CloudControl: Leveraging many public ChIP-seq control experiments to better remove background noise

Naozumi Hiranuma
Computer Science and Engineering, University of Washington
AC310 Paul G. Allen Center
Seattle, Washington
98195-2350
hiranumn@uw.edu

Scott Lundberg
Computer Science and Engineering, University of Washington
AC310 Paul G. Allen Center
Seattle, Washington
98195-2350
slund1@uw.edu

Su-In Lee
Computer Science and Engineering, University of Washington
AC536 Paul G. Allen Center
Seattle, Washington
98195-2350
suinlee@uw.edu

ABSTRACT

Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) is a widely used method to determine the binding positions of various proteins on the genome in a population of cells. A typical ChIP-seq protocol involves two experiments: one designed to capture target ChIP-seq signals ('target' experiment) and the other to capture background noise signals ('control' experiment). A peak calling algorithm then examines the difference between the target experiment data and control data to determine where the protein of interest binds along the genome. Our approach, named *CloudControl*, aims to improve the accuracy of peak calling by combining multiple control experiments from a publicly available source such as ENCODE to better remove background noise signals. To combine existing control experiment data we perform regression against a target experiment treating binned genome positions as samples (up to 32 million) and different control experiments as features (up to 455 through the ENCODE project). The regression fit is then used to generate a new control ChIP-seq data, which we refer to as CloudControl data. We use the following three metrics to evaluate the CloudControl data: (i) the presence of the known motifs for the corresponding target protein near called peaks, (ii) reproducibility among pairs of biologically replicated ChIP-seq experiments, and (iii) protein-protein physical interactions inferred from called peaks. In all three metrics, CloudControl data show superior performance over standard control tracks. This suggests that CloudControl can improve ordinary control tracks in the standard ChIP-seq protocol.

CCS Concepts

•**Applied computing** → *Bioinformatics;*

ACM-BCB '16 October 2–5, 2016, Seattle, WA, USA
© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Keywords

Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq), peak calling, regression

1. INTRODUCTION

Hundreds of millions of dollars have been invested recently in the ENCODE project to identify the genomic binding positions of various regulatory proteins in over a hundred cell types. Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) and related variants are by far the most common experimental procedures used to identify the binding positions of regulatory proteins.

As described in Section 2.1, separation of true protein binding signals from background noise is a challenging problem and has caught significant research attention from the computational biology community [22, 6, 15, 16, 7, 21]. The most standard approach to address this challenge is to obtain data from a *control* experiment intended to capture background noise signals (Section 2.2), and then compare this with data from a *target* experiment. Typical peak calling algorithms take these *target data* and *control data* (green-colored arrows in Figure 1a) and identify protein binding by looking for signal present only in the target experiment data [22, 6, 15, 16, 7] (Section 2.2).

In this paper, we present the CloudControl framework to *generate* control data for a particular target data by integrating hundreds of existing control data sets. The idea is that due to the stochastic nature of background noise, a single control experiment may not fully capture the true background noise model. CloudControl tackles this problem by finding an appropriate combination of a large number of publicly available control data sets from large scale projects such as ENCODE (Figure 1b) [4]. CloudControl takes as input a target ChIP-seq data and a set of control data sets (e.g., from ENCODE) and generates *CloudControl data* to be used as input for a peak calling algorithm (blue-colored arrows in Figure 1a). The generated CloudControl data is specifically tailored to a given target data, and we show it better removes background noise. This framework can easily be built into existing pipelines as it produces a CloudControl BAM file, which simply substitutes for the BAM file from a standard control experiment (Figure 1a).

The following results are shown in the subsequent sections of this paper. (i) Higher confidence peaks called with CloudControl data were more enriched for putative motif binding

