Published online 14 March 2019

Nucleic Acids Research, 2019, Vol. 47, No. 10 ***e58***
doi: 10.1093/nar/gkz156

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

Naozumi Hiranuma, Scott M. Lundberg and Su-In Lee*

Paul G. Allen School of Computer Science and Engineering, University of Washington, WA, USA, 98195-2350

Received October 16, 2018; Revised February 15, 2019; Editorial Decision February 20, 2019; Accepted February 28, 2019

ABSTRACT

ChIP-seq is a technique to determine binding locations of transcription factors, which remains a central challenge in molecular biology. Current practice is to use a 'control' dataset to remove background signals from a immunoprecipitation (IP) 'target' dataset. We introduce the *AIControl* framework, which eliminates the need to obtain a control dataset and instead identifies binding peaks by estimating the distributions of background signals from many publicly available control ChIP-seq datasets. We thereby avoid the cost of running control experiments while simultaneously increasing the accuracy of binding location identification. Specifically, AIControl can (i) estimate background signals at fine resolution, (ii) systematically weigh the most appropriate control datasets in a data-driven way, (iii) capture sources of potential biases that may be missed by one control dataset and (iv) remove the need for costly and time-consuming control experiments. We applied AIControl to 410 IP datasets in the ENCODE ChIP-seq database, using 440 control datasets from 107 cell types to impute background signal. 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. We also demonstrated that our framework identifies binding sites that recover documented protein interactions more accurately.

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 (1–3). The putative binding sites can then be used in downstream analysis (4,5), for example, to infer interactions among transcription factors (6–8), to semi-automatically annotate genomic regions (9,10) or to identify regulatory patterns that give rise to certain diseases such as cancer (11,12).

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 (13–21). Like other biological assays, ENCODE ChIP-seq guidelines recommend that researchers obtain two ChIP-seq datasets to help separate desirable signals from undesirable biases: (i) an IP (immunoprecipitation) target dataset to capture the actual protein binding signals using immunopreciptation and (ii) a control dataset to capture many potential biases (22). 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 1A), a single-dataset peak calling framework that replaces a control dataset with machine learning by inferring background signals 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 datasets. Many of them have an option to perform single-dataset analysis (i.e. IP dataset only) by determining the structure of background signals from the IP dataset itself; however, it is unlikely to be as accurate as 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 publicly available con-

*To whom correspondence should be addressed. Tel: +1 206 685 1418; Fax: +1 206 543 2969; Email: suinlee@cs.washington.edu

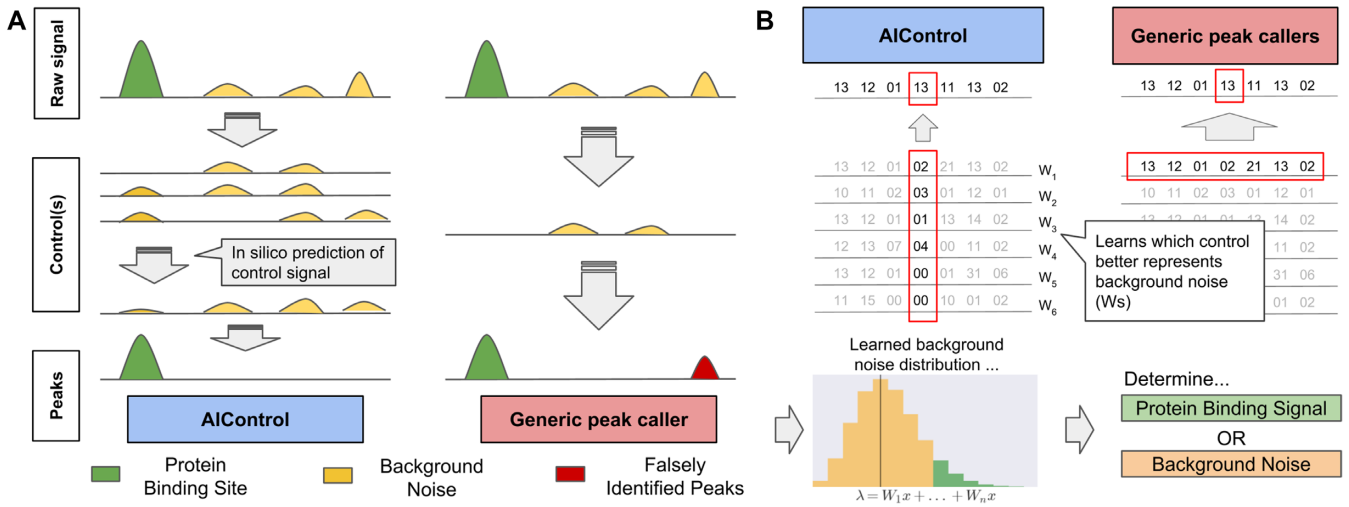

Figure 1. (**A**) An overview of the AIControl approach. A single control dataset may not capture different kinds of biases that give rise to background signal. AIControl more thoroughly removes background signal of ChIP-seq by using a large number of publicly available control ChIP-seq datasets (see 'Materials and methods' section). (**B**) Comparison of AIControl to other peak calling algorithms. (Left) AIControl learns appropriate combinations of publicly available control ChIP-seq datasets to impute background signal 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.

trol datasets (see 'Materials and methods' section for details).

Most popular peak callers—such as Model-based Analysis of ChIP-seq ver 2.0 (MACS2) (13) and Site Identification from Short Sequence Reads (SISSRs) (17)—learn local distributions of read counts from a matched control dataset in a nearby region (Figure 1B, 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 sources of biases, 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 (18).

AIControl's main innovations are 4-fold: (i) 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 nearby regions to learn the position-wise distributions, which may inaccurately estimate local structure of background signal. This feature also offers significant improvements over our previous work, CloudControl (16). CloudControl generates one synthetic control dataset based on publicly available control datasets and identifies peaks by using peak callers that rely on a large window of nearby signals for background estimation. Throughout this paper, we show that AIControl significantly improves peak calling quality relative to CloudControl. (ii) 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. (iii) Because AIControl integrates many control datasets, it can potentially capture more sources of biases compared to existing methods that use only one control (e.g. MACS2 and SISSRs). Most confounders—such as GC content and mappability—are likely present in some of the control datasets AIControl incorporates. See 'Modeling background signal' in 'Materials and methods' section for our mathematical formulation. (iv) AIControl does not need a matched control dataset. We incorporate 440 control ChIP-seq datasets from 107 cell types in the ENCODE database. By inferring local structure of background signal from the large amount of publicly available data, AIControl can identify peaks even in cell types without any previously measured control datasets. We demonstrate that our framework intelligently uses existing control datasets to estimate background distributions for IP datasets in unseen cell types in a cross-cell-type setup.

We evaluated the AIControl framework on 410 ChIP-seq 'IP datasets' available in the ENCODE database (23) (Supplementary Table S1) using 440 ChIP-seq 'control datasets' (Supplementary Tables S2 and S3). The IP datasets span across five cell types: K562, GM12878, HepG2, HeLa-S3 and HUVEC. Every cell type except for HUVEC is a cell line, and HUVEC is a primary cell (endothelial cell of umbilical vein). Results show the following: (i) AIControl outperformed other peak callers on identifying putative protein-binding sites based on sequence-based motifs (Figure 2). All competing peak callers used matching pairs of IP/control datasets, whereas AIControl did not—it used only IP datasets (no matching control) and publicly available control datasets. AIControl predicted putative binding sites well even when all control datasets from the same cell type were removed, which suggests that it reliably estimates background signals in a cross-cell-type manner when ChIP-seq is performed on an unseen cell type (Figure 4). (ii) PPIs were more accurately predicted from peaks called by

AIControl than from those called by all other methods in all tested cell types (Figure 5). (iii) Peaks identified by AIControl showed superior performance in motif enrichment and PPI recovery tasks when they were processed with the irreproducible discovery rate (IDR) pipeline (Supplementary Figure S1). (iv) AIControl exhibited strong performance on datasets that are not part of the ENCODE database (Figure 6). Our findings suggest that AIControl can remove the time and cost of running control experiments while simultaneously identifying binding site locations of transcription factors accurately.

