COMPUTATIONAL APPROACHES TO STUDYING GENE REGULATION USING CHROMATIN ACCESSIBILITY AND GENE EXPRESSION ASSAYS

Bryan C. Quach

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology in the School of Medicine.

Chapel Hill
2017

Approved by:

Terrence Furey

Greg Crawford

Samir Kelada

Karen Mohlke

Praveen Sethupathy

William Valdar

ABSTRACT

Bryan C. Quach: Computational approaches to studying gene regulation using chromatin accessibility and gene expression assays
(Under the direction of Terrence Furey)

The completion of the Human Genome Project marked the beginning of a new era in genomics characterized by significant improvements in high-throughput sequencing technology and the development of new sequencing-based assays to study a wide array of functional elements and biological properties at the genome-wide scale. These advancements were accompanied by the formation of large, multi-institutional consortia that produced publicly available data sets and functional genomic studies that broadened our understanding of the genome. Previously uncharacterized genomic regions became recognized as important components of gene regulation, but the broader knowledgebase of regulatory elements raised new questions to elucidate the growing complexity of gene regulation models. Additionally, quantitative trait loci (QTL) mapping approaches began taking advantage of quantitative sequencing data to study the impacts of genetic variation on molecular phenotypes such as gene expression at the genome-wide level. The popularity of high-throughput methods for studying gene regulation and transcription lead to a data deluge that necessitated new statistical methods and bioinformatics solutions for data management, processing, analysis, visualization, and interpretation. Specialized research areas emerged to better glean insights from sequencing data leading to new challenges and questions. In the following chapters, I present a novel machine learning framework for genomic footprinting, a concept focused on identifying transcription factor (TF) binding sites using chromatin accessibility sequencing data. I demonstrate that my framework outperforms existing methods for classifying TF binding sites via footprinting. In addition, I investigate characteristics of TF binding sites within chromatin accessibility data and assess technical factors that influence footprinting to provide an improved understanding of the strengths and limitations of using these data for TF binding site prediction. Through a separate study, I investigate the impact of a genotoxic chemical 1,3-butadiene on chromatin accessibility and gene expression in a population of genetically diverse mice. I perform expression QTL (eQTL) and chromatin accessibility QTL (cQTL) mapping in these mice and detect eQTLs and cQTLs in each tissue. In all, the work herein demonstrates multiple computational approaches to studying various gene regulatory relationships and provides insight on the efficacy of these approaches to inform future studies.

ACKNOWLEDGEMENTS

Firstly, I want to say thank you to my PhD adviser Terry Furey for being a great mentor and helping shape me as a scientist over the past five years. From troubleshooting broken code and effectively visualizing results to thinking through the biological interpretation of an analysis, your wisdom and intuition taught me the necessary skills and mindset to approach the small details of a problem while also seeing the big picture. I am thankful for the intentionality that you put into mentoring and teaching students and the sheer number of hours you have endured of me taxing your brain about research ideas and scientific musings during our meetings.

I also would like to thank my committee members Greg Crawford, Samir Kelada, Karen Mohlke, Praveen Sethupathy, and Will Valdar for all of your helpful advice and suggestions throughout various stages of this project. Each of you provided distinct and equally important perspectives to improve how I framed and approached problems. Additionally, without the help of members from Greg Crawford and Ivan Rusyn's labs, none of the 1,3-butadiene project research in chapter III would have been possible. I want to especially recognize Grace Chappell, Lauren Lewis, and Alexias Safi. The ATAC-seq and RNA-seq mouse data would not exist without their hard work in running the mouse experiments and prepping samples.

I also extend my gratitude to the current and former members of the Furey Lab for their helpful advice and encouragement whenever I hit roadblocks in my research. Special thanks to Karl Eklund, Jeremy Simon, and Paul Cotney for their work in building the Furey Lab sequencing data processing pipelines that have proven to be immense time savers. In addition, Greg Keele and Dan Oreper from the Valdar Lab provided software and expertise that expedited the analysis process for the QTL mapping detailed in chapter III. Financial support for this dissertation was provided by

NIGMS grant T32GM067553, NIEHS grant R01ES023195, and the Lineberger Comprehensive Cancer Center's University Cancer Research Fund.

Lastly, thank you to my friends and family near and far for all your love and support throughout my graduate training. Dad, thanks for always encouraging me in my educational pursuits and teaching me the life skills that have allowed me to survive under my own roof. Brad and Jen, thanks for playing host during my visits so that I could experience some of Europe and recharge my mind without breaking the bank. Matt and Luann, thank you for your love and encouragement and giving me two awesomely entertaining nephews to wrestle and build Legos with when I visit. Cherie, I am extremely grateful for your constant support and love. Thanks for always listening to how my day went and being so attentive. You always brighten my day with your spontaneous food adventures, and without you, I likely would have spent the last month of graduate school sustaining myself on ramen noodles.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ATAC-seq: assay for transposase-accessible chromatin

AUC: area under the curve

BD: 1,3-butadiene

bp: base pair

CC: Collaborative Cross

cDNA: complementary DNA

ChIP-seq: chromatin immunoprecipitation and sequencing

cQTL: chromatin accessibility quantitative trait locus

DeFCoM: Detecting Footprints Containing Motifs

DNA: deoxyribonucleic acid

DNase-seq: DNaseI sequencing

dsQTLs: DNaseI sensitivity quantitative trait locus

ENCODE: Encyclopedia of DNA Elements

eQTL: expression quantitative trait locus

FPR: false positive rate

FRiP: fraction of reads in peaks

GEO: gene expression omnibus

GTEx: Genotype-Tissue Expression

GWA: genome-wide association

HMM: hidden Markov model

LCL: lymphoblastoid cell line

LD: linkage disequilibrium

Mb: megabase

pAUC: partial area under the curve

PC: principal component

PCA: principal component analysis

ppm: particles per million

PWM: position weight matrix

QTL: quantitative trait locus

RBF: radial basis function

RNA: ribonucleic acid

SNP: single nucleotide polymorphism

SVM: support vector machine

TAMU: Texas A&M University

TF: transcription factor

TFBS: transcription factor binding site

TPM: transcripts per million

UCSC: University of California at Santa Cruz

UNC-CH: University of North Carolina at Chapel Hill

CHAPTER I

Introduction

The importance of gene regulation in cell development and biological homeostasis of living organisms has been well recognized [1–4]. Through the Human Genome Project [5], technological innovations and a broader understanding of genome organization and composition paved way for large-scale efforts in the genomics community to better understand functional genomic elements and the role of non-coding DNA in transcriptional regulation [6,7]. Although these efforts improved understanding of gene regulatory components such as promoters, enhancers, silencers, chromatin structure, and transcription factors, they also increased awareness of the complexity of regulatory dynamics and the interactions between the various components. Furthermore, follow-up studies to quantitative trait loci (QTL) mapping and genome-wide association (GWA) studies that detect trait-associated genetic variation contributed another layer of regulatory complexity by characterizing relationships between genetic variation and regulatory changes as intermediate mechanistic links between DNA sequence and phenotype [4,8]. The increasing availability of information, resources, methodologies, and technologies for studying gene regulation highlighted a growing opportunity and significance in further identifying regulatory elements and studying their roles in condition-specific contexts.

IDENTIFYING REGULATORY ELEMENTS GENOME-WIDE WITH HIGH-THROUGHPUT ASSAYS

Since the advent of Sanger sequencing, DNA sequencing technology continued to improve, and the introduction of massively parallel “next-generation” sequencing approaches revolutionized biological and biomedical science research by enabling the development of higher-throughput and more cost-effective alternatives to microarrays to assay biological properties such as transcription, nucleosome occupancy, chromatin interactions, transcription factor (TF) binding, and histone modifications genome-wide [9]. Although multiple different next-generation sequencing platforms exist, they share some commonalities in their approach. Each technology first requires the preparation of a sequence library through the ligation of oligonucleotide adapters to the ends of the DNA fragments to be sequenced. The fragments are then amplified and undergo a platform-specific sequencing reaction that allows the classification of each nucleotide. The ability for these reactions to occur simultaneously leads to the high-throughput that makes them massively parallel. The nucleotide readouts, referred to as reads, generate large quantities of data that then require the application of bioinformatics approaches for downstream processing, analysis, and interpretation. These next-generation sequencing platforms remain widely used, however newer sequencing platforms are being developed such as nanopore sequencing that rely on different sequencing chemistry and do not require fragment amplification [10].

From a simplified perspective, sequencing platforms all share the goal of accurately classifying the nucleotide sequence of the given fragments. The major distinctions in the sequencing-based methods for assaying different biological properties occur in isolating the relevant DNA or RNA. For example, Chromatin Immunoprecipitation Sequencing (ChIP-seq) aims to detect genomic locations of TF occupancy or histone modifications. To do this, binding proteins and genomic DNA are cross-linked, then the DNA is fragmented. Immunoprecipitation with a protein-specific antibody retrieves the protein-bound sequences that are then sequenced. Enrichment of reads mapping to a particular genomic location indicates TF occupancy (or histone modification) [11]. In DNaseI sequencing (DNase-seq), chromatin accessibility is assayed using the exonuclease DNaseI. Exposing genomic DNA to DNaseI results in the enzyme preferentially cutting DNA in more accessible, nucleosome-depleted regions. Following DNaseI digestion, size selected DNA fragments are sequenced and genomic regions with enrichment of mapped reads are classified as accessible chromatin regions [12]. In both ChIP-seq and DNase-seq, the biomolecule initially being isolated is DNA. With RNA sequencing (RNA-seq), RNA transcripts are initially isolated as opposed to DNA. For compatibility with sequencing platforms, these transcripts are typically converted to cDNA before sequencing, although some direct RNA sequencing approaches exist [13]. The reads from RNA-seq are mapped to their originating genes and can be analyzed to deduce estimates of RNA abundance.

With the three aforementioned methods, the diversity of biological properties related to gene regulation that can now be studied genome-wide created new opportunities for understanding their interactivity. ChIP-seq, DNase-seq, and RNA-seq among other methods were utilized by the Encyclopedia Of DNA Elements (ENCODE) project which sought to characterize all of the functional elements in the human genome [7] and in later stages also included the mouse genome [14]. In a 2012 report, the ENCODE project had produced 1,640 data sets in 147 different human cell types [15], and a 2014 mouse ENCODE publication comparing the mouse and human functional elements reported over 1,000 data sets in 123 mouse cell types and primary tissues [14]. Analyses by the ENCODE consortium found that 80.4% of the human genome is covered by at least one functional element. Of this fraction, RNA-associated elements and histone modifications comprised a large majority, and 15.2% of the coverage was attributed to DNaseI hypersensitive sites [15]. In comparisons with mouse functional elements, chromatin state landscapes and TF networks were found to be relatively stable between human and mouse [14]. Additionally, gene expression profiles were shown to be more consistent within tissue than within species [16]. To build upon the work by the ENCODE project, the more recent Roadmap Epigenomics Project constructed a collection of epigenomic profiles for 127 human tissues and cell types from adult and embryonic samples [17]. Analyses of these data showed associations between proximal and distal regulatory regions, histone marks, DNA methylation, chromatin accessibility, spatial organization, and gene expression that play important roles in cell type identity, development, and disease [17]. Taken together, the catalogue of genomic and epigenomic data and integrative analyses from these large-scale projects contributed new insights into the organization and regulation of human and mouse genes and the genome and continues to serve as an expansive public resource for biomedical research.

GENE EXPRESSION AND CHROMATIN ACCESSIBILITY AS QUANTITATIVE TRAITS

A fundamental challenge in genetics research is to understand genetic variation and its relationship to phenotypic variability. Efforts such as the International HapMap and 1000 Genomes Project extensively characterized common genetic variation across diverse human populations [18,19], and GWA studies have leveraged advancements in genotyping technology to link genetic variants to human traits and diseases. Although informative in many regards, these studies do not resolve the underlying biological mechanisms of discovered genotype-phenotype associations. For functional follow-up, data produced by the ENCODE and Roadmap Epigenomics consortia have served as valuable resources to refine lists of candidate GWAS variants and identify putative roles of non-coding variants [20], but these data still do not directly assess the impact of inter-individual variation on gene regulation and cellular behavior that results in the observed phenotypes.

A related but distinct approach from GWAS is expression QTL (eQTL) mapping. In eQTL mapping, gene expression levels are treated as quantitative traits and tested for associations with genetic variants. The first reported eQTL study analyzed over 1,500 genes and 3,312 genetic markers between two strains of *Saccharomyces cerevisiae* [21]. Since then, eQTL mapping has been performed in various contexts using model organisms and humans [22–24]. With RNA-seq (or gene expression microarrays) and current genotyping approaches, these analyses can include tens of thousands of genes, each regarded as an independent quantitative molecular phenotype. The Genotype-Tissue Expression (GTEx) Project pilot analysis demonstrated the utility of eQTL analyses by performing eQTL mapping in 9 human tissues and identifying eQTLs shared and unique to each. Significant eQTLs were compared to GWAS disease-related single nucleotide polymorphisms (SNPs) showing whole-blood specific eQTL enrichment for autoimmune-related GWAS variants [24]. This showed that by directly modeling the relationship between genetic variation and gene expression, eQTL mapping serves as a powerful tool to gain more insight into gene regulatory changes that can then be used to elucidate other genotype-phenotype links.

As a complementary approach to eQTL mapping, the genetic underpinnings of chromatin variation have been studied using sequencing-based assays. Kasowski *et al.* observed variation between lymphoblastoid cell lines (LCLs) from 19 individuals for histone modifications H3K27ac, H3K4me1, H3K4me3, H3K36me3, and H3K27me3. Work by McVicker *et al.* further assessed the genetic relationship to histone modifications by identifying SNPs significantly associated with variation of histone mark signals in LCLs derived from 10 unrelated individuals [25]. Similarly, Degner *et al.*, used DNase-seq to measure chromatin accessibility in 70 LCLs and detected 8,902 chromatin regions where chromatin accessibility was significantly associated with genotype, which they referred to as DNaseI sensitivity QTLs (dsQTLs). The dsQTLs discovered were found to be pre-dominantly local with enrichments for predicted TF binding sites. Sixteen percent of dsQTLs were also classified as eQTLs, and 55% of identified eQTLs were also dsQTLs. More recently, another genome-wide chromatin accessibility assay was developed called Assay for Transposase-Accessible Chromatin Using Sequencing (ATAC-seq) which relies on the Tn5 “tagmentation” process to fragment DNA at accessible chromatin regions and append adapters for sequencing [26]. Using ATAC-seq and genotype data from 24 European individuals, Kumasaka *et al.* reported 2,707 chromatin accessibility QTLs (cQTLs) which were also enriched for eQTLs and dsQTLs [27]. These QTL analyses using histone marks and chromatin accessibility data as quantitative traits demonstrate how chromatin assays can contribute to discovering associations between genotype and gene regulation that can ultimately inform physiologic or disease phenotype-genotype associations.

THE COLLABORATIVE CROSS AS A RESOURCE FOR GENETICS STUDIES

In human genetics and genomics studies, certain constraints limit the possible experimental designs that can be practically realized. As a proxy, various species such as *Danio rerio* (zebrafish), *Drosophila melanogaster* (fruit fly), and *mus musculus* (mouse) have been studied as model organisms to infer aspects of human biology [28–30]. In a 2002 review, Threadgrill *et al.* outlined propositions made by the Complex Trait Consortium to develop a mouse genetics resource for effective study of complex traits using QTL approaches [31]. The design and implementation of creating this resource became known as the Collaborative Cross (CC) [32]. The CC involved an international, multi-institutional effort to create a multiparent panel of recombinant inbred mouse strains derived from five classical inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, and NZO/HlLtJ) and three wild-derived strains (CAST/EiJ, PWK/PhJ, and WSB/EiJ) denoted as "founders". Because CC strains are inbred, they provide an advantage over human studies in that each strain can produce genetically identical individuals. This reduces the genotyping burden and allows for more sophisticated experimental designs to study multiple variables within the same population.

As described in [32], creating a CC strain requires a funnel breeding scheme that begins with the mating of the 8 founder strains in pairs. Two pairs from the resulting generation are then mated, and this process continues for subsequent generations until a final inbred CC strain is produced. By permuting the pairs in the initial generations, a large number of strains can be constructed. In an evaluation of the genome architecture of 350 CC strains, similar founder haplotype representation was observed when averaged across the CC lines, but deviations from expected frequencies were noted when focusing on specific genomic regions. Unlike many classical inbred strains, the CC population did not exhibit high levels of long-range linkage disequilibrium (LD). This type of LD has been reported to increase false positives in association mapping studies

[33]. As a proof-of-concept, Aylor *et al.* performed eQTL mapping using 156 incipient inbred CC lines (pre-CC) and detected 7,235 liver eQTLs at less than 1 megabase (Mb) resolution. A QTL study by Kelada *et al.* used 131 pre-CC lines to identify genetic associations with blood cell volume, white blood cell count, percentage of neutrophils, and monocyte number [34]. More recently, 45 CC strains were used to identify liver eQTLs and QTLs associated with treatment response to the drug tolvaptan. The study showed strain-specific variability in liver toxicity phenotypes and found several candidate susceptibility genes for tolvaptan drug-induced liver injury [35]. Each of these studies demonstrates the feasibility and power of the CC as a resource for QTL mapping and interrogating genetic factors in disease and complex traits.

The significant advancements in systems genetics and functional genomics have made the intricacies of gene regulation more apparent, fostering new hypotheses for how the contributing components interact [3,36,37]. The development of sequencing-based assays such as those used by ENCODE and the Roadmap Epigenomics Project made new types of analyses possible, but in doing so exposed new questions and challenges to address. Among these challenges is the development of bioinformatics approaches and statistical methods to manage, process, analyze, and interpret the vast quantities of biological data being generated. For instance, the development of DNase-seq and ATAC-seq for detecting accessible chromatin also led to observations that these methods could probe TF binding locations through an approach called footprinting [26,38], but the strengths and weaknesses of footprinting have not been well characterized. As previously mentioned, the utility of the Collaborative Cross for QTL mapping has been demonstrated, but the advantages of the CC can be further demonstrated by experimental designs and analyses that interrogate both chromatin accessibility and gene expression under varying environmental conditions.

