Published online 29 July 2017

Nucleic Acids Research, 2017, Vol. 45, No. 15 ***e145***
doi: 10.1093/nar/gkx594

Data exploration, quality control and statistical analysis of ChIP-exo/nexus experiments

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

[1]Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA, [2]Department of Public Health Sciences, Medical University of South Carolina, SC 29425, USA, [3]Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, WI 53726, USA, [4]Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA, [5]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI 53706, USA and [6]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53792, USA

Received April 13, 2017; Revised June 02, 2017; Editorial Decision June 27, 2017; Accepted July 12, 2017

ABSTRACT

ChIP-exo/nexus experiments rely on innovative modifications of 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 novel statistical quality control pipeline and accompanying `R/Bioconductor` 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 computed 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 *Escherichia coli*. `ChIPexoQual` analysis of these datasets resulted in guidelines for using these QC metrics across a wide range of sequencing depths and provided further insights for modelling ChIP-exo data.

INTRODUCTION

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 (36–100 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. In addition to these, several other high-resolution protocols have also been considered. In X-ChIP and ORGANIC (3,4), the DNA is fragmented by the application of endonuclease and exonuclease enzymes and then stabilized by sonication. The main difference between these two protocols is that in X-ChIP, the cells are crosslinked with formaldehyde and then the DNA is extracted by cell lysation, while the ORGANIC protocol achieves this step by nuclear isolation. Currently, ChIP-exo seems to be the more commonly adapted high-resolution protocol.

Figure 1A illustrates the differences between distinct ChIP-based protocols: ChIP-exo, ChIP-nexus, single-end (SE) ChIP-seq, paired-end (PE) ChIP-seq. The 5′ ends from

*To whom correspondence should be addressed. Tel: +1 608 263 4533; Fax: +1 608 262 0032; Email: keles@stat.wisc.edu
†These authors contributed equally to this work as first authors.

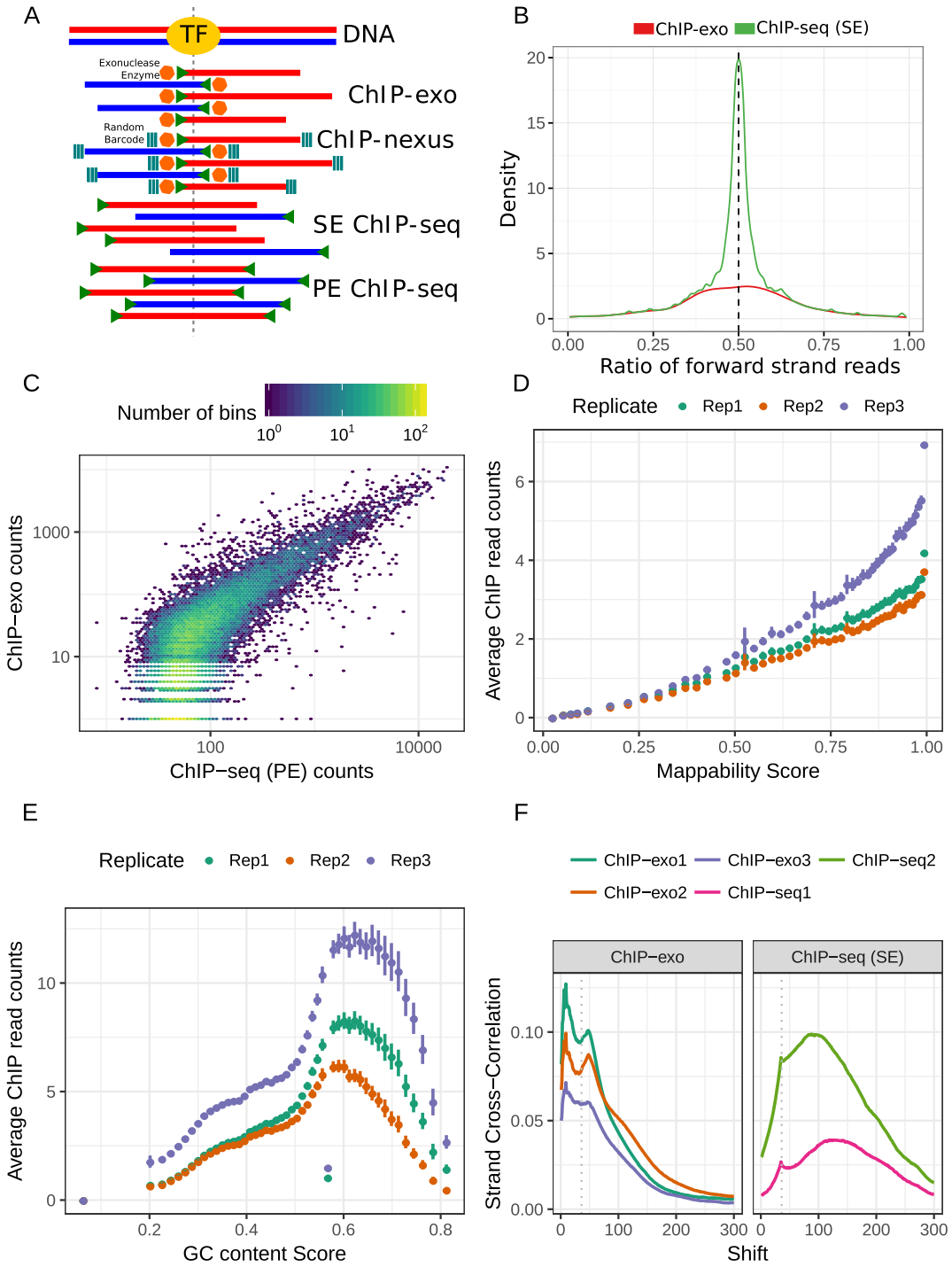

Figure 1. ChIP-seq versus ChIP-exo/nexus. (**A**) Processing of sonicated fragments bound by TF before immunoprecipitation and PCR amplification: For ChIP-exo, an exonuclease enzyme (orange hexagon) trims the 5′ ends of each DNA fragment to a fixed distance from the TF. For ChIP-nexus, a random barcode is added on the 3′ end, and transferred to the 5′ stopping base by self-circularization. For both ChIP-exo and SE ChIP-seq, an adaptor is ligated (green triangles) at the 5′ ends. The adaptors are ligated to both ends for PE ChIP-seq. (**B**) Forward Strand Ratio densities for SE ChIP-seq and ChIP-exo peaks. (**C**) Hexbin plot of PE ChIP-seq bin counts vs. ChIP-exo bin counts. (**D**) Mappability score vs. mean ChIP-exo read counts with error bands. E) GC-content vs. mean ChIP-exo read counts with error bands. (**F**) SCC curves for human CTCF from 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 (17) are shown in the right panel. The ChIP-exo curve shows local maxima at the motif and read lengths. SE ChIP-seq curves for both replicates are maximized at the fragment length and show local maxima at the read length.

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 the research community, features of ChIP-exo data, specially those pertaining to data quality, have not been investigated. First, DNA libraries generated by the ChIP-exo protocol are expected to be less complex than the libraries generated by ChIP-seq (5) because digestion by λ-exo aims to reduce the number of individual genomic positions, to which sequencing reads can map, to small regions located around the actual binding sites. Therefore, in high quality and 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 over-amplification 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 an important point with implications on the statistical analysis of ChIP-exo data. Specifically, currently available ChIP-exo specific statistical analysis methods (e.g. MACE (6), CexoR (7) and Peakzilla (8)) 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 (9–11) 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/Bioconductor` package named `ChIPexoQual`. The overall pipeline takes into account the characteristics of ChIP-exo/nexus data and addresses the critical shortcomings of the currently available QC pipelines that are not particularly tailored for ChIP-exo/nexus data (9–10,12–13). We apply this pipeline to a large collection of public and newly generated ChIP-exo/nexus data and we validate the QC pipeline by evaluating the samples for features that capture high signal to noise, such as occurrences of motifs recognized by the profiled DNA interacting protein and also utilize blacklisted regions as identified by the ENCODE consortium.

