STATISTICAL METHODS FOR CHIP-EXO/NEXUS QUALITY CONTROL, AND FINE-MAPPING OF MULTI-TRAIT SNPS IN HIGH LD BY USING NEXT-GENERATION SEQUENCING DATA

by

Rene Welch

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2018

Date of final oral examination: 15/08/18

The dissertation is approved by the following members of the Final Oral Committee:
Sündüz Keleş, Professor, Statistics and Biostatistics and Medical Informatics
Karl Broman, Professor, Biostatistics and Medical Informatics
Colin Dewey, Associate Professor, Computer Sciences and Biostatistics and Medical Informatics
Christina Kendziorski, Professor, Biostatistics and Medical Informatics
Qiongshi Lu, Assistant Professor, Biostatistics and Medical Informatics

To Nadya, nothing of this would have been possible without her.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Professor Sündüz Keleş who has guided me through my Ph.D. studies with her invaluable advice, her great patience, and kind support. Her passion for research and personal example have been a motivation through my PhD studies. Without her guidance and direction, this dissertation and my personal achievement in graduate school would not have been possible.

I thank my thesis committee members, Professor Karl Broman, Professor Colin Dewey, Professor Christina Kendziorski, and Professor Qiongshi Lu, for their reading of my dissertation and providing valuable comments. I would like to thank Professor Eric Johannsen, who generated interesting genomic data and shared sharp biological insights. It was great experience for me to work with his lab. I would also like to thank him for his time and valuable suggestions on my development as statistical practitioner.

I am grateful for all previous and current members in Keleş Research Group, who provided academically stimulating and exciting discussions on statistics, and computational biology. I want to thank friends in the Department of Statistics for their companion, emotional support, and fun times.

I would like to thank my partner, Nadya, who has been by my side for a long time sharing happy and hard times alike. It is her love and support that motivated me to work hard and grow every day.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Advances in the field of genomics over the last years have led to diverse types of measurements with vast amounts of next generation sequencing (NGS) data. As genomic measurements become faster and cheaper with higher-throughput technologies, larger quantities of data are being generated, and awaiting novel insights and discoveries. Multiple NIH-funded large consortia projects such as the Encyclopedia of DNA Elements, the Roadmap Epigenomics Mapping Consortium, and the Genotype-Tissue Expression Project have generated a myriad of data types of DNA accessibility (DNAse-seq and ATAC-seq), protein-DNA interaction (ChIP-seq, and ChIP-exo/nexus), RNA transcription (RNA-seq), and DNA methylation (Methyl-seq) among many other assays.

In this thesis we introduce two methods. First, we present `ChIPexoQual`, an `R/Bioconductor` quality control pipeline that enables 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 modeling ChIP-exo data.

Next, we introduce `FM-HighLD`, a fine-mapping method for variants in high linkage disequilibrium that, by using ChIP-seq based annotation data, is capable of calculating SNP probabilities for being causal for single or multiple traits by using annotation data. We provide a simple pipeline to calculate the allelic skew, a measure of allele-specific transcriptional activity calculated with in-vivo sequencing data. We evaluated `FM-HighLD` computationally by data-driven simulations, and to the best of our knowledge used for the first time Massively Parallel Reproducible Assays (MPRA) as gold-standards with which to evaluate causal predictions of eQTL

associations in lymphoblastoid cell lines of variants in high LD. These simulations resulted in a number of key findings regarding FM-HighLDs' performance. First, FM-HighLD outperforms state-of-the-art fine-mapping methods when the causal variants are randomly selected from the variants deemed as regulatory by Tewhey et al. Second, FM-HighLD exhibits higher recall rate when fewer causal variants are considered in a polygenic model. Finally, when the causal variants are selected in LD clusters composed by SNPs in high LD, FM-HighLD outperforms PAINTOR, and is comparable to CAVIAR.

Finally, in this thesis, we showcase two R/Bioconductor packages to facilitate the analysis of next generation sequencing data, and explore RNA-seq differential expression results. First, ChIPUtils is a package that calculates typical ChIP-seq quality control metrics, and generates a few visualizations to explore the data. Second, Segvis is a package for computing the genomic coverage of high throughout sequencing data along genomic segments, and allows the user multiple visualizations across different samples and peak sets. In other words, this package is capable of generating multiple visualization summarizing the binding patterns across individual or multiple peaks of different samples aligned to the genome.

1 INTRODUCTION

Advances in the field of genomics over the last years have led to diverse types of measurements with vast amounts of next generation sequencing (NGS) data. As genomic measurements become faster, cheaper, and more high-throughput, larger quantities of data are being generated, and awaiting novel insights and discoveries. Multiple-NIH funded large consortia projects such as the Encyclopedia of DNA Elements (The ENCODE Project Consortium (2012)), the Roadmap Epigenomics Mapping Consortium (Consortium et al. (2015)), and the Genotype-Tissue Expression Project (GTEx Consortium (2017); eGTEx Project (2017)) have generated diverse data types of DNA accessibility (DNAse-seq and ATAC-seq), protein-DNA interaction (ChIP-seq, and ChIP-exo/nexus), RNA transcription (RNA-seq), and DNA methylation (Methyl-seq) among many other assays.

Chromatin Immunoprecipitation coupled with next generation sequencing (ChIP-seq) has been the most popular high-throughput DNA-protein interaction profiling technology for the last years. In ChIP experiments, the signal is collected from protein binding fragments extracted with an antibody and then aligned to a reference genome after amplification and sequencing. This assay reveals locations in the genome with enriched activity of a protein of interest. Chromatin Immunoprecipitation followed by exonuclease digestion and next generation sequencing (ChIP-exo, Rhee and Pugh (2011)) and Chromatin Immunoprecipitation followed by exonuclease digestion, unique barcode, single ligation and next generation ligation (ChIP-nexus He et al. (2014)) are currently state-of-the-art high throughput assays for profiling protein-DNA interactions at a close to single base-pair resolution (Rhee and Pugh (2011),He et al. (2014)). Both present powerful alternatives to the popular ChIP-seq assay. ChIP-exo/nexus 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 is a further modification on the ChIP-exo protocol. ChIP-nexus aim 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 number of ligations in the standard ChIP-exo protocol from two to one, and adding unique, randomized barcodes to adaptors to enable monitoring of over-amplification. Both assays are expected to archive higher resolution when localizing regions in the genome enriched by a protein of interest.

In Chapter 2 we present `ChIPexoQual`, an `R/Bioconductor` quality control pipeline that enables 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 modeling ChIP-exo data.

Genome-wide association studies (GWAS) have detected thousands of robust and replicable genetic associations with multiple phenotypes. While a large number of variants reproducibly associate with several traits, these variants often map larger regions in linkage disequilibrium (LD). Therefore, the association between the most associated SNP and a trait can be indirect, because the lead SNP can be in high LD with the actual causal SNP for the trait. One implication of this observation is that when the correlation between two variants is close to one, it is statistically impossible to discern which SNP is the causal one. For that reason, it is imperative to include additional genomic information in the form of functional annotation data or eQTL data (Nica et al. (2010); Nicolae et al. (2010); Schaub et al. (2012); Grossman et al. (2013)). Recently, Massively Parallel Reporter Assays (MPRAs) have been modified to identify alleles that impact gene expression (Tewhey et al. (2016);

Ulirsch et al. (2016)). These assays allow to detect with increased sensitivity which variants modulate gene expression. This allows to experimentally detect the causal variants by the use of a high throughput assay.

The second chapter of this thesis presents a statistical method for detecting causal SNPs in a genomic region with highly correlated variants called `FM-HighLD`. Our method relies on the use of annotation data accompanied by association statistics and the LD matrix to discern the causal variants for both single and multiple traits. `FM-HighLD` explains the associations between summary statistics of causal SNPs and annotation data by selecting the causal variants and estimating this relationship simultaneously. We compared our model with the state-of-the-art fine-mapping methods `CAVIAR` (Hormozdiari et al. (2014)) and `PAINTOR` (Kichaev et al. (2014)) by using simulations, and analyzing eQTL data. We evaluated our eQTL analysis by using Massive Parallel Reporter Assay (MPRA) results as gold-standards, which represents the first application of MPRA in this setting. Chapter 3 includes more details of our approach, simulation results and case study results.

In Chapter 4 we introduce two R packages, `ChIPUtils` and `Segvis`, designed for the quick and easy computation of ChIP-seq quality control metrics, and visualization of high throughout sequencing data along genomic segments, respectively. The first package was developed simultaneously with `ChIPexoQual`. It calculates several quality control metrics without the need to install multiple software packages, which often depends on different programming languages. The second package was motivated by multiple projects where it was required to summarize the average ChIP-seq signal of multiple histone samples across a set of same-width peaks. We used `Segvis` to compare the average histone signal between peaks called with uniquely mapping reads and peaks called including multi-mapping reads (Zeng et al. (2015)), and to compare histone binding across EBNA3 ChIP-seq peaks (Wang et al. (2016)).

Finally, in Appendix A we end with a discussion about the development of `shiny` applications for the analysis of differentially expressed genes measured with RNA-seq data. We developed this application to allow our collaborators to explore results of a statistical analysis interactively. By sharing interactive documents, we

believe that we are enabling the discovery of future hypothesis.

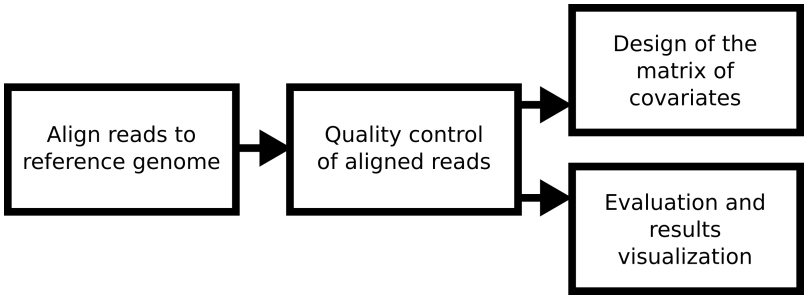

Figure 1.1: **Example diagram of a ChIP data pipeline.** After the ChIP data is generated. The first step is to align the reads to a reference genome. Next, we quantify the samples' quality. Finally, we perform two steps simultaneously: Design a matrix of features based on either the aligned reads or the results of another method (e.g. a peak caller), and visualize the aligned reads, and results of that analysis.

This dissertation is divided in three parts. In the first part, we introduce a quality control pipeline for ChIP-exo/nexus data. In this section we evaluate a diverse collection of ChIP-exo/nexus samples with the most frequently used ChIP-seq quality control metrics. In the second part, we present a statistical method for detecting single nucleotide polymorphisms (SNPs) in genomic regions with highly correlated variants. Finally in the last part, we introduce two `R` packages for the quick analysis and visualization of ChIP-seq/exo/nexus data. Figure 1.1 illustrates an example of a workflow followed when analyzing ChIP samples: Provided the ChIP samples are already aligned to a reference genome, we evaluate the samples with multiple quality control metrics with the objective of selecting a set of higher-quality samples for further analysis. The next step varies according to the objectives of the analysis, but typically it would be necessary to design a feature matrix with the aligned reads or with the results of other analysis, where the rows are based on a group of already pre-defined experimental units (e.g. regions of interest in the genome, genes, etc.) and the columns are different samples, and to visualize that matrix.

In Chapter 2 we introduce a quality control pipeline for ChIP-exo/nexus data, which is expected to be an improvement over the popular ChIP-seq assay. Next, in Chapter 3 we introduce a fine-mapping method to distinguish causal variants in high LD by the use of regulatory information, which is represented by a matrix where the rows are SNPs and the columns correspond one or multiple ChIP experiments. Finally, in Chapter 4 we showcase a package to calculate quality control metrics for ChIP-seq data, and a package to visualize high throughout sequencing data along genomic segments. The matrices matrices generated with high quality data (quantified by the QC metrics discussed in Chapters 2 or 4) could be explored by the use of `Segvis` or similar shiny applications to the one that is illustrated in Appendix A.

2 DATA EXPLORATION, QUALITY CONTROL AND STATISTICAL ANALYSIS OF CHIP-EXO/NEXUS EXPERIMENTS

This work was published in Nucleic Acids Research, 2017[1]

2.1 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 a close to single base-pair resolution (Rhee and Pugh (2011)). It presents a powerful alternative to the 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, He et al. (2014)) is a further modification of 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 number of ligations in the standard ChIP-exo protocol from two to one, and adding unique, randomized barcodes to adaptors to enable monitoring of over-amplification. In addition to these, several other high-resolution protocols have also been considered. In X-ChIP and ORGANIC (Kasinathan et al. (2014); Skene and Henikoff (2015)), 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

[1]Joint work with Dongjun Chung, Jeffrey Grass, Robert Landick and Sündüz Keleş, (Welch et al. (2017)).

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 adopted high-resolution protocol.

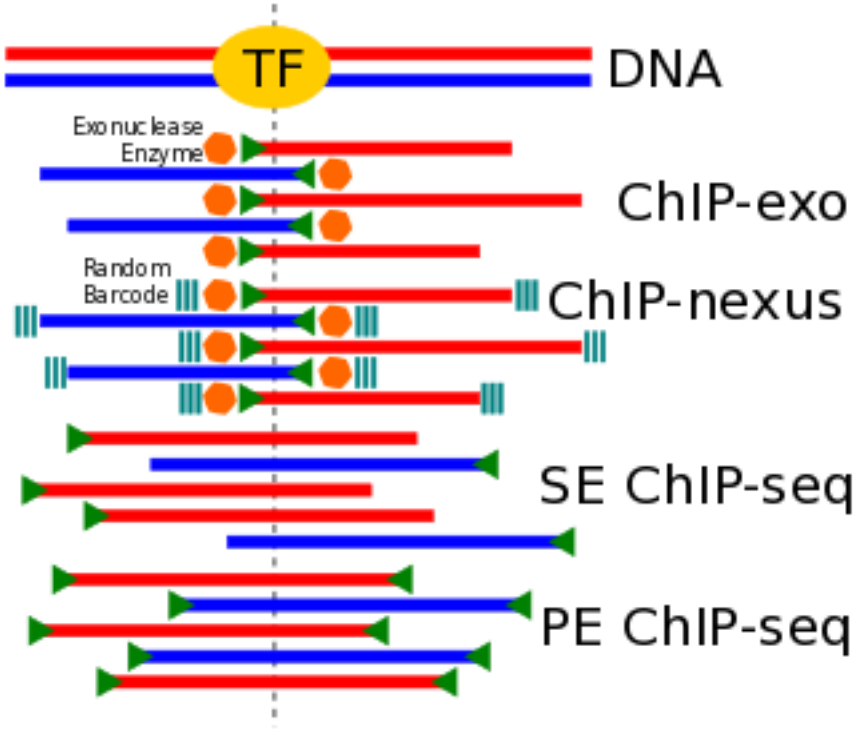

Figure 2.1: **Processing of sonicated fragments bound by TF before immunoprecipitation and PCR amplification.** For ChIP-exo, an exonuclease enzyme (orange hexagon) trims the 5′ ends of each DNA fragment to a fixed distance from the TF. For ChIP-nexus, a random barcode is added on the 3′ end, and transferred 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.

Figure 2.1 illustrates the differences between distinct ChIP-based protocols: ChIP-exo, ChIP-nexus, single-end (SE) ChIP-seq, and paired-end (PE) ChIP-seq. The 5′ ends from 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 (Mahony and Pugh (2015)) 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 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 in the statistical analysis of ChIP-exo data. Specifically, currently available ChIP-exo specific statistical analysis methods (e.g. MACE [Bardet et al. (2013)], CexoR [Madrigal (2015)] and Peakzilla [Wang et al. (2014)]) rely on the existence of peak-pairs formed by forward and reverse strand reads at the binding site. Finally, most of the current widely used ChIP-seq quality control (QC) guidelines (Landt et al. (2012); Marinov et al. (2014); Planet et al. (2011)) may not be directly applicable to ChIP-exo data.

2.2 Methods

We implemented our proposed QC pipeline with an **R/Bioconductor** package named `ChIPexoQual`.

`ChIPexoQual`: The method requires a set of N aligned reads from a ChIP-exo (or ChIP-nexus) experiment and performs the following steps:

1. Identify read islands, i.e., overlapping clusters of reads separated by gaps, from read coverage. The gaps are defined as the union of positions in the genome with fewer than h^* (by default $= 1$) aligned reads. The remaining islands can be interpreted as the natural partition of the genome determined by a ChIP-exo/nexus experiment.

2. Compute D_i, the number of reads in island i; U_i, the number of positions in island i; and W_i, the width of island i defined as the total number of bases in the islands, $i = 1, \cdots, I$.

3. For each island i, $i = 1, \cdots, I$ compute island statistics:

$$\begin{aligned} \text{ARC}_i &= \frac{D_i}{W_i} \\ \text{URC}_i &= \frac{U_i}{D_i} \\ \text{FSR}_i &= \#\text{ of fwd. strand reads aligning to island } i/D_i \end{aligned}$$

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 relationship depicted in the URC vs. ARC plot:

$$\text{URC}_i = \frac{\kappa}{\text{ARC}_i} + \gamma + \varepsilon_i$$

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:

1. *Subsampling analysis*. For high depth datasets (e.g. $\geqslant$ 60M reads for human and mouse samples), we subsample $N_1 < N_2 < \cdots < N$ reads, starting with $N_1 = 20$M reads and up to 50M reads in 10M increments as default, and apply steps 1 to 5 for each of the subsampled datasets.

2. *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 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:

1. FoxA1 experiments: we removed the islands with (i) reads residing only on one strand; (ii) $U_i \leqslant 15$; (iii) $D_i \leqslant 100$.

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

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 $3 \times \texttt{rl}$, where `rl` is the median read length of the experiment. 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 (He et al. (2014)).

We used `FIMO` (version 4.9.1,Grant et al. (2011)) 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 (Mathelier et al. (2016)), 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.

2.3 Results

We utilized a rich collection of publicly available ChIP-exo/nexus data from multiple organisms to build and evaluate our quality control pipeline (Table 2.2). These include: CTCF factor in human HeLa cell lines (Rhee and Pugh (2011)); ER factor in human MCF-7 cell lines (Serandour et al. (2013)); GR factor in IMR90, K562 and U2OS cell lines (Starick et al. (2015)); TBP factor in K562 cell lines (Venters and Pugh (2013)); H3 histone in *S. cerevisiae* where most, but not all of the tail was deleted ($\Delta 1 - 28$) (Rhee et al. (2014)). ChIP-nexus data included experiments from He et al. (2014) 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 2.1), SE and PE ChIP-seq. Collectively, we generated σ^{70} factor ChIP-exo, PE and SE ChIP-seq experiment under aerobic ($+O_2$) and anaerobic ($-O_2$) conditions in glucose minimal media. For simplicity, we named these experiments as E1, P1 and S1, respectively. Similarly, we generated σ^{70} factor ChIP-exo, and PE ChIP-seq experiment 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 2.2 and 2.3 summarize this comparison for one biological replicate of ChIP-exo and ChIP-seq experiments from the same biological conditions (ChIP-exo samples E1-1 from Table 2.1, PE ChIP-seq P1-1 and SE ChIP-seq S1-1 following the same convention).

Peak-pair assumption. We evaluated the peak-pair assumption, i.e., that 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-size of the binding site in the reverse strand. This observation is commonly utilized in designing statistical analysis methods for ChIP-exo data (Bardet et al. (2013); Madrigal (2015); Wang et al. (2014)). We considered the set of peaks identified in both the ChIP-seq and ChIP-exo samples as high quality peaks and calculated the proportions of forward strand reads (Figure 2.2). 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-strand protein-DNA interactions. Overall, such an imbalance is more likely to occur in low complexity libraries.

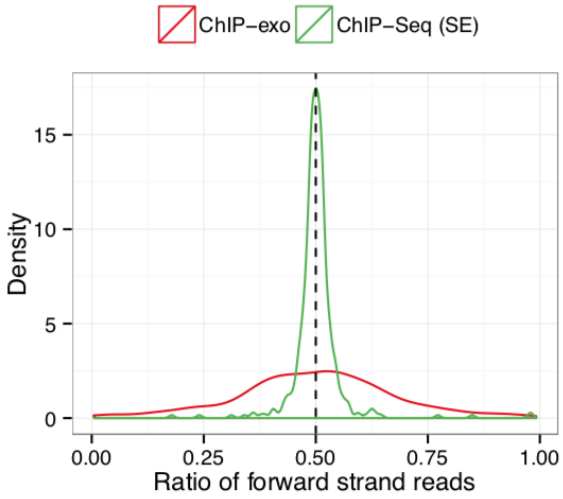

Figure 2.2: **Forward Strand Ratio densities for SE ChIP-seq and ChIP-exo peaks.**

Read distribution within signal and background regions. Using extended raw read counts within 150 bp non-overlapping intervals, i.e., bins interrogating the genome, Figure 2.3A 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 (data now shown).

Mappability and GC-content bias. We next evaluated ChIP-exo data of CTCF in HeLa cells (Rhee and Pugh (2011)) to investigate biases inherent to next generation sequencing experiments with eukaryotic genomes. Figures 2.3B and 2.3C display the bin-level average read counts against mappability and GC-content (Rozowsky et al. (2009) and Benjamin and Speed (2011), respectively). Each data point is obtained by averaging the read counts across bins with the same mappability or 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 (Benjamin and Speed (2011); Kuan et al. (2011); Valouev et al. (2008)). 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. The Mappability and GC-content are defined for the base i of the genome as:

$$m_i = \sum_{k=i-L+1}^{i+L-1} \frac{\delta_k}{2L+1} \quad g_i = \sum_{k=i-L+1}^{i+L-1} \frac{\tilde{g}_k}{2L+1}$$

where δ_k represents if nucleotide k can be mapped uniquely by a sequence of length L bp sequence starting at position k, and $\tilde{g}_k$ represents the occurrence of G or C at k-th position in a sequence of length L.

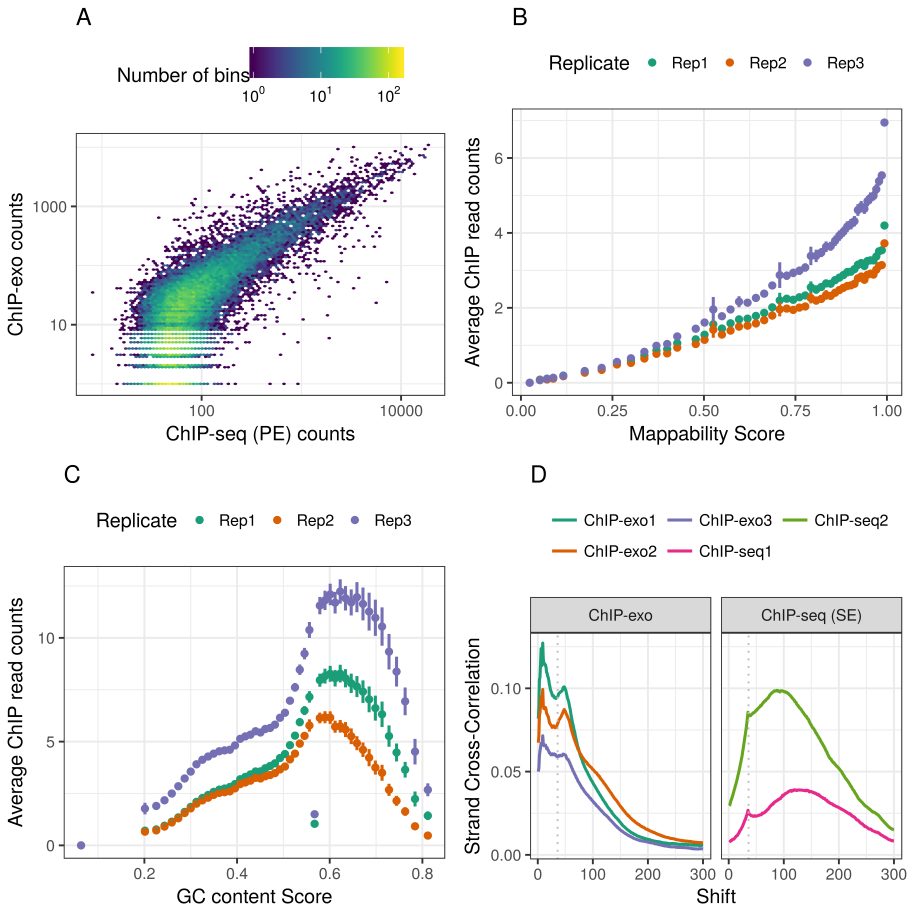

Figure 2.3: **ChIP-seq vs. ChIP-exo.** A) Hexbin plot of PE ChIP-seq bin counts vs. ChIP-exo bin counts. B) Mappability score vs. mean ChIP-exo read counts with error bands. C) GC-content vs. mean ChIP-exo read counts with error bands. D) SCC curves for human CTCF from HeLa cell lines. The ChIP-exo curve 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.