In chapter II, I introduce a novel method for TF binding site prediction, Detecting Footprints Containing Motifs (DeFCoM), that integrates DNase-seq or ATAC-seq data with ChIP-seq data and TF sequence motifs [39]. I use ENCODE data in conjunction with TF motif predictions to compare

DeFCoM to existing approaches and show that it outperforms other methods. I also evaluate current assumptions about chromatin accessibility signal characteristics at TF binding sites and assess the impact of technical factors on footprinting. In chapter III, I present an unpublished analysis that compares lung, liver, and kidney gene expression and chromatin accessibility for a control group of CC mice and mice exposed to the chemical 1,3-butadiene. I also characterize eQTLs and cQTLs in the three tissues to provide a basis for further studies investigating genetic associations with gene expression and chromatin accessibility in the CC population. In chapter IV, I discuss how my findings in Chapter II contribute to evaluating footprinting and integrating it into gene regulation studies, and I conclude the chapter discussing the significance of how my findings in analyzing CC mice contribute to interrogating environmental exposure and gene regulation in future CC studies.

CHAPTER II

DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter[1]

OVERVIEW

Identifying the locations of transcription factor binding sites is critical for understanding how gene transcription is regulated across different cell types and conditions. Chromatin accessibility experiments such as DNaseI sequencing (DNase-seq) and Assay for Transposase Accessible Chromatin sequencing (ATAC-seq) produce genome-wide data that include distinct "footprint" patterns at binding sites. Nearly all existing computational methods to detect footprints from these data assume that footprint signals are highly homogeneous across footprint sites. Additionally, a comprehensive and systematic comparison of footprinting methods for specifically identifying which motif sites for a specific factor are bound has not been performed.

Using DNase-seq data from the ENCODE project, I show that a large degree of previously uncharacterized site-to-site variability exists in footprint signal across motif sites for a transcription factor. To model this heterogeneity in the data, I introduce a novel, supervised learning footprinter called DeFCoM (Detecting Footprints Containing Motifs). I compare DeFCoM to nine existing methods using evaluation sets from four human cell-lines and eighteen transcription factors and show that DeFCoM outperforms current methods in determining bound and unbound

[1] A version of this work was previously published as Quach B, Furey TS. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. Bioinformatics. 2016;33: btw740.

motif sites. I also analyze the impact of several biological and technical factors on the quality of footprint predictions to highlight important considerations when conducting footprint analyses and assessing the performance of footprint prediction methods. Lastly, I show that DeFCoM can detect footprints using ATAC-seq data with similar accuracy as when using DNase-seq data.

INTRODUCTION

Chromatin dynamics vary based on developmental stage [40], cell type [41], and environmental stress [42]. Transcription factors (TFs) bind DNA in regions of accessible chromatin and play a central role in pre-transcriptional gene regulation. Understanding these interactions is critical in deciphering transcriptional regulation that defines cell identity in different contexts. DNase-seq [12] and ChIP-seq [43] identify regions of accessible chromatin and TF binding genome-wide, respectively. Notably, Hesselberth et al. observed that DNase-seq produces “footprints” at active TF binding sites characterized by a relative depletion of DNase-seq signal at these sites [44]. Thus, a single DNase-seq experiment captures high-resolution TF binding information for many TFs. As performing ChIP-seq for multiple TFs quickly becomes cost prohibitive, DNase-seq footprinting offers an enticing alternative.

Several computational footprint identification methods, which I will refer to as “footprinters”, have been developed [38,45–53]. These footprinters embrace one of two philosophies, which I denote as de novo and motif-centric footprinting (see Table 2.1 for an overview of methods). Models generated by de novo footprinters assume that there exist general data characteristics at footprint sites. These TF-agnostic models are used to predict all footprint sites, and then motif databases are queried to determine potential TFs bound in each individual footprint. In contrast, motif-centric footprinters first generate a set of candidate TF binding sites (TFBSs) based on a motif, and then predict at which motif sites a footprint exists, indicating active binding. Within each group, current methods exhibit similarities in approach. For instance, the de novo footprinters DBFP, HINT, and the HMM-based method described in [38] model footprints using probabilistic graphical models with similar state representations. FOS, Wellington, and DNase2TF are de novo footprinters that search for genomic locations akin to short inverse peaks. The motif-centric footprinters CENTIPEDE, msCentipede, and FLR utilize two-component mixture models to represent bound and unbound sites. In addition to DNase-seq data, some methods allow for the integration of complementary information such as histone modification status or distance from the nearest transcription start site. All these methods implicitly or explicitly assume there exists two distinct signal patterns in DNase-seq data that distinguish TF-bound and unbound sites. Except for msCentipede, footprinters expect that DNase-seq signal is highly homogeneous in both the bound and unbound groups and thus can be represented by a single model. This assumes TFs bind DNA in the same manner genome-wide, but TF binding behavior can vary across TFBSs [54].

More recently, Kahara and Lahdesmaki proposed a supervised classification approach, BinDNase, that learns TF-specific DNaseI cleavage patterns from training data to predict footprints in other data [46]. They show that their supervised approach often produced superior prediction accuracy over two unsupervised generative models, PIQ and CENTIPEDE. In contrast, Gusmao et al. conducted a systematic footprinter comparison and found most generative model footprinters outperformed BinDNase [55]. In their analysis, footprint detection accuracy was evaluated within a de novo footprinting framework based on overlap with ChIP-seq peak annotations. It is not clear how accurately this evaluates motif-centric footprinter performance.

Here, I conducted an in-depth, motif-centered analysis of DNaseI digestion signals and DNase-seq footprinters to provide a more complete understanding of strengths and weaknesses of current methods. I introduce a novel motif-centered method, Detecting Footprints Containing Motifs (DeFCoM), that approaches footprint identification using a nonlinear supervised classification framework. Importantly, DeFCoM is designed to capture variation in DNaseI signal within active footprints and unbound motif sites to enhance footprint classification accuracy, a consideration unaccounted for in previous footprinters. I compared the performance of DeFCoM against both de novo and motif-centric footprinting approaches across eighteen TFs in four cell-lines using data from the Encyclopedia of DNA Elements (ENCODE) Project [7] and show that DeFCoM outperforms existing approaches overall. In addition, I analyzed the variability in accuracy across multiple TFs and the effect of data quality and DNase-seq sequencing depth. Lastly, I show DeFCoM can detect footprints in data from Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) experiments with similar classification accuracy as with DNase-seq data

MATERIALS AND METHODS

Data and software

DNase-seq and ChIP-seq data (Tables 2.2 and 2.3) were obtained from the UCSC (University of California at Santa Cruz) ENCODE portal (https://www.genome.ucsc.edu/ENCODE/). ATAC-seq data for GM12878 [26] was obtained from GEO (Gene Expression Omnibus) using identification code GSE47753. The DAC Blacklisted Regions and Duke Excluded Regions for hg19 were downloaded from the UCSC Genome Database then combined into one set.

DeFCoM utilizes the Python packages PySam v0.9.0 and scikit-learn v0.17 [56]. The R package ROCR [57] was used for computing performance statistics and the ROC curves for the footprinters. F-Seq [58] was used to call peaks for DNaseI hypersensitive sites.

Generating cell-line specific motif sites

Sets of motifs labeled as active (TF-bound) or inactive (TF-absent) were generated as follows: 1) Transcription factor motif position weight matrices were downloaded from http://compbio.mit.edu/encode-motifs/ [59]. Motif occurrences were identified across the hg19 genome using FIMO (MEME v4.9.0) [60] with a genomic background nucleotide distribution pre-computed by FIMO and the parameters "--max-strand --max-stored-scores 1000000 --no-qvalue".

2) Predicted motif sites were removed if (i) they fell in ENCODE blacklisted regions, (ii) less than 10% of bases within a 200 bp window centered on the motif center had DNase-seq digestion data; (iii) they were less than 400 bp from chromosome boundaries; or (iv) there were ambiguous nucleotide calls within 400 bp of the motif site center. 3) Motif sites were annotated as active if they overlapped ChIP-seq peaks for that TF, or else they were labeled inactive. If multiple motif sites overlap the same peak region, only the site closest to the annotated point-source of the peak was retained. To further ensure inactive sites were not bound, I calculated ChIP-seq and input control signal enrichments, defined as s_{TF} - $s_{control,}$ where s_{TF} and $s_{control}$ are sequencing-depth normalized read density values in 200 bp windows centered on the motif. Inactive sites where $s_{TF} - s_{control} > 0$ were removed. Motif sets were created for 18 TFs (CEBPB, CHD2, CTCF, EP300, GABPA, JUN-D, MAFK, MAX, MYC, NRF1, RAD21, REST, RFX5, SRF, SP1, TAF1, TBP, USF2) in 4 human cell-lines (GM12878, H1-hESC, HepG2, and K562) except SP1 in K562 (no data).

Computing aggregate DNaseI digestion profiles

To create TF-specific summary statistics for each class of motif sites, I first generate the active and inactive motif site sets as detailed above. If multiple motifs exist for a TF, only one was chosen. For each class of motif sites, I constructed a matrix of DNaseI digestion frequencies where each row represents a unique motif site in the genome and each column represents a position within or flanking a motif site. All the rows were aligned based on the center of the motif site. DNaseI cut frequencies are denoted in DNase-seq data as the number of 5' read ends aligning at a given genomic position. To remove motif sites with spurious spikes in DNaseI activity, any rows of the matrix with a value exceeding 500 were removed. From these matrices all summary statistics were computed per column. For the aggregate DNaseI cut profiles, I used calculated mean cut frequencies. Likewise, per-column mean and standard deviations were computed to obtain coefficients of variation values.

DNaseI signal profiles and correlations

Aggregate DNaseI signal profiles were calculated for active and inactive motif sites for each TF in each cell type. DNaseI signal correlations for NRF1 were performed using only sites corresponding to the PWM (position weight matrix) "disc_1", for CHD2 using motif "disc_1", and for CEBPB using motif "known_1" (Figure 2.1) to ensure variability was not due to multiple motifs. Motif sites were extended 50 bp from the motif center and signal profiles were calculated. To remove sites with spurious spikes in DNaseI activity, motif regions with more than 500 DNase-seq reads were removed. Profiles were smoothed using 7 bp sliding windows to improve signal quality at sites with sparse signal. Aggregate mean DNaseI signal profiles for active and inactive sites were created using smoothed individual profiles. Pairwise Pearson correlation coefficients between active and inactive motif DNaseI profiles were used for complete-linkage hierarchical clustering followed by heatmap visualization.

DeFCoM feature extraction and training

DeFCoM (Detecting Footprints Containing Motifs) is an SVM (support vector machine)-based [61] supervised footprinter . Given a set of motif sites labeled as active or inactive for a given TF in a cell type/experimental condition, the SVM classifier is trained on features that are derived from DNase-seq data from the same cell type for each motif site. The trained model is used to predict active and inactive sites in a test set based only on DNase-seq data.

To train DeFCoM, motif site sets of size *m* and *n*, labeled as active or inactive respectively, were generated as described above (see *Generating cell-line specific motif sites*). The 5' end of each DNase-seq read was considered a digestion site. Initial active and inactive motif site DNaseI digestion count matrices, $D^{Active}{}_{ms}$ and $D^{Inactive}{}_{ns}$, were calculated, in which each row corresponded to a scaled DNaseI digestion profile consisting of the square root of the DNaseI digestion frequency at each position in an *s*-sized region centered on a motif site. For all the training and evaluation tests, *s*=200 bp regions were used. To account for spurious spikes in the data, any row in the matrix with a value greater than √500 was removed.

Intuitively, I wished to generate DNase digestion features in windows around a motif site, with smaller windows used near the motif site where the TF binds to allow for greater resolution, and progressively larger windows used at more distant regions. I also wanted to account for sparse or noisy DNaseI data. Given the region size *s*, I first defined varying-sized, non-overlapping, contiguous windows symmetric about the motif site center. Let *x*∈*{0,1,2,...,k}* index each window starting at the motif site center with the windows progressively increasing in size from 0 to *k*. I define *f(x)*, the size of window *x*, to be

$$f(x) = \begin{cases} x^2 + 5, & x < k \\ \left(\frac{s}{2} - g(x-1)\right) + x^2 + 5, & x = k \end{cases} \qquad (1)$$

$$g(x) = \begin{cases} g(x-1) + x^2 + 5, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad (2)$$

where the recursive function *g(x)* equals the sum total size of all windows up to and including window *x*. The total number of windows *k* that will span a region of size *s/2* can be calculated as follows:

$$\underset{k}{\operatorname{argmin}} \left(\frac{s}{2} - g(k)\right) \mid \frac{s}{2} - g(k) \geq 0 \qquad (3)$$

In equations 1 and 3, I use *s/2* because windows are symmetric about the motif center. For *s*=200, I defined 12 windows (6 on each side of the motif site center) with sizes 45, 21, 14, 9, 6, 5, 5, 6, 9, 14, 21, and 45. For each window, I computed the mean of the scaled DNaseI digestion counts and the slope of these counts across the window using D^{Active} and $D^{Inactive}$. This generated a feature vector *f* of length *4k*. To provide additional global features of the region *s,* I partitioned a 90 bp segment centered on the motif center into 3 windows, computed the mean and slope for these windows (6 features total), and calculated the mean cut frequency of a 150 bp region centered on the motif center (1 feature). Lastly, maximal absolute value scaling was used to scale each of the $4k + 7$ features to a [-1,1] range. This results in the final feature matrices F^{Active} and $F^{Inactive}$.

As part of the training process, DeFCoM selects between a linear and radial basis function (RBF) kernel SVM to use as the final classifier. To decide between the two SVM models, I bootstrapped 1000 samples 100 times from each of F^{Active} and $F^{Inactive}$ and applied 5-fold cross validation. I used the mean pAUCs (5% FPR) from the cross validations to select a model.

Training a soft-margin SVM requires the selection of a hyperparameter, which I denote as c, that specifies a tolerance threshold for the number of samples from either class that lie on the wrong side of the separating hyperplane. The higher the value of c, the more heavily misclassification is penalized during model training. Additionally, the RBF kernel contains a parameter that I denote as γ, which determines the distance of influence of the chosen support vectors. Higher values of γ specify a smaller distance of influence. For both the cross validation and cross cell-line tests, DeFCoM performs a grid search to find the best c and γ. The values used in the grid search were $c \in \{0.01, 0.1, 1, 10, 100, 1000, 10000\}$ and $\gamma \in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$.

For within cell-line tests, the SVM type (linear or RBF kernel) is pre-specified based on the analysis being performed. I applied 5-fold nested cross validation using annotated motif sites and DNase-seq data for the specified cell-line, and all evaluation statistics were computed for each fold then averaged across folds. In the cross cell-line setting, training the final SVM for DeFCoM is a two-stage process. First, a linear or RBF kernel SVM is chosen along with c and/or γ values. Then, a subset of 3000 samples from each class is chosen to train the selected SVM model. Because the number of total samples typically is much larger than these subsets, I select the SVM type and the c and γ values using a bootstrapping procedure. I take 1000 random samples from each motif site class 100 times, and for each bootstrap iteration, I apply 5-fold cross validation to both a linear and

RBF kernel SVM using the aforementioned grid of c and/or γ values. Following the bootstrapping, I compare the distributions of pAUCs generated by each SVM type using a two-sided Student's t-test. I selected the RBF kernel when there was a statistically significant difference ($\alpha \leq 0.01$) and the linear SVM otherwise. Following SVM type selection, I chose final c and/or γ values based on which values were selected the most frequently during the bootstrap procedure for the selected SVM type. To improve the computational efficiency of the SVM training phase, the chosen SVM was trained with 3000 randomly selected samples from each of F^{Active} and $F^{Inactive}$ to produce the final trained model.

For ATAC-seq data, the D^{Active} and $D^{Inactive}$ matrices were constructed using Tn5 transposase tagmentation events as opposed to DNaseI digestion frequencies. Tn5 tagmentation sites are denoted as 5' ATAC-seq read ends offset 5 bp downstream on the positive DNA strand and 4 bp upstream on the negative strand.

Footprinter implementations for comparative analysis

The footprinters BinDNase, CENTIPEDE, cut density, DNase2TF, HINT, FOS, msCentipede, PIQ, and Wellington (Table 2.1) were used to evaluate DeFCoM. These methods were chosen based on availability, compatibility with my evaluation framework, and their broad range of conceptually diverse approaches to footprinting. I outline below how these methods were applied in a motif-centric evaluation framework. Any footprinter not listed was applied with no modifications and default settings.

BinDNase

Similar to DeFCoM, BinDNase is a supervised footprinter. For the training phase of BinDNase, 3000 samples from each class of motif sites were randomly chosen. The remaining parameters were the same as described in [46].

CENTIPEDE

In implementing CENTIPEDE I used the default parameters specified by [50] with the exception that the prior included only PWM scores.

Cut Density

Cut density serves as a straightforward “baseline” model for footprinting. It simply sums the number of DNase-seq 5’ read ends that map within a specified genomic region. For each motif site in the evaluation sets, cut density was computed for regions spanning 50 bp upstream and downstream of the motif site center.

DNase2TF

We ran DNase2TF on motif sites that were extended by 100 bp in both directions to obtain an initial list of footprint calls. The “FDRs” parameter was set to 1 with default values for the other parameters. I filtered the footprints to only those that overlapped at least 75% of a motif site. If the footprint region is smaller than the motif site, then it was also retained regardless of percent overlap. For each motif site, I assigned it the score from the overlapping footprint. If multiple footprints correspond to a motif site, I selected the highest score. If no footprint is associated with the motif site then it was given the minimum possible score.

HINT

We applied HINT similarly to DNase2TF. Using default settings, an initial list of footprints was generated by evaluating motif sites that were extended by 100 bp in both directions. These were filtered to footprints smaller than their corresponding motif site and footprints overlapping at least 75% of a motif site. Motif sites were assigned scores using the same process as for DNase2TF.

FOS

FOS computes a score based on a depletion of reads within a central window of length *c* base pairs compared to a left and right flanking window each of length f base pairs. With each motif site, I calculated an FOS score for all combinations of c and f where c is an integer between 6 and z and $f \in \{3,4,...,11\}$. Let m represent the length of a motif, then $z=2*(21-m)$ when m is less than or equal to 18 and 6 otherwise. I aligned c to be centered over the motif site. I retained the highest score from all the calculations for a motif site. Sites FOS failed to score were given the lowest possible score.

Wellington

Similar to FOS, Wellington uses a center and flanking region to compute a score and call footprints. Wellington searches for footprints in a region using a combination of a 35 bp flank size and center sizes 11, 13, 15, 17, 19, 21, 23, and 25. I allowed Wellington to score sites using input regions that were 49 bp flanks from the center of the motif site. The maximum of the absolute values of scores was used as the footprint score for the associated motif site.