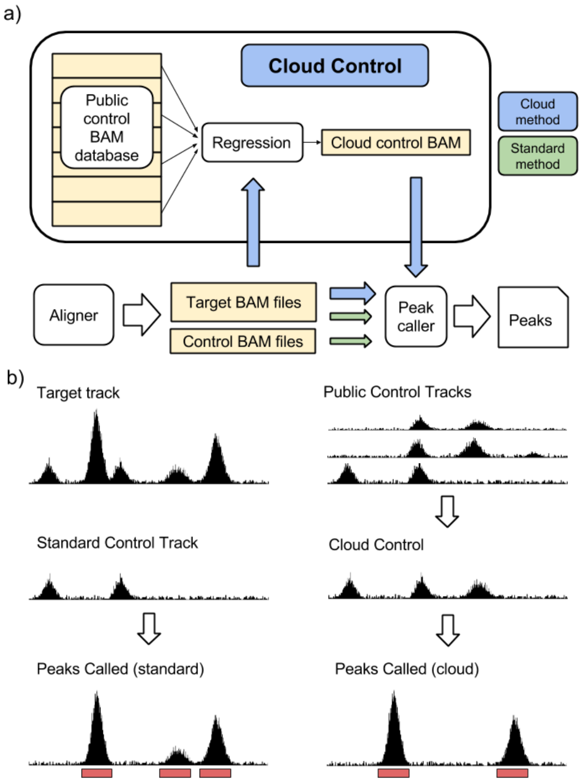

Figure 1: a) The CloudControl framework: It takes a target ChIP-seq data (as BAM files) as input and performs regression of publicly available control tracks against the target data. The output of the regression model (i.e., the regression fit) is used to generate the CloudControl data to be used as input data for a peak calling algorithm. b) A standard control track may not fully capture the true background model. This may leave some false positive peaks unfiltered (left). On the other hand, using multiple control data sets, CloudControl (right) has a better chance to remove false positive peaks that may not have been captured by a single control data set. The red bars indicate peaks called by a peak caller.

sites of the corresponding target proteins, especially within DNase-I hypersensitive sites (DHSs). (ii) The difference in reproducibility of peaks between replicate experiment pairs and non-replicate experiment pairs was more pronounced when CloudControl data were used. (iii) Known protein-protein interactions were better revealed when using the peaks called with CloudControl data than when using standard control data.

2. BACKGROUND

Here, we describe the ChIP-seq experimental procedure and how control experiments are done.

2.1 ChIP-seq experiment procedure

A typical ChIP-seq protocol targeting a regulatory protein (e.g., transcription factors, histone marks) consists of several steps [9, 2]. First, DNA in a population of cells is cross-linked with nearby proteins using formaldehyde. This is immediately followed by sonication to break the DNA into short segments of hundreds of nucleotides. The short sequences are then filtered and extracted using an antibody targeting the protein of interest. In the next step, the protein-DNA complexes captured by the antibody are separated. For directly bound proteins the resulting short DNA sequences are highly likely to contain the protein's binding site. In the high-throughput sequencing step, the first 30∼200 nucleotides of the filtered DNA fragments are sequenced. These sequenced reads can be mapped back to the reference genome using a DNA mapping program such as Bowtie or BWA [10, 11]. Since the reads can be mapped to both reverse and forward DNA sequence, this will produce many bimodal signals across the entire genome which are presumably enriched for the binding sites of the protein of interest. However, these peaks can also be generated by various sources of background noise.

Separation of true protein binding signals from background noise in ChIP-seq protocol has caught significant research attention from the computational biology community [22, 6, 15, 16, 7, 21]. Improvements in this process can greatly increase the quality of the downstream analysis of ChIP-seq experiments. There are many potential sources of noise in ChIP-seq experiments. For instance, antibodies may have low specificity to their target proteins, which increases the rate of random spurious binding events. This leads to the production of signals that do not contain true protein binding sites. There are also many undocumented gene duplication and deletion instances in the human genome, producing bumps of mapped read counts, which can be misinterpreted as binding signals [17]. The sonication process is also known to introduce bias which correlates with the accessibility to the genome, which can be captured by DNase-1 hypersensitivity sites [1].

2.2 Control ChIP-seq experiments

In order to remove background noise and other confounding factors, labs often run two sets of experiments: one experiment that targets the protein of interest, and the other control experiment that is designed to capture the background noise signal [9]. There are two major types of control experiments. "Input" controls are produced by following a typical ChIP-seq protocol but skipping the immunoprecipitation step, in which antibodies for the target protein are used to capture DNA near the target. On the other hand, "mock" controls are produced by using a control antibody that targets an irrelevant, non-nuclear antigen. A peak calling algorithm is then used to identify the regions enriched for peaks in the target experiment but not in the control experiment [22, 6, 15, 16, 7] (Figure 1).

3. METHODS & MATERIALS

3.1 Regression model

Here, we describe how we integrate many control data sets to generate a CloudControl data set for a given target data set. The basic idea of CloudControl is that we can better estimate the background noise signals within a target data set by using a large number of control data sets than by using a single control data set (Figure 1). In other words, CloudControl attempts to remove background noise as much as possible by using many control data sets. To achieve this goal, we can use a linear regression approach to approximate the target data based on a weighted combination of 445 control data sets available through the ENCODE project.

