DRAFT

Data exploration, quality control and statistical analysis of ChIP-exo 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 is a modification of the 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, ChIP-exo presents a number of unique challenges. We present a quality control pipeline to evaluate a number of key issues including strand imbalance, library complexity and enrichment. Assessment of these characteristics are facilitated through diagnostic plots and summary statistics calculated over regions of the genome with varying levels of coverage.

The pipeline explores and quantifies these aspects by partitioning the experiment reads into a collection of regions, calculating a series of summary statistics for each region, providing visualizations and calculating measures to globally asses the quality of a ChIP-exo experiment. We provide guidelines to distinguish libraries with low complexity from deeply sequenced experiments by the use of the QC pipeline. We demonstrate that the ChIP-exo QC pipeline it is also applicable to ChIP-nexus data, showing that those experiments present higher quality than ChIP-exo experiments under similar conditions.

We compared ChIP-exo with Paired End (PE) and Single End (SE) ChIP-Seq and found the following characteristics: First, although often assumed in ChIP-exo data analysis methods, the "peak pair" assumptions does not hold locally in actual ChIP-exo data. Second, we for the first time compared PE ChIP-Seq with ChIP-exo and found that at fixed sequencing depths, ChIP-exo provides higher sensitivity, specificity and spatial resolution than PE ChIP-Seq. Finally, 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 show higher sensitivity than PE ChIP-Seq.

Keywords: ChIP-exo; ChIP-nexus; ChIP-Seq; Quality Control; Spatial Resolution; Transcription Factor; Binding Site Identification on High-Res; Deconvolution

Background

ChIP-exo (Chromatin Immunoprecipitation followed by exonuclease digestion and next generation sequencing) Rhee and Pugh, 2011 is the state-of-the-art experiment developed to attain single base-pair resolution of protein binding site identification and it is considered as 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 bp in length) that the protein under study interacts with using random fragmentation of DNA and a protein-specific antibody. Then, exonuclease is introduced to trim the 5′ end of each DNA fragment to a fixed distance from the bound protein compared to ChIP-Seq. This step is unique to ChIP-exo and could potentially provide significantly higher spatial resolution compared to ChIP-Seq. Finally, high throughput sequencing of a small region (36 to 100 bp) at the 5′ end of each fragment generates millions of reads. Figure 1 illustrates the differences between ChIP-exo, Single End (SE) ChIP-Seq and Paired End (PE) ChIP-Seq: The 5′ ends of a ChIP-exo experiment are located more tightly around the binding proteins than in a ChIP-Seq experiment; in a PE ChIP-Seq experiment both ends are observed while in a SE ChIP-Seq experiment only the 5′ end.

ChIP-nexus (Chromatin Immunoprecipitation followed by exonuclease digestion, unique barcode, single ligation and next generation ligation) He et al., 2015 is a modification to the ChIP-exo protocol, where both sequencing adaptors are ligated at the end of the ChIP fragments. Then, after exonuclease digestion, DNA self-circularization with circLigase, and restriction enzyme cutting between the two adaptors, the final library is amplified. ChIP-nexus arises as an alternative to ChIP-exo as it provides the possibility of attaining higher resolution analysis and yields higher complexity libraries.

While the number of ChIP-exo data keeps increasing, characteristics of ChIP-exo data are not fully investigated yet. First, DNA libraries generated by the ChIP-exo protocol are expected to be less complex than the libraries generated by ChIP-Seq (Mahony et al., 2015). Second, although there are roughly the same amount of reads in both strands, locally there may be more reads in one strand than in the other. Finally, most of current ChIP-Seq quality control (QC) guidelines (Landt et al., 2012) may not be applicable on ChIP-exo, while there are not established QC pipelines for ChIP-exo; previous ChIP-exo analyses used ChIP-Seq samples to compare the resolution between experiments [1, 5, 6]. To address these challenges, we suggest a collection of quality control visualizations to interrogate these biases in a ChIP-exo experiment and globally asses the enrichment and library complexity of a ChIP-exo sample and a procedure to distinguish low complexity libraries from deeply sequenced experiments. This aspect is unique to ChIP-exo, since the exonuclease enzyme it is expected to digested the reads that are not bound to a transcription factor, therefore the number of bases where a ChIP-exo fragment could be potentially aligned is reduced. That way, in a high quality ChIP-exo experiment is possible to observe a large amount of reads to be aligned to unique position due to genomic signal instead of PCR artifacts.

In order to obtain the potential benefits of ChIP-exo on protein binding site identification, it is critical to use algorithms that could fully utilize information

available in ChIP-exo data. Rhee and Pugh, 2011 discussed that reads in the forward and reverse strand might construct peak pairs around bound proteins, of which heights were implicitly assumed to be symmetric. Based on this rationale, they used the "peak pair method" that predicts the midpoint of two modes of peak pairs as potential binding sites. Recently developed ChIP-exo data analysis methods, such as Mace (Wang et al, 20114), CexoR (Madrigal, 2015) and Peakzilla (Bardet et al., 2013), are also based on this peak pair assumption. However, appropriateness of such assumption was not fully evaluated in the literature yet. Furthermore, it is still unknown which factors could affect protein binding site identification using ChIP-exo data. In order to address this problem, we investigated various aspects of ChIP-exo data by contrasting them with their respective ChIP-Seq experiments.

Currently, research on statistical methods for ChIP-exo data is still in its very early stage. Although many methods have been proposed to identify protein binding sites from ChIP-Seq data (reviewed by Wilbanks and Facciotti, 2012 and Pepke and Wold, 2009), such as MACS (Zhang et al., 2008), CisGenome (Ji et al., 2008) and MOSAiCS (Kuan et al., 2009), these approaches might not fully utilize potentials of ChIP-exo data for high resolution identification of protein binding sites. Specifically these approaches reveal protein binding sites only in lower resolution, i.e., at an interval of hundreds to thousands of base pairs. Furthermore, they implicitly assume that there is only one "mode" or "predicted binding location" per this wide genomic interval. More recently, deconvolution algorithms such as Deconvolution (Lun et al., 2009[15]), GEM (Guo et al., 2012, an improved version of Guo et al., 2010) and PICS (Zhang et al., 2010) have been proposed to identify binding sites in higher resolution using ChIP-Seq data. However, most of them are still not tailored for ChIP-exo and PE and SE ChIP-Seq data in a unified framework and as a result, currently available methods are not appropriate for fair comparison between ChIP-exo and ChIP-Seq. To address these limitations, we developed and utilized an improved version of dPeak (Chung et al., 2013), a high resolution binding site identification (deconvolution) algorithm that we previously developed for PE and SE ChIP-Seq data, so that it can also handle ChIP-exo data. The dPeak algorithm implements a probabilistic model that accurately describes the ChIP-exo and ChIP-Seq data generation process.

Some of the key findings in this work are as follows. First, we demonstrate that the "peak pair" assumption of Rhee and Pugh, 2013 does not hold well in real ChIP-exo data. Second, we found that when analyzing ChIP-exo data and the Input control is not available, it is useful to adjust for GC content and mappability biases to improve peak calling and binding site identification. Third, we evaluated several methods to identify binding events and dPeak performs competitively respect to GEM and MACE when analyzing ChIP-exo data. Finally, when comparable number of reads is used for both ChIP-exo and ChIP-Seq , dPeak coupled with ChIP-exo data provides resolution comparable to PE ChIP-Seq and both significantly improve the resolution of protein binding identification compared to SE-based analysis with any of the available methods.

Results and discussion

Publicly available and novel datasets

We generated σ^{70} factor ChIP-exo, PE and SE ChIP-Seq experiments in *E. Coli* under aerobic ($+O_2$) and anaerobic ($-O_2$) conditions in glucose minimal media on the HiSeq2000 and Illumina GA IIx platforms. 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 where rifampicin was applied for 20 minutes and a control sample (without rifampicin being applied). We used these experimental designs for comparisons of ChIP-exo and PE ChIP-Seq assays of high resolution analysis and binding site identification. This comparison benefits from using σ^{70} for this study since is a transcription initiation factor of housekeeping genes in *E. Coli*. In this organism's genomes, many promoters contain multiple transcription start sites (TSS) and these TSS are often closely spaced ($10 \sim 150$ bp). These closely spaced binding sites are considered to be multiple "switches" that differentially regulate gene expression under diverse growth conditions [20].

Additionally, we gathered ChIP-exo data from diverse organisms: CTCF factor in human HeLa cell lines [1]; ER factor in human MCF-7 cell lines [5]; GR factor in IMR90, K562 and U2OS human cell lines [21]; TBP factor in human K562 cell lines [22]. Additionally, we also gathered the ChIP-nexus experiments provided by He et al., 2014: TBP in human K562 cell lines, MyC and Max in S2 *D. Melanogaster* cell lines and, Twist and Dorsal in *D. Melanogaster* embryo cell lines.

Comparison with ChIP-Seq data

We first compared various factors that could affect binding site identification between ChIP-exo and ChIP-Seq data by using the ChIP-exo reads from the first biological sample and first replicate grown under aerobic condition (first line from Table 1) and a SE ChIP-Seq replicate grown under the same conditions. In order to compare distribution of signal and background between ChIP-exo and ChIP-Seq data, we counted the number of extended reads mapping to a partition of the genome into non-overlapping bins. ChIP tag counts in ChIP-exo data were linearly related to ChIP tag counts in ChIP-Seq data for the regions with high ChIP tag counts (Upper part of Figure 2A). This implies that signals for potential binding sites are well reproducible between ChIP-exo and ChIP-Seq data. On the other hand, there was a clear difference in the background distribution between them (lower part of Figure 2A). Specifically, in ChIP-Seq data reads were almost uniformly distributed over background (non-binding) regions and the ChIP tag counts in there regions were significantly larger than zero. In contrast, in ChIP-exo data, there was larger variation in ChIP tag counts among background regions and ChIP tag counts were much lower in these regions compared to ChIP-Seq data. There were also large proportion of regions without any read in ChIP-exo data. These results indicate that for ChIP-exo data a much smaller portion of the genome is expected to be background. Therefore, methods that specifically model the background distribution must take into account that ChIP-exo's background is going to be composed by a smaller proportion of the reads in the experiment.

Using the same experiment, we next evaluated the "peak pair" assumption from Rhee and Pugh, 2011, i.e. a peak of reads in the forward strand is usually paired

with a peak of reads in the reverse strand that is located in the other site of the binding site. Note that currently available ChIP-exo data analysis methods, such as Wang et al., 2014, Madrigal 2015 and Bardet et al., 2013 rely on this assumption. In order to evaluate this assumption, we reviewed the proportion of reads in the forward strand in high quality ChIP-exo peaks such as at least one binding site is predicted in both ChIP-exo and ChIP-Seq data. We found that strands of reads were much less balanced in ChIP-exo data than in ChIP-Seq data in these regions with potential binding sites (Fig. 2B) and this indicates that the peak pair assumption might not hold in ChIP-exo data. This is caused by either the enzyme digestion or the strength at which the protein binds to the DNA: Although it is expected for the exonuclease enzyme to digest the ChIP fragments starting by their 5′ ends and stopping when finding a binding event, it may not be able to reach and digest every fragment in the ChIP sample. On the other hand, if the protein is not bound to both DNA strand, the enzyme may completely digest the fragment. Finally both effect are increased by the PCR amplification.