Effective sequencing depth

Signal-to-noise was measured using FRiP (fraction of reads in peaks) scores [62]. Peaks were called using F-Seq with default parameters, then the ratio of DNase-seq reads aligning within the top 50,000 peaks (ranked by F-Seq score) to the total aligned reads was calculated. This ratio was multiplied by the total aligned reads to obtain the effective sequencing depth.

Subsampled sequencing depth analysis

To compare DeFCoM's performance in two cell-lines with similar effective sequencing depths but different signal-to-noise ratios, I applied downsampling to both GM12878 and H1-hESC DNase-seq data. In each cell-line I used SAMTools to downsample the data to 25, 50, 75, and 100 million mapped reads. At each sequencing depth, I converted the labeled motif sites and DNase-seq data into feature vectors. I then used these feature vectors for 5-fold nested cross validation of

DeFCoM with the RBF kernel SVM. Lastly, the mean pAUCs (5% FPR) from the folds were computed for 18 transcription factors.

RESULTS

Aggregate DNaseI digestion profiles do not capture motif site heterogeneity

Aggregate mean DNaseI digestion profiles summarize positional DNaseI cleavage preferences at TFBSs. These profiles convey a single value at each position, thus they lack information regarding the variability in DNaseI activity at a given position across sites. Raj *et al.* showed that variation in DNaseI activity at TF-bound SP1 motif sites exceeded that expected under a multinomial model of DNaseI digestion signal [51]. To evaluate this more broadly, I determined positional variability in DNaseI digestion signal for multiple TFs (Figures 2.2A and 2.3). I stratified motif sites into active and inactive based on presence of corresponding ChIP-seq signal for the factor in the same cell type. I used these to evaluate two common assumptions held by several footprinting methods: 1) active TFBSs possess a general footprint pattern of local depletion in DNaseI digestion relative to flanking regions; and 2) inactive motif sites contain approximately uniformly distributed DNaseI digestion signal. For most factors, aggregate profiles for active sites clearly produced expected DNaseI digestion patterns, but with relatively large standard deviations. An investigation of individual binding sites clearly shows how sites deviate from the aggregate pattern (Figures 2.2C and 2.2D). In some cases, the previously characterized sequence preferences for DNaseI digestion [63] are visually apparent. For a minority of the TFs, the aggregate profile for active sites portrays a visually weak footprint or none at all (i.e. SRF, Figure 2.3). Overall, TFs exhibit aggregate profiles with consistently high coefficients of variation (Figure 2.4).

In spite of position-specific variability across motif sites, it is possible that DNaseI signal at individual sites resemble the aggregate profile in shape but not scale. To quantify the similarity of DNaseI digestion profiles at individual sites to the aggregate mean profiles, I calculated Pearson correlation coefficients between the aggregate profiles and every individual TFBS profile for CEBPB, CHD2, and NRF1 (Figure 2.5). Among the 3 TFs, 30-63% of the individual profiles did not correlate with the same class aggregate profile (Pearson's r < 0.1). Interestingly, I found that 17-51% of individual profiles from the active and inactive classes exhibited stronger positive correlations with the aggregate profile from the opposite class.

To further assess within and between class heterogeneity, I computed Pearson correlations between the top 2000 individual DNaseI digestion profiles, ranked based on the number of DNase-seq reads in a 100 bp window centered on the motif site, in the active and inactive classes for all three factors. I observed small clusters of highly correlated sites, implying possible subgroupings for DNaseI cleavage profiles within each class. I also found 34-53% of motif sites within each class exhibited negative or no correlation to each other (Pearson's r < 0) (Figures 2.2D and 2.6). Notably, 4-6% of correlations between sites from opposite classes had Pearson's r > 0.5. These analyses of variability in DNaseI digestion signal strongly indicate that aggregate mean profiles do not sufficiently capture the heterogeneity in DNaseI activity across motif sites.

We hypothesized that high correlations between sites from one class to the aggregate profile of the opposite class may be partially attributed to similarities in binding preferences for multiple TFs. Therefore, a motif site deemed inactive for a specific TF based on ChIP-seq data could be active for another TF with a similar motif. I assessed this by determining how many inactive motif sites overlapped ChIP-seq peaks for at least one other TF for each of 18 TFs in the K562 cell line. I found that this was the case for 8.85% of all inactive sites (Figure 2.7). For most TFs, the number of inactive motif sites was significantly larger than the number of active sites (Table 2.4). Thus, while the number of inactive sites overlapping another ChIP-seq peak was relatively small, these represented 0.41 to 32.21 times the total number of active motif sites for a TF. Footprint patterns at inactive sites that resemble active sites due to the binding of another factor highlights an important consideration and caveat when conducting motif-centric footprinting and evaluating the accuracy of footprint predictions. This also applies to de novo footprinting as it becomes an issue when annotating called footprints using motifs. A potential solution would be to exclude all motif sites overlapping ChIP-seq peaks for multiple TFs. However, this would remove 66%-100% of active sites for a TF. Additionally, this would require conducting a multitude of ChIP-seq experiments and disregards the fact that many TFs have binding partners.

Modeling data heterogeneity for footprinting

To account for the high variance in DNaseI activity at motif sites, I devised a novel supervised learning based footprint prediction framework called DeFCoM (Detecting Footprints Containing Motifs). DeFCoM trains an SVM using extracted features from DNaseI digestion profiles of motif sites labeled as active or inactive. In the training phase, DeFCoM applies a model selection procedure to choose between a linear kernel and nonlinear RBF kernel (Figure 2.8; see Materials & Methods). This allows DeFCoM to capture the complexity of the data when necessary with the RBF kernel, while avoiding over-fitting, a common problem in supervised learning, by choosing the linear kernel when that complexity is lacking. Once trained, the SVM uses features from DNaseI digestion profiles for new, unlabeled motif sites to determine which are active and inactive in another cell type/condition.

To assess DeFCoM's classification accuracy, I first performed 5-fold nested cross validation on 71 evaluation sets comprised of data from 18 transcription factors in the human cell-lines GM12878, H1-hESC, HepG2, and K562 generated by the ENCODE project. Secondly, I tested DeFCoM's ability to generalize across cell types by training models using data from one cell type and testing on an independent cell type. I also wanted to know whether using the RBF kernel increased accuracy given the demonstrated heterogeneity in these data. Therefore, for both sets of experiments, I used a linear and an RBF SVM and compared their classification performance. I will refer to these models as DeFCoM-linear and DeFCoM-RBF respectively. I calculated receiver operating characteristic (ROC) Area Under the Curve (AUC) values using all the data and also partial AUC (pAUC) values corresponding to partial ROC curves at a 5% false positive rate (FPR) cutoff.

When applied to the 71 data sets, DeFCoM-RBF performed better than a random classifier in all cases (Figure 2.9A). Notably, I observed a wide distribution of pAUC scores ranging from 0.096 to 0.981, but there was less variability in the full AUC scores (0.714-0.998). For the cross cell-line experiments, I expected that additional variability across the two data sets would decrease performance compared to the within cell-line cross validation tests. Indeed, I witnessed overall lower scores from the former but by a marginal amount (median pAUC decrease of 0.021) indicating there exist consistent footprint signals across cell types.

To determine whether using the nonlinear RBF kernel to model heterogeneity was warranted, I repeated the above experiments using the linear kernel. Overall, DeFCoM-RBF improved classification accuracy for all cell-lines in both experimental setups except for the cross cell-line case where the test set was derived from data in the K562 cell line (Figure 2.9B). I saw that the pAUC increased as much as 0.141 when using DeFCoM-RBF. However, the pAUC was essentially the same in 31% of cross validation tests and 41% of cross cell-line tests. This demonstrates that the RBF kernel can provide large gains in accuracy, but some factors or data sets may not possess enough DNaseI signal heterogeneity to benefit from more complex footprint modeling.

Interestingly, DeFCoM-linear performed substantially better on cross cell-line tests when training with GM12878 and evaluating with K562 data. This demonstrated the need for flexibility in model complexity. Therefore, I incorporated a model selection step during DeFCoM training to automatically determine the most appropriate kernel for a given test (see Materials & Methods). I found that with the exception of CTCF, my model selection procedure identified the better model in all cases in which there was a measurable difference between kernels (pAUC difference > 0.05; Figure 2.10). I also evaluated alternative methods for addressing cross cell-line applications of

DeFCoM and found the aforementioned approach produced the best results. Nevertheless, I describe the alternative procedures in the following section.

Variations for DeFCoM training in cross cell-line applications

To address the decrease in classification accuracy of DeFCoM when training in one cell-line and testing in another, I initially explored two methods in addition to the SVM model selection procedure.

Mitigating Data set Shift

Given the variety of factors involved in generating DNase-seq and ATAC-seq data as well as biological variability in the samples processed for sequencing, I considered the possibility that the DNase-seq and ATAC-seq data used for training DeFCoM may differ enough from the data being used during the classification phase of cross cell-line analyses to negatively impact classification performance. More formally, I hypothesized that the joint distribution between inputs into DeFCoM's RBF kernel SVM and the outputs produced by this SVM differed between the training and testing stage. This phenomena is more generally referred to in machine learning literature as data set shift [64].

To account for the possibility of data set shift, I trained a logistic regression model with data from GM12878 and K562 to obtain for each sample the probability that the sample was derived from GM12878, P(GM12878), and the probability that it was derived from K562, P(K562). If more than 25,000 motif sites existed in the active and inactive motif site sets for both cell-lines, I randomly selected 25,000 samples from each of the active and inactive motif site sets, totaling to 100,000 sites. These samples were converted into feature vectors, and assigned the class label "GM12878" or "K562". The labeled feature vectors were then used to train an L2-regularized logistic regression model. The regression model was then applied to feature vector representations of all the samples in both cell-lines to obtain P(GM12878) and P(K562) for each sample. The

GM12878 motif sites were then filtered to include only those for which P(K562) ≥ 0.4. These filtered motif sites were then used to train an RBF kernel SVM using 5-fold cross validation. Sample weights were included for the SVM training such that training samples more similar to the K562 test samples would receive a greater weight. I defined the weight to be P(K562)/P(GM12878). Table 2.5 provides the results of applying data set shift correction to DeFCoM for 17 transcription factors.

Sequencing Depth Matching

Another consideration related to cross cell-line analyses is the difference in sequencing depth between the training and testing set affecting DeFCoM performance. When the training data set comes from DNase-seq/ATAC-seq data with a lower sequencing depth than the test data, the dynamic range of DNaseI digestion frequencies at motif sites has the potential to be greater in the test set. Arguably, this could create another scenario where data set shift is a concern. Although I incorporate a square root transformation of the DNaseI digestion frequencies into the DeFCoM framework to mitigate dynamic range issues, I also tested if matching the sequencing depths between the training and testing data would improve DeFCoM's classification accuracy.

Using the subsampling feature in SAMTools (Li et al., 2009), I down-sampled the K562 DNase-seq data to match the GM12878 DNase-seq data sequencing depth. I then used the GM12878 and K562 data to generate the training and test set feature vectors respectively. With the GM12878 feature vectors I used 5-fold cross validation to train the RBF kernel SVM of DeFCoM, and I applied the trained model to the feature vector representations of the down-sampled K562 samples. Table 2.5 provides the results of this evaluation for 17 transcription factors. Compared to the model selection procedure, both the data set shift correction and down-sampling approaches produced worse classification performance.

Multiple variables impact motif-centric footprinting

In addition to addressing the heterogeneity of DNaseI signal at motif sites, my analyses provide insights into some variables that may affect motif-centered footprinting performance, though this is certainly not an exhaustive list of contributing factors. My observations suggest that the "footprintability" i.e., the quality of footprinting, of any particular data set is a function of several characteristics. I noted that features of the data from a particular cell-line and the specific TF being considered can contribute to footprintability. For instance, the pAUC is 0.36 higher on average in K562 compared to HepG2 for all cross validation experiments (Figure 2.9), suggesting that footprint signals in K562 are better overall. Within GM12878, the cross validation pAUC scores across TFs range from 0.210 to 0.915, highlighting the variability in footprintability across TFs. Lastly, pAUCs for CHD2 are higher than CEBPB in all cell types (Figure 2.11), suggesting active footprints for some factors are in general easier to discriminate than for others.

It is important to note that the four cell lines I use span a wide range of sequencing depths (Table 2.6). I wondered how closely footprintability was associated with total sequencing depth. Since the signal quality across data sets can widely vary, I also wondered whether the "effective" sequencing depth, based on the number of reads in DNaseI hypersensitive sites, was more important than simply the raw sequencing depth. I used mean pAUC values from DeFCoM's nested cross validation experiments for each TF across all cell lines to compare footprintability based on total and effective sequencing-depth. Overall, I found that for most factors, accuracy increased nonlinearly with respect to total sequencing depth, but not effective sequencing depth (Figure 2.12).

To better understand the trade-off between sequencing depth and signal quality, I focused on data from GM12878 and H1-hESC since they possess very different signal-to-noise ratios (0.19 versus 0.43 FRiP score). I performed 5-fold nested cross validation using DeFCoM and data from each cell line subsampled to 25, 50, 75, and 100 million aligned reads and calculated pAUCs for each. The effect of raw sequencing depth versus signal quality became more apparent when I assessed changes in pAUC at a fixed 5% FPR under this framework (Figure 2.13). As expected, the changes in pAUC vary by TF, but performance in the H1-hESC cell-line was less affected by increased sequencing depth. This suggests that for data with better signal-to-noise, informative DNaseI signals are present at lower sequencing depths, resulting in smaller improvements in footprintability with increased sequencing depth. I see the opposite in the GM12878 cell-line where increased sequencing depth substantially improves accuracy. When looking across sequencing depths at the number of H1-hESC active motif sites that are in the evaluation sets, I notice that more active sites meet the coverage filtering thresholds as sequencing depth increases. This shows that although much of the DNaseI signals may be present at lower sequencing depths, a higher sequencing depth can provide gains in sensitivity. The improvements in sensitivity will vary by TF, as evidenced by large increases for CTCF and RAD21 but significantly smaller increases for other TFs (Figure 2.14).

Interestingly, active footprints for some TFs were more accurately identified in GM12878 than H1-hESC at equivalent sequencing depths despite the reduced signal-to-noise. This may be due to the FRiP score serving as a global signal quality measure rather than at the level of individual TFs. To investigate this further, I analyzed the ratio of active motif sites to inactive sites for several TFs and found that many decreased drastically in GM12878 data with increasing sequencing depth compared to the same ratios in H1-hESC data (Figure 2.15A). For instance, in GM12878 for SP1 this ratio was 16.8 at a sequencing depth of 25 million reads but decreased to 0.55 at 100 million reads. In H1-hESC, I observed a much smaller ratio change from 0.48 to 0.10 for the same factor (Figure 2.15B). The large changes in active to inactive site ratios in GM12878 suggest that in data with lower signal-to-noise, the number of inactive sites is more affected by sequencing depth, at least based on my criteria. Across all 18 TFs in GM12878, I witnessed a -0.71 Pearson correlation on average between the active to inactive site ratios and pAUCs for a TF. In H1-hESC the mean correlation was -0.89. Overall, my results suggest that increasing sequencing depth to improve accuracy will primarily benefit noisy data sets, and that signal quality in data will affect accuracy by varying the number of inactive motif sites that are considered compared to the number of active motif sites.

DeFCoM outperforms other footprinters

To provide a comprehensive study of footprinting from a motif-centric perspective, I compared DeFCoM with nine competing footprinters: BinDNase, CENTIPEDE, cut density, DNase2TF, HINT, FOS, msCentipede, PIQ, and Wellington (Table 2.1). All methods were assessed based on their ability to correctly classify the same sets of motif sites for 18 TFs as active or inactive in the given cell-line. Partial AUCs (5% FPR) were calculated to compare the methods. For the supervised learning footprinters (DeFCoM and BinDNase), training was performed using data from K562 for test sets in GM12878, H1-hESC, and Hepg2, and in GM12878 for test sets in K562. To summarize performance across all data sets, I ranked each method by pAUC for each of the 71 tests and calculated their mean rank across all tests (Figure 2.16). DeFCoM ranked first in 25 of the 71 evaluation sets (34.7%) and second in an additional 29 test sets (40.3%). I see even better performance by DeFCoM when using pAUCs from within cell-line cross validation for the two supervised methods. DeFCoM ranked first 39 times (54.9%) and second 23 times (32.4%) (Figure 2.17). DeFCoM had the best mean rank for results from both the cross cell-line and cross validation tests followed by BinDNase and msCentipede. Interestingly, cut density, which simply predicts footprints based on the number of DNase-seq reads, had the 4th best mean rank despite not using any information about actual footprint signals (Figures 2.16B and 2.18). Previous studies witnessed similarly reasonable performance for this simple method [63,65], but Gusmao et al. showed that cut density's accuracy relative to other footprinters suffers at a 1% FPR [66]. In my study, cut density had the 5th best mean rank using pAUCs at a 1% FPR (Figure 2.19), still outperforming 5 other footprinters.

The improved classification accuracy of both DeFCoM and BinDNase over the unsupervised approaches highlights the utility of learning a discriminative model for motif-centric footprinting. Because DeFCoM defaults to a linear SVM model unless more complex modeling is required, I expect it to perform at least as well as BinDNase, which uses another type of linear model, logistic regression. Also, including the nonlinear RBF kernel enables DeFCoM to outperform BinDNase by as much as 0.0835 pAUC, though I note that the two footprinters have essentially the same accuracy for 59 of the 71 data sets (pAUC difference < 0.025). This increases to 65 of the 71 data sets using pAUC difference < 0.05 (Figure 2.20). BinDNase includes a computationally expensive greedy backward search to determine optimal features. Impressively, this shows that DeFCoM can achieve a similar or better accuracy than BinDNase using a set of predefined features that can be computed more efficiently. The greater overall performance of msCentipede relative to the other unsupervised footprinters indicates that modeling heterogeneity with an unsupervised method can produce comparable results to DeFCoM in some cases, though I note that for the factor TBP in HepG2, a model could not be learned in reasonable time (model training terminated after 60 days). For 48 of the 71 test sets, DeFCoM and msCentipede perform similarly (pAUC difference < 0.05), but using supervised learning affords DeFCoM better performance in 16 of the data sets (pAUC > 0.05), including a pAUC difference of 0.25 for the RAD21 test sets.