MATERIALS AND METHODS

ChIP-seq/exo/nexus datasets

E. coli *ChIP-exo and ChIP-seq samples.* For simplicity, we introduce some abbreviations for the *Escherichia coli* σ^{70} ChIP-exo (E), PE ChIP-seq (P), and SE ChIP-seq (S) samples. We denote the data generated in the first (second) batch as E1 (E2), P1 (P2) and S1 (S2). Summaries of the growth conditions and sample IDs for the ChIP-exo samples are included in Table 1. The SE and PE ChIP-seq samples generated under the same conditions share the same Id. convention. The procedures for sample preparation and sequencing are described in the supplement. The ChIP-exo experiments followed the protocol 7 described in (1).

Processing of the ChIP-exo and ChIP-nexus samples. We aligned the ChIP-exo/nexus samples in Table 2 by following the descriptions listed in their respective publications. When the alignment settings were not discernible in the original publication, we used `bowtie` (version 1.1.2) (14). We aligned the E1 samples of Table 1 with `bowtie -q -m 1 -l 55 -k 1 -5 3 -3 40 --best -S` and the E2 samples using `bowtie -q -m 1 -v 2 --best`. The average read lengths were 102 and 52 bp for the E1 and E2 samples, respectively. Hence, to make the alignments for both samples comparable, we trimmed 40 bp from the 3′ ends of the reads in the E1 samples. We trimmed 3 bp from the 5′ end to remove the adaptors in the E1 samples.

ChIP-exo and ChIP-seq peak calling with MOSAiCS to identify high signal peaks

`MOSAiCS` (15) is a model-based approach for the analysis of ChIP-seq and ChIP-exo data. We used `MOSAiCS` to identify sets of highly significant peaks for ChIP-exo and ChIP-seq under the GC + Mappability and InputOnly 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 of length 150 bp and counted the number of reads overlapping each interval. As is usually the practice with ChIP-seq analysis, each read was extended to the average fragment length of 150 bp toward the 3′ direction. To evaluate the strand imbalance, we identified a set of high signal peaks for ChIP-exo and SE ChIP-seq. The subset of these peaks for which `dPeak` (16) analysis identified one or more binding events were used in FSR assessments (Figure 1B and Supplementary Figure S1E).

Existing next generation sequencing data QC metrics and methods

We used the ChIP-seq QC metric definitions established by the ENCODE consortium (10,11), 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/keleslab/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 0–0.5 indicates severe bottlenecking, 0.5–0.8 moderate bottlenecking, 0.8–0.9 mild bottlenecking and 0.9–1 no bottlenecking.

In addition to ENCODE QC metrics, we considered `FASTQC` (version 0.11.5) and `htSeqTools` (version 1.16.0) (9) for assessing the overall quality of the ChIP-exo/nexus sequences. Collectively, these encompass all the metrics available for read-level data in `ChiLin` (13), which is another QC tool for ChIP-seq and DNase-seq, and `Q-nexus` (12), which is a ChIP-nexus analysis pipeline with QC features that are similar to that of `FASTQC`. The remaining metrics calculated by the `ChiLin` pipeline require the use of a peak calling algorithm or external data (such as DNase

Table 1. Summary of the *E. coli* σ^{70} ChIP-exo samples

Group	Growth	Treatment	Rep.	Id.	Depth	NSC	RSC	PBC	SSD
	Exp. $+O_2$	No Rif.	1	1	13 961 493	103.15	2.0193	0.1399	356.8525
ChIP-exo	Exp. $+O_2$	No Rif.	2	2	14 810 838	162.70	1.7805	0.1633	371.6857
(E1)	Stat. $+O_2$	No Rif.	1	3	16 108 774	153.51	1.8035	0.1353	402.3119
	Stat. $+O_2$	No Rif.	2	4	13 636 541	172.59	2.014	0.1532	400.2480
	Exp. $+O_2$	No Rif.	1	1	902 921	13.77	1.1270	0.2689	68.0992
ChIP-exo	Exp. $+O_2$	Rif. 20 min	1	2	1 852 124	17.91	1.5275	0.2590	96.9974
(E2)	Exp. $+O_2$	No Rif.	2	3	2 104 427	29.60	1.2844	0.2584	120.3401
	Exp. $+O_2$	Rif. 20 min	2	4	11 548 572	13.08	1.5122	0.1510	219.8427

Exp. stands for exponential and Stat. for stationary growth conditions. Rif. stands for *Rifampicin* treatment. Columns 7-10 depict QC metrics on these data: NSC: Normalized Strand Cross-Correlation; RSC: Relative Strand Cross-Correlation; PBC: PCR Bottleneck Coefficient; SSD: Standardized Standard Deviation.

Table 2. Summary of publicly available data used for development and evaluation of `ChIPexoQual`

Protocol	Organism	TF/histone	Cell type/treatment	Rep.	Depth	NSC	RSC	PBC	SSD
ChIP-exo	Human	CTCF	HeLa	1	23 576 694	22.82	1.2604	0.4654	0.8102
				2	20 947 081	13.79	1.1382	0.4292	0.7846
				3	37 688 587	20.51	1.2071	0.2744	0.6348
	Human	ER	MCF-7	1	9 289 835	19.87	1.0127	0.8082	0.6308
				2	11 041 833	21.48	1.0063	0.8024	0.7313
				3	12 464 836	18.72	1.0100	0.8203	0.7231
	Mouse	FoxA1	Liver	1	22 210 461	21.28	1.1104	0.6562	1.0728
				2	23 307 557	60.42	1.1604	0.7996	0.9790
				3	22 421 72	72.04	1.1975	0.1068	1.3861
	Human	GR	IMR90	1	47 443 803	8.86	1.3678	0.2978	0.8970
			K562	1	116 518 000	4.11	1.0441	0.0504	1.0708
			U2OS	1	3 255 111	10.05	1.0288	0.7714	0.3717
	Human	TBP	K562	1	61 046 382	12.01	1.1119	0.1232	0.7552
				2	94 314 770	7.93	1.0299	0.1681	2.8796
				3	114 282 270	9.25	1.1027	0.1464	2.9330
	S.Cerevisiae	H3	Tail deleted	1	35 951 922	6.80	1.0631	0.4435	24.1752
				2	32 568 539	4.37	1.2112	0.3902	17.2758
				3	21 600 382	11.91	1.1655	0.4774	29.0496
				4	11 030 924	14.92	1.1137	0.5381	30.9052
ChIP-nexus	D.Melanogaster	Dorsal	Embryo	1	8 863 170	7.27	1.0402	0.6766	1.6014
				2	10 003 562	7.19	1.0672	0.5656	1.6565
		Twist		1	18 244 203	5.82	1.1632	0.6592	4.6353
				2	52 546 982	5.27	1.1805	0.4549	7.1775
		Max	S2	1	18 320 743	3.60	1.3628	0.5178	2.8449
				2	24 965 642	3.47	1.0138	0.2124	5.2416
		MyC		1	7 832 034	5.92	1.0115	0.3935	3.3570
				2	22 824 467	5.76	1.0045	0.1879	6.9451
	Human	TBP	K562	1	33 708 245	32.16	1.1712	0.3102	2.2438
				2	129 675 001	32.70	1.2455	0.0492	4.5579

Columns 7-10 depict QC metrics on these data: NSC: Normalized Strand Cross-Correlation; RSC: Relative Strand Cross-Correlation; PBC: PCR Bottleneck Coefficient; SSD: Standardized Standard Deviation.

hypersensitive sites) and, therefore, are not utilized in our evaluations.