We evaluated ChIP-exo data for CTCF factor from human genome [1] to investigate issues specific to eukaryotic genomes for binding sites identification. Figures 2C and 2D 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. In Figure 2C it is shown that the ChIP-exo tag counts linearly increases with the mappability score and in Figure 2D it is shown that for GC - content below 0.6, the mean ChIP tag count increases and for GC - content greater than 0.6 it shows a decreasing trend. Benjamini and Speed, 2011 and Kuan et al., 2011 studied the presence of the mappability and GC - content biases in ChIP-Seq's background. It is not surprising to see these biases also present in ChIP-exo data, since ChIP-Seq and ChIP-exo signal seems to be linearly correlated for enriched regions (Figure 2A). Rozowsky et al., 2009 and Valouev et al., 2008 provide in depth analysis of the mappability and GC - content biases for ChIP-Seq respectively. Finally, these results indicate that binding site identification for ChIP-exo data benefits from using methods that take into account of apparent sequence biases such as mappability and GC content, mostly when an Input sample is not available.

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

We continued our exploration by investigating whether the current state-of-the-art QC pipelines for ChIP-Seq are suitable for ChIP-exo and ChIP-nexus. In Tables 1 and 2 we calculated a collection of the commonly used ChIP-Seq QC metrics using the ChIP-exo and ChIP-nexus experiments instead: Normalized Strand Cross-Correlation (NSC), Relative Strand Cross-Correlation and PCR Bottleneck Coefficient (PBC) defined as in `https://genome.ucsc.edu/ENCODE/qualityMetrics.html`.

DNA libraries generated by ChIP-exo and ChIP-nexus protocols are expected to be less complex than the libraries generated by ChIP-Seq, since the possible number of positions to which the reads can be aligned is being reduced due to the exonuclease digestion, considerable amounts of reads are being mapped to specific positions. This affects the interpretation of the PBC, since for ChIP-Seq low PBC

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 1 Current QC metrics applied to generated E. Coli σ^{70} samples. NSC stands for Normalized Strand Cross-Correlation, RSC stands for Relative Strand Cross-Correlation and PBC stands for PCR Bottleneck Coefficient.

Protocol	Organism	TF	Cell line	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 2 Current QC metrics applied to gathered data. NSC stands for Normalized Strand Cross-Correlation, RSC stands for Relative Strand Cross-Correlation and PBC stands for PCR Bottleneck Coefficient.

values indicate that the same read has been copied by the amplification process and aligned multiple times to the same position; while for ChIP-exo and ChIP-nexus when several reads are aligned to the same position are not necessarily the same read amplified, but several reads that their 5′ end was digested to the same position before the amplification step. It is of special importance to notice that for deeply sequenced ChIP-exo and ChIP-nexus experiments, the PBC values are quite low, which by following the ChIP-Seq guidelines it would indicate that those experiment show severe bottlenecking problems, which may incorrectly suggest that the positions with a large amount of aligned reads are being caused by PCR amplification rather than observed genomic signal.

The Strand Cross-Correlation (SCC) introduced by Kharchenko et al., 2008 is the most commonly used quality measure in ChIP-Seq. In general it measures how well the reads mapped to each strand are clustered around the locations where the proteins are binding to the DNA, and usually it is expected to observed two local maxima, one when the profiles are shifted by the average read length and another when the profiles are shifted by the unobserved fragment length. In a high quality ChIP-Seq dataset the last one is also the SCC global maxima. On the other hand, as a thought experiment in an idealized ChIP-exo experiment where the ChIP

fragments are digested until they found the binding protein and this protein bound to the whole motif sequence, then the SCC would a flat function with a jump at the motif length; since usually this is not the case, we also expect to observe a local maxima at the read length. Figure 3 shows the SCC curves for the CTCF factor from human genome: The ChIP-exo curve shows a 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 unobserved fragment length. Therefore, in ChIP-exo's cases both peaks are likely to be confounded. Hence measures based in the SCC such as the Normalized Strand Cross-Correlation (NSC) or the Relative Strand Cross-Correlation (RSC) are harder to interpret.

outline

1 ~~Available data, the motivation of this part is two fold:~~
   - ~~we want to list all data sets we obtained~~
   - ~~we want to emphasize that for the sig70 datasets we have a unique design that allows us to compare chip exo with SE and PE chipseq , the later one for the first time~~
2 ~~comparison with chip seq~~
3 ~~only in methods. list the align algorithm used~~
4 ~~chip seq qc metrics~~
5 chip exo qc pipeline, description and justification of the steps
   - need to justify the sampling step. for this we may use the depth table and add the factor to normalize one depth to another. explain the confounding of pcr amplification and the reduction of the number of possible mappable positions.
   - case study with FoxA1, analysis of enrichment vs library complexity and analysis and the strand imbalance tools. in methods we need to add which tests did we perform to show that the pipeline works. how do we expect from the diagnostics and how to interpret the plots.
   - analysis of why the pipeline works. in methods need to specify how and in results need to specify why it works, and what are the conclusions from it
6 high resolution analysis with chip exo and comparison with PE chip seq
   - justify what do we used dpeak (in methods need to list the other methods used). perhaps we can add gof plots in supplement. we also want to compare with the same method, therefore we can use the fact that dpeak have a unified model for PE and SE dpeak
   - after comparing with other methods we can compare with PE and SE ChIP-seq, this comparison gives also resolution and sensitivity
   - compare samples under same budget of reads

0.0.1 ChIP-exo Quality Control Pipeline

Figure 4 shows a flowchart for the ChIP-exo QC pipeline. In the first step, we partition the genome by keeping the non-digested ChIP-exo regions. Then, for each region it counts the number of fragments that compose the region and the number of positions to which the reads are being mapped to in each strand. With these values it calculates the following summary statistics:

$$\text{ARC} = \frac{\text{Nr. of reads in the region}}{\text{Width of the region}},$$
$$\text{URCR} = \frac{\text{Nr. of reads in the region mapped to exactly one position}}{\text{Nr. of reads in the region}},$$
$$\text{FSR} = \frac{\text{Nr. of fwd. strand reads in the region}}{\text{Number of reads in the region}}.$$

Finally it creates several visualizations designed to diagnose the quality level of a ChIP-exo sample. Figure 4A shows the typical behavior of the ARC vs. URCR plot. In general, the plot depicts two strong arms: One on the left with low ARC values and varying URCR which corresponds to ChIP-exo's background, regions that are usually composed by scattered reads that were not digested during the exonuclease step; and another one where the URCR decreases as the ARC increases, which corresponds to regions that are usually enriched and as the URCR decreases the library complexity does it as well, on the other hand high URCR values correspond to regions composed by position with relatively few fragments aligned to. Figures 4B and 4C analyze the strand imbalance bias in a ChIP-exo experiment: The first one depicts how quickly the regions exclusively formed by fragments in one strand are being filtered out as regions with higher depth are observed; and the second one shows how quickly the FSR's distribution approach the median, since in a high quality sample it is expected for the median to be approximately 0.5 and the enriched regions are going to be composed by fragments sequenced from both strands.

The ChIP-exo QC pipeline provides a model free framework to analyze the biases in a ChIP-exo experiment by taking advantage of the exonuclease enzyme that digests the non-enriched regions to partition the genome, calculates common ChIP-Seq QC metrics in ChIP-exo regions locally and allows the interpretation of these metrics by the use of diagnostic plots.

Enrichment and library complexity in ChIP-exo data

In ChIP-exo experiments, background fragments are often digested by the exonuclease enzyme, therefore the balance between the enrichment and library complexity of an experiment is a key factor determining the sample's data quality.

Using the Fox A1 in mouse liver cell lines from [5] and these two quantities, we explored the relationship between library complexity and experiment enrichment. In Figure 5A we present ARC vs. URCR plots for all three replicates. As a case of study, we compare the three plots to differentiate the quality of the three experiment. Hence, this might imply that the first replicate to have have more enriched regions and the third replicate's library complexity to be lower than the other two replicates library complexities. To verify this statement, we extracted the sequences around high confidence binding events and look for the FoxA1 motif using FIMO [27]. Figure 5B shows the number of candidate regions, which shows that the first replicate is being allocated into more enriched regions than the other ones. Figure 5C shows that for the first and third replicates, the FoxA1 motif is being detected in roughly the same proportion of sites, and finally in 5D we observe that the first and third replicate can detect the FoxA1 motif with the same significance, while the second replicate does not.

Strand imbalance in ChIP-exo data

The strand imbalance assessment is based in the observation that the enriched regions usually are composed of a higher quantity of reads, therefore we examined the FSR (defined as the ratio of number of forward stranded reads divided by the total number of reads in a given region) as the regions with lower depth are being filtered out. This indicator is of particular importance, as it evaluates the "peak pair" assumption that the original ChIP-exo paper suggested and multiple ChIP-exo data analysis methods rely on. For every ChIP-exo experiment, we calculated the global FSR and noticed that for all experiments is roughly 0.5, which means there are roughly the same amount of reads in both strands.

In order to asses the strand imbalance we created the visualization shown in Figure 6: Figure 6A presents the FSR's behavior as the lower depth regions are being filtered out, while Figure 6B) shows which percentage of the regions are composed by reads in both strand or only one (forward or backward). In a good data set, it would be expected that all quantiles shown to be quickly converging towards the median (in panel A) or the regions composed of reads in one strand being made of few fragments (in panel B). For each replicate, we divided the partitioned regions by asking whether they overlap a set of high quality ChIP-exo peaks, and then we tested (using the Wilcoxon rank sum test over the imbalance index defined in the methods section) if the strand imbalance's distribution is the same for both classes. For regions composed by a higher amount of reads, it is harder to distinguish their peaks by considering only the strand imbalance, hence in a better quality ChIP-exo experiment it is easier to distinguish enriched regions by the amount of reads in both strands. Similarly, we may consider that the strands of reads for background of a ChIP-exo experiment is more unbalanced than those for the enriched regions. In conclusion, Figure 6 shows that the global FSR does not represent the experiment's local strand imbalance, hence the "peak pair" assumption does not hold locally in ChIP-exo data.

Analysis of ChIP-nexus data with the QC pipeline

We applied the QC pipeline to the ChIP-nexus experiments and found that the overall pattern seen in the visualizations is similar to a high quality ChIP-exo experiment. Figure 7A illustrates a typical pattern of the ARC vs. URCR plot with ChIP-nexus data: Both arms seems to be more narrow and therefore the separation between them seems to be more noticeable, this seems reasonable since libraries generated by ChIP-nexus usually yields higher complexity than libraries generated by ChIP-exo. Additionally, while in ChIP-exo experiments is common to observe regions formed by undigested fragments in only one strand, we can observe that in ChIP-nexus experiment the occurrence of this type of regions is reduced. Figure 7B shows that in a ChIP-nexus experiment the FSR distribution approximates the median more quickly than in a ChIP-exo experiment and Figure 7C we observe that most of these single-stranded regions are formed by few fragments, which means that good ChIP-nexus experiment are less likely to show very imbalanced regions.

0.0.2 Evaluating ChIP-exo and ChIP-nexus data

We applied the QC pipeline to all ChIP-exo and ChIP-nexus datasets listed in Tables 2 and **??**, that way we partitioned the reads of each experiment into the

undigested regions by keeping the regions formed by at least one fragment. Figure 8A shows compares the slope of the following model:

$$\text{depth} = \beta \times (\#\text{ of unique positions}) + \epsilon$$

when sampling 1,000 regions and repeating this process 10,000 times for every ChIP-exo and ChIP-nexus experiment. The slope can be interpreted as the change in depth as number of unique position varies. This parameter measures the library complexity on a ChIP-exo experiment, since when it is high it means that several fragments are being aligned to the same position in the genome.

