DRAFT

Data Exploration, Quality Control, and Statistical Analysis of ChIP-exo/nexus Experiments

Rene Welch [1†], Dongjun Chung [6†], Irene Ong[3], Jeffrey Grass[3,4], Robert Landick[3,4,5] and Sündüz Keleş[1,2*]

[*]Correspondence: keles@stat.wisc.edu
[1]Department of Statistics, University of Wisconsin Madison, 1300 University Avenue, Madison, WI
Full list of author information is available at the end of the article
[†]These two authors contributed equally.

Abstract

ChIP-exo/nexus experiments present modifications on the commonly used ChIP-seq protocol for high resolution mapping of transcription factor binding sites. Although many aspects of the ChIP-exo data analysis are similar to those of ChIP-seq, these high throughput experiments present a number of unique quality control and analysis challenges. We develop a statistical quality control pipeline and accompanying `R` package, `ChIPexoQual`, to enable exploration and analysis of ChIP-exo and related experiments. `ChIPexoQual` evaluates a number of key issues including strand imbalance, library complexity, and signal enrichment of data from ChIP-exo/nexus experiments. Assessment of these properties are facilitated through diagnostic plots and summary statistics calculated over regions of the genome with varying levels of coverage.

We evaluated our QC pipeline with both large collections of public ChIP-exo/nexus data and multiple, new ChIP-exo datasets from *E. Coli*. `ChIPexoQual` analysis of these datasets resulted in guidelines for using these QC metrics across a wide range of sequencing depths and further insights for modeling ChIP-exo data. Finally, although ChIP-exo experiments have been compared to ChIP-seq experiments with single-end (SE) sequencing, we provide, for the first time, comparisons with paired-end (PE) ChIP-seq experiments. We illustrate that, at fixed sequencing depths, ChIP-exo provides higher sensitivity, specificity, and spatial resolution than PE ChIP-seq and both significantly outperform their SE ChIP-seq counterpart. Furthermore, we show that ChIP-exo and PE ChIP-seq are comparable in sensitivity for closely located binding events, but as the average distance between binding events increases, ChIP-exo exhibits higher sensitivity than PE ChIP-Seq. SK: The last sentence does not really make much sense?

Keywords: ChIP-exo; ChIP-nexus; ChIP-seq; Statistical Quality Control; Spatial Resolution; Transcription Factor; Binding Site Identification with High-Resolution; Deconvolution

Background

Chromatin Immunoprecipitation followed by exonuclease digestion and next generation sequencing (ChIP-exo) is currently one of the state-of-the-art high throughput assays for profiling protein-DNA interactions at or close to single base-pair resolution [1]. It presents a powerful alternative to popularly used ChIP-seq (Chromatin immunoprecipitation coupled with next generation sequencing) assay. ChIP-exo experiments first capture millions of DNA fragments (150 - 250 bps in length) that the protein under study interacts with using a protein-specific antibody and random fragmentation of DNA. Then, λ-exonuclease (λ-exo) is deployed to trim the $5'$ end of each DNA fragment to each protein-DNA interaction boundary. This step is unique to ChIP-exo and aims to achieve significantly higher spatial resolution compared to ChIP-seq. Finally, high throughput sequencing of a small region (36 to 100 bps) at the $5'$ end of each fragment generates millions of reads. Similarly, ChIP-nexus (Chromatin Immunoprecipitation followed by exonuclease digestion, unique barcode, single ligation and next generation ligation) [2] is a further modification on the ChIP-exo protocol. ChIP-nexus aims to overcome limitations of ChIP-exo by by yielding high complexity libraries with numbers of cells comparable to that of ChIP-seq experiments. This is achieved by reducing the numbers of ligations in the standard ChIP-exo protocol from two to one, and adding unique, randomized barcodes to adoptors to enable monitoring of overamplification. Figure 1 illustrates the differences between different ChIP-based protocols: ChIP-exo, single-end (SE) ChIP-seq, paired-end (PE) ChIP-seq, ChIP-nexus. The $5'$ ends of a ChIP-exo/nexus experiment are clustered more tightly around the binding sites of the protein than in a ChIP-seq experiment. In a PE ChIP-seq experiment, both ends are sequenced as opposed to only the $5'$ end in a SE ChIP-seq. SK: Note to Rene: ChIP-Nexus paper has a good description of what can go wrong with ChIP-exo; It seems like most of the read imbalance could also be due to ligation inefficiency. Although this probably does not explain why we have reads only from one strand in some regions. This is more likely explained by over digestion in one strand or single stranded TF-DNA interaction. Think a bit more on these issues and understand how the ligation could give rise to strand imbalance, we can ask Bob to have a closer look at the section where we have this discussion.

Although ChIP-exo/nexus protocols are being adopted by research community, features of ChIP-exo data, especially those pertaining to data quality, have not been investigated much. The key features of ChIP-exo/nexus that separate them from ChIP-seq are broadly as follows. First, DNA libraries generated by the ChIP-exo protocol are expected to be less complex than the libraries generated by ChIP-seq [3]) because digestion by λ-exo is expected to restrict the space of genomic positions that sequencing reads can map to, to small local regions around the actual binding sites. Therefore, in high quality and especially deeply sequenced ChIP-exo datasets, it is possible to observe large numbers of reads accumulating at a small number of bases due to actual signal rather than overamplification bias as commonly observed in ChIP-seq experiments. Second, although we expect approximately the same numbers of reads from both DNA strands at a given binding site, there may be locally more reads in one strand than in the other, owing to λ-exo efficiency, ligation efficiency, or other factors. This is a key point with implications on the statistical analysis of ChIP-exo data. Specifically, currently available ChIP-exo specific

statistical analysis methods (e.g., Mace [4], CexoR [5], and Peakzilla [6]) rely on the existence of peak-pairs formed by forward and reverse strand reads at the binding site. Finally, most of current widely used ChIP-seq quality control (QC) guidelines [7] may not be directly applicable to ChIP-exo data.

To address these challenges, we develop a suite of diagnostic plots and summary statistics and implement them in a versatile `R package` named `ChIPexoQual`. We apply this pipeline on a large collection of public and newly generated ChIP-exo/nexus data. We validate implications of the QC pipeline by evaluation of the samples for features that capture high signal to noise such as occurrences of motifs recognized by the profiled DNA interacting protein. Our analyses of this large collection of data revealed that the so-called ChIP-exo peak-pair assumption is subject to violations. To further address this and provide a platform where ChIP-exo and ChIP-seq experiments can be evaluated with comparable methods, we assess performances of recently developed methods suitable for ChIP-exo analysis, including dpeak [8] and GEM [9]. We observe that dPeak performs as good or better than the available ChIP-exo methods and provides a platform where PE and SE ChIP-seq can be compared with their ChIP-exo counterpart. Our comparisons of PE ChIP-seq with ChIP-exo interestingly highlights that while ChIP-exo outperforms PE ChIP-seq in terms of resolution and detection power, both are significantly better than SE ChIP-seq.

Results and discussion

Publicly available ChIP-exo/nexus and novel *E. coli* ChIP-seq/exo datasets

We utilized a rich collection of publicly available ChIP-exo/nexus data from multiple organisms to build and evaluate our quality control pipeline (Table 1). These include: CTCF factor in human HeLa cells [1]; ER factor in human MCF-7 cells [10]; GR factor in IMR90, K562, and U2OS human cells [11]; TBP factor in human K562 cells [12]. ChIP-nexus data included experiments from [2] profiling TBP in human K562 cells, MyC and Max in *D. Melanogaster* S2 cells, and Twist and Dorsal in *D. Melanogaster* embryo.

In order to have a setting where we can compare SE and PE ChIP-seq with their ChIP-exo counterpart, we profiled σ^{70} under a variety of conditions in *E. coli* with both ChIP-exo (Table 2) and PE ChIP-seq (SK: Table SXX). Collectively, we generated σ^{70} factor ChIP-exo, PE and SE ChIP-seq experiments under aerobic ($+O_2$) and anaerobic ($-O_2$) conditions in glucose minimal media. Similarly, we generated σ^{70} factor ChIP-exo, PE and SE (generated *in silico*) ChIP-seq experiments in *E. Coli* under aerobic ($+O_2$) conditions with and without rifampicin treatment. We refer to these latter set of experiments as time point 20 minutes and time point 0 minutes SK: Check for consistency with other parts of the paper and correct, ok to use different naming convention. We specifically used these experimental data for comparisons of ChIP-exo and PE ChIP-seq assays in identifying closely spaced binding events and their resolution, i.e., physical proximity of the predicted events to the actual binding sites. This comparison benefits from using σ^{70} which is a transcription initiation factor of housekeeping genes in *E. Coli*. Many *E. coli* promoters contain multiple transcription start sites (TSS). These TSSs are often closely spaced, i.e., within $10 \sim 150$ bps of each other, and are considered to be multiple "switches" that differentially regulate gene expression under diverse growth conditions [13].

ChIP-exo versus ChIP-seq: general features

Read distributions within signal and background regions. We first compared ChIP-seq and ChIP-exo in terms of data features that are well studied in ChIP-seq studies. Our σ^{70} ChIP-seq and ChIP-exo samples from *E. coli* are especially well suited for this task since they are all deeply sequenced compared to the genome size of *E. coli*. Figure 2 summarizes this comparison for one biological replicate of ChIP-exo and ChIP-seq experiments from the same biological conditions (samples x and y from Table 2 SK: Maybe add a column as sample number to this table to make referencing easy). Comparisons with other paired *E. coli* ChIP-seq and ChIP-exo samples led to similar conclusions (Supplementary Figure SK: SXXX). We summarized the extended raw read counts within 150 bps non-overlapping intervals, i.e., bins, interrogating the genome. Figure 2(A) depicts that, as expected, ChIP read counts from ChIP-exo and ChIP-seq are linearly correlated especially at high read counts. This indicates that signals for potential binding sites are well reproducible between ChIP-exo and ChIP-seq data. In contrast, there is a clear difference among the two data types for bins with low read counts, highlighting expected differences in the background read distributions of the two data types. SK: ***After seeing these types of plots from other samples (e..g., ER in MCF7, it is unlikely that what we are seeing from σ^{70} samples regarding background reads (the next few sentences) is typical. For CTCF and σ^{70}, the ChIP-exo samples have a wider dynamic range, but this could be due to differences in sequencing depth. *** Specifically, background reads from ChIP-seq are almost uniformly distributed over background (non-enriched) regions. In comparison, background regions in ChIP-exo show larger variation but have overall lower read counts. Furthermore, there are considerably large numbers of bins with zero read counts in ChIP-exo and non-zero read counts in ChIP-seq. Overall, these observations indicate that a much smaller portion of the genome is expected to be background in ChIP-exo compared to ChIP-seq and methods that specifically model the background read distribution might benefit from acknowledging this.

