In Silico Pooling of ChIP-seq Control Experiments

Guannan Sun[1], Rajini Srinivasan[2], Camila Lopez-Anido[2], Holly A. Hung[2], John Svaren[2], Sündüz Keleş[1,3]*

1 Department of Statistics, University of Wisconsin, Madison, Wisconsin, United States of America, **2** Waisman Center, Department of Comparative Biosciences, University of Wisconsin, Madison, Wisconsin, United States of America, **3** Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin, United States of America

Abstract

As next generation sequencing technologies are becoming more economical, large-scale ChIP-seq studies are enabling the investigation of the roles of transcription factor binding and epigenome on phenotypic variation. Studying such variation requires individual level ChIP-seq experiments. Standard designs for ChIP-seq experiments employ a paired control per ChIP-seq sample. Genomic coverage for control experiments is often sacrificed to increase the resources for ChIP samples. However, the quality of ChIP-enriched regions identifiable from a ChIP-seq experiment depends on the quality and the coverage of the control experiments. Insufficient coverage leads to loss of power in detecting enrichment. We investigate the effect of *in silico* pooling of control samples within multiple biological replicates, multiple treatment conditions, and multiple cell lines and tissues across multiple datasets with varying levels of genomic coverage. Our computational studies suggest guidelines for performing *in silico* pooling of control experiments. Using vast amounts of ENCODE data, we show that pairwise correlations between control samples originating from multiple biological replicates, treatments, and cell lines/tissues can be grouped into two classes representing whether or not *in silico* pooling leads to power gain in detecting enrichment between the ChIP and the control samples. Our findings have important implications for multiplexing samples.

Citation: Sun G, Srinivasan R, Lopez-Anido C, Hung HA, Svaren J, et al. (2014) *In Silico* Pooling of ChIP-seq Control Experiments. PLoS ONE 9(11): e109691. doi:10.1371/journal.pone.0109691

Editor: Noam Shomron, Tel Aviv University, Israel

Received December 3, 2013; **Accepted** August 7, 2014; **Published** November 7, 2014

Funding: This work was supported by National Institutes of Health Grants (HG006716 and HG007019) to S. K. and (NINDS (NS075269)) to J. S. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: keles@stat.wisc.edu

Introduction

Control experiments such as input DNA (whole cell extract) or ChIP with a non-specific antibody (e.g., IgG anti-serum) are commonly used to estimate background read distribution and are shown to be critical for identifying enriched regions in ChIP-seq experiments [1–4]. Standard ChIP-seq studies utilize paired controls by dividing the sample of interest into two parts, one of which is ChIPed with the antibody of interest and the other is used as the control sample. This practice typically limits the sequencing depths of control samples since resources are divided between ChIP and control samples and investigators are left wondering whether they have sufficient genomic coverage for their control samples. The issue of resources allocated to control samples remain highly significant in large scale ChIP-seq studies with 10s or 100s of diseased and non-diseased ChIPed individuals [5].

In [1], authors observed identification of more peaks when a ChIP-seq dataset was normalized against a more deeply sequenced input dataset. In a recent analysis [6], deeper sequencing of the input sample led to better detection specificity. The recent ENCODE ChIP-seq guidelines [7] suggest that the sequencing depth of the control experiment should be at least equal to and preferably larger than that of the ChIP sample.

We investigate *in silico* pooling designs, powered by the recent developments in sequencing technology, e.g., HiSeq 2000 of Illumina and multiplexing, for individually sequenced control samples. Although pooling of control samples generated within the same chromatin preparation step is commonly practiced, our investigation goes beyond this simple setting and inquires whether *in silico* pooling is a feasible option for input samples coming from different chromatin preparation steps. First, we establish that Pearson correlation between control samples within multiple biological replicates, multiple treatment conditions, and multiple cell/tissue types exhibit strong bi-modality indicating two classes, i.e., high and low, of within group correlations. Then, we show with multiple ChIP-seq experiments that *in silico* pooling designs increase the coverage for the control samples and lead to substantial power gain for the analysis of the ChIP-seq experiments within the high correlation class. Our computational experiments encompass datasets of wide range of genomic coverage from multiple organisms.

Materials and Methods

Mapped read files

We utilize a wide variety of datasets from worm *C. elegans*, rat, and human. Table 1 provides a list of these datasets.

ENCODE datasets. The publicly available human datasets are from four different groups participating in the ENCODE consortia [8]. These datasets consist of input control samples from biological replicates (with different chromatin preparation steps), multiple treatments within same cell line, and multiple cell lines. The sequencing depths of the samples vary between 1.8 and 46 million (M) reads. A detailed list of these datasets is available as Table S1 in File S1. We use these datasets for investigating similarities between input samples and focus on a small subset to evaluate the impact of *in silico* pooling on peak detection. We investigate *in silico* pooling on three classes of samples (Table 1).

Table 1. Description of the three classes of samples for *in silico* pooling analysis.

	Same treatment condition	Same tissue or cell line	Dataset
Multiple Biological Replicates Group (MBRG)	YES	YES	*C. elegans* Pha-4 (2 replicates) Rat Sox10 (2 replicates) *Hdac8 (K562) (2 replicates) Rat H3K27ac (2 replicates)
Multiple Treatments Group (MTG)	NO	YES	*c-Jun (K562) with IFN-Gamma treatment
Multiple Cell Lines Group (MCG)	YES	NO	*c-Fos (K562 vs. Huvec) *CTCF (Hct116 vs. Hpaf)

These three groups were evaluated separately for the feasibility and advantages of *in silico* pooling.
*: ENCODE dataset.
doi:10.1371/journal.pone.0109691.t001

