AIControl: Replacing matched control experiments with machine learning improves ChIP-seq peak identification

Nao Hiranuma[1], Scott M. Lundberg[1], and Su-In Lee[1,2]

[1] Paul G. Allen School of Computer Science and Engineering, University of Washington
[2] Department of Genome Sciences, School of Medicine, University of Washington
email: suinlee@cs.washington.edu

Abstract

Motivation: Accurately identifying the binding sites of regulatory proteins remains a central and unresolved challenge in molecular biology. The most commonly used experimental technique to determine binding locations of transcription factors is chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq). Because ChIP-seq is highly susceptible to background noise, the current practice obtains one matched "control" ChIP-seq dataset and estimates position-wise background distributions using ChIP-seq signals from nearby positions (e.g., within 5,000-10,000 bps). This approach poses the following four problems. (1) Incorporating such a large window of nearby positions may cause an inaccurate estimation of position-specific background distributions. (2) There are multiple ways to obtain control ChIP-seq datasets, but current peak calling methods constrain users to only one. (3) A single matched control dataset may not capture all sources of noise signals. (4) Generating a matched control dataset incurs additional time and cost.

Methods: We introduce the *AIControl* framework, which replaces matched control experiments by automatically learning weighted contributions of a large number of *publicly available control ChIP-seq datasets* to generate position-specific background noise distributions. We thereby avoid the cost of running control experiments while simultaneously increasing binding location accuracy. Specifically, AIControl can, (1) obtain a precise, position-specific background distribution (i.e., no need to use a window), (2) use machine learning to systematically select the most appropriate set of control datasets in a data-driven way, (3) capture noise sources that may be missed by one matched control dataset, and (4) remove the need for costly and time-consuming matched control experiments.

Results: We applied AIControl to 410 ChIP-seq datasets from tier 1 or 2 cell lines in the Encyclopedia of DNA Elements (ENCODE) database whose transcription factors have motif information. AIControl used 455 control ChIP-seq datasets from 107 cell lines and 9 laboratories to impute background ChIP-seq noise signals. Without using matched control datasets, AIControl identified peaks that were more enriched for putative binding sites than those identified by other popular peak callers that used a matched control dataset. Additionally, our framework reduced the reproducibility between two ChIP-seq datasets whose associated transcription factors had no documented interactions, which suggests that AIControl better removes reproducible confounding effects. Finally, we demonstrated that our framework improves the quality of downstream analysis by showing that binding sites it identified recover documented protein interactions more accurately.

Conclusion: AIControl removes the need to generate additional matched control data and accurately predicts binding events even when testing on cell lines not in the ENCODE's con-

trol datasets. The implementation can be accessed at `https://github.com/suinleelab/AIControl`.

Introduction

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is one of the most widely used methods to identify regulatory factor binding sites and analyze regulators' functions. ChIP-seq identifies the positions of DNA-protein interactions across the genome for a regulatory protein of interest by cross-linking protein molecules to DNA strands and measuring the locations of DNA fragment enrichment associated with the protein [2, 16, 25]. The putative binding sites can then be used in downstream analysis [31, 13], for example, to infer interactions among transcription factors [24, 29, 6], to semi-automatically annotate genomic regions [12, 15], or to identify regulatory patterns that give rise to certain diseases such as cancer [5, 3].

Identifying protein binding sites from signal enrichment data, a process called "peak calling," is central to every ChIP-seq analysis, and has thus been a focus of the computational biology research community [36, 20, 38, 11, 18, 14, 28, 30, 19]. Like other biological assays, ENCODE ChIP-seq guidelines recommend that researchers obtain two ChIP-seq datasets to help separate desirable signals from undesirable noise: (1) an *IP (immunoprecipitation) dataset* to capture the actual protein binding signals using immunopreciptation, and (2) a *control dataset* to capture background noise signals [21]. Peak calling algorithms compare IP and control datasets, locate peaks likely associated with true protein binding signals, and simultaneously minimize false positives. However, despite the guideline's recommendations, many ChIP-seq users perform experiments either without a matched control dataset or with a related control dataset from a public database in order to avoid the additional time and expense of generating control datasets.

Here, we introduce AIControl (Figure 1), a single-dataset peak calling framework that replaces a control dataset with machine learning by inferring background noise from publicly available control datasets on a large scale. As noted, most popular peak callers perform comparative ChIP-seq analysis using two datasets, – IP and control. Many of them have an option to perform single-dataset analysis (i.e., IP dataset only) by determining the background noise structure from the IP dataset; however, it is unlikely to be more accurate than when a control dataset is used. AIControl aims to estimate and simulate the true background distributions at each genomic position based on the weighted contribution of a large number of publicly available control datasets, where weights are learned from both the IP dataset and all control datasets used (see Methods for details).

Most popular peak callers, – such as Model-based Analysis of ChIP-seq ver 2.0 (MACS2) [38] and Site Identification from Short Sequence Reads (SISSRs) [28] –, learn local distributions of read counts in a matched control dataset in a nearby region (Figure 1(a), right). They then identify peaks by comparing observed read counts in the IP dataset to learned local background distributions across the genome. Several methods, – such as MOdel-based one and two Sample Analysis and inference for ChIP-Seq Data (MOSAiCs) and BIDCHIPS –, use a few predictors expected to represent background noise sources, such as GC content and read mappability. MOSAiCs performs negative binomial regression of an IP dataset using GC content, read mappability, and a matched control dataset as predictors [19]. Similarly, BIDCHIPS uses staged linear regression to combine GC content, read mappability, DNase 1 hypersensitivity sites, an input control dataset, and a mock control dataset [30].

AIControl's main innovations are four-fold than traditional approaches. (1) AIControl can learn position-specific background distributions at a finer resolution than traditional approaches by leveraging multiple weighted control datasets. Most other peak callers take a large window of

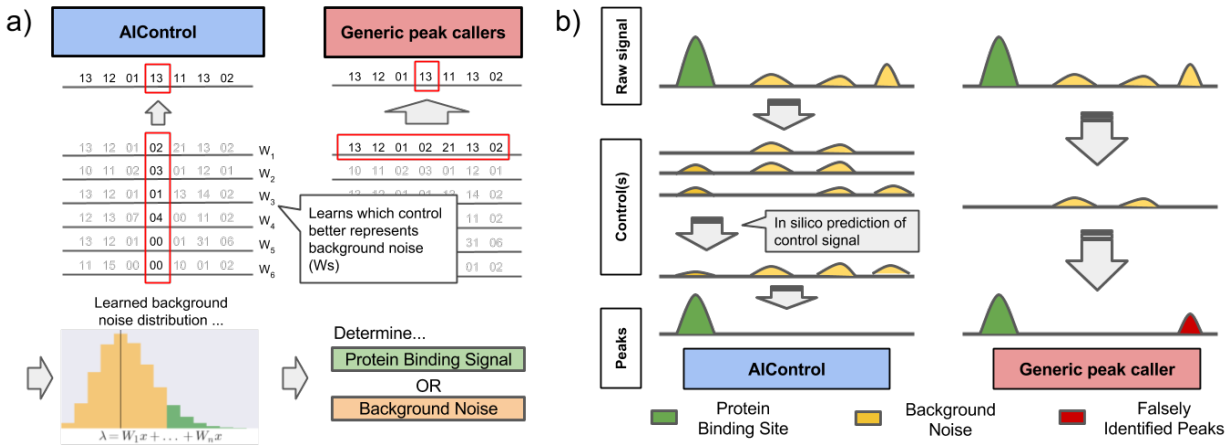


Figure 1: **(a) Comparison of AIControl to other peak calling algorithms.** (left) AIControl learns appropriate combinations of publicly available control ChIP-seq datasets to impute background noise distributions at a fine scale. (right) Other peak calling algorithms use only one control dataset, so they must use a broader region (typically within 5,000-10,000 bps) to estimate background distributions. (bottom) The learned fine scale Poisson (background) distributions are then used to identify binding activities across the genome. **(b) An overview of the AIControl approach.** A single control dataset may not capture all sources of background noise. AIControl more rigorously removes background ChIP-seq noise by using a large number of publicly available control ChIP-seq datasets (see Methods).

nearby regions to learn the position-wise distributions, which may inaccurately estimate local noise structure. This feature also offers significant improvements over our previous work, CloudControl [14]. CloudControl generates one synthetic control dataset based on publicly available control datasets and uses peak callers to identify peaks that rely on a large window of nearby signals for noise estimation. Throughout this paper, we show that AIControl significantly improves peak calling quality relative to CloudControl. (2) Existing peak callers require users to decide which control datasets to include. AIControl offers a systematic way to integrate a large number of publicly available control datasets. (3) Because AIControl integrates many control datasets, it can potentially capture more sources of noise signals compared to existing methods that use only one control (e.g., MACS2 and SISSRs). Additionally, GC content, mappability, or other confounders are likely to be present in some of the control datasets AIControl incorporates. See "Modeling background noise" in Methods for our mathematical formulation. (4) AIControl does not need a matched control dataset. We incorporate 455 control ChIP-seq datasets from 107 cell lines in the ENCODE database. By inferring local background noise structure from the large amount of publicly available data, AIControl can identify peaks even in cell lines without any previously measured control datasets. We demonstrate that our framework intelligently uses existing control datasets to estimate background distributions for IP datasets in brand new cell lines.

We evaluated the AIControl framework on 410 ChIP-seq IP datasets available in the ENCODE database [8] (Table S1 & S2) using 455 ChIP-seq control datasets (Table S3). Results show the following. (1) AIControl outperformed four other peak callers that used matching IP/control datasets when identifying putative protein binding sites based on overlaps between sequence-based motifs and DNase 1 hypersensitive regions. It predicted putative binding sites well even when all control datasets from the same cell line were removed, which suggests that AIControl is reliable even

when ChIP-seq is performed on a new cell line. (2) We investigated reproducibility between pairs of ChIP-seq datasets for transcription factors without any documented protein-protein interactions (PPIs). Reproducibility declined among pairs that were not supposed to have any reproducible true signal, suggesting that AIControl can successfully remove reproducible noise structures. (3) PPIs were more accurately predicted from peaks called by AIControl than from those called by all other methods across all tested cell lines. As a result, AIControl removes the time and cost of running control experiments while simultaneously increasing accuracy of identifying binding site locations of transcription factors.

Results

Peaks identified by AIControl are more enriched for known sequence motifs.

We compared AIControl to the following four alternative peak calling methods in terms of its enrichment for putative binding sites, the most widely used evaluation metric for peak-calling algorithms: MACS2 [38], SISSRs [28], SPP [18], and MACS2 + CloudControl [14]. To define putative binding sites without using ChIP-seq data, we identified sequence motifs in DNase 1 hypersensitive regions using FIMO from the MEME tool [1] and position weight matrices (PWMs) from the JASPAR database [17] (see Methods). MACS2, in particular, has been favored by the research community due to its simplicity and steady performance as validated by many comparative studies of peak calling algorithms [34, 36, 20]. To evaluate the enrichment for putative binding sites, we used negative log10 p-values of peaks to predict the presence of putative binding sites and measured the area under the precision-recall curves (AUPRCs). To ensure that each peak caller was tested on the same number of peaks, we measured AUPRC values on the n most significant peaks, where n is the minimum number of peaks called across all peak callers for each IP dataset (see Methods). This resulted in evaluating an average of 10,805 peaks per IP dataset for the entire hg19 genome (Figure S1).

Figure 2 and Table S4 show the AUPRCs achieved by the five peak callers for 410 IP datasets from five different cell lines: K562 (149), GM12878 (99), HepG2 (87), HeLa-S3 (60), and HUVEC (15). For all cell lines, AIControl yielded better AUPRCs than all other peak callers. Again, AIControl used only IP datasets without their matched control datasets, whereas other peak callers, except for CloudControl+MACS2, accessed both IP and their matched control datasets. To validate that we used these peak callers correctly, we further investigated the performance of all peak callers on five IP datasets for the RE1-Silencing Transcription factor (REST) measured in K562, for which we had quantitative polymerase chain reaction (qPCR) verified TF binding sites [27]. All five peak callers identified all eight qPCR-confirmed binding locations on chromosome 1 at a p-value threshold of 0.05.

Datasets that target certain transcription factors yielded more performance improvements than others. Figure S2 shows the improvement of the AIControl framework over MACS2 for datasets whose target transcription factors were available through more than 9 ChIP-seq datasets in K562. The performance of AIControl on datasets for JUN, MAX, and CTCF exhibited major improvements over MACS2, while AIControl showed a minor improvement ($< 5\%$ on average) for REST. The latter could be due to the antibody-transcription factor pair used for REST, known to be exceptionally specific and generates peaks highly enriched in sequenced reads [21]. This means there is not much background noise to better remove. There is no significant correlation between the relative performance of AIControl over MACS2 and the information content of the JASPAR PWMs (Figure S3).