Let $Y \in \mathcal{N}^{32,092,861\times 1}$ represent data from a target experiment. Each element Y_i represent the number of reads in the target experiment that map to the ith bin. Let $X \in \mathcal{N}^{32,092,861\times 455}$ represent the data matrix of read counts at all bins in the genome from 445 control experiments. Then, our goal is to find a vector $\beta \in \mathcal{R}^{455\times 1}$ that contains the weight values on these 445 control data sets, which make the linear combination, $X\beta$, that best approximates the target data set Y. This can be solved by using a linear regression model in which the 445 control data sets are features and $32,092,861$ genomic regions are treated as samples. This leads to the following optimization problem:

$$\underset{\beta}{\operatorname{argmin}} ||Y - X\beta||_2^2 + \lambda||\beta||_2^2. \qquad (1)$$

The solution for this optimization problem can be obtained by computing:

$$\beta = (X^T X - \lambda I)^{-1} X^T Y. \qquad (2)$$

Although the large number of samples should prevent our model from overfitting, we add a small ℓ_2 regularization parameter ($\lambda = 0.00001$) to ensure numerical stability. It is important to note that binning the read count data not only shrinks the data size (sample size) but also makes the samples (i.e., genomic regions) less dependent on each other.

After β is learned for a given target data set, we generate CloudControl data based on $X\beta$ – a linear combination of 445 control data sets, which can best explain the target data. We generate a BAM file based on $X\beta$ in the following way: For the ith bin, we sample $\hat{z_i}$ from a Poisson distribution centered at $(X\beta)_i$, the ith element of $X\beta$. We then randomly select $\hat{z_i}$ positions from the region that corresponds to the ith bin, and we generate a synthetic DNA read starting at each selected position.

Two regression models were trained separately for both forward and reverse complement reads to account for the bimodal shape of ChIP-seq signals. This produces a BAM file for a CloudControl data, which is based on linear combination of multiple control data sets. The BAM file can then be used as a control input for any peak calling algorithm (such as MACS2). In our experiments, we considered two widely used peak calling algorithms, MACS2 and SISSRs [22, 15, 6].

We investigated the performance of several regression models on data from a target experiment for RE1-Silencing Transcription factor (REST) in K562 (ENCODE ids ENCFF000NYK and ENCFF000NYN). The ridge regression, Poisson regression, and gradient boosting trees were evaluated by checking the presence of nearby putative REST binding sites predicted by FIMO [5]. The result showed that the choice of a regression model did not significantly affect the quality of peaks called with MACS2. This led us to choose the ridge regression, because it allows us to compute the optimal values of β from very large input data matrix efficiently using a closed form solution.

3.2 Data

The data used in our experiments were obtained through the Encyclopedia of DNA Elements (ENCODE) project by the US National Human Genome Research Institute (NHGRI) [4]. In particular, our analysis focused on the ChIP-seq target data sets that were measured on K562, GM12878, and HepG2 because these cell types contain the largest numbers of target experiment data. The number of target data sets used for our experiments was 217, 128 and 93 for K562, GM12878, and HepG2, respectively. We also obtained all available 455 control data sets across all cell types through the ENCODE project. The locations of DNase-1 hypersensitivity sites were obtained from the Epigenomics Roadmap project [3].

3.3 Peak calling

For peak-calling, we used Model-based Analysis of ChIP-seq (MACS2) [22] for most of our experiments. MACS2 assumes that the read counts in each genomic window are Poisson distributed. The mean λ for each window is learned by maximum likelihood estimation involving read counts of nearby windows in a control data set. A significant deviation of read counts in a target data set from the learned λ is considered a peak. We set the p-value threshold for peak calling to be 10^{-5}. MACS2 has been used by large scale projects such as the ENCODE project and the Epigenomics Roadmap project, making it one of the most popular choices for peak calling [3, 4]. We also explored another peak caller, Site Identification from Short Sequence Reads (SISSRss) to show the validity of CloudControl in pipelines that use peak-callers other than MACS2 [15, 6].

After peaks were called for each pair of target/control data, we constructed a vector of floats, in which each element corresponds to a non-ovelapping window of 1,000 base pairs and its value is equal to the most significant p-value of the peaks within the window. We also created a binary vector whose entries indicate whether there is a peak present or not in the corresponding genomic windows. For instance, in data from a human genome, there would be three million entries in the vector. These vectors are used in many of our evaluation metrics to represent the location of peaks as well as other information such as motif binding sites and DNase-1 hypersensitivity sites.