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 was raised for sequence duplication levels and k-mer

content representation, 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 λ-enzyme is expected to stop its digestion when it reaches the cross-linking protein.

The ENCODE consortium established empirical and widely used QC metrics on ChIP-seq data (Landt et al. (2012)). We evaluated how these metrics performed when applied to ChIP-exo/nexus samples, 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` (Landt et al. (2012); Marinov et al. (2014)). Tables 2.1 and 2.2 present these metrics for the collection of ChIP-exo/nexus datasets we consider in this paper.

The PBC is defined as the ratio of the number of genomic positions to which exactly one read maps and the number of genomic positions to which at least one read maps. The interpretation of this metric is affected by the reduced amount of possible genomic position to which the reads can align. Marinov et al. (2014) 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 number of positions to which the reads can align to are reduced due to the exonuclease digestion. For ChIP-seq samples, low PBC values (e.g. $\leqslant 0.5$) 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. (2011) presented in the R/Bioconductor package `htSeqtools` the Standarized Standard Deviation (SSD) as a metric to assess enrichment efficiency and to compare across samples. According to these guidelines established by the authors, higher values of this metric indicate high-quality. 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 samples 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 (samples 2 and 3) exhibit higher quality than the first ChIP-nexus replicate (data not shown). 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. (2008), 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 location of the protein-DNA interaction sites by calculating the Pearson correlation between forward and reverse strand 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 expect to maximize at the motif length indicating clustering of the forward and reverse strands around the binding site. This hinders the interpretation of the SCC for a ChIP-exo/nexus experiment since it is now maximized at an unobserved shorted fragment length that is confounded with the *"phantom peak"* at the read length. Carroll et al. (2014) 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 and suggested to calculated 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 2.3D 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. In a ChIP-exo experiment, the read length and the fragment length peaks in the SCC are confounded, as there are two local maxima very closed to each other. Furthermore, the former is close in proximity to the motif length; as a result, this may incorrectly suggest experiment to be marginally successful or even failed and renders QC metrics such as the Normalized Strand Cross-Correlation (NSC) or the

Relative Strand Cross-Correlation (RSC) harder to interpret. However, in the majority of the cases we examined, the profile itself seems informative about the enrichment signal in ChIP-exo/nexus experiments. These metrics are defined as:

$$\text{NSC} = \frac{\max_\delta \text{SCC}(\delta)}{\min_\delta \text{SCC}(\delta)}, \quad \text{RSC} = \frac{\max_\delta \text{SCC}(\delta) - \text{SCC}(\text{rl})}{\min_\delta \text{SCC}(\delta) - \text{SCC}(\text{rl})}$$

Details about the software we used to compute them are available in Chapter 4.

Group	Growth	Treatment	Rep.	Id.	Depth	NSC	RSC	PBC
ChIP-exo (E1)	Exp. $+O_2$	No Rif	1	1	13.9M	103.15	2.01	0.13
	Exp. $+O_2$	No Rif	2	2	14.8M	162.70	1.78	0.16
	Stat. $+O_2$	No Rif	1	3	16.1M	153.51	1.80	0.13
	Stat. $+O_2$	No Rif	2	4	13.6M	172.59	2.01	0.15
ChIP-exo (E2)	Exp. $+O_2$	No Rif	1	1	0.9M	13.77	1.12	0.26
	Exp. $+O_2$	Rif 20 min	1	2	1.8M	17.91	1.52	0.25
	Exp. $+O_2$	No Rif	2	3	2.1M	29.60	1.28	0.25
	Exp. $+O_2$	Rif 20 min	2	4	11.5M	13.08	1.51	0.15

Table 2.1: **Summary of the *E. coli* σ^{70} ChIP-exo.** Exp. stands for exponential and Stat. for stationary growth conditions. The last three columns depict ENCODE QC metrics on these data: NSC: Normalized Strand Cross-Correlation; RSC: Relative Strand Cross-Correlation; PBC: PCR Bottleneck Coefficient.

Seq.	Genome	TF	Cell	Rep.	Depth	NSC	RSC	PBC
ChIP-exo	hg19	CTCF	HeLa	-	48.4M	16.02	1.19	0.45
	hg19	ER	MCF-7	1	9.2M	19.87	1.01	0.80
				2	11.0M	21.48	1.00	0.80
				3	12.4M	18.72	1.01	0.82
	mm9	FoxA1	Liver	1	22.2M	21.28	1.11	0.65
				2	23.3M	60.42	1.16	0.79
				3	22.4M	72.04	1.19	0.10
	hg19	GR	IMR90	1	47.4M	8.86	1.36	0.29
			K562	1	116.5M	4.11	1.04	0.05
			U2OS	1	3.2M	10.05	1.02	0.77
	hg19	TBP	K562	1	61.0M	12.01	1.11	0.12
				2	94.3M	7.93	1.02	0.16
				3	114.2M	9.25	1.10	0.14
ChIP-nexus	dm3	Dorsal	embryo	1	8.8M	7.27	1.04	0.67
				2	10.0M	7.19	1.06	0.56
		Twist		1	18.2	5.82	1.16	0.65
				2	52.5	5.27	1.18	0.45
		Max	S2	1	18.3M	3.60	1.36	0.51
				2	24.9M	3.47	1.01	0.21
		MyC		1	7.8M	5.92	1.01	0.39
				2	22.8M	5.76	1.00	0.18
	hg19	TBP	K562	1	33.7M	32.16	1.17	0.31
				2	129.6M	32.70	1.24	0.04

Table 2.2: **Summary of publicly available data used for development and evaluation of `ChIPexoQual`.** The last three columns depict ENCODE QC metrics on these data: NSC: Normalized Strand Cross-Correlation; RSC: Relative Strand Cross-Correlation; PBC: PCR Bottleneck Coefficient.

ChIP-exo quality control pipeline `ChIPexoQual`

To address the limitations of available analytical exploration approaches discussed above, we developed `ChIPexoQual`. In Table 2.3, we compare `ChIPexoQual` against the existing tools previously discussed. 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 Serandour et al. (2013) and ChIP-nexus from He et al. (2014). Figure 2.4 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 2.4A 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 2.4B 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 the rate at which the proportion of islands exclusively composed of fragments on a single strand decreases as a function of the islands'

read depth. In a high quality sample, the proportion of islands with reads from only one strand is expected to decrease rapidly as we consider higher depth regions. In contrast, this proportion remains approximately constant in lower quality samples. The 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.

Aspect	`ChIPexoQual`	`ChiLin`	`ChIPQC`	`phantompeakQTS`	`htSeqTS`	`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	✓	✓	✓	✓		✓	✓

Table 2.3: **Comparison of the state-of-the-art quality control tools for ChIP-seq and ChIP-exo/nexus samples.** `phantompeakQTS` stands for `phantompeakQualTools` and `htSeqTS` stands for `ht-SeqTools`.

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, we processed the FoxA1 datasets with the following thresholds 1, 5, 25 and 50. 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 (Supplementary Figure B.1 to B.3). Hence, we decided to use the default threshold of $\mathtt{h}^* = 1$ when analyzing the FoxA1 samples.

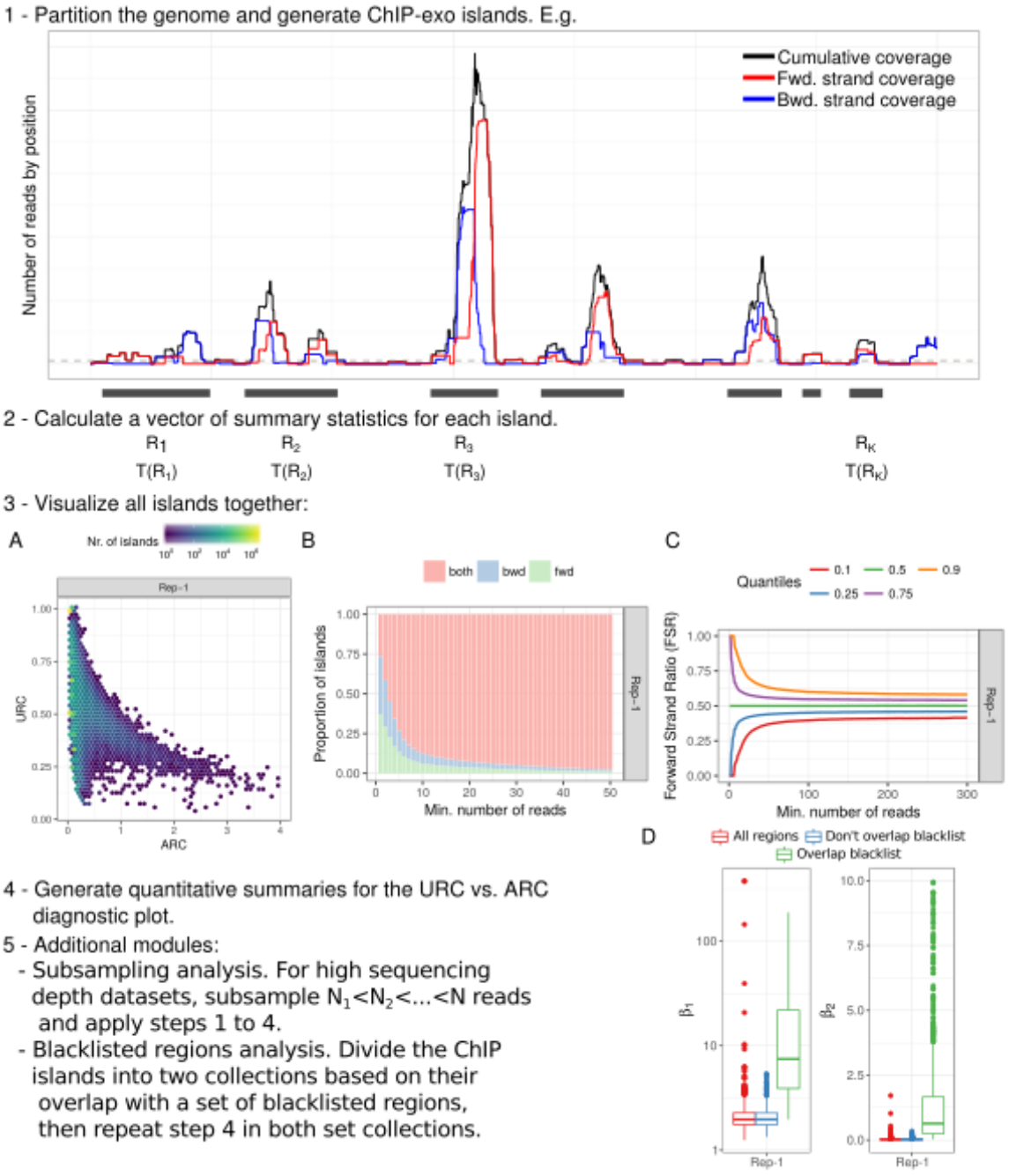

Figure 2.4: **ChIP-exo QC pipeline `ChIPexoQual`.** The ChIP-exo reads are partitioned into overlapping clusters of reads separated by gaps (step 1). Fr each region, the following summary statistics are calculated (step 2) and visualized (step 3): Average Read Coefficient (ARC), Unique Read Coefficient (URC) and Forward Strand Ratio (FSR). These statistics are visualized as: (A) URC versus ARC plot, (B) Region Composition plot, (C) FSR distribution plot, (D) Example of the Blacklisted region analysis module.

Figure 2.5A 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, indicating that the libraries for the first two replicates are more complex than that of the third replicate.

Figures 2.5B and C display the Read Composition and FSR distribution plots, which highlight specific problems with replicates 2 and 3. Figure 2.5B 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. FSR distributions of both of replicates 2 and 3 are more spread around their respective medians (Figure 2.5C). 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 2.5D 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.

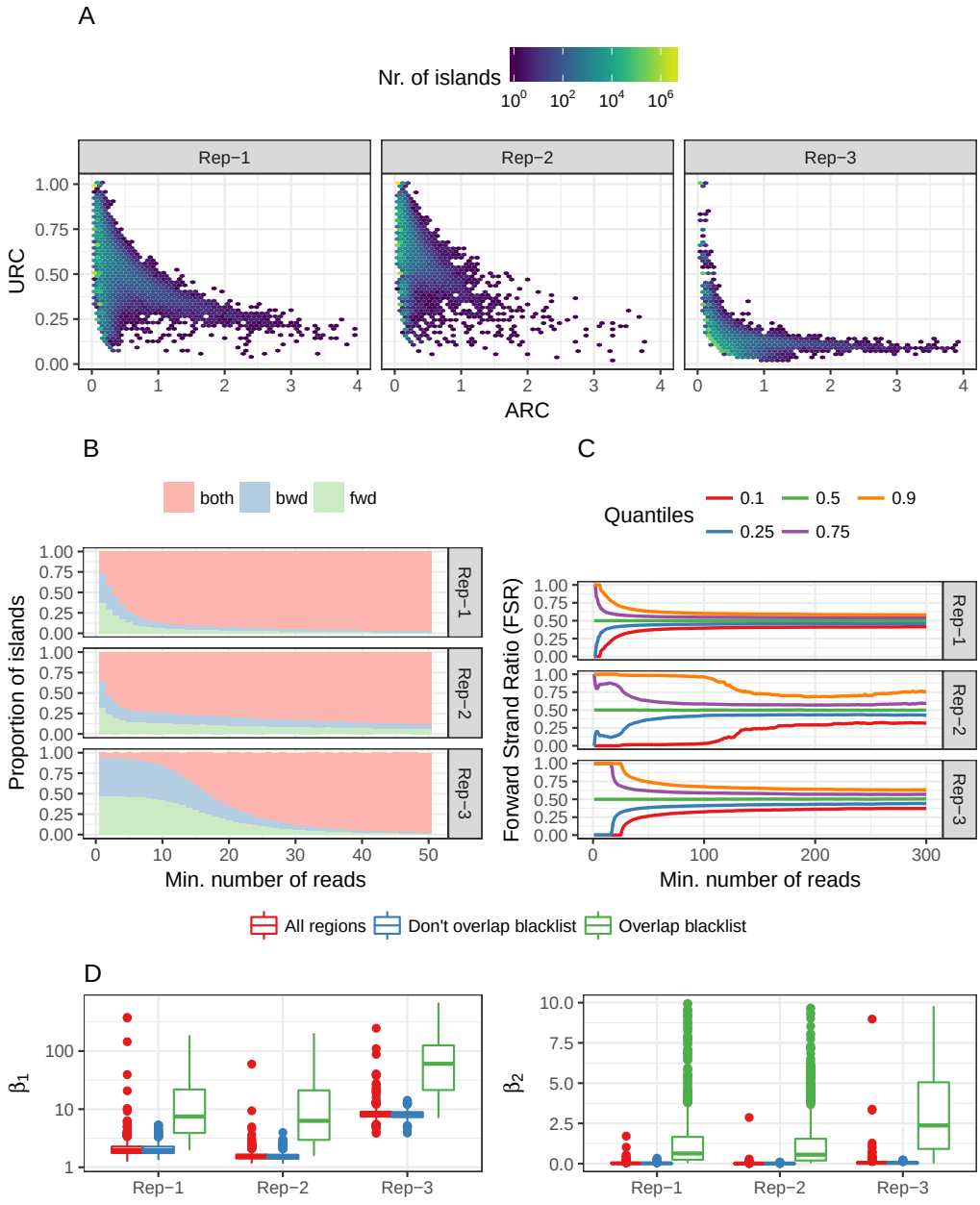

Figure 2.5: `ChIPexoQual` **diagnostic plots for the FoxA1 ChIP-exo data.** A) **URC** vs. **ARC** plot, B) Region Composition plot, C) **FSR** distribution plot comparison across three replicates and D) β_1 and β_2 scores stratified based on overlap with the blacklisted regions. In red, blue and green, we observe the β_1 and β_2 score distributions when calculated by sampling the islands among the set of all islands, the islands that didn't overlap a set of predefined blacklisted regions, and the islands that overlap the set of predefined blacklisted regions, respectively.

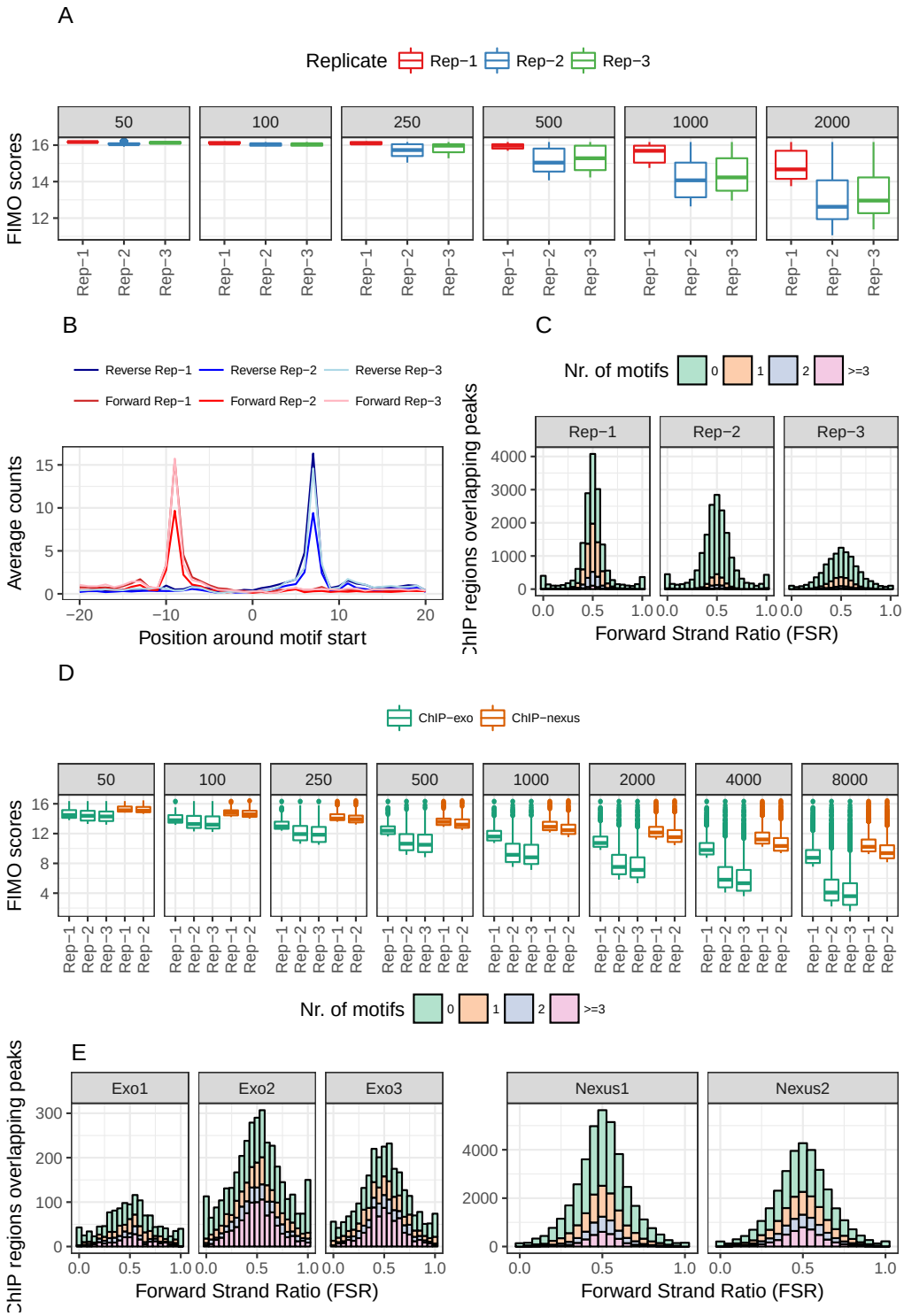

Figure 2.6: **Validation of `ChIPexoQual` pipeline with FoxA1 ChIP-exo.** (A-C) and TBP ChIP-exo/nexus (D-E). A) Comparison of the top 50 to 2000 FIMO scores for each replicate. B) FoxA1 mean coverage plots of the 5′ read ends centered around motif start positions separated by replicate and strand. C) FoxA1 FSR distribution of `ChIPexoQual` islands overlapping ChIP-exo peaks stratified by the number of motifs. D) Comparison of the top 50 to 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.

We conclude that replicate 1 is of 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 (Section 2.2) 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 2.6A compares the FIMO scores among the three replicates, not surprisingly confirming that the first replicate exhibits the highest quality.

Figure 2.6B 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 2.6C 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 2.1 and 2.2 with the `ChIPexoQual` QC pipeline. 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 ~ 60M to 115M reads, and replicate two of TBP ChIP-nexus in K562, with a sequencing depth of ~ 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 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 samples exhibit the pattern specific to high quality samples

(Supplementary Figure B.4). To confirm this implication, we compared the top FIMO scores (Grant et al. (2011))of the TBP motif for the ChIP-exo and ChIP-nexus replicates. Figure 2.6D 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.

Figure 2.6E 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 2.1 and 2.2) 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., $D_i = \beta_1 U_i + \beta_2 W_i + \epsilon_i$, where ϵ_i represents the random error term. As we discuss in Materials and Methods, this parametrization has a direct connection to $URC_i = \gamma + \frac{\kappa}{ARC_i} + \epsilon_i$, which aims to recapitulate the ARC_i relationship in the URC vs. ARC plots. Figure 2.7A displays estimated overall change in depth (β_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 β_1 in Figure 2.7A 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, we conclude that samples with estimated β_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 2.7B. 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 2.7B), β_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 and second replicate from the K562 ChIP-nexus experiments. Figure 2.7C 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 2.7D illustrates that the β_2 estimates remain approximately constant in ChIP-nexus sub-samples and sub-samples of the first ChIP-exo replicate, while they increase for the second and third 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.