Blacklisted regions in eukaryotic genomes

For the mm9, hg19, and dm3 genomes, we used the blacklists generated by the ENCODE consortium (17), available at https://sites.google.com/site/anshulkundaje/projects/blacklists. These lists consist of genomic segments for which next-generation sequencing experiments produce artificially high signal. These lists were empirically derived from large compendia of data generated by the ENCODE and modENCODE consortia, respectively.

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

We implemented our proposed QC pipeline with an `R/Bioconductor` package named `ChIPexoQual`, available at http://bioconductor.org/packages/release/bioc/html/ChIPexoQual.html. The analysis in this paper used version 1.0.0 of the `ChIPexoQual` package.

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

1. Identify read islands, i.e. overlapping clusters of reads separated by gaps, from read coverage. The gaps are defined as the union of positions in the genome with fewer than h^* (default $=1$) aligned reads. The remaining is-

lands can be interpreted as the natural partition of the genome determined by a ChIP-exo/nexus experiment.
2. Compute D_i, number of reads in island i; U_i number of positions in island i with at least one aligning read; and W_i, the width of island i defined as the total number of bases in the island, $i = 1, \cdots, I$.
3. For each island i, $i = 1, \cdots, I$, compute island statistics:

$$\text{ARC}_i = \frac{D_i}{W_i}, \quad \text{URC}_i = \frac{U_i}{D_i},$$

$$\text{FSR}_i = (\#\text{ of fwd. strand reads aligning to island } i)/D_i,$$

4. Generate diagnostic plots (i) URC vs. ARC plot; (ii) Region Composition plot; (iii) FSR distribution plot.
5. Randomly sample without replacement M (at least 500, default = 1000) islands and fit,

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

where ε_i denotes the independent error term. Repeat this process B (default = 1000) times and generate box plots of estimated β_1 and β_2.

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

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

is a re-parametrization of the following relationship from URC vs. ARC diagnostic plot:

$$\text{URC}_i = \frac{\kappa}{\text{ARC}_i} + \gamma + \varepsilon_i \quad (1)$$

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. Equivalently, $\beta_1 = 1/\gamma$ can be interpreted as the average number of aligned reads per unique position when the sequencing depth is large. To interpret $\beta_2 = -\kappa/\gamma$, we express κ as a function of ARC and URC and assume that γ is already estimated. Then,

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

$$\frac{\kappa}{\gamma} = \frac{U}{W} \times \frac{1}{\gamma} - \text{ARC} = \frac{U}{W} \times \lim_{D \to \infty} \frac{D}{U(D)} - \text{ARC},$$

where γ approximates the URC as the sequencing depth increases.

In a low quality experiment where reads accumulate in a few number of positions due to PCR amplification bias or other artifacts, several reads are expected to repeatedly align to the same collection of unique positions, making the term involving the limit diverge from ARC. In contrast, in a high-quality experiment, κ/γ is expected to converge to zero because the expression with the limit approximates ARC.

The ChIPexoQual pipeline is enriched by the following two additional modules that are utilized when the sequencing depth is high and/or blacklisted regions are available.

i. *Subsampling analysis.* For high depth datasets (e.g., $\geq 60M$ reads for human and mouse samples), we subsample $N_1 < N_2 < \cdots < N$ reads, starting with $N_1 = 20M$ reads and up to $50M$ reads in $10M$ increments as default, and apply steps 1 to 5 for each of the subsampled datasets.
ii. *Blacklisted regions analysis.* The islands identified by `ChIPexoQual` are separated into two different collections based on their overlap with a set of blacklisted regions. Then, the β_1 and β_2 scores are estimated for both collections and compared against the all island scores.

Motif analysis of FoxA1 and TBP enriched regions

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

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

These thresholds were empirically selected. To validate their robustness, we performed an analogous analysis by using the regions that overlapped a set of peaks (identified by MOSAiCS at FDR 5%) with width larger than $3 \times rl$, where rl is the median read length of the experiment (Supplementary Figures S34 and S35). The width filter was not applied to the TBP ChIP-exo samples, and accordingly to the ChIP-nexus samples for consistency, since they exhibited over-amplification (2).

We used `FIMO` (version 4.9.1) (18) to identify the FoxA1 and TBP motifs within each enriched region using the FoxA1 MA0148.1 and TBP MA0108.1 position weight matrices from the `JASPAR` database (19), respectively. For the FoxA1 experiments we used the default parameters and for the TBP experiments we considered all motifs identified with `FIMO` p.value < 0.05.

RESULTS

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 2). These include: CTCF factor in human HeLa cell lines (1); ER factor in human MCF-7 cell lines (20); GR factor in IMR90, K562 and U2OS human cell lines (21); TBP factor in human K562 cell lines (22); H3 histone in *S. cerevisiae* where most, but not all of the tail was deleted (Δ1-28) (23). ChIP-nexus data included experiments from (2) profiling TBP in human K562 cells, MyC and Max in *D. melanogaster* S2 cell lines, 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 ChIP-exo (Table 1), SE and PE ChIP-seq. Collectively, we generated σ^{70} factor ChIP-exo, PE and SE ChIP-seq experiments under aerobic ($+O_2$) and anaerobic ($-O_2$) conditions in glu-

cose minimal media. For simplicity, we named these experiments as E1, P1 and S1, respectively. Similarly, we generated σ^{70} factor ChIP-exo and PE ChIP-seq experiments in *E. coli* under aerobic ($+O_2$) conditions with and without rifampicin treatment. We also named these experiments E2 and P2, respectively.

ChIP-exo versus ChIP-seq: general features

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*. Figures 1B–C summarize this comparison for one biological replicate of ChIP-exo and ChIP-seq experiments from the same biological conditions (samples E1-1 from Table 1, P1-1 and S1-1 following the same Id. convention).

Peak-pair assumption. We evaluated the peak-pair assumption, i.e. a cluster of reads in the forward strand located on the left-hand-side of the binding site is usually paired with a cluster of reads located on the right-hand-side of the binding site in the reverse strand. This observation is commonly utilized in designing statistical analysis methods for ChIP-exo data (6–8). 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 1B and Supplementary Figures S1–S3). 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 more likely to occur in low complexity libraries.

Read distributions within signal and background regions. Using extended raw read counts within 150 bp non-overlapping intervals, i.e., bins interrogating the genome, Figure 1C 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 between the two data types for bins with low read counts, highlighting potential differences in the background read distributions of these data types. Comparisons with other paired *E. coli* ChIP-seq and ChIP-exo samples led to similar conclusions (Supplementary Figures S1–S3).

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 1D and E (Supplementary Figure S4) 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 (15,24–25). 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.

Existing high throughput sequencing quality control metrics applied to ChIP-exo/nexus data

We processed the ChIP-exo/nexus samples with FASTQC and observed that in 73.33% and 93.33% of the cases, at least a warning is raised for sequence duplication levels and kmer content representation (Supplementary Table S1), respectively. The former assumes that most sequences will occur only once in a diverse library and the latter assumes that any small fragment should not have a positional bias in its appearance within a library. Clearly, these assumptions are not appropriate for ChIP-exo/nexus data, as the exo-enzyme is expected to stop its digestion when it reaches the crosslinking protein.