MATERIALS AND METHODS

Methods

Modeling background signal. AIControl models background signals across the human genome as a linear combination of multiple different sources of confounding biases. 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 bias 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 $x_1, ..., x_n$:

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

$$\hat{y}_i = y_i + \epsilon_i \quad (2)$$

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 bias 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 its background signal $\hat{y}_t$, which is modeled as a weighted linear combination of $x_1, ..., x_n$ with a weight vector $\hat{w}_t \in \mathbb{R}^n$:

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

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, 440 ENCODE control datasets, summarized in Supplementary Tables S2 and S3) 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 weight vectors $w_1, ...w_m$:

$$\hat{w}_t = a_1w_1 + a_2w_2 + ... + a_mw_m. \quad (4)$$

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

$$\hat{y}_t = (a_1w_{11} + ... + a_mw_{m1}) \cdot x_1 + ... + (a_1w_{1n} + ... + a_mw_{mn}) \cdot x_n + \epsilon_t \quad (5)$$

$$= a_1 \cdot (w_{11}x_1 + ... + w_{1n}x_n) + ... + a_m \cdot (w_{m1}x_1 + ... + w_{mn}x_n) + \epsilon_t \quad (6)$$

$$= a_1y_1 + a_2y_2 + ... + a_my_m + \epsilon_t \quad (7)$$

$$= a_1\hat{y}_1 + a_2\hat{y}_2 + ... + a_m\hat{y}_m, \quad (8)$$

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-signal 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.

$$o_t = \mathsf{ProteinBindingSignal} + \mathsf{ReproducibleBackgroundSignal} + \mathsf{IrreproducibleNoise} \quad (9)$$

$$= p_t + y_t + \epsilon_t \quad (10)$$

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 440, which means that this problem is far from high-dimensional and unlikely to overfit.

Computing coefficients. 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:

$$\underset{a}{\operatorname{argmin}} ||o_t - Ya||_2^2 + \lambda||a||_2^2. \quad (11)$$

Here, Y is a g by m ($= 440$) 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:

$$\hat{a} = (Y^TY + \lambda I)^{-1}Y^To_t. \quad (12)$$

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 440 and g is 30 million (under the default setting where the size of bins is 100 base pairs (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. `ReproducibleBackgroundSignal` 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. Training on odd positions and predicting on even positions (and vice versa) are designed to further prevent any possible overfitting. Spearman's correlation values of learned coefficients are shown in Supplementary Figure S2. These values are generally above 0.8 for any pair in the same IP dataset, showing that learned sets of weights are consistent among forward, reverse, odd- and even-positioned data.

It is important to note that we need not to recompute $Y^T Y \in \mathbb{R}^{440\times440}$ for different IP datasets, because it remains constant when 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 $Y\hat{a}$.

Identifying peaks. Commonly used peak calling approaches identify a peak based on how far 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 (13,17). Usually, null distributions are semi-locally fit to signals from nearby regions (5,000–10,000 bps) in a matched control dataset.

Like many 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 of the null background distribution for the read count observed at the ith position of genome, c_{ti} ,in the target IP 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, $\hat{y}_{1i}, ..., \hat{y}_{mi}$, which are weighted differently with corresponding weights $a_1, ..., a_m$. The use of m control datasets (not just one matched control) lets us learn a higher resolution background distribution (Figure 1B). Finally, we introduce the minimum base count of 1 read to prevent $\hat{y}_t$ from being too small or 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 choose not to use our standard control dataset release and so do not have a sufficiently large number of background controls m.

We then calculate the P-value and fold enrichment of the observed count at each genomic bin based on the learned null background distribution and background count. 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 (and vice versa) for each forward and reverse strand. We then slide the locations of the P-values and enrichment values by $\frac{d}{2}$ and $-\frac{d}{2}$, for forward and reverse signals, respectively. d is defined as the expected distance between forward and reverse peaks; it is automatically estimated in our framework (see below). Finally, the smaller negative log10 P-value and fold enrichment of read counts between the slid forward and reverse signals at every position is output as a peak signal. This last step ensures that peaks have bimodal shapes as expected for transcription factor binding signals (13).

Estimating distance between forward and reverse peaks d. AIControl automatically estimates the distance between forward and reverse peaks similar to other peak callers. Specifically, for each dataset, we find the sliding distance d that minimizes the disagreement between the forward and reverse mapped reads. In particular, the disagreement is defined as follows:

$$\mathsf{disagreement}_d = \frac{1}{N}\sum_{i=0}^{N} |\mathsf{SlidForwardReads}_i - \mathsf{ReverseReads}_i|. \quad (13)$$

N is the number of bins in the hg38 genome, which is approximately 30 million with a bin size of 100 bps. We find d that minimizes the disagreement value with brute force search between $d = 0$ and $d = 400$. With a default binning size of 100 bps, there are only four options for d, and it is relatively fast to find the optimal d. The summary of d estimation for all 410 tested ENCODE IP datasets is shown in Supplementary Figure S3.

Merging contiguous bins with significant binding signal. AIControl assigns a P-value and fold enrichment of binding signal to each 100 bp genomic bin. As an optional post-processing step, the current implementation of AIControl can merge contiguous bins that have more significant binding signal than threshold (default is negative log10 P-value of 1.5, approximately P-value of 0.03) by taking the maximum P-value and fold enrichment values among them. The resulting peaks are output in a .narrowPeak format.

Data processing

Aligning BAM files. We describe BAM files used in this project in our prior work on ChromNet (6). Specifically, the raw FASTQ files were downloaded from the ENCODE database and were mapped to the UCSC hg38 genome with BOWTIE2 to ensure an uniform processing pipeline (24). We provide the full list of ENCODE experimental IDs used in this project in Supplementary Data S1.

Calling peaks with other methods. The version of MACS2 used in this paper was `MACS2 2.1.0.20150731` (13). The peaks were called with the following command: `'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 using the `'find.binding.positions'` function in its R package (15). Additionally, we downloaded SPP peaks from the ENCODE portal if they were available (we call it 'SPP-ENCODE'). For SISSRs, we used v1.4, and the peaks were also called with a P-value threshold of 0.05 with the following command: `'sissrs.pl -i chipseq_dataset.bed -b matched_control.bed -p 0.05 -s 3209286105'` (17). The peaks from CloudControl were obtained in conjunction with MACS2 using the same parameters as above (16). All resulting peak files can be viewed in Google Drive through our GitHub repository under the 'Paper' section (https://github.com/suinleelab/AIControl.jl).

Obtaining peaks that are optimally controlled with IDR. The ENCODE official pipeline for processing biological

replicate samples is to use SPP and IDR in combination (25). We also investigated the performance of peak callers in combination with the IDR process (Supplementary Figure S1). In particular, for peaks processed with SPP, we downloaded peak files tagged as 'optimal idr thresholded peak' from the portal website of ENCODE (23). If they were available on the hg19 genome, we used the UCSC liftover tool to convert peak locations from the hg19 to hg38 genome. For other peak callers (i.e. MACS2, SISSRs and AIControl), we used the Python implementation of idr (https://github.com/kundajelab/idr) to re-order peaks for each pair of biological replicates. We believe that the significant value thresholds we used (0.05 for MACS2 and SISSRs, and 0.03 for AIControl) are lenient enough to capture both reproducible and irreproducible signals that are required for the IDR process.

Storing large matrix of control signals efficiently. One of the challenges in implementing AIControl in a user-friendly manner is to find an efficient way of storing a massive amount control datasets. In particular, we have 440 ChIP-seq control datasets, and each of them is represented as a 30 million long vector, which stores read count for every 100 bp bin. Collectively, the control datasets are represented as a sparse, non-negative matrix of size 440 by 30 million. For this project, we developed our own file format to store the large matrix. First three 8 bit chunks of the file encode the following three parameters: (i) a number of control datasets (i.e. width of the matrix), (ii) a maximum possible value stored in the matrix (100, if duplicate reads are removed) and (iii) a data type (i.e. UInt8 or UInt16). Subsequent 8 or 16 bits, depending on the data type, are used for indicating an actual value in the current entry of matrix, if it is less than the predefined maximum value. Otherwise, it is used for indicating how many entries (value − the maximum value) to skip column-wise by filling in 0s. With this file format, we can compress all 440 control datasets to 4.6 GB. Given that typical BAM files are about 0.5–3.0 GB, we believe this makes our standard control dataset package compact enough for users to download.

Analysis pipeline

Standardizing peak signals. Different peak calling algorithms identify different numbers of peaks at a given significance value threshold and generate peaks with different widths. To eliminate the possibility that these differences in peak numbers or widths create biases, we standardized peak signals for each dataset as follows. (i) We bin peak signals by 1,000 bp windows. This creates vectors where each entry corresponds to the peak with the largest ranking measure (i.e. negative log10 *P*-values or signal values) among the peaks that fall into the corresponding bin. We thereby standardize peak width to 1,000 bps across all methods. (ii) For each dataset, we use the top n binned peaks for all peak callers, where n is the minimum number of binned peaks from all tested peak callers at their corresponding significance thresholds (see above). Choice of ranking measure matters. We used column 7 (signal value) of the narrowPeak format for SPP and AIControl, and column 8 (*P*-value) for MACS2. For SISSRs, its output does not follow the narrowPeak format, but we use *P*-values associated with peaks as the ranking measure. This process, which standardizes the number of peaks identified by different peak callers, results in an average of 21,470 genome-wide peaks per dataset (Supplementary Figure S4).