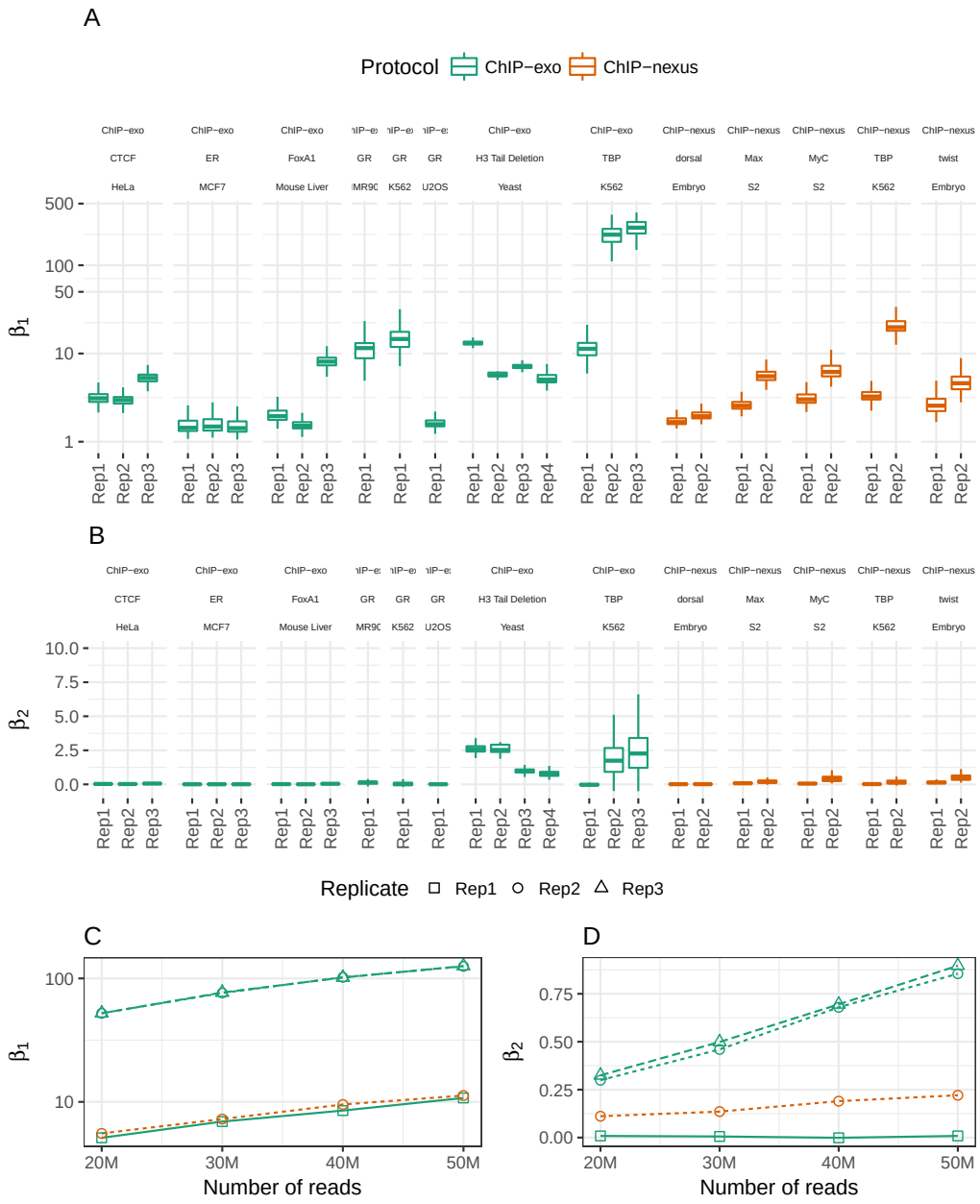

Figure 2.7: **Comparison of `ChIPexoQual` numerical summaries.** A) $\hat{\beta}_1$ and B) $\hat{\beta}_2$ for all eukaryotic ChIP-exo/nexus samples. C) Average estimated β_1 and D) β_2 for the ChIP-exo/nexus TBP samples in K562 cell lines when sub-sampling 20M to 50M reads.

2.4 Discussion

`ChIPexoQual` R package

We implemented `ChIPexoQual` as an R/Bioconductor package. `ChIPexoQual` utilizes a fast processing algorithm by parallel computing. Figure 2.8 provides `ChIPexoQual`'s processing times for a collection of samples representing different sequencing depths of the ChIP-exo/nexus experiments listed in Table 2.2 using four parallel threads on a server with 24 AMD 55Opteron 2.2GHz processors. This plot shows that `ChIPexoQual` requires between 125 and 640 seconds (80 and 420 when the aligned reads are already loaded into memory) for processing a ChIP-exo/nexus sample.

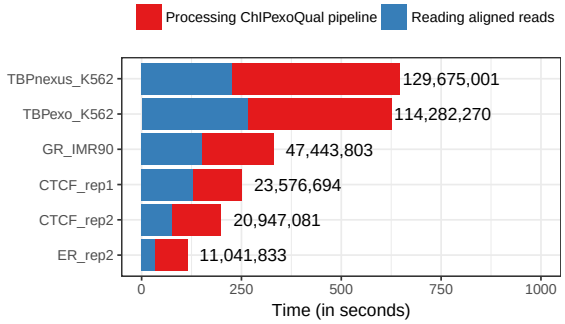

Figure 2.8: **Processing times for ChIP-exo/nexus samples representing different sequencing depths.**

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 is available from Bioconductor `http://bioconductor.org/packages/release/bioc/html/ChIPexoQual.html`.

3 MULTI-TRAIT FINE-MAPPING WITH INTEGRATED FUNCTIONAL ANNOTATION

3.1 Introduction

Genome-wide association studies (GWAS) have detected thousands of robust and replicable genetic associations with multiple phenotypes. While a large number of variants reproducibly associate with several traits, these variants often map larger regions in linkage disequilibrium (LD). Mounting evidence suggests that the majority of these variants are located in non-coding regions of the genome, thus it is necessary to understand the relationship of these variants with expression quantitative loci (eQTL), or functional annotation data (Nica et al. (2010); Nicolae et al. (2010); Schaub et al. (2012); Grossman et al. (2013)). Recently, Massively Parallel Reporter Assays (MPRAs) have been modified to identify alleles that impact gene expression (Tewhey et al. (2016); Ulirsch et al. (2016)). These assays allow to detect with increased sensitivity which variants modulate gene expression.

Multiple NIH funded large consortia projects such as the Encyclopedia of DNA Elements (The ENCODE Project Consortium (2012)), the Roadmap Epigenomics Mapping Consortium (Consortium et al. (2015)), and the Genotype-Tissue Expression Project (GTEx Consortium (2017); eGTEx Project (2017)) as well as individual-driven projects have generated diverse data types of DNA accessibility (DNAse-seq and ATAC-seq), protein-DNA interaction (ChIP-seq, and ChIP-exo/nexus), RNA transcription (RNA-seq), and DNA methylation (Methyl-seq) among many other assays. Several studies revealed that eQTL variants are associated with genetic variants that modify chromatin accessibility and TF binding (Degner et al. (2012); Maurano et al. (2012)). That is because TFs often cluster together on the same allele in regions of open chromatin (Reddy et al. (2012)).

Despite that several method had used different annotation data to solve a diverse array of problems as variant prioritization, multiple GWAS results integration with functional annotation, heritability partition based on GWAS summary statistics, variant interpretation with functional annotation (Chung et al. (2014); Finucane et al. (2015); Li and Kellis (2016); Shin and Keleş (2016); Reshef et al. (2017)), etc. Only `fgwas` (Pickrell (2014)), PAINTOR (Kichaev et al. (2014)), and `DAP` (Wen et al. (2016)) are fine-mapping methods that use summary statistics and annotation data to discern the causal variants (Schaid et al. (2018)).

In GWAS, the association between the most associated SNP and a trait can be indirect, because the lead SNP can be in high LD with the actual causal SNP for the trait. One implication of this observation is that when the correlation between two variants is close to one, is not statistically possible to discern which SNP is the causal one. Figure 3.1 shows an hypothetical example, where according to the test proposed by Udler et al. (2010) to distinguish two SNPs in LD, the population required increases exponentially as the correlation between them approaches one. When dealing with rare diseases, this becomes problematic as the cost of data collection can be very high for hard to reach populations, or in some cases the required number of patients may not even exist. Since LD patterns can be complex, fine-mapping methods are used to distinguish the causal SNP from the variants that are associated with it. For this work we focus on methods that use summary statistics, as the genotype data usually is not readily available or desirable to share. `CAVIAR` (Hormozdiari et al. (2014)) and `PAINTOR` (Kichaev et al. (2014); Kichaev and Pasaniuc (2015); Kichaev et al. (2016)) are two of the state of the art fine-mapping methods that use summary association statistics and the LD matrix to detect the causal variants. Although both methods use a multivariate normal model and latent variables to compute the causal-state probabilities for a group of SNPs, the main differences methods are that i) `PAINTOR` uses a logistic regression model to represent the association of the causal states and functional annotation data, and ii) `PAINTOR` is applicable to multi-trait problems as well. This model has modified to a Bayesian setting by integrating the SNP effects in `CAVIARBF` (Chen et al. (2015)), and to use stochastic shotgun search approximation in `FINEMAP` (Benner et al. (2016)). The first modification is currently is part of both the `CAVIAR` and `PAINTOR` software.

We introduce `FM-HighLD`, a single and multi-trait fine-mapping method for SNPs in high LD that uses *in-vivo* and/or sequence based annotation data. In contrast with `CAVIAR` or `PAINTOR` our method models the association statistics between the causal variants and one or multiple traits as function of annotation data, instead of assuming that the summary statistics of non-causal SNPs can be represented as linear combinations of the summary statistics of the causal SNPs. This is archived by considering the stronger assumption that after the causal variants are identified then the SNPs are independent, but the traits are not identically distributed. We evaluate the single-trait version of `FM-HighLD` by generating data-based simulations, and show that our method outperforms `PAINTOR` in a baseline and high LD scenario, while outperforming `CAVIAR` in the baseline scenario and performing comparably in the high LD settings. We carried out a comparison of `FM-HighLD` and `PAINTOR`

in multi-trait setting with eQTL fine-mapping analysis of eQTLs and utilized MPRA to benchmark the methods. We use multiple annotation datasets as allelic skew, overlapping peaks percentage in conjunction with sequence based or *in-vivo* features as peak overlaps. We demonstrate that our method is capable of detecting causal variants with increased accuracy.

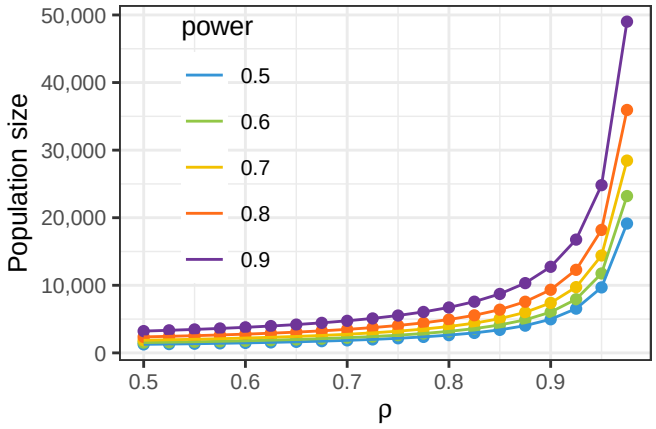

Figure 3.1: **Population size required to distinguish a causal SNP from its pair in LD.** We adapted the test proposed by Udler et al. (2010) to continuous traits. The population size required to distinguish the causal SNP among a pair of variants associated to a continuous trait increases exponentially as the correlation approaches one.

3.2 Methods

The FM-HighLD model.

`FM-HighLD` uses association statistics between variants and a single trait (or multiple traits) and annotation data to calculate the probability of variants being causal (Figure 3.2). `FM-HighLD` explains the associations between summary statistics of causal SNPs and annotation data by selecting the causal variants and estimating this relationship simultaneously.

Kreimer et al. (2017) summarized results from the Critical Assessments of Genome Interpretation (CAGI) eQTL challenge. First, they compiled the features used by all the

participants and classified them into 4 categories: Experimentally measured epigenetic properties, predicted epigenetic properties, locus properties, and k-mer frequencies. Second, they showed that classifying whether a variant is regulatory is significantly easier than classifying whether a regulatory variant is expression modulating. Finally, the highest ranked features for labelling regulatory variants are usually associated with open chromatin regions or the aggregation of multiple ChIP-seq experiments. This suggest that these features can be meaningful for building the annotation matrix.

Consider a genomic region with N regulatory variants clustered into M LD clusters of size q_m each. Lets assume that the N SNPs were tested for association with one or more traits, thus in the region there is a total of Q associations; in the single-trait case $N = Q$, and in the multi-trait case $N < Q$. Let $\mathbf{z} \in \mathbb{R}^Q$ denote the vector of z-scores for the Q associations from a single or multiple trait GWAS experiment. Let $\mathbf{A} \in \mathbb{R}^{Q \times p}$ denote an annotation score matrix for the SNPs tested, and without loss of generality assume that the entries of the vector $\mathbf{z}$ and matrix $\mathbf{A}$ are ordered by grouping the LD clusters together. `FM-HighLD` searches the combinatorial causal SNP space by minimizing the modified squared error function:

$$f(\mathbf{C}, \lambda) = \|\mathbf{C}^T(\mathbf{z} - \mathbf{A}\lambda)\|^2 \tag{3.1}$$

where $\mathbf{C}$ is the causal configuration matrix, which codes the causal variant in each LD cluster, i.e. groups of variants composed by the lead SNP and their correlated partners, and λ is the annotation coefficient vector. `FM-HighLD` is based on two assumptions:

1. If the SNP effects in a polygenic model are modeled as a function of the annotation data, then the association statistics of the causal variants are going to be functionally related to the annotation as well.

$$\mathbb{E}[\mathbf{C}^T\mathbf{z} \mid \mathbf{C}^T\mathbf{A}] = f(\mathbf{C}^T\mathbf{A}) + \epsilon$$

2. $\mathrm{Cov}(\epsilon) = \sigma^2 I$, i.e. once a causal SNP is selected per LD cluster, the rest of the correlations are ignored.

`FM-HighLD` assumes that every LD cluster can be represented by one of its members, thus the z-scores of those variants relate to the annotation scores under a linear model or a linear mixed model, for the single and multiple trait cases, respectively.

$$\text{Single trait: } \tilde{z}_m = \tilde{\mathbf{A}}_m \lambda + \epsilon_m, \quad m = 1, \cdots, M \tag{3.2}$$

$$\text{Multiple traits: } \tilde{z}_{tm} = \tilde{\mathbf{A}}_{tm} \lambda + \tilde{\mathbf{B}}_{tm} u_t + \epsilon_{tm}, \quad t = 1, \cdots, T, m = 1, \cdots, M_t \tag{3.3}$$

where:

- For the single-trait case:
  - In vector form: $\tilde{\mathbf{z}} = \mathbf{C}^T \mathbf{z}$ and $\tilde{\mathbf{A}} = \mathbf{C}^T \mathbf{A}$, where $\tilde{\mathbf{A}}_m$ corresponds to the annotation vector for the SNP selected as candidate causal.
  - M is the number of LD clusters.
  - $\epsilon_m \sim N(0, \sigma^2)$ for $m = 1, \cdots, M$.
- For the multi-trait case, for $t = 1, \cdots, T$:
  - In vector form $\tilde{\mathbf{z}}_t = \mathbf{C}_t^T \mathbf{z}_t$ and $\tilde{\mathbf{A}}_t = \mathbf{C}_t^T \mathbf{A}_t$, where $\tilde{\mathbf{A}}_{t,m}$ corresponds to the annotation vector for the SNP selected as candidate causal for trait t.
  - M_t is the number of active LD clusters for trait t.
  - $\epsilon_{tm} \sim N(0, \sigma^2)$ for $m = 1, \cdots, M_t$.
  - $u_t \sim N_h(0, \Sigma)$ is a vector of h random effects, with $\Sigma = \text{diag}(\sigma_1^2, \cdots \sigma_h^2)$.

Finally, let $w_m \sim \text{Ber}(\pi)$ and $w_{tm} \sim \text{Ber}(\pi)$ be indicators of the LD cluster m being causal in the single-trait case, or being causal for trait t in the multi-trait case. Thus, the generative model for the causal candidates are:

$$\begin{aligned}
&\text{Single trait: } \tilde{z}_m \sim \begin{cases} N(0, \sigma_0^2), & w_m = 0, \\ N(\tilde{\mathbf{A}}_m \lambda, \sigma^2), & w_m = 1. \end{cases} \\
&\text{Multiple traits: } \tilde{z}_{tm} \sim \begin{cases} N(0, \sigma_0^2), & w_{tm} = 0, \\ N(\tilde{\mathbf{A}}_{tm} \lambda, \sigma^2 + \tilde{\mathbf{B}}_{tm} \Sigma \tilde{\mathbf{B}}_{tm}^T), & w_{tm} = 1. \end{cases}
\end{aligned} \tag{3.4}$$

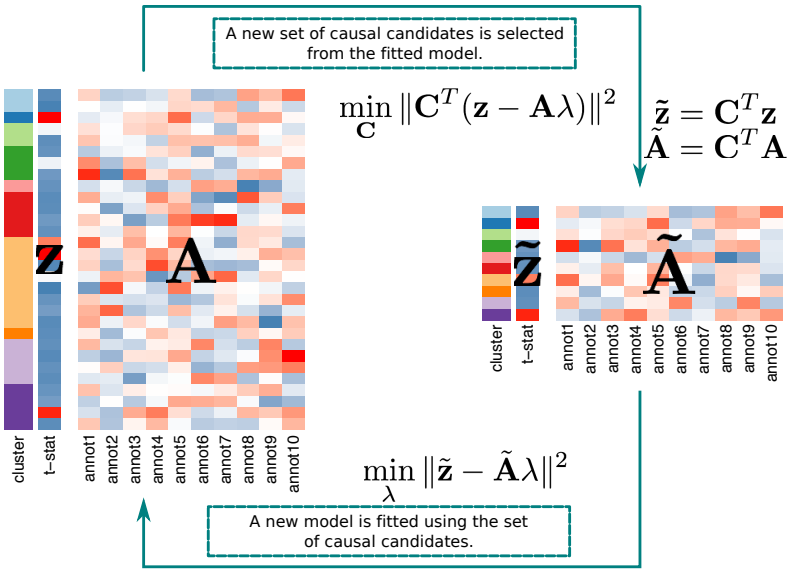

Figure 3.2: **FM-HighLD workflow for single-trait model.** `FM-HighLD` iterates two steps until convergence: i) Selects a group of candidate causal SNPs (one from each LD cluster) by looking for the variant best represented by the annotation; and ii) using the causal candidate SNPs, it models the association statistics from the variants to a single or multiple traits as a function of allele specific annotations.

The `FM-HighLD` algorithm

We rewrite equations 3.2 and 3.3 in their vector forms:

$$\mathbf{z} = \mathbf{A}\lambda + \epsilon,$$
$$\mathbf{z} = \mathbf{A}\lambda + \mathbf{B}\mathbf{u} + \epsilon.$$

We estimate the parameters of the model for single and multiple trait data using the Expectation- Maximization (EM) algorithm (Dempster et al. (1976)). In our EM implementation, we considered following important issues for efficient computation. Although there are explicit solutions for the regression coefficient λ in the single-trait case, there is not

an explicit solution for the causal configuration matrix $\mathbf{C}$. Furthermore, estimation of this matrix would require searching over $\prod_m q_m!$ different configurations, thus we employ the Expected Conditional-Maximization (ECM) algorithm (Meng and Rubin (1993)). By utilizing this algorithm, instead of comparing all possible configuration of the matrix $\mathbf{C}$, we compare only the q_m possible causal candidates for the m-th LD cluster, reducing that way the computation time to be $O(Q)$.

Pre-clustering of the variants into LD clusters: We grouped the regulatory SNPs into LD clusters by using the LD matrix as follows. We defined an undirected graph using the variants as vertices, and allowed an edge if the genomic distance between them was smaller than 1 MB and the squared correlation between them was greater or equal than 0.75. We used `igraph` (Csardi and Nepusz (2006)) to find the groups of maximal connected subgraphs, i.e. the groups of nodes for which there exists a path that connect all of them.

Initialization: For the single-trait (multi-trait case), we initialize the algorithm with a set of 1 candidate SNP per LD cluster (trait / LD cluster pair). Then, repeat until convergence:

1. **E-step**: Using the causal candidates, calculate the posterior probabilities for each LD cluster to contain a causal variant (contains a causal variant for trait t):

   $$\text{Single trait: } \gamma_m = P(w_m = 1 \mid \tilde{z}_m, \mathbf{A}_m) \propto \pi N(\tilde{z}_m \mid \tilde{\mathbf{A}}_m \lambda, \sigma^2)$$
   $$\text{Multiple traits: } \gamma_{tm} = P(w_{tm} = 1 \mid \tilde{z}_{tm}, \mathbf{A}_{tm}) \propto \pi N(\tilde{z}_m \mid \tilde{\mathbf{A}}_m \lambda, \sigma^2 + \tilde{\mathbf{B}}_{tm} \Sigma \tilde{\mathbf{B}}_{tm})$$

   where $N(x \mid \mu, \sigma^2)$ is the density function of Normal distribution with mean μ and variance σ^2.

2. **M-step**: We pool the posterior probabilities into a weight vector Γ. We estimate

   - For the single-trait case: The regression coefficient vector λ, and the error variance σ^2 are estimated by fitting a weighted linear model with the `lm` function.
   - For the multi-trait case: The regression coefficient vector λ, the random-effect variance matrix Σ and the error variance σ^2 are estimated by fitting a weighted linear mixed model with the `lmer` function (Bates et al. (2015)).

   With the regression coefficients estimated, we then select causal candidate variants by utilizing the following strategy:

- Single-trait case: For each LD cluster m, we pick the causal candidate as the SNP that is best represented by the annotation data:

$$\mathbf{C}_m[k] = \begin{cases} 1 & \text{if } k = \text{argmin}_{k'}(z_m[k'] - \mathbf{A}_m[k']\hat{\lambda})^2, \\ 0 & \text{otherwise.} \end{cases}$$

  where we denote the k-th entry (row) of the vector (matrix) with brackets. Then, we build the causal configuration matrix as a block-diagonal matrix $\mathbf{C} = \text{diag}(\mathbf{C}_1, \cdots, \mathbf{C}_M)$.

- Multi-trait case: For every trait t, we built a causal configuration matrix $\mathbf{C}_t$ following an analogous procedure as the single-trait case. Then, construct the causal configuration matrix as block-diagonal of the trait specific configuration matrices $\mathbf{C} = \text{diag}(\mathbf{C}_1, \cdots, \mathbf{C}_T)$.

This step is an heuristic that approximates the solution of a combinatorial problem, and it is known that the EM algorithm is guaranteed to converge to a local maxima, but not to global maxima. Therefore, we included a modification of this step, where the ℓ-th smallest variant is selected (by default $\ell = 2$) with probability p (by default $p = 0.05$) instead of the variant with minimum residual. We named this modified algorithm as `FM-HighLD r`.

The derivation of the modified EM algorithm is given in Appendix C.

Pooling multiple loci together in the Multi-trait model

.

The number of annotation covariates that can be used is bounded by the number of LD clusters. Thus, it is useful to pool multiple loci together to increase the number of LD clusters. For that purpose, we fitted a mixture of linear mixed models, where the q-th association statistic between trait t and variant m is generated from:

$$
\begin{aligned}
w_q &\sim \text{Ber}(\pi) \\
z_q \mid w_q = 0 &\sim N(0, \sigma_0^2) \\
z_q \mid w_q = 1 &\sim N(A_q^T \lambda, \sigma_q^2)
\end{aligned}
\tag{3.5}
$$

where the parameters in the $w = 1$ case are derived from a linear mixed model as defined by equation 3.3 using both fixed and random effects for the intercept and the allelic skew / percentage of overlapping peaks. Then, we grouped the data according to the signs of the estimated coefficients:

Group	Description
I	Both intercept and allelic skew coefficients $\geqslant 0$
II	Intercept coefficient < 0 and allelic skew coefficient $\geqslant 0$
III	Both intercept and allelic skew coefficients < 0
IV	Intercept coefficient $\geqslant 0$ and allelic skew coefficient < 0

Table 3.1: **Definition of groups to pool loci together for multi-trait analysis.**

Annotation data.

Peak overlaps annotation. ChIP-seq and DNase-seq peaks from GM12878, sequenced by the ENCODE Project Consortium were downloaded from `https://www.encodeproject.org/` (The ENCODE Project Consortium (2012)). ATAC-seq peaks from GM12878 were downloaded from the Gene Expression Omnibus with accession number GSE47753 (Buenrostro et al. (2013)). The peaks were centered around the summit to have fixed width of 500 bps.

