of the identified motifs 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 over other methods at different levels of label overlaps. We simulated 100 datasets with 6000 simulated sequences and varying the degree of overlap between two sets of labels: A, B and X,

Y from 50% to 99% (Figure 2*c*). We then compared SeqUnwinder with a simple multi-class classification approach (MCC) where each label was treated in isolation. In MCC training, therefore, each regulatory site is included in two separate training sets in accordance with its annotated labels. We also compared SeqUnwinder with DREME (Bailey, 2011), 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 (simulated) label assignments (Figure 1e and 1f). We used these measures to calculate the F1 score (harmonic mean of precision and recall) at different overlapping levels (Figure 2d).

Figure 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. However, 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) (Figure 2e and 2f)

Taken together, the synthetic data experiments demonstrate that SeqUnwinder provides the ability to discover sequence features associated with overlapping sets of genomic site labels.

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 ES cells efficiently converts the resident ES cells into functional spinal motor neurons (Mazzoni *et al*, 2013; Velasco *et al*, 2016). 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 (Velasco *et al*, 2016). We found that two of the ectopically expressed TFs, Isl1 & Lhx3, bind together at the vast majority of their targets during the programming process. We also found that this cooperative pair of TFs shifted their binding targets during programming. We therefore 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 48h programming process, and late sites were only bound during the final stage of programming.

In our previous work, we used standard *de novo* motif finders to characterize sequence features associated with each of the three dynamic binding categories (Velasco *et al*, 2016). We discovered motifs similar to the binding preferences of Oct4 and Zfp281 at early sites, while Onecut TF family motifs were enriched at late sites. However, it is possible that these sequence features are not specifically associated with binding dynamics, but rather reflect on coincident properties of the underlying genomic sites. For example, Oct4 and Zfp281 are both known regulators of pluripotency (Nichols *et al*, 1998; Pesce & Schöler, 2000; Wang *et al*, 2008; Fidalgo *et al*, 2011); the presence of related motifs at early Isl1/Lhx3 sites may merely be a secondary effect of a strong overlap between early-bound sites and regulatory regions that are active in ES cells.

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 for more details). Annotating Isl1/Lhx3 sites using both sets of labels (Isl1/Lhx3 binding dynamics and ES activity) results in six different subclasses (Figure 3a). The labels annotating binding dynamics at Isl1/Lhx3 sites overlap to varying degrees with the pre-existing activity status of these sites in ES cells. As can be seen from Figure 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 were more likely to be inactive in ES cells.

Using SeqUnwinder, we deconvolved several motif features associated with the various Isl1/Lhx3 binding site labels (Figure 3b). SeqUnwinder discovers motifs

similar to those bound by Oct4 and Zfp281. As suspected, SeqUnwinder finds an association between the Zfp281 motif and the ES-active label, while the Zfp281 motif is not associated with the early Isl1/Lhx3 binding label. Surprisingly, the Oct4 motif was highly associated with the early binding label, suggesting that Isl1/Lhx3 cooperates or competes with Oct4 binding at the early binding targets. To further test the association between Oct4 sites and early Isl1/Lhx3 binding activity, we profiled the binding of Oct4 in ES cells and at 12 hours after NIL induction. As shown in Figure 3c, Oct4 shows a preferential enrichment at early Isl1/Lhx3 sites, in line with SeqUnwinder's prediction. Therefore, by carefully labeling the sites with multiple sets of relevant annotations and using SeqUnwinder, we can assign Oct4 as a feature of early Isl1/Lhx3 binding sites and Zpf281 as a feature of ES-active sites.

SeqUnwinder also identifies a motif similar to that bound by the Onecut TF family as being highly associated with late binding but not with ES-inactive sites. As previously described in our earlier work, we characterized Onecut2 binding to be highly enriched at late Isl1/Lhx3 sites during iMN programming (Velasco *et al*, 2016) (Figure 3c). 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 (Velasco *et al*, 2016), supporting a model in which late Isl1/Lhx3 binding is dependent on Onecut TFs.

Our analysis of Isl1/Lhx3 binding during iMN programming serves as an example analysis scenario in which we are trying to find motif features associated with multiple overlapping annotation labels. As demonstrated, SeqUnwinder identifies motif features associated with the various labels, which can lead to

testable hypotheses about co-factors that serve mechanistic roles at subsets of binding sites. Interestingly, the motif features that are most highly associated with shared binding sites all correspond to homeobox motifs of the type bound by Isl1/Lhx3. 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.

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

To demonstrate the general applicability of SeqUnwinder to a broader range of TFs, we set out to characterize the sequence properties of sites that are bound by particular TFs across multiple conditions. The sequence properties of tissue-specific TF binding sites have been extensively studied (Heinz *et al*, 2010; Arvey *et al*, 2012; Setty & Leslie, 2015). 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. In other words, 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 (Gertz *et al*, 2013; Mahony *et al*, 2014).

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 16 sequence-specific TFs profiled in all 3 primary ENCODE cell-lines (K562, GM12878, and H1-hESC). Using MultiGPS (Mahony *et al*, 2014) on each TF's multi-condition dataset, we carefully curated sets of tissue-specific sites for 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 (Figure S1). 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. In summary, each examined TF's binding sites is categorized into 8 subclasses, each of which is composed of combinations of 6 distinct labels (Figure 4a).

We applied SeqUnwinder to each labeled sequence collection in order to characterize label-specific sequence features. We provide a detailed explanation for NRSF (REST), one of the examined TFs. Using MultiGPS, we identified a total of ~14,000 stringent binding events in NRSF ChIP-Seq datasets, which we categorized into the aforementioned subclasses (Figure S2a). Running SeqUnwinder on this collection of NRSF binding events, we identified several *de novo* motifs (Figure S2b). Interestingly, a *de novo* motif matching to the cognate NRSF motif had a high shared

(multi-condition) label-specific score. 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 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 for TEAD-like motif and K562-specific sites were enriched for GATA-like motif. In fact, GATA2 ChIP-Seq reads in K562 showed a striking enrichment at K562-specific NRSF binding sites (Figure S2a).

Similar results were observed for many of the examined factors. SeqUnwinder discovers motifs that match the TF's known cognate binding preference in 13 of 16 datasets. These cognate motifs are found to be highly associated with shared (multi-condition) sites for 11 of the 16 examined TFs (Figure 4b). Despite significant overlaps between shared sites and promoter proximal sites (Figure S1), the cognate motifs were not found to be predictive of proximal sites (Figure 4b). Further, the primary motif was not specifically predictive of cell-type specific binding sites for any of the examined TFs.