Figures 8B and 8C shows the parameters of the following model:

$$\text{URCR} = \frac{\kappa}{\text{ARC}} + \eta + \epsilon$$

considering only the regions composed by 10 or more aligned reads, that way the model reflects the enriched arm of the ARC vs. URCR visualization. The intercept η can be interpreted as the theoretical URCR for relatively high depth regions and the slope κ can be interpreted as the decay in URCR as the ARC increases. That way experiments with low complexity are showing low values for both parameters. We considered a linear model between the estimated parameters for each experiment and it's respective depth for which we were able to asses that there is not significant relationship between the estimated parameters and the interaction between depth and organism for each experiment. Hence, these parameters may be used as indicators of the overall quality of ChIP-exo experiment, but as with the SCC it is recommendable to also visualize the SCC curve.

0.1 High Resolution Binding Site Identification with dPeak and ChIP-exo

0.1.1 Comparison with ChIP-Seq data using dPeak

Figure 9 shows comparisons among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. We considered the RegulonDB data as ground truth, since those are the most recent annotation on *Escherichia Coli*. A RegulonDB annotation (Salgado et al, 2012 [20]) was considered to be identified if the distance from the closest dPeak binding site estimate was less that or equal to 20 bp. That way, the sensitivity is defined as the proportion of RegulonDB annotations identified in a peak and the resolution is defined as the minimum distance between a RegulonDB annotation and the closest dPeak binding site estimate. Figure 9A shows that the sensitivity increases as the mean distance between binding events increases. When the average distance is greater than 200 bp, dPeak identifies more than the 75% of the binding events in each peak, this is intuitive as the mean ChIP-Seq's fragment length is shorter than 200 bp, hence the read does not contribute to more than one binding event [19]. When the binding events in a peak are closer to each other, both ChIP-exo and PE ChIP-Seq are comparable, as the distance increases ChIP-exo identifies a higher proportion of the RegulonDB annotations; additionally SE ChIP-Seq is significantly less sensitive than both ChIP-exo and PE ChIP-Seq. Figure 9B shows that ChIP-exo and PE ChIP-Seq are comparable in resolution, while both protocols significantly outperform SE ChIP-Seq.

0.1.2 Systematic comparison of ChIP-Seq vs ChIP-exo under varying sequencing depth

Previously, ChIP-exo and SE ChIP-Seq have been compared only at a fixed depth level in the literature, while they did not include PE ChIP-Seq as well either. In order to address this limitation in previous studies, we sampled a fixed amount of reads for each of the ChIP-exo, PE ChIP-Seq and SE ChIP-Seq datasets of the σ^{70} samples (N reads for both ChIP-exo experiment and $N/2$ or N pairs for PE ChIP-Seq to assume more realistic situation of a fixed cost). For each sampled dataset we applied our lower-to-higher resolution pipeline by calling peaks with MOSAiCS [14] and then deconvolving the binding events by using dPeak [19]. For the ChIP-exo datasets we called peaks by using the GC-content and mappability models with MOSAiCS, since it's background is usually composed of scattered reads across the genome; and for the ChIP-Seq datasets we used their respective Input samples. Additionally, it is worth noting that for PE ChIP-Seq we sampled both ends of the fragment, hence for each sequencing depth we are sampling the half amount of pairs for PE ChIP-Seq than for ChIP-exo or SE ChIP-Seq.

Figures 10 shows the behavior of each data type in σ^{70} experiment under aerobic condition when comparable number of reads is used for all of ChIP-exo, SE ChIP-Seq and PE ChIP-Seq. In Figure 10A we show the number of candidate regions defined as the number of regions where a binding event was identified in a collection of high quality peaks; in Figure 10B we depict the number of binding events; in Figure 10C we show the number of identified targets, where a RegulonDB was considered identified if a binding event was identified in a 15 bp vicinity of it; and finally Figure 10D show the resolution defined as the distance from a RegulonDB annotation to the closest dPeak prediction. It is remarkable that even when the number of candidate peaks or the number of predicted events is lower for ChIP-exo, it outperforms both PE and SE ChIP-Seq in number of identified targets and resolution.

This may suggest that with ChIP-exo less false positive peaks are being called and that when the targets are being identified, dPeak estimates binding locations closer to the true location. Additionally, we can see that as the read depth increases, all four indicator seem to stabilize and hit a plateau earlier than the cases for ChIP-Seq, which may indicate that with ChIP-exo a smaller amount of reads is needed to identify the same number of targets than ChIP-Seq, but it may be also possible that this is an artifact occurring due to ChIP-exo's lower library complexity.

Figure S3 shows an analogous analysis but using the σ^{70} replicates with and without rif treatment. The left, middle and right columns show the fixed depth against the number of predicted events, identified targets and resolution being compared at a fixed depth level. The behavior of this quantities seems to be opposite to the one as in Figure 10, hence we used the ChIP-exo QC pipeline in the fixed depth ChIP-exo experiments. Figure 11 shows scatter plots of ARC vs URCR for several fixed sample sizes, as the fixed depth increases the two arm pattern becomes more distinctive while for lower depth, it seems that the majority of the sampled reads were aligned to enriched regions. On the other hand in Figures S4 to S7, we used the ChIP-exo QC pipeline on the samples that are outperformed by PE and SE ChIP-Seq. For a fixed low depth, we can see that the majority of the reads are

being aligned to non-enriched regions since the vertical arm seems to be stronger for all 2 conditions and replicates; for higher depth we can see low URCR values being predominant, which indicates that the majority of the regions being formed by few positions with a higher read concentration. In low complexity regions, the reads are being aligned to fewer positions but there is no control over the amount of reads mapped. Hence, those regions are more likely to being strand-imbalance which in turn may bias the binding site estimate and therefore decrease the number of identified targets or increase an experiments resolution.

0.1.3 dPeak outperforms competing methods in discovering closely spaced binding events from ChIP-exo and ChIP-Seq data

Figure 12 compares the resolution defined as the minimum distance between a RegulonDB annotation and a binding site predicted by either Peakzilla [9], MACE [7], GEM [16] or dPeak [19]. In a good dataset such as both of the ChIP-exo experiments under aerobic (panel 931 and 933) conditions, all the methods are comparable in resolution, and dPeak slightly outperforms the rest. On the other hand, for experiments dPeak identified binding sites with higher resolution than Mace and Peakzilla, but with low resolution than Gem. This may be due that the fact that Gem uses sequence information in addition to the aligned 5′ end counts that dPeak uses. [1]

Conclusions

We made a systematic exploration of several ChIP-exo experiments. We provided a list of factors that reflect the quality of a ChIP-exo experiment and we developed a QC pipeline which is capable of assessing the balance between the enrichment and the library complexity of a ChIP-exo experiment. Additionally, a set of diagnostics was established to assess the quality of a ChIP-exo experiment. While the QC pipeline only requires a set of aligned reads to give a global overview of a ChIP-exo experiment, this overview coincides with more elaborate analysis that is computationally more expensive to perform or requires additional inputs that may not be available, such as motif detection in a set of high quality regions or resolution analysis given a set of annotations as gold-standard.

We studied the shared biases between ChIP-exo and ChIP-Seq data, and noticed that for eukaryotic genomes the relationship between ChIP-Seq data and either the mappability or the GC content scores are still present in ChIP-exo. We also examined ChIP-exo's background and noticed that is significantly different from the ChIP-Seq one, since it consists of only a small quantity of fragments that was not digested by the exonuclease enzyme. Additionally, we showed that we have unbalanced number of reads in forward and reverse strands, and that in a lower quality ChIP-exo experiment those regions are going to be harder to differentiate from the possibly enriched regions.

To the extent of our knowledge, we made for the fist time a comparison between ChIP-exo and PE ChIP-Seq. Using a set of annotations as gold-standard, we showed that both protocols are comparable in resolution and that for regions with more than one binding site, ChIP-exo is more sensitive than both SE and PE ChIP-Seq.

[1] For here we may probably use only 933 as part of the main article and keep the rest for the supplement

We made a rigorous comparison between fixed depth ChIP-exo, PE ChIP-Seq and SE ChIP-Seq, and we probed that for sufficiently complex libraries, ChIP-exo experiments can outperform PE and SE ChIP-Seq in number of identified targets and resolution. The proposed ChIP-exo QC pipeline provides a rigorous, easily interpretable, computationally efficient framework to diagnose if the library complexity of a ChIP-exo experiment is adequate.

Methods

Considered data.

Growth conditions of generated data.

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

need to add the growth conditions

ChIP-exo and ChIP-nexus experiments.

We gathered a collection of ChIP-exo experiments spanning different cell lines and transcription factors: CTCF in HeLA cells [1], ER in MCF-7 cells [5], TBP in K562 cells [22], GR in IMR90, K562 and U2OS in K562 cells [21] and FoxA1 in mouse liver cells [5]. Additionally we used the ChIP-nexus experiments generated for He et al., 2015: TBP in human K562, MyC and Max factors in S2 cell lines in *D. Melanogaster*; Twist and Dorsal factors in *D. Melanogaster* embryo cell lines.

The read files were aligned following the instructions provided in their respective sources when available. Otherwise we used *bowtie* in its 1.1.2 version.

Comparison with ChIP-Seq data

Using the σ^{70} sample grown under aerobic condition for ChIP-exo and SE ChIP-Seq. We partitioned the *E.Coli* genome into non-overlaping bins of length 150, and for each bin we counted how many ChIP-exo and SE ChIP-Seq extended reads overlapped each bin. The reads were extended by 150 bp to their $3\prime$ directions.

To model the strand imbalance, we called a set of high quality peaks for ChIP-exo and SE ChIP-Seq (with GC content + Mappability and Input only models respectively), such that each peak contained at least one binding event (found by using **dPeak** with at most 5 binding events in each peak). For each peak, we calculated the FSR as:

$$\text{FSR} = \frac{\text{Nr. of fwd. strand reads mapped to the region}}{\text{Total nr. of reads mapped to the region}}$$

We estimated the FSR's densities by using R's *density* function using its default parameters, which computes gaussian kernel density estimates.

Definition of current ChIP-Seq QC guidelines.

We considered the definitions of ChIP-Seq QC guidelines provided in Landt et al., 2012 which are also listed in `https://genome.ucsc.edu/ENCODE/qualityMetrics.html`.

Strand Cross-Correlation.
The strand cross-correlation was proposed by Kharchenko et al., 2008 [26] and it may be one of the most used of the ChIP-Seq QC metrics. The SCC curve is defined as:

$$y(\delta) = \sum_c w_c r \left[n_c^+ \left(x + \frac{\delta}{2} \right), n_c^- \left(x - \frac{\delta}{2} \right) \right], \tag{1}$$

where $y(\delta)$ is the SCC for a strand shift δ, r is the Pearson correlation, w_c is the proportion of reads mapped to chromosome c and n_c^S is the read count vector for strand S and chromosome c. The following QC metrics are used to summarize the shape of the SCC curve:

$$\text{NSC} = \frac{\max_\delta y(\delta)}{\min_\delta y(\delta)}, \tag{2}$$

$$\text{RSC} = \frac{\max_\delta y(\delta) - y_{\text{bgd}}}{y_{\text{rl}} - y_{\text{bgd}}}. \tag{3}$$

where y_{bgd} is estimated as the background SCC level is defined as $y_{\text{bgd}} = \min_\delta y(\delta)$ and y_{rl} is the value of the SCC curve when the shift equals the experiment's read length. It is worth noticing that in Landt et al., 2012 the RSC is defined as:

$$\text{RSC} = \frac{\max_\delta y(\delta)}{y_{\text{rl}}}.$$

In ChIP-exo data, both definitions are equivalent since in a typical ChIP-exo experiment y_{bgd} is approximately zero.