Peak percentage annotation. Using peaks called by ENCODE (The ENCODE Project Consortium (2012)), we defined the peak percentage annotation for variant n as:

$$
\text{peak \%}_n = \left(\frac{\text{\# of ChIP-seq dataset in which SNP } n \text{ overlaps a peak}}{\text{\# of ChIP-seq datasets}} \right) \times 100 \tag{3.6}
$$

Allelic skew annotation. ChIP-seq and DNase-seq reads from GM12878, sequenced by the ENCODE Project Consortium were downloaded from `https://www.encodeproject.org/` (The ENCODE Project Consortium (2012)). ATAC-seq reads from GM12787 were downloaded from the Gene Expression Omnibus with accession number GSE47753 (Buenrostro et al. (2013)). Samples were aligned with `Bowtie2` (Langmead and Salzberg (2012)) to the hg19 genome where the regulatory variants were masked. The allele-specific alignments were determined with `SNPsplit` (Krueger and Andrews (2016)). This program tries to assign the reads to an exclusive allele, and the reads that can be aligned to both alleles are labelled as unassigned. Those reads were allocated to a random allele with probability 0.5. In order to make the allelic skew more representative of allele specific activity, we defined the allelic skew for the variant n:

$$\begin{aligned}\text{allelic skew}_n = &\log_2\left(1+\sum_a \text{alt}_{n,a}1(\text{alt}_{n,a}\geqslant \text{ref}_{n,a})\right)\\ &-\log_2\left(1+\sum_a \text{ref}_{n,a}1(\text{alt}_{n,a}\leqslant \text{ref}_{n,a})\right)\end{aligned} \tag{3.7}$$

where the sum ranges over all ChIP-seq TF data sets.

atSNP scores annotation. We calculated an annotation score matrix using `atSNP` (Zuo et al. (2015)) with JASPAR probability weight matrices (PWMs) (Khan et al. (2017)). `atSNP` quantifies the impact of SNPs, i.e., the likelihood that a given SNP disrupts or enhances the binding sites from a given set of position weight matrices characterizing the class of sequences TFs recognize. `atSNP` operates by scanning through subsequences overlapping with the SNP position with reference and SNP alleles for the best matches of both to a given PWM. It quantifies the significance of the best matches with the reference and SNP alleles by p-values. Then, the log ratio of the two p-values are defined as the `atSNP` annotation score, which empirically reflects the change in the ranks of the PWM matches of the alleles. SNPs likely to enhance or disrupt binding of given TF have large absolute `atSNP` scores for the corresponding PWM while SNPs with little potential impact on binding have scores close to zero.

DeepSea scores annotation. We calculated `DeepSea` annotations (Zhou and Troyanskaya (2015)) by providing VCF files of the regulatory SNPs to the `DeepSea`'s web browser

`http://deepsea.princeton.edu/job/analysis/create/`. `DeepSea` provides chromatin feature log fold change, computed by comparing the chromatin feature probabilities for sequences carrying the reference, and alternative alleles.

Modified atSNP and DeepSea annotations. We modified the sequence scores by multiplying the SNP score by a peak overlap indicator:

$$\mathbf{a}_{n,a} = \mathbf{s}_{n,a} 1(\text{SNP}_n \text{ overlaps a peak for assay } a).$$

These modified versions incorporate the actual *in vivo* chromatin measurements into *in silico* annotations.

Allelic skew imputation for SNPs with less that 5 aligned reads. Figure 3.3 exhibits that a small proportion of variants are covered by fewer than 5 reads. To keep these SNPs in our sample, we fitted a 5-nearest neighbors model using the GM12878 `DeepSea` fold change scores as covariates and the observed allelic skew as response.

To select the model, and the set of covariates we divided the data set into a train data set with 75% of the SNPs covered by 5 or more aligned reads, and kept the rest as an independent test data set. We evaluated the following models: k-nearest neighbors with $k = 5, 10, \cdots, 50$, elastic nets with all combinations in the grid formed by $\alpha = 0, 0.25, \cdots, 5$ and $\alpha = 0, 0.1, \cdots, 1$, and partial least squares with $n_{\text{comp}} = 1, 2, \cdots, 15$ using `caret` (Kuhn (2008)). We evaluated the models, and data set by comparing the total skew metric in an independent test data set:

$$\text{total skew} = \frac{\#(y_n^{\text{pred}} > 0)}{\#(y_n^{\text{obs}} > 0)} + \frac{\#(y_n^{\text{pred}} < 0)}{\#(: y_n^{\text{obs}} < 0)}$$

where y_n^* is the observed or predicted allelic skew for the n-th variant, respectively.

Single-trait simulation settings.

Using regulatory variants of the Massively Parallel Reported Assay, and their corresponding LD matrix estimates from the European (EUR) population of 1000 genomes. We generated a continuous phenotype from a polygenic model with $K = 5, 10$ causal SNPs as:

$$\mathbf{y} = \sum_{k \in \mathcal{C}(K)} \mathbf{x}_k \beta_k + \epsilon, \quad \epsilon \sim N(0, (1 - h_g^2)I).$$

Here, $\mathcal{C}(K)$ is the set of K causal SNPs, $\mathbf{y}$ is the phenotype vector, $\mathbf{x}_k$ is the vector of

centered and scaled genotypes and β_k is the effect size at the k-th causal SNP, and h_g^2 is the proportion explained of phenotypic variance. We fitted a mixture model with two components, means μ and $-\mu$, and same variance σ^2 to the allelic skew, and used these parameters to sample the SNP effects from the model $\beta_k \sim \pi N(-\mu, \sigma^2) + (1-\pi)N(\mu, \sigma^2)$. Where the estimated parameters are given by:

Parameter	Value
$\hat{\pi}$	0.8
$\hat{\mu}$	4
$\hat{\sigma}^2$	2

Table 3.2: **Estimated allelic skew parameters used to sample the SNP effects.**

We defined different rules to construct a set of K causal variants $\mathcal{C}(K)$. Then, we repeat independently for each simulation:

- **random**: We sample K different regulatory SNPs randomly without replacement.
- **high LD**: For every LD cluster with two or more SNPs, we calculate the average squared correlation as:

$$\text{ave}(\rho^2) = \frac{\sum_{i \neq j} \rho^2(\mathbf{x}_i, \mathbf{x}_j)}{\#(\text{SNP pairs})}$$

  where ρ is the sample correlation. Then, we sample K LD clusters weighted by its $\text{ave}(\rho^2)$ without replacement, and randomly select one causal SNP among its variants.

  For both sampling schemes, we generated the SNP effects from a Gaussian mixture model with two components

Measuring performance in single-trait simulations.

For every simulation, we calculated the area under the ROC and Precision-Recall curves using `precrec` (Saito and Rehmsmeier (2017)) to compare the posterior probabilities of a SNP being causal with labels indicating if the SNPs belong to the causal sets $\mathcal{C}(K)$.

Multi-trait real data analysis

We used the 2,335 regulatory SNPs the LCL dataset generated by Tewhey et al. (2016). The SNPs in this dataset were tested for association with 41 genes in average. We compared

`FM-HighLD` and `PAINTOR` (Kichaev et al. (2016)) using the eQTL association statistics from Tewhey et al, and the four annotation matrices that we built for the regulatory variants: Only allelic skew, allelic skew with atSNP scores, allelic skew with `DeepSea` log fold change scores, and allelic skew with ENCODE peak overlaps. We used a random intercept and a random allelic skew coefficient in the multi-trait `FM-HighLD` model, independently of the annotation matrix that we used.

Measuring the Multi-trait analysis of eQTL data.

We used the expression modulating (EM) labels resulted from the MPRA experiment by Tewhey et al. (2016) to evaluate the performance of our model. We used `precrec` (Saito and Rehmsmeier (2017)) to calculate the area under the ROC and Precision-Recall curves for every loci, separately. Since the EM labels are not gene specific, we defined the probability of a SNP being causal:

$$\mathrm{P}(\text{SNP } s \text{ is causal}) = \max_{g} \mathrm{P}(\text{SNP } s \text{ is causal for gene } g) \tag{3.8}$$

We define the Percentage of Identified loci, as the ratio between the number of correctly classified loci, divided by the total number of loci that were evaluated:

$$\%\ \text{Identified loci}(p, q) = \left(\frac{\#\ \text{of loci with AUROC} \geqslant p \text{ and AUPREC} \geqslant q}{\#\ \text{of evaluated loci}} \right) \times 100 \tag{3.9}$$

where AUROC and AUPREC are the area under the ROC and Precision-Recall curves, respectively.

3.3 Results

Peak overlaps are strongly associated with regulatory variants, but weakly associated with causal SNPs.

We used the dataset generated by Tewhey et al. (2016), which is composed by thousands of eQTL variants in regions associated with differential gene expression in a European (EUR) population. According to expression estimates from the Massively Parallel Reporter

Assay (MPRA), these variants were labelled as regulatory whether the associated reporter gene was expressed in any allele, and as expression modulating (EM) when the associated gene was expressed only in one allele. We first sought to determine whether the regulatory SNPs are associated with peak overlaps, Figure 3.3**A** shows that as expected the regulatory variants are significantly associated with peaks of different assays. We next interrogated whether expression modulating variants are associated with those peaks by conditioning over the regulatory SNPs. Figure 3.3**B** illustrates that the majority of the ChIP-seq peaks are not strongly associated with the EM variants, as the adjusted p.values were less than 10^{-2} for only few assays.

Allelic skew is strongly associated with EM variants.

Since peak overlaps don't have enough power to distinguish expression modulating among regulatory SNPs, we next sought to evaluate if Allele Specific Binding (ASB) events had enough power. For the variants that were covered by 5 or more mapped reads, we calculated Binomial tests and define an ASB event if the associated FDR < 5%. We quantified the percentage of regulatory variants that were covered by 5 or more aligned reads. Figure 3.3**C** exhibits that ATAC-seq is the only individual assay that covered more than 25% of those SNPs. We next grouped all ChIP-seq experiments to construct the allelic skew annotation defined as in equation 3.7, we conjecture that when adding the reads mapped to each allele, it results in artificial non-allele specific binding events (Figure 3.4).

We compared if the cumulative distribution of the absolute allelic skew was different for the EM variants, Figure 3.3**D** shows that variants with higher regulatory activity measured by the allelic skew are more likely to be causal. Supplementary Figure C.1 illustrates that for a robust threshold range the allelic skew is more strongly associated to EM variants than any of the peak overlaps.

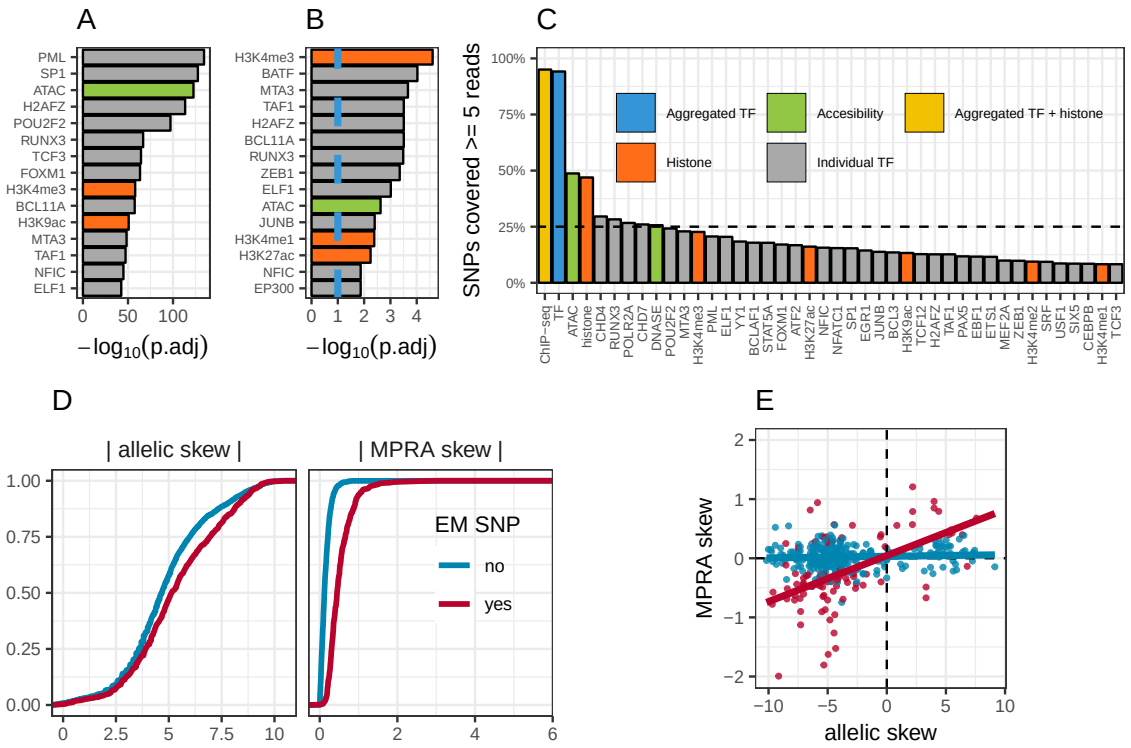

Figure 3.3: **Exploratory analysis of annotation data. A.** Adjusted p.values of χ^2 association tests between regulatory labels and peak overlaps. **B.** Adjusted p.values of χ^2 association tests between causal labels and peak overlaps (adjusted with BH) for different assays, we removed the assays with p.adj $= 1$. **C.** Percentage of causal SNPs coverage by 5 or more aligned reads. **D.** Cumulative distribution of the absolute allelic (left) and MPRA (right) skew for causal (red) and not causal SNPs (blue). **E.** Allelic skew vs. MPRA skew for the top 25 loci with highest conditional correlation for the EM SNPs.

Finally, we explored the relationship between the allelic and MPRA skew. We noticed that for some loci there is a different pattern when separating the EM from all the regulatory variants, for example we observe in Figure 3.3**E** that there is a clear increasing trend for the EM variants, but not for the rest when we aggregate the top 25 loci with highest conditional correlation of the expression modulating variants. In summary, we consider to aggregate the *in-vivo* information because approximately 75% of the regulatory SNPs are not covered

by 5 or more reads of almost any ChIP-seq sample, aggregating the ChIP-seq samples into an allelic skew feature reduces the dimension of the feature matrices and the number of features allowed by `FM-HighLD` is bounded by the number of LD clusters in the dataset, and the allelic skew is comparable to other features to discern the EM SNPs.

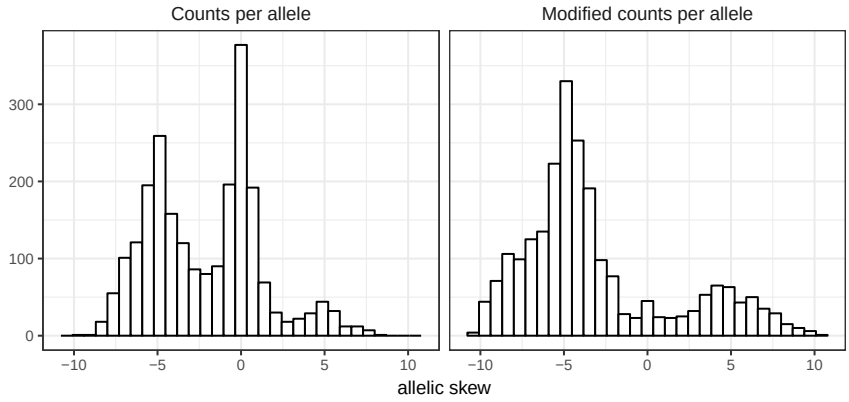

Figure 3.4: **Allelic skew histograms.** Comparison of two aggregating schemes to calculate the allelic skew: In the left panel, the allelic skew of each variant was computed by adding the reads mapping to each allele. In the right panel, the allelic skew for each variant was computed by adding only the read mapping to the allele where more reads are aligned.

Single-trait simulations.

We next explored the performance of our method when analyzing single-trait simulations. Using the LD matrix generated from the EUR population evaluated for the regulatory variants of the lymphoblastoid cell lines (LCL), we sampled a continuous phenotype from a polygenic model with $K = 5, 10$ SNPs under two sampling schemes: First, we selected K variants randomly (**random**); Second, for each loci we calculated its average ρ^2, then selected K random loci with probability proportional to its ave(ρ^2) from those loci with average ρ^2 greater than 0.75. This way, we are capable of ensuring that the K causal SNPs exhibit high correlations with multiple variants (**high LD**). Table 3.3 lists the mean and standard deviation of the ave(ρ^2) among the LD cluster containing the causal SNPs.

K	random	high LD
5	0.90 (0.05)	0.93 (0.02)
10	0.91 (0.03)	0.94 (0.01)

Table 3.3: **Moments of ave**(ρ^2)**.** Mean and standard deviation of the causal SNPs' ave(ρ^2).

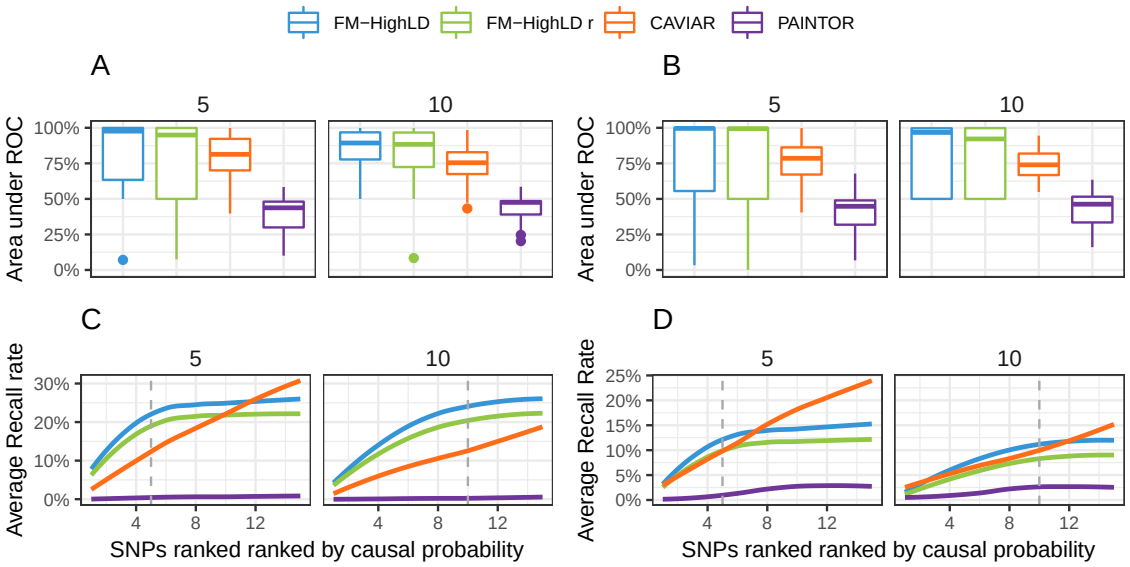

Figure 3.5: **Single-trait simulation results.** Simulated data of polygenic model with K SNPs. Area under the ROC curve for simulations where the K causal SNPs were (**A**, **random**) sampled randomly, or (**B**, **high LD**) sampled from LD cluster with average $\rho^2 \geqslant 0.75$. Average Recall Rate calculated for the top SNPs ranked by their probability of being causal for the (**C**) **random** and (**D**) **high LD** scenarios.

In our simulations, we compared the two versions of `FM-HighLD` (regular and with a randomization step called `FM-HighLD r`) to two other fine-mapping methods: `CAVIAR` (Hormozdiari et al. (2014)) and `PAINTOR` (Kichaev et al. (2014); Kichaev and Pasaniuc (2015); Kichaev et al. (2016)). Figures 3.5 **A** and **B** compare the area under the ROC curve distribution under the **random** and **high LD** sampling schemes, respectively. `FM-HighLD` outperforms both `CAVIAR` and `PAINTOR` when the trait it simulated from a smaller amount

of causal SNPs, this is not unexpected as the association statistics are the t-statistics are more likely to be proportional to the annotation for polygenic models with fewer variants (Supplementary Figure C.2). Figure 3.5 **C** exhibits the recall curves for the **random** scheme, where `FM-HighLD` outperforms the causal probability rank is smaller than the number of variants generating the trait. In other words `FM-HighLD` tends to assign a higher probability to the true causal variants. As expected, Figure 3.5**D** shows that the Recall rate is lower in average for all methods under the **high LD** scenario. In summary, these simulations show that `FM-HighLD` outperforms both `CAVIAR` and `PAINTOR` when the variants were randomly sampled among all the regulatory SNPs, and that both `FM-HighLD` and `CAVIAR` are comparable under the **high LD** scenario, while outperforming `PAINTOR`. These results highlight that `FM-HighLD` enables estimating the probability of being causal for variants with effects that depend on the annotation data.

Multi-trait eQTL analysis in LCL.

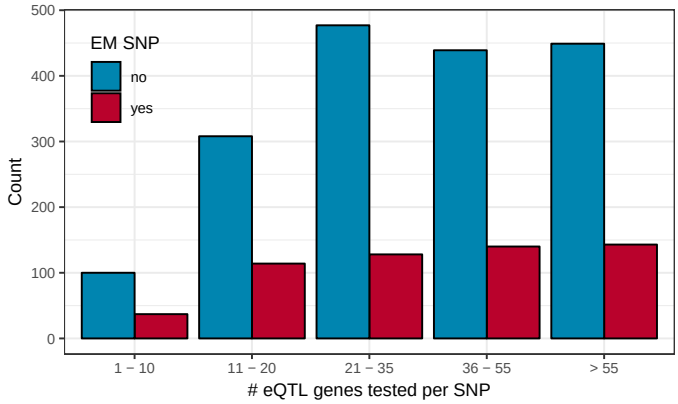

Figure 3.6: **Number of genes tested for association with a SNP.**

We analyzed a subset of the eQTL dataset constructed by Tewhey et al composed only by the variants that were deemed as regulatory, Kreimer et al. (2017) analyzed multiple methods to prioritize the regulatory variants as part of the CAGI 4th conference. Exploratory analysis indicated that every SNP was tested for association with 41 genes in average (Figure 3.6 displays the distribution of the number of genes tested per SNP stratified by EM label), as

Tewhey et al. constructed the MPRA by selecting the top associated variants and its LD partners, of a set of 3,642 eQTL genes in the EUR population of the Geuvadis RNA-seq of LCL dataset (Lappalainen et al. (2013)). To simplify the analysis, we partitioned the variants into different loci by aggregating the SNPs that were tested for at least one gene in common, and considered only the loci composed by 5 variants or more. This resulted into 144 loci, which in are composed in average by 2 LD clusters, with 1.7 SNPs and 0.4 causal variants in average. Figure 3.7**A** shows the chr7: 30 MB - 33 MB loci that was examined above, and **B** zooms into the chr7: 32.5 MB - 33 MB loci. This figure highlights various technical challenges: First, regulatory SNPs can be separated by more than 1MB, but still be regulating common genes. Second, the same variant is tested against multiple eQTL genes. Third, when zooming into the smaller loci Figure 3.7 exhibits that very few variants overlap with the ATAC-seq or histone ChIP-seq peaks. When coding this observation into overlap vectors, they might lack discriminative power to determine that which variants modulate certain eQTL genes. Finally, the LD matrix below exhibits that, as expected, these variants were selected to be in high LD. The bottom panel of Figure 3.7**C** compares the relationship between both *in-vivo* based annotations allelic skew and peak percentage, and the association statistics with multiple genes. In this example, we notice that only the EM variants exhibit more extreme values of activity even though multiple gene-SNP associations are not significant.