The first class, named **M**ultiple **B**iological **R**eplicates **G**roup (MBRG), considers biological replicates that share same treatment or cellular conditions. The second class is **M**ultiple **T**reatments **G**roup (MTG) and encompasses samples that are from same type of cell lines or tissues but have undergone different treatments. The third class, **M**ultiple **C**ell Lines **G**roup (MCG) represents the most heterogeneous group and consists of samples from different cell lines and tissues but with similar treatment conditions. ENCODE input control samples have multiple samples in each of these three categories. We use c-Jun input samples from the K562 cell line profiled before and 30 minutes and 6 hours after stimulation with IFN-gamma as an example of MTG. The number of uniquely mapping reads for these experiments vary between 13 M and 18 M. As an example of MCG, we use c-Fos input samples from K562 and Huvec cell lines both of which were generated at UC Davis. In addition, we used a c-Fos sample from Huvec cell line that was generated at Yale University to investigate the lab effect on pooling. For these samples, the number of uniquely mapping reads are 29 M, 20 M, and 45 M, respectively. As another example of MCG, we consider CTCF profiling in Hct116 and Hpaf cell lines. These samples have 21 M and 20 M uniquely mapping reads, respectively. In addition, we also utilize two replicates of Hdac8 profiling in K562 cell lines as an example of MBRG. In short, these experiments encompass multiple ENCODE samples in each of the three classes of samples that we are interested in evaluating *in silico* pooling for.

***C. elegans* Pha-4 dataset.** This modENCODE dataset [9] profiles transcription factor Pha-4 in the first stage of *C. elegans* larval development (L1) under starvation conditions with matching input controls. The number of uniquely mapping reads of two biological replicate controls are 11 M and 15 M. These high coverage experiments represent an example from the MBRG.

Rat Sox10 dataset. This dataset is also an example of MBRG and profiles Sox10 in spinal cords of two biological replicate rats. The numbers of uniquely mapping reads for the controls are 27 M and 88 M, respectively.

Rat H3K27ac dataset. This dataset is also an example of MBRG and profiles acethylation of H3K27 (H3K27ac), which commonly marks enhancers [10] and exhibits broader peak patterns compared to TFs, in peripheral nerves of two biological replicate rats. The numbers of uniquely mapping reads for the controls are 23 M and 26 M, respectively.

Similarity calculation

In order to define a similarity measure between any two control samples, we first binned the genome into 200*bp* intervals and counted the number of reads within each bin after reads are extended to the average fragment length. Then, for the ENCODE samples, we consider the ENCODE reported peak lists (downloaded from hgdownload.cse.ucsc.edu) for which these input samples were used as control across multiple ChIP experiments and obtain a master list of ChIP-enriched regions. We use this master list to generate a read count vector across bins spanning these enriched regions for each sample. Dissimilarity between any two samples is based on one minus the Pearson correlation between their read count vectors. Hierarchical clustering of the samples based on normalized Euclidean distance between sample read count vectors is available as Figure S4 in File S1. For the other datasets, we first identify ChIP-enriched regions using MOSAiCS [11] and follow the above procedure with these peak lists. As a result, the reported Pearson correlations are between bin-level read count vectors spanning ChIP-enriched regions of the genomes.

In silico pooling

For all the computational experiments, we randomly sample from the set of uniquely mapping reads to generate datasets of lower depths. *In silico* pooling simply combines the reads of different samples. We repeat all the pooling experiments five times to assess variability. The sub-sampling proportions vary for different datasets. For the *C. elegans* Pha-4 controls, we first randomly sub-sample 0.5%, 5%, 25%, and 50% of each of the biological matching control sample, generating a total of 8 control samples. We then treat these as multiplexed samples representing different coverage. We generate *in silico* pooling samples from these by pooling two sub-samples with similar depth between the

Table 2. Summary of the in silico pooling experiments for the Sox10 samples.

	# of reads from R1	# of reads from R2	Gold Standard
Pool 1	$n_1/2$	$n_2/2$	$n_1/2+n_2/2$ from R2 (GS_1)
Pool 2	$n_1/2$	$n_1/2$	n_1 from R2 (GS_2)
Pool 3	n_1	n_1	$2n_1$ from R2 (GS_3)

R1 and R2 denote the two biological replicates.
doi:10.1371/journal.pone.0109691.t002