Evaluating motif enrichment. We applied AIControl to 410 ChIP-seq IP 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 (26). We then used FIMO from the MEME software to search for the putative binding sites at the *P*-value threshold of 10^{-5} (27). The idea is that correctly identified peaks are likely in a region that contains the corresponding motif. Of course, motif enrichment alone may not be a reliable measure; thus in addition to examining the whole genome, we also focus on the regions where transcription factor binding occurs relatively more often. For instance, studies show that 98.5% of the transcription factor binding sites are positioned in DNase 1 hypersensitivity (DHS) regions (28).

Therefore, to increase the reliability of motif-based evaluation criteria, we focused our analysis on the following four regions when we performed motif enrichment-based evaluation: (i) the whole genome, (ii) DNase 1 hypersensitivity regions, (iii) regions that are 5,000 up- and downstream of the start sites of protein coding genes and (iv) regions that have more than 50% GC content. The DHS signals were downloaded from the portal website of the Roadmap Epigenomics project (29). The regions proximal to protein coding genes were obtained through BioMart (30). 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 each of aforementioned regions using varying thresholds of the significance of binned peaks. This led to a 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 the performance of peak calling methods. 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 type for (i) the whole genome (Figure 2 and Supplementary Figure S5), (ii) DHS regions (Supplementary Figure S6), (iii) up/downstream regions of protein-coding genes (Supplementary Figure S7) and (iv) high GC content regions (Supplementary Figure S8). For example, in Figures 2B and 4, each colored line 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 the IP datasets, which are sorted independently for different peak calling methods based on their y-axis values. We removed datapoints if any peak caller called <20 peaks to ensure the stability of AUPRC values (Supplementary Figure S4B). Peak callers with larger areas under the colored

line and above the dotted line, on average, better identified peaks supported by sequence motifs (Supplementary Tables S5 and S9). The AUPRC value for each peak caller on each dataset is shown in Supplementary Data S2 for the whole genome, Supplementary Data S3 for the DHS regions, Supplementary Data S4 for the gene proximal regions and Supplementary Data S5 for the GC rich regions.

Using area under PR curve of n *most significant peaks as an evaluation metric.* As described in 'Standardizing peak signals' section above, we analyze only the n most significant binned peaks, where n is determined by the minimum number of binned peaks called among all peak callers. We then generated the precision-recall curve for predicting the 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. Supplementary Figure S9 explains what the AUPRC metric captures for peak callers with different behaviors. Supplementary Figure S9A 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). Supplementary Figure S9B 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 500–15,000 most significant peaks depending on transcription factors. Our average choice for n is 21,470, much higher than the widely used threshold (Supplementary Figure S4A). This PR curve-based approach is equivalent to testing peak calling quality (precision and recall) at every possible rank threshold up to the minimum number of peak called, and we believe that this is better than using an arbitrary threshold.

Obtaining and evaluating on protein–protein interaction (PPI) matrix. The validated PPI interactions we used for evaluation were downloaded from the BioGrid website by 2018/2/5 (31). 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 type as follows. First, for each peak calling method, the inverse correlation matrix from all n IP datasets in the cell type of interest was computed using standardized peak signals (see 'Standardizing peak signals') after binarization. This resulted in a matrix of size n by n. 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 (Figure 5 and Supplementary Figure S10), like we did previously (6). Fold enrichment is defined as follows for given number of selected predicted interactions (x-axis of Figure 5 and Supplementary Figure S10):

$$\text{fold enrichment} = \frac{\text{\# of BioGrid-validated edges}}{\text{expected \# of validated by random}}. \quad (14)$$

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) (Figure 5 and Supplementary Figure S10). A larger area under the fold enrichment curve indicates superior performance similar to PR curves.

Measuring consistency among pairs 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 (31). The number of shared 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.

SOFTWARE AVAILABILITY

The Julia 1.0 implementation of the AIControl software and a thorough step-by-step guideline can be found at our GitHub repository: `https://github.com/suinleelab/AIControl.jl`. In the following four subsections, we described in detail some of the important steps to install and run our implementation and our user experience survey.

Converting an `.fastq` file to a sorted `.bam` file and aligning it to hg38

Users must align their input `.fastq` files to the hg38 genome from the UCSC repository, which can be found at `http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz` using `bowtie2` (24). Unlike other peaking callers, the unique core idea of AIControl is to leverage all available control datasets in public. This requires all data—both target ChIP-seq and public control datasets—to be mapped to the exact same reference genome. Currently, the control datasets that we provide are mapped to the UCSC hg38 genome. Therefore, for instance, if the target ChIP-seq dataset is mapped to a slightly different version of the hg38 genome, the AIControl pipeline will report an error. If users start with a `.bam` file that is already mapped, the recommended way of resolving this error is to use `bedtools`, which provides a way to convert `.bam` files back to `.fastq` files and realign to the correct version of the hg38. (see 'Step 3.1' on our GitHub repository at `https://github.com/suinleelab/AIControl.jl` for specific commands). The direct output of `bowtie2` is a `.sam` file. Users need to use `samtools` to convert it to a `.bam` file and sort it in lexicographical order (see 'Step 1' on our GitHub repository at `https:`

//github.com/suinleelab/AIControl.jl for a piped single-line command for alignment).

Downloading compressed control data files

AIControl requires users to download binned control datasets on their local systems. The compressed control data files for all 440 ENCODE control datasets are available through an FTP server for our project at https://dada.cs.washington.edu/aicontrol/. We have two separate files for signals mapped to forward and reverse strands. These files are 4.6 GB in total, and they occupy 13 GB when decompressed (see 'Control data files required for AIControl' on our GitHub repository at https://github.com/suinleelab/AIControl.jl).

Running the AIControl script

We generated a julia script file, `aicontrolScript.jl`, which performs the AIControl framework, which takes in the sorted `.bam` file and outputs a `.narrowPeak` file (see 'Step 3' on our GitHub repository at https://github.com/suinleelab/AIControl.jl for how exactly to execute it). For a full commentary of our major updates in Julia implementation, please refer to the 'Major Updates' section of our GitHub repository.

We provided error messages for several types of errors that could occur frequently in practice. These errors include (i) 'input ChIP-seq file missing error', (ii) 'compressed control file missing error' and (iii) 'genome mismatch error'. Specifically, for the 'genome mismatch error', we made an optional step that explains why that happens and how to resolve it in our GitHub repository (see 'Step 3.1'). If you have a problem running the AIControl pipeline, please refer to the 'Issues' page of our GitHub repository at https://github.com/suinleelab/AIControl.jl or e-mail suinlee@cs.washington.edu.

Verifying the AIControl pipeline

We conducted a survey for user experience on AIControl by asking 11 researchers in the computational biology field to run the AIControl pipeline from scratch, starting from Julia installation. All users were able to install and run AIControl successfully. We asked each of them (i) whether our step-by-step GitHub guideline was clear, (ii) whether they were able to install Julia with no issue when they followed just the guideline, (iii) whether they were able to run AIControl with no issue and (iv) which system they used to install and run AIControl on.

We verified that our pipeline can be installed and run on the following operating systems: CentOS 7, Ubuntu 18.04, Arch Linux, macOS Sierra, macOS Mojave and Windows 8. Although AIControl has been verified on many widely used systems (macOS, Windows and Linux), we recommend that users run AIControl on Unix-based systems (i.e. macOS or Linux), because the other peripheral software, such as `samtools` or `bowtie2`, are easier to install there through `conda` and `bioconda`.

RESULTS

Peaks identified by AIControl are more enriched for binding 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 (13), SISSRs (17), SPP (15) and MACS2 + CloudControl (16). To define putative binding sites without using ChIP-seq data, we identified sequence motifs using FIMO from the MEME tool (27) and position weight matrices (PWMs) from the JASPAR database (26) (see 'Materials and methods' section). 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 (20,21,32). To evaluate the enrichment for putative binding sites, we used ranking measures (negative log10 *P*-values or signal values, see 'Standardizing peak signals' section) of peaks to predict the presence of putative binding sites and measured the area under the precision-recall curves (AUPRCs) in the following four genomic regions: (i) the whole genome, (ii) DNase 1 hypersensitivity regions (DHS), (iii) 5,000 bps up- and downstream of protein coding gene start sites, (iv) regions with more than 50% GC content. 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 'Materials and methods' section). This process prevents peak callers that identify more peaks at a given threshold from having an unfair advantage. For the analyses across the whole genome, this resulted in an average n of 21,470 peaks per IP dataset for the entire genome (Supplementary Figure S4).