Next we assessed if high SeqUnwinder scores at multi-conditionally bound sites correspond to better quality and higher frequency motif instances. To do this we scanned the *de novo* identified cognate motifs and calculated peak-rate (fraction of peaks that have significant motif match) and hit-rate (total number of significant matches normalized by number of peaks) at all labels. As shown in Figures S3a and S3b, higher SeqUnwinder scores generally translate to high motif hit and peak rates

with few exceptions. However, for TFs USF1/2 and MAX, the motif peak-rates were less distinguishable across labels (Figure S3a) while the SeqUnwinder scores of the primary motif clearly indicate their high predictive power of the shared sites (Figure 4b). Specifically, for USF1, the *de novo* cognate motif hit-rates at shared and H1-hESC specific sites were both high at 0.54 and 0.41, respectively (Figure 4c). However, the fraction of hits in H1-hESC that had a central "CG" di-nucleotide was much less than that at shared sites (Figure 4c). Thus due to the position independent nature of PWMs, the absence of the preferred "CG" di-nucleotide of the *de novo* motif was not translated completely in the calculation of the motif hit and peak rates. By taking into account these higher order dependencies using a *k*-mer model, SeqUnwinder identifies the "CG"-containing cognate motif as being highly predictive of shared sites.

We also examined co-factor motifs associated with cell-type specific binding labels. Interestingly, we found IRF and RUNX motifs enriched at GM12878-specific binding sites for 3 and 9 of the 16 examined TFs, respectively. Similarly, the GATA motif was predictive of K562-specific binding for 13 out of the 16 examined TFs. A TEAD-like motif was predictive of H1-hESC specific sites for 11 of the 16 TFs (Figure 4d).

In summary, our analyses demonstrate the applicability of SeqUnwinder to the increasingly common problem of characterizing sequence features associated with cell-specific and cell-invariant TF binding. Uniquely, SeqUnwinder can implicitly account for location-dependent sequence composition biases by incorporating knowledge of extra layers of annotations. Our results further support the model that

high affinity cognate motif instances are a striking feature of multi-conditionally bound sites across a broad range of TFs.

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. To do this, we annotated a large collection of DNase I hypersensitive (DHS) sites with six cell-line labels depending on the enrichment of DNase-seq reads (Figure 5a). If we 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 (Shen *et al*, 2012). 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 (Figure 5a).

SeqUnwinder identifies several interesting motifs in this large collection of DHS sites, some of which were previously associated with specific cell-types (Figure 5b). For example, different parts of the CTCF motif were highly predictive of shared DHS sites. This result is consistent with previous finding suggesting largely invariant CTCF binding across cellular contexts (Cuddapah *et al*, 2009; Kim *et al*, 2007). RUNX, IRF and NF-κB motifs were enriched at GM12878 specific DHS sites. These motifs

were also discovered by SeqGL at GM12878 specific DHS sites (Setty & Leslie, 2015). GATA motifs, key regulators of Erythroid development (Han *et al*, 2016), were enriched at K562 specific DHS sites. SNAI and TEAD motifs were enriched at H1-hESC sites. TEAD motifs were previously shown to be enriched at ES specific DHS sites (Setty & Leslie, 2015), while SNAI class TFs are key regulators of epithelial to mesenchymal transition (Mistry *et al*, 2014). JUND and FOS motifs were enriched at HeLa-S3-specific DHS sites. HNF4A and various FOX motifs, which are known master regulator of hepatocytes (Alder *et al*, 2014; DeLaForest *et al*, 2011), were enriched at HepG2 specific DHS sites. Finally, motifs belonging to the ETS class of TFs were enriched at HUVEC specific DHS sites (Figure 5b). ETS factors have been shown to directly convert human fibroblasts to endothelial cells (Morita *et al*, 2015). 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 (Figure 4d). For example, IRF, GATA, and TEAD motifs associated with GM12878, K562, and H1-hESC specific DHSs were also predictive of cell-type specific binding for a majority 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*, 2011), 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 (Leslie & Kuang, 2004), di-mismatch kernel (Arvey *et al*, 2012), wild-card kernel (Leslie & Kuang, 2004; Setty & Leslie, 2015), and gkm-kernel (Ghandi *et al*, 2014)) 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 (Zhou *et al*, 2015; Chiu *et al*, 2016; Mathelier *et al*, 2016). More recently, deep learning models have also been harnessed to predict TF binding sites from unbound sites (Alipanahi *et al*, 2015).

In this manuscript, 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 two 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. We have shown that SeqUnwinder provides a unique ability to deconvolve

discriminative sequence features at overlapping sets of labels. SeqUnwinder leverages overlaps between labels and identifies features that are consistently shared across subclasses spanned by a label. SeqUnwinder is easy to use and takes as input a list of genomic coordinates and corresponding annotations and identifies interpretable sequence features that are enriched at a given label or at combinations of labels (subclasses). SeqUnwinder implements a multi-threaded version of the ADMM (Boyd *et al*, 2011) framework to train the model and typically runs in less than few hours for most datasets. SeqUnwinder can also be easily extended to incorporate different kinds of kernels and shape features.

We demonstrated the unique analysis abilities of SeqUnwinder using three analysis scenarios based on real ChIP-seq and DNase-seq datasets. 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 (Velasco *et al*, 2016). 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 (Velasco *et al*, 2016).

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. For example, Gertz *et al.*, showed that ER (estrogen receptor) binding sites bound in both ECC1 and T4D7, two human cancer cell lines, had high 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 (Figure 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.

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 (Pique-Regi *et al*, 2011; Kähärä & Lähdesmäki, 2015; Mathelier *et al*, 2016).

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 multiclass logistic regression classifier trained on subclasses of genomic sites. The predictive features for the classifier are *k*-mer frequencies in a fixed window around input loci, with *k* usually ranging from 4 to 6. 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. The overall objective function of SeqUnwinder is: -

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

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 subclass 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:

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

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 (Boyd *et al*, 2011). 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:

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

Where u_{np} is the scaled dual variable. The above sub-problem is solved using the LBFGS (limited-memory Broyden Fletcher Goldfarb Shanno) algorithm (Liu & Nocedal, 1989).

Sub-problem 2:

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

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

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

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

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:

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

and

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

Where $\epsilon^{abs}$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 (Boyd *et al*, 2011). 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 (Boyd *et al*, 2011).

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 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. These label-specific high scoring regions are called "hills". Label-specific hills are potentially enriched for sequence signals (e.g. TF binding motifs) that discriminate them from other genomic loci. Typically the hills are around 10-15bp in width. The hills are clustered based on their *k*-mer composition using K-means clustering. Where K is a user-defined input and is equivalent to the number of motifs that one expects at labeled genomic sites. MEME (Bailey & Elkan, 1994) 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 "$motif_x$".