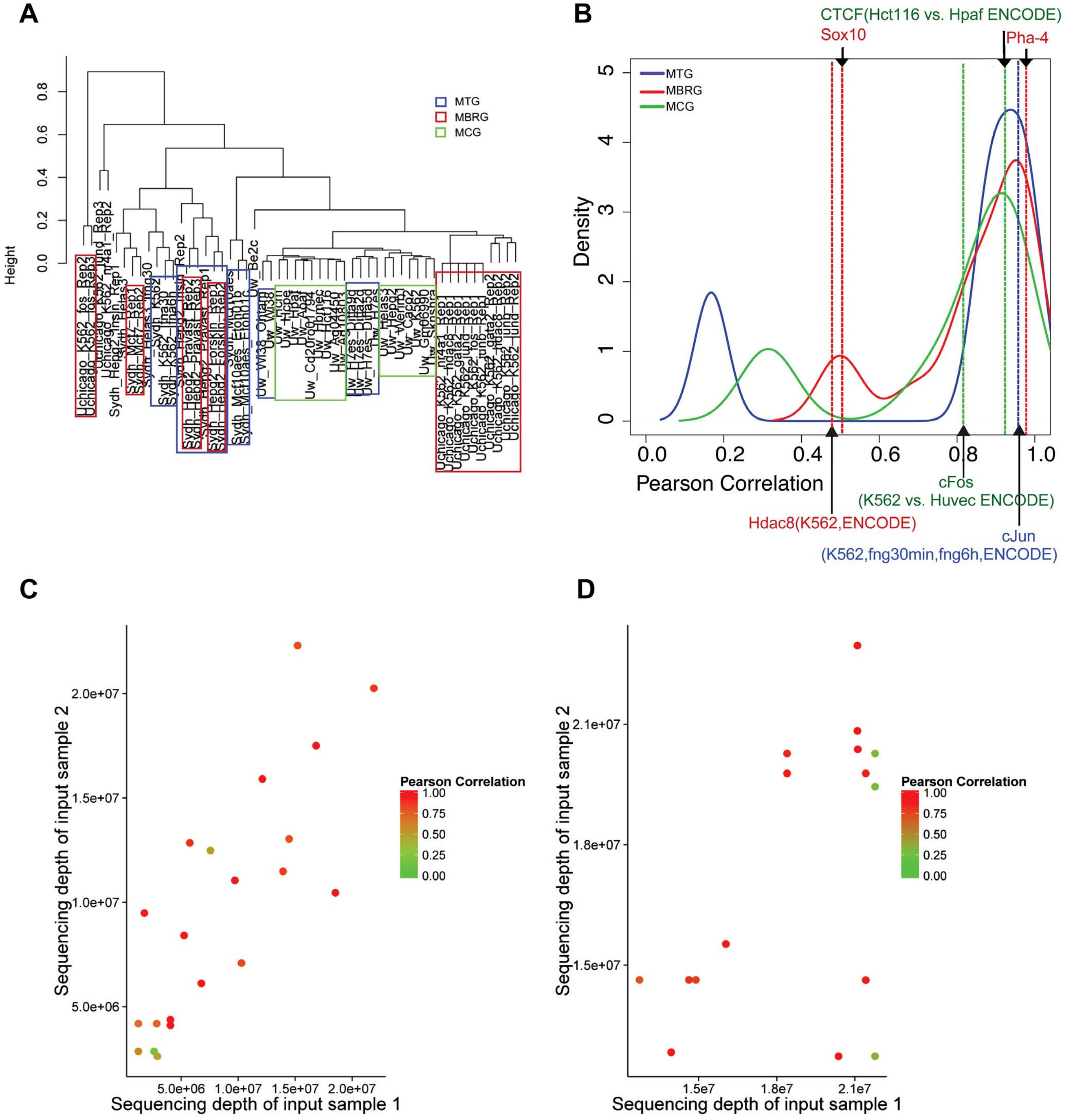

Figure 1. Exploratory analysis of ENCODE input samples. (a) Hierarchical clustering of 53 input samples from the ENCODE project. (b) Empirical density of the Pearson correlations within Multiple Biological Replicates (MBRG), Multiple Treatments (MTG), and Multiple Cell lines (MCG) Groups. (c) Effect of sequencing depth on the correlations within MBRG and MTG. (d) Effect of sequencing depth on the correlations within MCG. doi:10.1371/journal.pone.0109691.g001

biological replicates. This results in four *in silico* pooled samples with depths equivalent to 1%, 10%, 50%, and 100% of the full matching input sample, which we refer as the gold standard input sample.

For the rat Sox10 dataset, we use three types of *in silico* pooling and three types of gold standard inputs to accommodate the significant difference in the sequencing depths of the two biological replicates (Table 2). Let n_1 and n_2 denote the depths of the two replicates with $n_2 \approx 3n_1$. The first scheme pools half of each sample and uses an equal number of reads from replicate two as the gold standard. The second scheme pools $n_1/2$ reads from each of the replicates and compares with the gold standard of n_1 reads from the second sample. The third scheme pools n_1 reads from each sample and compares with $2n_1$ reads from the second replicate as the gold standard.

For the c-Jun dataset, we have four input samples with similar depths: two for the untreated sample and one for each of the two IFN-gamma stimulated samples. Therefore, we randomly sample a quarter from each of the input samples across three treatments for *in silico* pooling. For H3K27ac, c-Fos, and the other ENCODE datasets, we randomly sample half of the reads from each input sample for pooling.

Comparison between peak sets obtained with different control input samples

We perform all the peak identification with the two-sample MOSAiCS [11] model that compares the ChIP and the input samples. We illustrate the robustness of the results to the peak caller used by utilizing another commonly used peak caller, SPP [12] (Figure S1 in File S1). All the MOSAiCS fits are tuned so that the goodness-of-fit (GOF) plots exhibit good overall fits. We analyze each individual ChIP sample with its corresponding full size matching input sample to generate gold standard peak sets. We denote these peak sets by P_{match}, the peak sets obtained by lower depth matching input by $P_{match\_low}(r_1)$, and peak sets obtained by *in silico* pooling by $P_{pool}(r_2)$. Here, the depths of the lower depth and pooling samples are indexed by r_1 and r_2, where they each have r_1 and r_2 × (# of reads in the full matching input) reads, respectively. Our comparisons use P_{match} as the gold standard and compare the performances of P_{pool} and P_{match_low}. The overall design, unless specified otherwise, for k samples is that P_{match} is obtained by using full matching input control sample of depth n, for each of the k samples, P_{match_low} is by matching input control of depth n/k for each of the samples, and P_{pool} is by combining these k matching input controls to result in a pooled sample of depth n, i.e., $r_2 = r_1 \times k$. The evaluations are performed by comparing P_{match}, P_{pool}, and P_{match_low} for each individual ChIP sample and the overlap proportions, i.e., sensitivity, are averaged within each dataset. For example, for the rat Sox10 dataset, we perform a comparison between the three peaks sets for each of the two ChIP samples and report average. In addition, since each pooling experiment is repeated 5 times, the overall number of comparisons that the reported overlap proportions depend on vary from 10 to 20.

For the Sox10 dataset, we further use positive predictive value to compare the performances of P_{pool} and P_{match_low} with P_{match} (Figure S3(b) in File S1). In addition, we evaluate the consensus motif occurrence rates for the P_{match}, P_{pool}, and P_{match_low} peak sets of sequence specific TF Pha-4. Specifically, we use matrix scan utility of Regulatory Sequence Analysis Tools [13] with a p-value threshold of 0.01 to scan the peaks with the Pha-4 position specific weight matrix from JASPAR [14] (MA0546.1) (Figure S2 in File S1) and assess whether similar proportions of the P_{match}, P_{pool}, and P_{match_low} peaks have at least one motif.

