RESEARCH

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

Rene Welch [1†], Dongjun Chung [6†], 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 pose 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. Assessment of these features 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 modelling 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 for binding events separated by 200-400 bp ChIP-exo exhibits a significantly higher sensitivity.

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

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 popular 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 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 adaptors to enable monitoring of overamplification. Figure 1A 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.

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 ackage` named `ChIPexoQual`. We apply this pipeline on a large collection of public and newly generated ChIP-exo/nexus

data. We validate 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 reveal that the ChIP-exo peak-pair assumption, underlying most of the ChIP-exo analysis pipelines 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 and PE ChIP-seq (Table 2). 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. For simplicity, we named these experiments as E1, P1, and S1, respectively. 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. We also named these experiments E2, P2, and S2 for ChIP-exo, PE ChIP-seq and SE ChIP-seq, respectively. 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 1(B-E) summarizes this comparison for one biological replicate of ChIP-exo and ChIP-seq experiments from the same biological conditions (samples EB1-1 and P1-1 from Table 2). Comparisons with other paired *E. Coli* ChIP-seq and ChIP-exo samples led to similar conclusions (Supplementary Figure S1). We summarized the extended raw read counts within 150 bps non-overlapping intervals, i.e., bins, interrogating the genome. Figure 1B depicts that, as observed by others, 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 potential differences in the background read distributions of those data types.

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 1C). 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 85% 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 1E and 1F 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 to account for variability in the background read distribution.

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.

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

Protocol	Organism	TF	Cell type	Rep.	Depth	NSC	RSC	PBC
ChIP-exo	Human	CTCF	HeLa		48,478,450	16.02	1.1960	0.4579
				1	9,289,835	19.87	1.0127	0.8082
	Human	ER	MCF-7	2	11,041,833	21.48	1.0063	0.8024
				3	12,464,836	18.72	1.0100	0.8203
				1	22,210,461	21.28	1.1104	0.6562
	Mouse	FoxA1	Liver	2	23,307,557	60.42	1.1604	0.7996
				3	22,421,72	72.04	1.1975	0.1068
			IMR90	1	47,443,803	8.86	1.3678	0.2978
	Human	GR	K562	1	116,518,000	4.11	1.0441	0.0504
			U2OS	1	3,255,111	10.05	1.0288	0.7714
				1	61,046,382	12.01	1.1119	0.1232
	Human	TBP	K562	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.

Group	Growth	Treatment	Rep.	Id.	Depth	NSC	RSC	PBC
ChIP-exo (E1)	Exp. $+O_2$	No Rif.	1	1	13,961,493	103.15	2.0193	0.1399
	Exp. $+O_2$	No Rif.	2	2	14,810,838	162.70	1.7805	0.1633
	Stat. $+O_2$	No Rif.	1	3	16,108,774	153.51	1.8035	0.1353
	Stat. $+O_2$	No Rif.	2	4	13,636,541	172.59	2.014	0.1532
ChIP-seq PE (P1)	Exp. $+O_2$	No Rif.	1	1	27,665,432	4.01	3.9582	0.3869
	Exp. $-O_2$	No Rif.	1	2	44,707,340	3.56	3.3045	0.3134
ChIP-seq SE (S1)	Exp. $+O_2$	No Rif.	1	1	7,456,068	3.27	2.3863	0.5629
	Exp. $-O_2$	No Rif.	1	2	11,467,086	2.91	2.1362	0.5452
ChIP-exo (E2)	Exp. $+O_2$	No Rif.	1	1	902,921	13.77	1.1270	0.2689
	Exp. $+O_2$	Rif. 20 min	1	2	1,852,124	17.91	1.5275	0.2590
	Exp. $+O_2$	No Rif.	2	3	2,104,427	29.60	1.2844	0.2584
	Exp. $+O_2$	Rif. 20 min	2	4	11,548,572	13.08	1.5122	0.1510
ChIP-seq PE (P2)	Exp. $+O_2$	No Rif.	1	1	13,445,022	8.86	1.0541	0.9426
	Exp. $+O_2$	Rif. 20 min	1	2	16,538,920	7.03	1.0157	0.9378
	Exp. $+O_2$	No Rif.	2	3	11,642,722	10.77	1.0145	0.8891
	Exp. $+O_2$	Rif. 20 min	2	4	16,854,026	7.93	1.0048	0.9407
ChIP-seq SE (S2)	Exp. $+O_2$	No Rif.	1	1	6,722,511	9.01	2.8461	0.6632
	Exp. $+O_2$	Rif. 20 min	1	2	8,269,460	7.17	2.5168	0.5594
	Exp. $+O_2$	No Rif.	2	3	5,821,361	10.89	3.1291	0.6472
	Exp. $+O_2$	Rif. 20 min	2	4	8,427,013	8.12	2.6908	0.5895

Table 2 Summary of the E. coli σ^{70} ChIP-exo and ChIP-seq samples. Exp. stands for exponential and Stat. for stationary growth conditions. 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.

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 1D 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 S4 to S11. 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.

ChIP-exo quality control pipeline `ChIPexoQual`

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 2 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 (by the experiment's read length) SK: check: extended or not? What do we extend to in ChIP-exo? RW: right now, we don't extend. we use the coverage plot with the experiment's read length to partition the coverage into the regions. In the package I left it as a parameter but I don't is going to change a lot. For example, the ChIP-exo datasets read length is 45 and it's regions seems to be composed by scattered reads in both sides. So, I think that using a higher value will only made the regions wider but will not increase the number of unique position or the depth by a lot 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, URC_i, and FSR_i are computed for each region i. ARC_i denotes the *average read coefficient* and is defined as the ratio of the # of reads in island i (D_i) to the width of the island i (W_i); URC_i, *unique read coefficient*, quantifies the 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.

Figure 2A presents the typical behavior of the URC vs. ARC plot for a high quality ChIP-exo sample. In general, the plot depicts two strong arms. High URC

values correspond to regions with reads concentrated on a small number of positions. The left arm, with low ARC and varying URC 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 URC decreases as the ARC increases corresponds to regions that are usually ChIP enriched. RW: Finally we quantify the shape of the ARC vs. URC plot by the use of two estimated parameters: β_1 which represents the average number of reads aligned to the unique positions in large depth regions and β_2 which represents the overall change in depth as the width varies across a large set of regions. RW: Figures 2B and 2C presents the typical behavior of the Region Composition and Forward Strand Ratio distribution plots respectively, both quantify the strand imbalance as part of the QC pipeline. The former depicts how quickly the islands exclusively composed by fragments on 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. Even though not every region in a ChIP-exo experiment is perfectly balanced, the most enriched regions are expected to be composed by approximately the same quantity of reads in both strands.