Generation of synthetic datasets

To test SeqUnwinder in simulated settings, we generated various synthetic datasets. First, we generated 150bp long genomic sites by sampling sequences from a 2nd order Markov model of the human genome. We then randomly assigned labels to these binding sites at different frequencies. The overlap between the labels at the binding sites was varied from 0.5 to 0.99. Arbitrarily chosen TF binding motifs were

assigned to labels. A 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 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). 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

Processing ENCODE datasets

TF ChIP-seq datasets: We analyzed 16 TF ChIP-Seq ENCODE datasets in three primary cell-lines (GM12878, K562, and H1-hESC). The binding profiles for the factors were profiled using the MultiGPS software (Mahony *et al*, 2014). 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-specific" were required to be called in only one cell-type. In addition, cell-specific sites were also 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 the STAMP software (Mahony & Benos, 2007). The best matching hit with a p-value of less than 10e-5 was used to name the *de novo* identified motifs.

Acknowledgments

This work was supported by R01HD079682 NICHD (to EOM). The authors thank Dr. Frank Pugh, Dr. Ross Hardison, and members of the Center for Eukaryotic Gene Regulation at Penn State for helpful discussions.

Author Contributions

AK and SM conceived the study. AK designed and implemented the SeqUnwinder method and performed all analyses. SV performed iMN ChIP-seq experiments. SM and EOM supervised the work. AK and SM wrote the manuscript.

References:

Alder O, Cullum R, Lee S, Kan AC, Wei W, Yi Y, Garside VC, Bilenky M, Griffith M, Morrissy AS, Robertson GA, Thiessen N, Zhao Y, Chen Q, Pan D, Jones SJM, Marra MA & Hoodless PA (2014) Hippo Signaling Influences HNF4A and FOXA2 Enhancer Switching during Hepatocyte Differentiation. *Cell Rep.* **9:** 261–271

Alipanahi B, Delong A, Weirauch MT & Frey BJ (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33:** 831–838

Arvey A, Agius P, Noble WS & Leslie C (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.* **22:** 1723–1734

Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27:** 1653–1659

Bailey TL & Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2:** 28–36

Boyd S, Parikh N, Chu E, Peleato B & Eckstein J (2011) Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found Trends Mach Learn* **3:** 1–122

Chiu T-P, Comoglio F, Zhou T, Yang L, Paro R & Rohs R (2016) DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* **32:** 1211–1213

Cuddapah S, Jothi R, Schones DE, Roh T-Y, Cui K & Zhao K (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19:** 24–32

DeLaForest A, Nagaoka M, Si-Tayeb K, Noto FK, Konopka G, Battle MA & Duncan SA (2011) HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells. *Dev. Camb. Engl.* **138:** 4143–4153

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74

Ernst J & Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9:** 215–216

Fidalgo M, Shekar PC, Ang Y-S, Fujiwara Y, Orkin SH & Wang J (2011) Zfp281 functions as a transcriptional repressor for pluripotency of mouse embryonic stem cells. *Stem Cells* **29:** 1705–1716

Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, Cooper GM, Reddy TE, Crawford GE & Myers RM (2013) Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol. Cell* **52:** 25–36

Ghandi M, Lee D, Mohammad-Noori M & Beer MA (2014) Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10:** e1003711

Han GC, Vinayachandran V, Bataille AR, Park B, Chan-Salis KY, Keller CA, Long M, Mahony S, Hardison RC & Pugh BF (2016) Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution. *Mol. Cell. Biol.* **36:** 157–172

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H & Glass CK (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38:** 576–589

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA & Noble WS (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9:** 473–476

Kähärä J & Lähdesmäki H (2015) BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. *Bioinforma. Oxf. Engl.* **31:** 2852–2859

Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV & Ren B (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128:** 1231–1245

Lee D, Karchin R & Beer MA (2011) Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* **21:** 2167–2180

Leslie C & Kuang R (2004) Fast String Kernels Using Inexact Matching for Protein Sequences. *J Mach Learn Res* **5:** 1435–1455

Liu DC & Nocedal J (1989) On the limited memory BFGS method for large scale optimization. *Math. Program.* **45:** 503–528

Mahony S & Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* **35:** W253–W258

Mahony S, Edwards MD, Mazzoni EO, Sherwood RI, Kakumanu A, Morrison CA, Wichterle H & Gifford DK (2014) An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding. *PLoS Comput. Biol.* **10:** e1003501

Mathelier A, Xin B, Chiu T-P, Yang L, Rohs R & Wasserman WW (2016) DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst.* **3:** 278–286.e4

Mazzoni EO, Mahony S, Closser M, Morrison CA, Nedelec S, Williams DJ, An D, Gifford DK & Wichterle H (2013) Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. *Nat. Neurosci.* **16:** 1219–1227

Mistry DS, Chen Y, Wang Y & Sen GL (2014) SNAI2 Controls the Undifferentiated State of Human Epidermal Progenitor Cells. *Stem Cells* **32:** 3209–3218

Morita R, Suzuki M, Kasahara H, Shimizu N, Shichita T, Sekiya T, Kimura A, Sasaki K, Yasukawa H & Yoshimura A (2015) ETS transcription factor ETV2 directly converts human fibroblasts into functional endothelial cells. *Proc. Natl. Acad. Sci.* **112:** 160–165

Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, Schöler H & Smith A (1998) Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95:** 379–391

Pesce M & Schöler HR (2000) Oct-4: control of totipotency and germline determination. *Mol. Reprod. Dev.* **55:** 452–457

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y & Pritchard JK (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21:** 447–455

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, et al (2015) Integrative analysis of 111 reference human epigenomes. *Nature* **518:** 317–330

Setty M & Leslie CS (2015) SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLoS Comput. Biol.* **11:** e1004271

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV & Ren B (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature* **488:** 116–120

Velasco S, Ibrahim MM, Kakumanu A, Garipler G, Aydin B, Al-Sayegh MA, Hirsekorn A, Abdul-Rahman F, Satija R, Ohler U, Mahony S & Mazzoni EO (2016) A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells. *Cell Stem Cell*

Wang Z-X, Teh CH-L, Chan CM-Y, Chu C, Rossbach M, Kunarso G, Allapitchay TB, Wong KY & Stanton LW (2008) The transcription factor Zfp281 controls embryonic stem cell pluripotency by direct activation and repression of target genes. *Stem Cells* **26:** 2791–2799

Zhang Y, An L, Yue F & Hardison RC (2016) Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.* **44:** 6721–6731

Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordân R & Rohs R (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U. S. A.* **112:** 4654–4659

Figure Legends

Figure 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.

Figure 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 SeqUnwinder on simulated genomic sites in "A". **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.