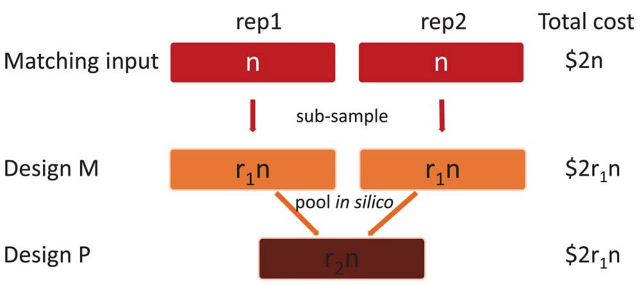

Figure 2. Matching (M) and *in silico* pooling (P) designs for $k = 2$ multiplexed samples. Designs M and P have the same cost since Design P simply pools reads from Design M *in silico* ($r_2 = 2 \times r_1$).
doi:10.1371/journal.pone.0109691.g002

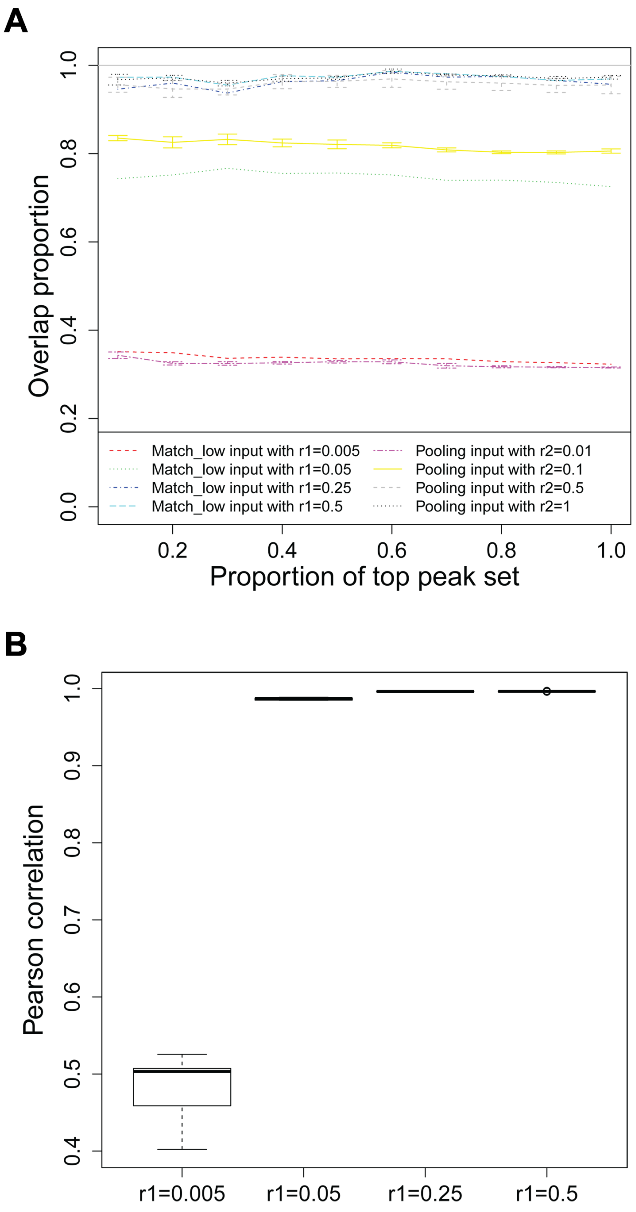

Figure 3. Evaluation of *in silico* pooling within MBRG with C. elegans Pha-4 experiments. (a) Overlap comparisons between $P_{pool}(r_2)$ and $P_{match\_low}(r_1)$ vs. P_{match}; (b) Correlations between sub-sampled inputs.
doi:10.1371/journal.pone.0109691.g003

Results

Clustering of multiple input samples from the ENCODE project

We hierarchically clustered 53 input control samples from the ENCODE project (Figure 1(a)) based on their bin-level input read count vectors over the master list of ChIP-enriched regions. Figure 1(a) reveals a strong lab effect where a few number of similar experiments (same cell type) performed by different labs cluster separately than their group. For example, although a few of the UChicago K562 inputs cluster with Snyder K562 input