ATAC-seq is comparable to DNase-seq for footprinting

Like DNase-seq, ATAC-seq assays for accessible chromatin and can generate visible footprints in aggregate accessibility profiles for active motif sites. Its low biological sample material requirement relative to DNase-seq makes it an appealing alternative when this is a limiting factor. I evaluated DeFCoM using GM12878 ATAC-seq data to determine its utility for motif-centric supervised footprinting. I applied 5-fold nested cross validation with the ATAC-seq data to train and test DeFCoM models for 18 TFs. The pAUC at 5% FPR and full AUC were averaged across the 5 folds from the cross-validation. I then repeated the nested cross validation with DNase-seq data on the same set of active and inactive sites (Figure 2.21). Despite the differences in sequencing depth of the DNase-seq (245 million reads) and ATAC-seq data (93 million reads), the pAUC and full AUC values are generally similar, with DeFCoM performing slightly better when using DNase-seq (mean pAUC difference = 0.072, mean AUC difference = 0.043). Overall this supports the feasibility of extending DeFCoM to experiments that use ATAC-seq.

DeFCoM as an open-source software package

Poor implementation and usability hinder the adoption of otherwise practical tools in the scientific community. With this in mind, I implemented DeFCoM to be an easy-to-use software package with a code-base that follows good software design principles. For both end-users and developers, I make my code freely accessible via a code repository (https://bitbucket.org/bryancquach/defcom) with extensive API documentation and a user guide. DeFCoM is the only supervised learning footprinter supported by thorough documentation to improve ease of use. I also include well-commented scripts to handle common data processing tasks for footprint analysis. DeFCoM is implemented in the Python programming language within an object-oriented framework that enhances modularity of the code for easy debugging, modification, and extension. Furthermore, because DeFCoM is a data-intensive method, I make use of scalable programming techniques such as batch processing and parallel computing to ensure feasibility for use on a modern desktop machine. As an open-source software package, I encourage the community to modify and adapt my code for further advancements in footprinting research

DISCUSSION

Our study provides novel insights into variables that affect identification of DNaseI footprints, and for the assessment of footprinter performance. Aggregate DNaseI digestion profiles do not represent well the footprint patterns seen at individual sites, thus footprinters that use models based on aggregate or general footprint signal patterns may suffer. Inactive motif sites for one TF may be bound by a TF that creates a footprint and thus be misclassified, at least for the original TF. This is a general challenge in the assessment of motif-centric approaches, but this does not necessarily reflect a weakness in these footprinters. The motif-centric footprinter is correctly identifying a footprint, though it mistakenly attributes it to the wrong factor. Arguably, this is better than spuriously identifying a footprint at a location where no factor is bound. This serves as an important consideration for both interpreting footprint predictions and assessing footprinters in a motif-based framework.

Heterogeneity in DNaseI digestion signals at motif sites exists, and I show that my DeFCoM footprinter benefits from being aware of this heterogeneity. At the same time, I also show that incorporating the flexibility to use more or less complicated models depending on the particular TF, cell line, and data set is important as well. DNase-seq and ATAC-seq footprint signals will vary based on biological and technical factors that influence the data. Footprinters that can model footprints well across this range of variability will obviously be more robust. Supporting this, msCentipede also models heterogeneity and was the best performing method that did not use supervised learning, though I found this method may be limited by unreasonable training times for specific data sets.

We show that determining appropriate sequencing depth for footprinting is not easy and is affected by many variables. I observed sequencing depth affected footprinter accuracy less when the DNase-seq data had a better signal-to-noise ratio, but I also witnessed variation in TF-specific footprintability at equivalent sequencing depths between cell-lines. Sung *et al.* provided evidence that DNA residence time plays a role in the clarity of a footprint signal [53]. Likewise, greater sequencing depth generally increased the number of sites where footprints were identified, but the benefit to individual factors varies. Biological variables such as these need to be further assessed on a per-TF basis in conjunction with technical factors to better realize which of these most strongly contribute to footprintability. This knowledge would help determine how to appropriately design footprinting experiments.

For footprinters such as DeFCoM that use supervised learning, the concordance between features of the training and test sets become important. Although this introduces added complexity, it can be leveraged to achieve more targeted results. For instance, high-confidence footprints in DNaseI hypersensitive sites could be identified by tailoring the training set to include only sites in areas of high DNaseI activity. Doing so would make the model more representative of these stronger footprint signals, though at the expense of generalizability to low signal regions. Potential variability between training and test sets should be minimal for situations in which data is generated from the same cell type for both but possibly under different experimental conditions.

A comprehensive evaluation of footprinting was reported in [66]. Though more rigorous than previous comparative analyses, their evaluation strategy was more informative for understanding footprinters in a *de novo* footprinting context. I provide a complementary footprinter evaluation from a motif-centric perspective. In my work, I focused on results at a 5% FPR to provide more practical insight on footprint detection accuracy at acceptable error rates. The ability of both DeFCoM and BinDNase to consistently outperform unsupervised footprinters, with the possible exception of msCentipede, further supports supervised learning-based methods. I note that my results contradict accuracy levels found in the previous evaluation for several footprinters. This demonstrates that evaluation methods can largely influence reported performance. The *de novo* footprinters DNase2TF and FOS performed poorly in my tests, because they failed to report a score for many of the motif sites in the test set. My results in conjunction with previous studies highlight the importance of evaluating a footprinter in the context for which it was designed.

ATAC-seq is quickly being adopted as it requires less biological starting material, and I show DeFCoM performs comparably with these data. As I learn more about the nuances of footprinting in both DNase- and ATAC-seq, I expect footprinters will adapt accordingly. In light of this, my implementation of DeFCoM in an open-source, modularized and object-oriented framework makes it conducive to modification and improvement. As such, I welcome and encourage collaborative efforts with others in the scientific community to address the needs of researchers as the field evolves.

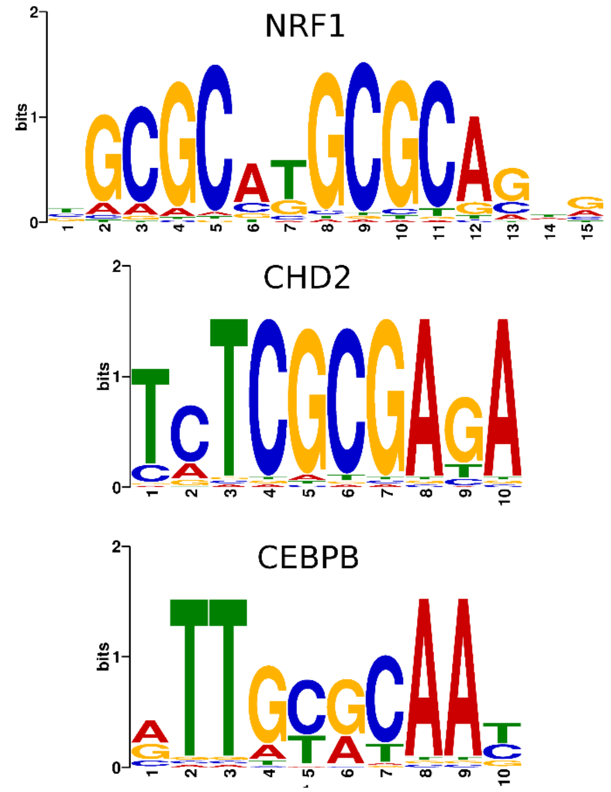

Figure 2.1. Motif logos for NRF1, CHD2, and CEBPB. Sequence logo representations of position weight matrices used to evaluate DNaseI signal profile heterogeneity.

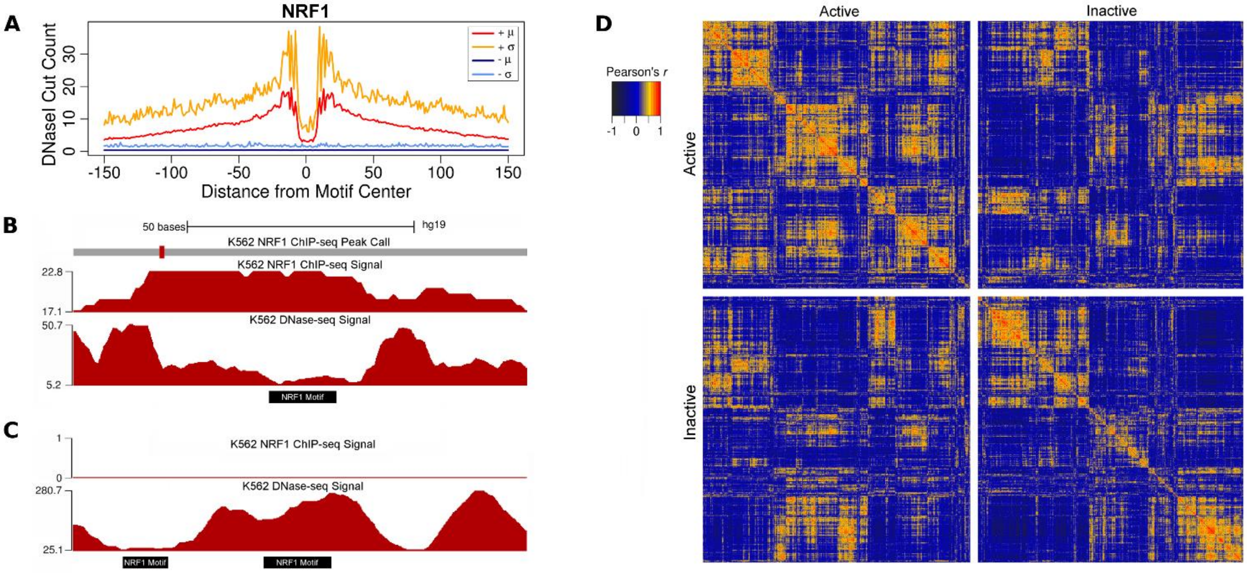

Figure 2.2 Within and between class variability in DNaseI digestion signal at motif sites. A) Per base means (μ) and standard deviations (σ) of DNaseI signal aggregated for NRF1 motif sites active (+) and inactive (-) in K562. B) K562 DNase-seq and ChIP-seq signal at an NRF1 motif site (Chr1:16,175,923-16,176,022) from the active class and C) two neighboring NRF1 inactive sites (Chr22:38,966,291-38,966,390). D) Pairwise Pearson correlations between the top 2000 NRF1 motif sites from the active and inactive class ranked by DNaseI digestion signal.

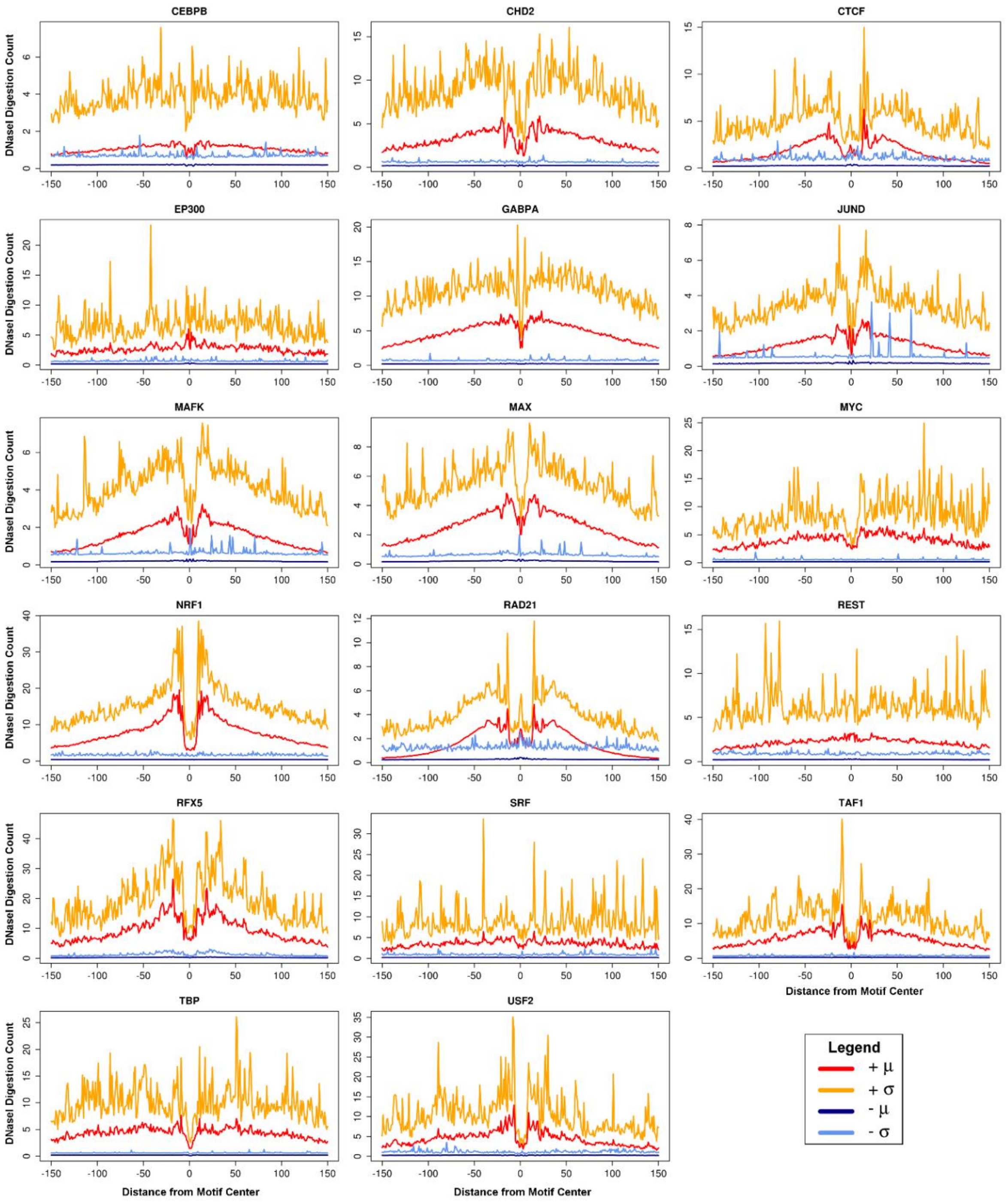

Figure 2.3. K562 DNaseI signal profiles. K562 aggregate mean (μ) and standard deviation (σ) DNaseI digestion profiles for the active (+) and inactive (-) motif site classes of 17 transcription factors.

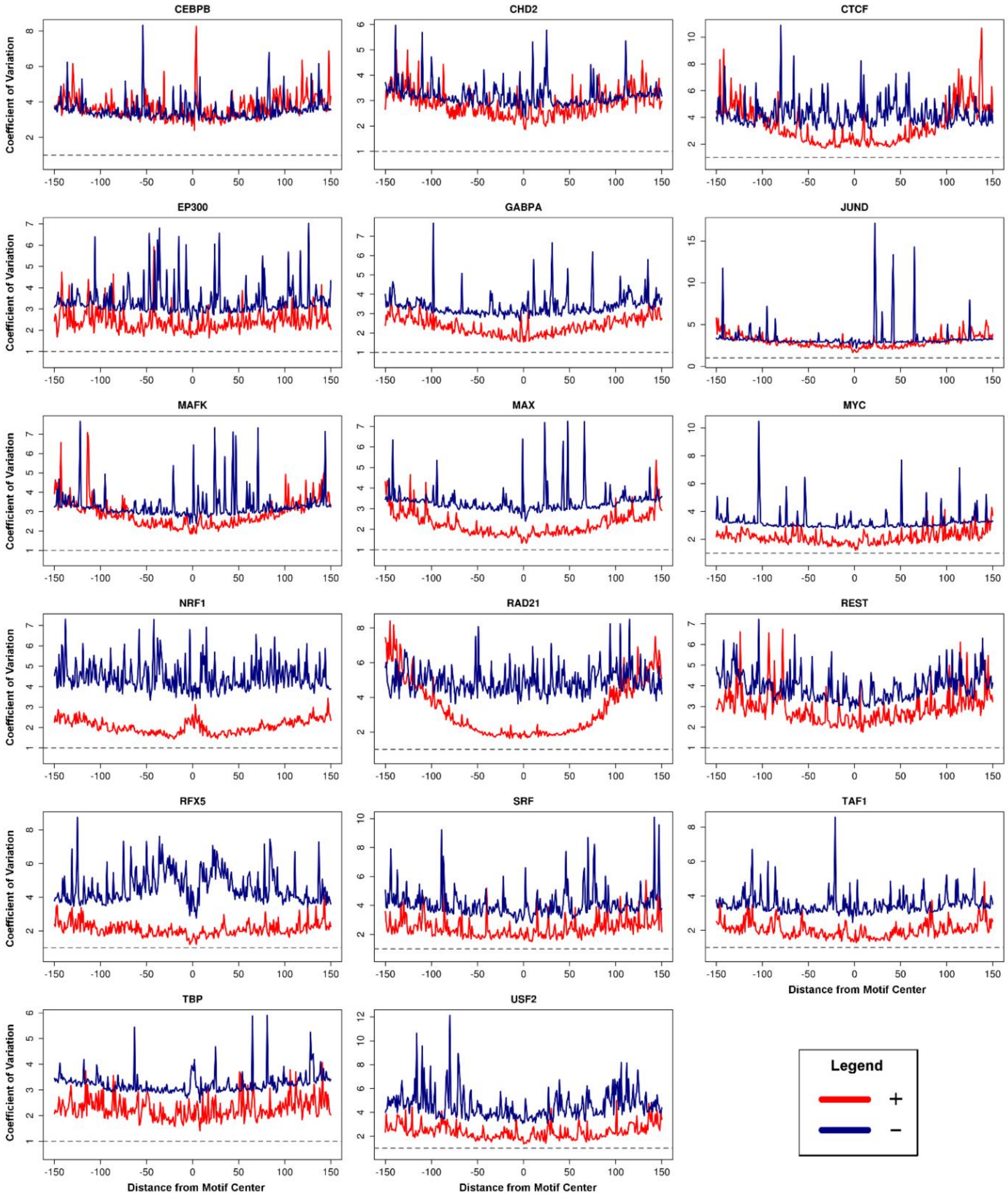

Figure 2.4. Coefficients of variation for K562 DNaseI digestion profiles. Coefficients of variation derived from K562 DNaseI digestion profiles for the active (+) and inactive (-) motif site classes of 17 transcription factors. The dashed horizontal gray line denotes a coefficient of variation of 1. Values above this signify that the standard deviation exceeds the mean