Peak-pair assumption. We next evaluated the peak-pair assumption, i.e., a cluster of reads in the forward strand is usually paired with a cluster of reads in the reverse strand that is located on the right-hand-side of the binding site, that is commonly utilized in designing statistical analysis methods for ChIP-exo data [4, 5, 6]. We considered the set of peaks identified in both the ChIP-seq and ChIP-exo samples as high quality peaks (Materials and Methods) and calculated the proportion of forward strand reads in these regions (Figure 2(B)). This plot reveals a higher level of strand imbalance for ChIP-exo compared to ChIP-seq. Potential reasons for this observation include ligation efficiency, efficiency of λ-exo digestion, and single-stranded protein-DNA interactions. Overall, such an imbalance is prevalent in SK: X% of the ChIP-exo samples used in this paper.

Mappability and GC-content bias. We next evaluated ChIP-exo data of CTCF in HeLa cells [1] to investigate biases inherent to next generation sequencing experiments with eukaryotic genomes. Figures 2(C) and 2(D) display the bin-level average read counts against mappability and GC-content. Each data point is obtained by averaging the read counts across bins with the same mappability of GC-content. These biases, increasing linear trend with mappability and non-linear trend with

GC-content, are similar to those observed in ChIP-seq datasets [14, 15, 16]. This observation indicates that analysis of ChIP-exo data should benefit from methods that take into account apparent sequencing biases such as mappability and GC content, mostly when an input control sample is not available.

Application of ENCODE ChIP-seq quality metrics to ChIP-exo and ChIP-nexus data

ENCODE consortium established empirical and widely used QC metrics on ChIP-seq data [7]. Currently, these constitute the state-of-the-art QC pipelines for these high throughput experiments. We evaluated how these metrics, namely Normalized Strand Cross-Correlation (NSC), Relative Strand Cross-Correlation (RSC), and PCR Bottleneck Coefficient (PBC) defined at `https://genome.ucsc.edu/ENCODE/qualityMetrics.html` [7], behave on ChIP-exo/nexus data in Tables 1 and 2.

Protocol	Organism	TF	Cell type	Rep.	Depth	NSC	RSC	PBC
ChIP-exo	Human	CTCF	HeLa		48,478,450	16.02	1.1960	0.4579
	Human	ER	MCF-7	1	9,289,835	19.87	1.0127	0.8082
				2	11,041,833	21.48	1.0063	0.8024
				3	12,464,836	18.72	1.0100	0.8203
	Mouse	FoxA1	Liver	1	22,210,461	21.28	1.1104	0.6562
				2	23,307,557	60.42	1.1604	0.7996
				3	22,421,72	72.04	1.1975	0.1068
	Human	GR	IMR90		47,443,803	8.86	1.3678	0.2978
			K562		116,518,000	4.11	1.0441	0.0504
			U2OS		3,255,111	10.05	1.0288	0.7714
	Human	TBP	K562	1	61,046,382	12.01	1.1119	0.1232
				2	94,314,770	7.93	1.0299	0.1681
				3	114,282,270	9.25	1.1027	0.1464
ChIP-nexus	D.Melanogaster	Dorsal	embryo	1	8,863,170	7.27	1.0402	0.6766
				2	10,003,562	7.19	1.0672	0.5656
		Twist		1	18,244,203	5.82	1.1637	0.6592
				2	52,546,982	5.27	1.1805	0.4549
		Max	S2	1	18,320,743	3.60	1.3628	0.5178
				2	24,965,642	3.47	1.0138	0.2124
		MyC		1	7,832,034	5.92	1.0115	0.3935
				2	22,824,467	5.76	1.0045	0.1879
	Human	TBP	K562	1	33,708,245	32.16	1.1712	0.3102
				2	129,675,001	32.70	1.2455	0.0492

Table 1 Summary of publicly available data used for development and evaluation of `ChIPexoQual`. The last three columns depict ENCODE QC metrics on these data: NSC: Normalized Strand Cross-Correlation; RSC: Relative Strand Cross-Correlation; PBC: PCR Bottleneck Coefficient.

Bio Sample	Condition	Treatment	Rep.	Depth	NSC	RSC	PBC
1	Aerobic	No Rif.	1	13,961,493	103.15	2.0193	0.1399
	Aerobic	No Rif.	2	14,810,838	162.70	1.7805	0.1633
	Anaerobic	No Rif.	1	16,108,774	153.51	1.8035	0.1353
	Anaerobic	No Rif.	2	13,636,541	172.59	2.014	0.1532
2	Aerobic	No Rif.	1	902,921	13.77	1.1270	0.2689
	Aerobic	No Rif.	1	1,852,124	17.91	1.5275	0.2590
	Aerobic	Rif. 20 min	2	2,104,427	29.60	1.2844	0.2584
	Aerobic	Rif. 20 min	2	11,548,572	13.08	1.5122	0.1510

Table 2 Summary of the E. Coli σ^{70} ChIP-exo and ChIP-seq samples. The last three columns depict ENCODE QC metrics on these data: NSC: Normalized Strand Cross-Correlation; RSC: Relative Strand Cross-Correlation; PBC: PCR Bottleneck Coefficient.

DNA libraries generated by ChIP-exo and ChIP-nexus protocols are expected to be less complex than the libraries generated by ChIP-seq because the numbers of positions to which the reads can align to are reduced due to the exonuclease digestion. This affects the interpretation of the PBC, which is defined as the ratio

of the number of genomic positions to which exactly one read maps to the number of genomic positions to which at least one read maps. For ChIP-seq samples, low PBC values (e.g., ≤ 0.5) indicate high levels of PCR amplification bias, i.e., PC bottleneck, unless the sequencing depth is high enough to sature all targets of the factor profiled. In contrast, for ChIP-exo/nexus, exonuclease digestion will lead to reads with same exact $5'$ end even before the PCR amplification step. We note that the PBC values are especially low for deeply sequenced ChIP-exo and ChIP-nexus samples; however, this does not automatically indicate severe bottlenecking as suggested by standard ChIP-seq guidelines.

The Strand Cross-Correlation (SCC), introduced by [17], is the most commonly used quality metric in assessing ChIP-seq enrichment quality. It aims to quantify how well the reads mapped to each strand are clustered around the locations of the protein-DNA interaction sites by calculating the Pearson correlation among forward and backward strands reads by shifting them across a range that covers both the read length of the experiment and the expected average fragment length. Typical SCC profiles exhibit two local maxima: at the average fragment length and the read length. In high quality experiments with clear ChIP enrichment, the average fragment length maximum coincides with the global maximum. In an idealized ChIP-exo experiment where the DNA fragments are digested to the boundaries of the protein-DNA interaction sites, we would expect the SCC profile to maximize at the motif length indicating clustering of the forward and reverse strand reads around the binding site. Figure 3 displays the SCC curves for the CTCF HeLa samples where the ChIP-exo curve shows local maxima at the motif and read lengths, while the SE ChIP-seq curves have a local maxima at the read length and a global maxima at the average fragment length. SCC profiles for other samples are available in Supplementary Figures SK: XXX. The read length and motif length maxima are often in close proximity of each other; as a result, this renders QC metrics such as the Normalized Strand Cross-Correlation (NSC) or the Relative Strand Cross-Correlation (RSC) harder to interpret; however, the profile itself seems informative about the enrichment signal in ChIP-exo/nexus experiments SK: this seems true at least based on my initial look at the plots, but we should be cautious in reporting this.

ChIP-exo quality control pipeline

We first present the overall pipeline and then discuss individual components with a case study using ChIP-exo data of FoxA1 from [10] and ChIP-nexus data from [2]. Figure 4 summarizes the 4-step pipeline. Given aligned reads from a ChIP-exo/nexus sample, the first step partitions the reference genome into islands by keeping the non-digested ChIP-exo regions. In step 2, the total number of extended reads SK: check: extended or not? What do we extend to in ChIP-exo? overlapping each island (D_i) and the number of unique island positions with at least one aligned read re recorded ((U_i). Then, three summary statistics ARC_i, URCR_i, and FSR_i are computed for each region i. ARC_i denotes the *average read coverage per base pair* and is defined as the ratio of the # of reads in island i (D_i) to the width of the island i (W_i); URCR_i, *unique read coverage ratio*, SK: Is unique read coverage ratio a good name for this? quantifies inverse of the effective coverage and is defined

as the ratio of the # of genomic positions with at least one aligned read within island i (U_i) to the # of reads in island i (D_i); and FSR_i denotes the proportion of forward strand reads. Step 3 of the pipeline generates several diagnostic plots aimed at quantifying ChIP enrichment and strand imbalance and step 4 generates quantitative summaries of these diagnostic plots SK: Aren't the linear model fit etc part of the pipeline?