Figure 3. **Sequence feature analysis at Lhx3 binding classes during iMN programming using SeqUnwinder. 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". **C)** ChIP-Seq profiles of Oct4 and Onecut2 ordered according to the binding classes defined in A).

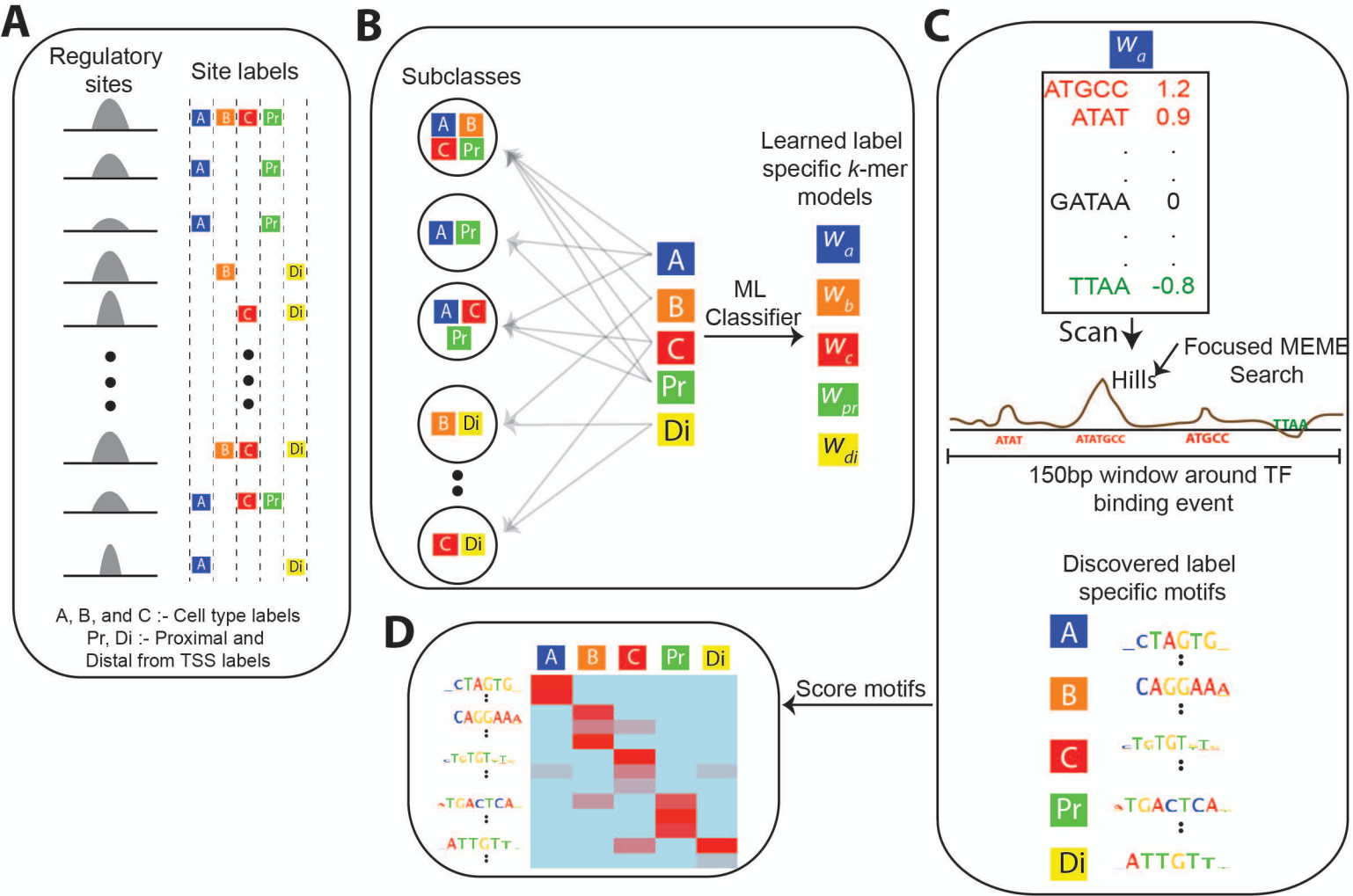
Figure 4. **SeqUnwinder analysis of sequence features at multi-condition TF binding sites for 16 ENCODE TFs. A)** Subclasses and labels at binding sites of 16 examined ENCODE TFs. **B)** Label specific scores of top scoring *de novo* motifs at shared sites. For 13 out of 16 examined TFs the top scoring shared motifs match to the cognate motifs of the TF. **C)** Comparison of PWM hit-rates and SeqUnwinder scores for USF1. **D)** Co-factor motifs identified by SeqUnwinder at cell-type specific sites across the 16 TFs.

Figure 5. **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".

Figure S1. Distance of TF binding events from annotated mRNA TSS for all 16 examined ENCODE TFs, stratified based on "shared" (black) or "cell line-specific" (yellow) labels. The X-axis represents the distance in "bp" in log-scale (natural logarithm).

Figure S2. Demonstration of SeqUnwinder's discovery of sequence properties at multi-conditionally bound NRSF binding sites. **A)** Heatmaps showing the NRSF ChIP-Seq reads at curated NRSF 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).

Figure S3. A) Rate of peaks that contain one or more motif instances, and **B)** total rate of motif instances for *de novo* identified motifs matching cognate binding preferences at labeled sites. Only 13 of 16 examined TFs, for which the *de novo* cognate motif was discovered as a discriminative feature, are shown here.


A
Regulatory sites
Site labels
A, B, and C :- Cell type labels
Pr, Di :- Proximal and Distal from TSS labels
B
Subclasses
Learned label specific k-mer models
ML Classifier
C
ATGCC 1.2
ATAT 0.9
GATAA 0
TTAA -0.8
Scan
Hills
Focused MEME Search
ATAT
ATATGCC
ATGCC
TTAA
150bp window around TF binding event
Discovered label specific motifs
Score motifs
D
CTAGTG
CAGGAAA
TGACTCA
ATTGTT