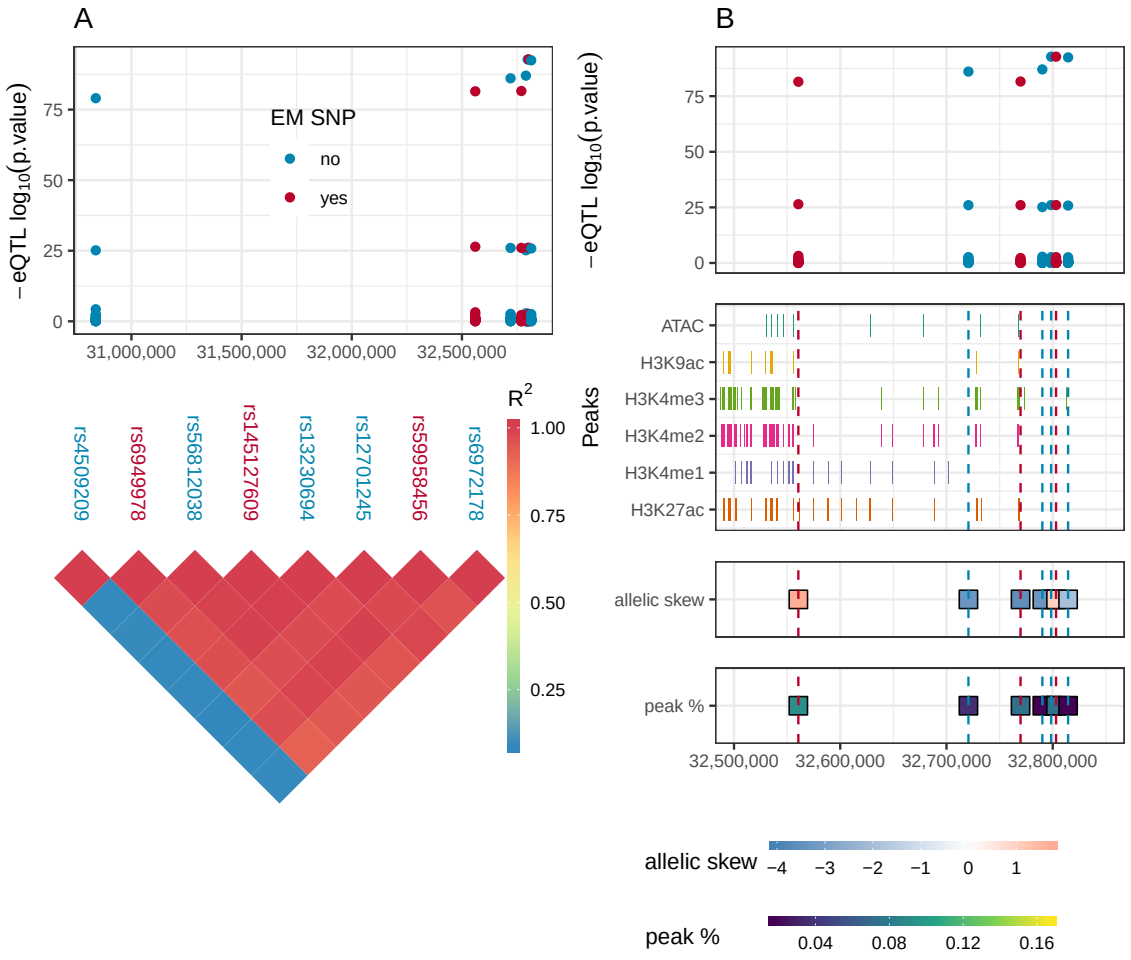

Figure 3.7: **Multi-trait data overview. A** (top) Manhattan plot for the chr7: 30 MB - 33 MB loci and (bottom) LD structure for the regulatory variant in the loci. **B** (top) Manhattan plot for the chr7: 32.5 MB - 33 MB loci; (middle) EM SNPs don't always overlap ATAC-seq, and histone ChIP-seq (H3K9ac, H3K4me1, H3K4me2, H3K4me3 and H3K27ac) peaks; (bottom) Heatmaps of allelic skew and peak percentage illustrating that EM variants exhibit increased activity on both annotation data.

We applied `FM-HighLD`, `FM-HighLD r` and `PAINTOR` to the eQTL dataset using 4 different sets of annotations used as fixed effects: Only allelic skew, allelic skew with atSNP

scores, allelic skew with DeepSea log fold change scores, and allelic skew with ENCODE peak overlaps, while using allelic skew as random effects as well. We first fitted linear mixed models using an intercept and allelic skew as both fixed and random effects, and then we pooled multiple loci together according to the signs of the estimated coefficients (equation 3.5 and Table 3.1) to increase the amount of annotation features to use, because this quantity is bounded by the number of LD clusters in the data.

We evaluated the fine-mapping methods for every loci by separate. Figures 3.8 and 3.9 depict the average area under the ROC and precision-recall curves when the false discovery rate is below 50%, respectively. FM-HighLD and FM-HighLD r outperform PAINTOR under both metrics. Furthermore, we observe that the majority of the models that include the Peak Percentage feature outperform the models that include the allelic skew, except in the models that use DeepSea annotation covariates. Finally, we observe that FM-HighLD r strategy tends to outperform the regular FM-HighLD.

Supplementary Figure C.4 exhibits that both FM-HighLD alternatives outperform PAINTOR, except in the case where only the allelic skew was used as annotation. At a first glance, it appears that this result contradicts the conclusions of the single-trait simulations. However, to evaluate the multi-trait data with PAINTOR, the variants tested for association with every gene are considered to be conditionally independent. That way, the computation of the likelihood is considerably simplified because the dimension of the LD matrices is smaller than in the single-trait case. We assessed that FM-HighLD r outperforms over FM-HighLD, this is occurring due to the fact that we are solving a mixed combinatorial-continuous optimization problem by approximating it as a continuous problem, which causes the algorithm to return the parameters of a local maxima, as guaranteed by the EM algorithm. Supplementary Figure C.5 exhibits the same pattern, where the allelic skew annotation is replaced with the percentage of ChIP-seq samples that at least one peak overlaps the variant. This is not surprising as both features exhibit are highly correlated (Supplementary Figure C.3), but with the allelic skew we can correctly determine the causal variants as this variable encodes more information than the peak percentage. To further elucidate the reasons why FM-HighLD outperforms PAINTOR, we compared the probabilities of the SNPs being causal for a certain trait by comparing the allelic skew with the t-statistic for the associations in the chr7: 30 MB - 33 MB. Figure 3.10 shows that: First, FM-HighLD assigns causal probabilities by trait-variant association. Second, this allows FM-HighLD to detect different groups of causal SNPs for different traits. Finally, PAINTOR tends to

assign high probabilities to almost all the traits tested for association with a given SNP, this is problematic because in some instances it could label as causal, SNPs that are not even significantly associated to a trait. This is not surprising, as in PAINTOR the annotation impacts the causal indicator latent variable by the use of a logistic regression model, while in our approach the association statistics are modeled as a function of annotation data for the candidate causal variants.

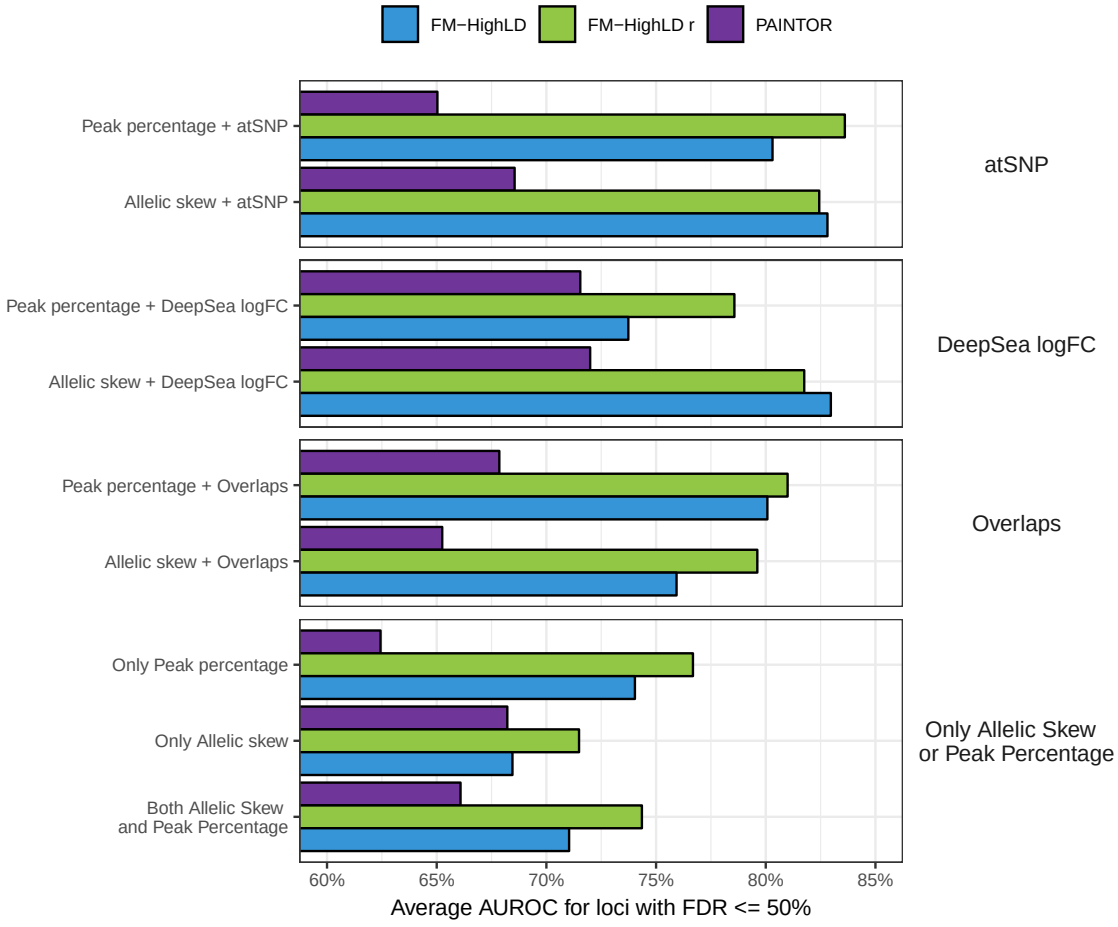

Figure 3.8: **Average AUROC over loci with FDR $\leqslant$ 50%.**

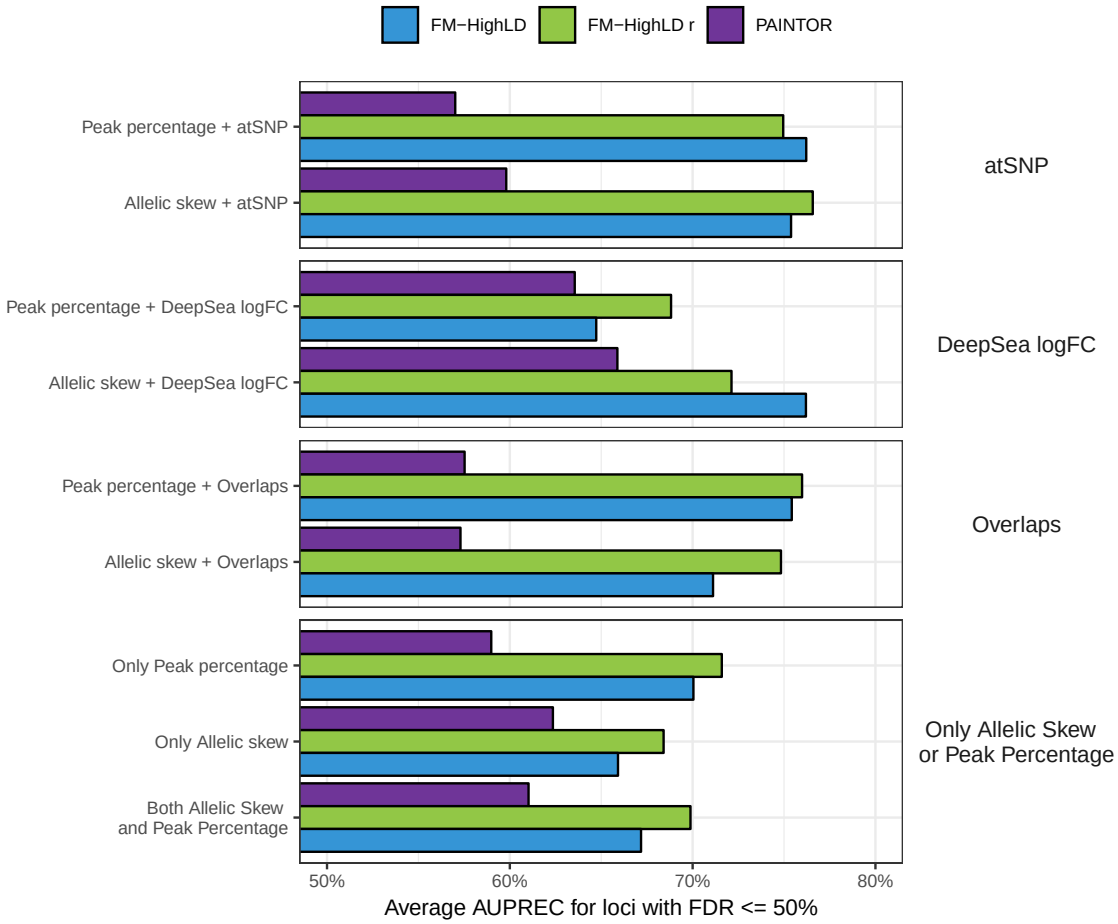

Figure 3.9: **Average AUPREC over loci with FDR $\leqslant$ 50%.**

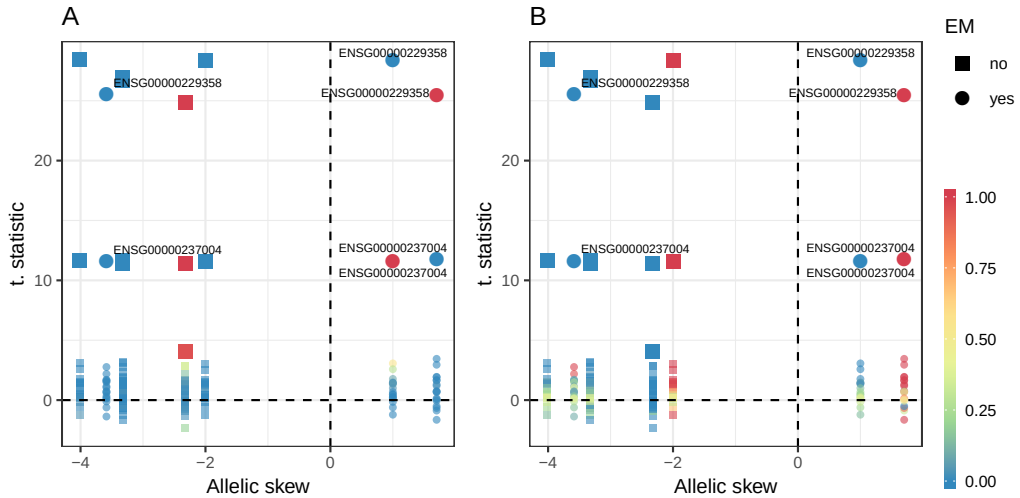

Figure 3.10: **Comparison of FM-HighLD and PAINTOR in chr7: 30 MB - 33 MB loci.** Allelic skew vs. t-statistic filled with probability of being causal estimated with `FM-HighLD` (**A**) and `PAINTOR` (**B**).

3.4 Discussion

We utilized for the first time MPRA data, which for this experiment provides a sequencing based gold-standard to evaluate fine-mapping methods by using real data, as contrasted with previous methods that were evaluated only by using simulated data. As with any other analysis, it is important to understand its limitations, thus we note that MPRAs may not always represent the neighboring chromatin structure of the variant because the assay isolates the sequence surrounding the variant from potential cofactors, this is argued by Tewhey et al. (2016) as they estimated the sensitivity of the assay to be in the $9\% - 24\%$ range. Therefore, it is possible that variants that are not labelled as EM by the MPRA are being assigned as causal by `FM-HighLD`.

We presented an integrative fine-mapping method for variants in high LD, named `FM-HighLD`, that is capable of detecting causal variants for single and multiple traits. The key ideas behind our framework are that given a set of putative variants is possible to discern the pattern that associates summary statistics with annotation data generated from multiple assays, and that if there is a mechanism recovered by the annotation data that explains the association statistics, then is possible to select a set of candidate causal SNPs. For the multi-trait case, `FM-HighLD` utilizes linear mixed models to account for the relatedness among different traits, that way the common independence and identically distributed multiple traits assumption is relaxed. This framework provides many directions for useful extensions: First, by including a prior distribution over the annotation coefficients is theoretically possible to add model selection capabilities; Second, we used ChIP-seq data to derive the allelic skew which would make our method only possible in immortalized cell lines for which several ChIP-seq assays have been generated. However, the facts that the allelic skew is correlated with ATAC-seq skew (Supplementary Figure C.6, $\hat{\rho} = 0.7372$) and the rapid generation of these samples would mean that our method to be applicable on tissue-derived cell lines; Third, since our framework is based on the joint continuous and combinatorial optimization of equation 3.1 it is possible to improve our predictions by the use of more sophisticated optimization techniques such as genetic algorithms. Finally, we note that with the rapid development of Hi-C assays, it could be possible in the future to design features in the future that code the relationship between and variant as part of the annotation data.

4 COMPUTATIONAL TOOLS FOR CHIP-SEQ DATA ANALYSIS

4.1 Introduction

Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) has become standard technology to investigate genome wide binding sites of transcription factors (TFs). To analyze these binding sites there are many tools for calling peaks, annotating those peaks to a curated/modeled partition of the genome, or discover regions that are differentially enriched. However, before starting to properly analyze the ChIP-seq data, one of the main complications is how to quantify the experiments' quality. Even though there are several pre-established QC metrics (Landt et al. (2012); Marinov et al. (2014); Planet et al. (2011)), it is necessary to pre-process the data to the specific software's required format, possibly install additional software environments (some of which the users might not be familiar with), or even re-code many of the suggested QC metrics (for example the metrics in ENCODE consortium (2011)). To our knowledge, the only R package that calculates these metrics is `ChIPQC` (Carroll et al. (2014)) which is focused into creating a set of quality report for the whole experiment, but it doesn't give the user the possibility of manipulating the computed values of the quality indicators.

In a downstream analysis, the software associated to the majority of the ChIP-seq data analysis method includes few visualization routines to accompany or diagnose the results generated by the method. For example, peak-caller methods such as `MOSAiCS` (Kuan et al. (2011)) or `CisGenome` (Ji et al. (2008)) include routines to plot the experiments' coverage across the peaks called by the model, but don't allow to manipulate the coverage estimated around those regions. Furthermore, there are not tools that allow the user to flexibly explore a collection of peaks, thus when summarizing the coverage across a collection of peaks it is required to calculate the coverage across the peaks and code a different routine for every subset of interest. Besides those utilities, other tools have been designed for exploratory analysis of peaks and overlapping aligned reads: First, the most commonly used is the genome browser (Kent et al. (2002)) which allows users to explore different types of genomic data along the whole genome, and observe the interaction between different types of tracks, but it seems to lack detail at base-pair resolution. Second, tools like `PeakAnalizer` (Salmon-Divon et al. (2010)), and `PAVIS` (Huang et al. (2013)) allow to visualize individual peaks,

but they seem to be focused on different tasks like peak-splitting or genome-annotation, respectively. Finally, tools like `ChIPseeker` (Yu et al. (2015)) are general suites of tools for downstream ChIP-seq analysis that allow the user to annotate peaks in genome, compare and visualize of peaks. However, it lacks flexibility when summarizing ChIP-seq signal across a set of annotated marks in either functional profiles, or heatmaps.

We introduce two `R` packages: `ChIPUtils`, and `Segvis`. These packages were motivated by common tasks that are required in ChIP-seq data analysis. The first package `ChIPUtils`, enables the user to compute the most frequently used ChIP-seq quality control metrics without the need to install multiple software environments, and the second package `Segvis`allows the user to produce multiple visualization of next-generation sequencing samples across single and multiple genomic regions.

This chapter is organized as follows. In Section 4.2, we describe the mathematical formulation for the routines implemented in `ChIPUtils` and `Segvis`. In Section 4.3) we exemplify the use of both packages to characterize open and closed complexes with *E. coli* ChIP-seq data.

4.2 Methods

ChIP-exo/nexus quality control with `ChIPexoQual`

In Chapter 2, we illustrated a case of study of `ChIPexoQual` and evaluated multiple ChIP-exo/nexus samples with the quality control pipeline. From January 2018 and August 2018, `ChIPexoQual` has been downloaded between 80 and 150 by distinct IPs. The download statistics are regularly updated in `http://bioconductor.org/packages/stats/bioc/ChIPexoQual/`.

ChIP-seq quality control with `ChIPUtils`

The state of the art quality control metrics for ChIP-seq data are the Strand Cross-Correlation (SCC) and operations over the SCC curve, the Fraction of Read in Peaks (FRIP), and the PCR Bottleneck Coefficient (PBC). The first two metrics represent the sample's enrichment, and the third metric measures the library complexity of a ChIP-seq experiment.

Strand Cross-Correlation

The Strand Cross-Correlation (Kharchenko et al. (2008)) measures how tightly the ChIP-seq aligned reads are located around unobserved binding events, and it is computed by calculating the correlation between the vectors of the 5′ ends aligned to the forward and reverse strands. For a shift δ, the SCC is defined as:

$$\mathrm{SCC}(\delta) = \sum_{c} w_c \, r\left[n_c^{+}\left(\frac{\delta}{2}\right), n_c^{-}\left(\frac{\delta}{2}\right)\right] \tag{4.1}$$

where $\mathrm{SCC}(\delta)$ is the Strand Cross-Correlation for a shift δ, r is the Pearson correlation, w_c is the proportions of the experiment's reads aligned to chromosome c, and $n_c^S(x)$ is the vector of 5′ ends count vector shift by x bps for strand S and chromosome c. For ChIP-seq data, the SCC is expected to have two local maxima, one when δ is equal to the experiment's average read length, and other one when δ is equal to the unobserved fragment length. The local maximum at the read length may be occasioned by the duplicated reads or not mappable regions, or simply by reads in both strand that are mapped in the same position and is referred as the "phantom peak". Related QC metrics to the SCC, are determined by the relative height of the “phantom peak” and the local maximum at the fragment length.

$$\mathrm{NSC} = \frac{\max_\delta \mathrm{SCC}(\delta)}{\min_\delta \mathrm{SCC}(\delta)}, \quad \mathrm{RSC} = \frac{\max_\delta \mathrm{SCC}(\delta) - \mathrm{SCC}(\mathrm{rl})}{\min_\delta \mathrm{SCC}(\delta) - \mathrm{SCC}(\mathrm{rl})} \tag{4.2}$$

where rl is the experiment's read length, where $\mathrm{SCC}(\mathrm{rl})$ represents the background's cross-correlation value. For both of these metrics, higher values are an indication of the ChIP-seq sample being enriched.

Fraction of Reads in Peaks

This metric also represents library enrichment, and to calculate requires a peak-calling algorithm to identify a set of peaks associated to the aligned reads. It is defined as the ratio between the aligned reads overlapping the set of peaks, and the total number of aligned reads. As expected, this metric is highly dependent on the peak-caller algorithm, and the tuning parameters used in that process. For that reason, this metric is comparable between two different datasets if and only if the tuning parameters and peak-caller method were the same.