Figure 4A presents the typical behavior of the URCR vs. ARC plot for a high quality ChIP-exo sample. In general, the plot depicts two strong arms. High URCR values correspond to regions with reads concentrated on a small number of positions. The left arm, with low ARC and varying URCR values, corresponds to background islands, regions that are usually composed of scattered reads that were not digested during the exonuclease step. The right arm where the URCR decreases as the ARC increases corresponds to regions that are usually ChIP enriched. SK: Not sure what this next sentence is referring to: Finally we quantify we quantify the relationship between library complexity by the use of two indexes that represent the change in the number of unique positions per regions and the change in of the width of a regions as the depth changes. Figures 4B and 4C aim to quantify the strand imbalance as part of the QC pipeline. The former depicts how quickly the islands exclusively formed by reads SK***I changed fragments to reads, but should this stay as fragments*** from a single strand are filtered out as islands with higher depths are observed. In a high quality sample, the proportion of islands with reads from only one strand is expected to decrease rapidly as we consider higher depth regions. In contrast, this proportion remains approximately constant in lower quality samples. The latter plot illustrates how quickly the quantiles of the FSR approaches to 0.5, the expected FSR value in high quality samples.

Application and validation of `ChIPexoQual` *with the FoxA1 ChIP-exo dataset*
SK: We should present the whole QC results and end with validation - panels from the two figures should be switched. We next use Fox A1 ChIP-exo datasets, with three biological replicates at comparable sequencing depths from mouse liver cells, to illustrate the QC pipeline. Figure 5A presents URCR vs. ARC plots for all three replicates. The first and third replicates exhibit a defined decreasing trend in URCR as the ARC increases. This indicates that these samples exhibit a higher ChIP enrichment than the second replicate. On the other hand, the overall URCR level from the first two replicates is higher than that of the third replicate level, elucidating that the libraries for the first two replicates are more complex than that of the third replicate.

Figure 6A and B display the strand-imbalance diagnostic plots and highlight specific problems with replicates 2 and 3. In particular, SK: I have a hard time interpreting B and C specifically, need to include something about their implications for replicates 2 and 3. SK: I also don't follow the discussion for panel C: Figure 6C compares the strand imbalance when ChIP-exo islands overlap with high quality peaks. It is noticeable that for regions composed by a large amount of reads, it is harder to distinguish their peaks by considering only the strand imbalance, hence in a high quality ChIP-exo experiment the background is expected to show a higher imbalance that the enriched regions. In conclusion, Figure 6 shows that the global

FSR does not accurately represent a ChIP-exo experiment's strand imbalance locally, hence the "peak pair" assumption does not completely hold in every ChIP-exo enriched region.

Overall, we conclude that replicate 1 is higher quality than both of replicates 2 and 3. We validate this observation with a motif analysis on the candidate binding regions identified from these replicates SK: Could we say binding events? or is the peak analysis on the whole binding region, i.e., peak? If so, we may want o change "candidate regions" to peaks? Figure 5B summarizes the total numbers of regions identified from each replicate. The lower number of enriched regions from replicate 2 is consistent with the lower ChIP enrichment pattern in the *UCRC* vs. *ARC* diagnostic plot. Scanning of these regions for the occurrence of FoxA1 sequence motif with the FIMO tool [19] indicates that the first replicate outperforms the other two in terms of percentage of candidate regions with the FoxA1 motif (Figure 5C). Furthermore, Figure 5D displays the average normalized read coverage around the actual motif locations in the candidate binding regions. These coverage plots reveal that the ChIP signal is more defined for the first and third replicates than the second one, indicating overall strength of the ChIP enrichment in these samples compared to the second replicate. Figures 5E, 5F, and 5G further highlight the overall quality of the identified motif sequences for each replicate and suggest that libraries with high library complexity (replicates 1 and 3) capture binding sites with better motif matches. SK: Worth thinking a bit more about the interpretation.

High sequencing depth may confound low-complexity library issues

We next evaluated every sample listed in Tables 1 and 2 with the `ChIPexoQual` QC pipeline (Supplementary Figures SK: XXX). A key, albeit not suprising, observation from large scale analysis is that the URCR vs. ARC plots typically display the three patterns captured in the FoxA1 study. We will refer to these as pattern I (FoxA1 replicate 1), II (FoxA1 replicate 2), and III (FoxA1 replicate 3), respectively. Pattern III where the two arms along ARC are not distinguishable can arise due to either low-complexity library or high sequencing depth. For example, all three replicates of the TBP ChIP-exo from K562, with sequencing depths between $\sim$ 60M to 115M reads, and replicate two of TBP ChIP-nexus in K562, with sequencing depth of $\sim$ 130M reads, exhibit this pattern. A simple but effective strategy to distinguish the two plausible scenarios from Pattern III is to apply the QC pipeline to sub-samples randomly generated from the full dataset at varying sequencing depths. We applied this strategy by sub-sampling 20M to 50M reads, a range that represents the sequencing depths of the human samples we are using in the paper, from the TBP samples. URCR vs. ARC diagnostics of these sub-samples (Supplementary Figures SK: XXX) indicate that of the four TBP samples with this pattern, replicates two and three of K562 ChIP-exo suffer from low-complexity library issues, whereas the other samples exhibit the pattern specific to high quality samples. SK: Add: Validation with motif on TBP ChIP-exo replicates.

Evaluation of a large collection of ChIP-exo and ChIP-nexus data with `ChIPexoQual`

We next performed an overall analysis of the `ChIPexoQual` QC pipeline results for the samples in Tables 1 and 2. We quantified the relationship between ARC

and URCR by fitting a reparametrized regression model of URCR as a function of ARC. Specifically, we considered $D_i = \beta_1 U_i + \beta_2 W_i + \varepsilon_i$, where ε_i represents the random error term. As we discuss in Materials and methods, this parametrization has a direct connection to $\text{URCR}_i = \frac{\kappa}{\text{ARC}_i} + \gamma + \epsilon_i$, which aims to recapitulate the relationship in the URCR vs. ARC plots. Figure 7A displays estimated overall change in depth ($\hat{\beta}_1$) as the number of positions with at least one aligned read varies across a large collection of ChIP-exo samples from eukaryotic genomes. The β_1 parameter can be interpreted as the limiting (i.e., large depth) SK: need a better name for: average read depth intensity URCR of a sample. As discussed earlier, high quality ChIP-exo samples are expected to have two arms in the URCR vs. ARC plots: one with low ARC and varying URCR and another with a decreasing URCRas ARC increases and stabilizes β_1. When the ChIP-exo sample is not deeply sequenced, high values of $\hat{\beta}_1$ in Figure 7A indicate that the library complexity is low. On the other hand, lower values correspond to higher quality ChIP-exo experiments. Taking into account the depths of these samples and visualizing all the diagnostic plots (SK: Supplementary Figures XXX), we conclude that samples with estimated β_1 values less than 10 seem to be high quality samples.

We interpret the β_2 parameter above as the bias for the average read coverage and display its estimates across all the eukaryotic samples in Figure 7B. Under perfect digestion by λ-exo, most of the reads aligned to binding regions are expected to accumulate around a binding event. SK: I don't follow the next argument: We should say something about the expected behavior of β_2 which suggest to be unlikely to observe as the sequencing depth increases to observe a reads being aligned to another position that was not previously covered. Low quality ChIP-exo experiment exhibit more scattered reads in both strands across the genome; therefore it is more likely to observe consider reads that align to position in the genome that were not previously covered. Although the third replicate of the TBP ChIP-exo experiment has comparable sequencing depth to the second replicate of the TBP ChIP-nexus experiment, the (Figure 7(B)), ARC bias is considerably higher for the ChIP-exo experiment. This potentially indicates that additional sequencing reads in comparison to replicates 1 and 2 are scattered around new positions instead of accumulating on the existing binding sites.

The interaction between these two parameters has implications regarding the quality of a ChIP-exo and ChIP-nexus sample. When either the adjusted ARC or the ARC bias is large SK: Not clear what large for β_2 means if we are only looking at a single sample owing to potentially the high sequencing depth of the sample, we suggest randomly sub-sampling reads to form samples of lower depth and evaluating the sub-samples with the QC pipeline. As an illustration, we revisit this strategy for the three replicates of TBP ChIP-exo in K562 [12] and second replicate from the K562 ChIP-nexus experiments [2]. Figure 7C exhibits an increasing trend in the estimated β_1 across varying sequencing depths for replicates 2 and 3 which we deem as lower quality than replicate 1. Furthermore, the estimates with lower depths are still higher than that of the replicate 1 and the overall trend of the ChIP-nexus sub-samples. Figure 7D illustrates that the *ARC* bias remains approximately constant in ChIP-nexus sub-samples and sub-samples of first replicate of ChIP-exo, while it increases for the second and third ChIP-exo replicates. This suggests that

these two ChIP-exo replicates have low library complexity and overall lower quality than the ChIP-nexus samples, regardless of the fact that all three experiments are deeply sequenced with more than 90M reads each. Furthermore, the `ChIPexoQual` diagnostic plots for each sub-sample (SK: Supplementary Figures XXX) illustrate that the two arms of the ARC versus UCRC plots are clearly visible in moderate depth sub-samples of TBP ChIP-nexus data. SK: Some more work/reorganization not to have too much repetition with this last paragraph and the section that follows right after FoxA1.

High-resolution binding event identification with statistical analysis of ChIP-exo data

Our QC pipeline `ChIPexoQual` operates on aligned read files and does not require any statistical analysis of the data such as identification of potential binding regions/events. This enables its broad and easy usability before any statistical analysis for identifying binding events from ChIP-exo/nexus data. We next evaluated recent analytical approaches for ChIP-exo data analysis and further compared ChIP-exo with PE ChIP-seq data on our *E. coli* samples.

Evaluation of available methods in discovering closely spaced binding events from E. coli ChIP-exo data

We specifically considered recently developed Peakzilla [6], MACE [4], GEM [9], and dPeak [8] in our evaluations with high quality *E.Coli* samples (samples SK: x and y from the aerobic condition). Among these methods, Peakzilla and MACE are specifically developed for ChIP-exo analysis, whereas both dPeak and GEM can identify high resolution binding events from ChIP experiments, rendering them suitable for ChIP-exo analysis as well.