Figure 2B compares the AUPRCs across the whole genome achieved by the five peak callers for 410 IP datasets across five different cell types: K562 (149), GM12878 (99), HepG2 (87), HeLa-S3 (60) and HUVEC (15). AIControl yielded better fold improvements of AUPRCs over that of baseline than the other peak callers (P-value < 0.0001 with the Wilcoxon signed-rank test on AIControl versus SISSRs on matched pairs of fold improvements). When the results are viewed separately for the five cell types, AIControl achieves the best performance in all cell types except for HUVEC (Supplementary Figure S5 and Supplementary Table S5). 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. AIControl continues to perform better on the motif enrichment task even when the analyses are restricted the aforementioned regions (2)–(4) of the genome (Supplementary Tables S6, S7, S8 and Supplementary Figures S6, S7, S8). 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 (33). All five peak callers

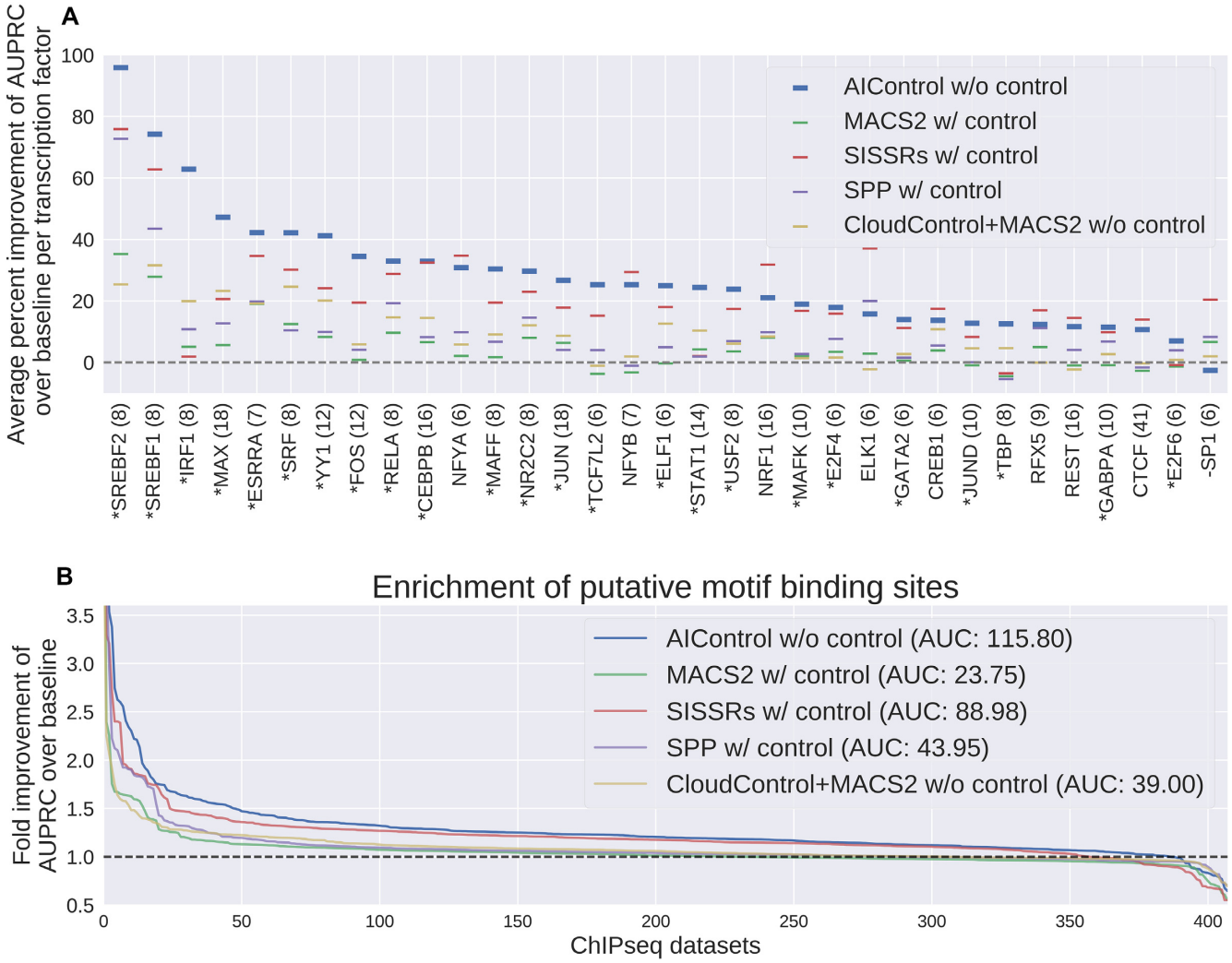

Figure 2. (**A**) Average percent improvement of area under precision recall curves (AUPRCs) per transcription factor. Peaks were identified using: (i) AIControl w/o control, (ii) MACS2 w/control, (iii) SISSRSs w/control, (iv) SPP w/control and (v) CloudControl + MACS2 w/o control. Only the transcription factors that were measured more than five times are shown (the number of measurements shown in parenthesis), and they are ordered by AIControl's performance. The transcription factors that AIControl performed the best on are shown with an asterisk (24 out of 33). The ones on which AIControl performed worst are shown with a minus sign (1 out of 33). (**B**) Relative performance of five peak calling methods compared to MACS2 without using a matched control dataset as a baseline (dotted line). 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 ranking measures associated with the peaks over the baseline (i.e. MACS2 without using a matched control dataset) across the whole genome. The *x*-axis shows the all 410 ENCODE ChIP IP datasets ordered by the fold improvement (*y*-axis) across all tested cell types; 149, 99, 60, 87 and 15 for K562, GM12878, HeLa-S3, HepG2 and HUVEC, respectively. Note that the ordering of datasets is different for each peak caller. Area between each line and the dotted baseline is shown in parenthesis.

identified all eight qPCR-confirmed binding locations on chromosome 1.

Datasets that target certain transcription factors yielded more performance improvements than others. Figure 2A shows the mean percent improvement of the five peak callers—three of which use matched control datasets—over baseline (i.e. MACS2 without matched control) for transcription factors that are measured more than five times across all 410 IP datasets. The AIControl framework outperformed all other peak callers in 24 out of 33 transcription factors without needing a matched control experiment. In particular, our framework exhibited major average improvements on transcription factors MAX, JUN and STAT1, over the best performing peak callers, while it showed decreased performance on SP1. Arvey *et al.* (34) examined cell-type-specific binding of patterns of 15 transcription factors from ENCODE data between K562 and GM12878. In the study, they observed more cell-type-specific peaks from transcription factors MAX, JUN, JUND, YY1 and SRF. In Figure 2A, AIControl performs the best on all of these factors, suggesting that our framework is able to well identify binding sites of transcription factors that exhibit differential binding patterns depending on target cell types.

We also investigated whether our analysis is affected by the quality of putative true binding sites by checking the relationship between the relative performance of AIControl over MACS2 across the whole genome and the information

contents of the JASPAR PWMs. However, we did not find any significant correlation ($r = 0.02$, P-value $= 0.84$, Supplementary Figure S11).

Interestingly, but not surprisingly, the more publicly available control datasets are incorporated, the better is the performance of AIControl. We picked 36 datasets that are associated with transcription factors measured more than 10 times across all tested cell types (i.e. YY1, CEBPB, STAT1, JUN, FOS, REST, MAX, CTCF and NRF1 from each tested cell type; some TF/cell-type pairs were missing due to data availability). When averaged over many random subsets of different sizes, we observed that the AIControl framework experiences monotonic but diminishing increase of the performance from no control to 440 control sets, peaking at 29% improvement with all 440 control sets (Supplementary Figure S12). While approximately 86% of the overall improvement is made with 200 control datasets, the additional 240 datasets still contribute positively (P-value < 0.001 for Wilcoxon signed-rank test, comparing the performance at 200 and 440 control pairwise). This result suggests that the improved performance of AIControl in fact depends on the inclusion of public control datasets.

We recognize that our own pipeline for SPP is different compared to that of the ENCODE consortium since it has extra read/peak filtering steps. Naturally, this results in slightly different sets of peaks. Therefore, we decided to compare AIControl directly against the ENCODE peaks (SPP-ENCODE) in order to assure that AIControl still holds an advantage. We downloaded a peak file from the ENCODE portal for each dataset, if it was available. Supplementary Figure S13 shows that SPP-ENCODE outperforms SPP potentially due to the extra filtering steps, and more importantly, AIControl outperforms SPP-ENCODE. We note that we did not include SPP-ENCODE to Figure 2 because the SPP-ENCODE data are available in only a subset of IP datasets (i.e. 123 out of 149 for K562; 65 out of 99 for GM12878; 46 out of 87 for HepG2; 18 out of 60 for HeLa-S3; 7 out of 15 for HUVEC). The IDs of SPP-ENCODE peak files downloaded are listed in Supplementary Data S6.

AIControl coefficients reflect cell type specificity but not lab specificity

AIControl learns the weights of contributions by all 440 ENCODE control datasets to estimate the background ChIP-seq signals for each IP dataset (Figure 1; also see 'Materials and methods' section). Figure 3 shows the magnitude of weights assigned to all 440 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 type (Figure 3). This is expected because known factors for background signals, such as sonication bias and DNA acid isolation, depend on cell types. On the other hand, when we sort the rows and columns based on lab, we see less significant pattern, except for the datasets from the Weissman lab (Supplementary Figure S14). This suggests that lab-specific batch effects are less significant than cell type-specific effects on the ENCODE ChIP-seq data. Although the control datasets from the same cell type 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 types. 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 types to identify peaks more accurately in GM12878.