Application and validation of `ChIPexoQual` *with the FoxA1 ChIP-exo dataset*
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 3A presents URC vs. ARC plots for all three replicates. The first and third replicates exhibit a defined decreasing trend in URC as the ARC increases. This indicates that these samples exhibit a higher ChIP enrichment than the second replicate. On the other hand, the overall URC 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.

SK: I have a hard time interpreting B and C specifically, need to include something about their implications for replicates 2 and 3.

Figure 3B and 3C display the Read Composition and FSR distribution plots, which highlight specific problems with replicates 2 and 3. RW: In Figure 3B, we can observe decreasing trends in the proportions of regions formed by fragments in one exclusive strand. In a high quality experiment its is expected to observe an exponential decay in the proportion of single stranded regions, while in lower quality experiments the trend may be linear of even constant (Supplement Figure S25). In the FSR distribution plot, both replicate 2 and 3 FSR's distributions are more spread around their respective medians. The approximation of the 0.1 and 0..9 quantiles to the median indicated the aforementioned lower enrichment in the second replicate and the low complexity in the third one.

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? RW: I used a set of high quality regions, i.e. using the ChIP-exo pipeline I filtered them by npos, FSR and depth. Figure S26 show the overlap between these regions and the peaks (roughly speaking these regions are almost a proper subset of the peaks). Figure 4A 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 UCR vs. ARC diagnostic plot. Scanning of these regions for the occurrence of FoxA1 sequence motif with the FIMO tool [18] indicates that the first replicate outperforms the other two in terms of percentage of candidate regions with the FoxA1 motif (Figure 4B).

Furthermore, Figure 4C 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. Figure 4D further highlights the overall quality of the identified motif sequences for each replicate and suggest that libraries with high library complexity (replicates 1 and RW: 2) capture binding sites with better motif matches. RW: Figure 4E compares the top FIMO scores among the three replicates, not-surprisingly confirming that the first explicate exhibits the highest quality. Finally, Figure 4F compares the Strand Imbalance by considering the regions overlapping with their respective sets of ChIP-exo peaks. Not surprisingly, for the replicate 2 we can see that the Imbalance distribution differ among both classes, which is caused by the aforementioned lower enrichment of that sample; for the remaining replicates we can see that the Imbalance is more similar when regions composed by more reads are considered. This suggests that enriched regions are more likely to be balanced, therefore this suggest that the "peak-pair" assumption may not hold in non-enriched regions.

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 S12 to S22). A key, albeit not surprising, observation from large scale analysis is that the URC 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 subsamples 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. URC vs. ARC diagnostics of these sub-samples (Supplementary Figures S23 to S25) 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. RW:

To confirm this, we compared the top FIMO scores [18] of the TBP motif calculated with the ChIP-exo and ChIP-nexus replicates. Figure 5E illustrates that the scores of the ChIP-nexus replicates and the first ChIP-exo replicate are higher than the other two samples when considering the top scores. As expected the ChIP-nexus scores are comparable among and show better performance than the three ChIP-exo experiment and among those three replicates, replicate 1 outperforms the other two.

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 URC by fitting a reparametrized regression model of URC 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 $\text{URC}_i = \frac{\kappa}{\text{ARC}_i} + \gamma + \epsilon_i$, which aims to recapitulate the relationship in the URC vs. ARC plots. Figure 5A 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) for the URC of a sample. As discussed earlier, high quality ChIP-exo samples are expected to have two arms in the URC vs. ARC plots: one with low ARC and varying URC and another with a decreasing URC as ARC increases and stabilizes β_1. When the ChIP-exo sample is not deeply sequenced, high values of $\hat{\beta}_1$ in Figure 5A 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 (Supplementary Figures S12 to S22), we conclude that samples with estimated β_1 values less than 10 seem to be high quality samples.

RW: We interpret the β_2 as the overall change in depth as the width varies and display its estimates across all the eukaryotic samples in Figure 5B. Under perfect digestion by λ-exo, most of the reads aligned to binding regions are expected to accumulate around a binding event. RW: In a high quality experiment the overall variation in depth is going to be small as the overall width of the regions changes, as this change is not going to involve the majority of the reads as those are located tightly around the binding sites. On the other hand, in low quality experiments regions are usually composed of a fixed proportion of reads aligned to a small number of unique positions; hence the overall change in depth as the width varies is proportional to this fixed proportion. SK: I don't follow the next argument: We should say something about the expected behavior of β_2 RW: Informally, this is what I tried to say: If we have an enriched region and the width is being reduced, then the depth of the region is not going to change a lot since the majority of the reads are localized around the binding sites and not the extremes. On the other hand, in a non-enriched region, by reducing its width then several reads may have been lost (in this case I am thinking about a regions of width w_i, with k reads aligned to any of the u_i positions, in this case, since the region is not enriched we observe $u_i << w_i$ and $d_i = ku_i$. Finally, $\frac{\Delta d_i}{\Delta w_i} \approx k/h$, where $h = \Delta w_i \leq rl$.)

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 5B), $\hat{\beta}_2$ 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 $\hat{\beta}_1$ is large or $\hat{\beta}_2$ is different from zero 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 5C 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 5D illustrates that the β_2 estimates 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 (Supplementary Figures S23 to S25) illustrate that the two arms of the ARC versus URC 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 E1-1 and E1-2). 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 6A and 6B 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 6A) and dPeak, on average, has slightly better resolution than the rest in the second replicate (Figure 6B).