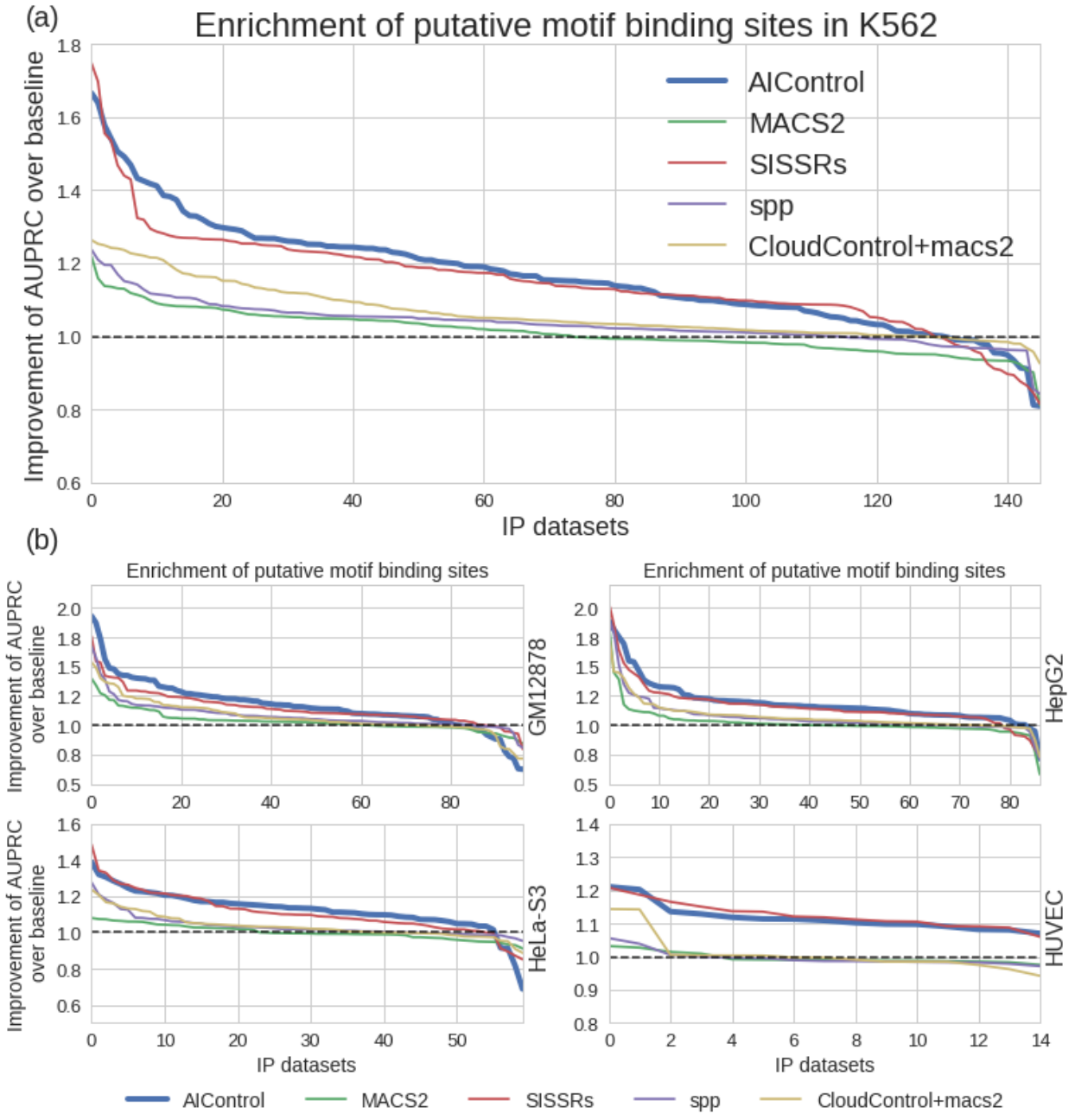


Figure 2: **Relative performance of five peak calling methods compared to MACS2 without using a matched control dataset as a baseline (dotted line)**. Peak regions were identified using: AIControl, MACS2, SISSRSs, SPP, and CloudControl + MACS2. The y-axis shows the fold improvement of the area under the precision-recall curves (AUPRCs) for predicting the presence of putative binding sites with significance values associated with the peaks over the baseline (i.e., MACS2 without using a matched control dataset). The x-axis shows the ENCODE ChIP IP datasets in each cell line ordered by the fold improvement (y-axis) for: (a) 121 IP datasets from K562, and (b) other tier 1 ENCODE cell lines (i.e., GM12878, HepG2, HeLa-S3, and endothelial cell of umbilical vein (HUVEC)).

AIControl weights reflect cell line specificity but not lab specificity.

AIControl learns the weights of contributions by all 455 ENCODE ChIP-seq control datasets to estimate the background ChIP-seq signals for each IP dataset (Figure 1; also see Methods). Figure 3 shows the magnitude of weights assigned to all 455 control datasets (columns) for each of the 410 IP datasets (rows). A clear block diagonal pattern emerges when we sort the rows and columns based on cell line (Figure 3) rather than lab that produced the data (Figure S4). This is expected because known factors for background noise signals, such as sonication bias and DNA acid isolation, depend on cell lines. The result also confirms that there is no significant lab-specific batch effects in the ENCODE ChIP-seq data. Although the control datasets from the same cell line as the IP dataset are more likely to have large weight magnitudes (Figure 3), it is important to note that AIControl learns to put high weights on some of the other biologically similar cell lines. For example, the green box in Figure 3 indicates the weights of control datasets measured in GM12892 and learned for the IP datasets in GM12878. Both are B-lymphocyte cell types, and AIControl learns to leverage information from both cell lines to identify peaks more effectively in GM12878.

AIControl performs competitively even when control datasets from the same cell lines or the same labs are removed.

The results described in the previous subsection indicate that AIControl can leverage background noise information from biologically similar cell lines. A natural question is whether AIControl can correctly identify peaks in an IP dataset from a new cell line that is not included in the public control datasets AIControl uses. Another important question is whether AIControl performs well for an IP dataset generated in a lab that did not generate the control datasets AIControl uses. To address these questions, we compared the following four settings: (1) AIControl with all 455 control datasets, (2) AIControl without control datasets from the same cell line as the IP dataset, (3) AIControl without control datasets from the same lab as the IP dataset, and (4) SISSRs without control datasets. We chose SISSRs because it performs better than all other peak callers except for AIControl (Figure 2). As expected, Figure 4 shows that AIControl continues to perform well in setting (3) compared to (1), and that AIControl's performance does not decline in setting (2) for most cell lines except K562. For example, when AIControl is used for IP datasets in GM12878, it leverages information from other B-lymphocyte cell lines (e.g. GM12892) (Figure S5). The decline in K562 has two likely reasons. First, the largest number (53 of 455, or 11.9%) of ENCODE control datasets are from the K562 cell line (Table S3). Second, K562 control datasets may be unique; unlike GM12878, which has a similar B-lymphocyte driven cell line from the same study, K562 may not have other cell lines that capture its background ChIP-seq signals well. However, these problems are naturally resolved over time as the number of control datasets in public databases continues to grows. That said, the performance of AIControl remains competitive with other peak calling methods in all cell lines including K562 and significantly improves relative to the baseline (i.e., MACS2 without a matched control dataset).

AIControl reveals protein physical interactions better than alternative methods.

Lundberg et al. (2016) showed that the *chromatin network* (i.e., a network of transcription factors (TFs) that co-localize in the genome and interact with each other) can be inferred by estimating the conditional dependence network among multiple ChIP-seq datasets. The authors showed that the inverse correlation matrix computed from a set of ChIP-seq datasets can capture many of the known physical protein-protein interactions (PPIs) from the BioGrid database [33]. Here, we use

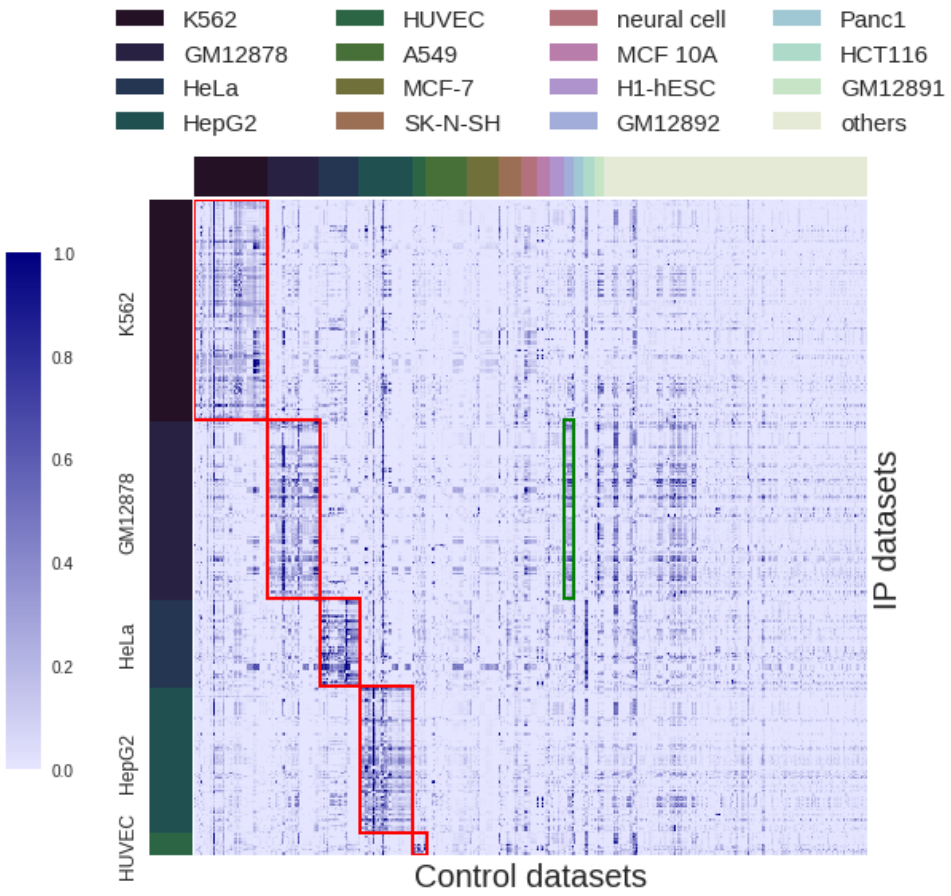


Figure 3: **Magnitudes of normalized weights the AIControl model assigns to 455 ENCODE control datasets (columns) for each of the 410 IP datasets (rows)** (Table S1 and S3). The red rectangles indicate the weights of the control datasets measured in the same cell line as the IP datasets. The green rectangle indicates the weights for control datasets measured in GM12892, which AIControl learned to estimate background ChIP-seq signals for IP datasets measured in GM12878. Both are B-lymphocyte cell lines.