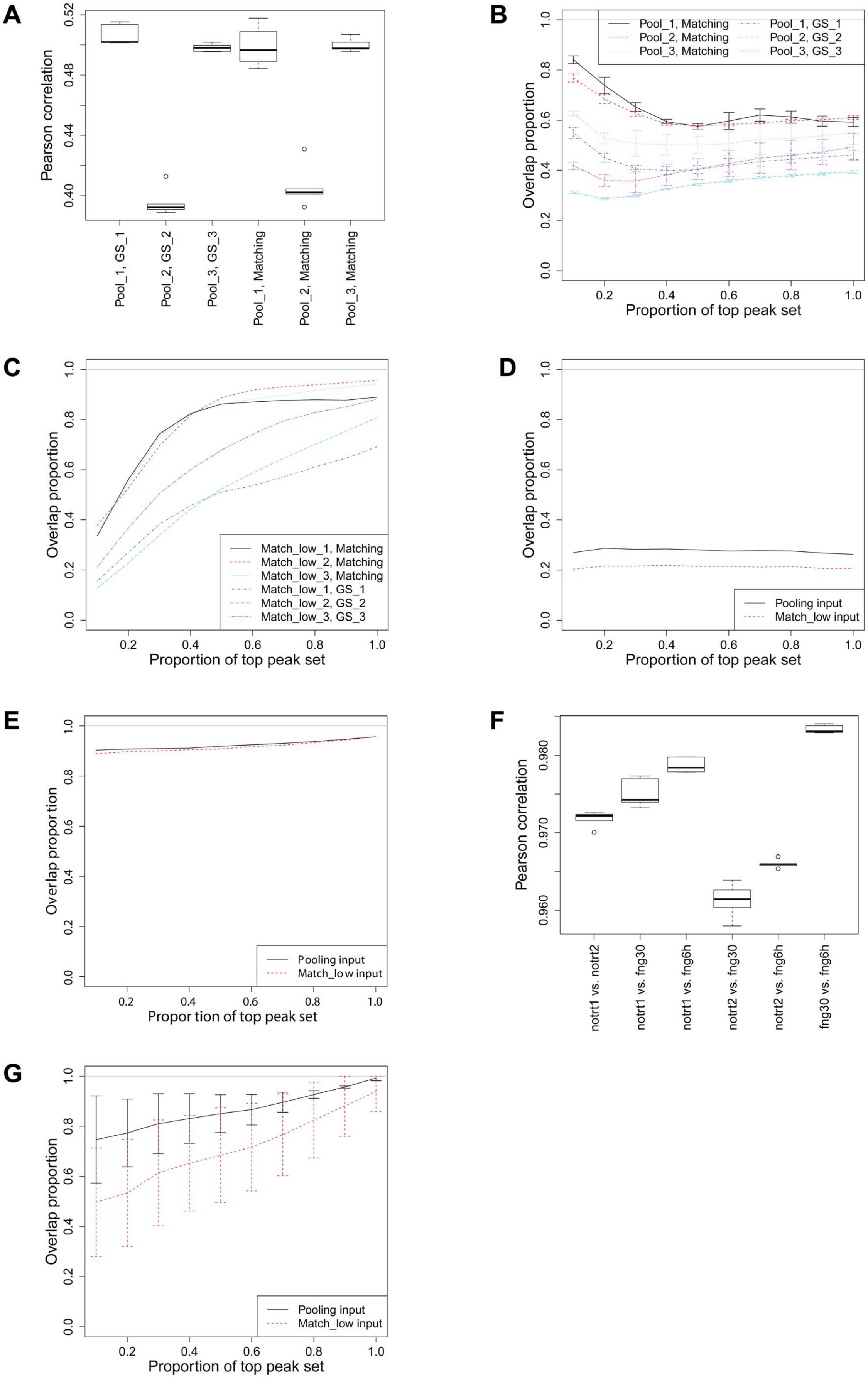
A
Pearson correlation
0.40
0.44
0.48
0.52
Pool_1, GS_1
Pool_2, GS_2
Pool_3, GS_3
Pool_1, Matching
Pool_2, Matching
Pool_3, Matching
B
Overlap proportion
Proportion of top peak set
Pool_1, Matching
Pool_2, Matching
Pool_3, Matching
Pool_1, GS_1
Pool_2, GS_2
Pool_3, GS_3
C
Overlap proportion
Proportion of top peak set
Match_low_1, Matching
Match_low_2, Matching
Match_low_3, Matching
Match_low_1, GS_1
Match_low_2, GS_2
Match_low_3, GS_3
D
Overlap proportion
Proportion of top peak set
Pooling input
Match_low input
E
Overlap propor tion
Propor tion of top peak set
Pooling input
Match_lo w input
F
Pearson correlation
0.960
0.970
0.980
notrt1 vs. notrt2
notrt1 vs. fng30
notrt1 vs. fng6h
notrt2 vs. fng30
notrt2 vs. fng6h
fng30 vs. fng6h
G
Overlap proportion
Proportion of top peak set
Pooling input
Match_low input
0.0
0.2
0.4
0.6
0.8
1.0

Figure 4. Evaluation of *in silico* pooling within MBRG (rat Sox10 and ENCODE Hdac8 datasets) and within MTG (c-Jun in K562 with three levels of treatment). (a) Sox10 correlations between two sub-sample inputs for all three pooling schemes based on peaks identified by using various gold standard input sets, where '(x, y)' refers to sub-sample inputs corresponding to pooling schemes x based on peaks by input y. (b) Sox10 overlap comparisons between P_{pool} and P_{match}, where '(x, y)' refers to comparison between P_x and P_y. (c) Sox10 overlap comparisons between P_{match_low} and P_{match}. (d) Hdac8 overlap comparisons between P_{pool} and P_{match_low} vs. P_{match}. (e) H3K27ac overlap comparisons between P_{pool} and P_{match_low} vs. P_{match}. (f) c-Jun correlations between four sub-sample inputs. (g) c-Jun overlap comparisons between P_{pool} and P_{match_low} vs. P_{match}.
doi:10.1371/journal.pone.0109691.g004

samples, majority of them cluster with the rest of the UChicago input samples. However, within each lab, biological replicates tend to cluster together as opposed to multiple treatments or cell line/tissue types with a few exceptions. Hepg2 input sample ("Sydh_hepg2_Pravast_rep1"), which has a much lower sequencing depth than the other two replicate samples of the same condition, cluster slightly more closely with another set of Hepg2 samples than these two biological replicates. The low sequencing depth of this sample is a likely explanation for this result. Although Pearson correlation is a scale-free measurement of similarity, it is affected by sequencing depths of samples, as is illustrated below and Figure 1(c). This clustering reveals a higher overall similarity within MBRG compared to MTG and MCG. Hierarchical clustering based on normalized Euclidean distance leads to the same conclusions (Figure S4 in File S1).

Next, we set out to assess the distribution of Pearson correlations within each of the biological replicates, treatments, and cell lines groups. Figure 1(b) depicts the empirical density of the Pearson correlations within each group. We observe a bi-modal density within each group and the samples can be divided into two classes by using a Pearson correlation threshold of 0.6 to 0.8. These bi-modal densities reveal that within each of MBRG, MTG, and MCG, there are pairs of samples with both *high* and *low* Pearson correlations. The median correlation of the low class for MBRG is higher than those of MTG and MCG, confirming the higher level of overall similarity for MBRG. We also investigate the effect of the sequencing depths on the correlation between samples. Figure 1(c) illustrates that when the sequencing depths are low, the correlations between samples that are biological replicates or are exposed to different treatments on the same cell/tissue types are lower than the correlations observed at higher sequencing depths. However, for samples from more diverse conditions, i.e., different cell types, low correlations arise irrespective of sequencing depths (Figure 1(d)).