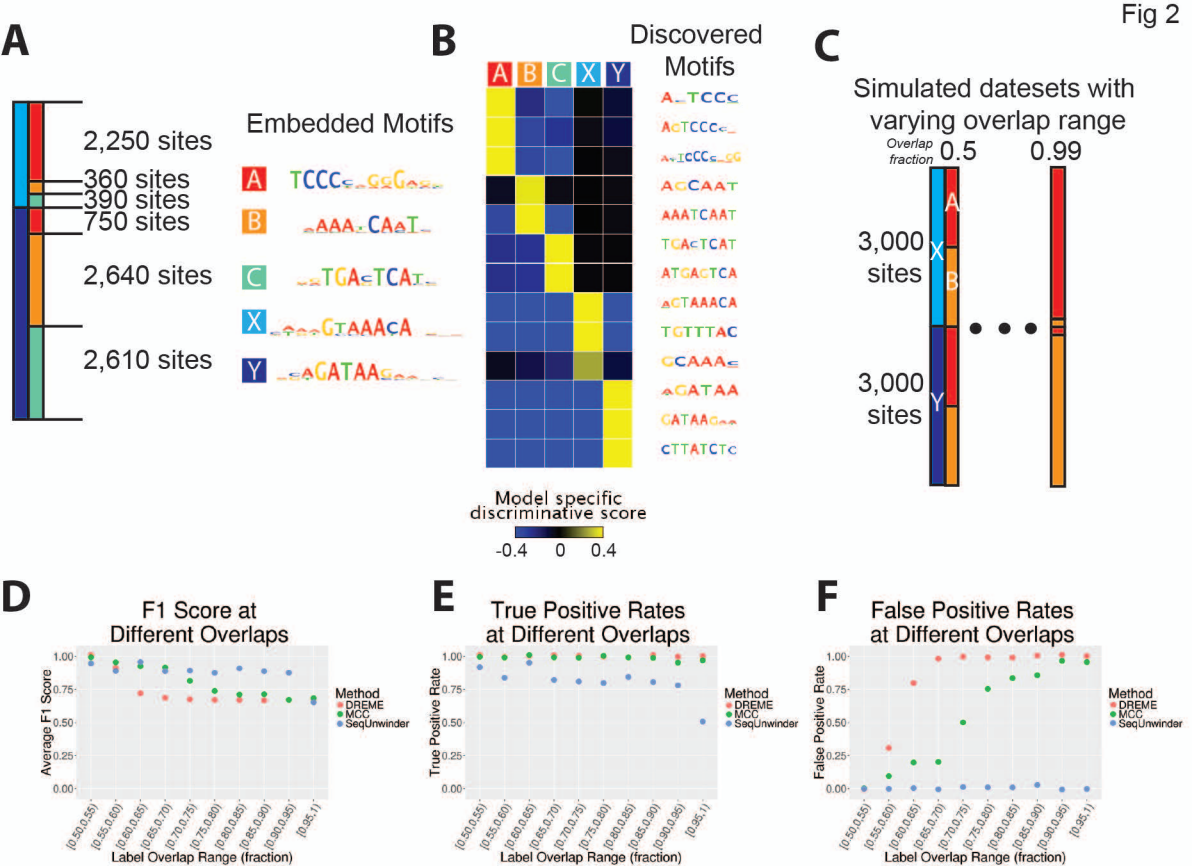

A
2,250 sites
360 sites
390 sites
750 sites
2,640 sites
2,610 sites
Embedded Motifs
A
B
C
X
Y
B
A B C X Y
Discovered Motifs
Model specific discriminative score
-0.4 0 0.4
C
Simulated datasets with varying overlap range
Overlap fraction
0.5
0.99
3,000 sites
3,000 sites
D
F1 Score at Different Overlaps
Average F1 Score
Label Overlap Range (fraction)
E
True Positive Rates at Different Overlaps
True Positive Rate
Label Overlap Range (fraction)
F
False Positive Rates at Different Overlaps
False Positive Rate
Label Overlap Range (fraction)
Method
DREME
MCC
SeqUnwinder

Fig 3

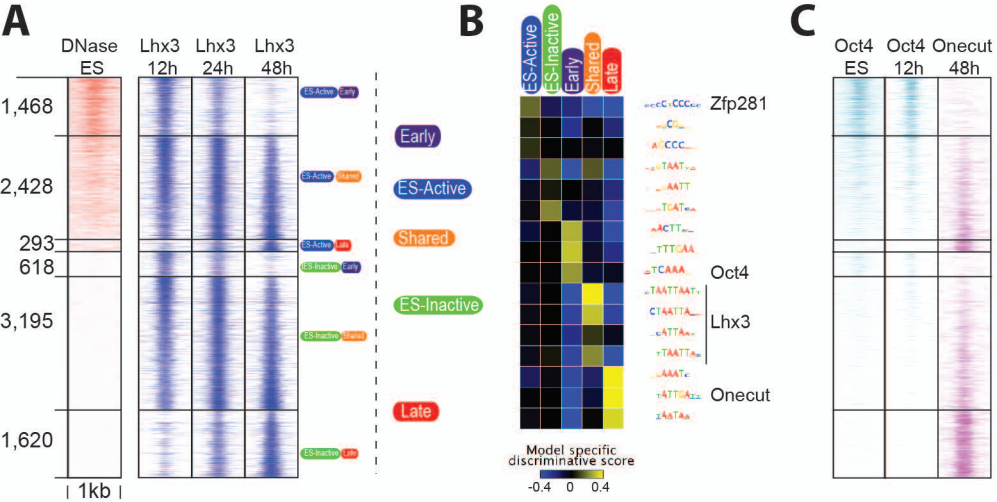
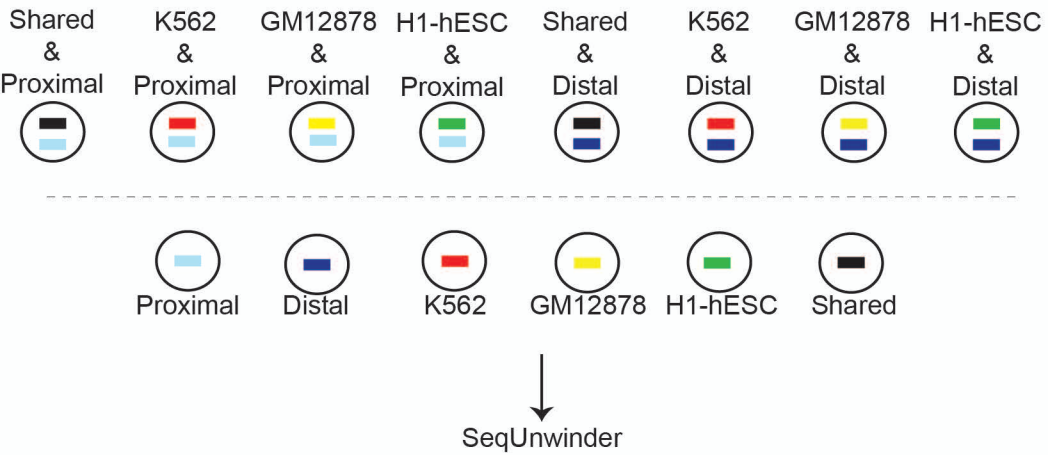