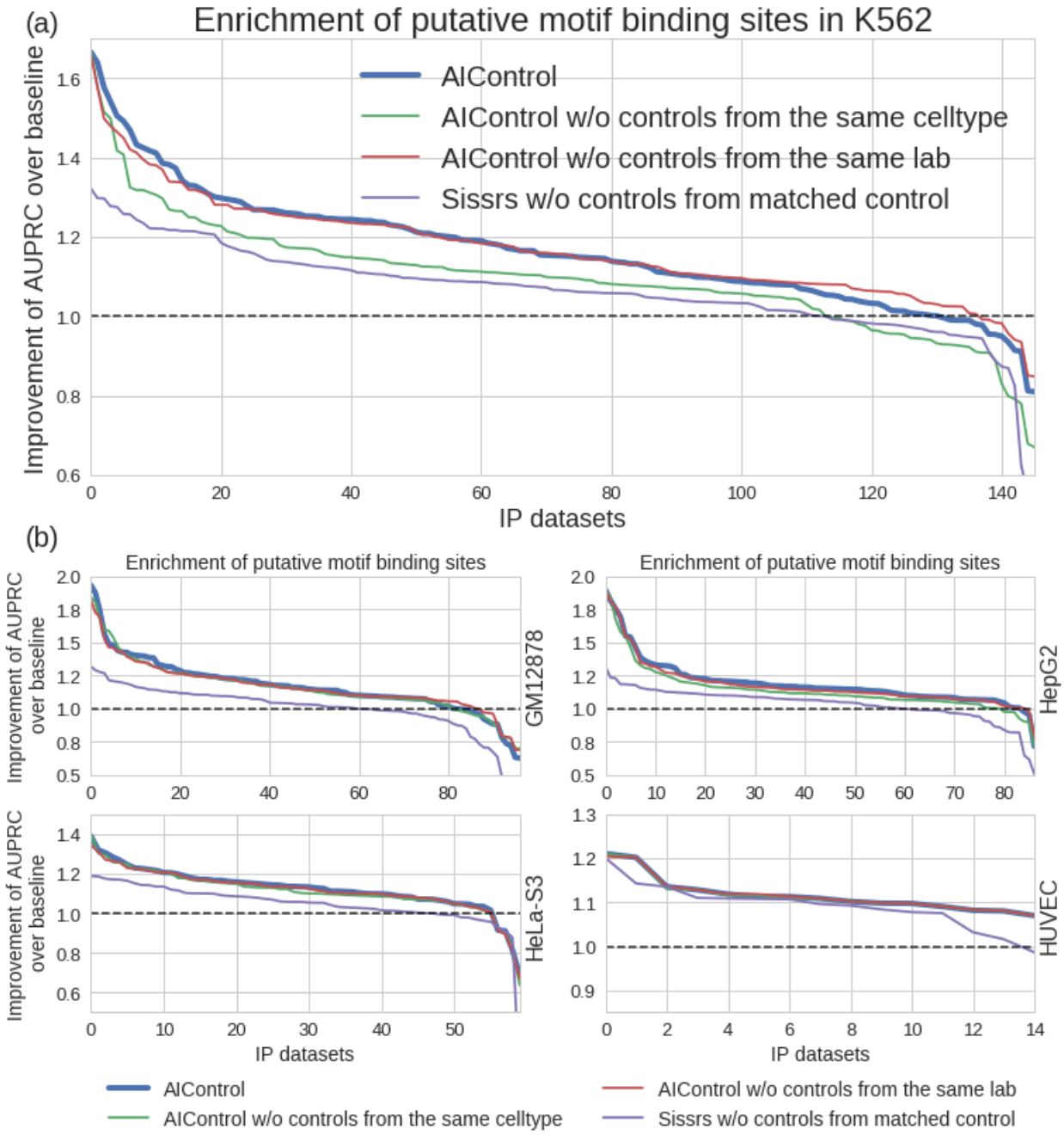


Figure 4: **Relative performance of AIControl when control datasets form the same cell lines or the same labs have been removed.** We compare across four settings: AIControl with all 455 controls, AIControl without control datasets from the same cell lines as IP datasets, AIControl without control datasets from the same labs, and SISSRs without any control sample. As in Figure 2, the y-axis shows the fold improvement of the AUPRCs for predicting the presence of putative binding sites compared with the baseline (i.e., MACS2 without using a matched control dataset) for (a) 121 IP datasets measured in K562, and (b) IP datasets measured in these tier 1 ENCODE cell lines: GM12878, HepG2, HeLa-S3, endothelial cell of umbilical vein (HUVEC).

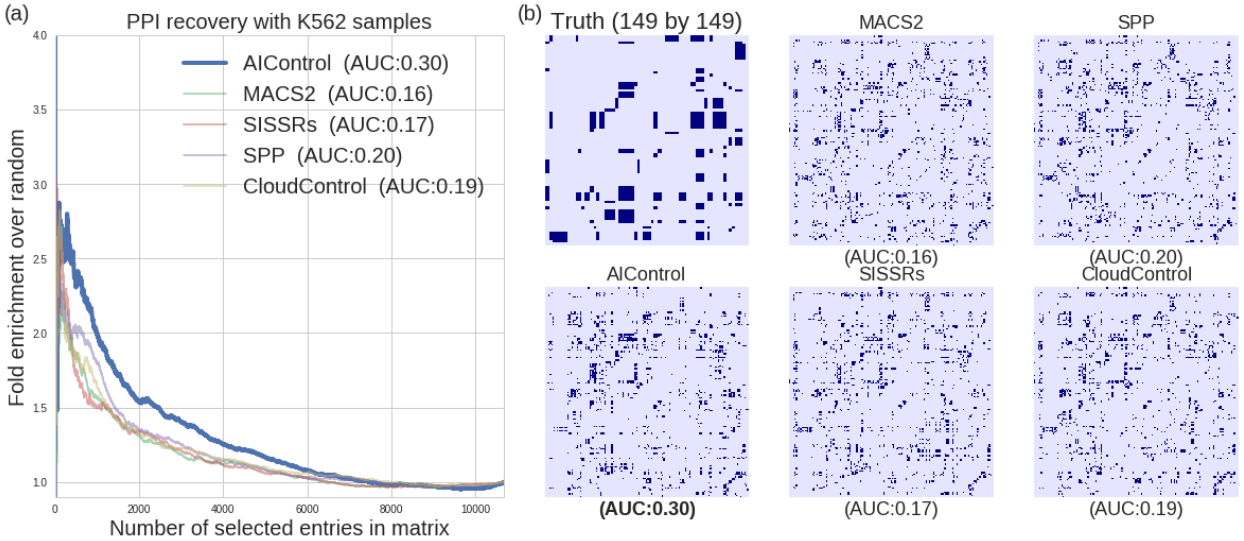


Figure 5: **Relative performance of AIControl compared to other peak callers on PPI task in K562.** (a) Enrichment of BioGrid-supported interactions of transcription factors in the inverse correlation networks inferred from 149 K562 IP datasets. Peak signals are obtained from : AIControl, MACS2, SISSRs, SPP, and CloudControl+MACS2. The network edges that connect the same transcription factors are not considered in this analysis. (b) Documented interactions among regulatory proteins (top left) and heatmaps of inverse correlation networks (rest). The network edges that refer to pairs of the same transcription factors are not shown.

the same evaluation criteria: significance of the overlap between BioGrid-supported PPIs and the network estimates inferred based on the peaks called by AIControl or by alternative methods.

Figure 5(a) shows the fold enrichment of true positive predictions over random ones with respect to the number of network edges considered (x-axis) (i.e., sorted based on the magnitude of entries in inverse correlation matrices), as revealed by Lundberg et al. (2016). Areas under the enrichment curves indicate that AIControl performs considerably better at revealing known PPIs than other methods in K562 (Figure 5) and the other cell lines (Figure S6). Notably, although SISSRs exhibits competitive performance against AIControl in the motif enrichment task (Figures 2 and 4), AIControl outperforms SISSRs by nearly two-fold in identifying previously known PPIs. In particular, AIControl predicted BioGrid-supported interactions – such as JUN/STAT1, E2F6/MAX, IRF1/STAT1, FOS/JUN and GATA2/JUN – better than other peak callers (Table S6). Additionally, we performed the same enrichment analysis on the other four cell lines: GM12878, HepG2, HeLa-S3, and HUVEC. We observed that the improved performance of AIControl in terms of the area under the enrichment curve consistently generalizes to all cell lines (Figure S6). Figure 5(b) shows the inverse correlation matrices generated from the peaks called by 5 different peak callers as well as the PPIs documented in BioGrid database (labeled as 'Truth'). Note that AIControl constructs a chromatin network that best overlaps with the groundtruth (i.e., BioGrid PPIs) relative to other methods.

AIControl uniquely recovers 15 TFs (Table S7). Among these interactions, two pairs (MEF2A/TBP and CEBPB/SRF) are known to interact with each other according to the BioGrid database. Although other interactions are not currently in the database, some studies suggest potential interactions between the pairs. For example, multiple studies suggest a functional interaction between RUNX1 and JUND; this interaction plays a key role in regulating the tissue inhibitor of metalloproteinase 1 (TIMP1) [4], and treating the SEM cell line with RUNX1 siRNA represses JUND expression [37]. Interactions among CEBPB, NFY, and other transcription factors were also thought to play a functionally important roles in the Hypoxia Inducing Factor (HIF) transcriptional response [10]. These predicted interactions, unique to AIControl, may serve as potential targets for discovering previously uncharacterized PPIs.

Although we showed that AIControl better recovers known PPIs, it is important to note that the truth matrix likely contains some false positives and potentially many false negatives. First, the truth matrix is constructed using information drawn from all available cell lines. Second, some interactions might still be undocumented in the BioGrid database. Further, our prediction from ChIP-seq data is more likely to recover interactions near DNA strands. Despite these uncertainties, the finding that AIControl recovers PPIs more effectively in all cell lines suggests that using this framework can improve the quality of downstream analysis that follows ChIP-seq experiments.

AIControl better removes common background noise among datasets.

One of the most frequently used quality measures for biological experiments is the *reproducibility* of a pair of replicate datasets. A pair of replicate datasets should capture the exact same signals; thus, a pair with better quality should be more reproducible. On the other hand, the quality of background noise removal can be assessed by measuring the *irreproducibility* of a pair of unrelated datasets. We define an "unrelated" pair as a pair of datasets that: (1) are in the same cell line, and (2) target unrelated transcription factors without any documented PPI in BioGrid. As described in Methods, AIControl models ChIP-seq experiments as follows:

$$\text{ChIP-seq data} = \text{protein-binding-signal} + \text{background-noise} + \text{irreproducible-noise}$$

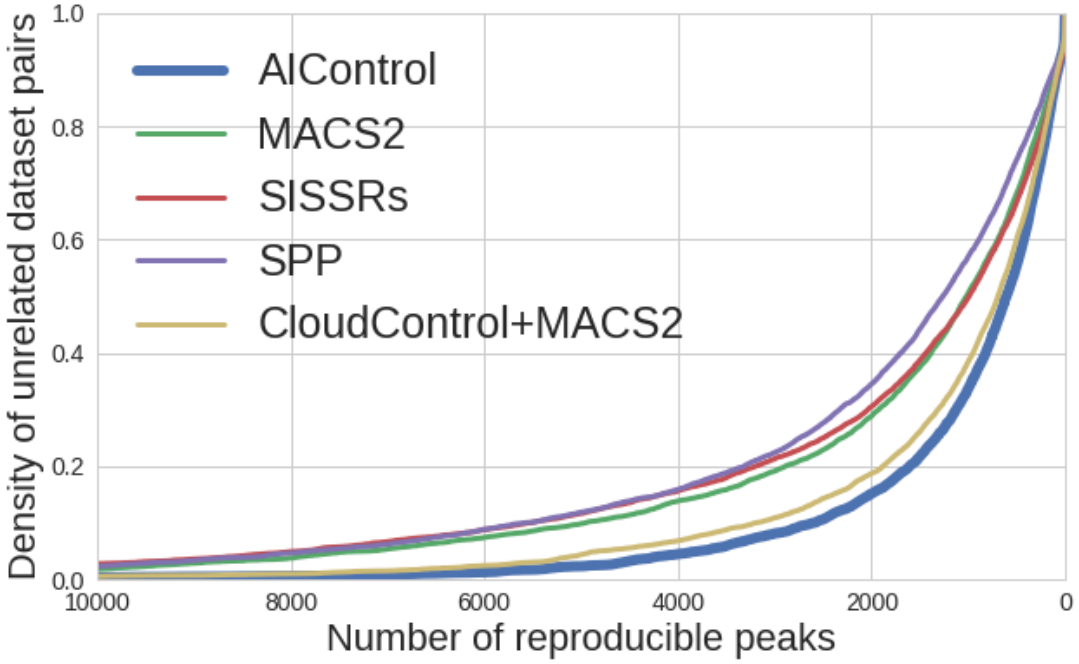


Figure 6: **Cumulative density plot of the number of reproducible peaks for 9,310 pairs of unrelated datasets in K562 processed by five peak callers**: AIControl, MACS2, SISSRs, SPP, and CloudControl+MACS2. Note that the x-axis decreases from left to right. The smaller area under the curve indicates that the evaluated pairs are less reproducible (i.e., better performance).

Because the protein-binding-signal of a pair of datasets that target 'unrelated' transcription factors should not be reproducible, the only term that could be reproduced is background-noise. However, if a peak caller perfectly removes all background noise, it should ideally leave no peak that is reproducible between a pair of unrelated datasets. Thus, peak callers better able to remove background noise should have fewer reproducible peaks for unrelated datasets. *Reproducible* peaks are defined as peaks that appear in the same genomic locations between a pair of datasets.