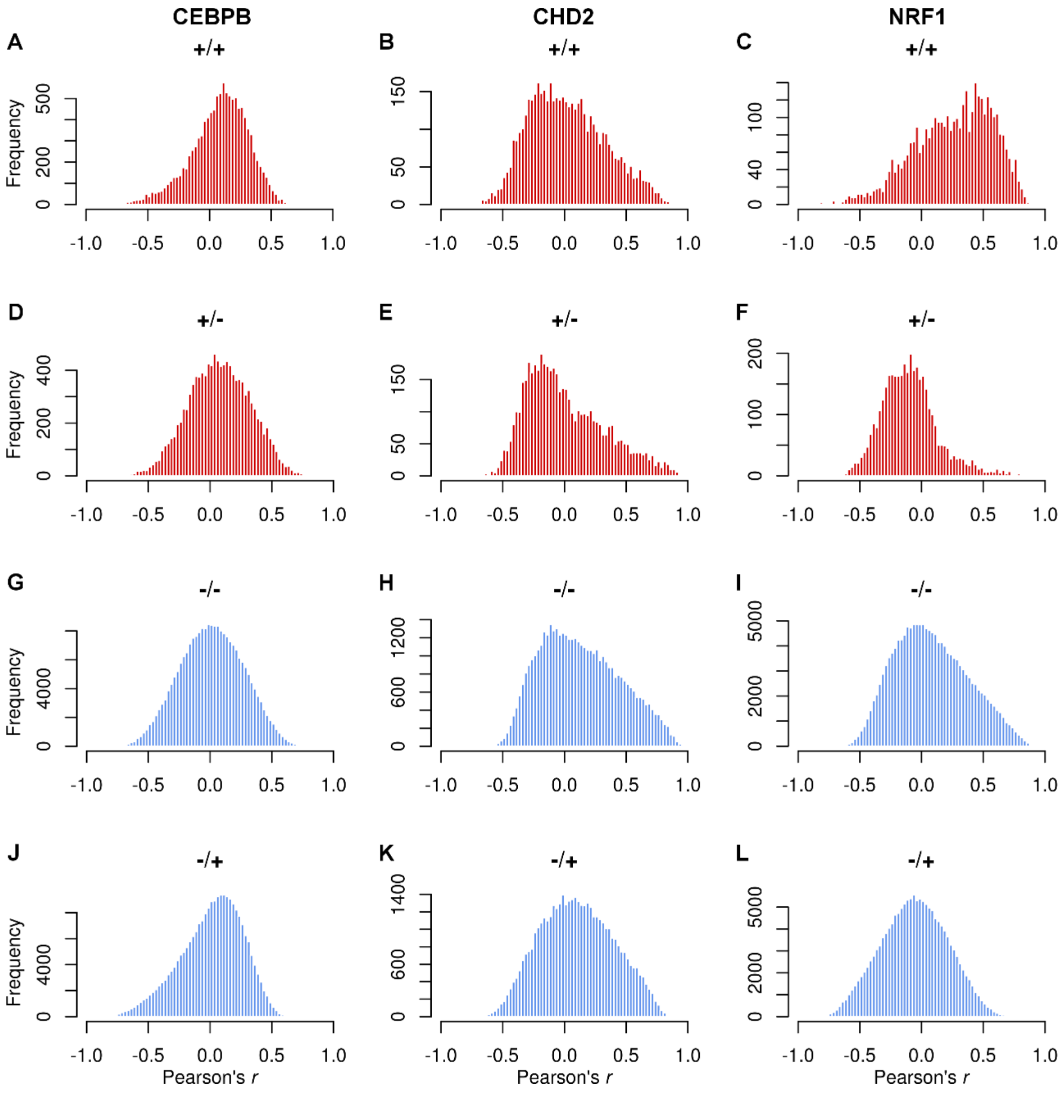

Figure 2.5. Correlations between aggregate and individual DNaseI digestion profiles.

Histograms conveying the spread in Pearson correlation coefficients i) between K562 DNaseI signal at active motif sites and the active class aggregate mean DNaseI cut profile for A) CEBPB, B) CHD2, and C) NRF1. ii) between K562 DNaseI signal at active motif sites and the inactive class aggregate mean DNaseI cut profile for D) CEBPB, E) CHD2, and F) NRF1. iii) between K562 DNaseI signal at inactive motif sites and the inactive class aggregate mean DNaseI cut profile for G) CEBPB, H) CHD2, and I) NRF1. iv) between K562 DNaseI signal at inactive motif sites and the active class aggregate mean DNaseI cut profile for J) CEBPB, K) CHD2, and L) NRF1.

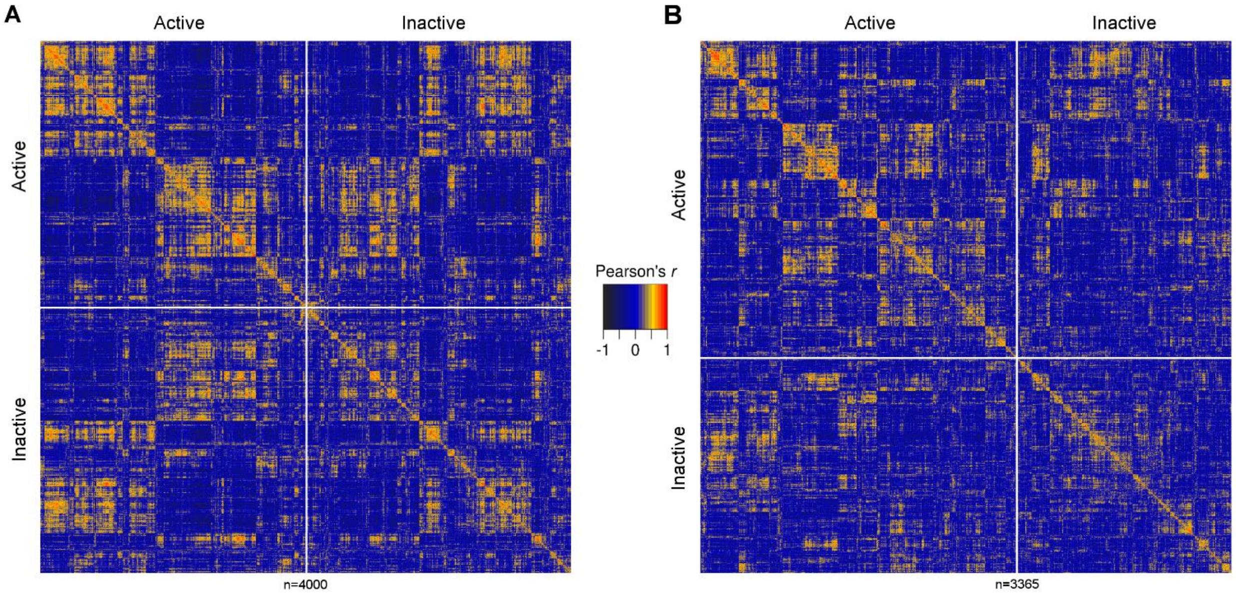

Figure 2.6. Pairwise correlation heatmaps for individual DNaseI digestion profiles.

Correlations between individual DNaseI digestion profiles at A) CEBPB and B) CHD2 motif sites in K562. After filtering active and inactive NRF1 motif site sets for the top 2000 sites ranked by total DNaseI digestion events (for CHD2 the inactive set had less than 2000 sites), we applied hierarchical clustering to the DNaseI signal profiles based on pairwise Pearson correlations and visualized the correlation values as heatmaps.

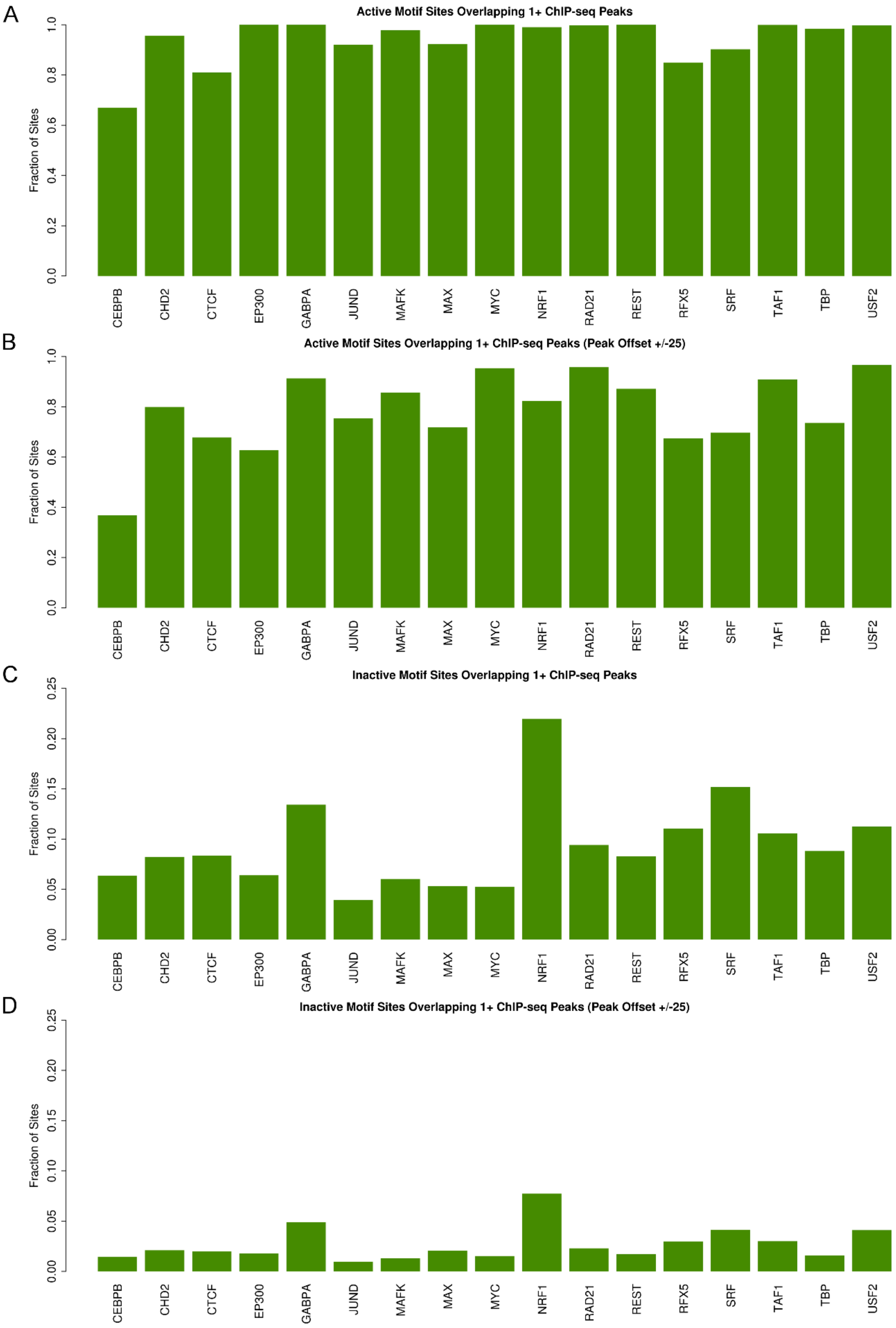
A
Active Motif Sites Overlapping 1+ ChIP-seq Peaks
Fraction of Sites
0.0
0.2
0.4
0.6
0.8
1.0
CEBPB
CHD2
CTCF
EP300
GABPA
JUND
MAFK
MAX
MYC
NRF1
RAD21
REST
RFX5
SRF
TAF1
TBP
USF2
B
Active Motif Sites Overlapping 1+ ChIP-seq Peaks (Peak Offset +/-25)
Fraction of Sites
0.0
0.2
0.4
0.6
0.8
1.0
CEBPB
CHD2
CTCF
EP300
GABPA
JUND
MAFK
MAX
MYC
NRF1
RAD21
REST
RFX5
SRF
TAF1
TBP
USF2
C
Inactive Motif Sites Overlapping 1+ ChIP-seq Peaks
Fraction of Sites
0.00
0.05
0.10
0.15
0.20
0.25
CEBPB
CHD2
CTCF
EP300
GABPA
JUND
MAFK
MAX
MYC
NRF1
RAD21
REST
RFX5
SRF
TAF1
TBP
USF2
D
Inactive Motif Sites Overlapping 1+ ChIP-seq Peaks (Peak Offset +/-25)
Fraction of Sites
0.00
0.05
0.10
0.15
0.20
0.25
CEBPB
CHD2
CTCF
EP300
GABPA
JUND
MAFK
MAX
MYC
NRF1
RAD21
REST
RFX5
SRF
TAF1
TBP
USF2

Figure 2.7. Motif site overlap with ChIP-seq peaks. A) The fraction of K562 active motif sites for a transcription factor that overlap a ChIP-seq peak for another factor (100 transcription factors were included). B) The same analysis as in (A) but using a 50 bp window centered on the ChIP-seq peak offset instead of the full peak region. C) The fraction of K562 inactive motif sites for a transcription factor that overlap a ChIP-seq peak for another factor. D) The same analysis as in (C) but using a 50 bp window centered on the ChIP-seq peak offset instead of the full peak region.

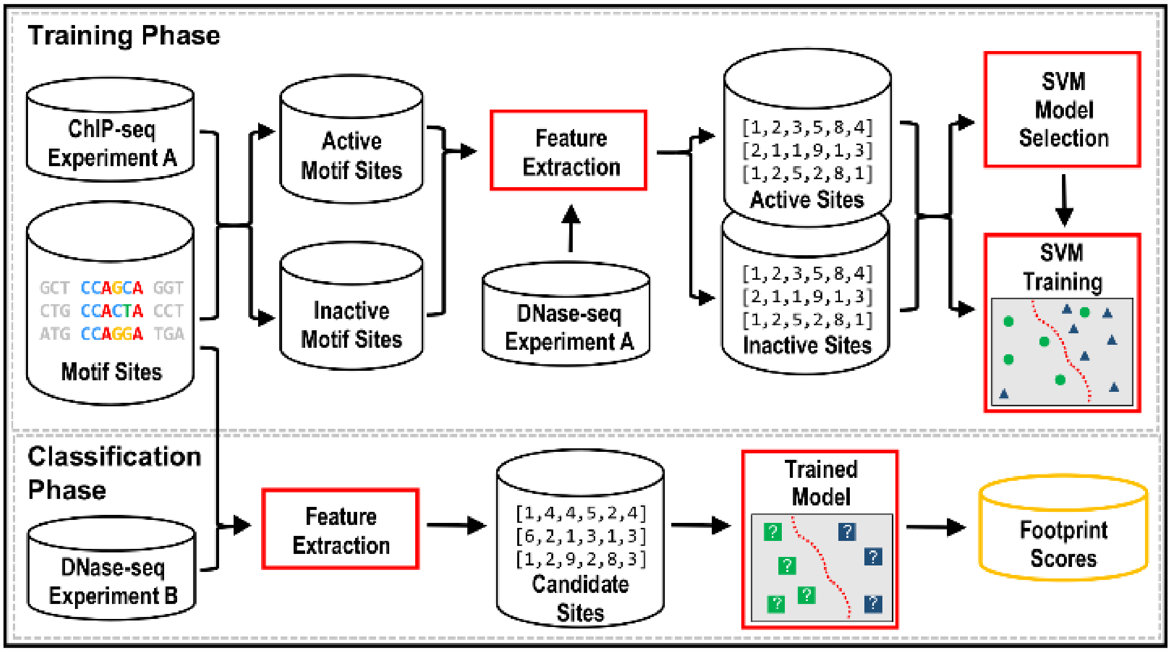

Figure 2.8. Overview of the DeFCoM classification framework. In the training phase, active and inactive motif site sets are constructed using ChIP-seq data. Corresponding DNase-seq data is used to produce DNaseI digestion profiles for each motif site. These profiles are converted into feature vectors that go into model selection and SVM training. The trained SVM model can then be used to classify motif sites as active or inactive in a different experiment or condition for which DNase-seq data are available, without the need ChIP-seq data.

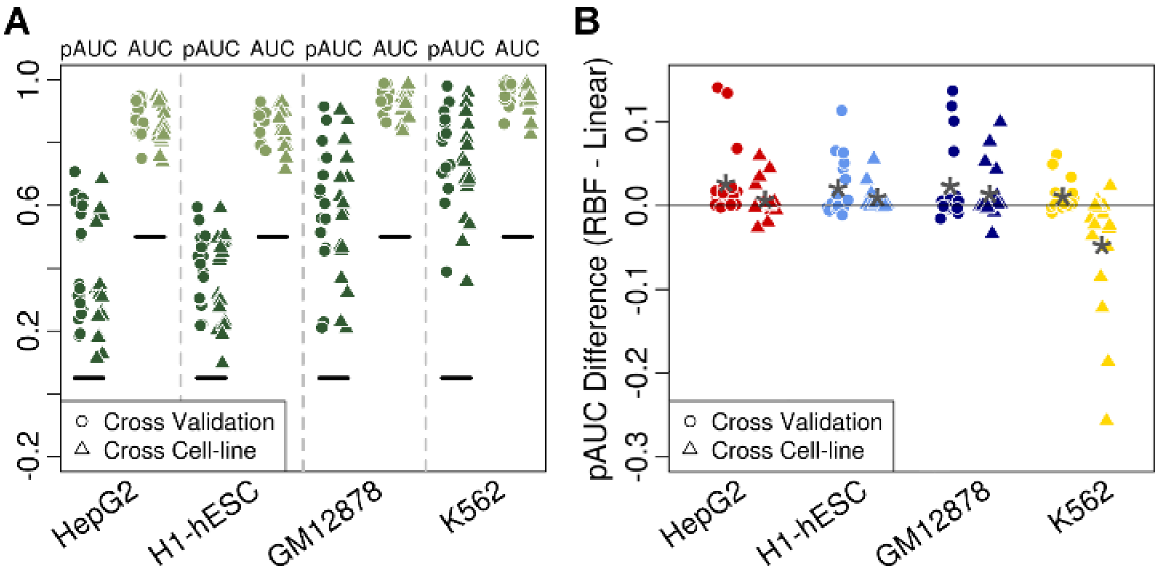

Figure 2.9. Comparison of DeFCoM model variants. A) Partial (5% FPR) and full AUCs from evaluations of DeFCoM-RBF for 18 TFs in 4 cell-lines. Black horizontal lines signify values if classifications were random. B) Comparison of DeFCoM to DeFCoM-linear by differences in pAUCs for the same test sets as A.