AIControl retains its performance in a cross-cell-type setting where control datasets from the same cell types or the same labs are excluded from the background set

The results described in the previous subsection indicate that AIControl leverages information about background signal from biologically similar cell types. A natural question is whether AIControl can correctly identify peaks in an IP dataset from an unseen cell type that is not included in the public control datasets that AIControl uses. This tests AIControl's ability to estimate, in a cross-cell-type manner, background signals in an unknown cell type from background signals in known cell types. 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 settings: (i) AIControl with all 440 control datasets except for matched controls, (ii) AIControl without control datasets from the same cell type as the IP dataset, (iii) AIControl without control datasets from the same lab as the IP dataset, (iv) AIControl without control datasets from the same lab or the same cell type as the IP dataset, (v) SISSRs with matched control datasets and (vi) SISSRs without matched control datasets. We chose SISSRs because it is the best competitor in terms of identifying presence of motif sequences (Figure 2).

Figure 4 shows that AIControl with different patterns of excluding control datasets are still able to outperform SISSRs with matched control datasets in all cell types except for HUVEC, for which we only have a small number of datasets ($n = 15$). Notably, the exclusion of control datasets from the same cell type (i.e. settings (ii) and (iv)) has a larger impact on the performance than the lab-based exclusion (i.e. setting (iii)). This indicates that lab specific biases are less significant and easier to learn in a cross-lab setting. For settings (ii) and (iv), we observed the largest decrease in the performance of AIControl in K562, followed by HepG2 and HeLa-S3. On the other hand, in GM12878, AIControl was able to leverage information from other B-lymphocyte cell lines (i.e. GM12892) (Supplementary Figure S15).

The decline in K562, followed by HepG2 and HeLa-S3, has two likely reasons. First, the largest number (48 of 440, or 10.9%) of ENCODE control datasets are from the K562 cell line (Supplementary Table S3). Second, the structure of background signals in K562, HepG2 and HeLa-S3 cell types may be unique because of their abnormal karyotypes. Regions with multiplication, deletion and copy number variations that are not documented in the reference genome can display signals that are locally proportional to alterations in the abnormal karyotype. This makes it harder for AIControl to estimate background signals in abnormal regions without having access to the controls from the same

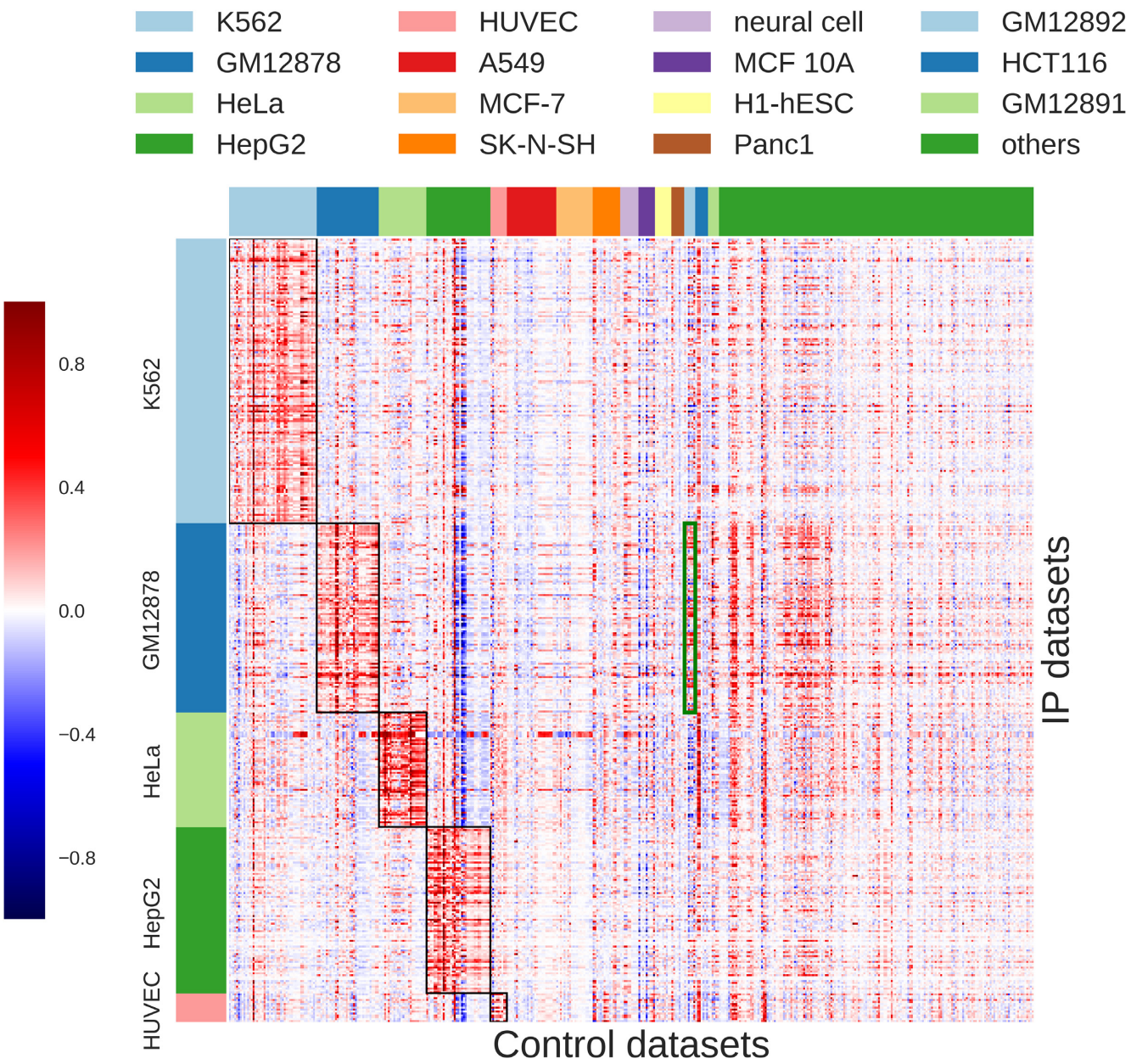

Figure 3. Normalized weights that the AIControl framework assigns to 440 ENCODE control datasets (columns) for each of the 410 IP datasets (rows) (Supplementary Tables S1 and S3). The black rectangles indicate the weights of the control datasets measured in the same cell type as the IP datasets. The 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 types.

cell type. On the other hand, in GM12878 and HUVEC, which is known to have a normal karyotype, the performance of AIControl did not drop even without having access to control datasets from the same cell types.

We further investigated whether including additional control datasets with abnormal karyotypes hurt the performance of AIControl because they are less informative in certain regions, or including control datasets from just normal karyotypes is better even if it results in a significantly reduced number of datasets. For the GM12878 datasets, we compared the performance (i) using all 440 control datasets against the performance and (ii) using only 93 control datasets from the related GM cell lines, which have relatively normal karyotypes. Matching control datasets were not used. We observed that the performance slightly increases even when control datasets with abnormal karyotypes were included (Supplementary Figure S16, P-value $= 0.02$ with Wilcoxon signed-rank test). This is expected since our model should be able to put appropriate weights on control datasets, based on how informative they are for imputing common background signal in an IP dataset, especially given that we use 30 million genomic positions as samples that provide strong statistical power. While Supplementary Figure S12 from the previous section already