Figure 6 shows the number of reproducible peaks for 9,310 pairs of unrelated datasets in K562 processed by five peak callers: AIControl, MACS2, SISSRs, SPP, and CloudControl+MACS2. For AIControl, the lower area under the curve demonstrates that peaks it identified are generally less reproducible between a pair of unrelated datasets. For other peak callers, a larger percentage of unrelated dataset pairs contained more reproducible peaks, suggesting that their noise-removal process was not as complete as AIControl's.

AIControl performs well on the IP datasets outside the ENCODE database.

The recommended protocol strictly regulates ChIP-seq experiments in the ENCODE database. However, external labs do not always adhere precisely to this protocol. To assure that the AIControl framework retains strong performance on a ChIP-seq IP dataset that is not a part of the ENCODE database, we performed peak calling on all four IP datasets obtained from the study by Schmidt et al. [31] with available motif information, using the following peak calling frameworks: AIControl,

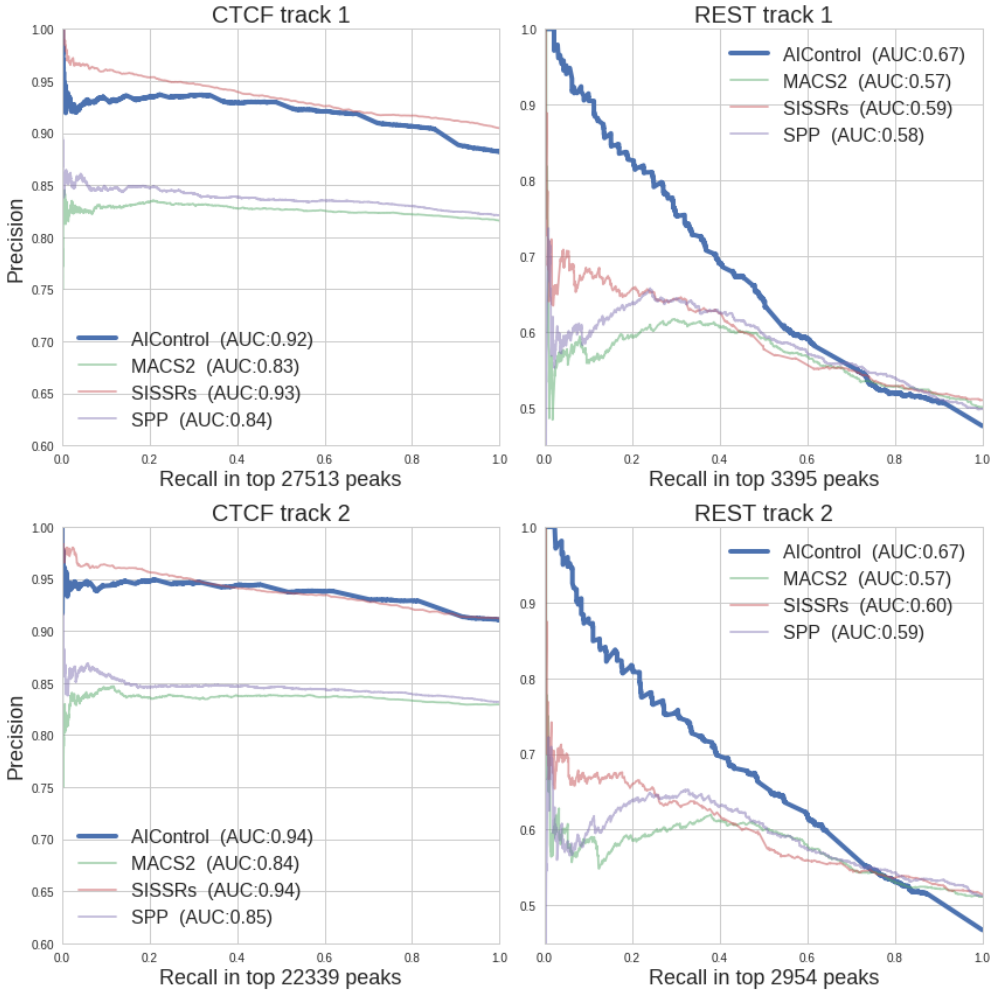
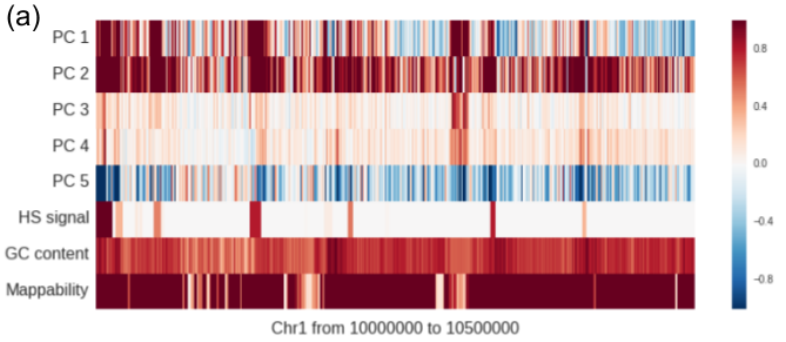


Figure 7: **Enrichment of putative motif binding sites in DNase 1 hypersensitivity sites in all datasets with available motif information from Schmidt et al. [31].** Both datasets on the left target CTCF, and both on the right side target the REST transcription factor.



(b)

	Spearman correlation			Variance captured (%)
PCs	HS	GC	MP	
1	**0.28**	0.35	**0.31**	41.3
2	0.21	**0.53**	0.12	5.16
3	0.08	0.22	0.03	2.88
4	-0.13	0.17	-0.17	2.76
5	0.03	0.07	0.05	1.95
Total				**54.5**

Figure 8: **Principal component analysis of K562 control signals.** (a) Normalized signals from 53 K562 control datasets projected onto 5 principal components and known noise factors on chromosome 1 from 10,000,000 to 10,500,000 base pairs. Signals are binned into 100 base pair bins. (b) Spearman correlation between projected control signals and known noise factors: DNase 1 hypersensitivity (HS), normalized GC content (GC), and read mappability values (MP).

MACS2, SISSRs, and SPP. These four IP datasets are not recorded in the ENCODE database.

Figure 7 shows the performance of AIControl and the other peak calling algorithms using precision-recall (PR) curves. Unlike Figures 2 and 4, where we collectively showed the performance on many IP datasets using area under the PR curve (AUPRC) values, here we show an individual PR curve for each IP dataset. The PR plot indicates that AIControl consistently performs well on four datasets in the K562 cell line that target two different transcription factors (CTCF and REST). IP datasets on PU1 and GATA1 are available from Schmidt et al. However, they are not included in this analysis due to the lack of motif PWMs for *H. sapiens* in the JASPAR database. SPP and MACS2 perform worse on the CTCF datasets, but AIControl remains competitive with SISSRs. In addition, AIControl shows significantly more enrichment of true positives in the highly ranked peaks over all other methods on the REST datasets. Note that MACS2, SISSRs, and SPP use the designated control datasets provided by Schmidt et al., while AIControl learns background signal distributions from the ENCODE dataset.

Principal components are associated with potential noise sources.

Control datasets are supposed to capture background noise that is also present in corresponding IP datasets. Many studies suggest that this background noise is a combination of multiple different noise sources, for example, GC content, sonication bias, and platform-specific biases [30, 19, 22]. The AIControl framework assumes that observed background noise in a control dataset can be represented as the weighted sum of many different known or unknown noise sources (see Methods).

Figure 8 exhibits Spearman's correlation coefficients between potential noise sources and K562 control datasets projected on the first five principal components. Open chromatin regions (HS) and read mappability (MP) are captured by the first principal component, while GC content (GC) is well captured by the second principal component. Notably, the first five principal components collectively capture only 54.5% variance, suggesting the likelihood of other noise sources that contribute to the observed background noise. AIControl implicitly learns the contributions from unobserved noise sources; thus, we believe that it improves performance relative to other peak identification

methods.

Discussion

Accurately identifying the locations of regulatory factor binding events remains a core, unresolved problem in molecular biology. AIControl offers a framework for processing ChIP-seq data to identify binding locations of transcription factors without requiring a matched control dataset.

AIControl makes key innovations over existing systems. (1) It learns position-specific distribution of background noise at much finer resolution than other methods by using publicly available control datasets on a large scale (see Methods). Our evaluation metrics show that using finer background distributions improved enrichment of putative TF binding locations, recovery of known protein protein interactions, and reduction of reproducible background noise between a pair of datasets. (2) AIControl systematically integrates control datasets from a public database (e.g., ENCODE) without any user input. Its ability to learn background noise extends to datasets obtained in new cell lines without any previously measured control datasets. We obtained 455 control ChIP datasets from 107 cell lines in the ENCODE database, and AIControl learns to statistically combine them to estimate background noise in an IP dataset in any cell line. We showed that our performance on new cell lines exceeds that of established baselines. AIControl's performance is also generalizable to datasets from labs outside the ENCODE project. (3) The mathematical model of AIControl accounts for multiple noise sources due to its integration of control datasets at a large scale (see Methods). On the other hand, some noise sources may not be fully captured by existing methods that use only one matched control dataset [38, 28] or account for only a specific set of noise sources [19, 30]. (4) Finally, AIControl reduces the time and cost incurred by generating a matched control dataset since it does not require a control to perform rigorous peak calling.

We demonstrated the effectiveness of AIControl by conducting a large-scale analysis on the peaks identified in 410 ENCODE ChIP-seq datasets from five major ENCODE cell lines for 54 different transcription factors (Table S1). We showed that AIControl has better motif sequence enrichment compared to other peak callers within predicted peak locations that are identified in DNase 1 hypersensitivity regions. However, this metric measures only direct interactions between transcription factors and DNA. Therefore, we also evaluated the performance of AIControl with two other metrics: PPI enrichment analysis and reduction of reproducible noise structure. We observed that AIControl is superior to other peak callers even without any matched control samples. With these three metrics, we showed that our framework's single-dataset peak identification performs better than other established baselines with matched controls datasets.

AIControl satisfies many of the properties favored by the comparative analysis of peak calling algorithms [34, 36, 20]. This includes the use of local distributions that are suitable for modeling count data and the ability to combine ChIP-seq and input signals in a statistically principled manner. One caveat of our default implementation is that we bin signals into 100 base pair windows in order to perform genome-wide regression; most transcription factors show signals wider than 100 base pairs. Thus, we believe that our resolution is sufficient to conduct accurate downstream analysis. Users can tune the size of binning windows if they prefer higher resolution.

ChIP-seq is one of the most widely used techniques for identifying protein binding locations. However, conducting a set of two ChIP-seq experiments can be resource intensive. By removing the cost of obtaining control datasets, we believe that AIControl can lead to more accurate ChIP-seq signals without expending additional resources. The Julia implementation and accompanying files of AIControl can be accessed at https://github.com/suinleelab/AIControl.

Methods

Modeling background noise

AIControl models the background noise signal across the genome as a linear combination of multiple different noise sources. In particular, let us denote a control ChIP-seq dataset i as $y_i \in \mathbb{R}^g$, where g represents the number of binned regions across the whole genome. Let us also denote the signals from n noise sources as $x_1, ..., x_n \in \mathbb{R}^g$. For example, x_1 may represent the GC content across the whole genome. Then, we model each control dataset y_i as a linear combination of $x_1, ..., x_n$:

$$y_i = w_{i1}x_1 + w_{i2}x_2 + ... + w_{in}x_n \tag{1}$$

$$\hat{y}_i = y_i + \epsilon_i \tag{2}$$

Here, ϵ_i represents irreproducible noise in a control dataset i, and $\hat{y}_i$ represents an observed control dataset i. Each control dataset is modeled as a specific linear combination of n noise sources, and $w_i = \langle w_{i1}, w_{i2}, ..., w_{in} \rangle \in \mathbb{R}^n$ corresponds to a specific control dataset i. These weight vectors of all control datasets are *not* observed.