The ENCODE consortium established empirical and widely used QC metrics on ChIP-seq data (10). We evaluated how these metrics, namely PCR Bottleneck Coefficient (PBC), Normalized Strand Cross-Correlation (NSC), and Relative Strand Cross-Correlation (RSC) defined at https://genome.ucsc.edu/ENCODE/qualityMetrics.html (10,11). Tables 1 and 2 present these metrics for the collection of ChIP-exo/nexus datasets we consider in this paper.

Marinov *et al.* (11) discussed that highly complex ChIP-seq libraries can become exhausted by deep sequencing. Hence, the PBC is expected to decrease as the sequencing depth increases. This effect is expected to be more severe in ChIP-exo/nexus as DNA libraries generated by those protocols are expected to be less complex than the libraries generated by ChIP-seq because the numbers of positions to which the reads can align to are reduced due to the exonuclease digestion. This affects the interpretation of the PBC, which is defined as the ratio of the number of genomic positions to which exactly one read maps to the number of genomic positions to which at least one read maps. For ChIP-seq samples, low PBC values (e.g., $\leq$0.5) indicate high levels of PCR amplification bias, i.e. PCR bottleneck, unless the sequencing depth is high enough to saturate 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.

Planet *et al.* (9) presented in the `R/Bioconductor` package `htSeqTools` the Standardized Standard Deviation (SSD) as a metric to assess enrichment efficiency and to compare across samples. According to the guidelines established by the authors, higher values of this metric indicates high-quality. We calculated the SSD coefficient for all the ChIP-exo/nexus samples (Tables 1 and 2). Detailed examination of these results reveals a key shortcoming of this metric as the propensity to label samples with low library complexity as higher quality because the reads in such sam-

ples align to fewer positions in the genome. For example, when comparing the ChIP-exo/nexus TBP samples, the use of this metric suggests that the deeply sequenced ChIP-exo samples (replicates 2 and 3) exhibit higher quality than the first ChIP-nexus replicate. This is in contrast to evaluation of these datasets with an independent, motif-based metric as we discuss below.

The Strand Cross-Correlation (SCC), introduced by Kharchenko *et al.* (26), is a 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 between 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, the SCC profile is expected to maximize at the motif length indicating clustering of the forward and reverse strand reads around the binding site. This hinders the interpretation of SCC for a ChIP-exo/nexus experiment since it is now maximized at an unobserved shorter fragment length that is confounded with the *'phantom peak'* at the read length. Carroll et al. (27) studied the impact of blacklisted regions and duplicated reads when calculating the SCC for ChIP-exo data. The authors showed that there is a dramatic effect in the SCC profile when removing duplicated reads but the effect of removing the blacklisted regions may be specific in few positions of the SCC profile and suggested to calculate the SCC using only aligned reads that overlap the experiment's set of peaks but don't overlap a set of predefined blacklisted regions. Several biases are introduced into the computation of this modified SCC, because it requires the use and tuning of a peak calling algorithm. Furthermore, in a lower quality experiment, the peaks may not correspond to actual binding sites. Figure 1F displays the SCC curves for the CTCF HeLa samples where the ChIP-exo curve actually shows local maxima at 12 bp and the read length, while the SE ChIP-seq curves have an expected 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 S5 to S14. In ChIP-exo experiments, the read length and the fragment length peaks in the SCC are confounded. Furthermore, the former is close in proximity to the motif length; as a result, this may incorrectly suggest experiments to be marginally successful or even failed (e.g. Supplementary Figure S8) and renders QC metrics such as the Normalized Strand Cross-Correlation (NSC) or the Relative Strand Cross-Correlation (RSC) harder to interpret. However, in majority of the cases we present, the profile itself seems informative about the enrichment signal in ChIP-exo/nexus experiments.

ChIP-exo quality control pipeline `ChIPexoQual`

To address the limitations of available analytical exploration approaches discussed above, we developed `ChIPexoQual`. In Table 3, we compare `ChIPexoQual` against the existing tools discussed above. We highlight that `ChIPexoQual` provides a global view of both library enrichment and complexity, and detailed diagnostic plots for the balance between the two. We first present the overall pipeline and then discuss individual components with a case study using ChIP-exo data of FoxA1 from (20) and ChIP-nexus data from (2). Figure 2 summarizes the 4-step pipeline and the two additional modules. Given aligned reads from a ChIP-exo/nexus sample, the first step partitions the reference genome into islands representing overlapping clusters of reads separated by gaps by removing the regions with fewer than h^* aligned reads. In step 2, the total number of reads overlapping each island (D_i) and the number of island positions with at least one aligned read (U_i) are recorded. 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 number 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 number of genomic positions with at least one aligned read within island i (U_i) to the number 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. 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. As a result, this arm depicts the balance between library enrichment and complexity. Low URC in this arm corresponds to regions composed by reads concentrated in a smaller number of positions.

We quantify the shape of the URC versus ARC 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. These parameters are estimated by sampling experiments on the original samples. We provide further details on how to obtain these later in the paper where we apply the pipeline to a large collection of ChIP-exo/nexus experiments. Figure 2B and C present the typical behavior of the *Region Composition* and *Forward Strand Ratio (FSR) distribution* plots, both of which quantify the strand imbalance as part of the QC pipeline. The *Region Composition plot* depicts how quickly the ratio of islands exclusively composed of fragments on a single strand among the islands with comparable read depth decreases as a function of read depth of the island. In a high quality sample, the proportion of islands with reads from only one strand is expected