PCR Bottleneck Coefficient.
The PCR Bottleneck coefficient is a measure of library complexity in ChIP-Seq data:

$$\text{PBC} = \frac{\text{Nr. of positions to which exactly one unique mapping read is aligned}}{\text{Nr. of positions to which at least one unique mapping read is aligned}}$$

For human and mouse genome, the ENCODE project states that a PBC value in the 0 - 0.5 range indicates severe bottlenecking, in the 0.5 - 0.8 range moderate bottlenecking, in the 0.8 - 0.9 range indicates mild bottlenecking and in the 0.9 - 1 range indicates that there is no presence of bottlenecking.

These QC guidelines were calculated with the **ChIPUtils** package in its 0.99.0 version (available in `https://github.com/welch16/ChIPUtils`). This package provides an easy to use interface to calculate basic quality control metrics and diagnostic plots for ChIP-Seq data.

ChIP-exo quality control pipeline.

The ChIP-exo QC pipeline requires a set of aligned reads from a ChIP-exo (or ChIP-nexus) experiment. Then:

1 Partitions the experiment's coverage by keeping the regions formed by one or more aligned reads.
2 For each region, it counts the number of reads that are being aligned to the region and the number of positions to which at least one mapping read is aligned in each strand.
3 For each region, it calculates the following statistics:

$$\text{ARC} = \frac{\text{Total nr. of reads mapped to the region}}{\text{Width of the region}}$$
$$\text{URCR} = \frac{\text{Total nr. of position that at least one read is aligned in the region}}{\text{Total nr. of reads mapped to the region}}$$
$$\text{FSR} = \frac{\text{Nr. of fwd. strand reads mapped to the region}}{\text{Total nr. of reads mapped to the region}}$$

4 The ChIP-exo QC pipeline creates visualizations to summarize the summary statistics and calculates the contribution of the number of unique position and the width to the number of reads per region by randomly sampling M regions to fit the model:

$$\text{depth} = \beta_1 \text{npos} + \beta_2 \text{width} + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$, β_1 and β_2 are the contribution from the number of unique positions and width to the total number of reads per regions respectively; this process is repeated independently N times.

We implemented this ChIP exo QC pipeline in the *ChIPexoQual* R package, additionally it generates visualizations as seen in Figure 4. We used its 1.0 version which is available in `https://github.com/welch16/ChIPexoqual`.

Motif analysis of FoxA1 enriched regions

For each replicate, we used the ChIP-exo QC pipeline to partition the mouse genome into a set of regions with their respective summary statistics; we filtered them into collections of high quality regions by:

1 Removing the regions formed by reads on only one strand.
2 Removing regions with reads aligned to at most 15 unique positions.
3 Removing regions with less that 100 reads.

Then we used **FIMO** on its 4.9.1 version [27] to find the FoxA1 motif (with id MA0148.1 in the core JASPAR database).

Imbalance test of FoxA1 replicates

For each replicate, we used the ChIP-exo QC pipeline to partition the mouse genome into a set of regions with their respective summary statistics. We filtered the regions with reads in only one strand and we transformed the FSR into an *Imbalance index* that is zero when the region is perfectly balanced 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 onto two classes by considering whether it overlaps a ChIP-exo peak of not. For each replicate we called peaks with the GC content + Mappability model from MOSAiCS [14] (version 2.9.7) with bin size and fragment length of 200 bp. We called peaks with an FDR of 5%, a threshold of 100 and a maximum gap size of 200 bp. We further filtered the peaks by keeping only the peaks with an average ChIP count of 200 extended reads.

To show that the class that don't overlap with peaks exhibits heavier tails, we used a Wilcoxon test over the *Imbalance index*.

GC - content and Mappability

To define the mappability score we follow the definition from Rozowsky et al., 2009:

$$m_i = \sum_{k=i-L+1}^{i+L-1} \frac{\delta_k}{2L-1}.$$

where δ_i is the indicator if the base at coordinate i can be mapped uniquely by a 32 bp sequence at position i, and L is the expected fragment length. GC - content score is defined analogously, where δ_i represents the occurrence of a G or C at the i-th position in the genome.

The mappability and GC - content scores for a bin are defined as the average of the scores across the nucleotides in the bin.

High resolution analysis with ChIP-exo

We considered RegulonDB [20] annotations as gold-standard and considered an annotation as being identified if the distance to a estimated binding events is less or equal than 20 bp. We defined the resolution as the distance from an annotation to its closest predictions and the sensitivity as the fraction of identified annotations in a genomic region.

Method comparison for ChIP-exo.

We compared dPeak, Chung et al., 2013; GEM, Guo et al., 2012; MACE, Wang et al., 2014 and Peazilla, Bardet et al., 2013 for the ChIP-exo data analysis. For the dPeak algorithm we used the R package **dPeak** version 2.0.1 which is available from `https://github.com/dongjunchung/dpeak`. For the GEM algorithm, we used it's Java implementation version 2.6 which is available from `http://groups.csail.mit.edu/cgs/gem/`. For the Mace algorithm, we used it Python implementation version 1.2, which is available from `http://dldcc-web.brc.bcm.edu/lilab/MACE/docs/html/`. For the Peakzilla algorithm, we used the version available in `https://github.com/steinmann/peakzilla`. Candidate regions for **dPeak** and **GEM** were identified for each replicate of ChIP-exo data using the **MOSAiCS** algorithm [14] (one sample analysis using false discovery rate of 0.01%) implemented as an

R package **mosaics** version 2.9.7 (available from *bioconductor*). We further filtered out candidate regions by using the 300 peaks with higher average ChIP tag count to avoid potential false positive 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. We downloaded **CexoR** [8] on its 1.8 version from *bioconductor* but were unable to use it for the σ^{70} experiments.

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 model for ChIP-exo and the Input only model for PE and SE ChIP-Seq). Then, we deconvolved the peak into binding events with dPeak (version 2.0.1) by considering a maximum of 5 binding sites on each peak. To avoid false positives we only considered ChIP-exo peaks with average ChIP counts greater than 3000 that overlapped both the SE and PE ChIP-Seq peaks, we considered other cutoff values but still obtained results similar to what we presented in this paper.

Saturation analysis of ChIP-exo, PE ChIP-Seq and SE ChIP-Seq.
We sub-sampled N fragments for both ChIP-exo and SE ChIP-Seq protocols. For PE ChIP-Seq we sub-sampled N pairs or $N/2$ fragments. For each seed, we called peaks using MOSAiCS [14] (GC content+Mappability for ChIP-exo and Input only for SE and PE ChIP-Seq) for the maximum sample size and to avoid false positives we considered only the top 500 peaks for each data protocol. We defined the number of candidate regions as the number of top sample peaks such that a binding events was estimated using the sampled reads and the dPeak's model; the number of predicted events is the total quantity of binding events estimated using the dPeak's model; the number of identified targets are number of gold-standard annotations within 15 bp from an estimated binding events; and the resolution is defined as the minimum distance from a gold-standard annotation to an estimated binding event. We repeated this analysis for ten seeds and reported the median between all those values.

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. 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;.
5. 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;.

6. Rhee HS, Pugh F. ChIP-exo A method to identify genomic location of DNA-binding proteins at near single nucleotide accuracy. Current Protocols in Molecular Biology. 2012;.
7. 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;.
8. Madrigal P. CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates. EMBnetjournal. 2015;.
9. 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;.
10. Wilbanks E, Facciotti M. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. PLOS One. 2012;.
11. Pepke S, Wold B, Ali M. Computation for ChIP-seq and RNA-seq studies. Nature. 2009;.
12. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson D, Bernstein B, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008;.
13. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-Seq data. Nature biotechnology. 2008;.
14. 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;.
15. Lun DS, Sherrid A, Weined B, Sherman DR, Galagan JE. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-Seq data. Genome Biology. 2009;.
16. 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;.
17. Guo Y, Papachristoudis G, Altshuler RC, Gerber GK, Jaakkola TS, Gifford DK, et al. Discovering homotypic binding events at high spatial resolution. Bioinformatics. 2010;.
18. Zhang X, Robertson G, Krzwinski M, Ning K, Droit A, Jones S, et al. PICS: Probabilistic Inference for ChIP-Seq. Biometrics. 2010;.
19. 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;.
20. 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;.
21. 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;.
22. Venters BJ, Pugh F. Genomic organization of human transcription initiation complexes. Nature. 2013;.
23. Benjamin Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research. 2011;.
24. Rozowsky J, Euskirchen G, Auerbach R, Zhang Z, Gibson T, Bjornson R, et al. PeakSeq enables systematic scoring of ChIP-Seq experiments relative to controls. Nature, Biotechnology. 2009;.
25. 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;.
26. Kharchenko P, Tolstorukov M, Park P. Design and analysis of ChIP-Seq experiments for DNA-binding proteins; 2008.
27. Grant C, Bailey T, Noble WS. FIMO: Scanning for occurrences of a given motif;.
28. Mendenhall EM, Bernstein BE. DNA-protein interactions in high definition. Genome Biology. 2012;.
29. 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;.
30. 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;.
31. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature Reviews: Genetics. 2012;.

1 Figures

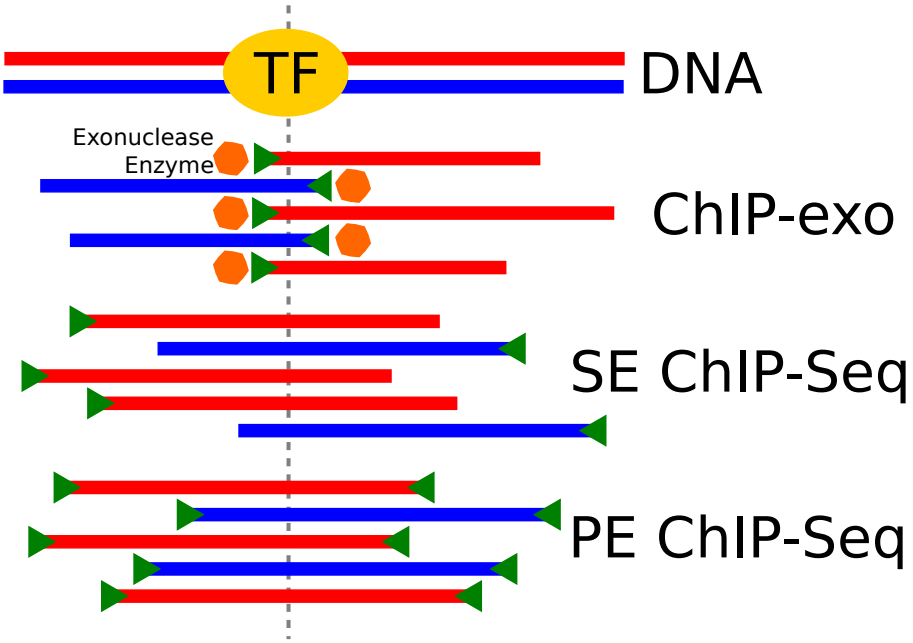


Figure 1 A TF is bound to the forward (red) and backward (blue) strand of the DNA. Then, is sonicated: For ChIP-exo a exonuclease enzyme (orange hexagon) trims the $5\prime$ ends of each DNA fragment to a fixed distance from the bound protein, finally is subjected to Immunoprecipitation and amplification.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