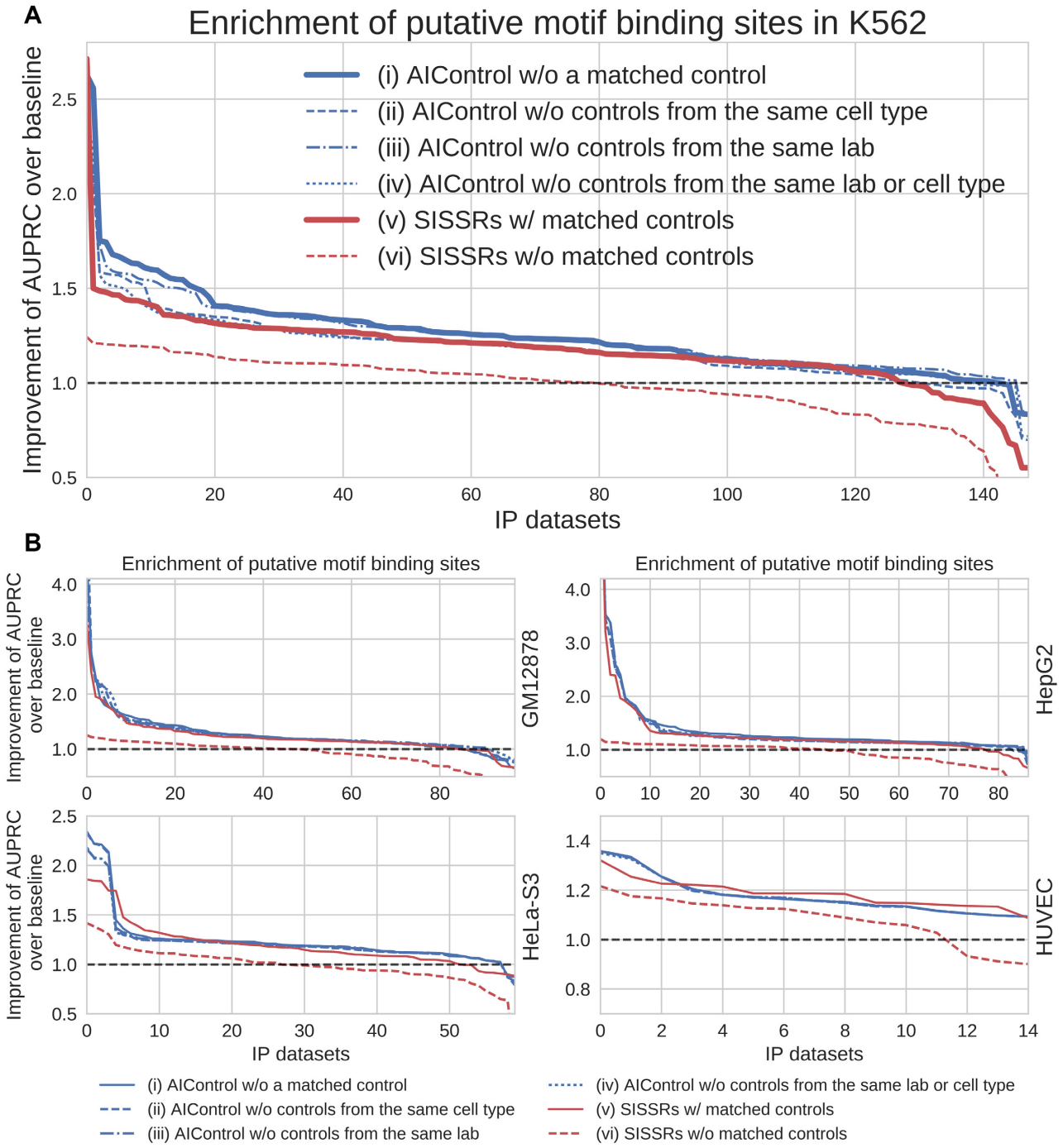

Figure 4. Relative performance of AIControl when control datasets from the same cell types or the same labs have been excluded from the background set. We compare across six settings: (i) AIControl with all 440 control datasets except for matched controls, (ii) AIControl without control datasets from the same cell type as the IP dataset, (iii) AIControl without control datasets from the same lab as the IP dataset, (iv) AIControl without control datasets from the same lab or the same cell type as the IP dataset, (v) SISSRs with matched control datasets and (vi) SISSRs without matched control datasets. 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**) 149 IP datasets measured in K562, and (**B**) IP datasets measured in the tier 1 ENCODE cell types: GM12878, HepG2, HeLa-S3, HUVEC.

shows that more control datasets are always better (assuming you do not know which controls you need), this result on GM cell types shows that even control datasets from abnormal cell types improve the performance on cell types with a normal karyotype, thanks to the ability of AIControl to properly integrate background signal structures.

The performance of AIControl indeed dropped when the control datasets from the same cell types are not available. However, it is important to note that our framework, without controls from the same cell type, identified peaks that are better associated with sequence motifs than other peak callers with matched control datasets from the same cell type (Supplementary Tables S5 and S9). This suggests that our framework can successfully estimate the structure of background signals in one cell type by leveraging information from other cell types in a 'cross-cell-type' manner.

AIControl reveals transcription factor interactions better than alternative methods

One of the many downstream use cases of ChIP-seq data is to learn interactions among regulatory factors by observing their co-localization patterns on genome (35–37). In particular, Lundberg *et al.* (6) 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 (31). Here, we use 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. Note that the interactions between datasets that target the same transcription factor and the self-interactions in the diagonal entries are included in this analysis.

Figure 5A 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.* (6). Areas under the enrichment curves indicate that AIControl performs better at revealing known PPIs than other methods in K562 (Figure 5). In particular, AIControl ranked more true BioGrid-supported interactions—for example, JUN/STAT1, E2F6/MAX, IRF1/STAT1 and GATA2/JUN—above the threshold (defined as the number of true interactions) than other peak callers (Supplementary Table S10). Additionally, we performed the same enrichment analysis on the other four cell types: 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 other cell types (Supplementary Figure S10).

Figure 5B visualizes the inverse correlation matrices generated from the peaks called by five 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 ground truth (i.e. BioGrid PPIs) relative to other methods. Additionally, we compared AIControl against SPP peaks downloaded from ENCODE ('SPP-ENCODE') and SPP peaks generated with our own pipeline. Similar to the result in the motif enrichment task, AIControl continues to exhibit better performance at recovering PPIs (Supplementary Figure S17).

Supplementary Table S11 shows top 10 TF interactions that are uniquely suggested by AIControl. Although these interactions are not currently in the database, some studies suggest potential interactions between the pairs. For example, 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 (38). 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 types. 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 accurately in all cell types suggests that using this framework can improve the quality of downstream analysis that follows ChIP-seq experiments.

AIControl better removes common background signal among datasets

One of the most frequently used quality measures for biological experiments is the consistency of a pair of replicate datasets, which can be measured by the number of shared peaks. A pair of replicate datasets should capture the exact same signals; thus, a pair with better quality should share more peaks with each other. On the other hand, the quality of background signal removal can be assessed by measuring the inconsistency in a pair of unrelated datasets. We define an 'unrelated' pair as a pair of datasets that (i) is in the same cell type and (ii) targets unrelated transcription factors without any documented PPI in BioGrid. As described in 'Materials and methods' section, AIControl models ChIP-seq experiments as follows:

$$\begin{aligned}\texttt{ChIPseqData} = {} & \texttt{ProteinBindingSignal} \\ & + \texttt{ReproducibleBackgroundSignal} \\ & + \texttt{IrreproducibleNoise}\end{aligned}$$

For a pair of unrelated datasets, we assume that there is no `ProteinBindingSignal` that gives rise to shared peaks. The only source of shared peaks in an unrelated pairs is `ReproducibleBackgroundSignal`. If a peak caller perfectly removes `ReproducibleBackgroundSignal`, it should ideally leave no peak that is shared between a pair of unrelated datasets. Thus, peak callers better able to remove common background signal should have fewer shared peaks for unrelated datasets. This metric alone is not perfect, because a pair of completely random peaks can achieve

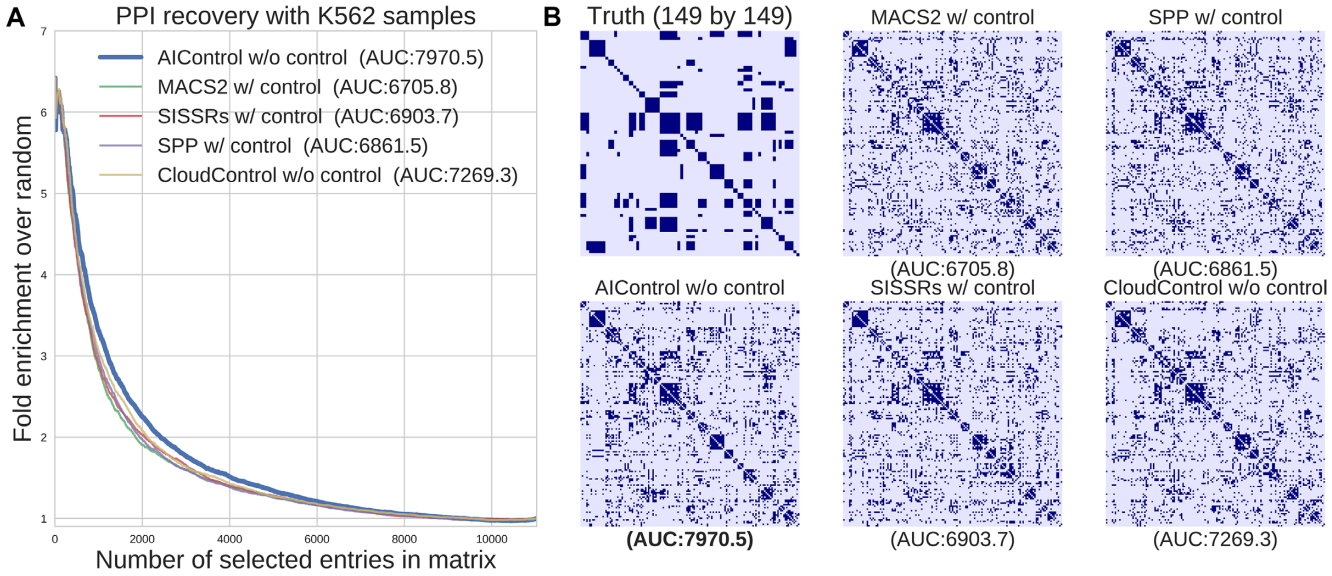

Figure 5. Performance of AIControl compared to other peak callers on the PPI recovery task in K562. (**A**) Enrichment of BioGrid-supported interactions of transcription factors in the inverse correlation networks inferred from 149 K562 IP datasets. Peak signals are obtained from: (i) AIControl w/o control, (ii) MACS2 w/control, (iii) SISSRSs w/control, (iv) SPP w/control and (v) CloudControl + MACS2 w/o control. (**B**) Documented interactions among regulatory proteins (top left) and heat maps of inverse correlation networks (rest). The heat maps are binarized to show top 3,583 interactions, which is equal to the number of true interactions.

good results as well. However, we believe that, in combination with the motif enrichment and PPI recovery task, this metric highlights an important aspect of noise removal process.