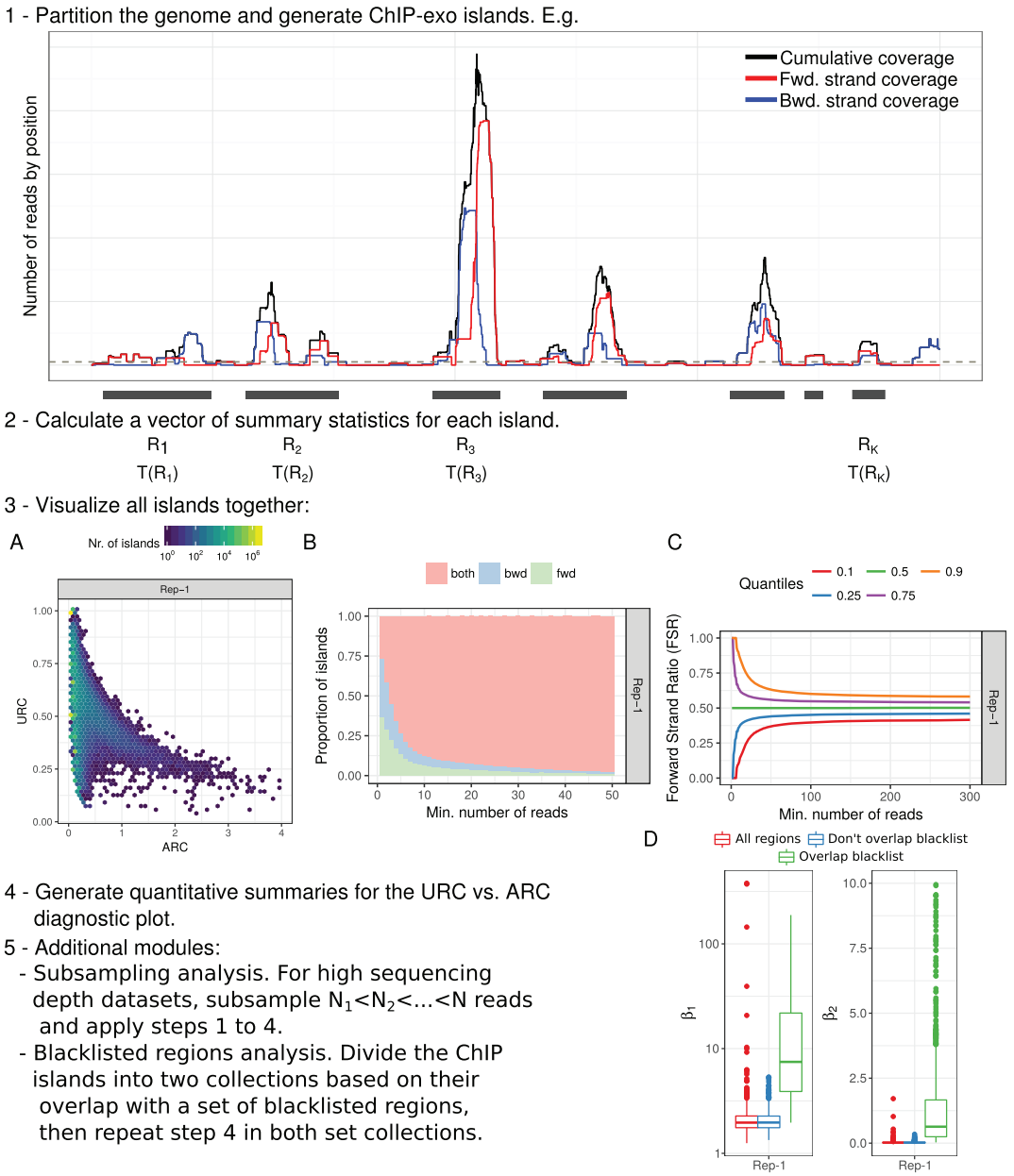

Figure 2. ChIP-exo QC pipeline `ChIPexoQual`. The ChIP-exo reads are partitioned into overlapping clusters of reads separated by gaps (step 1). For each region, the following summary statistics are calculated (step 2) and visualized (step 3): Average Read Coefficient (ARC), Unique Read Coefficient (URC) and Forward Strand Ratio (FSR). These statistics are visualized as: (**A**) URC versus ARC plots, which presents the overall balance between library complexity and enrichment; there are two arms, one with low ARC and varying URC and one where the URC decreases as the ARC increases. (**B**) Region Composition plot, which shows the strand composition for all the regions formed by a minimum number of reads. (**C**) FSR distribution plot, which illustrates the FSR's distribution as the depths of the islands get larger. D) Example of the Blacklisted region analysis module. Both $\hat{\beta}_1$ and $\hat{\beta}_2$ scores are significantly higher for islands overlapping the blacklisted regions, and robust to the removal of them.

to decrease rapidly as we consider higher depth regions. In contrast, this proportion remains approximately constant in lower quality samples. The *Forward Strand Ratio distribution 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 have approximately equal numbers of reads in both strands.

Application and validation of `ChIPexoQual` with the FoxA1 ChIP-exo dataset. We next illustrate the proposed QC pipeline using FoxA1 ChIP-exo datasets, which were profiled at comparable sequencing depths in three biological replicates of mouse liver cells. We first investigated various thresholds for partitioning the mouse genome using these ChIP-exo samples. We specifically considered small thresholds because larger thresholds are likely to partition wider regions into smaller ones, discard parts of wide regions, and ignore background regions completely. With this in mind,

Table 3. Comparison of state-of-the-art quality control tools for ChIP-seq and ChIP-exo/nexus samples

Aspect\Tool	ChIPexoQual	ChiLin	ChIPQC	phantompeakqualtools	htSeqTools	FASTQC	Q-nexus
Pipeline tailored to ChIP-exo/nexus experiments	✓						✓
Global view of library enrichment	✓	✓	✓	✓	✓		
Global view of library complexity	✓	✓	✓	✓	✓		✓
Balance between library enrichment and amplification	✓						
Peak-pair assumption diagnostic by dynamic analysis of strand imbalance	✓						
Analysis of subsampled experiment to determine overall quality	✓	✓					
Explicit analysis of blacklisted regions	✓	✓					
Sequence quality scores distributions		✓				✓	
Analysis of over-represented kmers and sequences		✓				✓	
Analysis of duplicated reads	✓	✓	✓	✓		✓	✓

we processed the FoxA1 datasets with the following thresholds 1, 5, 25 and 50 (Supplementary Figure S15). We observed that, in a high-quality experiment, if multiple thresholds are small and close to each other, then the partitions are similar and the distributions of the proposed metrics are similar as well. Hence, we decided to use the default threshold of $h^* = 1$ when analyzing the FoxA1 samples.

Figure 3A presents URC versus 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, elucidating that the libraries for the first two replicates are more complex than that of the third replicate.

Figures 3B and C display the Read Composition and FSR distribution plots, which highlight specific problems with replicates 2 and 3. Figure 3B exhibits apparent decreasing trends in the proportions of regions formed by fragments in one exclusive strand. High quality experiments tend to show exponential decay in the proportion of single stranded regions, while for the lower quality experiments, the trend may be linear or even constant (Supplement Figure S21). FSR distributions of both of replicates 2 and 3 are more spread around their respective medians (Figure 3C). The rate at which the 0.1 and 0.9 quantiles approach the median indicate the aforementioned lower enrichment in the second replicate and the low complexity in the third one.

In addition to step 4, when a set of blacklisted regions is available we divide the ChIP-exo/nexus islands into two groups based on whether or not they overlap the blacklisted regions. Figure 3D illustrates that, first, β_1 and β_2 scores are robust to existence of islands in the blacklisted regions. Second, for the islands overlapping the blacklisted regions, both summary metrics are significantly higher in both the overall level and variance. Therefore, this stratified analysis further indicates that the β_1 and β_2 scores provide good overall assessments of the datasets and can clearly separate blacklist regions.

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. A conservative approach to identify high quality binding regions (Materials and Methods) reveals 7014, 1855, and 2187 regions for replicates 1, 2 and 3, respectively. The lower number of enriched regions from replicate 2 is consistent with the lower ChIP enrichment pattern in the URC vs. ARC diagnostic plot. Figure 4A compares the FIMO scores among the three replicates, not-surprisingly confirming that the first replicate exhibits the highest quality.

Figure 4B 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 slightly 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 4C compares FSR distributions of the ChIP islands overlapping the union of the peaks across the three replicates and highlights that the samples largely satisfy the 'peak-pair' assumption because peaks with at least one motif tend to be more strand-balanced. Furthermore, samples with lower library complexity appear to exhibit heavier FSR tails.

High sequencing depth may confound low-complexity library issues. We evaluated every sample listed in Tables 1 and 2 with the ChIPexoQual QC pipeline (Supplementary Figures S16–S27). A key observation from this large scale analysis is that the URC versus ARC plots typically display one of 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 a sequencing depth of $\sim$ 130M reads, exhibit this pattern.