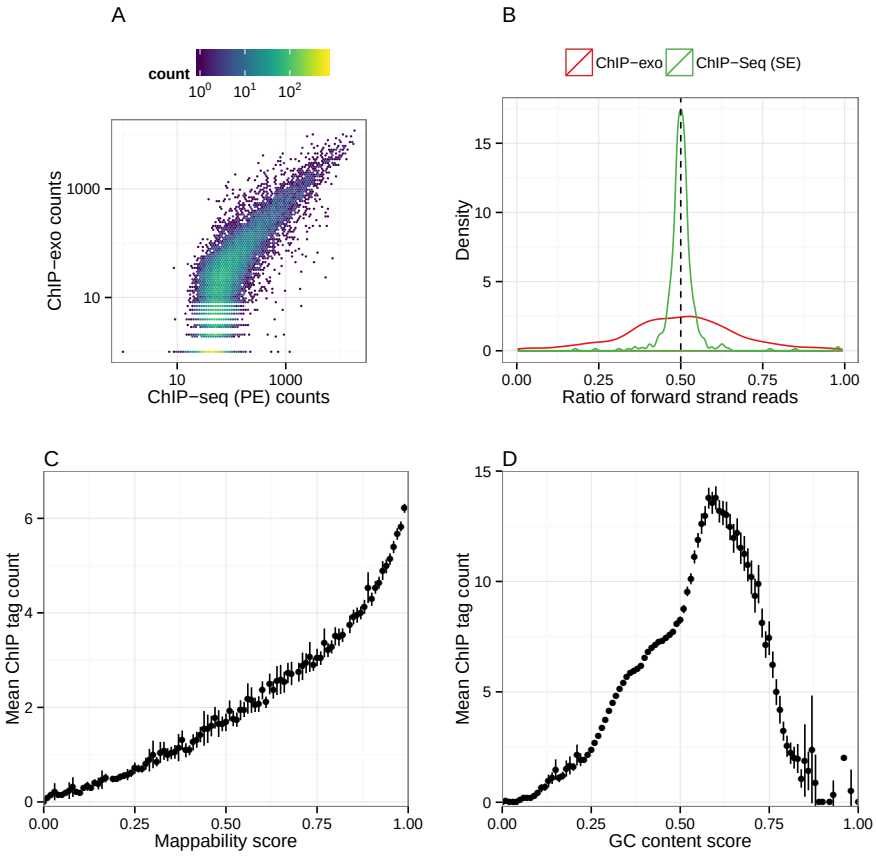


Figure 2 A) Hexbin plot of PE ChIP-Seq bin counts vs ChIP-exo bin counts. B) Forward Strand Ratio densities for SE ChIP-Seq and ChIP-exo peaks. C) Mappability score vs mean ChIP tag counts with 0.95 confidence bands. D) GC - content score vs mean ChIP tag counts with 0.95 confidence bands.

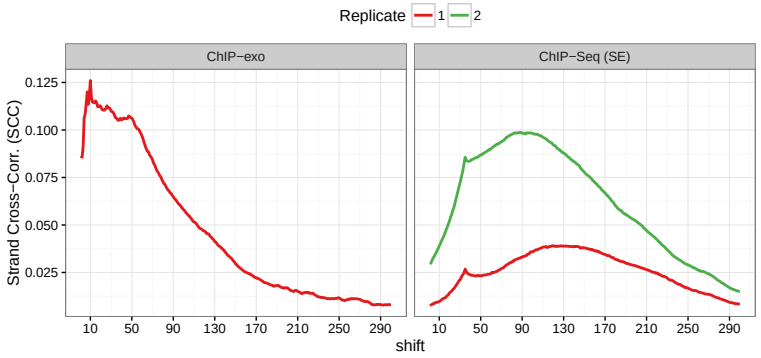


Figure 3 SCC curves for human CTCF on HeLa cell lines. The SCC curve for the ChIP-exo sample from [1] is shown in the left panel, and the SCC for ChIP-Seq samples from **https://www.encodeproject.org/experiments/ENCSR000AOA/** are shown in the right panel. The ChIP-exo curve shows local maxima at the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

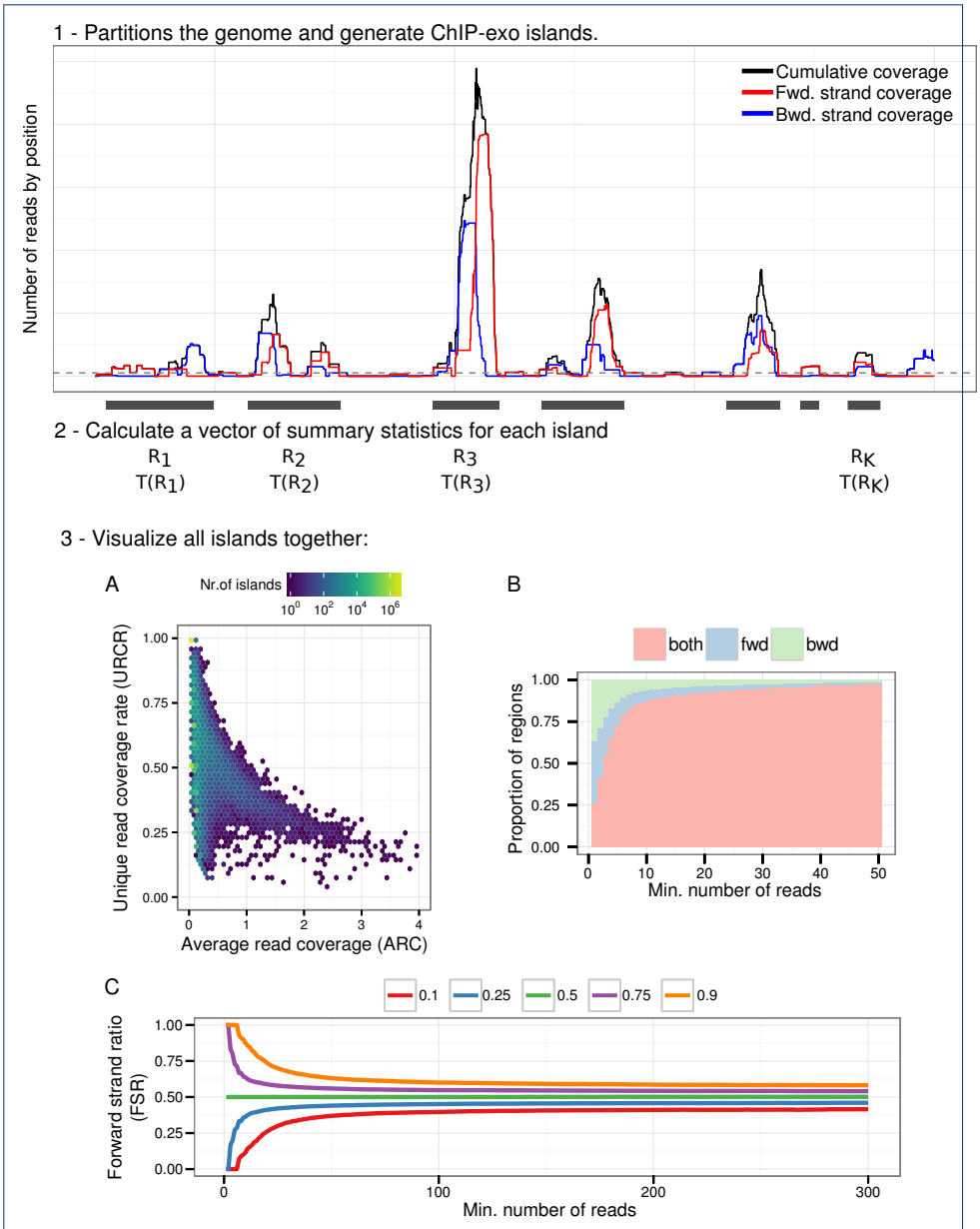


Figure 4 The ChIP-exo reads are partitioned by keeping the regions in the genome formed by the undigested fragments. For each region, we calculate a collection of summary statistics, which are visualized as A) ARC vs. URCR, this plot presents a global view of the balance between library complexity and 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.

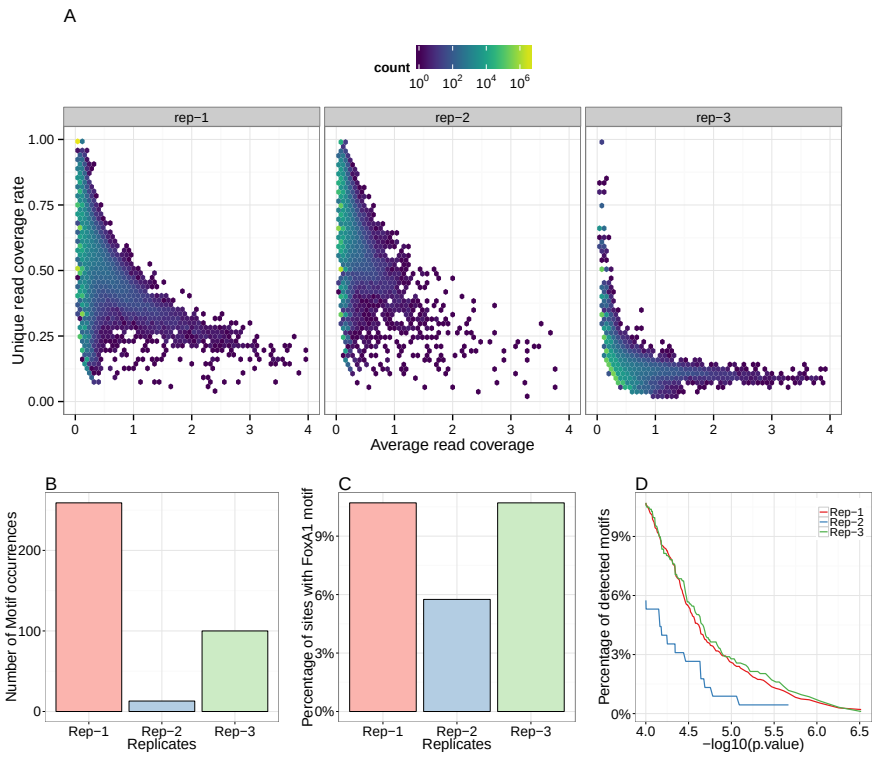


Figure 5 Using the mouse FoxA1 experiment from [5]: A) Hexbin plots of ARC against URCR, there is a slight separation into two strong arms, one corresponds to low ARC and varying URCR, and for the other URCR decreases as ARC increases. B) Number of candidate sites for each replicate. C) Percentage of candidate sites where the FoxA1 motif was detected. D) Cumulative proportion of detected motifs by replicate.