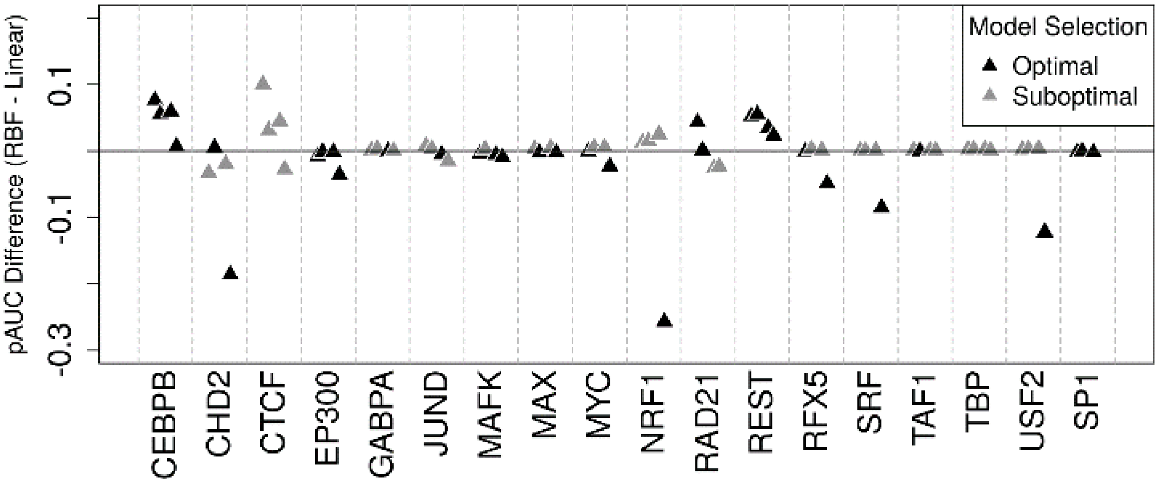

Figure 2.10. DeFCoM training phase model selection performance. Assessment of when the model selection procedure chooses the better SVM type (linear vs. RBF kernel) during the training phase of cross cell-line tests for 18 TFs. Optimal and suboptimal denote whether the model selection procedure chose the SVM type that produced the higher pAUC (FPR 5%) values.

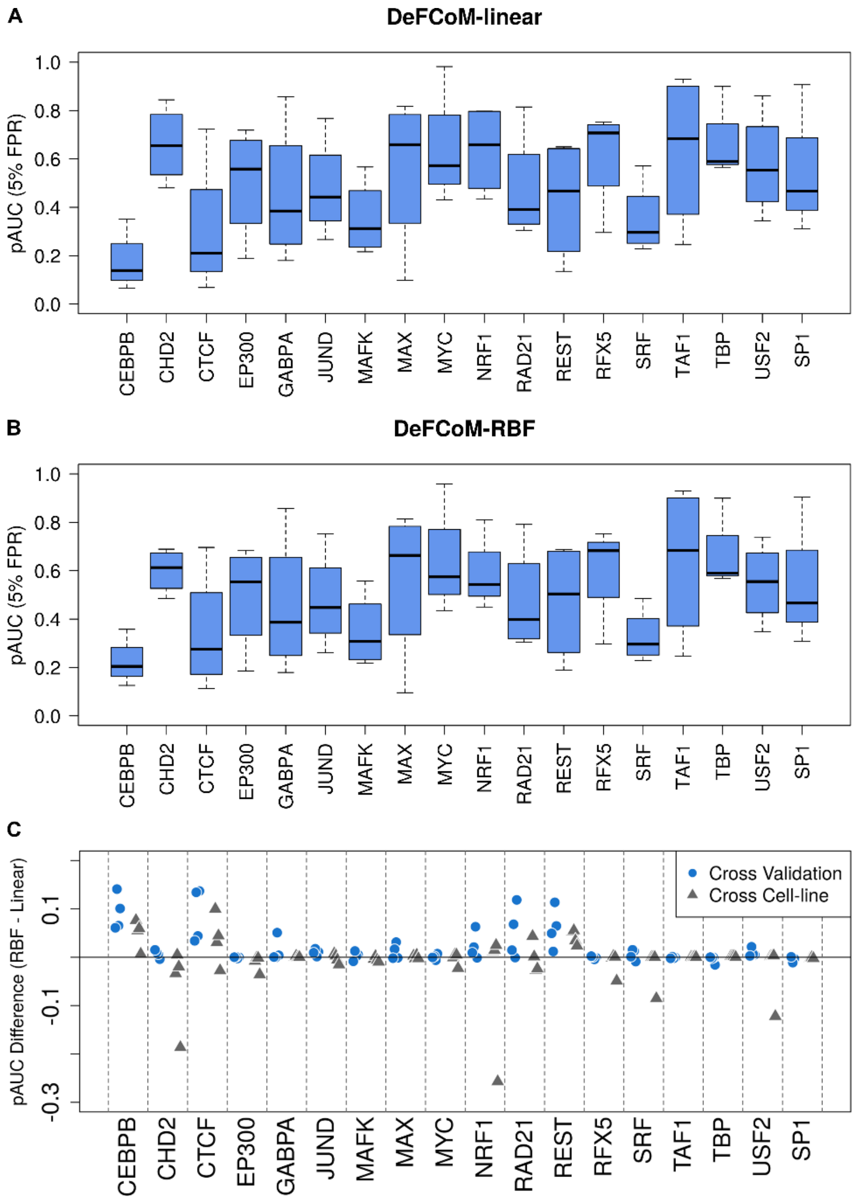

Figure 2.11. Classification performance of DeFCoM-linear vs. DeFCoM-RBF. Distribution of pAUC (5% FPR) scores by transcription factor from cross cell-line tests for A) DeFCoM-linear and B) DeFCoM-RBF. C) Comparison of DeFCoM to DeFCoM-linear by pAUC difference (5% FPR) for 18 TFs in cross validation and cross cell-line evaluations.

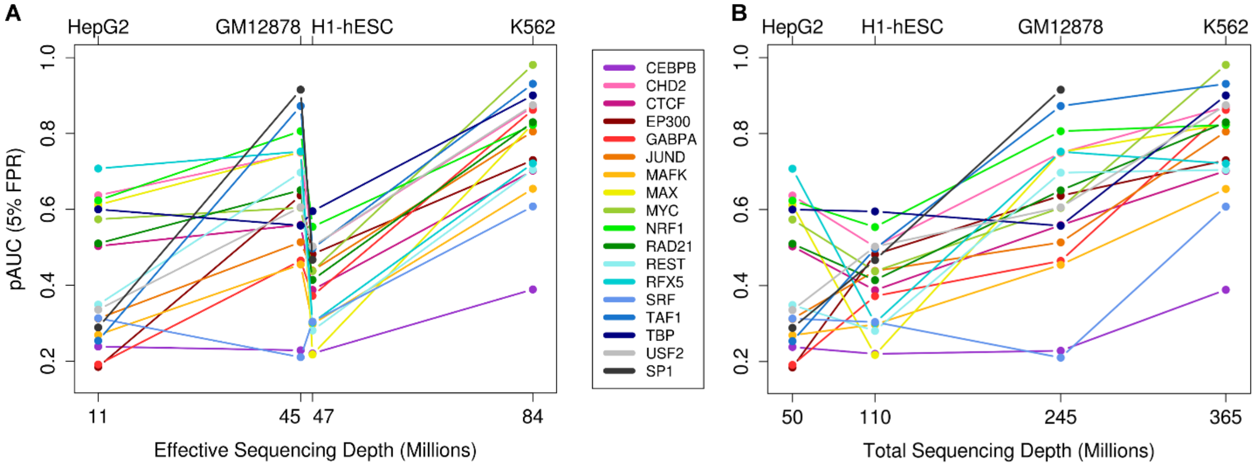

Figure 2.12. Comparisons of DeFCoM classification performance with effective and total sequencing depth. Comparison of DeFCoM classification performance at A) effective sequencing depths for the four cell-lines used and B) total sequencing depths for the same cell-lines. Notably, pAUC values vary widely across TFs and poorly correlate with sequencing depth for most transcription factors.

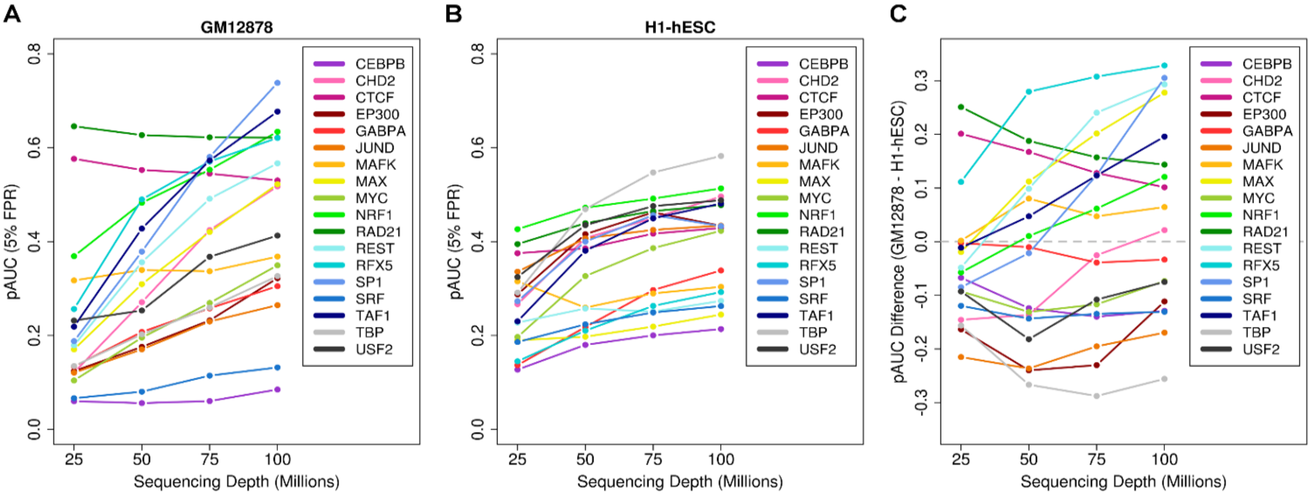

Figure 2.13. DeFCoM classification performance on subsampled data. Comparison of DeFCoM classification performance with A) subsampled GM12878 and B) H1-hESC DNase-seq data. H1-hESC possesses a higher signal-to-noise ratio and is affected less by increased sequencing depth. C) The pAUC difference between GM12878 and H1-hESC for each TF at the four subsampled sequencing depths

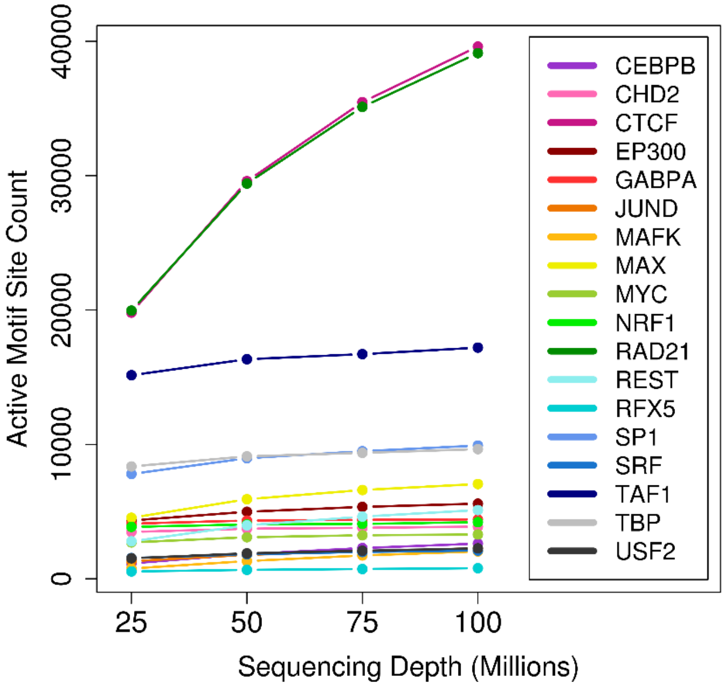

Figure 2.14. Active motif site set size at various sequencing depths. Number of motif sites that meet filtering criteria in the active set across sequencing depths in H1-hESC.

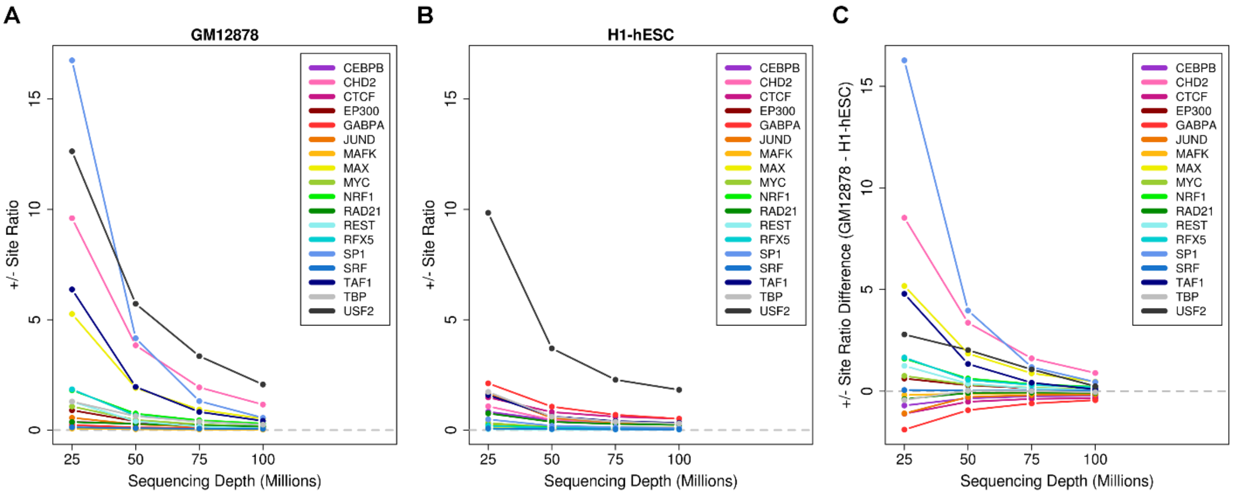

Figure 2.15. Active to inactive motif site set size ratio at various sequencing depths.

Comparison of the ratio of active (+) motif sites to inactive (-) motif sites across four subsampled sequencing depths in A) GM12878 and B) H1-hESC DNase-seq data. C) The difference in ratios between GM12878 and H1-hESC for each TF at the four subsampled sequencing depths.

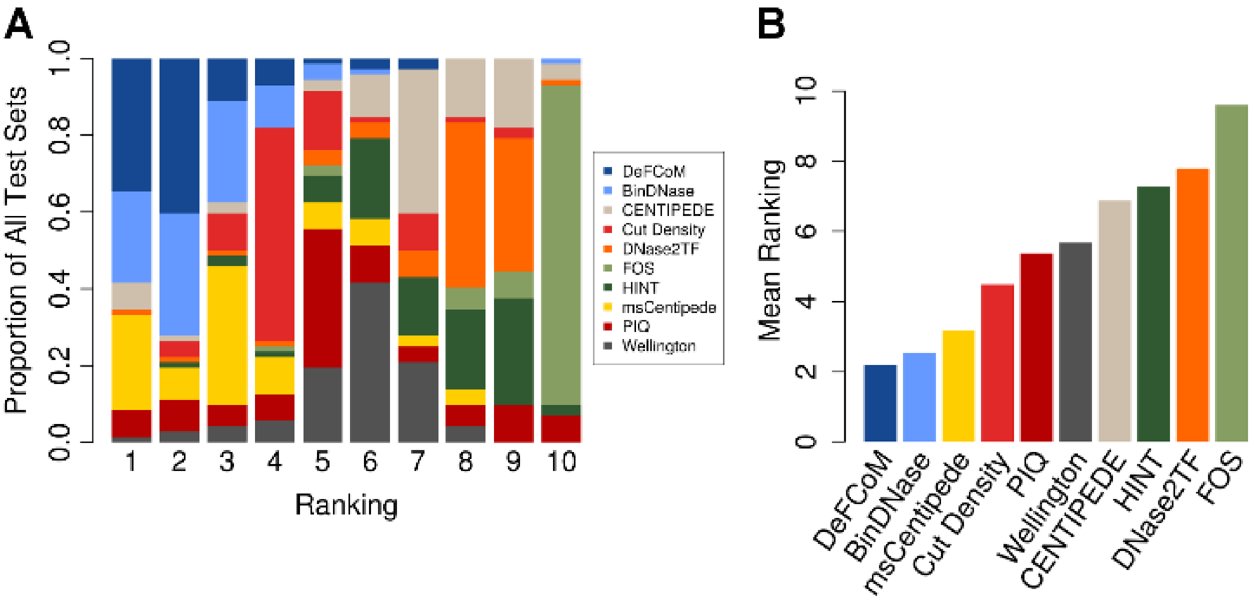

Figure 2.16. Performance ranking of footprinters. A) Frequency at which each footprinter obtains a particular rank (based on 5% FPR pAUC) for all 71 evaluation sets. B) Mean rank, derived from A, of each footprinter.