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 6C and 6D compare the densities of these parameters estimated by dPeak for *E. coli* ChIP-exo and SE ChIP-seq data (sample E1-1 and S1-1, respectively). 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 [19], 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 7 summarizes the comparisons of the three data types for σ^{70} under aerobic condition (samples E1-1, P1-1 and S1-1) as a function of sequencing depth. For these computational experiments, we used σ^{70} RegulonDB [13] binding events as gold standard. Figure 7A displays the number of candidate regions (i.e., peaks) where at least one binding event was identified whereas Figure 7B depicts the number of identified binding events, i.e., each candidate region can harbour multiple binding events. In terms of the number of binding events, ChIP-exo ranks on the conservative side. In Figure 7C, 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 7D 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 8 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 depths are comparable or all the samples have depths sufficient enough for

saturation RW: The samples have sufficient depth for saturation: ChIP-exo is 14M, PE ChIP-seq is 27M and SE ChIP-seq is 7.5M. Since the main point is to compare ChIP-exo and PE-ChIP-seq I think this plots are good as PE ChIP-seq has twice as many reads as ChIP-exo. About the comparison with SE ChIP-seq I don't it works very well, as the # of ChIP-exo fragments is almost double the amount of SE ChIP-seq. So, I think we have the following options: A) Keep it with the disclaimer that the comparison against SE ChIP-seq is not as important as the other, B) Remove the SE ChIP-seq part but keep ChIP-exo vs. PE ChIP-seq or C) Send the whole figure / section to the supplement.. 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 8A 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 8B 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 `ChIPexoQual` package is available at `https://github.com/keleslab/ChIPexoQual`.

Materials and methods

ChIP-seq/exo/nexus datasets

E. Coli ChIP-exo and ChIP-seq samples

RW: *Growth conditions.* All strains were grown in MOPS minimal medium supplemented with 0.2% glucose [20] at $37°C$. E1 samples were sparged with a gas mix of 1% CO_2, 30% O_2 and 69% N_2; P1 and S1 aerobic samples with a gas mix of 95% N_2 and 5% CO_2 (anaerobic) or 70% N_2, 5% CO_2o and 25% O_2 (aerobic); E2 and

P2 samples with a gas mix of 5% CO_2, 25% O_2 and 70% N_2, additionally samples E2-2, E2-4, P2-2 and P2-4 were treated with *Rifampicin* by 20 minutes. P1-2 and S1-2 samples were harvested during a stationary growth ($OD_{600} \geq 1.0$ for 1 hour). The remaining samples were harvested during a mid growth ($OD_{600} = 0.3$ using a Perkin Elmer Lambda 25 UV/Vis Spectrophotometer). WT *E. Coli* K-12 MG1655 was used for the E1, P1 and S1 experiments and WT *E. coli* K-12 RL3000 was used for the E2 and P2 experiments.

RW: *Library preparation and sequencing.* For ChIP-seq experiments, 10 ng of immunoprecipitated and purified DNA fragments from the aerobic and anaerobic σ^{70} samples (one biological sample for both aerobic and anaerobic growth conditions), along with 10 ng of input control (two biological replicates for anaerobic Input and one biological sample for aerobic Input), were submitted to the University of Wisconsin-Madison DNA Sequencing Facility for ChIP-seq library preparation. Samples were sheared to 200 - 500 nt during the IP process to facilitate library preparation. All libraries were generated using reagents from the Illumina Paired End Sample Preparation Kit (Illumina) and the Illumina protocol *"Preparing Samples for ChIP Sequencing of DNA"* (Illumina part # 11257047 RevA) as per the manufacturer's instructions, except products of the ligation reaction were purified by gel electrophoresis using 2% SizeSelect agarose gels (Invitrogen) targeting 275 bp fragments. RW: need part about enzyme digestion After library construction and amplification, quality and quantity were assessed using an Agilent DNA 1000 series chip assay (Agilent) and QuantIT PicoGreen dsDNA Kit (Invitrogen), respectively, and libraries were standardized to 10μM. RW: For Chip-exo?For PE ChIP-seq data, cluster generation was performed using an Illumina cBot Paired End Cluster Generation Kit (v3). Paired reads, 36 bp run was performed for each end, using 200 bp v3 SBS reagents and CASAVA (the Illumina pipeline) v 1.8.2, on the HiSeq2000. For SE ChIP-seq data, cluster generation was performed using an Illumina cBot Single Read Cluster Generation Kit (v4) and placed on the Illumina cBot. A single read, 32 bp run was performed, using standard 36 bp SBS kits (v4) and SCS 2.6 on an Illumina Genome Analyzer IIx. Base calling was performed using the standard Illumina Pipeline version 1.6.

RW: The S2 group samples were built *in silico*, by randomly sampling one of the two ends in a read pair from the respective PE ChIP-seq experiment in the P2 group with the same conditions.

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) [21]. The E1 samples were aligned by using `bowtie -q -m 1 -l 55 -k 1 -5 3 -3 40 --best -S` and the E2 samples were aligned by using `bowtie -q -m 1 -v 2 --best`.

ChIP-exo and ChIP-seq peak calling with MOSAiCS

RW: `MOSAiCS` [15] is a model-based approach for the analysis of ChIP-exo and ChIP-seq data. We used `MOSAiCS` to identify sets of high quality peaks for ChIP-exo and ChIP-seq under the GC + Mappability and Input Only modes for background estimation, respectively. Subsequently, we called peaks with a 5% FDR and a threshold of at least 100 extended fragments.

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

We partitioned the *E. coli* genome into non-overlapping 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. The subset of these peaks for which `dPeak` analysis identified one or more binding events were used in FSR assessments (Figure 1C and Figure S1).

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.

RW: I did a quick check and the plot looks more like fig. S1: The ChIP-seq density seems to be more localized around 0.5 (expected), and the ChIP-exo density seems to become more flat. I think, this happens because when we remove peaks with more than 1 TFBS, we are also removing peaks where at least two binding sites are located close to each other, which makes harder for the λ enzyme to reach those regions, so as the distance between the binding events get shorter we are observing regions that behave more like ChIP-seq in than ChIP-exo. in 0A we can see that the mean distance between binding sites is below 100 for the majority of the peaks.

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.

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:

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

where W_i denotes the width of island i.

4 Generate diagnostic plots (i) URC vs. ARC plot; (ii) Region Composition plot; (iii) FSR distribution plot.

5 Randomly sample M (at least 1000) 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:

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

with $\beta_1 = 1/\gamma$ and $\beta_2 = -\kappa/\gamma$. In this setting, γ can be considered as the large-depth URC_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 URC and assuming that γ is already estimated, we can observe the following identities:

$$\kappa = \frac{U}{W} - \gamma\text{ARC}$$
$$\frac{\kappa}{\gamma} = \frac{1}{\gamma}\frac{U}{W} - \text{ARC}$$

This is important: γ approximates the URC 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:

$$\frac{U}{W} \times \frac{1}{\hat{\gamma}} = \frac{U}{W} \times \lim_{D\to\infty} \frac{D}{U(D)} \sim \text{ARC} \tag{2}$$

Therefore, $\beta_2 = -\kappa/\gamma$ is interpreted as the ARC bias as the sequencing depth increases. RW: In a high quality experiment γ is estimated by a positive value, additionally for a given region in the aforementioned high quality experiment reads align to the majority of the possible mappable positions, therefore we can consider that the first term in (2) converges to the observed ARC of that region. On the other hand, if we consider a non-enriched region in low quality experiment, then the reads are aligned to few unique positions, showing that κ/γ theoretically diverges.