Supplementary Figure S18 shows the 'sharedness' of peaks for 9,310 pairs of unrelated datasets in K562 processed by five peak callers: AIControl, MACS2, SISSRs, SPP and CloudControl+MACS2. The *y*-axis indicates the proportion of unrelated datasets that have less than a particular number of shared peaks, which is represented in the *x*-axis. The smaller area under the curve demonstrates that a peak caller generally identifies fewer shared peaks between a pair of unrelated datasets, and AIControl exhibits the smallest area under its curve. For other peak callers, a larger percentage of unrelated dataset pairs contained more shared peaks, suggesting that their bias-removal process was not as thorough as that of AIControl's.

AIControl is compatible with the irreproducible discovery rate (IDR) framework

We investigated the performance of AIControl in a situation where biologically replicated samples are available. In particular, the ENCODE consortium uses the irreproducible discovery rate (IDR) framework to adaptively rank and select peaks based on the rank consistency/reproducibility of signals among biological replicates (25). The ENCODE official pipeline uses SPP in combination with IDR to identify and reorder peaks in ChIP-seq datasets. In order to evaluate the effect of IDR on the peak callers, for datasets where biological replicates are available, we performed IDR analysis after calling peaks with AIControl, MACS2 and SISSRs. For SPP, we directly downloaded peaks processed with IDR from the ENCODE website.

Supplementary Figure S1A shows the superior performance of AIControl+IDR in the motif sequence identification task across five tested cell types (i.e. K562, GM12878, HepG2, HeLa-S3 and HUVEC) compared to that of other peak callers with IDR. Negative log10 IDR values were used as a ranking measure. Needless to say, other peak callers have access to matched control datasets, whereas AIControl does not. We also observed that AIControl+IDR better predicts protein–protein interactions in the K562 cell type than other peak callers when they are used in combination with IDR (Supplementary Figure S1B).

AIControl retains its performance on 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 generalizes strong performance on a ChIP-seq IP dataset that is not a part of the ENCODE database, we performed peak calling on 14 IP datasets that are obtained from 8 independent studies (39–46), which are not part of the ENCODE database and are only on the GEO or ArrayExpress database. We only analyzed the datasets whose target motif PWMs for *H. sapiens* are available in the JASPAR database. The AUPRC values were measured across the whole genome. The information of the 14 datasets are summarized in Supplementary Table S4. We compared the following peak calling frameworks: AIControl, MACS2, SISSRs and SPP.

Figure 6 shows the performance of AIControl on all 14 external datasets in comparison to other peak callers. Individual PR curves are shown in Supplementary Figure S19. AIControl retains its strong performance on all datasets except for the one that targets SP1 in the HEK293 cell

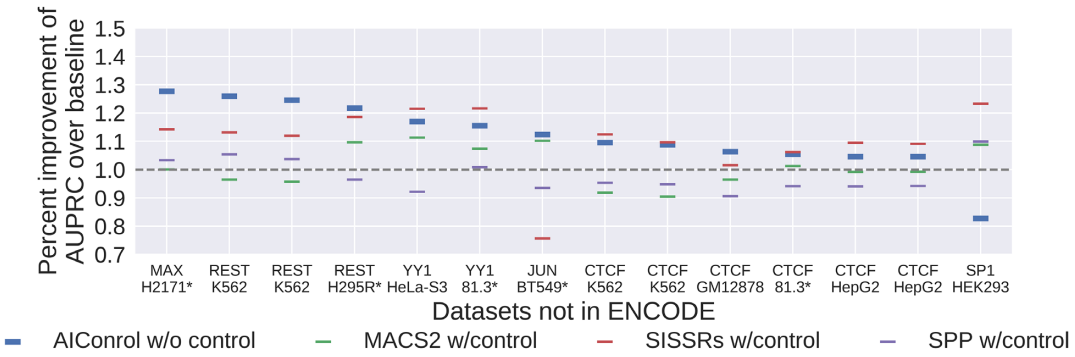

Figure 6. Relative performance of five peak calling methods on external datasets. Peak regions were identified using: (i) AIControl w/o control, (ii) MACS2 w/control, (iii) SISSRSs w/control and (iv) SPP w/control. The *y*-axis shows the percent improvement of the area under the precision-recall curves (AUPRCs) for predicting the presence of putative binding sites with ranking measures associated with the peaks over the baseline (i.e. MACS2 without using a matched control dataset). The *x*-axis shows the non-ENCODE ChIP IP datasets ordered by the percent improvement achieved by AIControl (*y*-axis). Datasets measured in cell types that are not in the 440 ENCODE control set are shown with asterisks.

type. This is consistent with the result shown in Figure 2A, where AIControl exhibits the worst performance on the datasets with SP1. Most notably, 5 out of the 14 external datasets were measured in the cell types that were not included in the pool of control datasets used for AIControl. All datasets had matched control datasets, which were used by MACS2, SISSRs and SPP, but not by AIControl. The fact that AIControl without matched control datasets retained strong performance suggests that it is able to integrate background control signals in a cross-cell-type setting even in a case where the datasets are not as highly regulated as that of ENCODE database.

Principal components of the control datasets are associated with potential bias sources

Control datasets capture background signals that are also present in corresponding IP datasets. Many studies suggest that these background signals are combinations of multiple different sources of biases, for instance, GC content, sonication biasand platform-specific biases (18,19). The AIControl framework assumes that observed background signals in a control dataset can be represented as the weighted sum of many different known or unknown bias sources (see 'Materials and methods' section).

Supplementary Figure S20 shows Spearman's correlation coefficients between potential bias sources and K562 control datasets projected on the first five principal components. Open chromatin regions (HS) and read mappability (MP) are similar to the first principal component, while GC content (GC) is similar to the second principal component. Notably, the first five principal components collectively capture only 54.05% variance, which suggests other bias sources are likely to exist that contribute to the observed background signal. AIControl implicitly learns the contributions from unobserved sources of biases; this is one of the reasons that AIControl can call more accurate peaks 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. (i) It learns position-specific distribution of background signal at much finer resolution than other methods by using publicly available control datasets on a large scale (see 'Materials and methods' section). Our evaluation metrics show that using finer background distributions improved enrichment of putative TF-binding locations and recovery of known protein–protein interactions. (ii) AIControl systematically integrates control datasets from a public database (e.g. ENCODE) without any user input. Its ability to learn background signals extends to datasets obtained in unseen cell types without any previously measured control datasets. We obtained 440 ChIP control datasets from 107 cell types in the ENCODE database, and AIControl learns to statistically combine them to estimate background signals in an IP dataset in any cell type. We showed that our performance on unseen cell types exceeds that of established baselines. AIControl's performance is also generalizable to datasets from labs outside the ENCODE project. (iii) The mathematical model of AIControl accounts for multiple sources of biases due to its integration of control datasets at a large scale (see 'Materials and methods' section). On the other hand, some sources of biases may not be fully captured by existing methods that use only one matched control dataset (13,17) or account for only a specific set of biases. (18,19). (iv) 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 the 410 ENCODE ChIP-seq datasets from five major ENCODE cell types for 54 different transcription factors (Supplementary Table S1). We showed that AIControl has better motif sequence enrichment compared to other peak callers within predicted peak locations. However, this metric measures only direct interactions between transcription factors and DNA. Thus, we evaluated the performance of AIControl with another metric: PPI enrichment analysis. In this

metric, we also observed that AIControl is superior to other peak callers even without any matched control samples. In conclusion, 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 (20,21,32). 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. There are several future extensions for our framework. (i) Our default implementation bins IP and control datasets into 100 bp windows in order to perform fast genome-wide regression. Because most transcription factors show signals wider than 100 bps, we believe that our resolution is sufficient to conduct accurate downstream analysis. The Julia implementation can be accessed at `https://github.com/suinleelab/AIControl.jl`, and the accompanying files can be accessed through the Google Drive link under the 'Paper' section on the GitHub page. (ii) Since our framework learns weights that are globally applied to all genomic positions, it performed relatively worse on estimating background signal for cell types with abnormal karyotype in a cross-cell-type setting. In future, this could be resolved by automatically detecting karyotype abnormality and learning different sets of weights for those regions. (iii) Unlike other peak callers, the unique core idea of AIControl is to leverage available control datasets in public. This requires all datasets (both user-provided ChIP-seq target dataset and public control datasets) to be mapped to the exact same version of reference genome. The public control datasets are currently mapped to the hg38 human genome assembly from the UCSC repository, which can be found at `http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz`. Therefore, our framework also expects user-provided ChIP-seq datasets in the UCSC hg38 space. If an user wants to use our framework for identifying peaks on the different genome assembly (e.g. hg19) or on the different version of hg38, they must start by remapping reads to the UCSC version of the hg38 assembly.