A
DNase ES
Lhx3 12h
Lhx3 24h
Lhx3 48h
1,468
2,428
293
618
3,195
1,620
1kb
ES-Active Early
ES-Active Shared
ES-Active Late
ES-Inactive Early
ES-Inactive Shared
ES-Inactive Late
Early
ES-Active
Shared
ES-Inactive
Late
B
ES-Active
ES-Inactive
Early
Shared
Late
Zfp281
Oct4
Lhx3
Onecut
Model specific discriminative score
-0.4 0 0.4
C
Oct4 ES
Oct4 12h
Onecut 48h

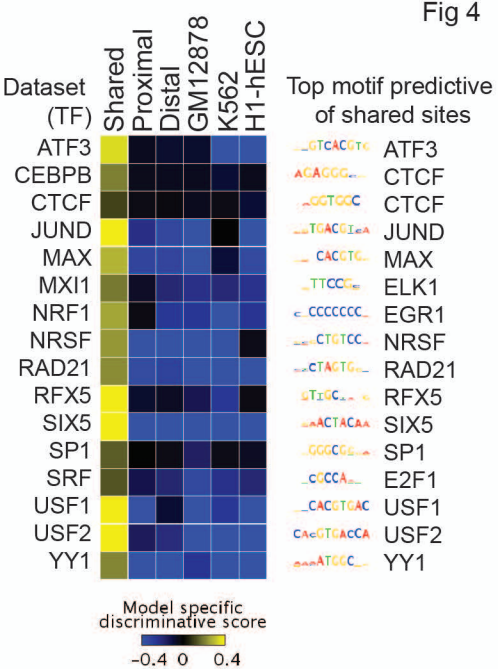
Fig 4

A

TF binding site classes for ENCODE factors



B



C

De novo USF1 motif

Cis-bp USF1 motif

	Peak -rate	Peak (CG) -rate
Shared	0.54	0.41
K562	0.32	0.24
GM12878	0.27	0.21
H1hESC	0.41	0.12

D

Co-factor motif	GM12878	K562	H1-hESC
RUNX	3/16	0/16	0/16
IRF	9/16	0/16	0/16
GATA	0/16	13/16	0/16
TEAD	0/16	0/16	11/16

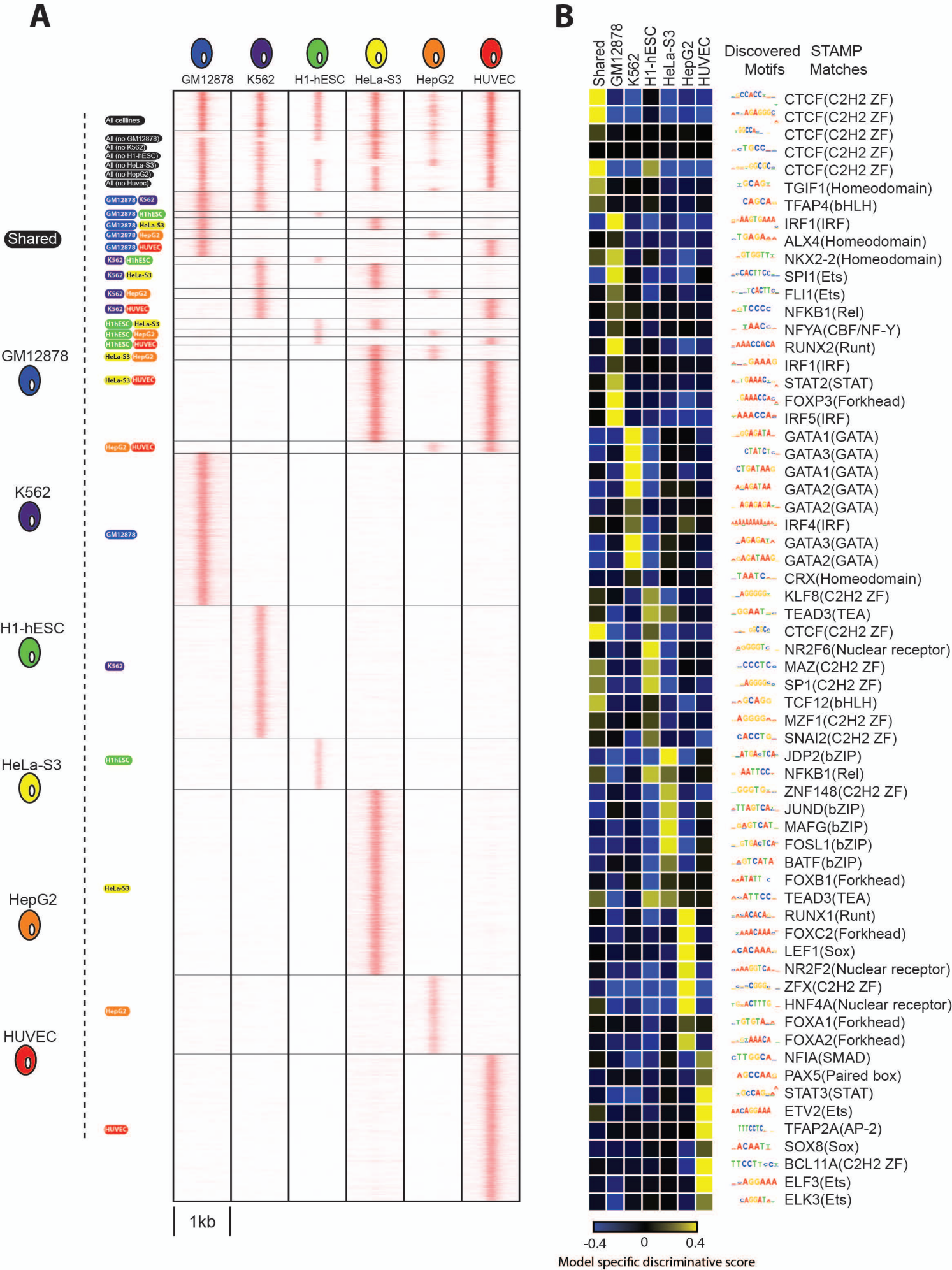
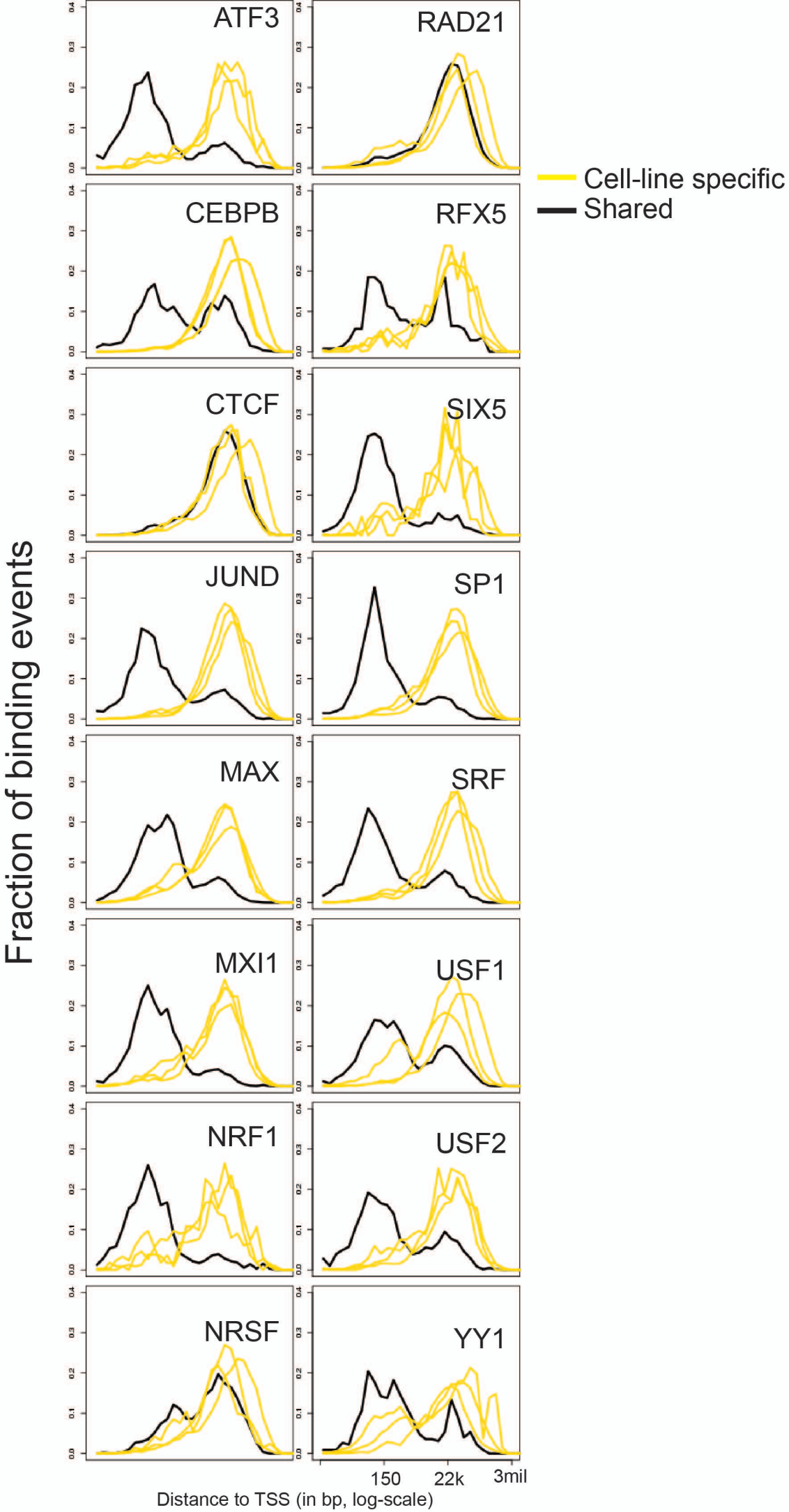