SK: ***

Motif analysis of FoxA1 and TBP enriched regions

For each ChIP-exo/nexus sample, we used the ChIP-exo QC pipeline to partition its respective genome into a set of islands with their respective summary statistics. We then filtered them into collections of high quality regions by:

- FoxA1 experiments: (i) removing the islands with reads residing only on one strand; (ii) removing the islands with $U_i \leq 15$; (iii) removing islands with $D_i < 100$.
- TBP experiments: (i) removing the islands with reads residing only on one strand; (ii) removing islands with $W_i < 50$ or $W_i \geq 2000$ bp; (iii) removing islands with the $D_i <= \text{median}_j D_j$.

These thresholds were empirically selected and the overall conclusions were robust to their variation.

We used `FIMO` (version 4.9.1) [18] to identify the FoxA1 and TBP motifs within each enriched regions using the FoxA1 MA0148.1 and TBP MA0108.1 position weight matrices from the `JASPAR` database [22] respectively. For the FoxA1 experiments we used the default parameters and for the TBP experiments we considered all motifs found with p.value less than 0.01.

Testing for strand imbalance in FoxA1 ChIP-exo replicates

We partitioned the mouse genome into a set of islands by using the ChIP-exo QC pipeline. Subsequently, we filtered the islands with reads in only one strand. We define an `Imbalance`, that is zero when the region is composed by the same amount of reads in both strands and infinity when consists of reads in one strand exclusively, as:

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

We divided the ChIP-exo regions into two classes depending on whether they overlap the respective replicate set of high quality ChIP-exo peaks and compared the two classes `Imbalance`'s distribution.

SK: Is this showing the imbalance index is different between peak and non-peak regions? To show that the class that don't overlap with peaks exhibits heavier tails, we used a Wilcoxon test over the *Imbalance index*.

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 σ^{70} site was deemed as identified if there exists 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`). RW: For each method, we calculated the resolution by considering only the annotations overlapping any of the top 400 peaks called by `MOSAiCS`. Default tuning parameters were used during model fitting for all methods, except for GEM where the peaks were explicitly provided as candidate regions. 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 it requires more than one replicate to call ChIP-exo peaks.

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

SK: Next paragraph could use more editing.

RW: We sub-sampled 10K to 100K (in 10K increments) reads from each of the σ^{70} ChIP-exo, PE and SE ChIP-seq datasets. Specifically, for every N reads in ChIP-exo and SE ChIP-seq, $N/2$ read pairs were sampled from PE ChIP-seq to operate under fixed sequencing costs. For each dataset, we called peaks using `MOSAiCS` [15] and estimated TFBS using `dPeak` [8] by considering at most 5 binding sites for each peak. 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 different seeds and reported the median across there experiments.

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

RW: 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 deconvolved the peaks into binding events with `dPeak` 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.

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. Grant C, Bailey T, Noble WS. FIMO: Scanning for occurrences of a given motif;.
19. Zhang Q, Zeng X, Younkin S, Kawli T, Snyder MP, Keleş S. Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. BMC Bioinformatics. 2016;.
20. Neidhardt FC, Bloch PL, Smith DF. Culture Medium for Enterobacteria. Journal of Bacteriology. 1974;.
21. Langmead B, Trapnell C, Mihal P, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;.
22. Mathelier A, Fornes O, Arenillas DJ, Chen Cy, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Research. 2016;.
23. Robinson DG, Storey JD. subSeq: Determining appropiate sequencing depth through efficient read subsampling. Bioinformatics. 2014;.
24. Mendenhall EM, Bernstein BE. DNA-protein interactions in high definition. Genome Biology. 2012;.
25. 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;.
26. 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;.
27. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature Reviews: Genetics. 2012;.
28. 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;.
29. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2011;.

1 Figures

RW: The enumeration of some figures is a little bit off, I am going to fix it after we finish the text