Figures 8A and 8B compare binding events identified with all the four methods in terms of resolution, where the resolution is defined as the distance from a RegulonDB [13] annotation to its closest prediction. The resolutions of all the methods are comparable for the first replicate (Figure 8(A)) and dPeak, on average, has slightly better resolution than the rest in the second replicate (Figure 8(B)).

dPeak model has two parameters that further elucidate characteristics of ChIP-exo data. The δ parameter of dPeak quantifies the average distance from the $5'$ ends of the reads to the binding event they are profiling and the σ parameter is a measure of dispersion of $5'$ ends around the binding events. Figures 8C and 8D compare the densities of these parameters estimated by dPeak for *E. coli* ChIP-exo and SE ChIP-seq data. As expected, both are lower for ChIP-exo than for SE ChIP-Seq, indicating that dPeak, although originally developed for high-resolution binding event identification from ChIP-seq, can learn this information.

Saturation analysis with ChIP-exo and ChIP-seq: Systematic comparison of ChIP-seq vs ChIP-exo under varying sequencing depths

Establishing competitive performance of dPeak compared to other ChIP-exo analysis methods enable comparison of ChIP-exo and ChIP-seq with a unified analysis framework using dPeak. Previous comparisons of ChIP-exo and SE ChIP-seq were all performed without controlling for the sequencing depths of the samples. More importantly, although it is well established that PE ChIP-seq leads to better resolution that SE ChIP-seq in terms of binding site identification SK: Cite Qi's BMC

Bioinformatics paper, previous work does not have an in depth comparison of ChIP-exo and PE ChIP-seq. In order to address these limitations of the previous studies, we performed a sub-sampling experiment by sampling fixed numbers of reads from each of the σ^{70} ChIP-exo, PE ChIP-seq, and SE ChIP-seq datasets. Specifically, for every N reads in ChIP-exo and SE ChIP-seq, $N/2$ read pairs were sampled for PE ChIP-seq to operate under fixed sequencing costs.

Figure 9 summarizes the comparisons of the three data types for σ^{70} under aerobic condition as a function of sequencing depth. For these computational experiments, we used σ^{70} RegulonDB [13] binding events as gold standard. Figure 9A displays the number of candidate regions (i.e., peaks) where at least one binding event was identified whereas Figure 9B depicts the number of identified binding events, i.e., each candidate region can harbor multiple binding events. In terms of the number of binding events, ChIP-exo ranks on the conservative side. In Figure 9C, we display the number of correctly identified binding events, where we consider a RegulonDB event as correctly identified if an estimated binding event is identified within 15 bps of it. Finally, Figure 9D presents the resolution defined as the distance from a RegulonDB binding event to the closest dPeak prediction. These comparisons indicate that although PE ChIP-seq performs much better than SE ChIP-seq in terms of identifying closely spaced binding events, ChIP-exo outperforms PE ChIP-seq at comparable depths.

Figure 10 shows comparisons among dPeak analysis of full ChIP-exo, PE ChIP-seq and SE ChIP-seq σ^{70} datasets SK: Again, make sure these ChIP-seq samples are introduced before and numbering system will help to refer each sample. SK: Are the sequencing depths comparable? We are making a big deal about matching depths and then jumping into this analysis. The only way these results would work is that either the depts are comparable or all the samples have depths sufficient enough for saturation. Utilizing RegulonDB binding events as ground truth, we computed sensitivity as the proportion of RegulonDB events identified by dPeak in each analysis and the resolution as the minimum distance between a RegulonDB event and the closest dPeak binding event prediction. Figure 10A illustrates that as the mean distance between binding events increases, sensitivity for all data types increase. Consistent with the sub-sampling experiments, both PE ChIP-seq and ChIP-exo significantly outperform SE ChIP-seq and ChIP-exo exhibits more power than PE ChIP-seq in deconvolving binding events. Figure 10B highlights that ChIP-exo and PE ChIP-seq are comparable in resolution with these deeply sequenced samples, while both protocols significantly outperform SE ChIP-seq.

Conclusions

We presented a systematic exploration of several ChIP-exo/nexus experiments. We provided a list of factors that reflect the quality of a ChIP-exo experiment and developed a QC pipeline, named `ChIPexoQual`. `ChIPexoQual` takes as input aligned reads and automatically generates several diagnostic plots and summary measures that enable assessing enrichment and library complexity. Our analysis of several datasets indicated that while the QC pipeline only requires a set of aligned reads to give a global overview of the quality of a given ChIP-exo dataset, implications of the diagnostic plots and the summary measures align well with more elaborate analysis

that is computationally more expensive to perform and/or requires additional inputs that may not be available, such as motif occurrences in a set of high quality regions or resolution analysis based on a gold-standard.

To the best of our knowledge, we also provide the first systematic comparison between ChIP-exo and PE ChIP-seq datasets using our *E. coli* σ^{70} samples. This comparison revealed that PE ChIP-seq compares much more competitively with ChIP-exo compared to SE ChIP-seq. However, overall, ChIP-exo provides the best performance in terms of deconvolving closely spaced binding events and resolution. The pname is available at `https://github.com/keleslab/ChIPexoQual`.

Materials and methods

ChIP-seq/exo/nexus datasets

E. coli ChIP-exo and ChIp-seq samples

We generated ChIP-exo and PE ChIP-seq samples from σ^{70} in *E.Coli*. For each PE ChIP-seq experiment, an *in silico* SE version was obtained by randomly sampling one of the two ends in a read pair.

need to add the growth conditions. SK: Add something based on dpeak paper, we can ask Jeff to check.

Processing of the ChIP-exo and ChIP-nexus samples

We aligned the read files of the samples listed in Table 2 either following the directions in their original publications when available or with `bowtie` (version 1.1.2) [**?**]. SK: include bowtie comment to show how parameters were set.

Generation of a set of high signal regions from *E. coli* samples to assess strand imbalance

We partitioned the *E.Coli* genome into non-overlaping intervals, i.e., bins, of length 150 bps and counted the number of reads overlapping each bin. As is usually the practice with ChIP-seq analysis, each read was extended to the average fragment length of 150 bps towards the $3'$ direction. To evaluate the strand imbalance, we identified a set of high quality peaks for ChIP-exo and SE ChIP-seq by analyzing both with `mosaics` [15] under the GC content + Mappability and Input only modes for background estimation, respectively. The subset of these peaks for which `dPeak` analysis identified one or more binding events were used in FSR assessments (Figure SK: Fig num). SK: Do the plots change if we consider only peaks with one binding event? One could argue that when there are multiple events, there might be more background reads, which could skew the FSR distribution.

ENCODE ChIP-seq QC metric guidelines

We used the ChIP-seq QC metric definitions established by [7] and described in detail at `https://genome.ucsc.edu/ENCODE/qualityMetrics.html`. These QC metrics were calculated with the **ChIPUtils** package (version 0.99.0 from `https://github.com/welch16/ChIPUtils`). Empirical data from the ENCODE project suggests the following guidelines for interpretation of the QC metrics for human and mouse genomes: a PBC value between 0 to 0.5 indicates severe bottlenecking, 0.5 to 0.8 moderate bottlenecking, 0.8 to 0.9 mild bottlenecking, and 0.9 - 1 no bottlenecking.

SK: I am in favor of skipping these definitions of ENCODE metrics for the paper - you might want to keep it for your dissertation-

ChIP-exo quality control with `R` package `ChIPexoQual`

We implemented our proposed QC pipeline with an `R` package named `ChIPexoQual`, available at `https://github.com/welch16/ChIPexoqual`. The analysis in this paper used version 1.0 of the `ChIPexoQual` package.

`ChIPexoQual`: The package takes in as input a set of aligned reads from a ChIP-exo (or ChIP-nexus) experiment and performs the following steps.

1. Identify read islands, i.e., overlapping clusters of reads separated by gaps, from read coverage.
2. Compute D_i, number of reads in the island i , and U_i, number of island i positions with at least one aligning read, $i = 1, \cdots, I$.
3. For each island i, $i = 1, \cdots, I$, compute island statistics:

$$\text{ARC}_i = \frac{D_i}{W_i}, \quad \text{URCR}_i = \frac{U_i}{D_i},$$
$$\text{FSR}_i = (\#\text{ of forward strand reads aligning to island } i)/D_i,$$

   where W_i denotes the width of island i.
4. Generate diagnostic plots (i) URCR vs. ARC plot; (ii) SK: Name this FSR plot; (iii) SK: Name this FSR plot as in Figure 4.
5. Randomly sample M (at least SK: ?) islands and fit,

$$D_i = \beta_1 U_i + \beta_2 W_i + \varepsilon,$$

   where ε denotes the independent error term. Repeat this process B times and generate box plots of estimated β_1 and β_2.

Interpretation of the linear model in the QC pipeline The linear model

$$D_i = \beta_1 U_i + \beta_2 W_i + \varepsilon_i$$

re-parametrization of the following relationship from $UCRC$ vs. ARC diagnostic plot:

$$\text{URCR}_i = \frac{\kappa}{\text{ARC}_i} + \gamma + \epsilon_i \tag{1}$$

with $\beta_1 = 1/\gamma$ and $\beta_2 = -\kappa/\gamma$. In this setting, γ can be considered as the large-depth URCR_i, i.e., the limiting ratio between the number of positions with at least one mapping read and depth as the depth tends to infinity.