4. RESULTS

4.1 CloudControl improves the motif enrichment of high confidence peaks in hypersensitive regions.

The JASPER CORE database stores the position weight matrices (PWM) of many transcription factors across different species [13, 14]. For the tracks whose target protein's PWM for human are available on JASPER, we ran a motif scanner program FIMO to determine the putative binding sites of these motifs [5]. Then used the putative motif binding locations as ground truth for the corresponding target tracks. We evaluated the peaks generated by a peak caller for the following treatments on control data:

- **cloud** - CloudControl data (Figure 1) were used for each target track.
- **standard** - A designated control data set was used for each target track.
- **no-control** - No control data were used.

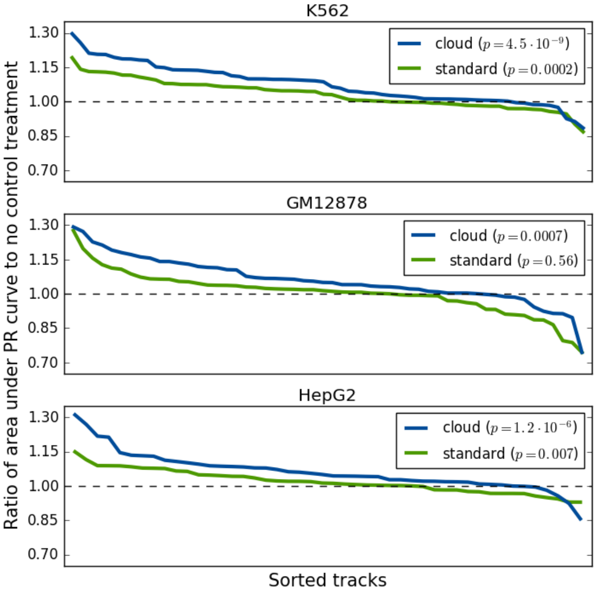

Figure 2: Performance of peak calling using either a standard control or cloud control. Peaks are considered correct when putative motif binding sites are nearby in DNase-I hypersensitivity sites. The number of target tracks examined are 62, 54, 46 for K562, GM12878, and HepG2, respectively. The motif binding sites were predicted by FIMO using position weight matrices from JASPAR CORE. The area under the PR curve for each target ChIP-seq data set with the cloud (blue) or standard controls (green) was compared against no control. The x-axis represents different ChIP-seq target tracks sorted by their relative performance over no control. 20,000 peaks with the highest confidence values from MACS2 were used for this analysis.

It has been reported that 98.5% of the transcription factor binding sites are located in the open chromatin regions defined by DNase-I hypersensitivity sites (DHSs) [18]. Therefore, our analysis focused on the peaks that are located near DHSs. For each ChIP-seq experiment, we evaluated the top 20,000 peaks. Given 3 million possible positions for peaks to appear, this would amount to 0.66% of the whole genome.

Figure 2 shows the relative performance of the standard and cloud over the peaks called without any control data (no-control) for K562, GM12878, and HepG2. For each target track measured in K562, we measured the area under precision-recall (PR) curve for predicting the presence of putative protein binding sites of the target transcription factor using the p-values of peaks called by MACS2 as predictors. The use of standard controls significantly improved the quality of the peaks called in K562 and HepG2 over the no-control treatment, but surprisingly not in GM12878 ($p = 0.0002$ and $p = 0.007$ for K562 and HepG2, respectively). On the other hand, the cloud treatment achieved significant improvement over the no-control treatment in all three cell types ($p = 4.5 \cdot 10^{-9}$, $p = 0.0007$, and $p = 1.2 \cdot 10^{-6}$ for K562, GM12878, and HepG2, respectively).

More importantly, the cloud control significantly outperformed the standard control in all three cell types (Figure 3), showing that the use of CloudControl likely improves the quality of called peaks over the traditional use of control tracks ($p = 0.0007$, $p = 0.0028$, and $p = 0.0018$ for K562, GM12878, and HepG2, respectively).