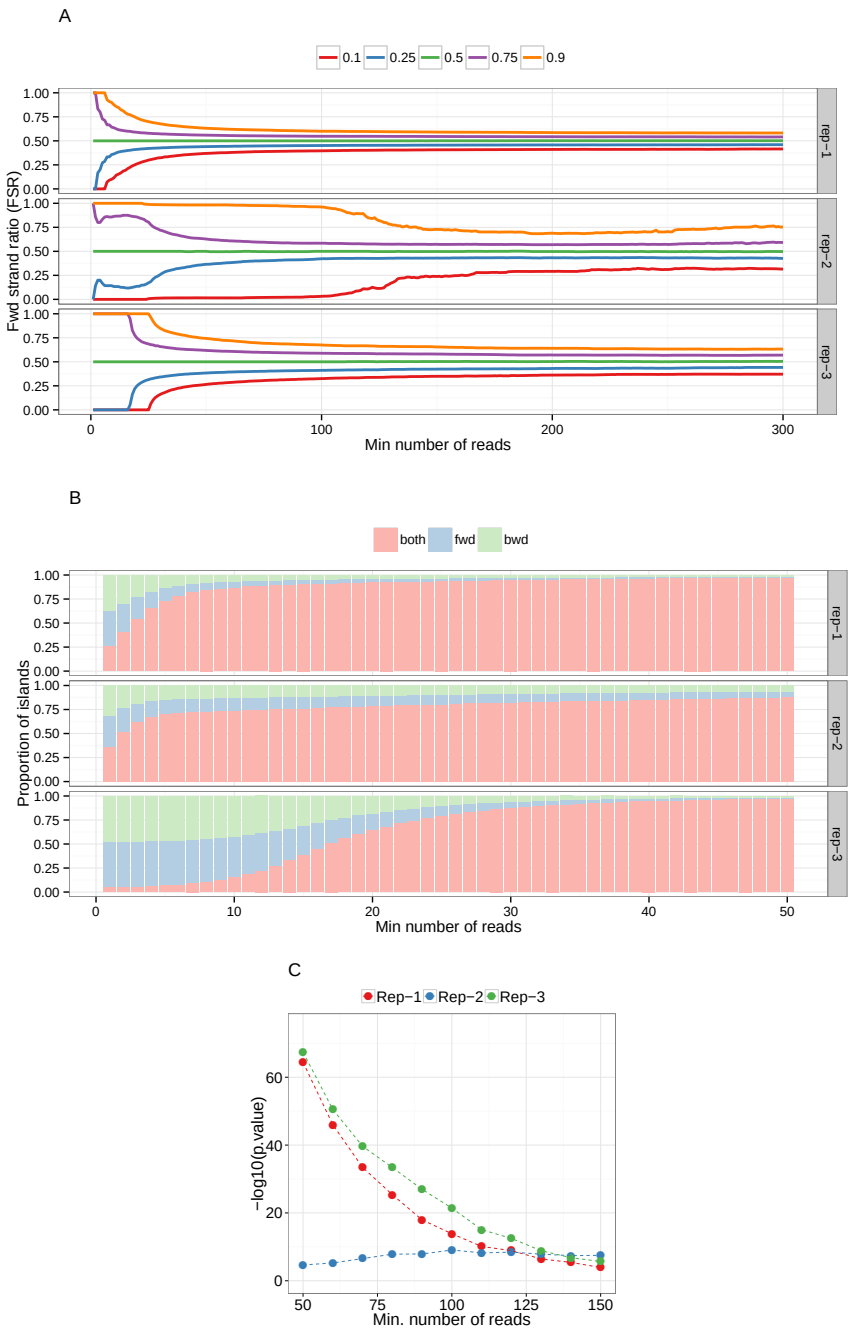


Figure 6 Strand imbalance QC plots for the same data as in Figure 5A. A) FSR distribution quantiles as the lower depth regions are being filtered out, all quantiles approach to the median as the lower bound increases. B) Stacked histogram with the proportion of regions that are formed by two strands or only one, in a good sample the single-stranded regions are going to be filtered out quickly as in the middle row. C) $-\log_{10}(\text{p.value})$ of testing if the imbalance distributions differs when ChIP-exo regions overlap their peaks.

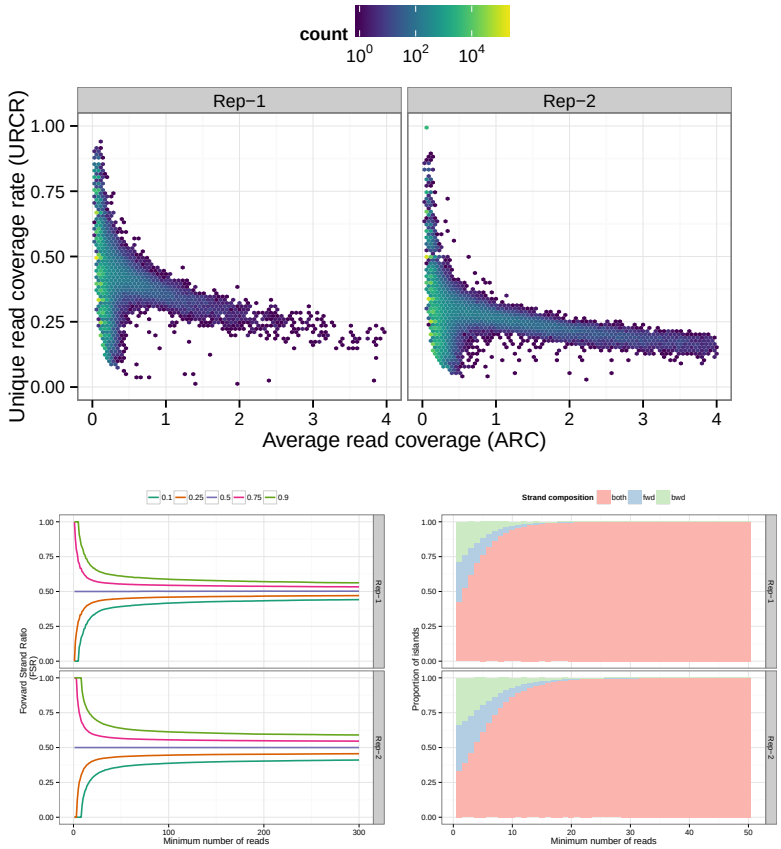


Figure 7 We applied the QC pipeline on ChIP-nexus data. A) ARC vs. URCR hexbin plots, B) FSR distribution quantiles and C) Stacked histogram with the proportion of regions composed by fragments on a single strand for the MyC experiment in S2 cell lines in *D. Melanogaster* from He et al., 2015 [2]

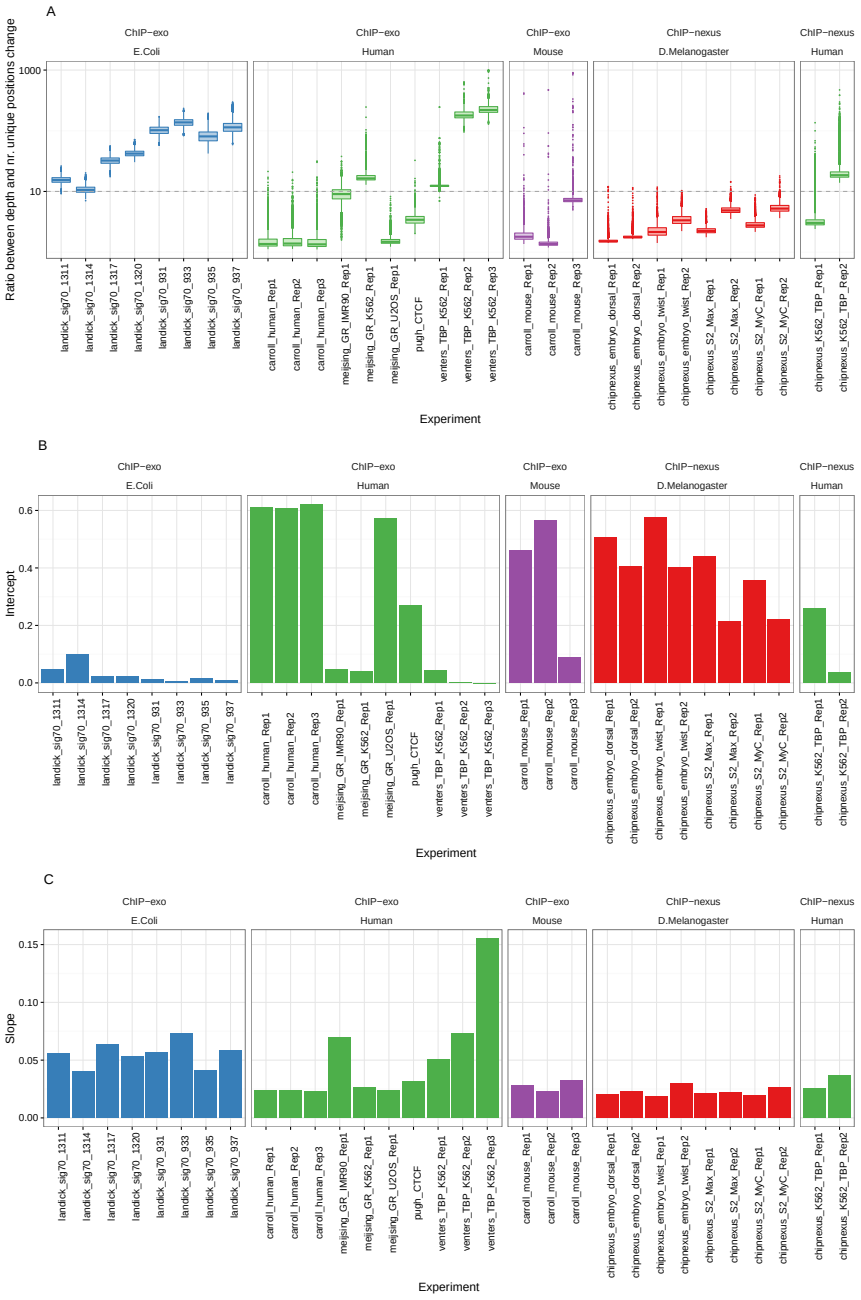


Figure 8 A) Comparison of change in depth as the number of unique position obtained by sampling 1,000 regions and fitting the model depth $= \beta \times$ (# unique positions), and repeating this process 10,000 times. B) Intercept and C) Slope in the the URCR $= k/\text{ARC} + b$ model for the regions formed by more than 10 aligned positions.

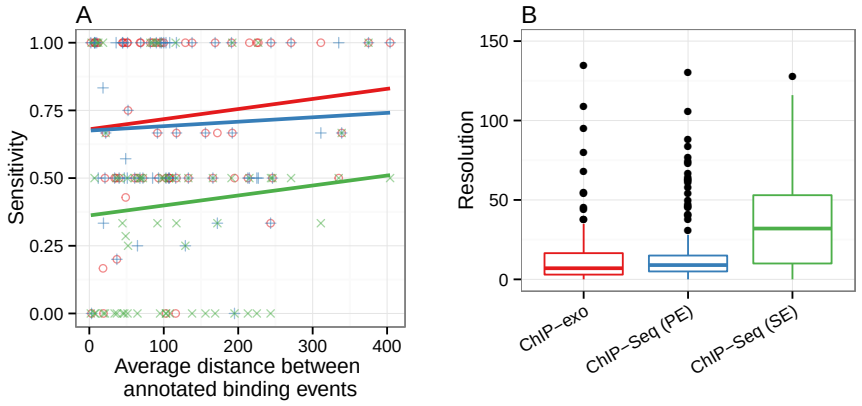


Figure 9 Comparison of A) sensitivity and B) resolution between ChIP-exo and ChIP-Seq data. Sensitivity is defined as the proportion of RegulonDB annotations identified using each data. Resolution is defined as the distance between RegulonDB annotation and its closest prediction.