A simple but effective strategy to distinguish the two plausible scenarios from Pattern III is to apply the QC pipeline to sub-samples randomly generated from the full dataset at varying sequencing depths (*sub-sampling analysis* module). We applied this strategy by sub-sampling 20M to 50M reads in 10M increments, a range that represents the sequencing depths of the human samples we are using in this paper, from the TBP samples. URC vs. ARC diagnostics of these sub-samples (Supplementary Figures S30 to S33) indicate that, among the four TBP samples with this pattern, replicates two and three of K562 ChIP-exo suffer from low-complexity library issues, whereas the other sam-

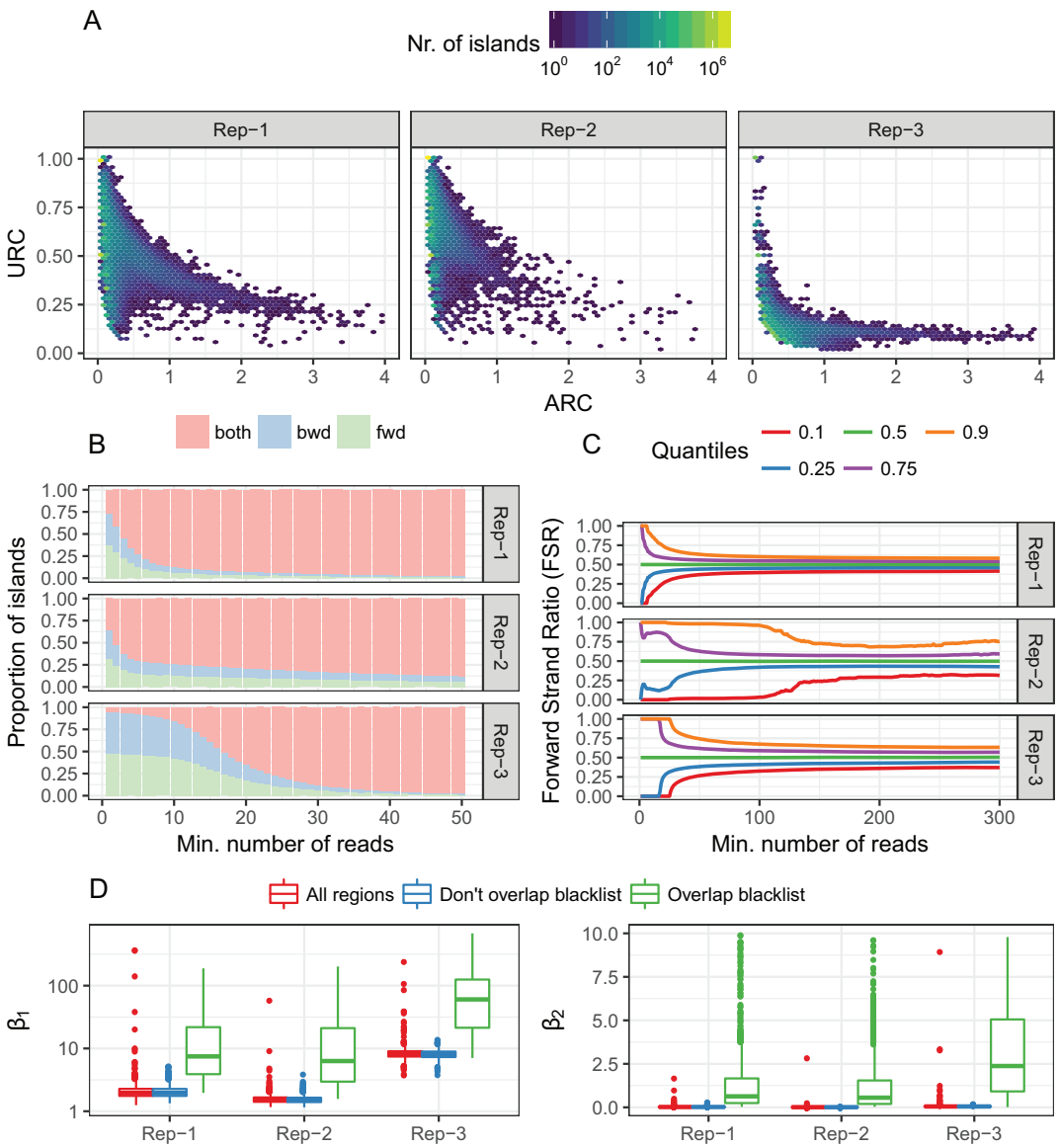

Figure 3. `ChIPexoQual` diagnostic plots for the FoxA1 ChIP-exo data (20). (**A**) URC versus ARC plot, (**B**) Region Composition plot, (**C**) FSR distribution plot comparison across three replicates and (**D**) β_1 and β_2 scores stratified based on overlap with the blacklisted regions.

ples exhibit the pattern specific to high quality samples. To confirm this implication, we compared the top FIMO scores (18) of the TBP motif for the ChIP-exo and ChIP-nexus replicates. Figure 4D illustrates that the first ChIP-exo replicate and ChIP-nexus replicates identify binding events with consistently better motif matches than the other ChIP-exo replicates. This implication on overall quality is further confirmed by the large separation of the β_1 and β_2 scores between regions that do and do not overlap with the blacklist regions for these high quality samples (Supplementary Figures S28-S29).

Figure 4E compares the FSR distributions of ChIP islands overlapping the union of peaks across all TBP samples by stratifying them with respect to TBP motif occurrence. Overall, while the peaks in high quality experiments are more likely to have a motif occurrence if they are balanced, many strand-unbalanced peaks with motifs are also identified. Specifically, the proportion of peaks with FSR smaller than 0.3 or larger than 0.7 varied between 0.38-0.43 and 0.20-0.22, for ChIP-exo and the ChIP-nexus experiments, respectively. This further confirms the conclusion of the `ChIPexoQual` QC pipeline.

Summary statistics for the URC versus ARC diagnostic plot. We next utilized QC pipeline results for all the samples (Tables 1 and 2) and quantified the relationship between ARC and URC by fitting a reparametrized regression model of URC as a function of ARC. Specifically, we considered a model of read depth (D_i) on the number of positions with at least one aligned read (U_i) and the width of the island (W_i), i.e., $D_i = \beta_1 U_i + \beta_2 W_i + \varepsilon_i$, where ε_i represents the random error term. As we discuss in Materials and Methods, this parametrization has a direct connection