For a particular target IP dataset t, AIControl attempts to estimate the background noise signal $\hat{y}_t$, which is modeled as a weighted linear combination of $x_1, ..., x_n$ with a weight vector $\hat{w}_t \in \mathbb{R}^n$:

$$\hat{y}_t = \hat{w}_{t1}x_1 + \hat{w}_{t2}x_2 + ... + \hat{w}_{tn}x_n + \epsilon_t. \tag{3}$$

Below, we show that we can estimate $\hat{y}_t$ without explicitly learning $\hat{w}_t$ and $x_1, ..., x_n$. The idea is that we can view a set of weight vectors $w_1, ...w_m \in \mathbb{R}^n$ from m publicly available control datasets (here, 455 ENCODE datasets) as a spanning set of $\mathbb{R}^n$ (or a large subset of it) provided that $n << m$. Thus, we can model $\hat{w}_t$ as a linear combination of $w_1, ...w_m$:

$$\hat{w}_t = a_1w_1 + a_2w_2 + ... + a_mw_m. \tag{4}$$

Plugging equation (4) into equation (3) leads to:

$$\begin{aligned}
\hat{y}_t &= (a_1w_{11} + ... + a_mw_{m1}) \cdot x_1 + ... + (a_1w_{1n} + ... + a_mw_{mn}) \cdot x_n + \epsilon_t && (5)\\
&= a_1 \cdot (w_{11}x_1 + ... + w_{1n}x_n) + ... + a_m \cdot (w_{m1}x_1 + ... + w_{mn}x_n) + \epsilon_t && (6)\\
&= a_1y_1 + a_2y_2 + ... + a_my_m + \epsilon_t && (7)\\
&= a_1\hat{y}_1 + a_2\hat{y}_2 + ... + a_m\hat{y}_m, && (8)
\end{aligned}$$

where ϵ_t represents the total irreproducible noise. This shows that $\hat{y}_t$ can be represented as a weighted linear combination of a large number of m control datasets. To learn the coefficient vector, $a = \langle a_1, ..., a_m \rangle$, we could do a linear regression of a true background-noise vector for IP dataset t, y_t, against $\hat{y}_1, ..., \hat{y}_m$; however, y_t is not observed. Instead, we regress the observed signal of the IP dataset, o_t, against $\hat{y}_1, ..., \hat{y}_m$ given that o_t can be decomposed as follows.

$$\begin{aligned}
o_t &= \textsf{protein-binding-signal} + \textsf{background-noise} + \textsf{irreproducible-noise} && (9)\\
&= p_t + y_t + \epsilon_t && (10)
\end{aligned}$$

The idea is that in theory, m control datasets, $\hat{y}_1, ..., \hat{y}_m$, should contain no information about p_t and ϵ_t; therefore, we can determine the coefficient vector a by regressing o_t against $\hat{y}_1, ..., \hat{y}_m$ unless we overfit. Here, the sample size is millions, and the number of variables is 455, which means that this problem is far from high-dimensional and unlikely to overfit.

Computing coefficients

To further reduce the chance of overfitting, we regularize AIControl by applying the L2 ridge penalty on the coefficient vector $a = \langle a_1, ..., a_m \rangle$. This leads to the following objective function:

$$\underset{a}{\text{argmin}} ||o_t - Ya||_2^2 + \lambda ||a||_2^2. \tag{11}$$

Here, Y is a g by m ($= 455$) matrix, where each column i corresponds to $\hat{y}_i$, the ith observed control dataset. Using the closed form solution of ridge regression, we can efficiently compute the coefficient vector, a:

$$\hat{a} = (Y^T Y - \lambda I)^{-1} Y^T o_t. \tag{12}$$

Because this regression problem involves a large number of samples (i.e., is far from being high-dimensional), we chose a small regularization coefficient $\lambda = 0.00001$ to ensure numerical stability. Since the dimension of Y is g by m, where m is 455 and g is 30 million (under the default setting where the size of bins is 100 bps), we are unlikely to require strict regularization to prevent overfitting.

When implementing AIControl, we learned separate models for signals mapped to forward/reverse strands and even/odd positions, which results in four coefficient vectors $\hat{a}$ per target IP dataset. Background-noise is estimated separately for forward and reverse strands as $\hat{y}_{\text{forward}}$ and $\hat{y}_{\text{reverse}}$ by applying the coefficients learned at even positions to calculate $\hat{y}$ at odd positions, and vice versa. Spearman's correlation values of learned coefficients are shown in Figure S7. These values are generally above 0.8 for any pair in the same IP dataset.

It is important to note that we need not to recompute $Y^T Y \in \mathbb{R}^{455 \times 455}$ for different IP datasets because it remains constant given that the same set of control datasets is reused. To estimate $\hat{y}_t$, we need only two passes through the whole genome: the first to compute $Y^T o_t$, and the second to calculate $\hat{a}Y$.

Identifying peaks

Commonly used peak calling approaches identify a peak based on how much its read count at a particular genomic region diverges from the null distribution (typically, Poisson, Zero-inflated Poisson, or negative binomial distribution) that models background signal without protein-binding events [38, 28]. Usually, null distributions are semi-locally fit to signals from nearby regions (5000 – 10000 bps) in a matched control dataset.

Like other peak callers, AIControl uses the Poisson distribution to identify peak locations; however, null background distributions are learned at much finer scale. In particular, we use the following probabilistic model to model the null background distribution for the read count observed at the ith position of genome c_{ti} in the target dataset t:

$$\begin{aligned} c_{ti} &\sim \mathsf{Poisson}(\lambda = \mathsf{maximum}(\hat{y}_{ti}, 1)) \\ &= \mathsf{Poisson}(\lambda = \mathsf{maximum}(a_1 \hat{y}_{1i} + a_2 \hat{y}_{2i} + ... + a_m \hat{y}_{mi}, 1)), \end{aligned}$$

where $\hat{y}_1, ..., \hat{y}_m$ represent m publicly available control datasets, and $a_1, ..., a_m$ are estimated using equation (12). This approach can be viewed as fitting a Poisson distribution to count data at each

genomic bin i, $\hat{y}_{1i}, ..., \hat{y}_{mi}$, that are weighted differently with corresponding weights $a_1, ..., a_m$. The use of m control datasets (not just one matched control) lets us learn background distribution more locally (Figure 1(a)). Finally, we introduce the minimum base count of 1 read to prevent $\hat{y}_t$ from being negative since the coefficient vector a can contain negative values. In our implementation, users have an option to include nearby b bins to learn the null background distribution in case they do not have a sufficiently large enough m.

We then calculate the p-value of the observed count at each genomic bin based on the learned null background distribution. To this point, peak identification processes are completed separately for forward and reverse strands; we use $a_1, .., a_m$ learned from even-numbered regions to identify peaks at odd-numbered regions for each forward and reverse strand. We then combine the p-values from forward and reverse strands by sliding the locations of the p-values by $\frac{d}{2}$ and $-\frac{d}{2}$, respectively. d is defined as the expected distance between forward and reverse peaks; it is a user-specifiable parameter set to 200 bps as a default. Finally, the more significant negative log10 p-value between the slide forward and reverse signals at every position is output as a peak signal. This last step ensures that peaks have the bimodal shapes expected for transcription binding signals [38].

Aligning BAM files

We descibe BAM files used in this project in our work on ChromNet [24]. Specifically, the raw FASTQ files were downloaded from the ENCODE database and mapped to the hg38 genome with BOWTIE2 to ensure a uniform processing pipeline [22]. We provide the full list of ENCODE experimental IDs used in this project in Supplementary Data 1.

Calling peaks with other methods

The version of MACS2 used in this paper was MACS2 2.1.0.20150731 [38]. The peaks were called with arguments macs2 callpeak -f BAM -t chipseq_dataset –control matched_control -q 0.05. The version of SPP used was 1.13, and the peaks were called with an FDR threshold of 0.05 [18]. For SISSRs, we used v1.4, and the peaks were also called with an FDR threshold of 0.05 [28]. The peaks from CloudControl were obtained in conjunction with MACS2 using the same parameters as above [14].

Standardizing peak signals

Different peak calling algorithms identify different numbers of peaks at a given p-value threshold and generate peaks with different widths. To eliminate the possibility that these differences in peak numbers or widths create bias, we standardized peak signals for each dataset as follows. (1) We bin peak signals by 1,000 base pair windows. This creates vectors where each entry corresponds to the peak with the most significant p-value among the peaks that fall into the corresponding bin. We thereby standardize peak width to 1,000 bps across all methods. (2) For each dataset, we use the top n peaks for all peak callers, where n is the minimum number of peaks identified from all tested peak callers at a p-value cutoff of 0.05. This process, which standardizes the number of peaks identified by different peak callers, results in an average of 10,805 genome-wide peaks per dataset (Figure S1).

Evaluating motif enrichment

We applied AIControl to 410 ChIP-seq datasets from the ENCODE database for which we could find motif information. For each IP dataset, we obtained a probability weight matrix (PWM) of binding sites for its target transcription factor from the JASPAR database [17]. We then used FIMO from the MEME software to search for the putative binding sites at the p-value threshold of 10^{-5} [1]. Motif enrichment alone may not be a reliable measure of true transcription factor binding events. Studies show that 98.5% of the transcription factor binding sites are positioned in DNase 1 hypersensitivity (DHS) regions [35]. Therefore, to increase the reliability of motif-based evaluation criteria, we focused on DHS regions when we performed motif enrichment-based evaluation. After standardizing peak signals across all methods (as described in the previous subsection), for each peak calling method, we predicted the presence of putative binding sites in DHS regions using varying thresholds of the significance of peaks. This led to a standard precision-recall curve for predicting the presence of putative binding sites when the significance level of the peak varies (i.e., x-axis in the standard precision-recall curve). We then used the area under the precision-recall curve (AUPRC) to assess peak calling method performance. We computed the AUPRC using Riemann sum approximation.

We used 'waterfall plots' to collectively visualize the AUPRCs of all peak callers for all IP datasets in each cell line. For example, in Figures 2 and 4, each color corresponds to the performance of a particular peak calling algorithm on IP datasets. The y-axis measures the ratio of AUPRC given by the corresponding peak calling methods to AUPRC given by the baseline method, i.e., MACS2 without a control dataset (also represented by the dotted line). The x-axis corresponds to an IP dataset, which are sorted independently for different peak calling methods based on their y-axis values. AIControl's larger area under the colored line and above the dotted line relative to alternative methods indicates its superior performance in identifying peaks supported by sequence motifs (Table S4 and S5).

Using area under PR curve of n most significant peaks as an evaluation metric

As described in 'standardizing peak signals' above, we analyze only n most significant peaks, where n is determined by the minimum number of peaks called from all peak callers. We then generated the precision-recall curve for predicting presence of putative binding sites using peak significance values and used the AUPRC to assess peak calling quality. Here, we aim to justify the use of AUPRC as an evaluation metric. Figure S8 explains what the AUPRC metric captures for peak callers with different behaviors. Figure S8(a) shows a precision-recall curve when a peak caller performs well at selecting true peaks in top n (captured by area A) but poorly at ordering them in top n (captured by area B). Figure S8(b) shows an example for the opposite case, in which a peak caller perform poorly at placing true peaks in top n but well at ordering them in top n. Both quantities measured by the area A and B are important for high quality peak calling. In practice, researchers use $500 \sim 5000$ most significant peaks depending on transcription factors. Our average choice for n is 10,805, much higher than the widely used threshold (Figure S1).

Obtaining true protein-protein interaction (PPI) matrix

The validated PPI interactions we used for evaluation were downloaded from the BioGrid website by 2018/2/5 [33]. We used only the PPIs in Homo sapiens from BIOGRID-ORGANISM-Homo_sapiens-3.4.157.mitab.txt. Because the interactions are recorded in terms of Entrez ID in BioGrid, the uniprot IDs of the targeted transcription factor of ENCODE IP datasets were

converted to Entrez ID using the Uniprot Mapping Tool from http://www.uniprot.org/mapping/.

PPIs were estimated for each cell line as follows. First, for each peak calling method, the inverse correlation matrix from all n IP datasets in the cell line of interest was computed using standardized peak signals (see 'Standardizing peak signals'). This resulted in a matrix of size n by n. Then, the matrix entries that correspond to an interaction between the same transcription factor were removed from evaluation. Finally, the magnitudes of the inverse correlation values were used as predictors for PPIs.

To visualize the quality of predictions, we used fold enrichment plots (Figures 5 and S6), like we did previously [24]. Fold enrichment is defined as follows for given number of selected predicted interactions (x-axis of Figures 5 and S6):

$$\text{fold enrichment} = \frac{\text{\# of BioGrid-validated edges}}{\text{expected \# of validated by random}}. \tag{13}$$

This value has been shown to reflect both type 1 and type 2 errors. We plot the fold enrichment value (y-axis) against the number of predicted interactions selected (x-axis) (Figures 5 and S6). A larger area under the fold enrichment curve indicates superior performance similar to PR curves.

Conducting reproducibility analysis for a pair of unrelated IP datasets

This analysis used 9,310 pairs of IP datasets in K562 that target unrelated transcription factors. Here, 'unrelated' means that the pair of transcription factors has no documented PPIs based on the BioGrid database [33]. The number of reproducible peaks between a pair of datasets is computed as follows: First, we binarize the standardized peak signals for each dataset. Then, we counted the number of non-zero entries at the intersection between two datasets. This gives us the number of peaks in the same genomic locations between a pair of datasets.