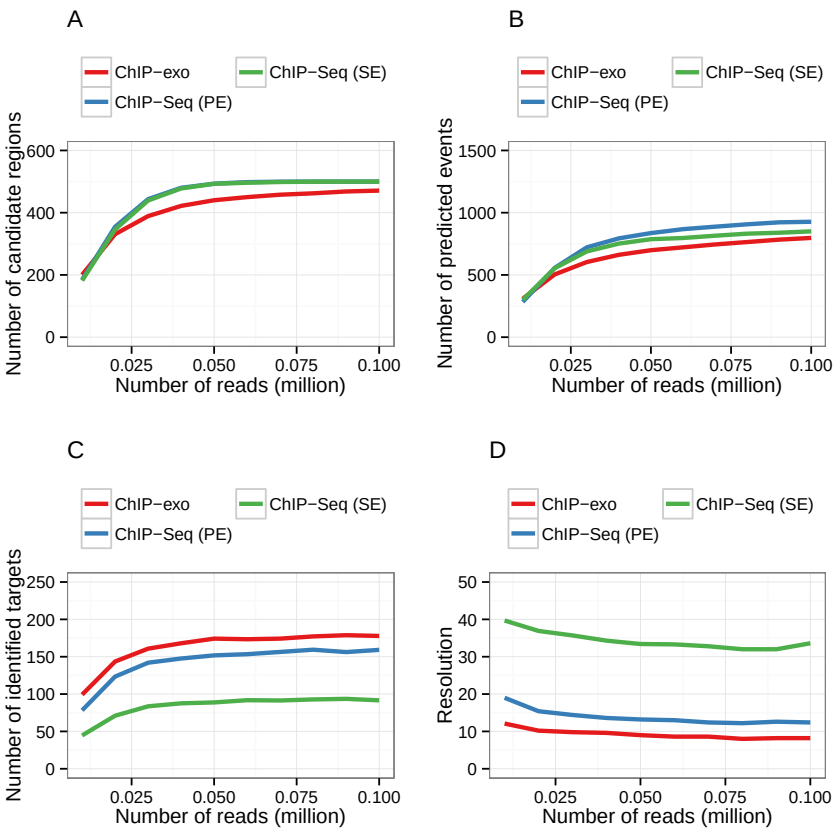


Figure 10 Comparison of the number of A) candidate regions, B) predicted events, C) identified targets and D) resolution among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. RegulonDB annotations are considered as a gold standard. A gold standard binding events was deemed identified if a binding event was estimated at a $\pm$ 15 vicinity of it.

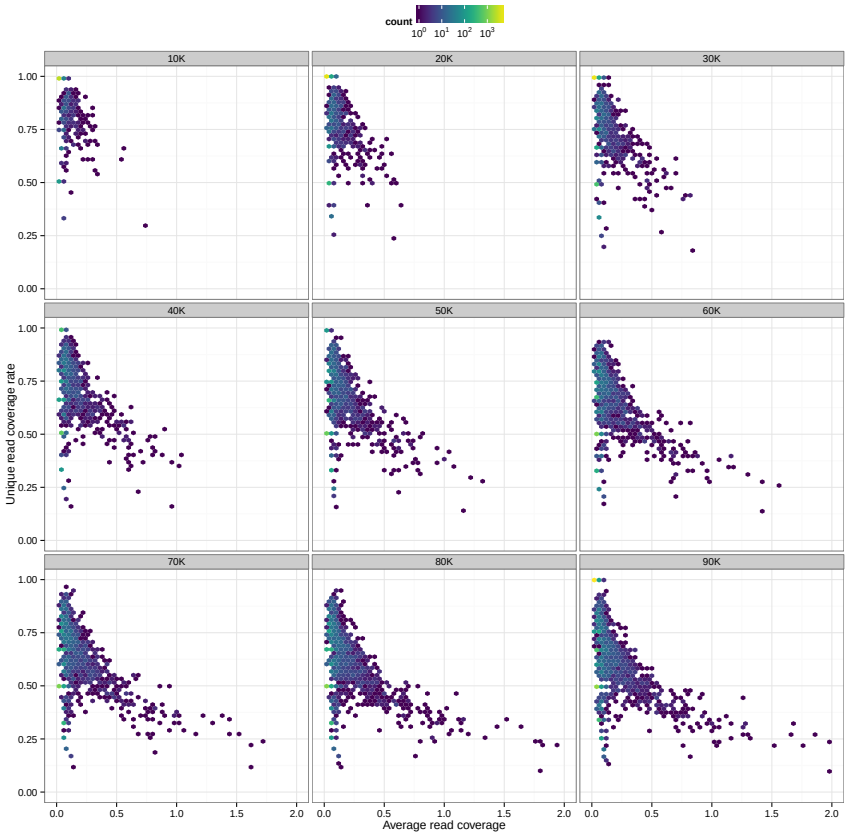


Figure 11 Hexbin plots of ARC vs URCR of the σ^{70} ChIP-exo experiment under aerobic condition when 10K to 90K reads are being sampled.

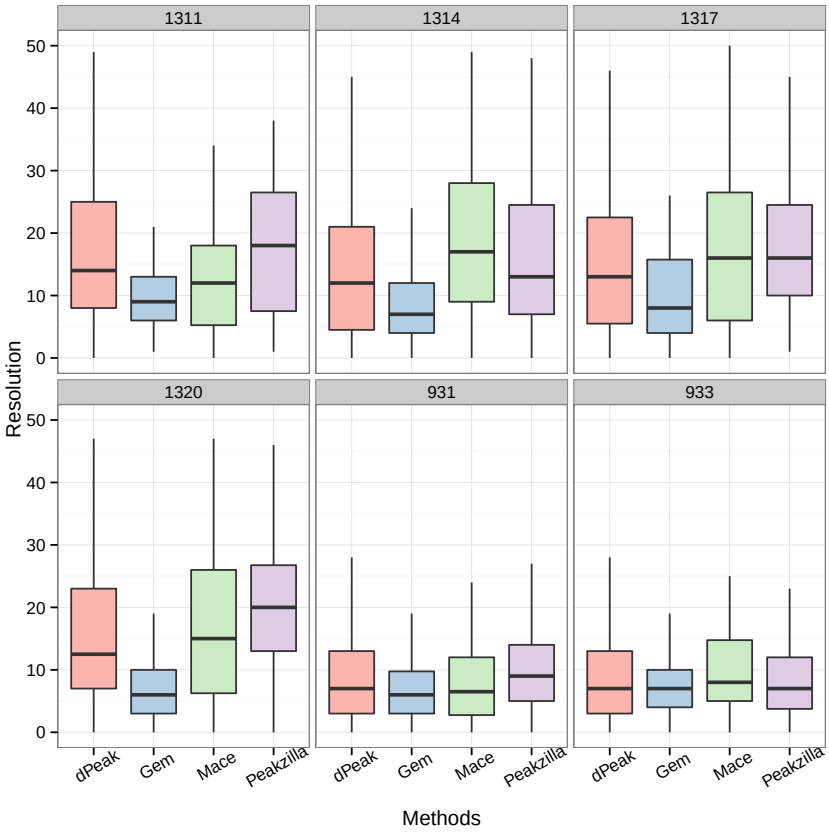


Figure 12 Comparison of the resolution between dPeak, Gem, Mace and Peakzilla methods. Resolution is defined as the minimum distance between a RegulonDB annotation and a predicted binding event.

Supplement

Note: I am not sure whether to include several figures on this supplement

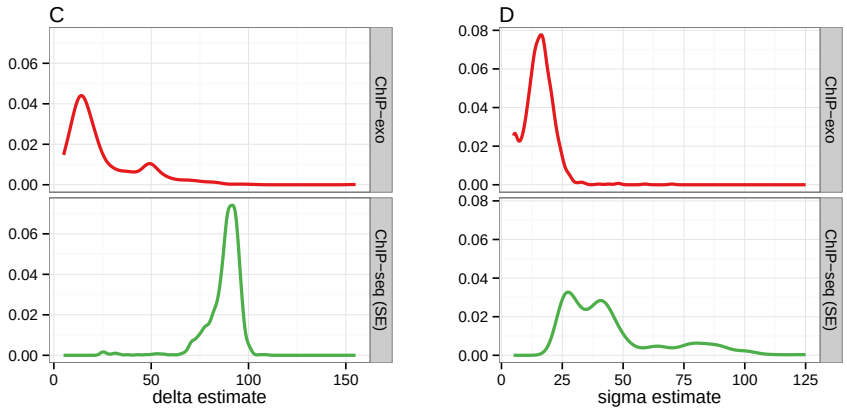


Figure S1 C) δ parameter in dPeak measures average distance of the reads to their respective binding site. In ChIP-exo data, reads were located much closer to the binding site than in SET ChIP-Seq. D) σ parameter measure the dispersion of reads around each binding site. In ChIP-exo data, reads showed less variation around the their respective binding sites compared to SET ChIP-Seq.

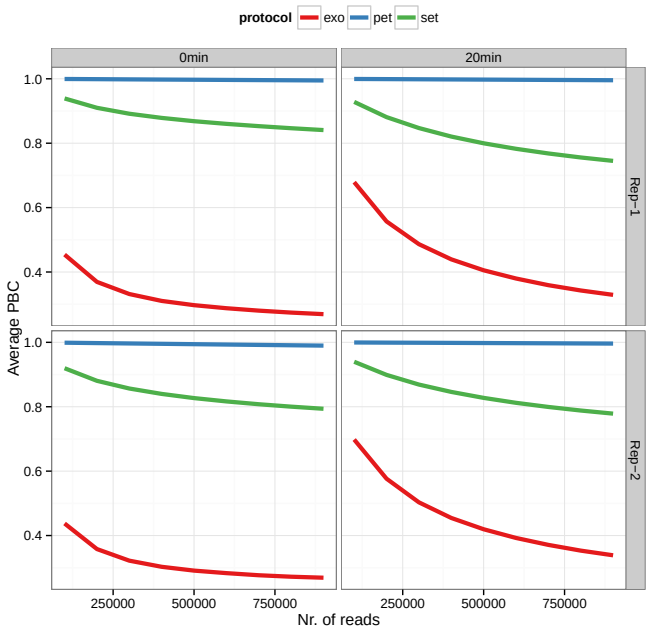
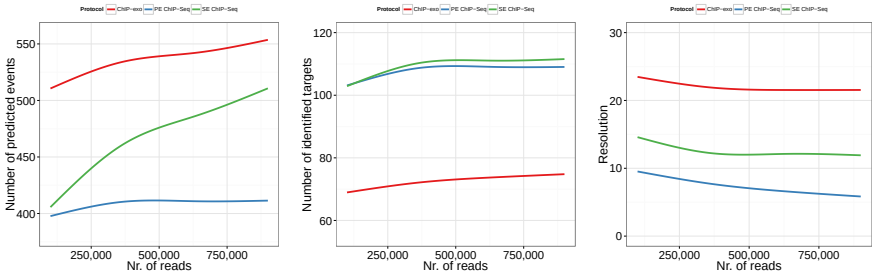
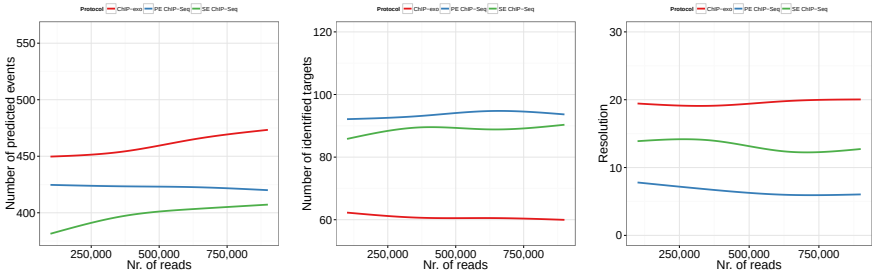


Figure S2 Average PBC (among all seeds) of the sampled ChIP-exo, PE ChIP-Seq and SE ChIP-Seq experiments under the rif treatment conditions used for saturation analysis.