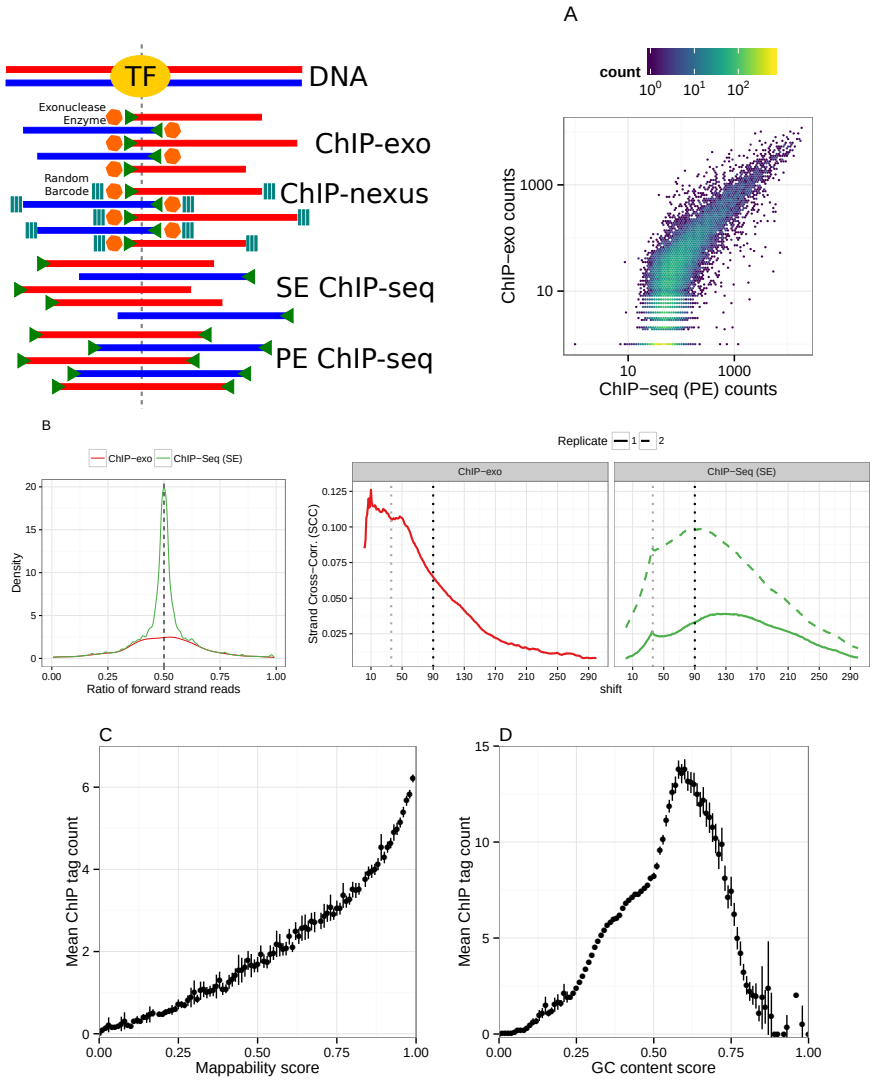

Figure 1 A) A TF is bound to the forward (red) and backward (blue) strand of the DNA. Then, is sonicated: For ChIP-exo a exonuclease enzyme (orange hexagon) trims the $5\prime$ ends of each DNA fragment to a fixed distance from the bound protein, finally is subjected to Immunoprecipitation and amplification. For ChIP-nexus a random barcode is added on the $3\prime$ end, and transferred to the $5\prime$ stopping base by self-circularization, after the library is amplified and the reads aligned, reads with identical barcode and start position are removed. 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. B) Hexbin plot of PE ChIP-seq bin counts vs ChIP-exo bin counts. C) Forward Strand Ratio densities for SE ChIP-seq and ChIP-exo peaks. D) 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 [29] 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. E) Mappability score vs mean ChIP tag counts with 0.95 confidence bands. F) GC - content score vs mean ChIP tag counts with 0.95 confidence bands.

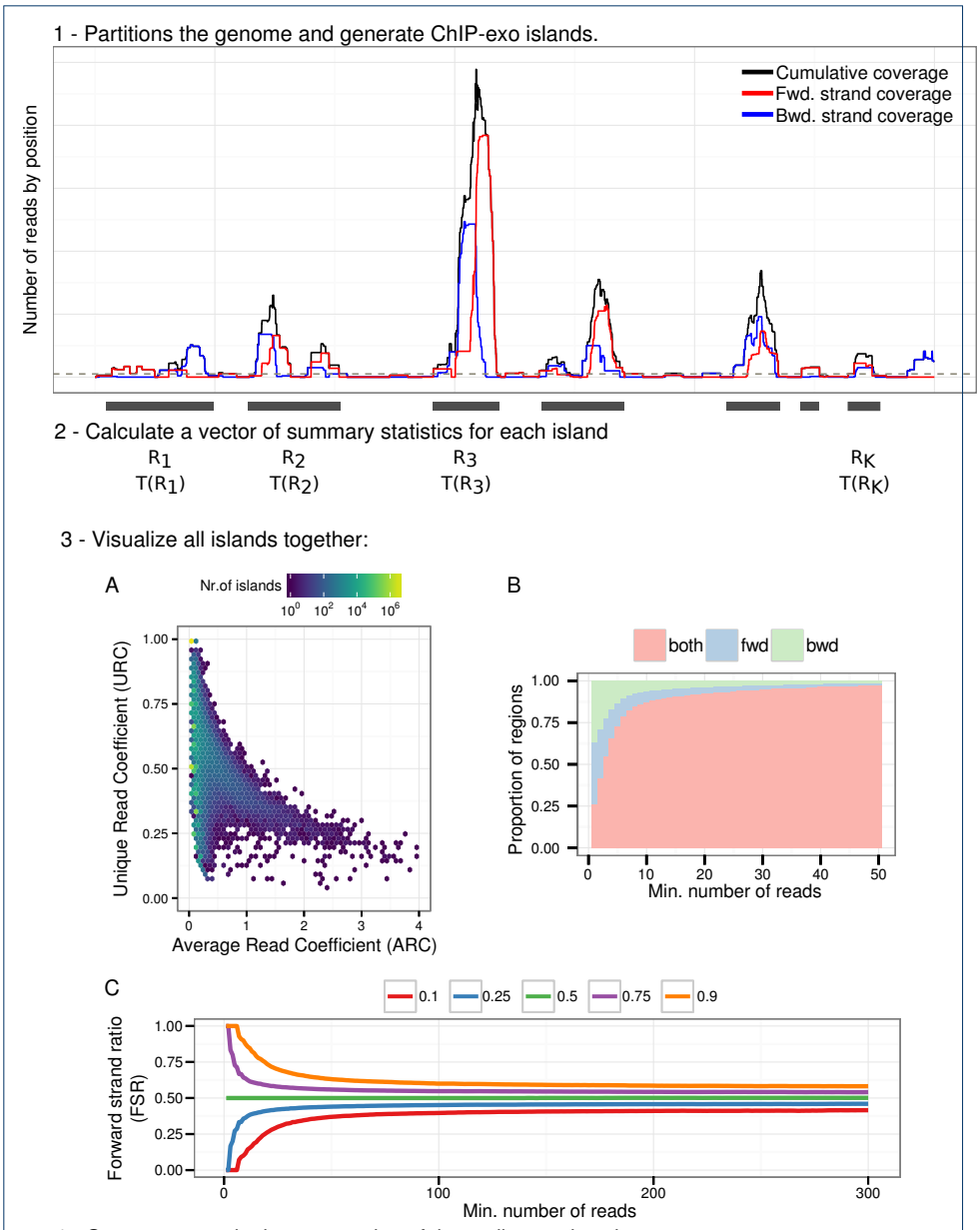

Figure 2 RW: The ChIP-exo reads are partitioned by keeping the regions in the genome composed by the undigested fragments. For each region the following summary statistics are calculated: Average Read Coefficient (ARC), Unique Read Coefficient (URC) and Forward Strand Ratio (FSR). These statistics are visualized as: A) ARC vs. URC plots, which presents the overall balance between library complexity and enrichment; there are two arms, one with low ARC, that corresponds to regions formed by few aligned positions and one where the URC decreases as the ARC increases, that corresponds to potentially enriched regions. B) Region Composition plot, which shows the strand composition for all the regions formed by a min. number of reads; low depth regions are commonly formed by undigested reads in one unique strand, while high depth regions are usually formed by reads in both. C) FSR distribution plot, it illustrates the FSR's distribution as as higher depth regions are being considered: In a high quality experiment the FSR's distribution is tightly centered around the FSR's median.