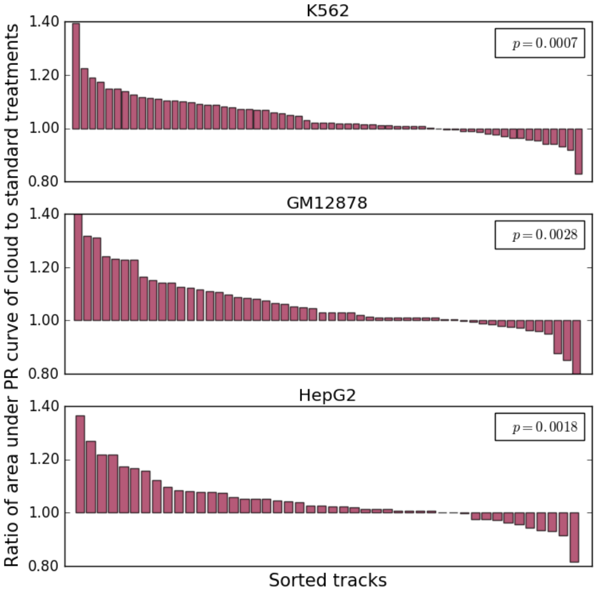

Figure 3: A waterfall plot showing the relative performance of the peaks called with cloud controls over standard controls for predicting putative motif binding sites nearby in DNase-1 hypersensitivity sites. The number of tracks examined are 62, 54, 46 for K562, GM12878, and HepG2, respectively. Each bar represents the ratio of area under the PR curve for a particular ChIP-seq track with the cloud treatment to that of the standard treatment. 20,000 peaks with the highest confidence values were used. 44 out of 62 tracks were improved when using cloud in K562. 41 out of 54 tracks were improved in GM12878. 35 out of 46 tracks were improved in HepG2. The regression weights β learned for the tracks that showed the best and second best improvement in GM12878 and the best improvement in HepG2 are shown in Figure 6 (ENCFF00OBH, ENCFF002EBQ, and ENCFF002EDT, respectively)

Recall that the area under the PR curve is calculated based on the top 20,000 peaks ranked by significance. Figure 4 shows the effect of considering different numbers of peaks. The cloud treatment is statistically superior to the standard treatment up to top 50,000 peaks. 50,000 peaks amount to approximately 1.6% of the human genome, which we assume as an upper-bound for the number of binding sites for transcription factors that are not broad-source (e.g. ZNF217 and histone marks).

One of the advantages of CloudControl is its flexibility to fit into any pipeline of sequence aligners, peak callers, and subsequent data analysis. In order to demonstrate this claim, an identical evaluation was performed on the peaks called by SISSRs in K562 [15, 6] (Figure 5). Similarly to the result of MACS2, the cloud treatment significantly outperformed the standard treatment. These results support our hypothesis that the use of an appropriate combination

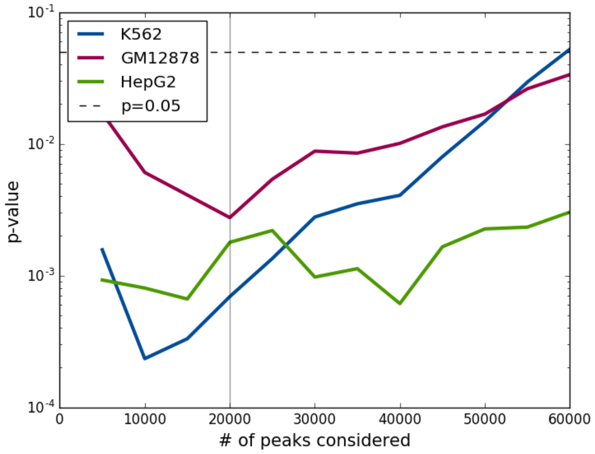

Figure 4: Statistical significance of the improvement of cloud control over standard control. P-values are calculated using different numbers of peaks considered in DNase-I hypersensitivity sites, using a paired student's t test (one sided). The null hypothesis is that there is no difference between the means of the area under PR curves for the cloud and standard treatment for the given number of peaks. The results shown in Figure 2 and Figure 3 focused on the top 20,000 peaks (indicated with the gray line).

of many control tracks allowed us to remove that noise more effectively.

4.2 Linear combinations of standard controls capture more noise in DNase-I hypersensitivity regions

Figure 6 shows the magnitudes of the weights β assigned to all 455 ENCODE control tracks by CloudControl for three ChIP-seq experiments: ENCFF00OBH, ENCFF002EBQ, and ENCFF002EDT. These are the experiments for which the cloud control improved the standard control the most in GM12878 and HepG2 (Figure 2). Notably, the regression model did not assign all the weight to the control tracks measured in the same cell type or in the same lab where the target experiments were performed. This highlights the importance of including many controls, even those measured in different cell types or labs. One can imagine using a single control as drawing one instance of background error from the true error model, so a single control track likely does not capture some aspects of the true error model. CloudControl utilizes a cloud of publicly available ChIP-seq control experiments to overcome the variance of the background noise.