A
GM12878
K562
H1-hESC
HeLa-S3
HepG2
HUVEC
All celllines
All (no GM12878)
All (no K562)
All (no H1-hESC)
All (no HeLa-S3)
All (no HepG2)
All (no Huvec)
Shared
1kb
B
Shared
Discovered Motifs
STAMP Matches
CTCF(C2H2 ZF)
TGIF1(Homeodomain)
TFAP4(bHLH)
IRF1(IRF)
ALX4(Homeodomain)
NKX2-2(Homeodomain)
SPI1(Ets)
FLI1(Ets)
NFKB1(Rel)
NFYA(CBF/NF-Y)
RUNX2(Runt)
STAT2(STAT)
FOXP3(Forkhead)
IRF5(IRF)
GATA1(GATA)
GATA3(GATA)
GATA2(GATA)
IRF4(IRF)
CRX(Homeodomain)
KLF8(C2H2 ZF)
TEAD3(TEA)
NR2F6(Nuclear receptor)
MAZ(C2H2 ZF)
SP1(C2H2 ZF)
TCF12(bHLH)
MZF1(C2H2 ZF)
SNAI2(C2H2 ZF)
JDP2(bZIP)
ZNF148(C2H2 ZF)
JUND(bZIP)
MAFG(bZIP)
FOSL1(bZIP)
BATF(bZIP)
FOXB1(Forkhead)
RUNX1(Runt)
FOXC2(Forkhead)
LEF1(Sox)
NR2F2(Nuclear receptor)
ZFX(C2H2 ZF)
HNF4A(Nuclear receptor)
FOXA1(Forkhead)
FOXA2(Forkhead)
NFIA(SMAD)
PAX5(Paired box)
STAT3(STAT)
ETV2(Ets)
TFAP2A(AP-2)
SOX8(Sox)
BCL11A(C2H2 ZF)
ELF3(Ets)
ELK3(Ets)
-0.4
0
0.4
Model specific discriminative score

Fig S1


ATF3
RAD21
CEBPB
RFX5
CTCF
SIX5
JUND
SP1
MAX
SRF
MXI1
USF1
NRF1
USF2
NRSF
YY1
Cell-line specific
Shared
Fraction of binding events
150
22k
3mil
Distance to TSS (in bp, log-scale)