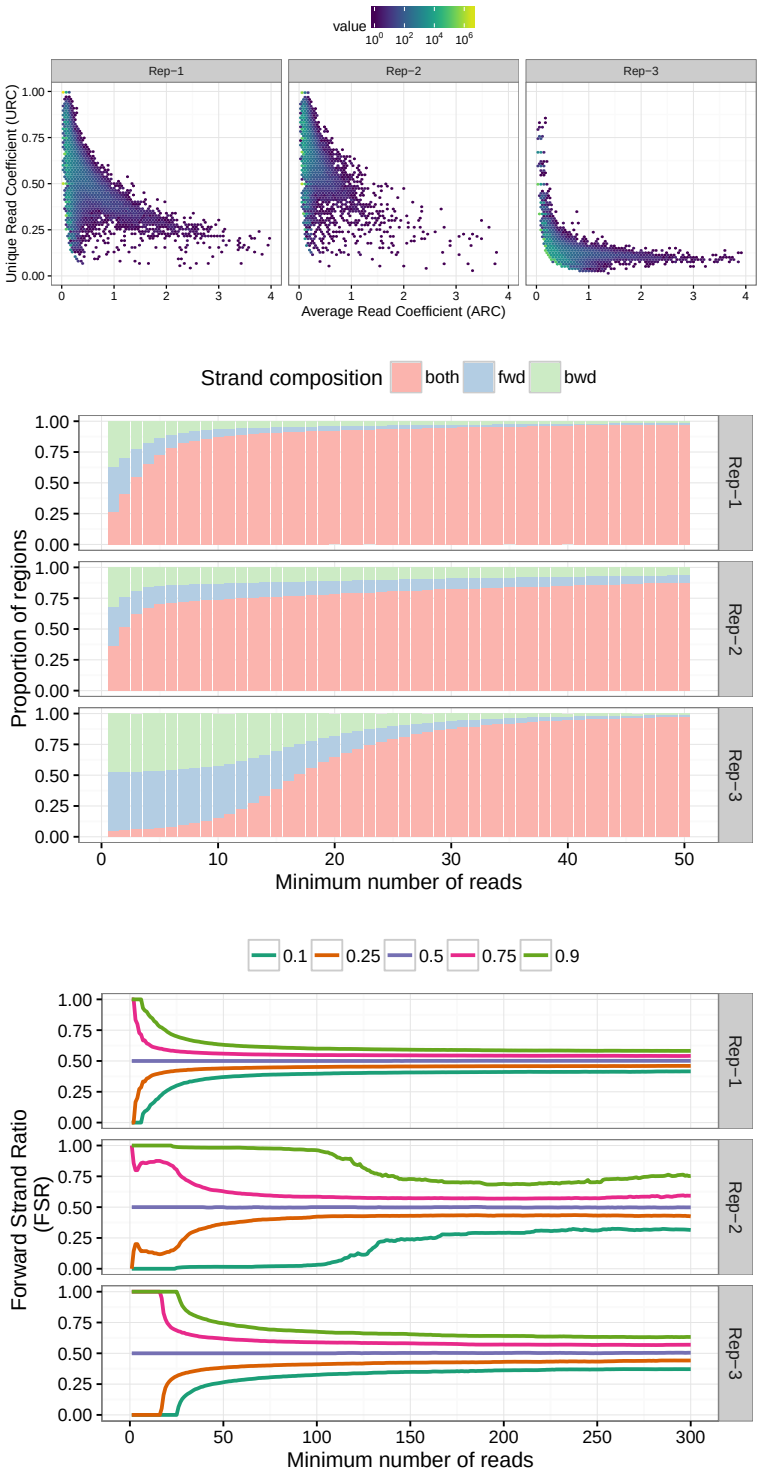

Figure 3 RW: Diagnostic plots generated by `ChIPexoQual`. Comparison of A) ARC vs. URC plot, B) Region Composition plot and C) FSR distribution plot between the three FoxA1 replicates in mouse liver tissue from [10].