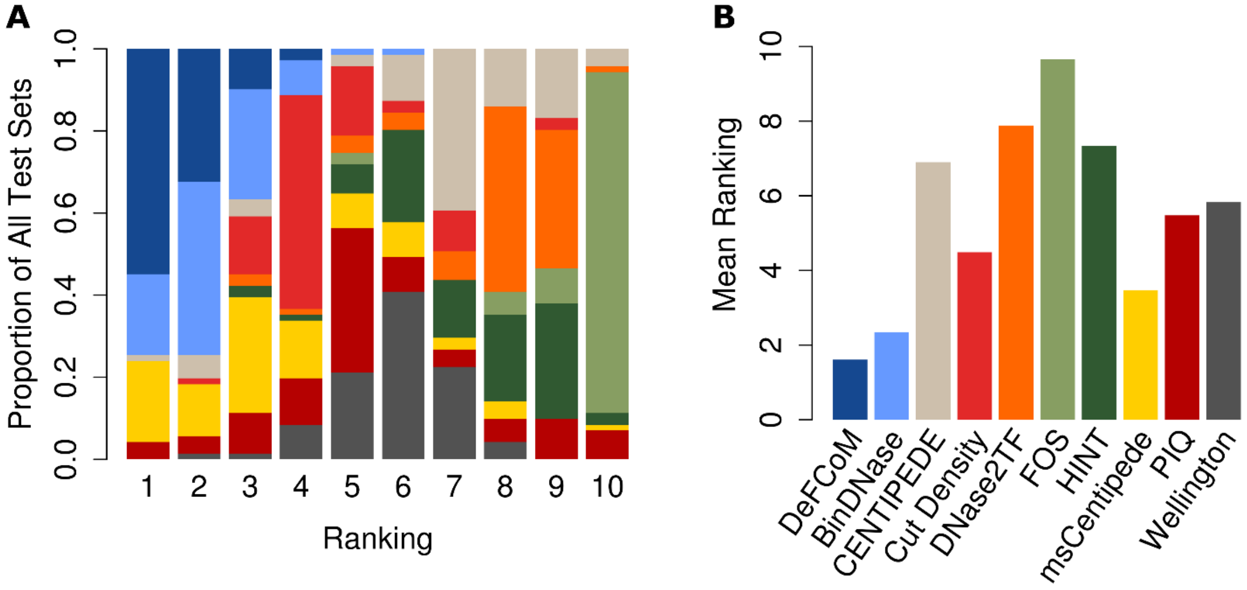

Figure 2.17. Performance ranking of footprinters using cross validation results. Comparison of footprinters when DeFCoM and BinDNase mean pAUCs from cross-validation are used. A) Frequency at which each footprinter obtains a rank (based on 5% FPR pAUC) for all 71 evaluation sets. B) Mean rank, derived from A, of each footprinter

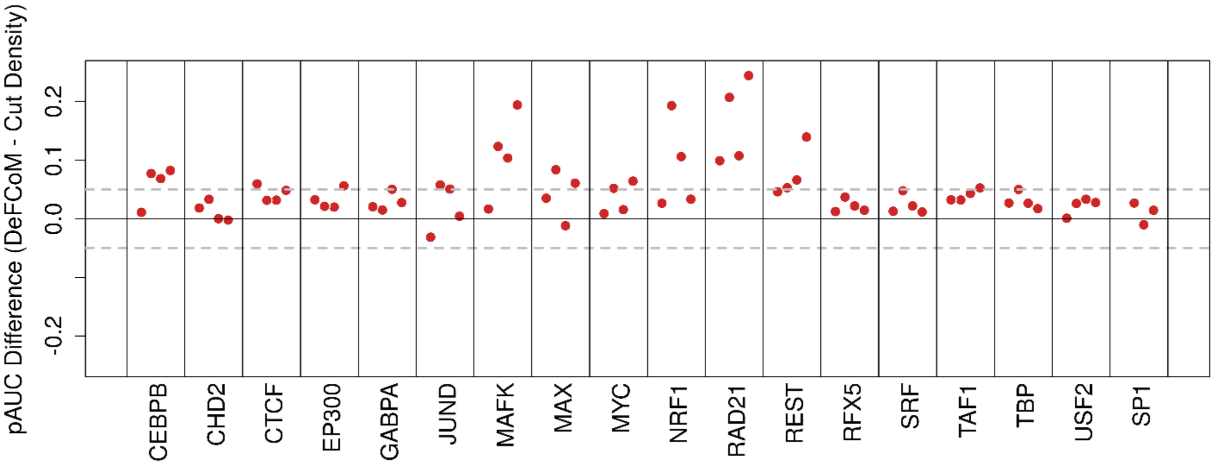

Figure 2.18. Partial AUC comparison between DeFCoM and Cut Density. Comparison between DeFCoM and Cut Density pAUCs (5% FPR) for 71 test sets from 18 transcription factors and 4 cell-lines. Gray, horizontal dashed lines are at the -0.05 and 0.05 pAUC difference.

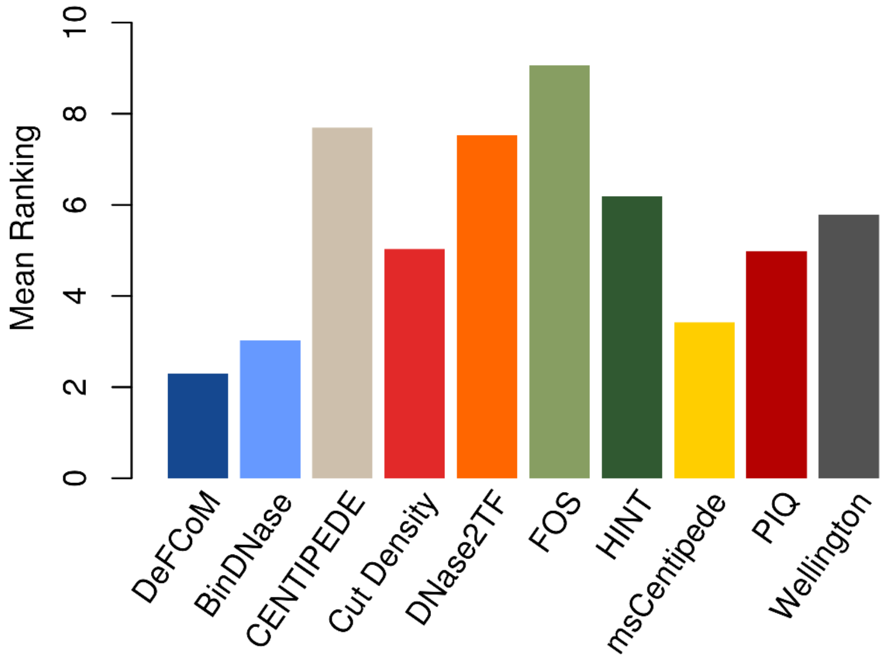

Figure 2.19. Performance ranking of footprinters at a 1% FPR cutoff for pAUCs. Comparison of footprinters by mean rank of 71 test sets. Mean ranks are based on pAUCs at a 1% FPR.

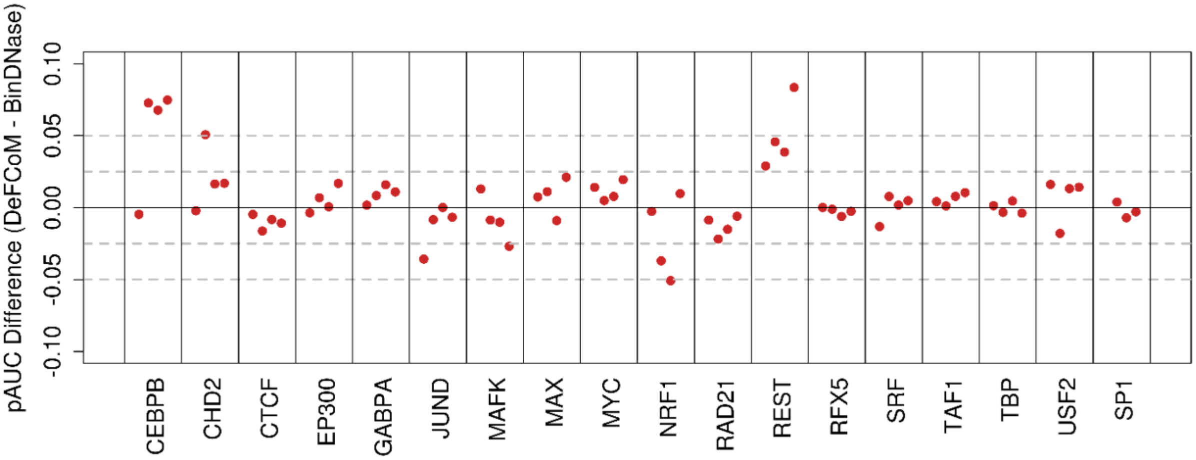

Figure 2.20. Partial AUC comparison between DeFCoM and BinDNase. Comparison between DeFCoM and BinDNase pAUCs (5% FPR) for 71 test sets from 18 transcription factors and 4 cell-lines. Gray, horizontal dashed lines are at the -0.05, -0.025, 0.025 and 0.05 pAUC differences.

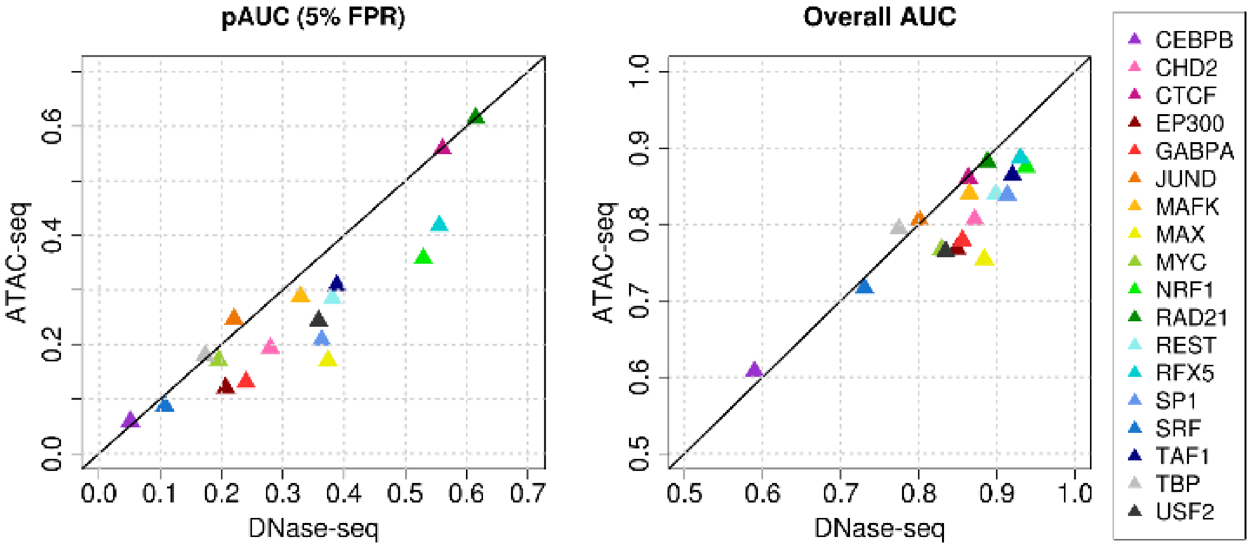

Figure 2.21. Comparison between using DNase-seq and ATAC-seq with DeFCoM. Comparison between using GM12878 ATAC-seq and DNase-seq data with DeFCoM. Partial AUC (left) and full AUC (right) results from cross-validation tests for 18 TFs

Footprinter Name	Author	Footprinter Type	Classification Algorithm	Included in Comparison
Boyle et al.	[38]	Motif-centric	Probabilistic Graphical Model (Hidden Markov Model)	No
DBFP	[45]	De novo	Probabilistic Graphical Model (Dynamic Bayesian Network)	No
DeFCoM	-	Motif-centric	Support Vector Machine*	Yes
BinDNase	[46]	Motif-centric	Logistic Regression*	Yes
CENTIPEDE	[50]	Motif-centric	Bayesian hierarchical mixture model	Yes
Cut Density	-	Motif-centric	Window-based summary statistic	Yes
DNase2TF	[53]	De novo	Window-based summary statistic	Yes
FLR	[67]	Motif-centric	Mixture model	No
FOS	[48]	De novo	Window-based summary statistic	Yes
HINT	[68]	De novo	Probabilistic Graphical Model (Hidden Markov Model)	Yes
Millipede	[47]	Motif-centric	Logistic Regression	No
msCentipede	[51]	Motif-centric	Bayesian multi-scale model	Yes
PIQ	[52]	Motif-centric	Gaussian process and expectation propagation	Yes
Wellington	[49]	De novo	Binomial test	Yes

*=Supervised learning method

Table 2.1. Summary of footprinters. Names of existing footprint detection methods and characteristics of their approach. The last column indicates if they were included in the method classification performance comparison.

Cell-line	Files
GM12878	wgEncodeOpenChromDnaseGm12878AlnRep1.bam wgEncodeOpenChromDnaseGm12878AlnRep2.bam wgEncodeOpenChromDnaseGm12878AlnRep3.bam wgEncodeOpenChromDnaseGm12878AlnRep4.bam wgEncodeOpenChromDnaseGm12878AlnRep5.bam
H1-hESC	wgEncodeOpenChromDnaseH1hescAlnRep1.bam wgEncodeOpenChromDnaseH1hescAlnRep2.bam
HepG2	wgEncodeOpenChromDnaseHepg2AlnRep1.bam wgEncodeOpenChromDnaseHepg2AlnRep2.bam wgEncodeOpenChromDnaseHepg2AlnRep3.bam
K562	wgEncodeOpenChromDnaseK562AlnRep1V2.bam wgEncodeOpenChromDnaseK562AlnRep2V2.bam wgEncodeOpenChromDnaseK562AlnRep3V2.bam

Table 2.2 ENCODE DNase-seq data files. File names for DNase-seq data used. Obtained from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromDnase/.

Transcription Factor	Files*
CEBPB	wgEncodeAwgTfbsHaibGm12878Cebpbsc150V0422111UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibHepg2Cebpbsc150V0416101UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibK562Cebpbsc150V0422111UniPk.narrowPeak.gz wgEncodeAwgTfbsSydhH1hescCebpbIggrabUniPk.narrowPeak.gz wgEncodeHaibTfbsGm12878Cebpbsc150V0422111AlnRep1.bam wgEncodeHaibTfbsGm12878Cebpbsc150V0422111AlnRep2.bam wgEncodeHaibTfbsHepg2Cebpbsc150V0416101AlnRep1.bam wgEncodeHaibTfbsHepg2Cebpbsc150V0416101AlnRep2.bam wgEncodeHaibTfbsK562Cebpbsc150V0422111AlnRep1.bam wgEncodeHaibTfbsK562Cebpbsc150V0422111AlnRep2.bam wgEncodeSydhTfbsH1hescCebpbIggrabAlnRep1.bam wgEncodeSydhTfbsH1hescCebpbIggrabAlnRep2.bam
CHD2	wgEncodeAwgTfbsSydhGm12878Chd2ab68301IggmusUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhH1hescChd2IggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhHepg2Chd2ab68301IggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhK562Chd2ab68301IggrabUniPk.narrowPeak.gz wgEncodeSydhTfbsGm12878Chd2ab68301IggmusAlnRep1.bam wgEncodeSydhTfbsGm12878Chd2ab68301IggmusAlnRep2.bam wgEncodeSydhTfbsH1hescChd2IggrabAlnRep1.bam wgEncodeSydhTfbsH1hescChd2IggrabAlnRep2.bam wgEncodeSydhTfbsHepg2Chd2ab68301IggrabAlnRep1.bam wgEncodeSydhTfbsHepg2Chd2ab68301IggrabAlnRep2.bam wgEncodeSydhTfbsK562Chd2ab68301IggrabAlnRep1.bam wgEncodeSydhTfbsK562Chd2ab68301IggrabAlnRep2.bam
CTCF	wgEncodeAwgTfbsBroadGm12878CtcfUniPk.narrowPeak.gz wgEncodeAwgTfbsBroadH1hescCtcfUniPk.narrowPeak.gz wgEncodeAwgTfbsBroadHepg2CtcfUniPk.narrowPeak.gz wgEncodeAwgTfbsBroadK562CtcfUniPk.narrowPeak.gz wgEncodeHaibTfbsH1hescCtcfsc5916V0416102AlnRep1.bam wgEncodeHaibTfbsH1hescCtcfsc5916V0416102AlnRep2.bam wgEncodeHaibTfbsHepg2Ctcfsc5916V0416101AlnRep1.bam wgEncodeHaibTfbsHepg2Ctcfsc5916V0416101AlnRep2.bam wgEncodeHaibTfbsK562CtcfcPcr1xAlnRep1V2.bam wgEncodeHaibTfbsK562CtcfcPcr1xAlnRep2V2.bam wgEncodeSydhTfbsGm12878Ctcfsc15914c20StdAlnRep1.bam wgEncodeSydhTfbsGm12878Ctcfsc15914c20StdAlnRep2.bam
EP300	wgEncodeAwgTfbsBroadK562P300UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibGm12878P300Pcr1xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibH1hescP300V0416102UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibHepg2P300V0416101UniPk.narrowPeak.gz wgEncodeHaibTfbsGm12878P300Pcr1xAlnRep1.bam wgEncodeHaibTfbsGm12878P300Pcr1xAlnRep2.bam wgEncodeHaibTfbsH1hescP300V0416102AlnRep1.bam wgEncodeHaibTfbsH1hescP300V0416102AlnRep2.bam wgEncodeHaibTfbsHepg2P300V0416101AlnRep1.bam

	wgEncodeHaibTfbsHepg2P300V0416101AlnRep2.bam wgEncodeSydhTfbsK562P300IggrabAlnRep1.bam wgEncodeSydhTfbsK562P300IggrabAlnRep2.bam
GABPA	wgEncodeAwgTfbsHaibGm12878GabpPcr2xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibH1hescGabpPcr1xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibHepg2GabpPcr2xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibK562GabpV0416101UniPk.narrowPeak.gz wgEncodeHaibTfbsGm12878GabpPcr2xAlnRep1.bam wgEncodeHaibTfbsGm12878GabpPcr2xAlnRep2.bam wgEncodeHaibTfbsH1hescGabpPcr1xAlnRep1.bam wgEncodeHaibTfbsH1hescGabpPcr1xAlnRep2.bam wgEncodeHaibTfbsHepg2GabpPcr2xAlnRep1.bam wgEncodeHaibTfbsHepg2GabpPcr2xAlnRep2.bam wgEncodeHaibTfbsK562GabpV0416101AlnRep1.bam wgEncodeHaibTfbsK562GabpV0416101AlnRep2.bam
JUN-D	wgEncodeAwgTfbsHaibH1hescJundV0416102UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibHepg2JundPcr1xUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhGm12878JundUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhK562JundIggrabUniPk.narrowPeak.gz wgEncodeHaibTfbsH1hescJundV0416102AlnRep1.bam wgEncodeHaibTfbsH1hescJundV0416102AlnRep2.bam wgEncodeHaibTfbsHepg2JundPcr1xAlnRep1.bam wgEncodeHaibTfbsHepg2JundPcr1xAlnRep2.bam wgEncodeSydhTfbsGm12878JundIggrabAlnRep1.bam wgEncodeSydhTfbsK562JundIggrabAlnRep2.bam
MAFK	wgEncodeAwgTfbsSydhH1hescMafkIggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhHepg2Mafkab50322IggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhK562Mafkab50322IggrabUniPk.narrowPeak.gz wgEncodeSydhTfbsGm12878MafkIggmusPk.narrowPeak.gz wgEncodeSydhTfbsGm12878MafkIggmusAlnRep1.bam wgEncodeSydhTfbsGm12878MafkIggmusAlnRep2.bam wgEncodeSydhTfbsH1hescMafkIggrabAlnRep1.bam wgEncodeSydhTfbsH1hescMafkIggrabAlnRep2.bam wgEncodeSydhTfbsHepg2Mafkab50322IggrabAlnRep1.bam wgEncodeSydhTfbsHepg2Mafkab50322IggrabAlnRep2.bam wgEncodeSydhTfbsK562Mafkab50322IggrabAlnRep1.bam wgEncodeSydhTfbsK562Mafkab50322IggrabAlnRep2.bam
MAX	wgEncodeAwgTfbsHaibK562MaxV0416102UniPk.narrowPeak.gz wgEncodeAwgTfbsSydhGm12878MaxIggmusUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhH1hescMaxUcdUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhHepg2MaxIggrabUniPk.narrowPeak.gz wgEncodeHaibTfbsK562MaxV0416102AlnRep1.bam wgEncodeHaibTfbsK562MaxV0416102AlnRep2.bam wgEncodeSydhTfbsGm12878MaxIggmusAlnRep1.bam wgEncodeSydhTfbsGm12878MaxIggmusAlnRep2.bam wgEncodeSydhTfbsH1hescMaxUcdAlnRep1V2.bam wgEncodeSydhTfbsH1hescMaxUcdAlnRep2V2.bam wgEncodeSydhTfbsHepg2MaxIggrabAlnRep1.bam wgEncodeSydhTfbsHepg2MaxIggrabAlnRep2.bam
MYC	wgEncodeAwgTfbsSydhH1hescCmycIggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhK562CmycUniPk.narrowPeak.gz