Availability of implementation and evaluation pipeline

The Julia 5.0 implementation of AIControl can be found at https://github.com/suinleelab/AIControl. The binned data for all 455 ENCODE control datasets is available through Google Drive on our project website at https://sites.google.com/a/cs.washington.edu/suinlee/projects/aicontrol.

Supplementary Information

Supplementary Figures

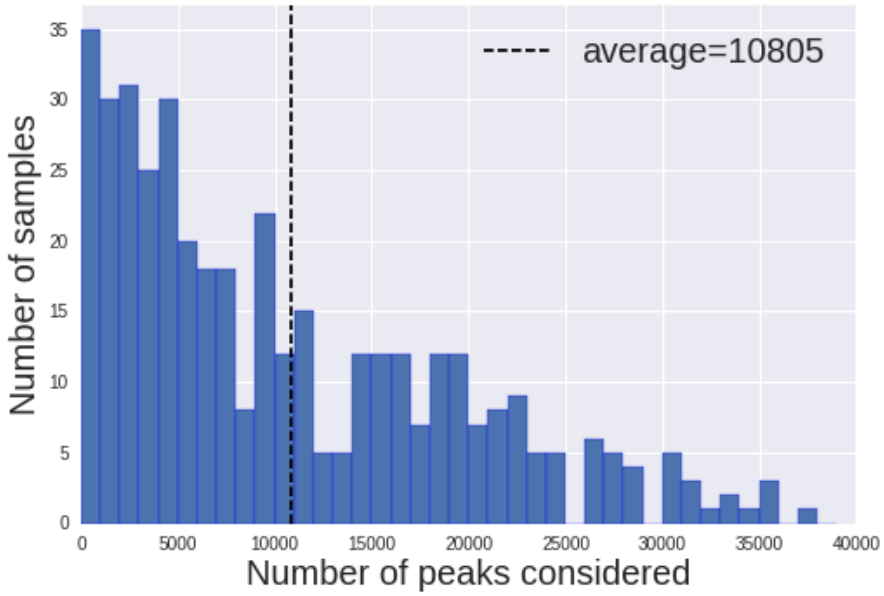


Figure S1: A histogram of the numbers of peaks considered by our evaluation metrics. For each IP dataset, the minimum number of peaks identified among the five peak callers was used for downstream evaluations (see Methods).

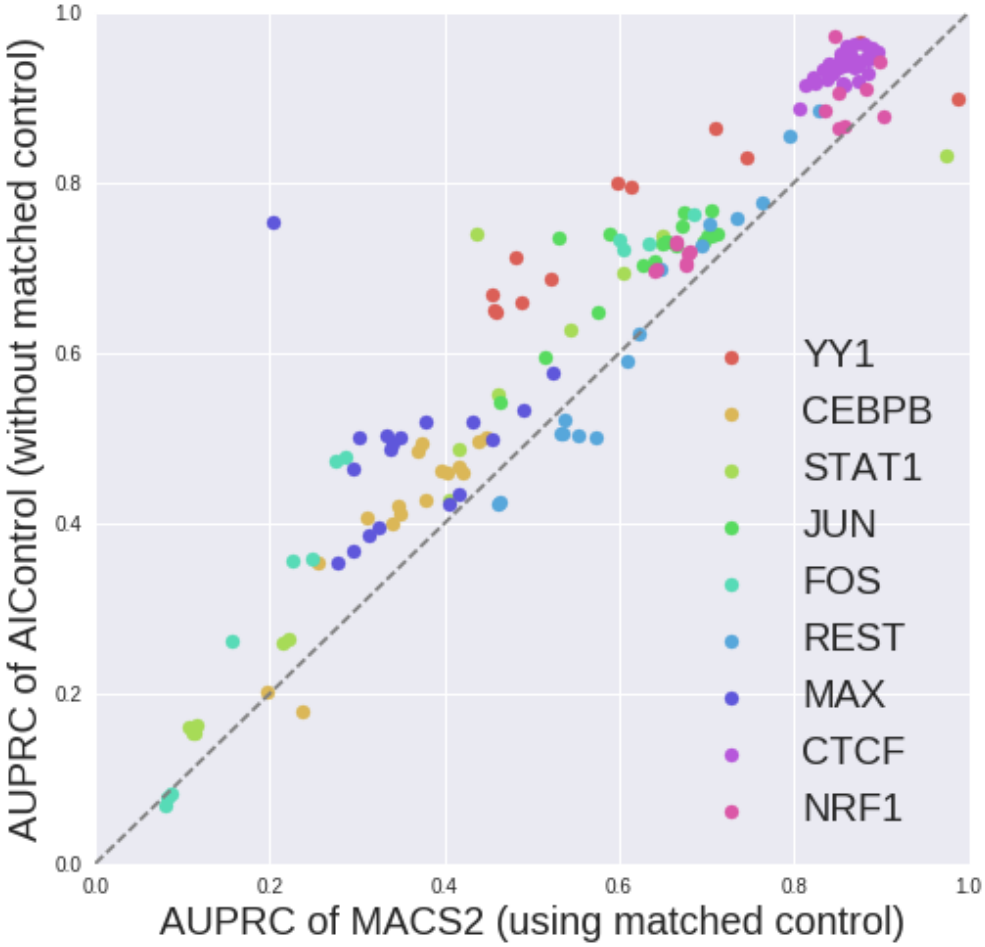


Figure S2: Enrichment of putative binding sites in DNase 1 hypersensitivity regions for ENCODE K562 datasets that target frequently measured regulatory proteins ($N > 9$): CEBPB, STAT1, JUN, REST, MAX, CTCF, NRF1.

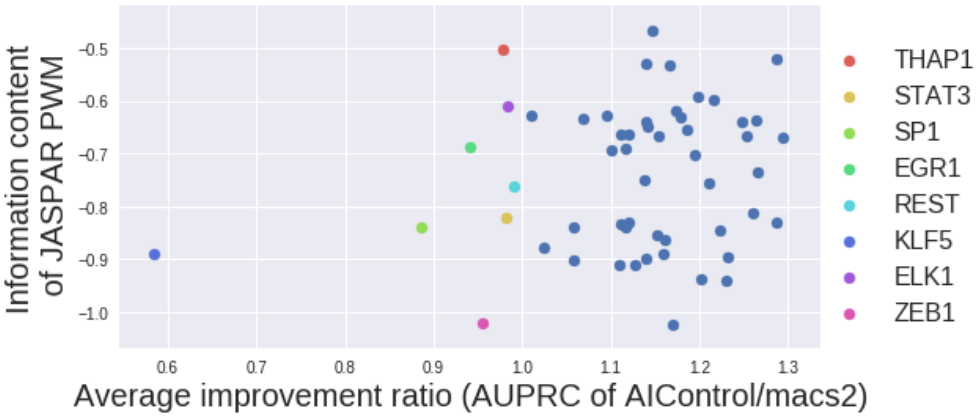


Figure S3: Information content (sum of mean relative entropy per base pair position) of PWMs plotted against the ratio of improvement from MACS2 to AIControl on the AUPRC metric for predicting transcription factor binding events. Each point represents a transcription factor. In particular, AIControl performed worse on the following regulatory factors: THAP1, STAT3, SP1, EGR1, REST, KLF5, ELK1, and ZEB1.

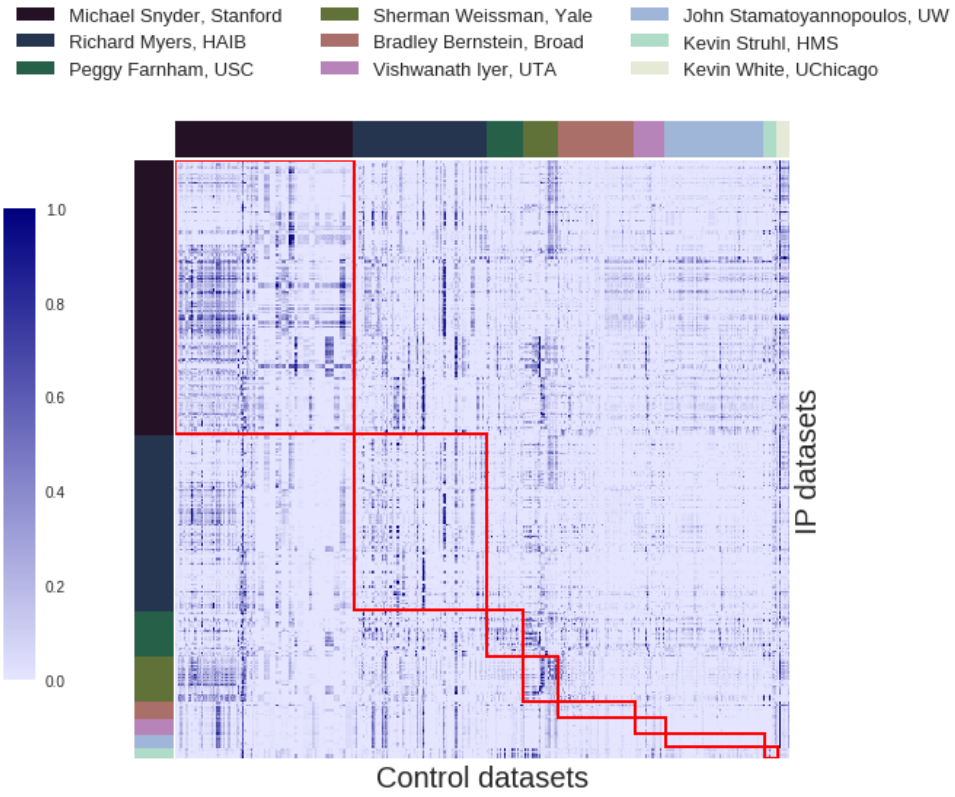


Figure S4: The magnitudes of normalized weights assigned to 455 ENCODE control datasets (columns) for each of the 410 IP datasets (rows) by the AIControl model (Tables S1 and S2). The red rectangles indicate the weights for the control datasets measured in the same lab as the target datasets.

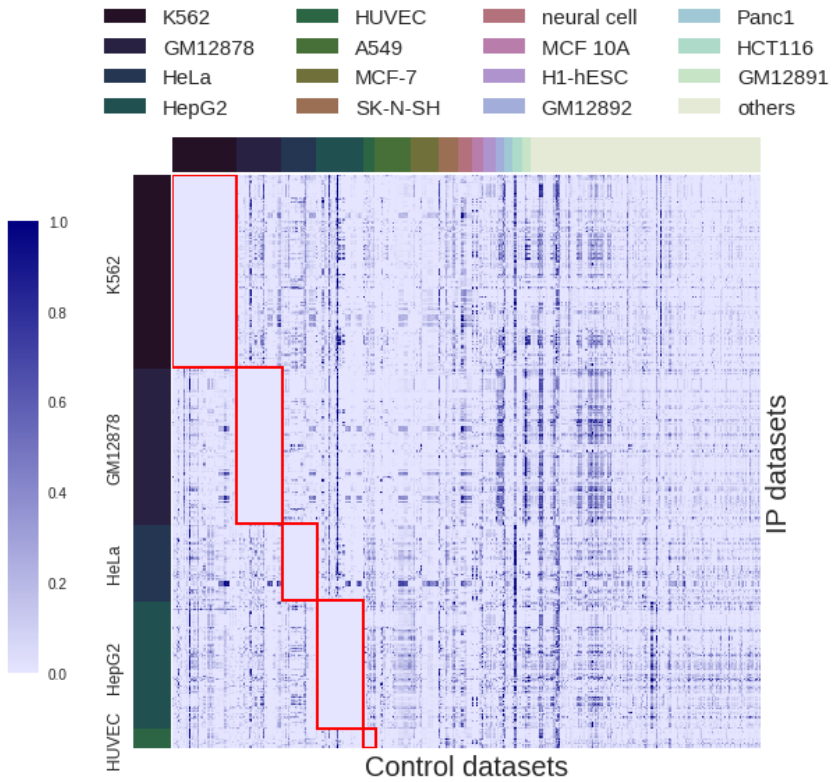


Figure S5: The magnitudes of normalized weights assigned to 455 ENCODE control datasets (columns) for each of the 410 IP datasets (rows) by the AIControl model (Tables S1 and S3) when control datasets from the matched cell line are excluded. The red rectangles indicate the weights of the control samples measured in the same cell line as the IP datasets.