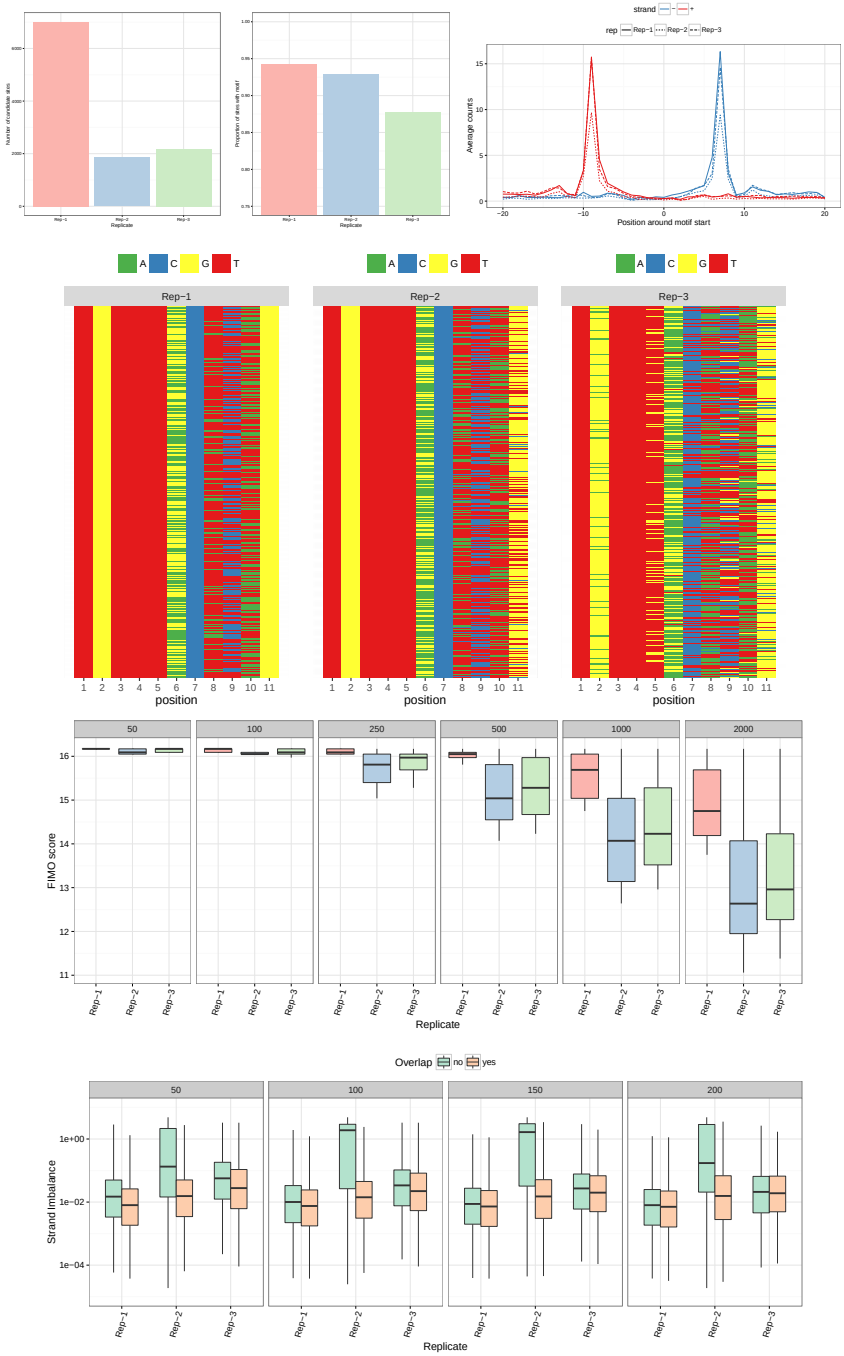

Figure 4 Using the mouse FoxA1 experiment from [10] and the FoxA1 motif with MA0148.1 id in the JASPAR database: RW: A) Number of high quality regions where the FoxA1 motif was searched. B) Proportion of candidate regions with motif. C) FoxA1 Average Coverage plots centered around motif start positions separated by replicate and strand. D) Sequence composition of sites with motif for each replicate. E) Comparison of the top 50, 100, 250, 500, 1000 and 2000 FIMO scores for each replicate. F) Strand Imbalance comparison of regions with at least 50, 100, 150 and 250 reads overlapping with ChIP-exo peaks for each replicate.

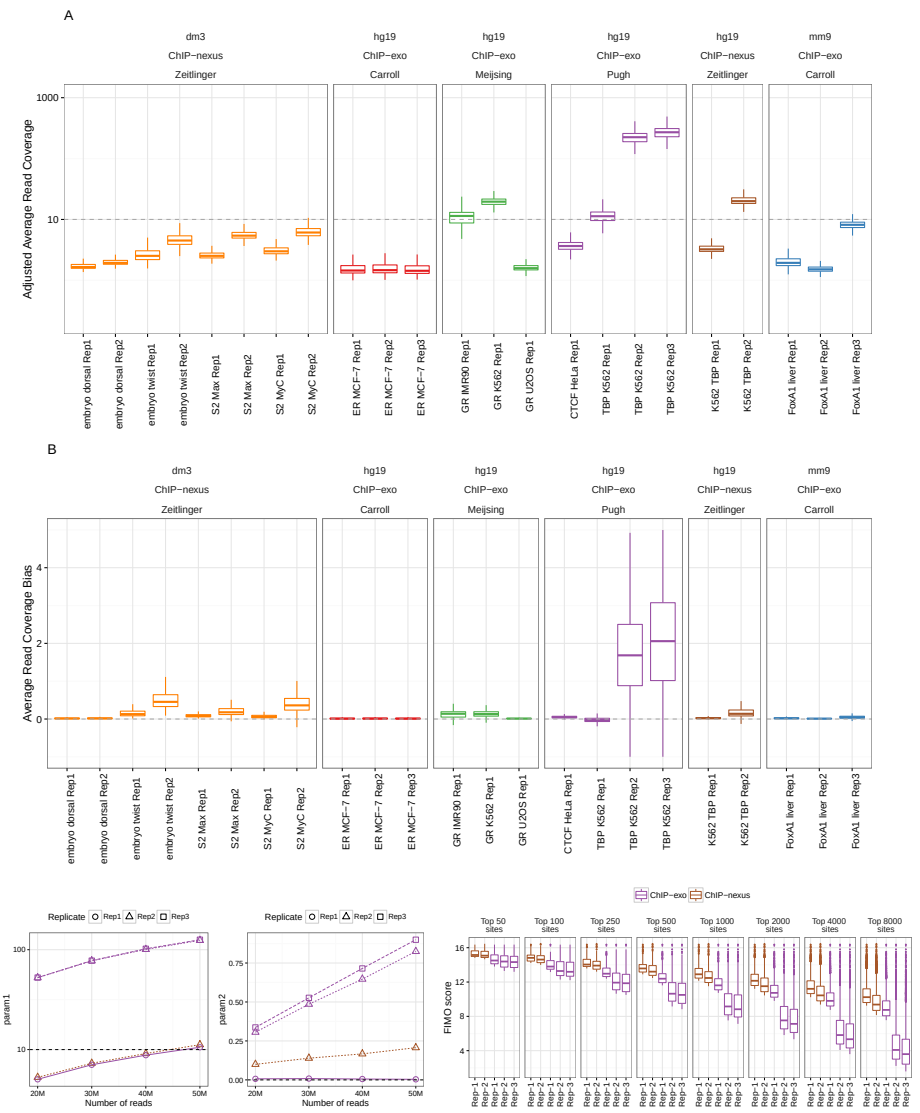

Figure 5 RW: Comparison of all estimated A) β_1 and B) β_2 for all eukaryotic ChIP-exo/nexus samples. Comparison of the average estimated A) β_1 and B) β_2 for the ChIP-exo/nexus TBP samples in K562 cell lines when sub-sampling 20M to 50M reads. E) Comparison of the top 50, 100, 250, 500, 1000, 2000, 4000 and 8000 FIMO scores for each TBP experiment. The scores were calculated with the TBP MA0108.1 motif in the JASPAR database.