The background noise for ChIP-seq experiments is a mixture of many factors, including the properties of the human genome and the biases introduced during the experimental procedures. In particular, it has been previously known that the sonication process preferentially makes breakage in the regions of genome that are more physically accessible, such as DNase-I hypersensitivity sites (DHSs) [9]. Figure 7 shows the enrichment of background signals of standard and cloud control experiments in DHSs across three different cell types. The standard controls have approximately 4-fold enrichment in the hypersensitivity regions compared to random expectation, whereas the cloud controls have approximately have 9-fold enrichment. This shows that the cloud control tracks remove signals in the hypersensitivity regions more actively than standard control. It is important to note that both the true background noise and the true binding sites for most transcription factors are known to be enriched in DHSs [18]. Combined with the result from the previous section, these results show that CloudControl preferentially removes the background noise more actively in DHSs, which improves the quality of called peaks inside DHSs.

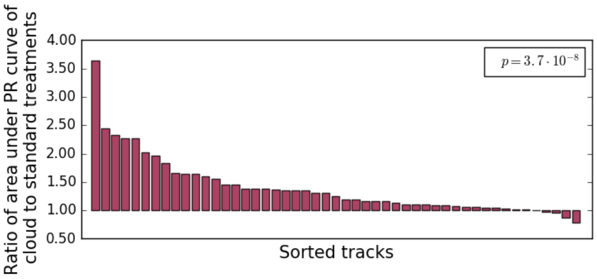

Figure 5: A waterfall plot showing the relative performance of the peaks called with cloud controls over standard controls in K561 using SISSRs. 49 tracks were examined. Each bar represents the ratio of area under the PR curve for a particular ChIP-seq track with the cloud treatment to that of the standard treatment. 20,000 peaks with the highest confidence values were used. 45 out of 49 tracks were improved when using cloud control compared with standard control.

4.3 CloudControl captures the difference in reproducibility between replicate and non-replicate pairs better.

Reproducibility is often used to assess the quality of ChIP-seq tracks and compare the performance of peak calling algorithms. The premise is that a pair of tracks that target the same protein in the same cell type should exhibit signals that are highly correlated [20, 8]. The ENCODE project uses the irreproducibility discovery rate (IDR) to evaluate the quality of ChIP-seq experiments [4, 12]. This measure is commonly used to evaluate a pair of biologically replicated experiments, and it is based on the assumption that the background noise signal is not correlated between the pair. However, our goal is to evaluate the quality of control tracks, not the target tracks themselves, and such an assumption might invalidate our evaluation. Instead, for each pair of experiments, we considered a binary classification problem, where we predict the presence of a peak in 1,000 base pair windows of one experiment by the p-values associated with the corresponding windows in the other experiment. We then measured the reproducibility of replicate pairs and non-replicate pairs using area under the PR curve. This essentially evaluates how recoverable an experiment is using peak signals from another experiment. If a pair of tracks target the same protein in the same cell type, the values should be high. On the other hand, if a pair does not target the same protein, then the value should in general be low.

The results are summarized in Figure 8. In all three cell

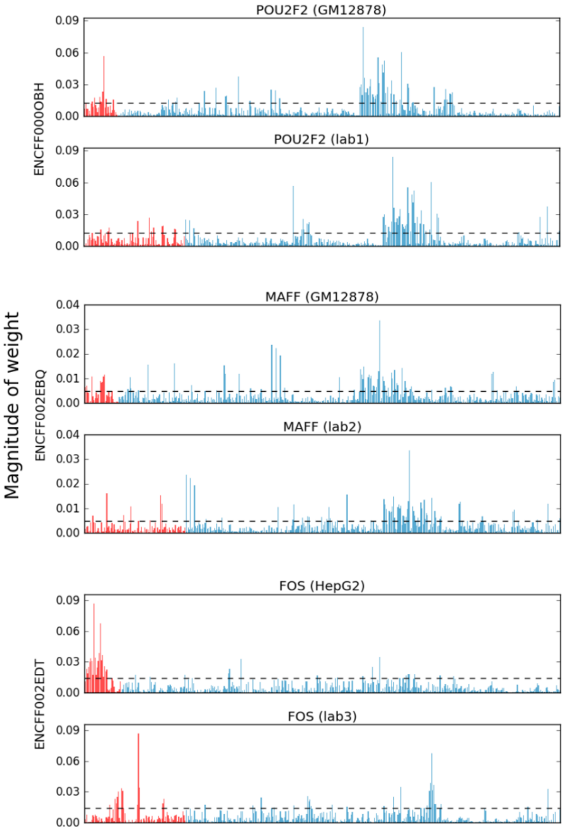

Figure 6: Bar-plots summarizing the magnitudes of the weights assigned to the ENCODE control tracks by the regression. The x axis represents 455 different control tracks. For each ChIP-seq experiment (ENCFF00OBH, ENCFF002EBQ, and ENCFF002EDT), the red bars in the top sub-figure indicate the controls from the same cell type as its target track. Similarly, the red bars in the bottom sub-figure indicate the controls from the same lab. The dotted line indicates the magnitude of the weight assigned to the standard control track for each experiment.