PCR Bottleneck Coefficient

The PCR Bottleneck Coefficient (PBC, Marinov et al. (2014)), is defined as the ratio between the number of genomic positions to which exactly one reads maps and the total number of genomic positions to which at least one unique mapping reads maps. The PBC of mouse and human experiments can be interpreted by following the guidelines stated in the ENCODE project's website (`https://genome.ucsc.edu/ENCODE/qualityMetrics.html`): $0 \leqslant \text{PBC} \leqslant 0.5$ indicates severe bottlenecking, $0.5 < \text{PBC} \leqslant 0.8$ determines moderate bottlenecking, $0.8 < \text{PBC} \leqslant 0.9$ is mild bottlenecking, and $0.9 < \text{PBC} \leqslant 1.0$ is no bottlenecking. Very low values can indicate a technical problem, such as PCR bias, or a biological finding, such as a very rare genomic feature. However, this metric can be greatly impacted by the length of the experiments' genome, or the space of possible positions where reads can map.

List of `ChIPUtils` functions for ChIP-seq data analysis.

Table 4.1 lists some of the available function in `ChIPUtils`.

Function	Objective
ChIPdata	Creates a `ChIPdata` object from a SE or PE ChIP-seq sample.
PBC	Calculates the **PBC**.
SCC	Calculates the **SCC**.
NSC	Calculates the **NSC**.
RSC	Calculates the **RSC**.
createBins	Bins a genome of interest in same width intervals, and counts the number of aligned reads overlapping each bin.

Table 4.1: **Summary of `ChIPUtils` functions.**

`ChIPUtils` is available at `https://github.com/welch16/ChIPUtils`.

List of `Segvis` functions for ChIP-seq peaks visualization

`Segvis` contains the following methods for the visualization of regions in the genome:

Function	Objective
readBedFile	Parses the content of a bed file into a GRanges object.
SegvisData	Summarizes the ChIP-seq samples' coverage across genomic regions.
plot_region	Plots the coverage of the ChIP-seq samples across a specific region.
DT_region	Returns a data.table with the coverage of the ChIP-seq samples across a specific region.
plot_profile	Plots the summarized signal of the experiment across a set of genomic regions with the same width.
DT_profile	Returns a data.table with the summarized signal of the experiment across a set of genomic regions with the same width.
plot_heatmap	Plots a heat map of the ChIP-seq experiments' normalized signal across a set of genomic regions with the same width.

Table 4.2: **Summary of `Segvis` functions.**

`Segvis` is available at `https://github.com/welch16/Segvis`. In section 4.3 we show examples of typical output of these functions.

4.3 Results

Quality control metrics of ChIP-seq data in *e. Coli* with `ChIPUtils`

We first calculated quality control metrics with `ChIPUtils` for SE and PE ChIP-seq data of σ^{70} in the *e. Coli* genome. Table 4.3 shows that as expected the PE datasets are of better quality than the SE data. This is not surprising, as the SE data was generated by sampling one end of each paired read in the PE data, i.e. the PE datasets double the sequencing depth of the SE datasets.

Group	Growth	Treatment	Rep.	Id.	Depth	NSC	RSC	PBC
ChIP-seq PE	Exp. $+O_2$	No Rif.	1	1	13.4M	8.86	1.05	0.94
	Exp. $+O_2$	Rif. 20 min	1	2	16.5M	7.03	1.01	0.93
	Exp. $+O_2$	No Rif.	2	3	11.6M	10.77	1.01	0.88
	Exp. $+O_2$	Rif. 20 min	2	4	16.8M	7.93	1.00	0.94
ChIP-seq SE	Exp. $+O_2$	No Rif.	1	1	6.7M	9.01	2.84	0.66
	Exp. $+O_2$	Rif. 20 min	1	2	8.2M	7.17	2.51	0.55
	Exp. $+O_2$	No Rif.	2	3	5.8M	10.89	3.12	0.64
	Exp. $+O_2$	Rif. 20 min	2	4	8.4M	8.12	2.69	0.58

Table 4.3: **Summary of the *E. coli* σ^{70} ChIP-seq samples.** Exp. stands for exponential and Stat. for stationary growth conditions. The last three columns depict ENCODE QC metrics on these data: NSC: Normalized Strand Cross-Correlation; RSC: Relative Strand Cross-Correlation; PBC: PCR Bottleneck Coefficient.

The statistics from Table 4.3 can be calculated with the `ChIPUtils::SCC`, `NSC`, `RSC` and `ChIPUtils::PBC` methods, respectively. However, it is recommended to visualize the SCC curve before calculating the NSC and RSC (Landt et al. (2012)).

Constructing a set of high quality PE ChIP-seq peaks

Since the SE ChIP-seq data was artificially built by sampling one end of every pair of aligned reads in the PE ChIP-seq dataset, we consider only the PE data to call peaks. We called peaks with `MOSAiCS` (Kuan et al. (2011)) for the σ^{70} and included a set of β'_f samples, for which the same treatments were applied. Table 4.4 shows number of peaks called for each sample.

IP	Condition	Rep. 1	Rep. 2
σ^{70}	Rif. 0 min	1989	2045
β'_f	Rif. 0 min	1079	1106
σ^{70}	Rif. 20 min	2313	2348
β'_f	Rif. 20 min	2477	2433

Table 4.4: **Number of peaks called with `MOSAiCS` for each sample.**

We next proceed to build a set of high quality peaks. For that purpose, we i) considered only the intersection of all 8 peak sets, ii) pulled the replicates for each pair of condition / IP samples, as the correlation (Table 4.5) is high enough, iii) we evaluated the peak signal by the value at the peak summit normalized by sequencing depth, and iv) kept only the peaks with signal above the 90% quantile (Figure 4.1).

IP	**Condition**	ρ
σ^{70}	Rif. 0 min	0.9736
σ^{70}	Rif. 20 min	0.9843
β'_f	Rif. 0 min	0.9933
β'_f	Rif. 20 min	0.9783

Table 4.5: **Correlation of σ^{70} and β'_f samples by rif. treatment.**

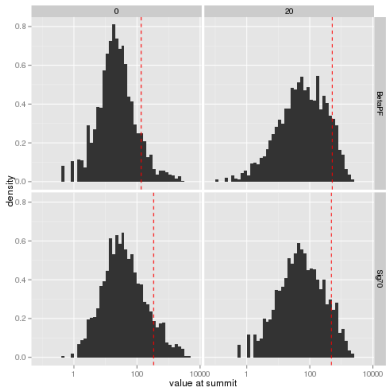

Figure 4.1: **Signal at peak summit for common σ^{70} and β'_f peaks.**

Examples of plots generated with `Segvis`

We next used `Segvis` to explore the set of high quality peaks. The `plot_region` method returns a `ggplot` object (Wickham (2016)), thus we can modify it to obtain publication ready figures (Figure 4.2). In both figures, we can see that as expected both σ^{70} samples exhibit a punctuated shape. On the other had, we can notice that for the β'_f samples the pattern differs whether the rifampicin (rif.) treatment was applied: there is an elongated shape when no treatment is applied, but a punctuated shape in the other case. This occurs

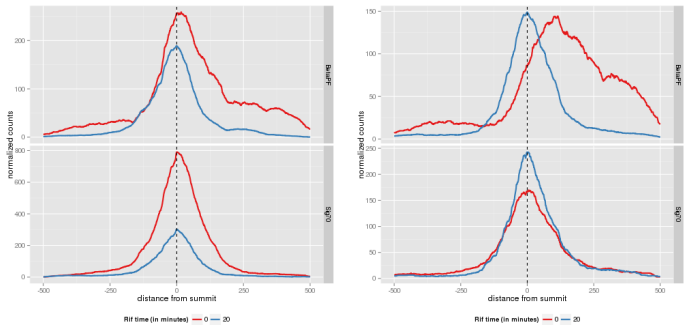

Figure 4.2: **Two peak examples comparing the binding patterns of σ^{70} and β'_f samples.** The genomic coordinate was centered toward the σ^{70} peak summit positions.

because the rif. treatment is expected to inhibit the transcription process performed by the β'_f IP.

After observing those patterns, we asked whether the same behavior was shared for a wider set of peaks. For that purpose, we used the `plot_profile` method on the peaks that had only one binding site (the binding sites were estimated with `dPeak` Chung et al. (2013)). Figure 4.3 exhibits the mean over all the peaks with only 1 binding site, as expected the average peaks are also punctuated for every sample, except the β'_f under no rif. treatment which exhibits elongated tails towards both directions, as peaks were called in a strand agnostic fashion.

Finally, we can examine the binding pattern in each peak separatedly. The method `plot_heatmap` allows clustering the peaks relative to one sample. Figure 4.4 shows the σ^{70} and β'_f peaks under both rif. conditions clustered by the `hclust` function, using the `ward.D2` strategy over the β'_f-rif 0 sample. It is clear that only for that replicate, there are groups of peaks with a strong elongated signal.

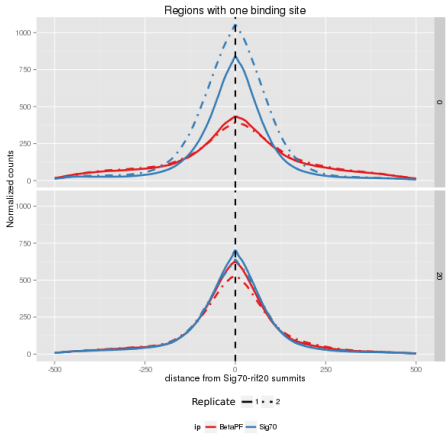

Figure 4.3: **Average profile for σ^{70} and β'_f peaks with only 1 binding site, separated by replicate.**

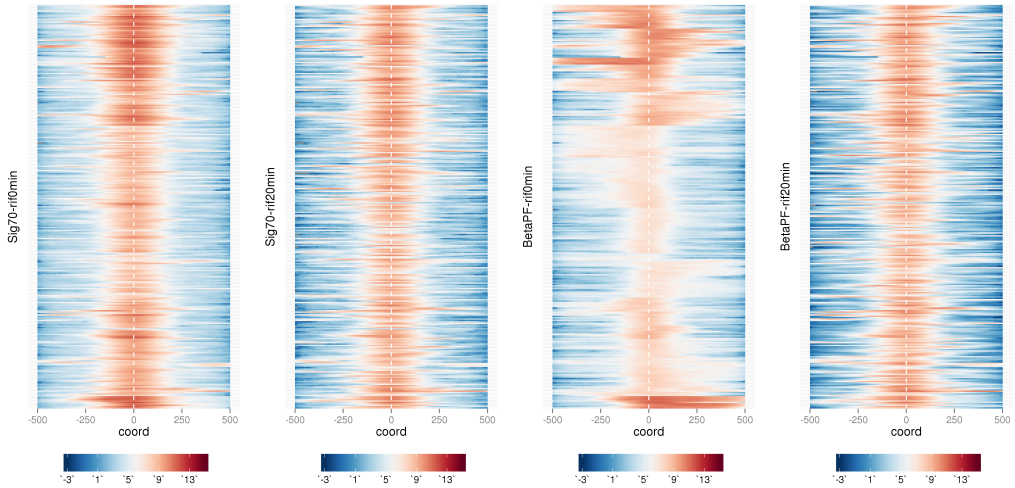

Figure 4.4: **Heatmap of the σ^{70} and β'_f peaks.** Clustered according to the β'_f sample when rif. treatment wasn't applied with `hclust` method and `ward.D2` strategy.

4.4 Discussion

We have introduced two `R` packages `ChIPUtils` and `Segvis` designed to facilitate the analysis of ChIP-seq and ChIP-exo/nexus data. The first, `ChIPUtils`, contains multiple utility

function to analyze these assays, from which the most used functions are the quality control metrics, as to the best of our knowledge there isn't an R package with these functions that can be integrated into an analysis pipeline. The second, `Segvis`, allows the user to quickly visualize multiple samples across the genome.

`Segvis` extends the Bioconductor's `GenomicRanges::GRanges` class, thus is not only capable of generating the plots shown in Section 4.3, but to integrate functions as `subset` to explore ChIP data interactively. These capabilities were inspired by our work with Wang et al. (2016), where our collaborators were interested in the binding patterns of histone ChIP-seq data over logical combinations of of EBNA2, EBNA3, and RBPJ peak overlaps. In other words, they asked whether the average histone profiles vary when different combinations of RBPJ and EBNA proteins bind to the genome.

5 CONCLUSIONS AND FUTURE WORK

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

In Chapter 2, we present a systematic exploration of several ChIP-exo/nexus datasets. By a rigorous exploration of the multiple characteristics of ChIP-exo/nexus samples, 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 main findings of this work are: First, we showed that commonly used quality control procedures / pipelines are not accurate for quantifying the quality of a given ChIP-exo/nexus sample. Second, the assumptions followed by quality control metrics that focus on library complexity and read distribution are challenged by ChIP-exo/nexus data because of the number of possible positions where a ChIP-exo/nexus read can map is dramatically reduced compared to the amount of possible positions where a ChIP-seq read can map. Third, we show that the strand imbalance is more likely to occur in low complexity libraries, thus it is a useful indicator for detecting low quality ChIP-exo samples. Finally, we provide a diagnostic plot and estimate a quality index by the analysis of the relationship between the enrichment and library complexity of a ChIP-exo/nexus sample.

5.2 Multi-Trait Fine-Mapping with Integrated Functional Annotation

In Chapter 3 we introduce an integrative fine-mapping method for variants in high LD, named `FM-HighLD`, that is capable of detecting causal variants for single and multiple traits. The key ideas behind our framework are that given a set of putative variants is possible to discern the pattern that associates summary statistics with annotation data generated from multiple assays, and that if there is a mechanism recovered by the annotation data that explains the association statistics, then is possible to select a set of candidate causal SNPs. For the multi-trait case, `FM-HighLD` utilizes linear mixed models to account for the relatedness among different traits, that way the common independence and identically distributed multiple traits assumption is relaxed. For this project we utilized for the first time MPRA data, which for this experiment provides a sequencing based gold-standard to evaluate fine-mapping methods by using real data, as contrasted with previous methods that were evaluated only by using simulated data. As with any other protocol, it is important to understand its limitations, thus we note that MPRAs may not always represent the neighboring chromatin structure of the variant because the assay isolates the sequence surrounding the variant from potential cofactors, this is argued by Tewhey et al. (2016) as they estimated the sensitivity of the assay to be in the $9\% - 24\%$ range.

The `FM-HighLD` framework provides many directions for useful extensions: First, by including a prior distribution over the annotation coefficients is theoretically possible to add model selection capabilities, this is specially important as the amount of data generated by NIH-consortia keeps increasing; Second, since our framework is based on the joint continuous and combinatorial optimization of equation 3.1 it is possible to improve our predictions by the use of more sophisticated optimization techniques such as genetic algorithms. Third, the number of covariates used in the annotation matrix can be increased by pooling multiple loci together. An improvement over our current framework could be to pool loci together according to the whole annotation matrix (e.g. by using hierarchical clustering over the SNPs and cutting the tree near the root). Finally, another extension is to increase the number of causal candidate variants per LD cluster.

In this project, we used ChIP-seq data to derive the allelic skew and peak percentage features, which would make our method only possible in immortalized cell lines for which

several ChIP-seq assays have been generated. However, the facts that the allelic skew is correlated with ATAC-seq skew (Supplementary Figure C.6, $\hat{\rho} = 0.7372$) and the rapid generation of these samples would mean that our method to be applicable on tissue-derived cell lines. With the rapid development of Hi-C assays, it could be possible in the future to design features in the future that code the relationship between and variant as part of the annotation data.

5.3 Computational tools for high-throughput data analysis

In Chapter 4 and Appendix A, we illustrated the use of computational tools for exploring high-throughput data. In Chapter 4 we introduced two `R` packages `ChIPUtils` and `Segvis` designed to facilitate the analysis of ChIP-seq and ChIP-exo/nexus data. The first, `ChIPUtils`, contains multiple utility function to analyze these assays, from which the most used functions are the quality control metrics, as to the best of our knowledge there isn't an R package with these functions that can be integrated into an analysis pipeline. The second, `Segvis`, enables the user to quickly visualize multiple samples across the genome. We showed that by using `Segvis` it is possible to summarize and visualize the binding patterns of multiple ChIP-seq assays across a set of pre-defined genomic regions (e.g. the context around SNPs of interest, or a TF-enriched ChIP-seq peaks). In Appendix A we illustrated an example of using a `shiny` application to explore results of an RNA-seq experiment. As more high-throughput data is becoming available, the structures to integrate data from multiple sources are becoming more complicated, thus we consider that the development and use of interactive tools emerge as a powerful alternative for the analysis and discovery of new hypothesis.

A USING SHINY APPLICATIONS FOR RNA-SEQ RESULTS EXPLORATION

A.1 Introduction

RNA-seq revolutionized the way in which transcripts are measured. However, interactive RNA-seq data analysis can be complicated without a knowledge on a programming language. In our experience, scientists rely on tools like Excel to explore their data. This task turns out to be incredibly difficult due to the large dimensions of the data. For example, after performing the typical steps of gene quantification, and keeping only genes with a certain amount of mapping reads, our collaborators' data is a table of approximately 16K genes and 12 samples.

In this chapter, we analyze RNA-seq data generated by the Johannsen Lab. The design is as follows:, 12 RNA-seq samples were generated from 3 cell lines, and 2 different treatment were applied. In other words, only 2 samples were available from each cell / treatment combination. Figure A.1 illustrates the design that utilized to generate the experiment. To allow our collaborators to explore their data, we propose the use of a `flexdashboard` (Iannone et al. (2018)) , enriched with `shiny` (Chang et al. (2018)) applications that summarize standard visualizations.

A.2 Methods

RNA-seq data analysis

Processing of the RNA-seq samples. We used `RSEM` (version 1.3, Li and Dewey (2011)) along with `Bowtie 2` (version 2.2.6, Langmead and Salzberg (2012)) to quantify the gene expression in the UCSC Genes track. Table A.1 states the total number of reads, the number of mapped reads and the percentage of aligned reads.

Differential expression analysis with DESeq2. We used `DESeq2` (version 1.2, Love et al. (2014)) to analyze our data. This package requires a matrix of count data, and the design formula to fit a negative binomial model for each gene, where the variance is estimated by borrowing information across all genes. We used the estimated counts by `RSEM`, and

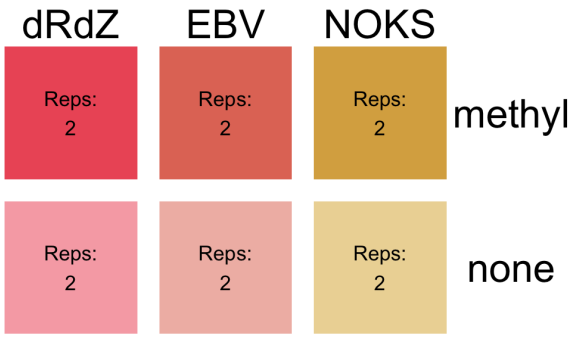

Figure A.1: **Diagram illustrating the design of the experiment.** 2 samples were generated for every treatment cell lines combination. Three cell lines were utilized Normal Oral Keratinocytes-Spontaneously (NOKS), and EBV infected NOKS (EBV), and an EBV mutant type (DRDZ). The samples were treated with methyl cellulose to induce differentiation.

removed all genes with less than 20 aligned reads. We considered a design with both cell and treatment effects, and the interaction between them, in other words we used the `R` formula:

```
~ cell + treatment + cell:treatment
```

Then, we used contrasts to test for the genes that were differentially expressed due to the treatment in a specific cell. Finally, we corrected for multiple hypothesis testing wit the Benjamini-Hochberg method.

A.3 Results

We provide a `flexdashboard` application, because this package makes available layouts that provides multiple related visualizations. For this analysis, we wanted to compare

Cell	Treatment	Rep.	# reads	# aligned	% aligned
ΔRΔZ	methyl	1	149,330,196	130,510,225	87.4%
ΔRΔZ	methyl	3	134,541,921	117,260,433	87.2%
ΔRΔZ	none	1	227,943,109	189,832,684	83.3%
ΔRΔZ	none	3	154,476,614	132,754,861	85.9%
EBV	methyl	1	128,908,667	114,943,279	89.2%
EBV	methyl	2	114,468,638	89,222,773	77.9%
EBV	none	1	162,214,708	146,977,179	90.6%
EBV	none	2	102,633,281	77,580,138	75.6%
NOKS	methyl	1	126,552,671	110,406,019	87.2%
NOKS	methyl	2	103,556,957	79,731,912	77.0%
NOKS	none	1	113,120,599	93,038,479	82.2%
NOKS	none	2	113,187,367	86,665,304	76.6%

Table A.1: **Alignment statistics for RNA-seq samples.**

the different results of the three contrast tests that we made. Thus, we used the following visualizations: i) an `hexbin` plot comparing the log fold changes between any pair of cell lines, ii) a volcano plot comparing the estimated log fold change vs. the p.value for each gene, iii) an MA plot (Irizarry et al. (2003)), and iv) a heatmap of the estimated `rlog` transformation (provided by `DESeq2`, the heatmap is interactive, as it was generated with `d3heatmap`, Cheng et al. (2018)); where for the last three, it is possible to select the cell line used for the contrast. Figure A.2 shows a screenshot of the dashboards main page.

In the left bottom corner of screenshot A.2, we can see three buttons: **Load**, **Save** and **Clean**, which can be used to operate gene list from previous sessions, or to search gene lists of interest.

Interactive analysis

We next show a quick example of how to use this dashboard to explore the data. Lets say, that we pick some genes in the volcano plots for each cell line. Those are coded in different colors: ΔRΔZ in blue, EBV in green, and NOKS in yellow. Figure A.3 illustrates the volcano plots for the EBV and NOKS contrasts, respectively.

Then, the user may be curious about those genes behavior in the MA plots or the hexbin

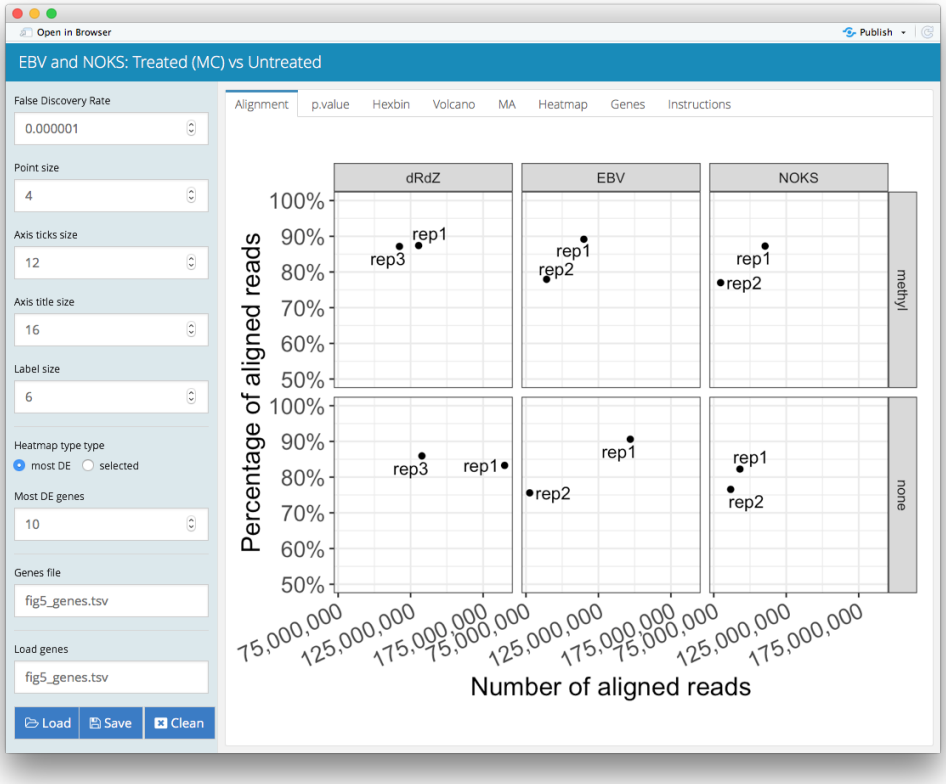

Figure A.2: **Screenshot of the MC vs. Untreated dashboard.** Starting tab of the `flexdashboard` application we generated to analyze the RNA-seq data.

plots. Figure A.4 illustrates the same group of genes, but observed into an MA plot for the ΔRΔZ contrast, and in the log fold change comparison of EBV vs. ΔRΔZ contrasts.