SK: ***β_2 needs more clarification. On the other hand, to interpret $\beta_2 = -\kappa/\gamma$, by expressing κ as a function of ARC and URCR and assuming that γ is already estimated, we can observe the following identities:

$$\kappa = \frac{\text{npos}}{\text{width}} - \gamma \text{ARC}$$
$$\frac{\kappa}{\gamma} = \frac{1}{\gamma}\frac{\text{npos}}{\text{width}} - \text{ARC}$$

This is important: γ approximates the URCR as the sequencing depth increases, which implies that $-\kappa/\gamma$ can be interpreted as the large sequencing depth bias of the ARC since as the depth increases, the first term of κ/γ is going to approximate the average read coverage:

$$\frac{\text{npos}}{\text{width}} \times \lim_{d\to\infty} \frac{d}{\text{npos}} \sim \text{ARC}$$

Therefore, $\beta_2 = -\kappa/\gamma$ is interpreted as the ARC bias as the sequencing depth increases. SK: ***

Motif analysis of FoxA1 enriched regions

For each FoxA1 ChIP-exo replicate, we used the ChIP-exo QC pipeline to partition the mouse genome into a set of islands with their respective summary statistics. We then filtered them into collections of high quality regions by (i) removing the islands with reads residing only on one strand; (ii) removing the islands with $U_i \leq 15$; (iii) removing islands with $D_i < 100$. These thresholds were empirically selected and the overall conclusions were robust to their variation. We used `FIMO` (version 4.9.1, SKinclude command line to show parameters) [19] to identify the FoxA1 motif within each enriched region using FoxA1 position weight matrix MA0148.1 from the JASPAR database [**?**].

Testing for strand imbalance in FoxA1 ChIP-exo replicates

SK: This paragraph could benefit from a bit more polishing.

We used the ChIP-exo QC pipeline to partition the mouse genome into a set of islands with their respective summary statistics. We filtered the islands with reads in only one strand and transformed the FSR into an *imbalance index* that is zero when the resulting region is perfectly balanced in terms of forward and reverse strands reads and infinity when it consists of reads in one strand exclusively:

$$\text{Imbalance index} = -\log_{10}(4 \times \text{FSR} \times (1 - \text{FSR}))$$

We divided the ChIP-exo regions into two classes depending on whether or not they overlap ChIP-exo peaks, where the we identified the peaks with `mosaics` using the GC content + Mappability background model [15] (version 2.9.7) and at a false discovery rate (FDR) level of 0.05 and minimum read threshold of 100. The bin size and fragment length for MOSAiCS runs were set to 200 bps. We further filtered the resulting peaks by keeping only the peaks with an average extended ChIP read count of 200 and used a Wilcoxon test to assess the difference in imbalance index between the two groups of ChIP-exo regions.

High resolution analysis with ChIP-exo

In all the evaluations using *E. coli* samples, we considered RegulonDB [13] σ^{70} sites as gold-standard. A *sig* site was deemed as identified if there existed a binding event within its 20 bps proximity. We defined the resolution as the distance from an σ^{70} site to its closest predicted binding event and the sensitivity as the fraction of correctly identified σ^{70} sites in a genomic region.

Statistical methods for ChIP-exo

We compared dPeak [8] (`https://github.com/dongjunchung/dpeak`, version 2.0.1), GEM [9] (`http://groups.csail.mit.edu/cgs/gem/`, version 2.6), MACE [4] (`http://dldcc-web.brc.bcm.edu/lilab/MACE/docs/html/`, version 1.2), peakzilla [6] (`https://github.com/steinmann/peakzilla`).

SK: Too many repeats of mosaics version etc, try to minimize.

Candidate regions for `dPeak` and `GEM` were identified for each replicate of ChIP-exo data using the **MOSAiCS** algorithm [15] (one sample analysis using false discovery rate of 1%) implemented as an `R` package **mosaics** (version 2.9.7 from Bioconductor). We further filtered out candidate regions by using the top 400 peaks with highest average ChIP read counts to avoid potential false positives based on the exploratory analysis. These regions were also explicitly provided to the GEM algorithm as candidate regions. Default tuning parameters were used during model fitting for all methods. Although we were able to download `CexoR` [5] (version 1.8 from *bioconductor*), we were unable to use it for the σ^{70} experiments because of the SK: *** issues.

dPeak analysis of σ^{70} ChIP-exo and ChIP-seq data

We compared the estimated binding events predicted by the **MOSAiCS** + **dPeak** pipeline using reads generated by ChIP-exo, PE and SE ChIP-seq protocols. We called peaks at a 5% FDR level using **MOSAiCS** (the GC content + Mappability background model for ChIP-exo and the Input only model for PE and SE ChIP-seq). Then, we deconvolved the peaks into binding events with **dPeak** (version 2.0.1) by considering a maximum of 5 binding events within each peak. To avoid false positives, we only considered ChIP-exo peaks with average ChIP read count greater than 3,000 that overlapped both the SE and PE ChIP-seq peaks. Repeating the analysis with other cutoff values led to similar conclusions.

Saturation analysis of ChIP-exo, PE and SE ChIP-seq

SK: Next paragraph could use more editing. We sub-sampled N reads for both ChIP-exo and SE ChIP-seq protocols. For PE ChIP-seq, we sub-sampled $N/2$ pairs equaling to a total of N reads. For each sub-sample, we identified peaks using MOSAiCS [15] (GC content + Mappability background model for ChIP-exo and Input only for SE and PE ChIP-seq) SK: not sure: for the maximum sample size. To avoid potential false positives, we considered only the top 500 peaks for each data protocol. We defined the number of candidate peaks as the number of top sample peaks with at least one predicted dPeak binding event; the number of predicted events is the total number of dPeak predicted binding events; the number of identified σ^{70} sites as the number of gold-standard σ^{70} sites within 15 bps from

an estimated binding event; and the resolution as the minimum distance from a gold-standard σ^{70} site to an estimated binding event. We repeated this analysis for ten random sub-sampling experiments and reported the median across these experiments.

SK: Yeah, Materials and Methods has too many repeats, needs to be organized a bit and cleaned up. Figure captions need to be modified and made more easy follow.

Author details
[1]Department of Statistics, University of Wisconsin Madison, 1300 University Avenue, Madison, WI. [2]Department of Biostatistics and Medical Informatics, University of Wisconsin Madison, 600 Highland Avenue, Madison, WI. [3] Great Lakes Bioenergy Research Center, University of Wisconsin Madison, 1552 University Avenue, Madison, WI. [4] Department of Biochemistry, University of Wisconsin Madison, 433 Babcock Drive, Madison, WI. [5] Department of Bacteriology, University of Wisconsin Madison, 1550 Linden Drive, Madison, WI. [6] Department of Public Health Sciences, Medical University of South Carolina, 135 Cannon Street, Charleston, SC.

References

1. Rhee HS, Pugh F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011;.
2. He Q, Johnston J, Zeitlinger J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nature Biotechnology. 2014;.
3. Mahony S, Franklin PB. Protein-DNA binding in high-resolution. Critical Reviews in Biochemistry and Molecular Biology. 2015;.
4. Wang L, Chen J, Wang C, Uusküla-Reimand L, Chen K, Medina-Rivera A, et al. MACE: model based analysis of ChIP-exo. Nucleic Acids Research. 2014;.
5. Madrigal P. CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates. EMBnetjournal. 2015;.
6. Bardet AF, Steinmann J, Bafna S, Knoblich JA, Zeitlinger J, Stark A. Identification of transcription factor binding sites from ChIP-Seq data at high resolution. Bioinformatics. 2013;.
7. Landt S, Marinov G, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-Seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research. 2012;.
8. Chung D, Park D, Myers K, Grass J, Kiley P, Landick R, et al. dPeak, High Resolution Identification of Transcription Factor Binding Sites from PET and SET ChIP-Seq Data. PlOS, Computational Biology. 2013;.
9. Guo Y, Mahony S, Gifford DK. High resolution genome wide binding event finding and Motif discovery reals transcription factor spatial bindings constraints. PLOS, Computational Biology. 2012;.
10. Serandour A, Gordon B, Cohen J, Carroll J. Development of and Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. Genome Biology. 2013;.
11. Starick SR, Ibn-Salem J, Jurk M, Hernandez C, Love MI, Chung HR, et al. ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. Genome Research. 2015;.
12. Venters BJ, Pugh F. Genomic organization of human transcription initiation complexes. Nature. 2013;.
13. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo Js, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more;.
14. Benjamin Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research. 2011;.
15. Kuan PF, Chung D, Pan G, Thomson JA, Stewart R, Keleş S. A Statistical Framework for the Analysis of ChIP-Seq Data. Journal of the American Statistical Association. 2009;.
16. Valouev A, Johnson D, Sundquist A, Medina C, Anton E, Batzoglou S, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature, Methods. 2008;.
17. Kharchenko P, Tolstorukov M, Park P. Design and analysis of ChIP-Seq experiments for DNA-binding proteins; 2008.
18. Carroll T, Liang Z, Salama R, Stark R, de Santiago I. Impact of artifact removal on ChIP quality metrics in ChIP-Seq and ChIP-exo data. Frontiers in Genetics, Bioinformatics and Computational Biology. 2014;.
19. Grant C, Bailey T, Noble WS. FIMO: Scanning for occurrences of a given motif;.
20. Mendenhall EM, Bernstein BE. DNA-protein interactions in high definition. Genome Biology. 2012;.
21. Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;.
22. Bolstad B, Irizarry R, Âstrand M, Speed T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;.
23. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature Reviews: Genetics. 2012;.
24. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference in Intelligent Systems for Molecular Biology. 1994;.