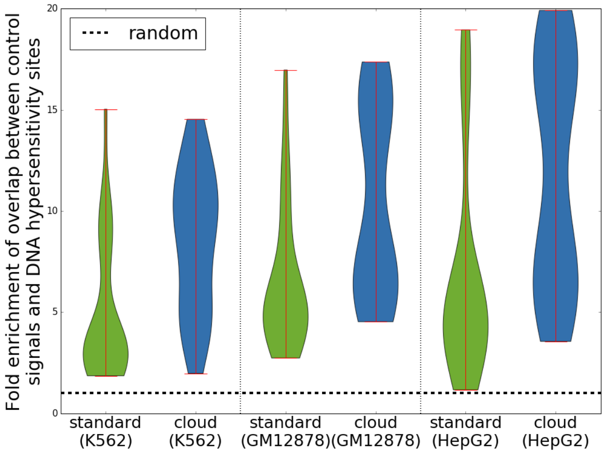

Figure 7: A violin plot showing the fold enrichment of overlap between control signals and DNase-I hypersensitivity sites relative to random expectation. The number of tracks examined are 53, 39, 34, 36, 35, 38 for the standard K562, cloud K562, standard GM12878, cloud GM12878, standard HepG2, and cloud HepG2 treatment, respectively. Here, the the number of control tracks are different between standard and cloud for the same cell type. The reason is that standard treatment uses one control track for each target track, whereas control treatment uses multiple control tracks for each target track. The total number of control tracks used for the same set of target tracks can be different between standard and cloud. Green bars represent standard controls and blue bars represent cloud controls.

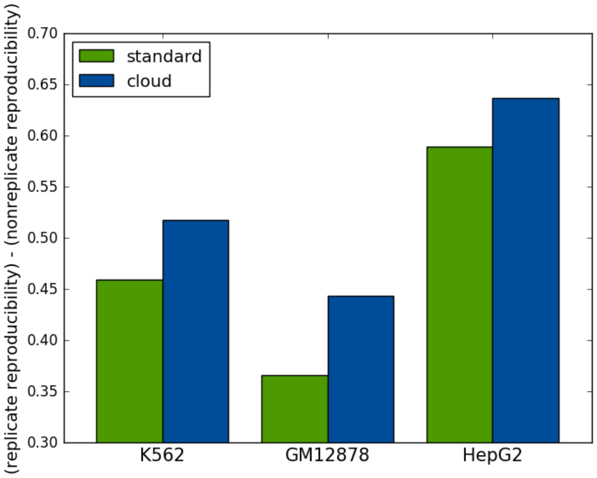

Figure 8: Difference in reproducibility between replicate pairs and non-replicate pairs. The number of tracks examined are 50, 24, 66 for the K562, GM12878, and HepG2 cell types, respectively. The reproducibility value of a pair of tracks is defined as the mean of the area under PR curves for predicting the presence of peaks in one track by the p-values of peaks in the other track.

types, the difference in the reproducibility values of replicate and non-replicate tracks were more pronounced when the cloud controls were used ($p < 0.0001$ for all cell types). Interestingly, in the K562 and HepG2 cell types, the reproducibility among the replicate pairs remained similar but the reproducibility among the non-replicate pairs decreased. On the other hand, in the GM12878 cell type, the reproducibility among the replicate pairs increased whereas the reproducibility among the non-replicate pairs remained unchanged. This is likely caused by the higher variability of control experiments performed in the GM12878 cell type compared to the K562 and HepG2 celltype. Good control experiments should make a pair of non-replicate target experiments dissimilar and a pair of replicate target experiments more similar. The trend shown in Figure 8 indicates that CloudControl can improve the quality of the current standard peak calling procedure.

4.4 Protein-protein interactions are better recovered by CloudControl

Table 1: Area under PR curves for predicting protein-protein interactions with positional correlation of peaks between pairs of tracks. The number of tracks investigated are 195, 119, and 84 for K562, GM12878 and HepG2, respectively. The cell types where the cloud treatment significantly improved over the standard treatment are marked with asterisks (p<0.0001) for both K562 and HepG2 cell types.