In silico pooling within MBRG: *C. elegans* Pha-4 dataset

We use the deeply sequenced *C. elegans* Pha-4 experiments from modENCODE [15] for an in depth analysis of *in silico* pooling between two biological replicates. Our computational experiments are aimed at deciphering when *in silico* pooling is an advantageous strategy for ChIP-seq experiments over multiple experimental units, e.g., people, strains, animals, under same treatment and cellular conditions. We specifically compare two types of designs as illustrated in Figure 2. In design M, the replicates are multiplexed to generate matching input samples of size $r_1 \times$ (# of reads in the full matching input control sample) reads for each of two biological replicates. We varied r_1 as 0.005, 0.05, 0.25, and 0.5. 0.5 M reads in *C. elegans* correspond to $\sim$13 M reads on the mappable human genome and $\sim$11 M reads on the mappable rat genome; hence $r_1 = 0.05$ roughly corresponds to $\sim$0.5 M *C. elegans* reads and 13 M human reads. Then, each ChIP sample is analyzed with respect to its corresponding matching input sample which is smaller in size than the gold standard matching input sample. In design P, the multiplexed control samples are pooled *in silico*, generating a control sample with two times the depth of the individual samples, and each ChIP sample is analyzed with respect to this pooled sample. Figure 3(a) compares peaks identified using each of these different depth control samples across the two designs with the gold standard P_{match} as a function of the top $10,20,\cdots,100\%$ of the peak set. Peaks within each set are ordered with respect to their ChIP enrichment posterior probabilities reported by MOSAiCS. We observe that when $r_1 = 0.05$ leading to $P_{pool}(0.1)$, the overlap between P_{pool} and P_{match} is greater than 0.8, and when $r_1 \geq 0.25$, the overlap between P_{pool} and P_{match} is approaching 1. This indicates that *in silico* pooling, which only sequences $0.05 \times n$ reads for every biological replicate and yields a total of $0.1 \times n$ input control reads, generates the same set of ChIP-enriched regions as the matching input design where each biological replicate has n reads (a total $2 \times n$ input control reads). This is roughly a factor of 20 decrease in the overall cost of sequencing the input library for these set of experiments.

We observe the most significant improvement due to *in silico* pooling when the replicate control samples have depths as small as 0.1/2-th of the full matching control sample. At lower depth, neither $P_{match\_low}(r_1)$ nor $P_{pool}(r_2)$ is able to capture P_{match}; however, $P_{pool}(0.1)$ performs significantly better than $P_{match\_low}(0.05)$ indicating that *in silico* pooling is able to identify peaks in the same rank order as the P_{match}.

For lower depth with $r_1 = 0.005$, P_{pool} and P_{match_low} perform comparably, indicating that even for cases when pooling does not lead to improvements, it does not deteriorate the performance. For higher depths, P_{match_low} and P_{pool} perform very similarly to P_{match} indicating that these depths have reached saturation in terms of information usable for estimating background read distribution of the ChIP sample. Figure S2 in File S1 illustrates that with $r_1 = 0.25$, P_{pool} and P_{match_low} perform similarly to P_{match} in terms of Pha-4 consensus motif occurrence rate. Next, we investigate the Pearson correlations between the input controls of biological replicates for varying r_1. Figure 3(b) summarizes pairwise correlations between the replicates for different r_1 values. The correlation pattern exhibits a significant drop for $r_1 = 0.005$ indicating that pooling is not likely to increase the effective depth when individual input control samples are only 0.5% of the full matching input control sample. The value of the Pearson correlation between biological replicates at which *in silico* pooling provides significant improvement agrees well with the empirical cut-off we derive for dividing samples within the MBRG into two in Figure 1(b).

In silico pooling within MBRG: Rat Sox10 dataset

Because the two replicates have significantly different sequencing depths for the Sox10 experiments, we generate gold standard matching input controls based on the replicate with the larger depth (Table 2) (denoted as GS_1, GS_2, GS_3 in Figures 4(a), (b), and (c)) and use the ChIP sample from this replicate for evaluations. Results based on the actual full depth matching inputs (denoted as Matching) are also presented for comparison. Figure 4(a) displays correlations between the sub-sampled inputs based on peaks identified by using various gold standard input sets. For example, (pool_1, GS_1) denotes correlations between an input sample of size $n_1/2$ reads from replicate 1 and another input