1 Figures

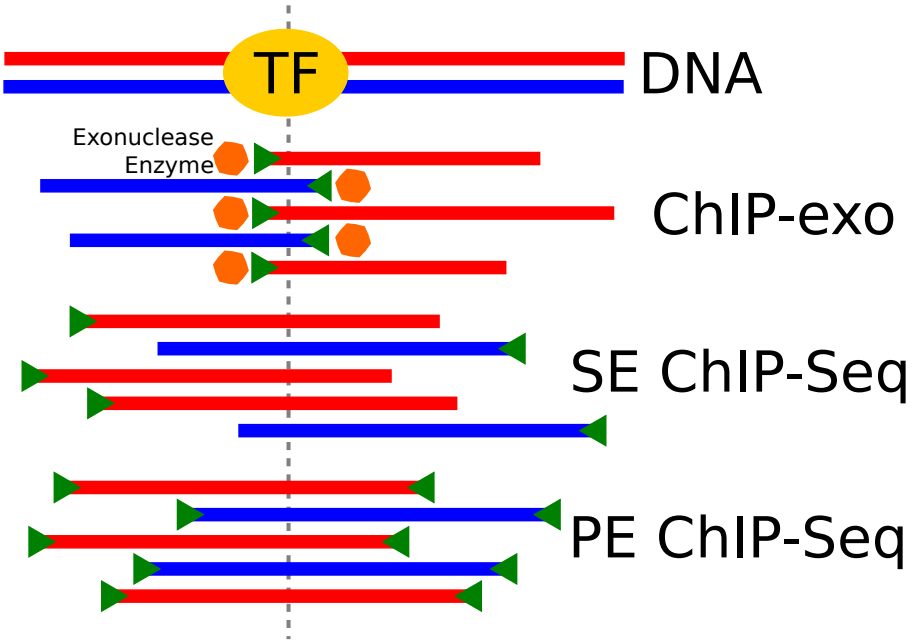

Figure 1 A TF is bound to the forward (red) and backward (blue) strand of the DNA. Then, is sonicated: For ChIP-exo a exonuclease enzyme (orange hexagon) trims the $5\prime$ ends of each DNA fragment to a fixed distance from the bound protein, finally is subjected to Immunoprecipitation and amplification. For both ChIP-exo and SE ChIP-Seq an adapter is ligated (green triangles) at the $5\prime$ ends, while for PE ChIP-Seq is ligated to both ends.

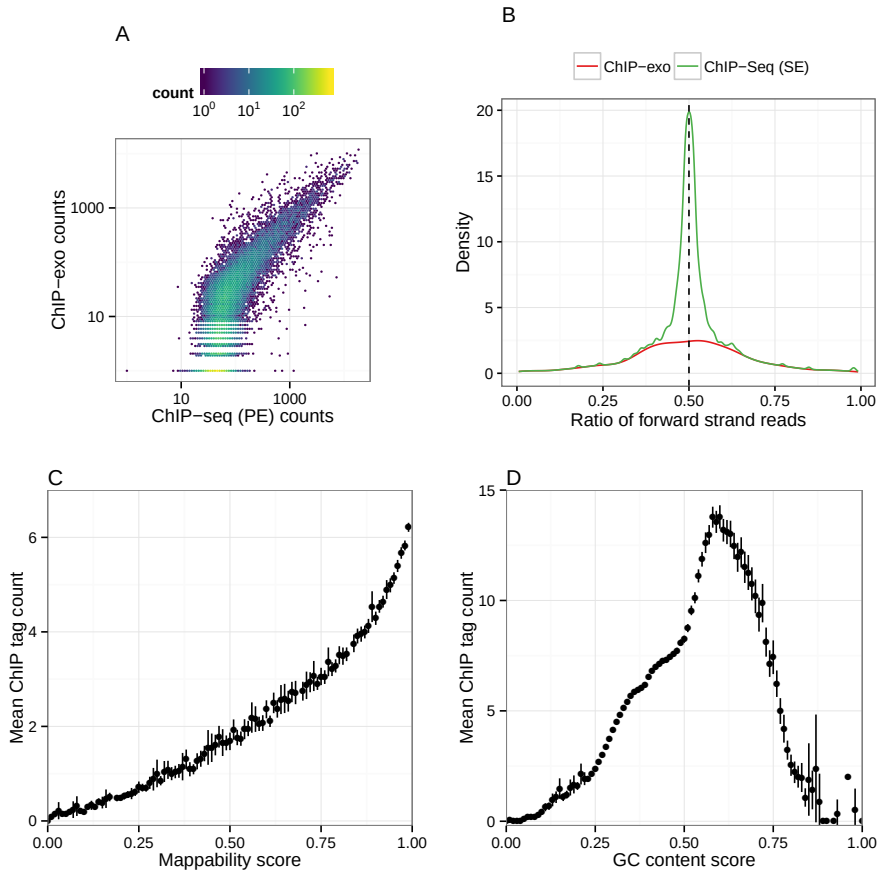

Figure 2 A) Hexbin plot of PE ChIP-Seq bin counts vs ChIP-exo bin counts. B) Forward Strand Ratio densities for SE ChIP-Seq and ChIP-exo peaks. C) Mappability score vs mean ChIP tag counts with 0.95 confidence bands. D) GC - content score vs mean ChIP tag counts with 0.95 confidence bands.

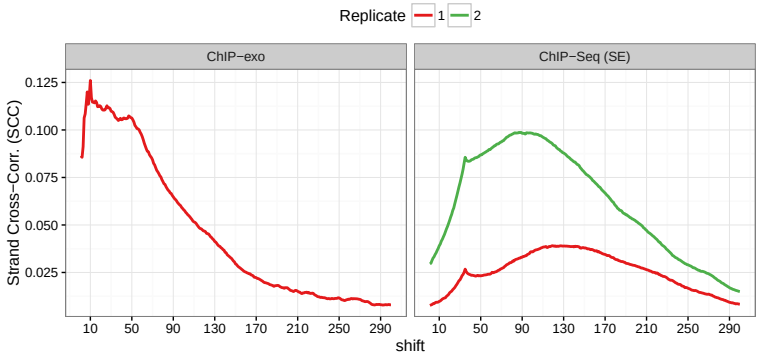

Figure 3 SCC curves for human CTCF on HeLa cell lines. The SCC curve for the ChIP-exo sample from [1] is shown in the left panel, and the SCC for ChIP-Seq samples from **https://www.encodeproject.org/experiments/ENCSR000AOA/** are shown in the right panel. The ChIP-exo curve shows local maxima at the motif and read length. Both SE ChIP-Seq curves are maximized at the fragments length and show a local maxima at its read length.

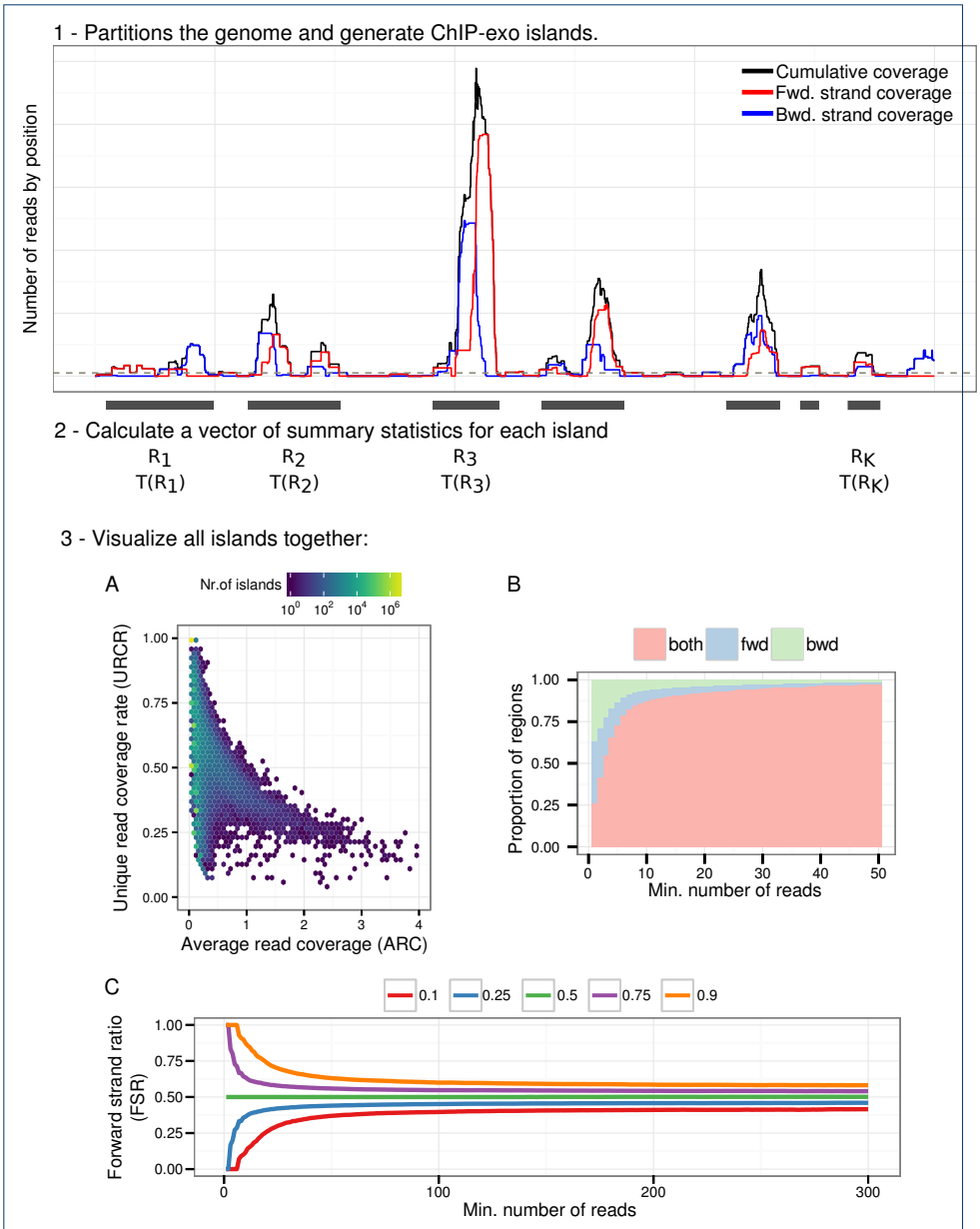

Figure 4 The ChIP-exo reads are partitioned by keeping the regions in the genome formed by the undigested fragments. For each region, we calculate a collection of summary statistics, which are visualized as A) ARC vs. URCR, this plot presents a global view of the balance between library complexity and enrichment. There are two arms: First, one with low ARC, which corresponds to regions formed by few aligned positions; Second, where the URCR decreases as the ARC increases. B) Min. depth vs Proportions of regions, this plot shows the strand composition for all the regions formed by a min. number of reads. Regions with low depth tend to be formed by undigested reads from one unique strand, while regions with higher signal are usually formed by reads in both strands. C) Min depth vs. FSR, this plot depicts how quickly the FSR's distribution approximate the median. In a high quality sample, the median is around 0.5, and the other quantiles reach that value quickly.