Treatment	K562	GM12878	HepG2
cloud	**0.1700*****	**0.1135*****	0.1473
standard	0.1353	0.1074	0.1472
no-control	0.1273	0.0877	0.0990

Using the binary vectors indicating the location of peaks, we computed a correlation matrix, where high Pearson correlation between track a and b suggests the interaction between target transcription factors of a and b. Three correlation matrices were created for peaks called with controls generated by CloudControl (cloud), standard controls (standard) and without any control (no-control). These correlation matrices were then compared to protein-protein interactions (PPIs) documented in the BioGrid database to evaluate the performance of the cloud, standard, and no-control treatments in revealing previously known PPIs [19]. Table 1 shows the area under precision recall (PR) curve for predicting PPIs with correlation matrices, using the BioGrid-supported PPIs as the ground truth. The area under the PR curves obtained using the cloud controls were significantly larger than those of no-control in all three cell types. The cloud controls performed significantly better than the standard controls in the K562 and GM12878 cell types. They performed at least as good as standard controls in HepG2.

Although the results seem promising, PPI enrichment as an evaluation metric is not as robust as motif presence that was presented in the previous section. First, the BioGrid database does not document all extant protein-protein interactions in all cell types. The number of interactions documented are likely biased towards the interactions that involve popularly studied transcription factors and cell lines. Second, PPIs in BioGrid do not always happen in close proximity to chromosomes. Because of the nature of ChIP-seq, we were only able to estimate PPIs that occur near chromosomes. These factors contributed to overall low values of area under PR curve presented in this analysis. However, despite these confounding factors, the general upward trend of the improvement in PPI enrichment by the cloud treatment suggests that control tracks generated by CloudControl can be used to substitute for and possibly outperform existing designated control tracks.

5. CONCLUSIONS

The current control procedure for ChIP-seq experiments uses only a single control track to remove background noise. However, a single control track might not recover the true background sufficiently, due to the stochastic nature of the background noise. To combat this problem, CloudControl regresses publicly available control tracks against a target track to determine the combination of control tracks that best estimate the true background model. One of the advantages of CloudControl is that it can easily be plugged into any existing pipeline because it produces BAM file that can simply substitute for ordinary control tracks. We have shown that the peaks called with the control tracks generated by CloudControl are superior to the peaks called with ordinary control tracks in terms of the following three aspects: (i) predicting putative motif binding sites for their target proteins, (ii) making biologically replicate experiments more similar and non-replicate experiments more dissimilar, and (iii) recovering known protein-to-protein interactions.

6. REFERENCES

[1] R. K. Auerbach, G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi, P. Lefrançois, K. Struhl, M. Gerstein, and M. Snyder. Mapping accessible chromatin regions using Sono-Seq. *Proceedings of the National Academy of Sciences*, 106(35):14926–14931, 2009.

[2] A. Barski, S. Cuddapah, K. Cui, T.-y. Roh, D. E. Schones, and Z. Wang. Resource High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, 2007.

[3] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, L. Arthur, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, S. Tarjei, J. A. Thomson, and T. S. Mikkelsen. The NIH Roadmap Epigenomics Mapping Consortium. *Naure Biotechnology*, 28(10):1045–1048, 2013.

[4] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[5] C. E. Grant, T. L. Bailey, and W. S. Noble. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[6] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acid Research*, 36(16):5221–5231, 2008.

[7] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, 26(12):1351–1359, 2008.

[8] T. D. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, and L. L. Elo. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC genomics*, 10:618, 2009.

[9] S. G. Landt, G. K. Marinov, A. Kundaje, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, 22(9):1813–1831, 2012.

[10] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(R25), 2009.

[11] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[12] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, 5(3):1752–1779, 2011.

[13] A. Mathelier, O. Fornes, D. J. Arenillas, C.-y. Chen, G. Denay, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 2015.

[14] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, et al. JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42:D142–D147, 2014.

[15] L. Narlikar and R. Jothi. ChIP-Seq Data Analysis: Identification of Protein-DNA Binding Sites with SISSRs Peak-Finder. *Methods Mol Biol*, 802:1–17, 2012.

[16] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology*, 27(1):66–75, 2009.

[17] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews Genetics*, 15(2):121–32, 2014.

[18] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

[19] M. Tyers, A. Breitkreutz, C. Stark, T. Reguly, L. Boucher, and B.-J. Breitkreutz. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(1):D535–539, 2006.

[20] E. G. Wilbanks and M. T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, 5(7), 2010.

[21] H. Xu, L. Handoko, X. Wei, C. Ye, J. Sheng, C. L. Wei, F. Lin, and W. K. Sung. A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, 26(9):1199–1204, 2010.

[22] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008.