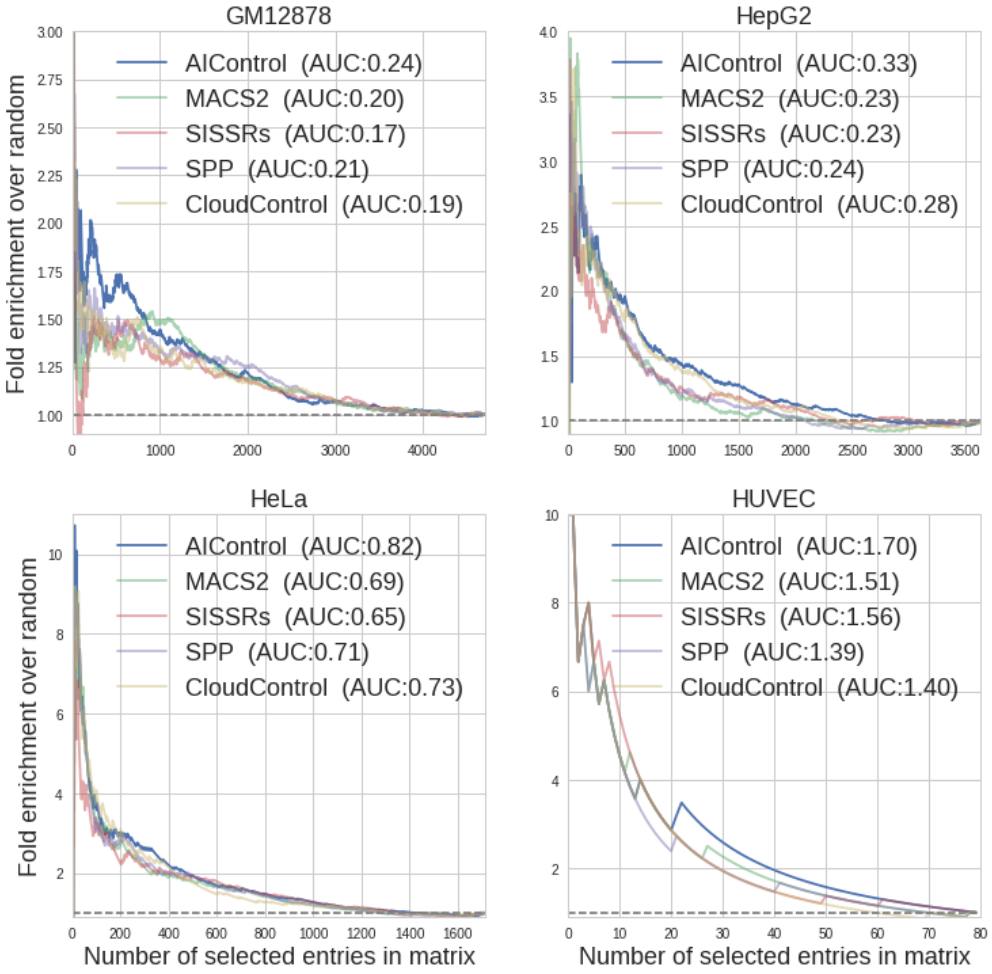


Figure S6: Enrichment of BioGrid-supported interactions between transcription factors in the inverse-correlation networks inferred from peak signals obtained from five different peak calling frameworks (i.e., AIControl, MACS2, SISSRs, SPP, and CloudControl+MACS2) in four ENCODE tier 1 or 2 cell lines (i.e., GM12878, HeLa-S3, HepG2, and HUVEC).

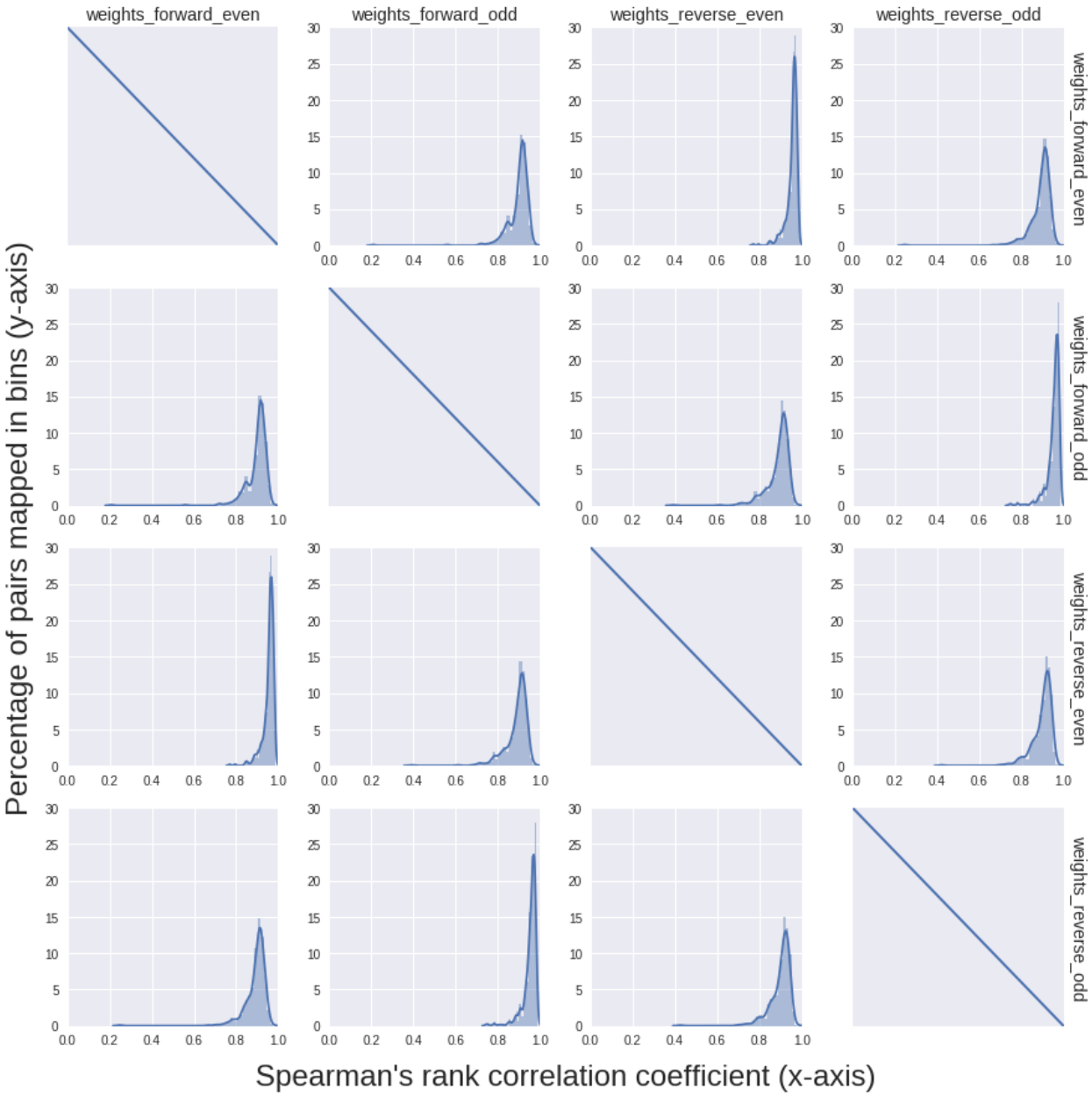


Figure S7: Histograms of Spearman correlation values for all possible pairs among 4 weight vectors learned across all IP datasets. The smooth lines are generated by performing kernel density estimates with bandwidth determined by Lundberg et al (2016) [32].

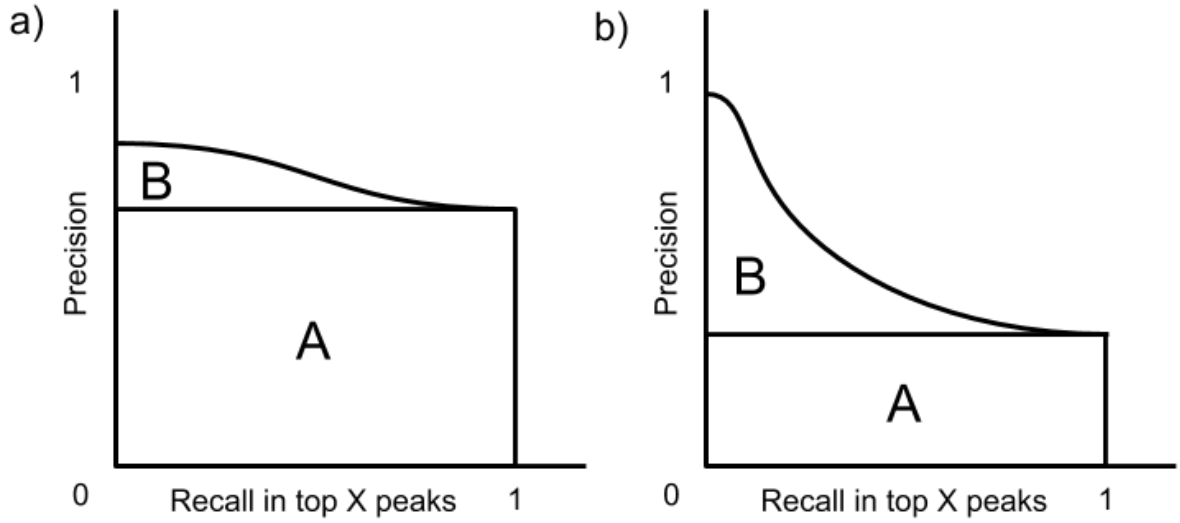


Figure S8: (a) An example PR curve for a peak caller that performs well at selecting true peaks in top n (captured by area A) but poorly at ordering them in top n (captured by area B). (b) An contrasting example PR curve for a peak caller that performs poorly at placing true peaks in top n but well at ordering them in top n. Both area A and B are important for high quality peak calling.

Supplementary Tables

Table S1: Summary of all ENCODE IP datasets processed and evaluated by AIControl broken down by cell line. This summarizes the full listing of all 410 IP ChIP-seq datasets. See Supplementary Data 1 for the full list. The "Datasets" column represents how many datasets were included in the analysis as IP and control experiments, and the "transcription factors" column indicates how many unique transcription factors were included.

Cell line	Datasets	Transcription factors
K562	149	40
GM12878	99	36
HepG2	87	32
HeLa-S3	60	25
HUVEC	15	5
Total	**410**	**54**

Table S2: Summary of all ENCODE control datasets processed and evaluated by AIControl broken down by lab. This summarizes the full listing of all control datasets. See Supplementary Data 1 for the full list.

Lab	Datasets
Michael Snyder, Stanford	132
Richard Myers, HAIB	101
John Stamatoyannopoulos, UW	71
Bradley Bernstein, Broad	55
Peggy Farnham, USC	26
Sherman Weissman, Yale	25
Vishwanath Iyer, UTA	22
Kevin White, UChicago	14
Kevin Struhl, HMS	9
Total	**455**

Table S3: Summary of all ENCODE control datasets processed and evaluated by AIControl broken down by cell line. This summarizes the full listing of all control datasets. Supplementary Data 1 for the full list.

cell line	Datasets	cell line	Datasets
K562	53	**keratinocyte**	3
HepG2	37	**GM18951**	3
GM12878	36	**skeletal muscle myoblast**	2
A549	31	**T47D**	2
HeLa-S3	26	**osteoblast**	2
MCF-7	20	**SK-N-MC**	2
SK-N-SH	15	**IMR-90**	2
H1-hESC	10	**foreskin fibroblast**	2
neural cell	10	**WERI-Rb-1**	2
MCF 10A	9	**Loucy**	2
endothelial cell of umbilical vein	9	**erythroblast**	2
HCT116	7	**Karpas-422**	2
Panc1	7	**astrocyte**	2
GM12892	7	**myotube**	2
GM12891	6	**DOHH2**	2
fibroblast of lung	5	**LNCaP clone FGC**	2
GM18526	5	**Oci-Ly-3**	2
Oci-Ly-7	5	**Oci-Ly-1**	2
GM18505	4	**GM08714**	2
HEK293	4	**WI38**	2
GM19099	4	**mononuclear cell**	2
CD14-positive monocyte	4	**T-cell acute lymphoblastic leukemia**	2
B cell	4	**NB4**	2
GM19193	4	**fibroblast of dermis**	2
GM15510	4	**SUDHL6**	2
PFSK-1	4	**NT2/D1**	2
cardiac mesoderm	4	**GM12875**	2
GM10847	4	**cardiac fibroblast**	2
mammary epithelial cell	3	**GM12864**	2
HL-60	3	**GM12865**	2
ECC-1	3	**Other celltypes**	46
Total			**455**

Table S4: A table that summarizes Figure 2 by calculating the area under the water fall plots. Any performance above baseline positively contributes to the area under curve values. Alternatively, any performance below baseline negatively contributes to the values. The evaluation is constrained in DNase 1 hypersensitivity regions.

cell line	AIControl	MACS2	SISSRs	SPP	CloudControl
K562	**24.37**	1.69	22.18	5.22	9.36
GM12878	**16.37**	3.33	14.30	7.93	6.81
HepG2	**16.48**	1.75	14.00	6.09	6.44
HeLa-S3	**7.42**	0.07	6.56	2.02	1.99
HUVEC	1.75	-0.06	**1.87**	-0.06	0.12

Table S5: A table that summarizes Figure 4 by calculating the area under the water fall plots. Any performance above baseline positively contributes to the area under curve values. Alternatively any performance below baseline negatively contributes to the values. The evaluation is constrained in DNase 1 hypersensitivity regions.

cell line	AIControl	AIControl w/o controls from the same cell line	AIControl w/o controls from the same lab	SISSRs w/o matched controls
K562	24.37	14.34	**24.79**	9.46
GM12878	**16.37**	15.41	15.95	-0.78
HepG2	**16.48**	12.31	15.34	2.76
HeLa-S3	**7.42**	6.71	7.04	2.47
HUVEC	**1.75**	1.74	**1.75**	1.37

Table S6: A list of true PPIs that are supported by the BioGrid Database, and the number of edges recovered by the six peak callers: AIControl (A), MACS2 (B), SISSRs (C), SPP (D), CloudControl+MACS2 (E), MACS2 without a matched control dataset (F). An edge is considered as "recovered", if it is in the 1,371 most significant predictions. The threshold value (1,371) is equal to the number of positive entries in the truth matrix. The PPIs mentioned in the main text are highlighted in bold.