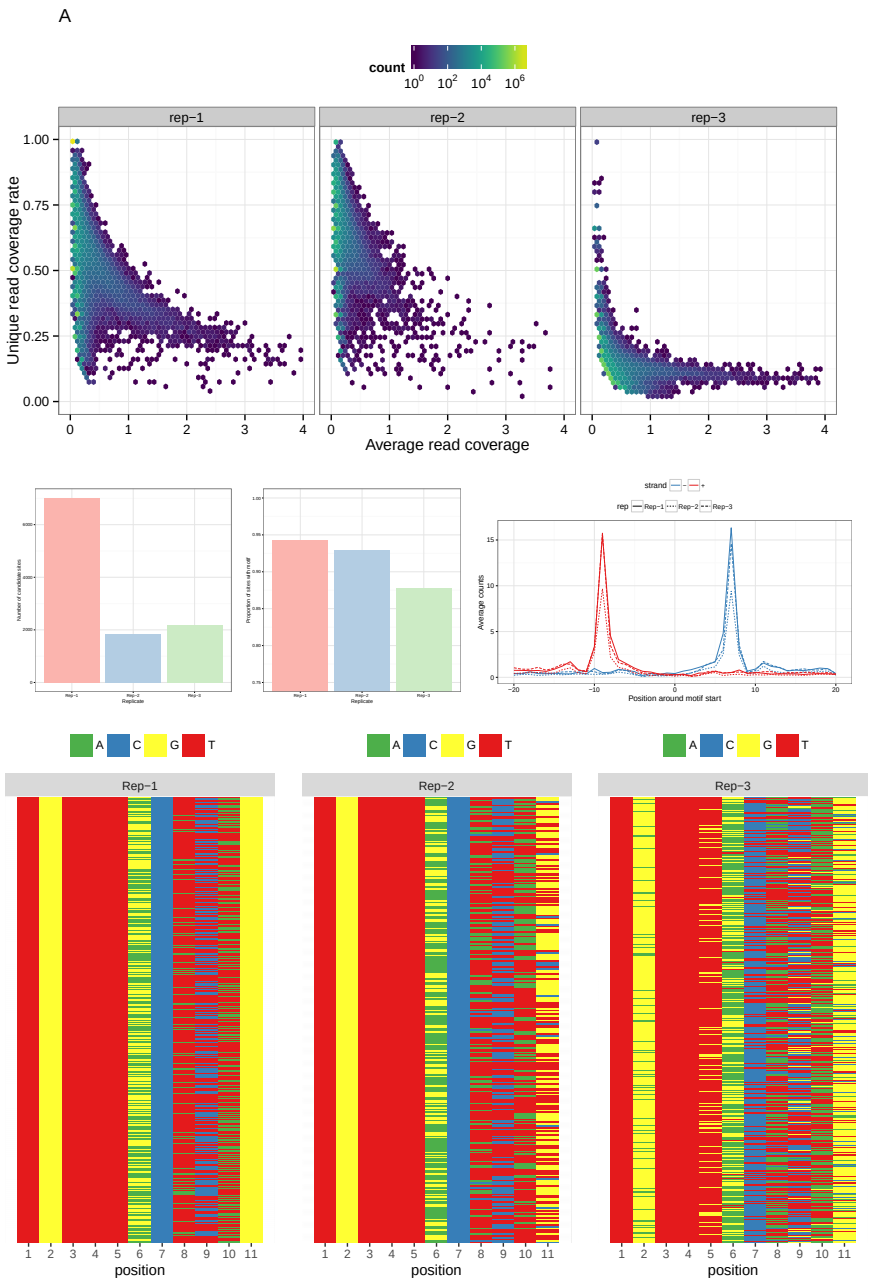

Figure 5 Using the mouse FoxA1 experiment from [10]: A) Hexbin plots of ARC against URCR, there is a slight separation into two strong arms, one corresponds to low ARC and varying URCR, and for the other URCR decreases as ARC increases. B) Number of candidate sites for each replicate. C) Percentage of candidate sites where the FoxA1 motif was detected. D) Average coverage around FoxA1 motif. Base distribution for matched sequence for Rep-1 (E), Rep-2 (F) and Rep-3 (G).

need to add labels for each figure and think about order, in the supplement we have ecdf from fimo score or pval, and overlaps with peaks called by mosaics.
Also, need to change labels from repk to Rep-k

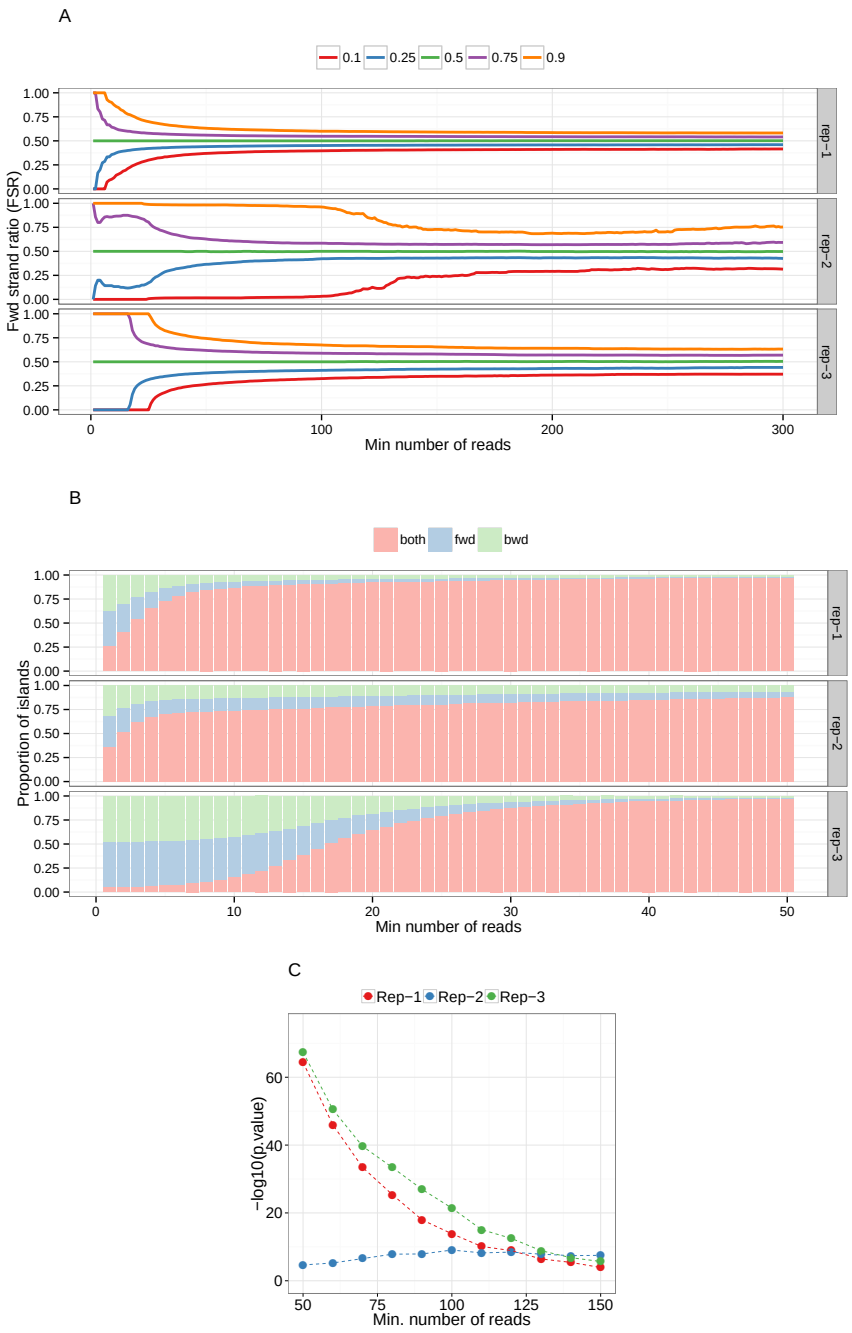

Figure 6 Strand imbalance using the mouse FoxA1 experiment from [10]: A) FSR distribution quantiles as the lower depth regions are being filtered out, all quantiles approach to the median as the lower bound increases. B) Stacked histogram with the proportion of regions that are formed by two strands or only one, in a good sample the single-stranded regions are going to be filtered out quickly as in the middle row. C) $-\log_{10}$(p.value) of testing if the imbalance distributions differs when ChIP-exo regions overlap their peaks.