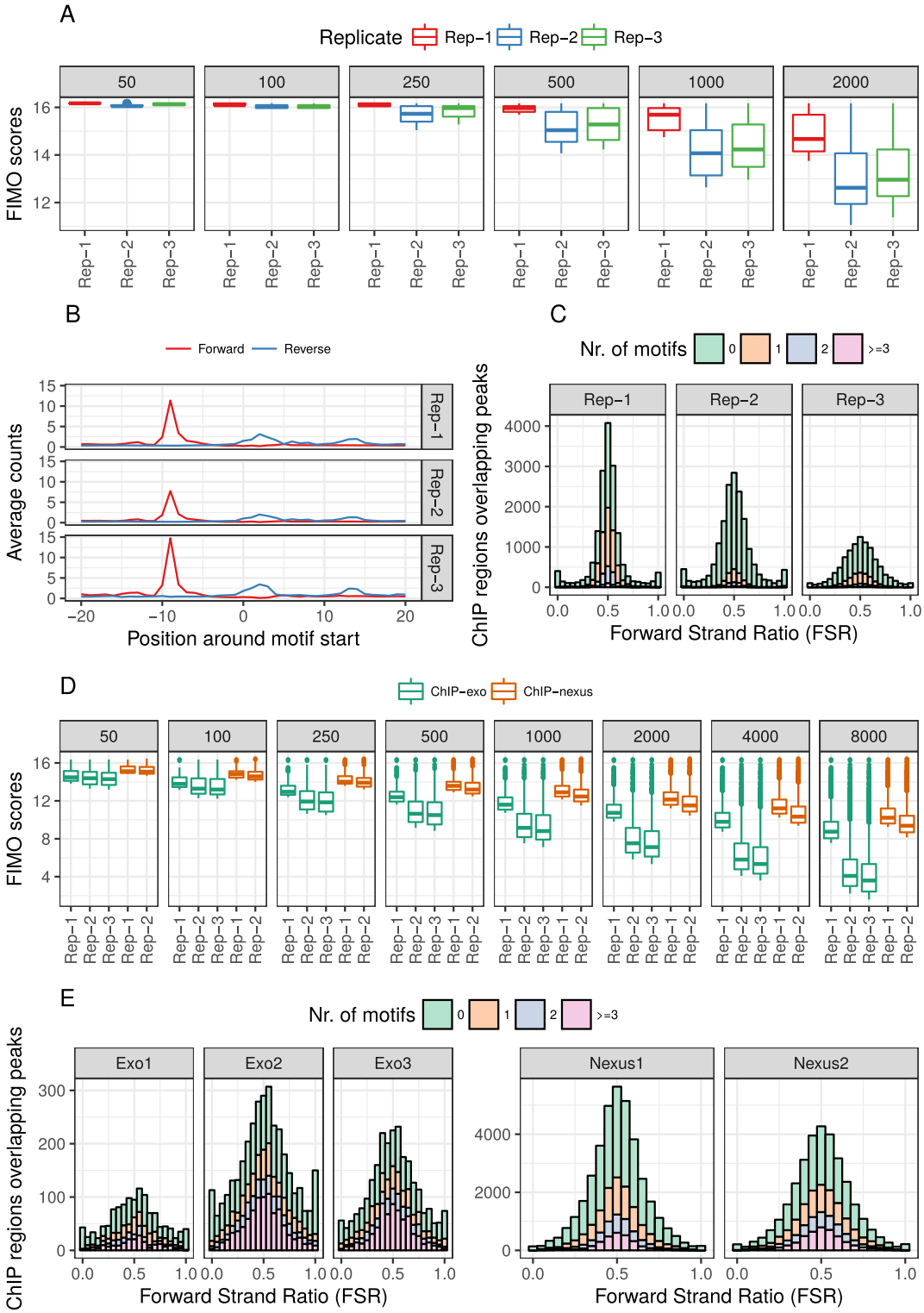

Figure 4. Validation of the `ChIPexoQual` pipeline with FoxA1 ChIP-exo (A–C) and TBP ChIP-exo/nexus (D, E) data. (**A**) Comparison of the top 50, 100, 250, 500, 1000 and 2000 FIMO scores for each replicate. (**B**) FoxA1 average coverage plots of the 5′ read ends centered around motif start positions separated by replicate and strand. (**C**) FoxA1 FSR distribution of `ChIPexoQual` islands overlapping ChIP-exo peaks stratified by the number of motifs. (**D**) Comparison of the top 50, 100, 250, 500, 1000, 2000, 4000 and 8000 FIMO scores for each TBP ChIP-exo/nexus sample. (**E**) TBP FSR distribution of `ChIPexoQual` islands overlapping ChIP-exo/nexus peaks stratified by the number of motifs.

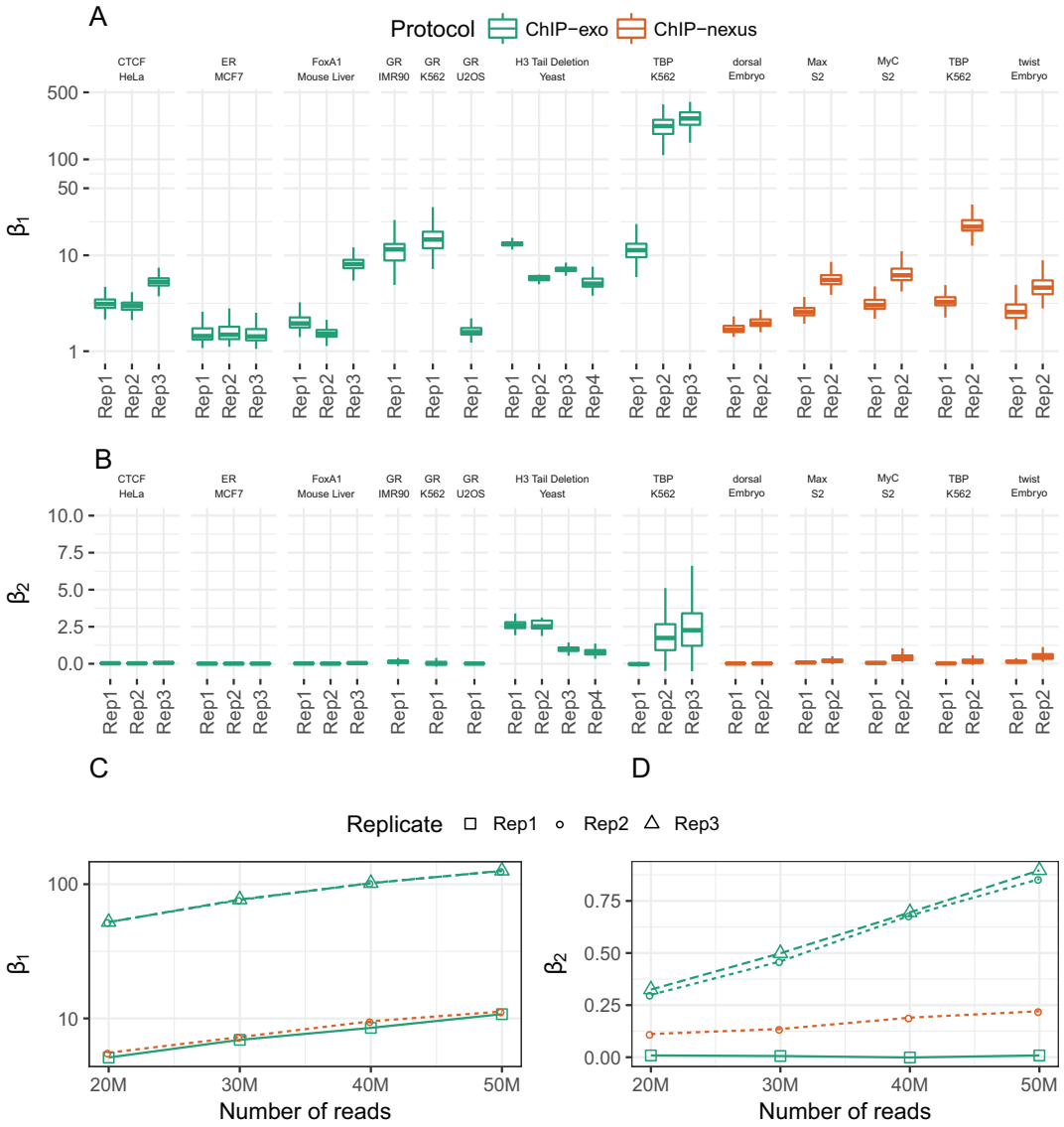

Figure 5. Comparison of `ChIPexoQual` numerical summaries. (**A**) $\hat{\beta}_1$ and (**B**) $\hat{\beta}_2$ for all eukaryotic ChIP-exo/nexus samples. (**C**) Average estimated β_1 and (**D**) β_2 for the ChIP-exo/nexus TBP samples in K562 cell lines when sub-sampling 20M to 50M reads.

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 γ parameter can be interpreted as the limiting (i.e., large depth) URC of a sample. As discussed earlier, high quality ChIP-exo samples are expected to have two arms in the URC versus 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. In contrast, 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 S16–S27), we conclude that samples with estimated $\hat{\beta}_1$ values <10 seem to be high quality samples.

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 binding events. In a high quality sample, the overall variation in depth is expected to be small as the overall widths of the regions change. This is because the majority of reads are expected to be located tightly around the binding sites and, as a result, the region width should not significantly affect its depth. In contrast, low quality sample 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. For example, although the third replicate of the TBP ChIP-exo experiment has comparable

sequencing depth to the second replicate of the TBP ChIP-nexus experiment (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. In summary, samples with estimated β_2 values close to zero can be considered as high quality samples.