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.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge the following people for testing the AIControl software and giving suggestions on improving its usability: Ayse Berceste Dincer, Joseph D. Janizek, Gabriel Erion, Pascal Sturmfels, and Nicasia Beebe-Wang from Prof. Su-In Lee's lab; Mehran Karimzadeh from Prof. Michael M. Hoffman's lab; Timothy Durham, and Jacob Schreiber from Prof. William Noble's lab; Arpit Mishra, Eric Waddell, and Stephanie L. Battle from Prof. R. David Hawkins' lab; David Read from Prof. Cole Trapnell's lab; and Daniel C. Jones, Alex Okeson, and Erin H. Wilson from Paul G. Allen School of Computer Science & Engineering.

FUNDING

This work was supported by National Science Foundation CAREER [DBI-1552309, DBI-1355899]; American Cancer Society [127332-RSG-15-097-01-TBG]; and National Institutes of Health [R35 GM 128638]. Funding for open access charge is from: NSF CAREER [DBI-1552309].
Conflict of interest statement. None declared.

REFERENCES

1. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
2. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
3. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
4. Ernst,J. and Kellis,M. (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.*, **33**, 364–376.
5. Schmidt,D., Wilson,M.D., Ballester,B., Schwalie,P.C., Brown,G.D., Marshall,A., Kutter,C., Watt,S., Martinez-Jimenez,C.P., Mackay,S. *et al.* (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**, 1036–1040.
6. Lundberg,S.M., Tu,W.B., Raught,B., Penn,L.Z., Hoffman,M.M. and Lee,S.I. (2016) ChromNet: learning the human chromatin network from all ENCODE ChIP-seq data. *Genome Biol.*, **17**, 82.
7. Ng,F.S., Ruau,D., Wernisch,L. and Göttgens,B. (2016) A graphical model approach visualizes regulatory relationships between genome-wide transcription factor binding profiles. *Brief. Bioinform.*, 162–173.
8. Chorley,B.N., Campbell,M.R., Wang,X., Karaca,M., Sambandan,D., Bangura,F., Xue,P., Pi,J., Kleeberger,S.R. and Bell,D.A. (2012) Identification of novel NRF2-regulated genes by ChIP-Seq: influence on retinoid X receptor alpha. *Nucleic Acids Res.*, **40**, 7416–7429.
9. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
10. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
11. Bottomly,D., Kyler,S.L., McWeeney,S.K. and Yochum,G.S. (2010) Identification of β-catenin binding regions in colon cancer cells using ChIP-Seq. *Nucleic Acids Res.*, **38**, 5735–5745.
12. Berger,M.F., Lawrence,M.S., Demichelis,F., Drier,Y., Cibulskis,K., Sivachenko,A.Y., Sboner,A., Esgueva,R., Pflueger,D., Sougnez,C. *et al.* (2011) The genomic complexity of primary human prostate cancer. *Nature*, **470**, 214–220.
13. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
14. Diaz,A., Park,K., Lim,D.A., Song,J.S. *et al.* (2012) Normalization, bias correction, and peak calling for ChIP-seq. *Stat. Appl. Genet. Mol. Biol.*, **11**, 9.
15. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
16. Hiranuma,N., Lundberg,S. and Lee,S.I. (2016) 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, Seattle, pp. 191–199.
17. Narlikar,L. and Jothi,R. (2012) ChIP-Seq data analysis: identification of Protein–DNA binding sites with SISSRs peak-finder. *Next Gen. Microarray Bioinform.: Methods Protocols*, 305–322.
18. Ramachandran,P., Palidwor,G.A. and Perkins,T.J. (2015) BIDCHIPS: bias decomposition and removal from ChIP-seq data clarifies true binding signal and its functional correlates. *Epigenetics Chromatin*, **8**, 33.
19. Kuan,P.F., Chung,D., Pan,G., Thomson,J.A., Stewart,R. and Keleş,S. (2011) A statistical framework for the analysis of ChIP-Seq data. *J. Am. Stat. Assoc.*, **106**, 891–903.
20. Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, **5**, e11471.
21. Laajala,T.D., Raghav,S., Tuomela,S., Lahesmaa,R., Aittokallio,T. and Elo,L.L. (2009) A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, **10**, 618.
22. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
23. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
24. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
25. Li,Q., Brown,J.B., Huang,H., Bickel,P.J. *et al.* (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.
26. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R. *et al.* (2017) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.
27. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
28. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
29. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
30. Kinsella,R.J., Kähäri,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P., Kerhornou,A. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, **2011**, bar030.
31. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
32. Thomas,R., Thomas,S., Holloway,A.K. and Pollard,K.S. (2016) Features that define the best ChIP-seq peak calling algorithms. *Brief. Bioinform.*, **18**, 441–450.
33. Mortazavi,A., Thompson,E.C.L., Garcia,S.T., Myers,R.M. and Wold,B. (2006) Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. *Genome Res.*, **16**, 1208–1221.
34. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome Res.*, **22**, 1723–1734.
35. Perner,J., Lasserre,J., Kinkley,S., Vingron,M. and Chung,H.R. (2014) Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling. *Nucleic Acids Res.*, **42**, 13689–13695.
36. Zhou,J. and Troyanskaya,O.G. (2014) Global quantitative modeling of chromatin factor interactions. *PLoS Comput. Biol.*, **10**, e1003525.
37. Van Steensel,B., Braunschweig,U., Filion,G.J., Chen,M., van Bemmel,J.G. and Ideker,T. (2010) Bayesian network analysis of targeting interactions in chromatin. *Genome Res.*, **20**, 190–200.
38. Dengler,V.L., Galbraith,M.D. and Espinosa,J.M. (2014) Transcriptional regulation by hypoxia inducible factors. *Crit. Rev. Biochem. Mol. Biol.*, **49**, 1–15.
39. Schmidl,C., Rendeiro,A.F., Sheffield,N.C. and Bock,C. (2015) ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat. Methods*, **12**, 963–965.
40. Schwalie,P.C., Ward,M.C., Cain,C.E., Faure,A.J., Gilad,Y., Odom,D.T. and Flicek,P. (2013) Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biol.*, **14**, R148.
41. Schmidt,D., Schwalie,P.C., Ross-Innes,C.S., Hurtado,A., Brown,G.D., Carroll,J.S., Flicek,P. and Odom,D.T. (2010) A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.*, **20**, 578–588.
42. Zhao,C., Qiao,Y., Jonsson,P., Wang,J., Xu,L., Rouhi,P., Sinha,I., Cao,Y., Williams,C. and Dahlman-Wright,K. (2014) Genome-wide profiling of AP-1–regulated transcription provides insights into the invasiveness of triple-negative breast cancer.*Cancer Res.*, **74**, 3983–3994.
43. Doghman,M., Figueiredo,B.C., Volante,M., Papotti,M. and Lalli,E. (2013) Integrative analysis of SF-1 transcription factor dosage impact on genome-wide binding and gene expression regulation. *Nucleic Acids Res.*, **41**, 8896–8907.
44. Lin,C.Y., Lovén,J., Rahl,P.B., Paranal,R.M., Burge,C.B., Bradner,J.E., Lee,T.I. and Young,R.A. (2012) Transcriptional amplification in tumor cells with elevated c-Myc. *Cell*, **151**, 56–67.
45. Michaud,J., Praz,V., Faresse,N.J., JnBaptiste,C.K., Tyagi,S., Schütz,F. and Herr,W. (2013) HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy. *Genome Res.*, **23**, 907–916.
46. Völkel,S., Stielow,B., Finkernagel,F., Stiewe,T., Nist,A. and Suske,G. (2015) Zinc finger independent genome-wide binding of Sp2 potentiates recruitment of histone-fold protein Nf-y distinguishing it from Sp1 and Sp3. *PLoS Genet.*, **11**, e1005102.