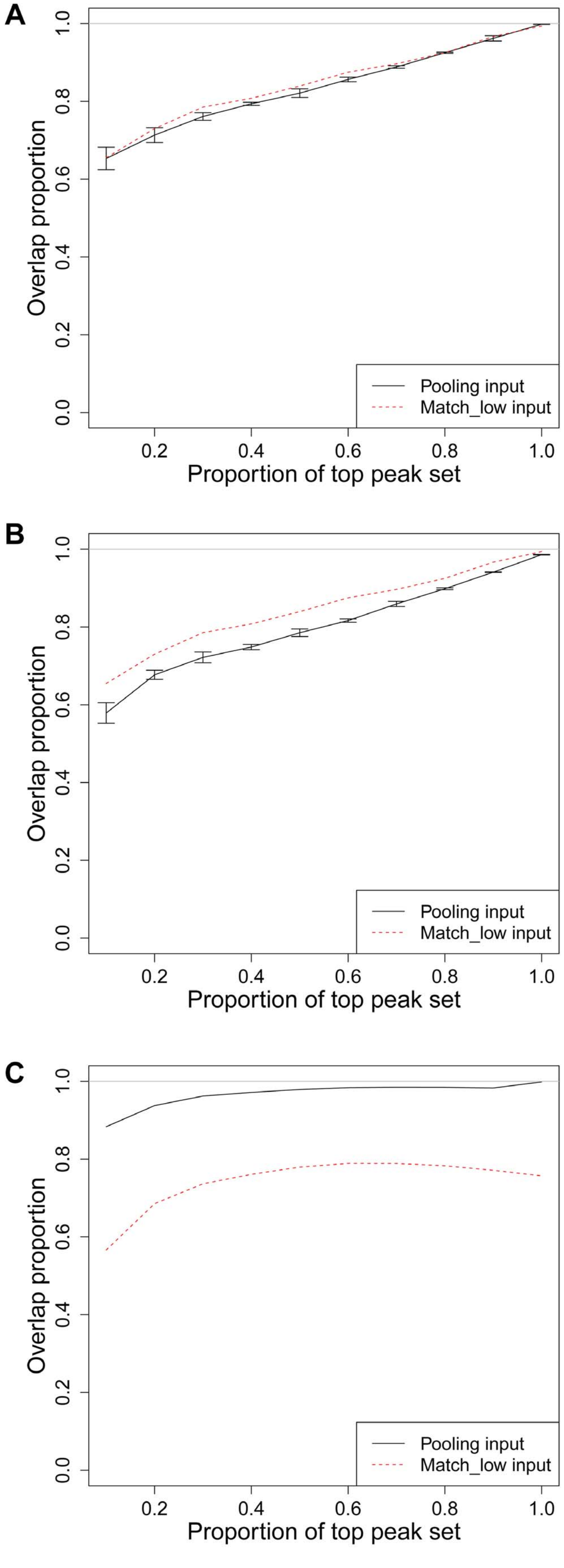

Figure 5. Evaluation of *in silico* pooling within MCG with ENCODE samples. a) c-Fos overlap comparisons between P_{pool} and P_{match_low} vs. P_{match} in cell lines K562 vs. Huvec. (b) c-Fos overlap comparisons between P_{pool} (P_{match_low}) and P_{match} in cell line Huvec with inputs from two different labs. (c) CTCF overlap comparisons between P_{pool} and P_{match_low} vs. P_{match} in cell lines Hct116 and Hpaf.
doi:10.1371/journal.pone.0109691.g005

sample of size $n_2/2$ reads from replicate 2. The correlations are based on bin-level data spanning top 10% of the ChIP-enriched regions that are identified when the ChIP replicate 2 is analyzed with its matching input of size $n_1/2+n_2/2$ reads from input replicate 2 (denoted as GS_1). Similarly, (pool_1, matching_1) denotes correlations between an input sample of size $n_1/2$ reads from replicate 1 and another input sample of size $n_2/2$ reads from replicate 2 where the ChIP-enriched regions are defined by analyzing both ChIP replicates with their original full matching input samples. Figure 4(a) illustrates that regardless of the pooling design and the gold standard used for the matching input sample, the correlations between sub-sampled input samples are lower than the empirical 0.6 to 0.8 threshold that partitions the Pearson correlations from the ENCODE samples into two groups. The effect of these low correlations is imminent on the comparison of P_{pool} with P_{match} for all the three pooling schemes with different gold standards. Figure 4(b) compares peaks from the three pooling schemes to peaks from two types of matching inputs and reveals that none of the pooling schemes perform as well as the corresponding gold standard or matching input P_{match}. We also compare P_{match_low} of replicate 2 corresponding to three pooling approaches with peaks from different gold standards (Figure 4(c)), e.g., Match_low_3 has n_1 reads from replicate 2 as opposed to $2n_1$ reads in GS_3 and n_2 reads in Matching. The comparison between Figures 4(b) and 4(c) indicate that P_{pool} still performs better than P_{match_low} in terms of identifying peaks in the same rank order as in P_{match} for the top 20% of the peaks. However, the overall sensitivity of P_{pool} is lower than P_{match_low} since the number of peaks including true positives and false positives in P_{match_low} is significantly higher than in P_{pool} (Figure S3(a) in File S1). Figure S3(b) in File S1 compares the positive predictive values between P_{pool_1} and $P_{match\_low\_1}$ to the corresponding P_{match}, as well as P_{pool_2} and $P_{match\_low\_2}$ to the corresponding P_{match}. Neither of them have high positive predictive values and P_{match_low} is slightly below P_{pool} due to too many false positives. Therefore, when the correlation between two replicates is lower than the empirical 0.6 to 0.8 threshold, increasing the sequencing depths with further sequencing might be a better strategy.

In silico pooling within MBRG: randomly selected ENCODE Hdac8 and histone H3K27ac datasets

We randomly pick one of the ENCODE datasets (Hdac8 profiling in K562) that exhibits low Pearson correlation across samples within the MBRG. The correlation between the two replicates of the Hdac8 dataset is only 0.48, which classifies this sample into low correlation class in the empirical density of the Pearson correlations (Figure 1(b)). Figure 4(d) displays P_{match}, P_{pool}, and P_{match_low} comparisons of these samples. Although, both P_{pool} and P_{match_low} exhibit very little overlap with P_{match}, P_{pool} is significantly better than P_{match_low}. This suggests that if an experiment yields multiplexed samples of low depth, the optimal procedure is to increase the depth of each biological replicate input control sample; however, when this is not feasible, an analysis based on pooled control samples is better than using low depth matching input for each replicate.

As a final illustration for MBGR, we use the rat H3K27ac ChIP-seq dataset for two biological replicates. Figure 4(e) displays the comparisons of P_{match}, P_{pool}, and P_{match_low}. P_{pool} and P_{match_low} perform almost as good as P_{match} since the Pearson correlation between sub-sampled input samples is extremely high (0.99).

In silico pooling within MTG: ENCODE c-Jun (K562) dataset

The sub-sampled inputs of the c-Jun dataset exhibit very high correlations despite their low sequencing depths ($\sim$ 20 M mappable reads, Figure 4(f)). Figure 4(g) compares P_{match} with P_{pool} and P_{match_low} and reports averages of overlapping proportions when each of the three ChIP samples representing different treatments are analyzed with respect to matching, pooling, low depth matching input samples. The overlap performances agree well with the observed high correlations. P_{pool} compares well with P_{match} and is significantly better and less variable than P_{match_low}.