The interaction between β_1 and β_2 has implications regarding the quality of ChIP-exo and ChIP-nexus samples. When either $\hat{\beta}_1$ 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 apply this strategy for the three replicates of TBP ChIP-exo in K562 (22) and second replicate from the K562 ChIP-nexus experiments (2). Figure 5C reveals a much higher $\hat{\beta}_1$ (and larger than 10) for replicates 2 and 3 compared to replicate 1 and both ChIP-nexus samples. Figure 5D illustrates that the β_2 estimates remain approximately constant in ChIP-nexus sub-samples and sub-samples of first replicate of ChIP-exo, while they increase 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 S30–S33) illustrate that the two arms of the ARC vs. URC plots are clearly visible in moderate depth sub-samples of TBP ChIP-nexus data. Similarly, Supplementary Figure S32 illustrates that, as expected, the suggested subsampling strategy is also effective for the E1 and E2 samples, which are deeply sequenced, relative to the *E. coli* genome.

`ChIPexoQual` R package

We implemented `ChIPexoQual` as an R/Bioconductor package. `ChIPexoQual` utilizes a fast processing algorithm by parallel computing. Supplementary Figure S36 provides `ChIPexoQual`'s processing times for a collection of samples representing different sequencing depths of the ChIP-exo/nexus experiments listed in Table 2 using four parallel threads on a server with 24 AMD 55Opteron 2.2GHz processors. This plot shows that `ChIPexoQual` requires between 125 and 640 s (80 and 420 when the aligned reads are already loaded into memory) for processing a ChIP-exo/nexus sample.

CONCLUSION

We presented a systematic exploration of several ChIP-exo/nexus datasets. We provided a list of factors that reflect the quality of a ChIP-exo/nexus experiment and developed an easy to use QC pipeline, implemented into an R/Bioconductor package called `ChIPexoQual`. `ChIPexoQual` takes aligned reads as input and automatically generates several diagnostic plots and summary measures that enable assessing enrichment and library complexity.

Our analysis of several datasets indicated that the QC pipeline only requires a set of aligned reads to provide a global overview of the quality of a given ChIP-exo dataset. The 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 often may not be available, such as motif occurrences in a set of high quality regions or resolution analysis based on a gold-standard.

The `ChIPexoQual` package (version `1.0.0`) is available from Bioconductor (http://bioconductor.org/packages/release/bioc/html/ChIPexoQual.html). The Bioconductor version does not currently include the blacklist sub-module. A stable version (version `0.99.15`) with this additional submodule is available at https://github.com/welch16/ChIPexoQual/tree/devel.

DATA AVAILABILITY

Escherichia coli ChIP-exo sequence and processed data are available under the NCBI's Gene Expression Omnibus (28) and are accessible through GEO series accession number GSE84830 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84830).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

R.W. acknowledges the funding provided by CONACYT.
Authors' contributions: R.W. and S.K. developed the `ChIPexoQual` pipeline. RW implemented the `ChIPexoQual` pipeline and D.C. implemented the `dPeak` package. R.W. and D.C. performed the analysis. J.G. and R.L. performed the *E. coli* sequencing experiments. R.W. and S.K. wrote the manuscript. All authors approved the final draft.

FUNDING

National Institutes of Health (NIH) [HG003747 and HG007019 to S.K.] (in part); NIH [GM38660 to R.L.]; CONACYT [215196 to R.W.]. Funding for open access charge: NHGRI.
Conflict of interest statement. None declared.

REFERENCES

1. Rhee,H.S. and Pugh,F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
2. He,Q., Johnston,J. and Zeitlinger,J. (2014) ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat. Biotechnol.*, **33**, 395–401.
3. Kasinathan,S., Orsi,G.A., Zentner,G.E., Ahmad,K. and Henikoff,S. (2014) High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods*, **11**, 203–209.
4. Skene,P.J. and Henikoff,S. (2015) A simple method for generating high-resolution maps of genome-wide proteing binding. *eLIFE*, **e09225**, 1–9.
5. Mahony,S. and Franklin,P.B. (2015) Protein-DNA binding in high-resolution. *Crit. Rev. Biochem. Mol. Biol.*, **50**, 269–283.

6. Wang,L., Chen,J., Wang,C., Uusküla-Reimand,L., Chen,K., Medina-Rivera,A., Young,E.J., Zimmermann,M.T., Yan,H., Sun,Z. *et al.* (2014) MACE: model based analysis of ChIP-exo. *Nucleic Acids Res.*, **42**, e156.
7. Madrigal,P. (2015) CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates. *EMBnet. J.*, **21**, e837.
8. Bardet,A.F., Steinmann,J., Bafna,S., Knoblich,J.A., Zeitlinger,J. and Stark,A. (2013) Identification of transcription factor binding sites from ChIP-Seq data at high resolution. *Bioinformatics*, **29**, 2705–2713.
9. Planet,E., Attolini,C.S.-O., Reina,O. and Rosell,D. (2011) htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics*, **28**, 589–590.
10. Landt,S., Marinov,G., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B., Bickel,P., Brown,J., Cayting,P. *et al.* (2012) ChIP-Seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
11. Marinov,G.K., Kundaje,A., Park,P.J. and Wold,B.J. (2014) Large-Scale Quality Analysis of Published ChIP-seq Data. *G3: Genes Genomes Genetics*, **4**, 209–223.
12. Hansen,P., Hecht,J., Ibn-Salem,J., Menkuec,B.S., Roskosch,S., Truss,M. and Robinson,P.N. (2016) Q-nexus: a comprehensive and efficient analysis pipeline designed for ChIP-nexus. *BMC Genomics*, **17**, 1–15.
13. Qin,Q., Mei,S., Wu,Q., Sun,H., Li,L., Taing,L., Chen,S., Li,F., Liu,T., Zang,C. *et al.* (2016) ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. *BMC Bioinformatics*, **17**, 1–13.
14. Langmead,B., Trapnell,C., Mihal,P. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, 1–10.
15. Kuan,P.F., Chung,D., Pan,G., Thomson,J.A., Stewart,R. and Keleş,S. (2009) A statistical framework for the analysis of ChIP-Seq data. *J. Am. Stat. Assoc.*, **106**, 891–903.
16. Chung,D., Park,D., Myers,K., Grass,J., Kiley,P., Landick,R. and Keleş,S. (2013) dPeak: high resolution identification of transcription factor binding sites from PET and SET ChIP-Seq data. *PLoS Comput. Biol.*, **9**, e1003246.
17. The ENCODE Project Consortium. (2011) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
18. Grant,C., Bailey,T. and Noble,W.S. (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
19. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
20. Serandour,A., Gordon,B., Cohen,J. and Carroll,J. (2013) Development of and Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biol.*, **14**, 1–9.
21. Starick,S.R., Ibn-Salem,J., Jurk,M., Hernandez,C., Love,M.I., Chung,H.-R., Vingron,M., Thomas-Chollier,M. and Meijsing,S.H. (2015) ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Res.*, **25**, 825–835.
22. Venters,B.J. and Pugh,F. (2013) Genomic organization of human transcription initiation complexes. *Nature*, **502**, 53–58.
23. Rhee,H.S., Bataille,A., Zhang,L. and Pugh,F. (2014) Subnucleosomal structures and nucleosome assymetry across a genome. *Cell*, **159**, 1377–1388.
24. Benjamin,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, 1–14.
25. Valouev,A., Johnson,D., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Meth.*, **5**, 829–834.
26. Kharchenko,P., Tolstorukov,M. and Park,P. (2008) Design and analysis of ChIP-Seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
27. Carroll,T., Liang,Z., Salama,R., Stark,R. and de Santiago,I. (2014) Impact of artifact removal on ChIP quality metrics in ChIP-Seq and ChIP-exo data. *Front. Genet. Bioinformatics Comp. Biol.*, **5**, 75.
28. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.