PPI		of edges recovered						PPI		of edges recovered					
TF1	TF2	A	B	C	D	E	F	TF1	TF2	A	B	C	D	E	F
CTCF	ZBTB33	0	1	0	2	1	2	NFYA	NFYB	7	6	7	6	7	6
NFYB	TBP	3	1	1	2	2	1	JUND	MEF2A	1	0	0	1	1	0
MAX	USF1	10	9	8	9	9	9	JUND	YY1	2	3	2	1	0	2
CEBPB	CREB1	1	1	1	1	1	1	SP1	TBP	5	2	7	4	4	1
JUN	**STAT1**	**27**	**14**	**13**	**19**	**12**	**19**	CREB1	JUN	6	8	9	9	10	8
E2F6	**MAX**	**21**	**15**	**15**	**17**	**18**	**16**	SP1	SRF	2	4	4	3	3	3
REST	SP1	6	7	8	9	10	11	SP1	SREBF1	4	7	6	0	0	6
SREBF1	USF1	0	0	0	0	0	0	FOSL1	JUN	22	17	20	19	16	16
ESRRA	SP1	6	5	8	6	7	6	ELF1	NFYA	0	1	0	1	1	0
CEBPB	JUN	6	3	6	3	5	4	**FOS**	**JUN**	**25**	**17**	**17**	**20**	**20**	**16**
JUN	TBP	4	2	0	0	2	0	EGR1	SP1	7	6	4	4	4	2
ELK1	SRF	4	3	4	3	3	3	FOS	JUND	0	0	0	0	0	0
CEBPB	SRF	1	0	0	0	0	0	FOS	MAFF	0	1	0	0	0	0
FOS	USF2	6	4	4	2	5	3	CEBPB	SP1	2	0	2	0	1	0
MAFF	NRF1	0	0	0	0	1	0	SP1	YY1	3	4	3	4	3	2
CEBPB	NRF1	0	1	2	0	0	1	JUN	MEF2A	11	5	5	8	8	6
FOS	TBP	0	2	0	1	0	0	**IRF1**	**STAT1**	**28**	**15**	**18**	**20**	**25**	**21**
MEF2A	TBP	2	0	0	0	0	0	JUN	JUND	18	15	14	14	14	14
CREB1	YY1	2	4	3	4	3	3	FOS	STAT1	3	1	0	1	1	1
NFYB	SP1	4	5	2	4	5	3	NFYA	SRF	0	0	0	2	0	0
CREB1	JUND	4	3	4	3	4	2	FOS	USF1	1	1	0	1	2	1
CEBPB	RUNX1	2	0	3	0	0	0	JUN	RUNX1	2	0	2	0	1	1
IRF1	IRF2	10	9	8	10	9	7	FOSL1	USF1	1	2	1	1	1	0
CEBPB	EGR1	0	0	0	0	0	0	JUN	MAX	10	5	7	6	11	7
EGR1	JUND	2	1	3	2	2	0	ELF1	SP1	2	5	2	5	3	2
GABPA	YY1	3	2	1	4	4	3	ELF1	RUNX1	0	0	0	0	0	0
GATA2	**JUN**	**15**	**7**	**10**	**9**	**11**	**9**	FOS	RUNX1	0	0	0	0	0	0
USF1	USF2	4	3	3	3	3	3	NFYA	SP1	6	5	4	4	5	5
CREB1	NFYA	0	2	1	3	1	1	SREBF1	YY1	2	2	1	1	0	1
MAX	TBP	7	4	4	5	4	4	GABPA	SP1	1	2	2	3	1	2
IRF1	SP2	3	3	2	5	5	5	JUN	NFYA	0	1	0	0	0	1
MAX	YY1	12	11	6	7	7	7	CTCF	YY1	2	3	7	5	2	5
REST	TBP	2	3	2	2	2	2	JUN	SP1	1	1	3	3	2	2

Table S7: A list of protein protein interactions that AIControl uniquely predicted in top 1371. This threshold (1371) is equal to the number of true entries in the truth matrix.

Protein 1	Protein2	Ranking in predictions	Biogrid documentation	Paper reference
NFYB	NR2C2	664	False	
GATA2	USF1	729	False	[7, 26]
STAT1	ZNF263	735	False	
MEF2A	TBP	759	True	
CEBPB	NFYA	805	False	[10]
CREB1	USF1	822	False	[23]
E2F4	MEF2A	897	False	
E2F4	GATA2	941	False	
MEF2A	USF1	996	False	[7]
ZBTB33	ZNF263	1142	False	
ESRRA	STAT1	1278	False	[9]
CEBPB	SRF	1294	True	
MAFF	TBP	1331	False	
JUND	RUNX1	1352	False	[4, 37]
CTCF	GATA2	1360	False	

References

[1] Timothy L Bailey et al. "MEME SUITE: tools for motif discovery and searching". In: *Nucleic acids research* 37.suppl_2 (2009), W202–W208.

[2] Artem Barski et al. "High-resolution profiling of histone methylations in the human genome". In: *Cell* 129.4 (2007), pp. 823–837.

[3] Michael F Berger et al. "The genomic complexity of primary human prostate cancer". In: *Nature* 470.7333 (2011), pp. 214–220.

[4] Marie Bertrand-Philippe et al. "Regulation of tissue inhibitor of metalloproteinase 1 gene transcription by RUNX1 and RUNX2". In: *Journal of Biological Chemistry* 279.23 (2004), pp. 24530–24539.

[5] Daniel Bottomly et al. "Identification of β-catenin binding regions in colon cancer cells using ChIP-Seq". In: *Nucleic acids research* (2010), gkq363.

[6] Brian N Chorley et al. "Identification of novel NRF2-regulated genes by ChIP-Seq: influence on retinoid X receptor alpha". In: *Nucleic acids research* (2012), gks409.

[7] Jessica J Connelly et al. "GATA2 is associated with familial early-onset coronary artery disease". In: *PLoS genetics* 2.8 (2006), e139.

[8] ENCODE Project Consortium et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), pp. 57–74.

[9] Huxing Cui et al. "Eating disorder predisposition is associated with ESRRA and HDAC4 mutations". In: *The Journal of clinical investigation* 123.11 (2013).

[10] Veronica L Dengler, Matthew D Galbraith, and Joaquín M Espinosa. "Transcriptional regulation by hypoxia inducible factors". In: *Critical reviews in biochemistry and molecular biology* 49.1 (2014), pp. 1–15.

[11] Aaron Diaz et al. "Normalization, bias correction, and peak calling for ChIP-seq". In: *Stat Appl Genet Mol Biol* 11.3 (2012), p. 9.

[12] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature methods* 9.3 (2012), pp. 215–216.

[13] Jason Ernst and Manolis Kellis. "Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues". In: *Nature biotechnology* 33.4 (2015), pp. 364–376.

[14] Naozumi Hiranuma, Scott Lundberg, and Su-In Lee. "CloudControl: Leveraging many public ChIP-seq control experiments to better remove background noise". In: *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM. 2016, pp. 191–199.

[15] Michael M Hoffman et al. "Segway: a dynamic Bayesian network method for segmenting genomic data". In: (2009).

[16] David S Johnson et al. "Genome-wide mapping of in vivo protein-DNA interactions". In: *Science* 316.5830 (2007), pp. 1497–1502.

[17] Aziz Khan et al. "JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework". In: *Nucleic acids research* 46.D1 (2017), pp. D260–D266.

[18] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. "Design and analysis of ChIP-seq experiments for DNA-binding proteins". In: *Nature biotechnology* 26.12 (2008), pp. 1351–1359.

[19] Pei Fen Kuan et al. "A statistical framework for the analysis of ChIP-Seq data". In: *Journal of the American Statistical Association* 106.495 (2011), pp. 891–903.

[20] Teemu D Laajala et al. "A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments". In: *BMC genomics* 10.1 (2009), p. 1.

[21] Stephen G Landt et al. "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia". In: *Genome research* 22.9 (2012), pp. 1813–1831.

[22] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4 (2012), pp. 357–359.

[23] Ting-Chi Lu et al. "Retinoic acid utilizes CREB and USF1 in a transcriptional feed-forward loop in order to stimulate MKP1 expression in human immunodeficiency virus-infected podocytes". In: *Molecular and cellular biology* 28.18 (2008), pp. 5785–5794.

[24] Scott M Lundberg et al. "ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data". In: *Genome biology* 17.1 (2016), p. 1.

[25] Tarjei S Mikkelsen et al. "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells". In: *Nature* 448.7153 (2007), pp. 553–560.

[26] Meredith L Moore, Edwards A Park, and Jeanie B McMillin. "Upstream stimulatory factor represses the induction of carnitine palmitoyltransferase-Iβ expression by PGC-1". In: *Journal of Biological Chemistry* 278.19 (2003), pp. 17263–17268.

[27] Ali Mortazavi et al. "Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire". In: *Genome research* 16.10 (2006), pp. 1208–1221.

[28] Leelavati Narlikar and Raja Jothi. "ChIP-Seq data analysis: identification of Protein–DNA binding sites with SISSRs peak-finder". In: *Next Generation Microarray Bioinformatics: Methods and Protocols* (2012), pp. 305–322.

[29] Felicia SL Ng et al. "A graphical model approach visualizes regulatory relationships between genome-wide transcription factor binding profiles". In: *Briefings in Bioinformatics* (2016), bbw102.

[30] Parameswaran Ramachandran, Gareth A Palidwor, and Theodore J Perkins. "BIDCHIPS: bias decomposition and removal from ChIP-seq data clarifies true binding signal and its functional correlates". In: *Epigenetics & chromatin* 8.1 (2015), p. 1.

[31] Dominic Schmidt et al. "Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding". In: *Science* 328.5981 (2010), pp. 1036–1040.

[32] David W Scott. *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons, 2015.

[33] Chris Stark et al. "BioGRID: a general repository for interaction datasets". In: *Nucleic acids research* 34.suppl_1 (2006), pp. D535–D539.

[34] Reuben Thomas et al. "Features that define the best ChIP-seq peak calling algorithms". In: *Briefings in bioinformatics* 18.3 (2016), pp. 441–450.

[35] Robert E Thurman et al. "The accessible chromatin landscape of the human genome". In: *Nature* 489.7414 (2012), p. 75.

[36] Elizabeth G Wilbanks and Marc T Facciotti. “Evaluation of algorithm performance in ChIP-seq peak detection”. In: *PloS one* 5.7 (2010), e11471.

[37] Adam C Wilkinson et al. “RUNX1 is a key target in t (4; 11) leukemias that contributes to gene activation through an AF4-MLL complex interaction”. In: *Cell reports* 3.1 (2013), pp. 116–127.

[38] Yong Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome biology* 9.9 (2008), p. 1.