In silico pooling within MCG: ENCODE c-Fos (K562 vs. Huvec) and CTCF (Hct116 vs. Hpaf) datasets

Figure 5(a) reports P_{match}, P_{pool}, and P_{match_low} comparisons of the c-Fos analysis in K562 and Huvec cell lines. The reported overlap proportions are averages of the analysis of the K562 and Huvec c-Fos ChIP samples with varying input controls. Both P_{pool} and P_{match_low} have rankings of the enriched regions different than that of P_{match}; however, both controls are able to identify enriched regions captured by P_{match}. The correlations between the two sub-sampled inputs vary around 0.83 across different sub-sampling experiments and support pooling. However, as illustrated in Figure 5(a), both the individual matching sub-sample inputs and the *in silico* pooling inputs perform equally well. Since c-Fos is profiled in Huvec cells by two different labs, we also evaluated the possibility of pooling input samples generated by different labs. The correlations between low depth input samples generated from the input samples of two different labs varied around 0.82 across different sub-sampling experiments. Figure 5(b) illustrates that even though both P_{pool}, obtained from combining input samples from two labs, and P_{match_low} are able to identify peaks of P_{match}, there is a significant difference between P_{pool} and P_{match_low} in terms of the ranking of the peaks. In particular, pooling input samples across different labs seems to generate a worse ranking than that of using a low depth matching input.

As another example of *in silico* pooling within MCG, we randomly picked the CTCF ChIP-seq datasets from Hct116 and Hpaf cells from the set of MCG samples with high Pearson correlation, e.g., correlation between the full matching input samples of these cells is 0.93. Figure 5(c) displays the comparisons of P_{match}, P_{pool}, and P_{match_low}. P_{pool} which has twice the depth of P_{match_low} performs much better than P_{match_low} and is almost as good as P_{match}.

Discussion

We investigated the effect of *in silico* pooling of control samples within MBRG, MTG, and MCG across multiple datasets with varying levels of genomic coverage from multiple organisms. Our computational experiment based on vast amounts of ENCODE data show that pairwise Pearson correlations between control samples originating from MBRG, MTG, and MCG exhibit a strong bi-modality indicating two classes representing whether or not *in silico* pooling may lead to power gain in detecting enrichment between the ChIP and the control samples. We demonstrate that *in silico* pooling designs increase the coverage of control samples and perform as well as the matching input within the high correlation class using multiple ChIP-seq experiments. Comparison of the input samples with simple Pearson correlation across candidate enrichment regions provided consistent results across the large numbers of datasets we have considered. Although we considered more formal testing approaches (e.g., differential enrichment analysis of the input samples with DBChIP [16] and irreducible discovery rate (IDR) [17]) for comparing two or more input samples, the overall insufficient depths of the input samples (median depths ranged between 0 to 28 for the *C. elegans* samples; 2 to 4 for the rat Sox10 samples, and is 2 for the ENCODE c-Fos samples) resulted in discrete p-values with many ties and prohibited robust application of these approaches.

Our findings have important implications for multiplexing control samples in ChIP-seq experiments, especially for samples from MBRG. Specifically, they endorse starting with relatively low coverage of multiplexed control samples and calculating their pairwise Pearson correlations for candidate enrichment regions that may be obtained by an initial comparison of ChIP and control samples. If the pairwise correlations belong to the empirical high correlation class, multiplexed samples can be pooled as the *in silico* pooling input and used for peak detection. Otherwise, increasing the coverage of control samples by further sequencing is essential to gain power in detecting enrichment between the ChIP and the control samples. Although the multiplexing control samples with relatively low coverage are only available control samples for candidate enrichment region selection, our experiments based on *C. elegans* Pha-4 and ENCODE Hdac8 data show that the pairwise input sample correlations are very comparable for candidate enrichment regions that are obtained by full matching input and lower depth inputs.

Supporting Information

File S1 Supporting figures and tables.
(PDF)

Acknowledgments

This work was supported by National Institutes of Health Grants (HG006716 and HG007019) to S.K. and (NINDS (NS075269)) to J.S.

Author Contributions

Conceived and designed the experiments: GS SK. Performed the experiments: RS CL HH JS. Analyzed the data: GS SK. Contributed reagents/materials/analysis tools: GS SK. Wrote the paper: GS SK.

References

1. Ho J, Bishop E, Karchenko P, Negre N, White K, et al. (2011) ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. BMC Genomics 12: 134.
2. Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M (2008) Modeling ChIP sequencing in silico with applications. PLoS Computational Biology 4: e1000158.
3. Blahnik KR, Dou L, O'Geen H, McPhillips T, Xu X, et al. (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. Nucleic Acids Research 38: e13.
4. Nix DA, Courdy SJ, Boucher KM (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC bioinformatics 9: 523.

5. RoadmapEpigenome. The NIH Roadmap Epigenomics Mapping Consortium. http://www.roadmapepigenomics.org/.
6. Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, et al. (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. Nature Methods 9: 609–12.
7. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research 22: 1813–1831.
8. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.
9. W Zhong PWS (2006) Genome-wide prediction of *C. elegans* genetic interactions. Science 311: 1481–4.
10. Vicky W, Zhou BEB, Goren A (2010) Charting histone modifications and the functional organization of mammalian genomes. Nature Reviews Genetics: 718.
11. Kuan PF, Chung D, Pan G, Thomson JA, Stewart R, et al. (2011) A statistical framework for the analysis of ChIP-Seq data. Journal of the American Statistical Association 106: 891–903.
12. Kharchenko PV, Tolstorukov M, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotechnology 6: 1351–1359.
13. Turatsinze J, Thomas-Chollier M, Defrance M, Van Helden J (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nature Protocols 3: 1578–1588.
14. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, et al. (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Research.
15. The modENCODE Project Consortium (2009) Unlocking the secrets of the genome. Nature 459: 927–30.
16. Liang K, Keleş S (2012) Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics 28: 121–122.
17. Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics 5: 1752–1779.