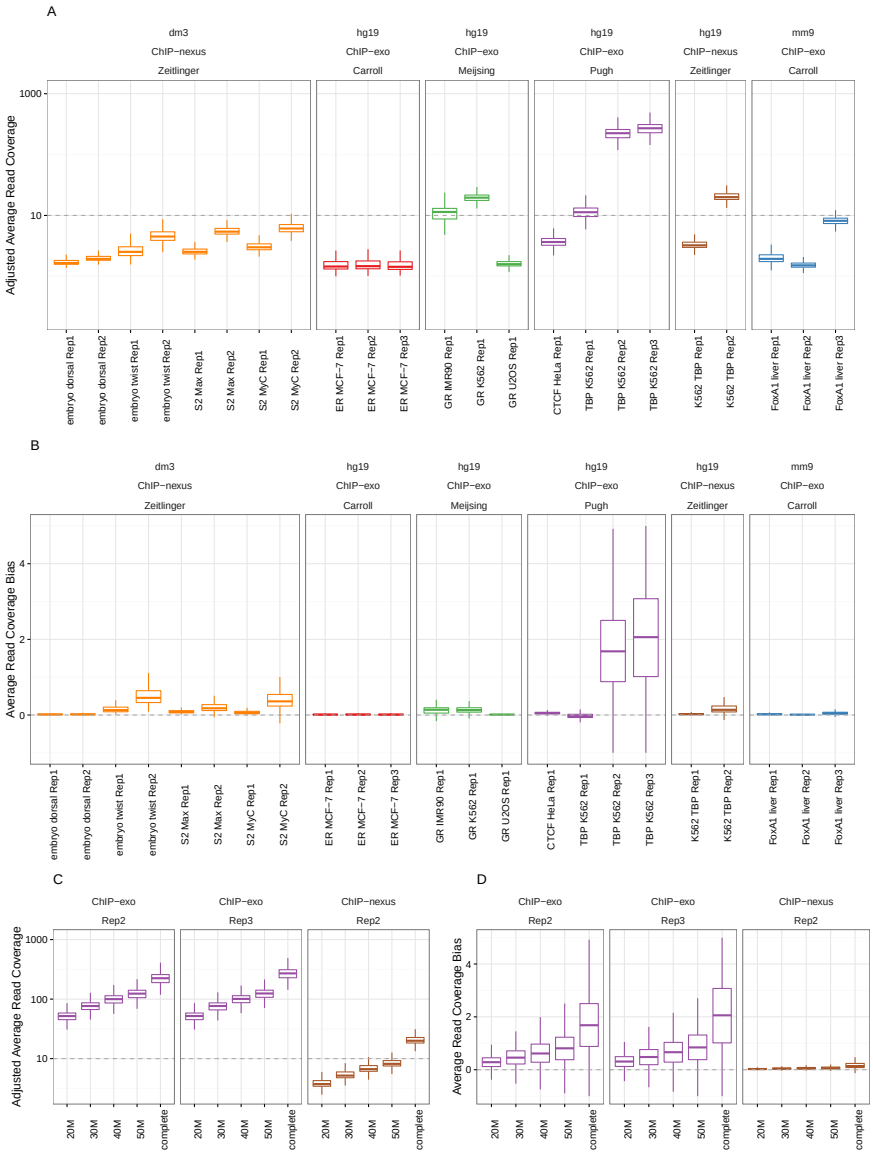

Figure 7 Comparison of adjusted Average Read Coverage and ARC bias for eukaryotic genome ChIP-exo and ChIP-nexus experiment. A) Boxplot of adjusted ARC. B) Boxplot of large sequencing depth ARC's bias. Comparison of sub-sampled ChIP-exo and ChIP-nexus experiments of TBP factor in human K562 cell lines: C) Adjusted ARC and D) ARC's bias.

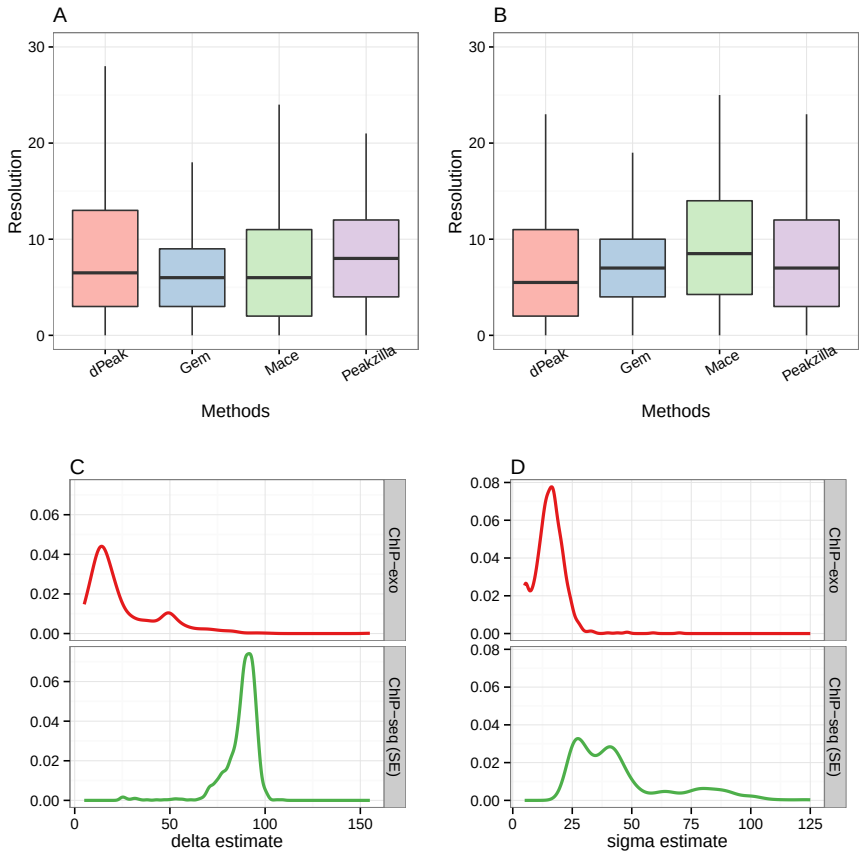

Figure 8 Comparison of the resolution between dPeak, Gem, Mace and Peakzilla methods with *E.Coli* aerobic experiment from the first biosample A) Replicate 1 and B) Replicate 2. Resolution is defined as the minimum distance between a RegulonDB annotation and a predicted binding event. C) δ parameter in dPeak measures average distance of the reads to their respective binding site. In ChIP-exo data, reads were located much closer to the binding site than in SET ChIP-Seq. D) σ parameter measure the dispersion of reads around each binding site. In ChIP-exo data, reads showed less variation around the their respective binding sites compared to SET ChIP-Seq.

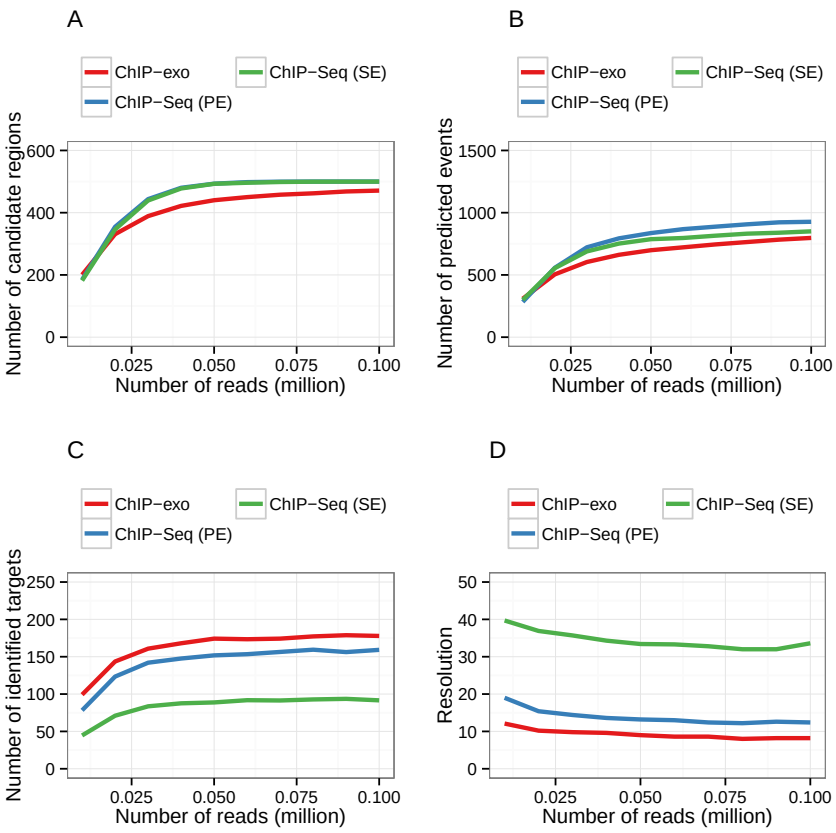

Figure 9 Comparison of the number of A) candidate regions, B) predicted events, C) identified targets and D) resolution among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. RegulonDB annotations are considered as a gold standard. A gold standard binding events was deemed identified if a binding event was estimated at a $\pm$ 15 vicinity of it.

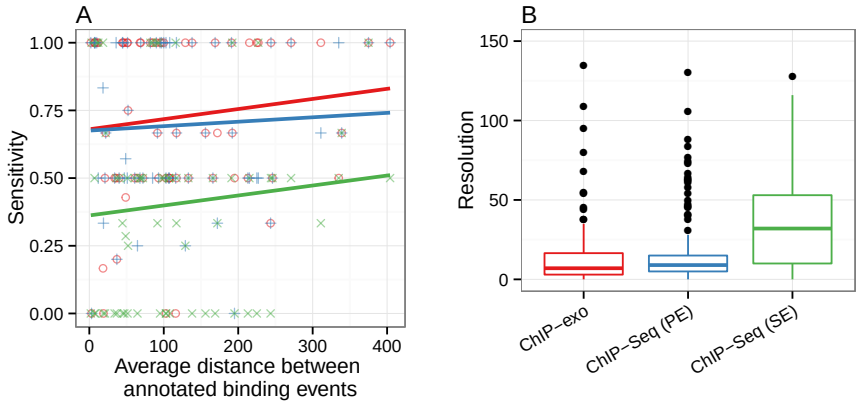

Figure 10 Comparison of A) sensitivity and B) resolution between ChIP-exo and ChIP-Seq data. Sensitivity is defined as the proportion of RegulonDB annotations identified using each data. Resolution is defined as the distance between RegulonDB annotation and its closest prediction.