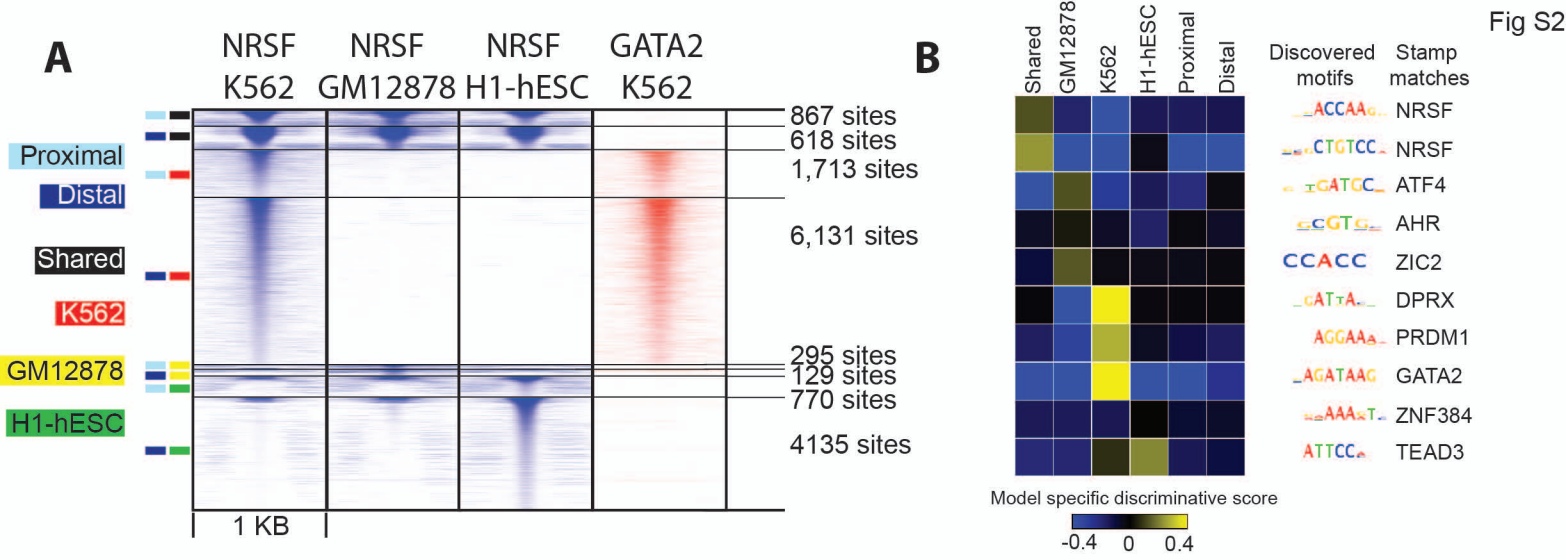


Fig S2
A
NRSF K562
NRSF GM12878
NRSF H1-hESC
GATA2 K562
Proximal
Distal
Shared
K562
GM12878
H1-hESC
867 sites
618 sites
1,713 sites
6,131 sites
295 sites
129 sites
770 sites
4135 sites
1 KB
B
Shared
GM12878
K562
H1-hESC
Proximal
Distal
Discovered motifs
Stamp matches
NRSF
NRSF
ATF4
AHR
ZIC2
DPRX
PRDM1
GATA2
ZNF384
TEAD3
Model specific discriminative score
-0.4
0
0.4

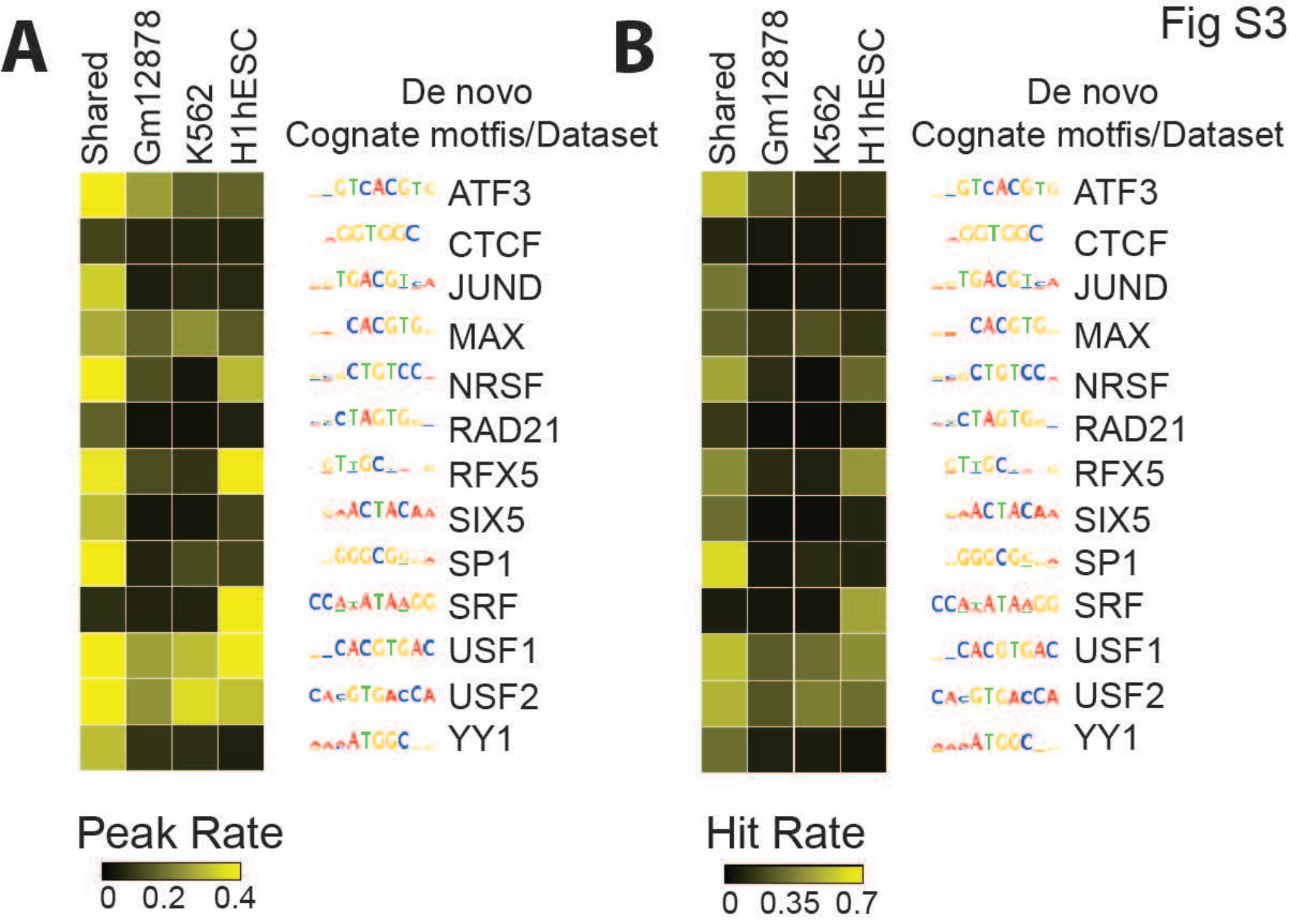


Fig S3
A
Shared
Gm12878
K562
H1hESC
De novo
Cognate motfis/Dataset
ATF3
CTCF
JUND
MAX
NRSF
RAD21
RFX5
SIX5
SP1
SRF
USF1
USF2
YY1
Peak Rate
0 0.2 0.4
B
Shared
Gm12878
K562
H1hESC
De novo
Cognate motfis/Dataset
ATF3
CTCF
JUND
MAX
NRSF
RAD21
RFX5
SIX5
SP1
SRF
USF1
USF2
YY1
Hit Rate
0 0.35 0.7