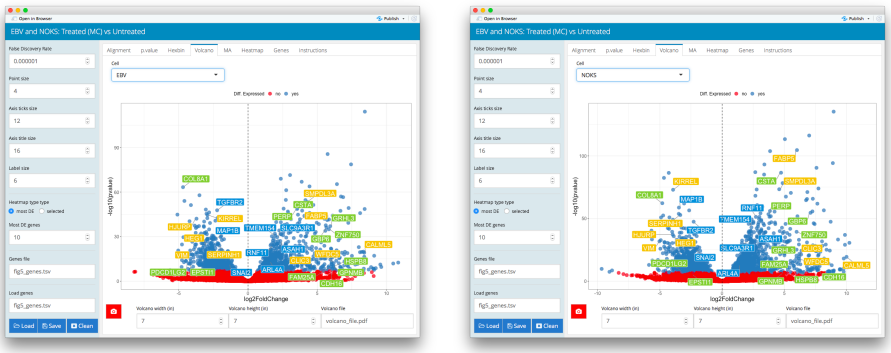

Figure A.3: **Volcano plots for EBV (left) and NOKS (right) contrasts.** Two screenshots of genes selected by the user, using the volcano plots. The contrasts of origin is coded by the colors: ΔRΔZ in blue, EBV in green, and NOKS in yellow.

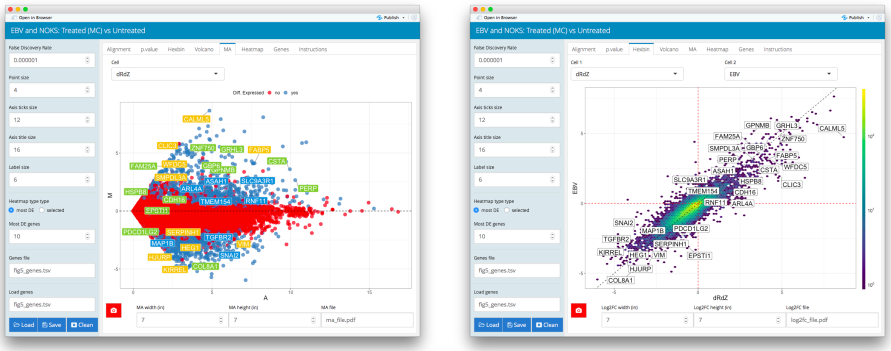

Figure A.4: **MA and hexbin plot.** Two screenshots of genes selected by the user, MA plot on the left and hexbin plot on the right.. The contrasts of origin is coded by the colors: ΔRΔZ in blue, EBV in green, and NOKS in yellow in the MA plot only.

A.4 Discussion

We propose to use interactive documents to allow our collaborators to explore the data as well. These app is not meant to replace a serious statistical analysis, but to facilitate the result exploration by i) make the analysis interactive, and available to our collaborators,

and ii) enhance the discovery of future hypothesis by sharing the process of exploring the results.

B APPENDIX OF DATA EXPLORATION, QUALITY CONTROL AND STATISTICAL ANALYSIS OF CHIP-EXO/NEXUS EXPERIMENTS

ChIP-exo/nexus datasets

Processing of the ChIP-exo and ChIP-nexus samples. We aligned the ChIP-exo/nexus samples in Table 2.2 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 Langmead et al. (2009)). We aligned the E1 samples in Table 2.1 with 'bowtie -q -m 1 -l 55 -k 1 -5 3 -3 40 –best' and the E2 samples using 'bowtie -q -m 1 -v 2 –best'. The average read length were 102 and 52 bps for the E1 and E2 samples, respectively. Hence, we 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 (Kuan et al. (2011)) 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 *E. coli* 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 towards 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‘ (Chung et al. (2013)) analysis identified one of more binding events were used in FSR assessments (Figure 2.2).

Supplementary figures

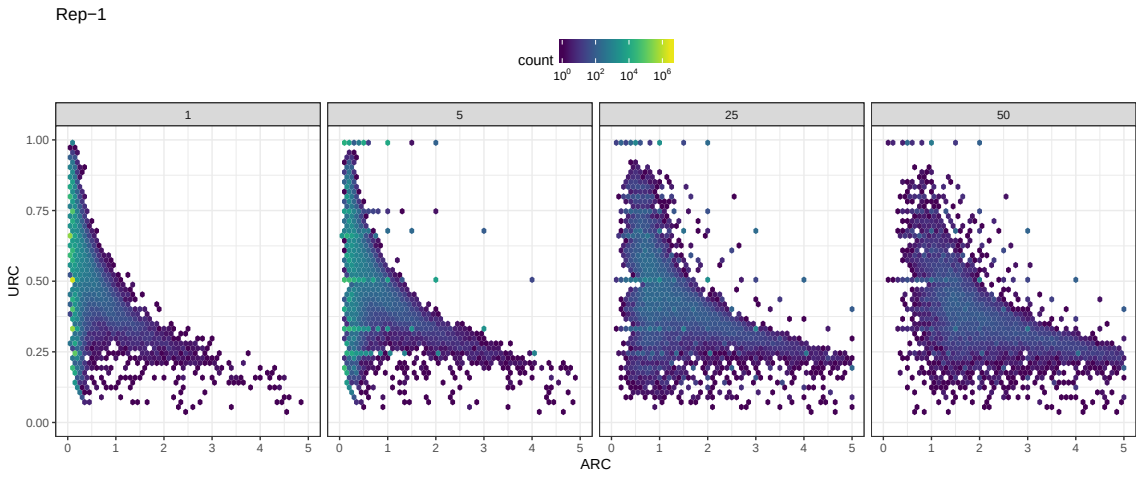

Figure B.1: **ARC vs. URC plot for FoxA1 replicate 1 in mouse liver.** ARC vs. URC plot obtained by using $h = 1, 5, 25, 50$

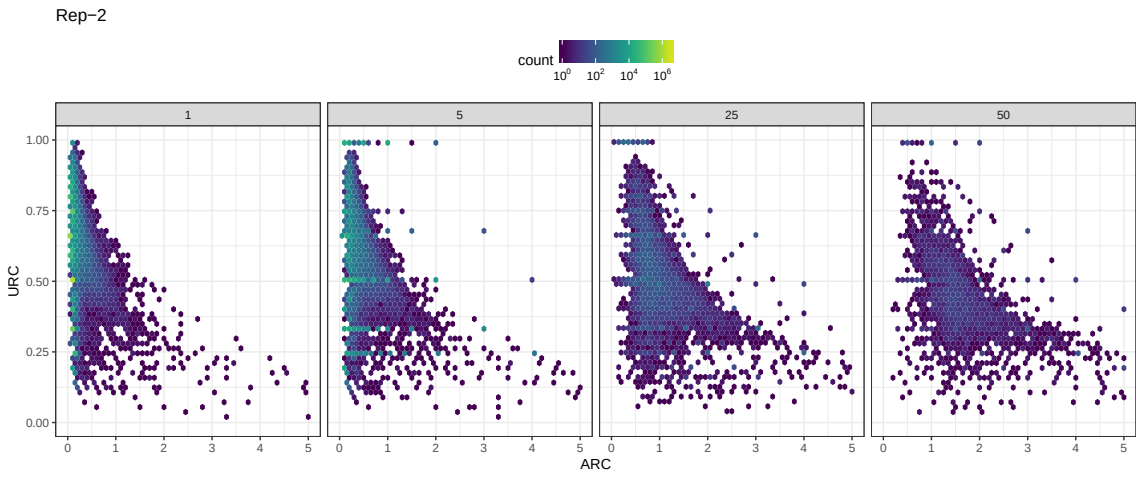

Figure B.2: **ARC vs. URC plot for FoxA1 replicate 2 in mouse liver.** ARC vs. URC plot obtained by using $h = 1, 5, 25, 50$

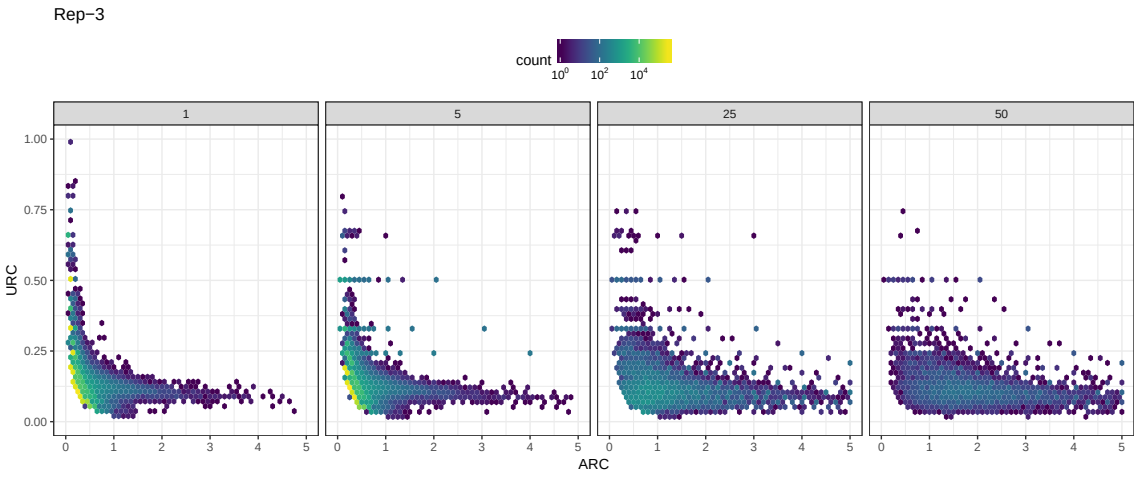

Figure B.3: **ARC vs. URC plot for FoxA1 replicate 3 in mouse liver.** ARC vs. URC plot obtained by using $h = 1, 5, 25, 50$

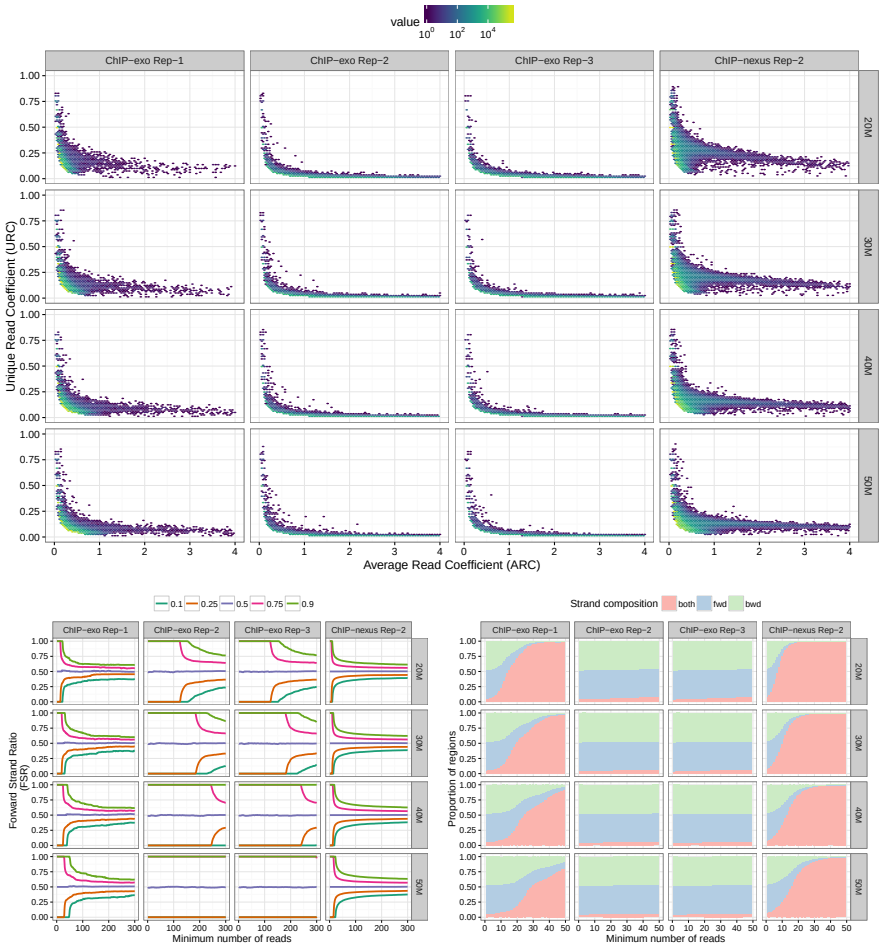

Figure B.4: **ChIPexoQual applied to subsampled TBP ChIP-exo/nexus samples.** ChIP-exo QC pipeline diagnostics applied to ChIP-nexus and ChIP-exo experiments of TBP factor in K562 human cell lines when sampling 20M to 50M reads from each experiment.

C APPENDIX OF EVALUATING MULTI-TRAIT FINE-MAPPING METHODS FOR VARIANTS IN HIGH LD WITH MASSIVELY PARALLEL REPORTER ASSAY AND ALLELE SPECIFIC ANNOTATION DATA

Derivation of the FM-HighLD algorithm

In this section, we show how the ECM was used to maximize the likelihood of the **FM-HighLD** model. Following the generative model described in equation 3.4. In both cases, we can rewrite it as:

$$
\begin{aligned}
w_q &\sim \text{Ber}(\pi) \\
\tilde{z}_q &\sim \begin{cases} N(0, \sigma_0^2), & w_q = 0 \\ N(\tilde{\mathbf{A}}_q, \sigma_q^2), & w_q = 1 \end{cases}
\end{aligned}
$$

for $q = 1, \cdots, Q$ where Q is the total number of associations and the variance can be a unique value $\sigma_q^2 = \sigma^2$ for the single trait case, or be different in each association $\sigma_q^2 = \sigma^2 + \tilde{\mathbf{B}}_q \Sigma \tilde{\mathbf{B}}^T$ in the multiple trait case.

The **FM-HighLD** framework is a mixture of regression models (linear models for single-trait or linear mixed models for multi-trait), thus the E-step require to calculate the posterior probabilities of the causal indicator variables $w_q, q = 1, \cdots, Q$ given the data, which are calculated by using the Bayes theorem:

Single trait: $\gamma_m = P(w_m = 1 \mid \tilde{z}_m, \mathbf{A}_m) \propto \pi N(\tilde{z}_m \mid \tilde{\mathbf{A}}_m \lambda, \sigma^2)$

Multiple traits: $\gamma_{tm} = P(w_{tm} = 1 \mid \tilde{z}_{tm}, \mathbf{A}_{tm}) \propto \pi N(\tilde{z}_m \mid \tilde{\mathbf{A}}_m \lambda, \sigma^2 + \tilde{\mathbf{B}}_{tm} \Sigma \tilde{\mathbf{B}}_{tm})$

Then, the M-step require to maximize the conditional expectation of the model's complete likelihood, which in the **FM-HighLD** case is given by:

$$
\begin{aligned}
Q(\Theta;\Theta^0) &= \mathbb{E}_{\Theta^0}\left[\log L_c(\Theta) \mid \mathbf{z}, \mathbf{A}\right] \\
&= \sum_{q=1}^{Q} \gamma_q \log \pi N(\tilde{z}_q \mid \tilde{\mathbf{A}}_q \lambda, \sigma_q^2) + (1-\gamma_q)\log[(1-\pi)N(\tilde{z}_q \mid 0, \sigma_0^2)] \\
&= f(\pi) + g(\lambda, \sigma^2)
\end{aligned}
$$

where σ^2 denotes all the variance parameters, and $\Theta = \left(\lambda, \pi, \sigma^2\right)$. Thus, the function Q is maximized when both f and g functions are maximized.

1. We derive the function f and set it equal to zero:

$$
\begin{aligned}
f(\pi) &= \sum_{q=1}^{Q} \gamma_q \log \pi + (1-\gamma_q)\log(1-\pi) \\
f'(\pi) &= \sum_{q=1}^{Q} \frac{\gamma_q}{\pi} - \frac{1-\gamma_q}{1-\pi} \\
f'(\pi) = 0 \iff (1-\pi)\sum_{q=1}^{Q} \gamma_q &= \pi\left(Q - \sum_{q=1}^{Q}\gamma_q\right) \iff \hat{\pi} = \frac{1}{Q}\sum_{q=1}^{Q}\gamma_q
\end{aligned}
$$

2. To maximize the function g, we notice that:

$$
g(\lambda, \sigma^2) = C(\sigma^2) - \sum_{q=1}^{Q}(1-\gamma_q)\frac{\tilde{z}_q^2}{2\sigma_0^2} - \sum_{q=1}^{Q}\gamma_q\frac{(\tilde{z}_q - \tilde{\mathbf{A}}_q\lambda)^2}{2\sigma_q^2}
$$

where $C = C(\sigma^2)$ depends only on the variance parameters. We rewrite the last term in their vector form:

$$
\sum_{q=1}^{Q}\gamma_q\frac{(\tilde{z}_q - \tilde{\mathbf{A}}_q\lambda)^2}{2\sigma_q^2} = (\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\lambda)^T\mathbf{D}(\tilde{\mathbf{z}} - \tilde{\mathbf{A}}\lambda)
$$

where $\mathbf{D} = \text{diag}\left(\frac{\gamma_1}{2\sigma_1^2}, \cdots \frac{\gamma_q}{2\sigma_q^2}\right)$, indicating that we can use the `lm` or `lme4:lmer` functions to fit a weighted regression model depending if there is only one or multiple traits to estimate λ and the variance parameters, except the background noise σ_0^2, which is estimated by:

$$\hat{\sigma}_0^2 = \frac{\sum_{q=1}^{Q}(1-\gamma_q)\tilde{z}_q^2}{Q - \sum_{q=1}^{Q}\gamma_q}$$

We are going to focus in the single-trait case to derive the closed form for the causal configuration matrix $\mathbf{C}$, as the multi-trait case performs the same strategy for all traits by separate. Without loss of generality, we assume that all variant in a LD cluster are grouped together, and define $\mathbf{r} = \mathbf{z} - \mathbf{A}\lambda$, then:

$$\mathbf{r} = \begin{bmatrix} \mathbf{z}_1 - \mathbf{A}_1\lambda \\ \mathbf{z}_2 - \mathbf{A}_2\lambda \\ \vdots \\ \mathbf{z}_M - \mathbf{A}_M\lambda \end{bmatrix}, \text{ and } \mathbf{C} = \begin{bmatrix} \mathbf{c}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{c}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{c}_M \end{bmatrix}$$

where each vector $\mathbf{c}_m \in \{0,1\}^{q_m}$ for $m = 1, \cdots, M$. Then, lets notice that maximizing the likelihood is equivalent to minimizing the squared error:

$$\min_{\mathbf{C}} \|\mathbf{C}^T(\mathbf{z} - \mathbf{A}\lambda)\|^2 = \min_{\mathbf{C}} \|\mathbf{C}^T\mathbf{r}\|^2 = \min_{\mathbf{c}_m \in \{0,1\}^{q_m}} \sum_{m=1}^{M} (\mathbf{c}_m \mathbf{r}_m)^2 \tag{C.1}$$

where the last equality follows from the identity:

$$\mathbf{C}^T\mathbf{r} = \begin{bmatrix} \mathbf{c}_1^T & 0 & \cdots & 0 \\ 0 & \mathbf{c}_2^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{c}_M^T \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_M\lambda \end{bmatrix}$$

Since equation C.1 is a sum of only non-negative terms, minimizing it is equivalent to minimize each term by separate. Thus, we notice that for every $m = 1, \cdots, M$:

$$(\mathbf{c}_m^T \mathbf{r}_m) = \left(\sum_{k=1}^{q_m} c_m[k] r_m[k] \right)^2 = \sum_{k=1}^{q_m} c_m[k] r_m^2[k]$$

because all entries of the $\mathbf{c}_m$ vectors are either 0 or 1. This implies that:

$$\min_{\mathbf{c}_m \in \{0,1\}^{q_m}} (\mathbf{c}_m \mathbf{r}_m)^2 = \min_{k=1,\cdots,q_m} r_m^2[k].$$

Supplementary figures

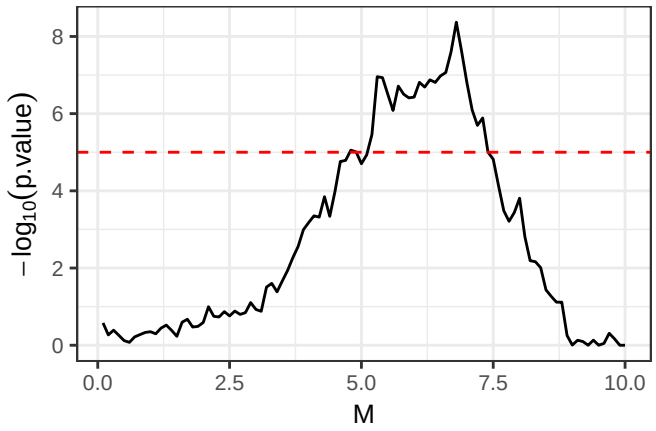

Figure C.1: **Binarized allelic skew.** p.values of χ^2 tests between expression modulating labels and binarized allelic skew. The binarized allelic skew is defined by the indicator function $1(|\text{ allelic skew }| \geqslant M)$, where M is a threshold ranging from 0 to 10. For $5 \leqslant M \leqslant 7.5$, the binarized allelic skew is more associated with the causal SNPs than peak overlaps for any assay.

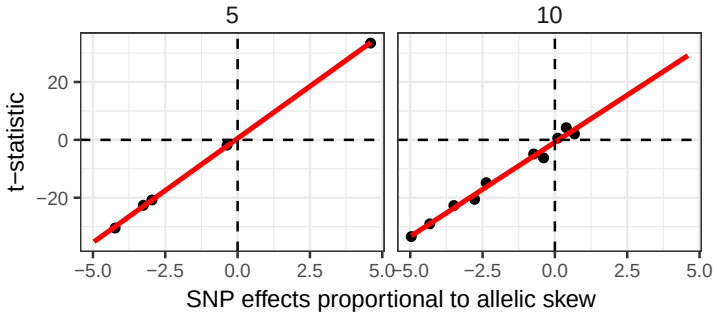

Figure C.2: **Comparison between SNP effects and t-statistics of a GWAS experiment with K causal SNPs in a polygenic model.** We simulated a continuous phenotype from a polygenic model with K variants, where their effects are defined as simulated allelic skew for the SNP. The t-statistics were calculated for each variant by separate.

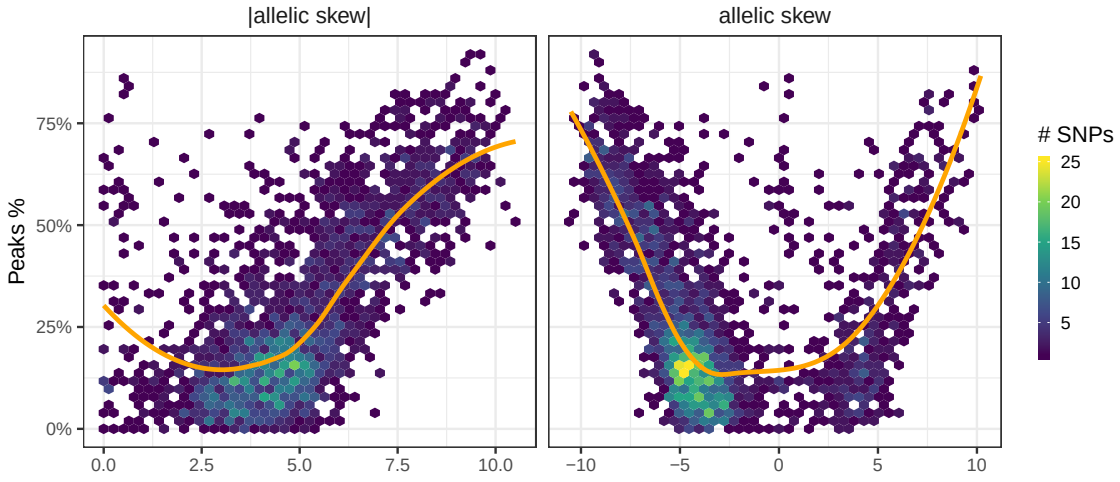

Figure C.3: **Comparison of allelic skew and peak percentage.** Comparison of the allelic skew and its absolute value with the percentage of ChIP-seq assay such that any of their peaks overlap the variant.