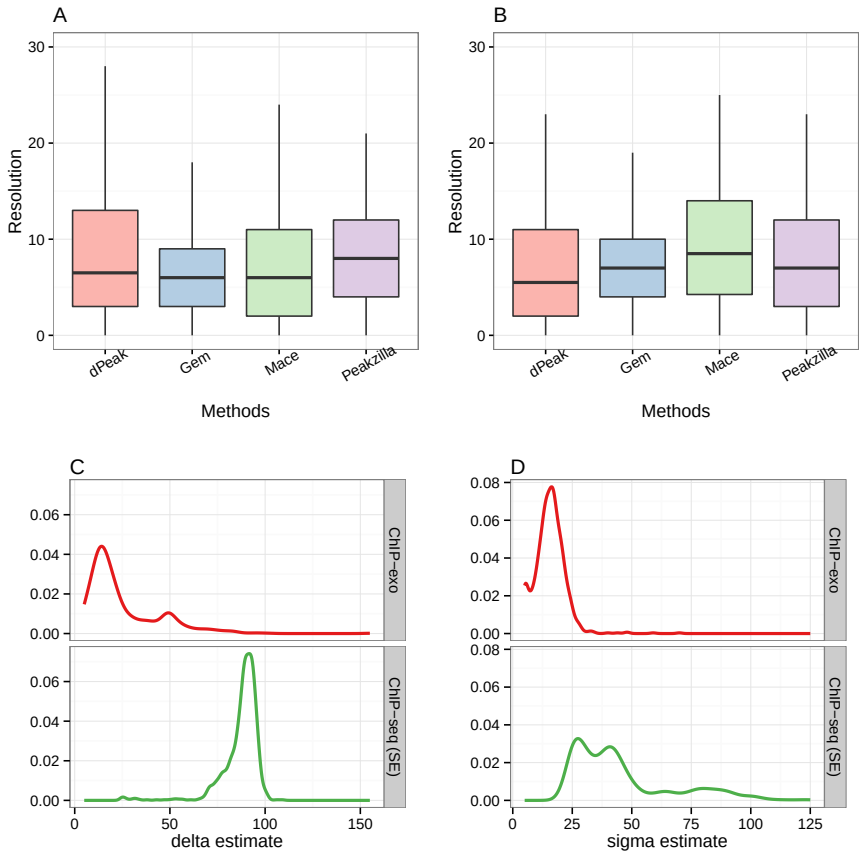

Figure 6 RW: Resolution comparison between dPeak, Gem, Mace and Peakzilla using the A) E1-1 and B) E1-2 *E.coli* ChIP-exo experiments. Estimated C) δ and D) σ parameter densities for ChIP-exo (E1-2 sample) and SE ChIP-seq (S1-1 sample).

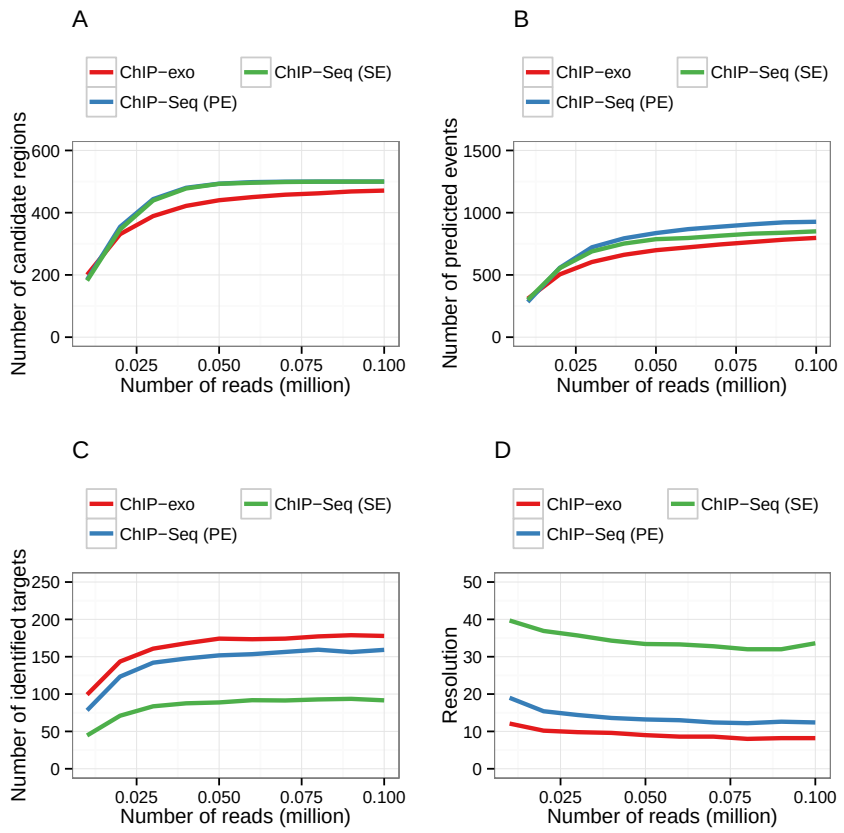

Figure 7 Comparison of the number of A) candidate regions, B) predicted events, C) identified targets and D) resolution among ChIP-exo, PE ChIP-seq and SE ChIP-seq.

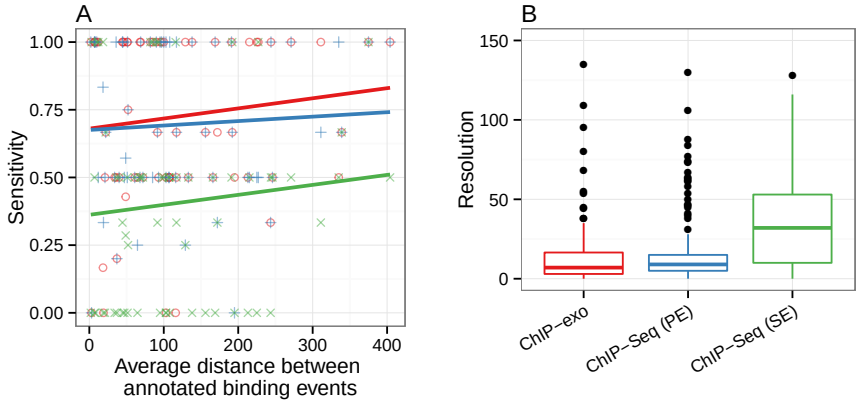

Figure 8 Comparison of A) sensitivity and B) resolution between ChIP-exo and ChIP-seq data.