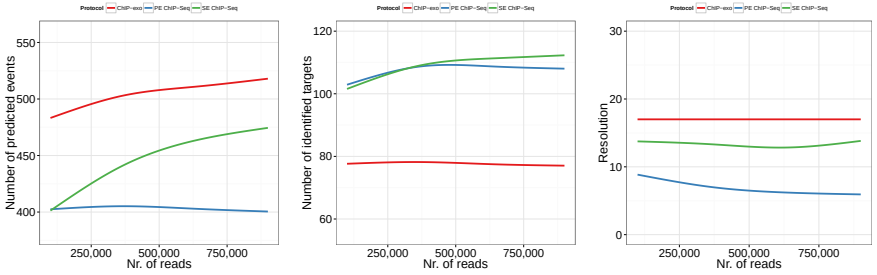
- Rep-1 and rif-0min:



- Rep-1 and rif-20min:



- Rep-2 and rif-0min:



- Rep-2 and rif-20min:

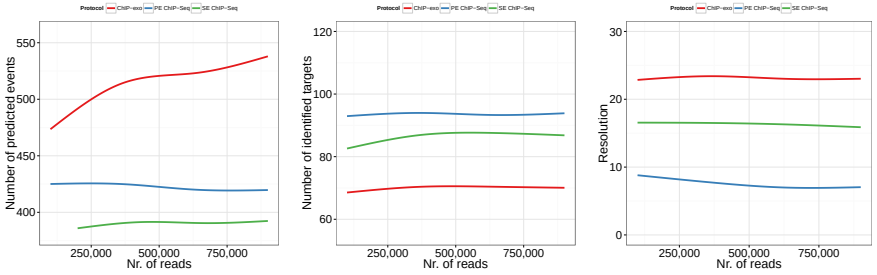


Figure S3 Comparison of the number of predicted events (left), identified targets (middle) and resolution (right) among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. RegulonDB annotations are considered as gold standard. A RegulonDB binding events was deemed identified if a binding event was estimated at a $\pm$ 15 vicinity of it.

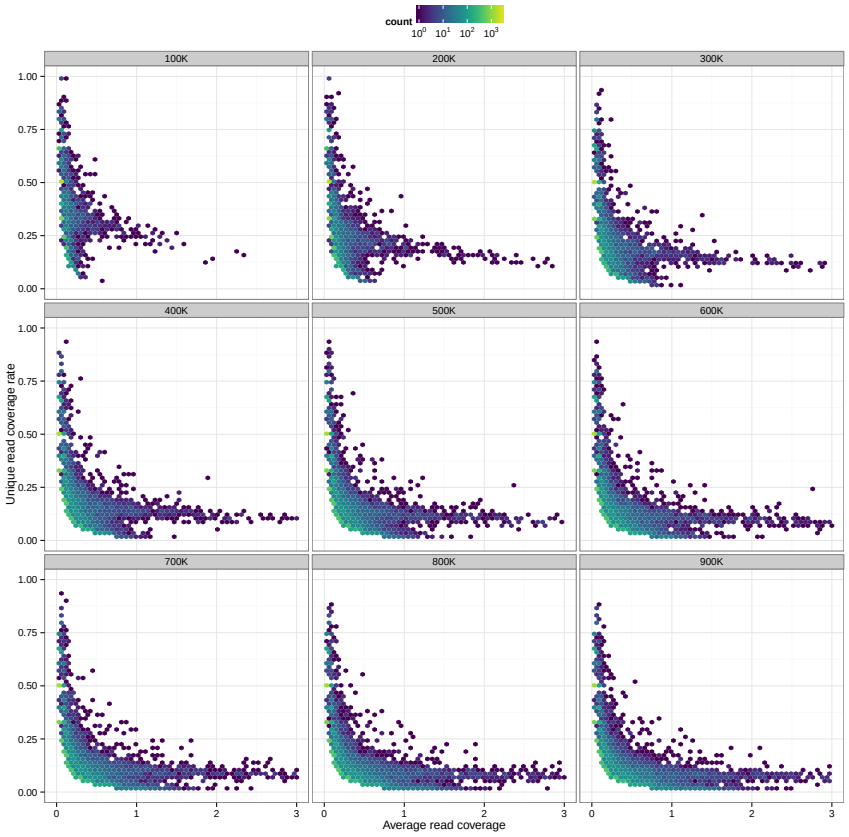


Figure S4 ARC vs URCR hexbin plots of Rep-1 and rif-0min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.

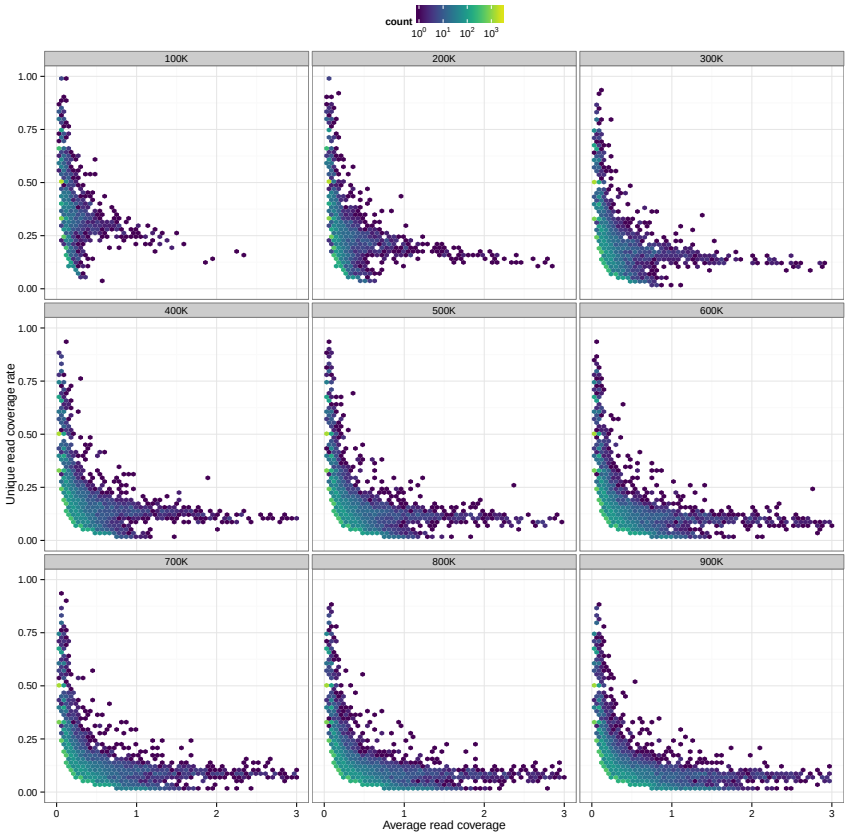


Figure S5 ARC vs URCR hexbin plots of Rep-1 and rif-20min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.

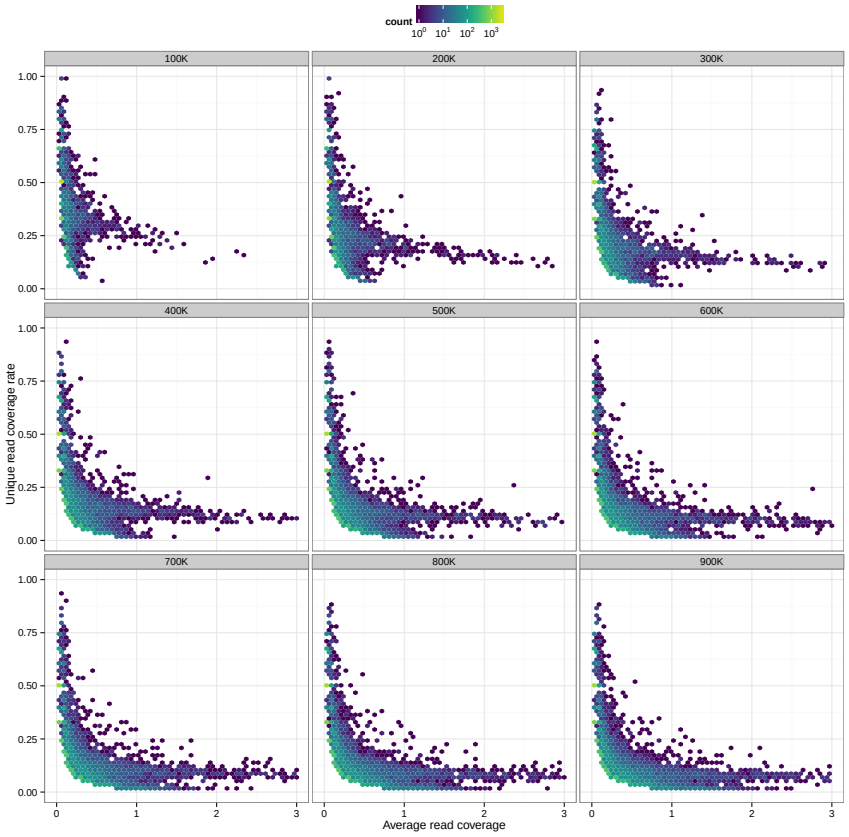


Figure S6 ARC vs URCR hexbin plots of Rep-2 and rif-0min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.

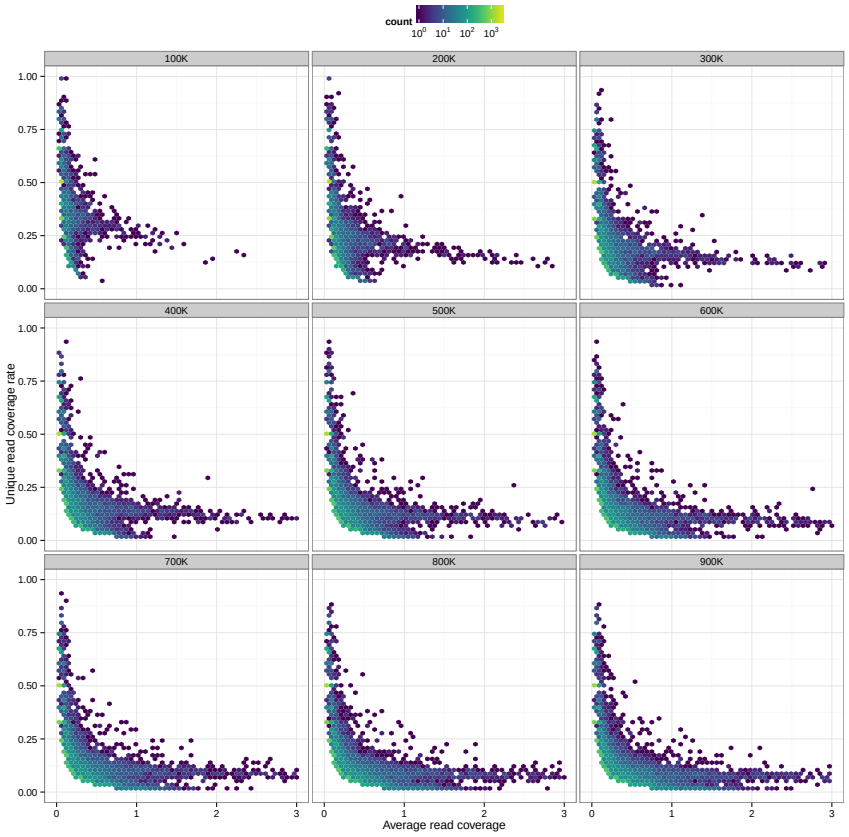


Figure S7 ARC vs URCR hexbin plots of Rep-2 and rif-20min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.