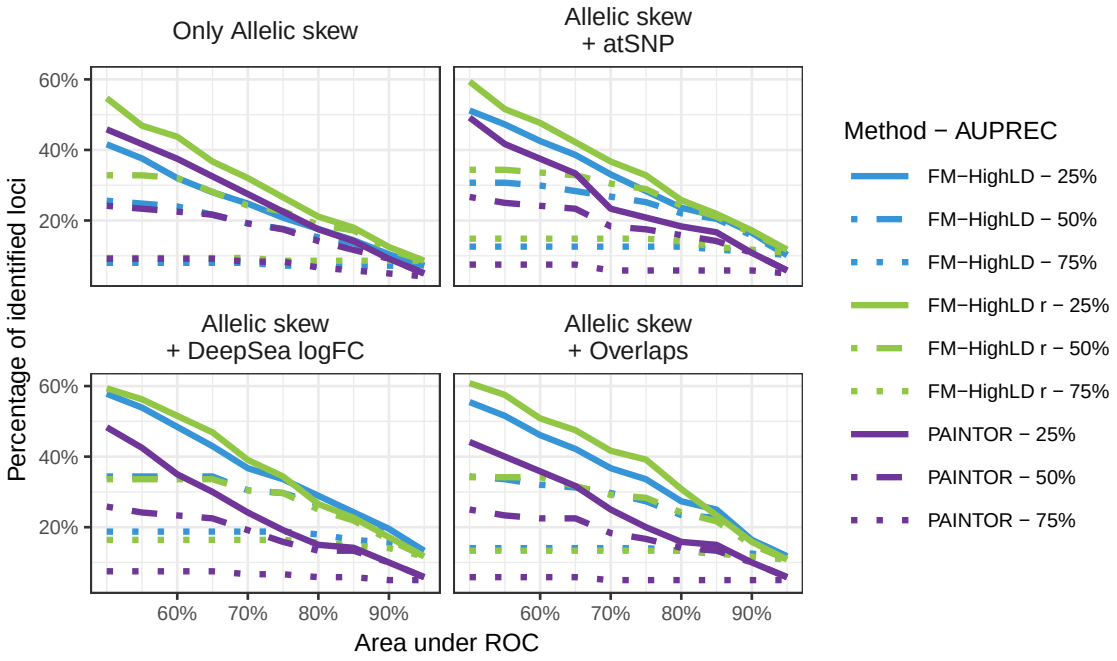

Figure C.4: **% of Identified loci comparison between FM-HighLD and PAINTOR.** % Identified loci(p, q) is defined as the ratio between the number of loci with AUROC $\geqslant p$ and AUPREC $\geqslant q$ and the amount of loci evaluated where AUROC and AUPREC are the areas under the ROC and Precision-Recall curves, respectively.

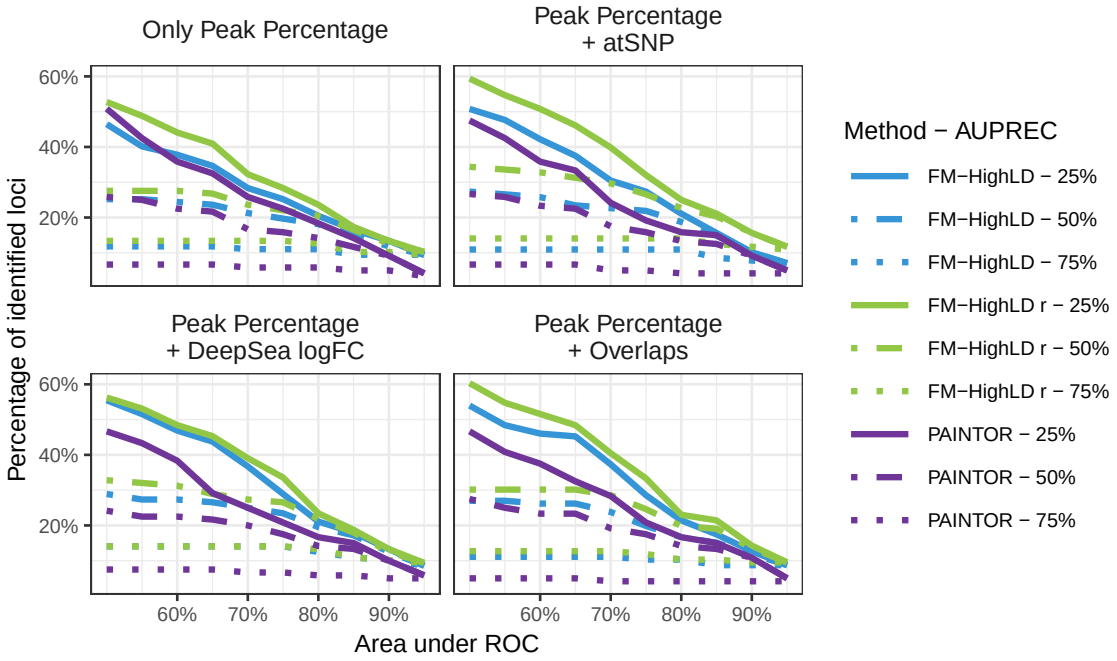

Figure C.5: **% of Identified loci comparison between FM-HighLD and PAINTOR.** % Identified loci(p, q) is defined as the ratio between the number of loci with AUROC $\geqslant p$ and AUPREC $\geqslant q$ and the amount of loci evaluated where AUROC and AUPREC are the areas under the ROC and Precision-Recall curves, respectively.

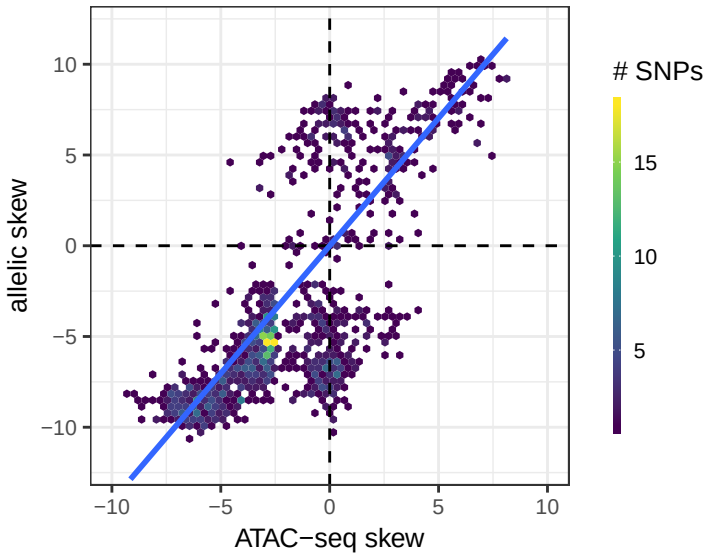

Figure C.6: **Comparison of ATAC-seq skew vs. allelic skew.**

REFERENCES

Bardet, Anaïs F, Jonas Steinmann, Sangeeta Bafna, Juergen A Knoblich, Julia Zeitlinger, and Alexander Stark. 2013. Identification of transcription factor binding sites from ChIP-Seq data at high resolution. *Bioinformatics* 29(21):2705–2713.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effect Models Using lme4. *Journal of Statistical Software* 67(1):1–48.

Benjamin, Yuval, and Terence P Speed. 2011. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* 40(10):e72.

Benner, Christian, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. 2016. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32(10):1493–1501.

Bioinformatics, Babraham. FASTQC. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* 10(12): 1213–1218.

Carroll, Thomas, Ziwei Liang, Rafik Salama, Rory Stark, and Ines de Santiago. 2014. Impact of artifact removal on ChIP quality metrics in ChIP-Seq and ChIP-exo data. *Frontiers in Genetics, Bioinformatics and Computational Biology* 15:75.

Chang, Wingston, Joe Cheng, JJ Allaire, Yihui Xie, Jonathan McPherson, RStudio, jQuery Foundation, Mark Otto, Jacob Thornton, Alexander Farkas, Scott Jehl, Stefan Petre, Andrew Rowls, Dave Gandy, Kristopher M Kowal, Denis Ineshin, Sami Samhuri, John Fraser, John Gruber, Ivan Sagalaev, and R Core Team. 2018. Shiny. `https://cran.r-project.org/web/packages/shiny/index.html`.

Chen, Wenan, Beth R Larrabee, Inna G Ovzyannikova, Richard B Kennedy, Iana H Haralambieva, Gregory A Poland, and Daniel J Schaid. 2015. Fine Mapping Causal Variants

with an Approximate Bayesian Method Using Marginal Test Statistics. *G3: Genes, Genomes, Genetics* 200(3):719–736.

Cheng, Joe, Tal Galili, Michael RStudio, Bostock, and Justin Palmer. 2018. d3heatmap. `https://cran.r-project.org/web/packages/d3heatmap/index.html`.

Chung, Dongjun, Dan Park, Kevin Myers, Jeffrey Grass, Patricia Kiley, Robert Landick, and Sündüz Keleş. 2013. dPeak, High Resolution Identification of Transcription Factor Binding Sites from PET and SET ChIP-seq Data. *PLOS Computational Biology* 9(10):e1003246.

Chung, Dongjun, Can Yang, Cong Li, Joel Gelernter, and Hongyu Zhao. 2014. GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleitropy and Annotation. *PLOS Genetics* 10(11):e1004787.

Consortium, Roadmap Epigenomics, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518:317–330.

Csardi, Gabor, and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal* 1695(5):1–9.

Degner, Jacob F, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482:390–394.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin. 1976. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B* 39(1): 1–38.

eGTEx Project. 2017. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nature, Genetics* 49:1664–1670.

ENCODE consortium. 2011. ENCODE QC metrics. `https://genome.ucsc.edu/ENCODE/qualityMetrics.html`.

Finucane, Hilary K, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yair Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzi Zang, Kyle Farh, Stephan Ripke, Felix R Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R B Perry, Yukinoria Okada, Soumya Raychaudhuri, Mark J Daly, Nick Patterson, Benjamin M Neale, and Alkes L Price. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature, Genetics* 47:1228–1235.

Grant, Charles, Timothy Bailey, and William S Noble. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017–1018.

Grossman, Sharon R, Kristian G Andersen, Ilya Shlyaakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, Daniel J Park, Dustin Griesemer, Elinor K Karlsson, Sunny H Wong, Moran Cabili, Richard A Adegbola, Rameshwar N K Bamezai, Adrian V S Hill, Fredrik O Vannberg, John L Rinn, 1000 Genomes Project, Eric S , Lander, Stephen F Schaffner, and Pardis C Sabeti. 2013. Identifying recent adaptations in large-scale genomic data. *Cell* 152(4):703–713.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* 550:204–213.

He, Qiye, Jeff Johnston, and Julia Zeitlinger. 2014. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology* 33:395–401.

Hormozdiari, Farhad, Emrah Kostem, Eun Y Kang, Bogdan Pasaniuc, and Eleazar Eskin. 2014. Identifying causal variants at loci with multiple signals of association. *Genetics* 198(2):497–508.

Huang, Weichun, Rasiah Loganantharaj, Bryce Schroeder, David Fargo, and Leping Li. 2013. PAVIS: a tool for peak annotation and visualization. *Bioinformatics* 29(23):3097–3099.

Iannone, Richard, JJ Allaire, Barbara Borges, RStudio, Abdullah Keen IO, Almsaeed, Jonas Mosbech, Noel Bossart, Lea Verou, Dmitry Baranovskiy, Sencha Labs, Bojan Djuricic, Tomas Sardyha, Bryan Lewis, Carson Sievert, Joshua Kunst, Ryan Hafen, Bob Rudis, and Joe Cheng. 2018. flexdashboard. `https://cran.r-project.org/web/packages/flexdashboard/index.html`.

Irizarry, Rafael, Bridget Hobbs, Francois Collin, Yasmin Beazer-Barclay, Kristen Antonellis, Uwe Scherf, and Terence Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2):249–264.

Ji, Hongkai, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. 2008. An integrated software system for analyzing ChIP-chip and ChIP-Seq data. *Nature biotechnology* 26:1293–1300.

Kasinathan, Sivakanthan, Guillermo A Orsi, Gabriel E Zentner, Kami Ahmad, and Steven Henikoff. 2014. High-resolution mapping of transcription factor binding sites on native chromatin. *Nature Methods* 2:203–209.

Kent, W James, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. 2002. The Human Genome Browser at UCSC. *Genome Research* 12:996–1006.

Khan, Aziz, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Cheneby, Shubhada Kulkarni, Ge Tan Damir

Baranasic, David J Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoit Ballester, Wyeth W Wasserman, Francois Parcy, and Anthony Mathelier. 2017. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acid Research* 46(D1):D260–D266.

Kharchenko, Peter, Michael Tolstorukov, and Peter Park. 2008. Design and analysis of ChIP-Seq experiments for DNA-binding proteins. *Nature, Biotechnology* 26:1351–1359.

Kichaev, Gleb, and Bogdan Pasaniuc. 2015. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *The American Journal of Human Genetics* 97:260–271.

Kichaev, Gleb, Megan Roytman, Ruth Johnson, Eleazar Eskin, Sara Lindstrom, Peter Kraft, and Bogdan Pasaniuc. 2016. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* 33(2):248–255.

Kichaev, Gleb, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc. 2014. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLOS Genetics* 10(10): e1004722.

Kreimer, Anat, Aoyang Zeng, Matthew D Edwards, Yuchun Guo, Kevin Tian, Sunyoung Shin, Rene Welch, Michael Wainberg, Rahul Mohan, Nicholas A Sinnot-Armstrong, Yue Li, Göksen Eraslan, Talal Bin Amin, Ryan Tewhey, Pardis C Sabeti, Jonathan Goke, Nikola S Mueller, Manolis Kellis, Anshul Kundaje, Michael A Beer, Sündüz Keleş, David K Gifford, and Nir Yosef. 2017. Predicting gene expression in massively parallel reporter assays: A comparative study. *Human Mutation Variation, Informatics, and Disease* Special Issue: The Critical Assessment of Genome Interpretation(38):1240–1259.

Krueger, Felix, and Simon R Andrews. 2016. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Research* 5:1479.

Kuan, Pei Fen, Dongjun Chung, Guangjin Pan, James A Thomson, Ron Stewart, and Sündüz Keleş. 2011. A Statistical Framework for the Analysis of ChIP-seq data. *Journal of the American Statistical Association* 106(495):891–903.

Kuhn, Max. 2008. Building predictive models in R using the Caret package. *Journal of Statistical Software* 28(5):1–26.

Landt, Stephen G, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J Hartemink, Michael M Hoffman, Vishwanath R Iyer, Youngsook L Jung, Subhradip Karmakar, Manolis Kellis, Peter V Kharchenko, Qunhua Li, Tao Lio, X Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M Myers, Peter J Park, Michael J Pazin, Marc D Perry, Debasish Raha, Timothy E Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John A Stamatoyannopoulos, Michael Y Tolstorukov, Kevin P White, Simon Xi, Peggy J Farnham, Jason D Lieb, Barbara J Wold, and Michael Snyder. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* 22:1813–1831.

Langmead, Ben, and Steven L Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359.

Langmead, Ben, Cole Trapnell, Pop Mihal, and Steven L Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10:R25.

Lappalainen, Tuuli, Michael Sammeth, Marc R Friedländer, Peter A ć Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Puliakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, The Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralk Sudbrak, Ángel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas MEitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506–511.

Li, Bo, and Colin N Dewey. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12(1):323.

Li, Yue, and Manolis Kellis. 2016. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichment across multiple complex human diseases. *Nucleic Acids Research* 44(18):e144.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15(12):550.

Madrigal, Pedro. 2015. CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates. *EMBnet.journal* 21:e837.

Mahony, Shaun, and Franklin Pugh. 2015. Protein-DNA binding in high-resolution. *Critical Reviews in Biochemistry and Molecular Biology* 50(4):269–283.

Marinov, Georgi K, Anshul Kundaje, Peter J Park, and Barbara J Wold. 2014. Large-Scale Quality Analysis of Published ChIP-seq data. *G3 Genes, Genomes, Genetics* 4(2):209–223.

Mathelier, Anthony, Oriol Fornes, David J Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, Allen W Zhang, François Parcy, Boris Lenhard, Albin Sandelin, and Wyeth W Wasserman. 2016. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* 44(D1):D110–D115.

Maurano, Matthew T, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfield, Anthony Raubitscheck, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337:1190–1195.

Meng, Xiao-Li, and Donald B Rubin. 1993. Maximum Likelihood Estimation via the ECM algorithm: A general framework. *Biometrika* 80(2):267–278.

Nica, Alexandra C, Stephen B Montgomery, Antigone S Dimas, Barbara E Stranger, Claude Beazley, Inês Barroso, and Emmanouil T Dermitzakis. 2010. Candidate Causal Regulatory

Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLOS Genetics* 6(4):e1000895.

Nicolae, Dan L, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dilan, and Nancy J Cox. 2010. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLOS Genetics* 6(4):e1000888.

Pickrell, Joseph. 2014. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *American Journal of Human Genetics* 94:559–573.

Planet, Evarist, Camille Stephan-Otto Attolini, Oscar Reina, and David Rosell. 2011. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* 28(4):589–590.

Reddy, Timothy, Jason Gertz, Pauli Florencia, Katerina Kucera, Katherine E Varley, Kimberly M Newberry, Georgi K Marinov, Ali Mortazavi, Brian A Williams, Lingyun Song, Gregory E Crawford, Barbara Wold, Huntington F Willard, and Richard M Myers. 2012. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research* 22:860–869.

Reshef, Yakir A, Hilary K Finucane, David R Kelley, Alexander Gusev, Dilan Kotliar, Jacob C Ulirsch, Farhad Hormozdiari, Joseph Nasser, Luke O'Connor, Bryce van de Geijn, Po-Ru Loh, Shari Grossman, Gaurav Bhatia, Steven Gazal, Pier F Palamara, Luca Pinello, Nick Patterson, Ryan P Adams, and Alkes L Price. 2017. Detecting genome-wide directional effecs of transcription factor binding on polygenic disease risk. *bioRxiv*.

Rhee, Ho Sung, Alain R Bataille, Liye Zhang, and B Franklin Pugh. 2014. Subnucleosomal Structures and Nucleosome Assymetry across a Genome. *Cell* 159:1377–1388.

Rhee, Ho Sung, and Franklin Pugh. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 147:1408–1419.

Rozowsky, Joel, Ghia Euskirchen, Raymond Auerbach, Zhengdong Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark Gerstein. 2009. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nature Biotechnology* 27:66–79.

Saito, Takaya, and Marc Rehmsmeier. 2017. Precrec: fast and accurate precision-recall ad ROC curve calculations in R. *Bioinformatics* 33(1):145–147.

Salmon-Divon, Mali, Heidi Dvinge, Kairi Tammoja, and Paul Bertone. 2010. PeakAnalyzer: Genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* 11(1):415.

Schaid, Daniel J, Wenan Chen, and Nicholas B Larson. 2018. From genome-wide associations to candidates causal variants by statistical fine-mapping. *Nature Reviews Genetics* 19: 491–504.

Schaub, Marc A, Alan P Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. 2012. Linking disease associations with regulatory information in the human genome. *Genome Research* 22:1748–1759.

Serandour, Aurelien, Brown Gordon, Joshua Cohen, and Jason Carroll. 2013. Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biology* 14(12):R147.

Shin, Sunyoung, and Sündüz Keleş. 2016. Annotation Regression for Genome-Wide Association Studies with an Application to Psychiatric Genomic Consortium Data. *Statistics in Biosciences* 9(1):50–72.

Skene, Peter J, and Steven Henikoff. 2015. A simple method for generating high-resolution maps of genome-wide protein binding. *eLIFE* 4:e09225.

Starick, Stephan R, Jonas Ibn-Salem, Marcel Jurk, Céline Hernandez, Michael I Love, Ho-Ryun Chung, Martin Vingron, Morgane Thomas-Chollier, and Sebastiaan H Meijsing. 2015. ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Research* 25(6):825–835.

Tewhey, Ryan, Dylan Kotliad, Daniel S Park, Brandon Liu, Sarah Winnicki, Steven K Reilly, Kristian G Andersen, Tarjei S Mikkelsen, Eric S Lander, Stephen F Schaffner, and Pardis C Sabeti. 2016. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reported Assay. *Cell* 65(6):1519–1529.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.

Udler, Miriam S, Jonathan Tyrer, and Douglas F Easton. 2010. Evaluating the Power to Discriminate Between Highly Correlated SNPs in Genetic Association Studies. *Genetic Epidemeology* 34(5):463–468.

Ulirsch, Jacob C, Satish K Nandakumar, Li Wang, Felix C Giani, Xiaolan Zhang, Peter Rogov, Alexandre Melnikov, Patrick McDonel, Ron Do, Tarjei S Mikkelsen, and Vijay G Sankaran. 2016. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* 65(6):1530–1545.

Valouev, Anton, David Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard Myers, and Arend Sidow. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods* 5(9):829–834.

Venters, Bryan J, and Franklin Pugh. 2013. Genomic organization of human transcription initiation complexes. *Nature* 502(7469):53–58.

Wang, Anqi, Rene Welch, Bo Zhao, Tram Ta, Keleş Sündüz, and Eric Johannsen. 2016. Epstein-Barr Virus Nuclear Antiger 3 (EBNA3) Protein Regulate EBNA2 Binding to Distinct RBPJ Genomic Sites. *Journal of Virology* 90(6):2906–2919.

Wang, Liguo, Junshend Chen, Chen Wang, Liis Uusküla-Reimand, Kaifu Chen, Alejandra Medina-Rivera, Edwin J Young, Michael T Zimmermann, Huihuang Yan, Zhifu Sun, Yuji Zhang, Stephen T Wu, Haojie Huang, Michael D Wilson, Jean-Pierre A Kocher, and Wei Li. 2014. MACE: model based analysis of ChIP-exo. *Nucleic Acids Research* 42(20):e156.

Welch, Rene, Dongjun Chung, Jefrey Grass, Robert Landick, and Sündüz Keleş. 2017. Data exploration, quality control and statistical analysis of ChIP-exo/nexus experiments. *Nucleic Acids Research* 45(15):e145.

Wen, Xiaoquan, Yeji Lee, Francesca Luca, and Roger Pique-Regi. 2016. Efficient Integrative Multi-SNP association Analysis via Deterministic Approximation of Posteriors. *The American Journal of Human Genetics* 98(6):1114–1129.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yu, Guangchuang, Li-Gen Wang, and Qing-Yu He. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* 31(14): 2382–2383.

Zeng, Xin, Bo Li, Rene Welch, Constanza Rojo, Ye Zheng, Colin N Dewey, and Sündüz Keleş. 2015. Perm-Seq: Mapping Protein-DNA Interactions in Segmental Duplication and Highly Repetitive Regions of Genomes with Prior-Enhanced Read Mapping. *PLOS, Computational Biology* 11(10):e1004491.

Zhou, Jian, and Olga G Troyanskaya. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* 12(10):931–934.

Zuo, Chandler, Sunyoung Shin, and Sunduz Keles. 2015. atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* 31(20):3353–3355.