	wgEncodeAwgTfbsUtaGm12878CmycUniPk.narrowPeak.gz wgEncodeAwgTfbsUtaHepg2CmycUniPk.narrowPeak.gz wgEncodeOpenChromChipGm12878CmycAlnRep1.bam wgEncodeOpenChromChipGm12878CmycAlnRep2.bam wgEncodeOpenChromChipHepg2CmycAlnRep1.bam wgEncodeOpenChromChipHepg2CmycAlnRep2.bam wgEncodeOpenChromChipHepg2CmycAlnRep3.bam wgEncodeSydhTfbsH1hescCmycIggrabAlnRep1.bam wgEncodeSydhTfbsH1hescCmycIggrabAlnRep2.bam wgEncodeSydhTfbsK562CmycIggrabAlnRep1.bam wgEncodeSydhTfbsK562CmycIggrabAlnRep2.bam
NRF1	wgEncodeAwgTfbsSydhGm12878Nrf1IggmusUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhH1hescNrf1IggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhHepg2Nrf1IggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhK562Nrf1IggrabUniPk.narrowPeak.gz wgEncodeSydhTfbsGm12878Nrf1IggmusAlnRep1.bam wgEncodeSydhTfbsGm12878Nrf1IggmusAlnRep2.bam wgEncodeSydhTfbsH1hescNrf1IggrabAlnRep1.bam wgEncodeSydhTfbsH1hescNrf1IggrabAlnRep2.bam wgEncodeSydhTfbsHepg2Nrf1IggrabAlnRep1.bam wgEncodeSydhTfbsHepg2Nrf1IggrabAlnRep2.bam wgEncodeSydhTfbsK562Nrf1IggrabAlnRep1.bam wgEncodeSydhTfbsK562Nrf1IggrabAlnRep2.bam
RAD21	wgEncodeAwgTfbsHaibGm12878Rad21V0416101UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibH1hescRad21V0416102UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibHepg2Rad21V0416101UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibK562Rad21V0416102UniPk.narrowPeak.gz wgEncodeHaibTfbsGm12878Rad21V0416101AlnRep1.bam wgEncodeHaibTfbsGm12878Rad21V0416101AlnRep2.bam wgEncodeHaibTfbsH1hescRad21V0416102AlnRep1.bam wgEncodeHaibTfbsH1hescRad21V0416102AlnRep2.bam wgEncodeHaibTfbsHepg2Rad21V0416101AlnRep1.bam wgEncodeHaibTfbsHepg2Rad21V0416101AlnRep2.bam wgEncodeHaibTfbsK562Rad21V0416102AlnRep1.bam wgEncodeHaibTfbsK562Rad21V0416102AlnRep2.bam
REST	wgEncodeAwgTfbsHaibGm12878NrsfPcr1xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibH1hescNrsfV0416102UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibHepg2NrsfV0416101UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibK562NrsfV0416102UniPk.narrowPeak.gz wgEncodeHaibTfbsGm12878NrsfPcr1xAlnRep1.bam wgEncodeHaibTfbsGm12878NrsfPcr1xAlnRep2.bam wgEncodeHaibTfbsH1hescNrsfV0416102AlnRep1.bam wgEncodeHaibTfbsH1hescNrsfV0416102AlnRep2.bam wgEncodeHaibTfbsHepg2NrsfV0416101AlnRep1.bam wgEncodeHaibTfbsHepg2NrsfV0416101AlnRep2.bam wgEncodeHaibTfbsK562NrsfV0416102AlnRep1.bam wgEncodeHaibTfbsK562NrsfV0416102AlnRep2.bam
RFX5	wgEncodeAwgTfbsSydhGm12878Rfx5200401194IggmusUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhH1hescRfx5200401194IggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhHepg2Rfx5200401194IggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhK562Rfx5IggrabUniPk.narrowPeak.gz wgEncodeSydhTfbsGm12878Rfx5200401194IggmusAlnRep1.bam

	wgEncodeSydhTfbsGm12878Rfx5200401194IggmusAlnRep2.bam wgEncodeSydhTfbsH1hescRfx5200401194IggrabAlnRep1.bam wgEncodeSydhTfbsH1hescRfx5200401194IggrabAlnRep2.bam wgEncodeSydhTfbsHepg2Rfx5200401194IggrabAlnRep1.bam wgEncodeSydhTfbsHepg2Rfx5200401194IggrabAlnRep2.bam wgEncodeSydhTfbsK562Rfx5IggrabAlnRep1.bam wgEncodeSydhTfbsK562Rfx5IggrabAlnRep2.bam
SRF	wgEncodeAwgTfbsHaibGm12878SrfPcr2xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibH1hescSrfPcr1xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibHepg2SrfV0416101UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibK562SrfV0416101UniPk.narrowPeak.gz wgEncodeHaibTfbsGm12878SrfPcr2xAlnRep1.bam wgEncodeHaibTfbsGm12878SrfPcr2xAlnRep2.bam wgEncodeHaibTfbsH1hescSrfPcr1xAlnRep1.bam wgEncodeHaibTfbsH1hescSrfPcr1xAlnRep2.bam wgEncodeHaibTfbsHepg2SrfV0416101AlnRep1.bam wgEncodeHaibTfbsHepg2SrfV0416101AlnRep2.bam wgEncodeHaibTfbsK562SrfV0416101AlnRep1.bam wgEncodeHaibTfbsK562SrfV0416101AlnRep2.bam
SP1	wgEncodeAwgTfbsHaibGm12878Sp1Pcr1xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibH1hescSp1Pcr1xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibHepg2Sp1Pcr1xUniPk.narrowPeak.gz wgEncodeHaibTfbsGm12878Sp1Pcr1xAlnRep1.bam wgEncodeHaibTfbsGm12878Sp1Pcr1xAlnRep2.bam wgEncodeHaibTfbsH1hescSp1Pcr1xAlnRep1.bam wgEncodeHaibTfbsH1hescSp1Pcr1xAlnRep2.bam wgEncodeHaibTfbsHepg2Sp1Pcr1xAlnRep1.bam wgEncodeHaibTfbsHepg2Sp1Pcr1xAlnRep2.bam
TAF1	wgEncodeAwgTfbsHaibGm12878Taf1Pcr1xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibH1hescTaf1V0416102UniPk.narrowPeak.gz wgEncodeAwgTfbsHaibHepg2Taf1Pcr2xUniPk.narrowPeak.gz wgEncodeAwgTfbsHaibK562Taf1V0416101UniPk.narrowPeak.gz wgEncodeHaibTfbsGm12878Taf1Pcr1xAlnRep1.bam wgEncodeHaibTfbsGm12878Taf1Pcr1xAlnRep2.bam wgEncodeHaibTfbsH1hescTaf1V0416102AlnRep1.bam wgEncodeHaibTfbsH1hescTaf1V0416102AlnRep2.bam wgEncodeHaibTfbsHepg2Taf1Pcr2xAlnRep1.bam wgEncodeHaibTfbsHepg2Taf1Pcr2xAlnRep2.bam wgEncodeHaibTfbsK562Taf1V0416101AlnRep1.bam wgEncodeHaibTfbsK562Taf1V0416101AlnRep2.bam
TBP	wgEncodeAwgTfbsSydhGm12878TbpIggmusUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhH1hescTbpIggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhHepg2TbpIggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhK562TbpIggmusUniPk.narrowPeak.gz wgEncodeSydhTfbsGm12878TbpIggmusAlnRep1.bam wgEncodeSydhTfbsGm12878TbpIggmusAlnRep2.bam wgEncodeSydhTfbsH1hescTbpIggrabAlnRep1.bam wgEncodeSydhTfbsH1hescTbpIggrabAlnRep2.bam wgEncodeSydhTfbsHepg2TbpIggrabAlnRep1.bam wgEncodeSydhTfbsHepg2TbpIggrabAlnRep2.bam wgEncodeSydhTfbsK562TbpIggmusAlnRep1.bam wgEncodeSydhTfbsK562TbpIggmusAlnRep2.bam

USF2	wgEncodeAwgTfbsSydhGm12878Usf2IggmusUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhH1hescUsf2IggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhHepg2Usf2IggrabUniPk.narrowPeak.gz wgEncodeAwgTfbsSydhK562Usf2IggrabUniPk.narrowPeak.gz wgEncodeSydhTfbsGm12878Usf2IggmusAlnRep1.bam wgEncodeSydhTfbsGm12878Usf2IggmusAlnRep2.bam wgEncodeSydhTfbsH1hescUsf2IggrabAlnRep1.bam wgEncodeSydhTfbsH1hescUsf2IggrabAlnRep2.bam wgEncodeSydhTfbsHepg2Usf2IggrabAlnRep1.bam wgEncodeSydhTfbsHepg2Usf2IggrabAlnRep2.bam wgEncodeSydhTfbsK562Usf2IggrabAlnRep1.bam wgEncodeSydhTfbsK562Usf2IggrabAlnRep2.bam

*File prefixes denote the following base URLs:
wgEncodeAwg = http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/
wgEncodeHaib = http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/
wgEncodeSydh = http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/
wgEncodeUchicago = http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUchicagoTfbs/
wgEncodeOpenChrom = http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromChip/
wgEncodeUw = http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwTfbs/

Table 2.3. ENCODE ChIP-seq data files. File names for ChIP-seq data used in model training and comparisons of footprinters.

Transcription Factor	GM12878 Active	GM12878 Inactive	H1-hESC Active	H1-hESC Inactive	HepG2 Active	HepG2 Inactive	K562 Active	K562 Inactive
CEBPB	1695	96482	2732	10517	5612	4027	11438	229030
CHD2	7096	25202	3885	15856	3335	1599	4884	44050
CTCF	40146	661473	40910	83590	32083	113093	49546	875895
EP300	3481	125945	5655	95419	11998	55446	1319	663038
GABPA	5892	177426	4427	9436	8662	119616	10456	191479
JUND	1605	68708	2327	9776	6908	2475	20622	236287
MAFK	592	74172	2134	19569	3127	4780	11309	224351
MAX	10282	109679	7167	167547	9621	6646	35351	272022
MYC	3276	84147	3314	50435	3131	7310	3875	220364
NRF1	5293	57488	4216	51703	1821	61654	3863	151948
RAD21	32581	668688	40368	180678	32535	46745	31817	987156
REST	5029	332257	5282	289916	9065	84687	11713	1182481
RFX5	2583	48576	797	15346	2646	2213	1103	237826
SP1	11923	162337	10009	102770	11448	88642	-	-
SRF	5295	197928	2139	67446	2278	33010	2983	339381
TAF1	11143	155336	17261	57067	14577	74334	12783	417206
TBP	7341	121048	9695	35519	6998	9172	10052	358025
USF2	2877	7031	2293	1416	1988	544	1721	28189

Table 2.4. Active and inactive motif site set sizes.

Transcription Factor	Down-sampled K562 Data pAUCs (5% FPR)	Dataset Shift Correction pAUCs (5% FPR)
CEBPB	0.27	0.33
CHD2	0.56	0.84
CTCF	0.64	0.68
EP300	0.57	0.60
GABPA	0.76	0.85
JUND	0.68	0.75
MAFK	0.62	0.55
MAX	0.72	0.73
MYC	0.95	0.98
NRF1	0.47	0.77
RAD21	0.69	0.83
REST	0.65	0.15
RFX5	0.63	0.64
SRF	0.49	0.51
TAF1	0.90	0.93
TBP	0.66	0.89
USF2	0.63	0.77

Table 2.5. Classification performance for DeFCoM training phase variants. Partial AUC values for two variants of DeFCoM assessed on 17 TFs. Models were trained on K562 data and tested on GM12878 data sets.

Cell-line	FRiP Score	Total Sequencing Depth (Millions of reads)	Effective Sequencing Depth (Millions of reads)
GM12878	0.186284	245	45
H1-hESC	0.427376	110	47
HepG2	0.230881	50	11
K562	0.231837	365	84

Table 2.6. DNase-seq signal quality and sequencing depth statistics.

CHAPTER III

Characterizing molecular variation in Collaborative Cross mice at multiple levels of 1,3-butadiene exposure

OVERVIEW

It has been well demonstrated that genetic variation plays a large role in phenotypic variability. Genotype can influence modifications to the gene regulatory landscape that in turn affect transcription, protein expression, and ultimately a phenotype. Despite this established view of information flow within cells, a great challenge remains in uncovering which biological processes and components are at work within a particular context. A major goal in toxicogenomics is to understand these molecular relationships in response to toxicant exposure. For this study, I performed a descriptive analysis of how genetic variation and toxicant exposure relate to changes in chromatin organization and gene expression. Using the Collaborative Cross, a genetically diverse panel of multi-parent recombinant inbred mouse strains, I analyzed gene expression and chromatin accessibility data for lung, liver, and kidney tissue from 50 Collaborative Cross strains across three levels of exposure to 1,3-butadiene (BD), a gas used for the production of rubber and polymers. I also incorporated genetic data to perform quantitative trait loci (QTL) mapping of gene expression (eQTL) and chromatin accessibility (cQTL). From these analyses I observed tissue-specific differences in variability of gene expression and accessible chromatin in response to BD with lung exhibiting the largest differences. Additionally, I report eQTLs and cQTLs detected for each tissue in each of the three BD treatment groups and find most associations to be local for both eQTLs and cQTLs. In lung and kidney, "hotspot" genomic regions enriched for cQTLs were found, and we identified Collaborative Cross founder strain haplotypes as candidates for driving these hotspot associations.

INTRODUCTION

Chemical exposure can have distinct effects within individuals across tissues and cell types as well as across individuals [69,70]. To understand these differences at the molecular level, researchers are taking advantage of high-throughput assays for measuring various aspects of gene regulation, metabolism, and protein expression in relation to toxicant exposure [71]. Despite the use of multi-omics approaches for toxicology studies, the integration of these complementary data types with genotype information to gain a holistic view of toxicity susceptibility remains a challenge.

Of particular interest in this study is the DNA damage-inducing chemical 1,3-butadiene (BD). At room temperature, BD is an industrial gas and is mainly used in synthetic rubber and polymer production. These butadiene-based polymers are integrated into many commercial products such as automobiles, footwear, and plastics [72]. Additionally, BD is generally found at low concentrations in the environment and is also a component of tobacco smoke [73,74]. When inhaled, lung and liver microsomes metabolize BD into epoxide intermediates that react with DNA to form DNA adducts [75,76]. Importantly, mice and rat chronic inhalation studies have shown that BD exposure causes tumor formation in several tissues as a consequence of this DNA damage [77,78].

At the epigenetic level, changes in bulk DNA methylation and histone modification levels in response to BD have been observed for mouse lung and liver tissues, but significant changes were not observed in kidney [69,79]. In liver, these epigenetic marks were measured across a genetically diverse group of 7 inbred mouse strains (NOD/ShiLtJ, CAST/EiJ, A/J, WSB/EiJ, PWK/PhJ, C57BL/6J, and 129S1/SvImJ), and strain-specific epigenetic variation was found [79]. In this study, I further investigated the relationship between genetic variation, molecular variability, and BD exposure using the Collaborative Cross (CC), a genetically diverse population of inbred mouse strains with haplotypes inherited from the 7 aforementioned mouse strains plus NZO/H1LtJ. With gene expression and chromatin accessibility data for mice from 50 CC strains, I assessed global trends in variation of these molecular readouts across and within lung, liver, and kidney tissue at three BD exposure levels (control and two concentrations of BD). Through this analysis, I report tissue-specific differences in BD response. With available haplotype data, I performed gene expression quantitative trait loci (eQTL) and chromatin accessibility QTL (cQTL) mapping and provide an initial characterization of significant associations for each tissue and treatment group.

MATERIALS AND METHODS

Animals and 1,3-butadiene exposure

Male Collaborative Cross (CC) mice, obtained from UNC-CH (Chapel Hill, NC, USA), were housed in sterilized cages in a temperature-controlled (24°C) room with a 12/12-hr light/dark cycle and access to purified water and NIH-31 pelleted diet (Purina Mills, Richmond, IN, USA). Mice were randomly assigned to a control group or one of two experimental groups that I denote "625 ppm" and "1500 ppm". At approximately 10 weeks old, following a two-week acclimation period, mice were placed in cylindrical metal mesh exposure chambers for 6 hours a day, Monday-Friday, spanning a two-week period. Exposure chambers for control group mice emitted clean air, and 625 ppm group and 1500 ppm group exposure chambers contained an average concentration of 624±72 ppm and 1464±196 ppm of BD gas respectively. Immediately following the final exposure, mice were euthanized by exsanguination following deep nembutal (100 mg/kg intraperitoneal injection) anesthesia, and lungs, livers, and kidneys were excised, snap-frozen in liquid nitrogen, and stored at –80°C for subsequent processing. The animals were treated humanely and with regard for alleviation of suffering, and all procedures were approved by the Institutional Animal

Care and Use Committee at UNC-CH. Experimental procedures and preparation of mice samples were performed by the Rusyn Lab at Texas A&M University (TAMU).

Collaborative Cross reference genomes and transcriptomes

Alignment and processing of sample data from RNA-seq and ATAC-seq required CC strain-specific reference genomes and transcriptomes that I denote as "pseudo-genomes" and "pseudo-transcriptomes" respectively. Pseudo-genomes in FASTA file format and corresponding MOD files were downloaded from the CC resource website (http://csbio.unc.edu/CCstatus/index.py?run=Pseudo) for Build 37. A Build 37 MOD file provides a CC strain-specific mapping between genomic positions from a CC strain's pseudo-genome and the mm9 (C57BL/6J) genomic coordinate space. To construct pseudo-transcriptomes for each CC strain, I used the appropriate MOD file to convert mm9 RefSeq gene annotations into strain-specific gene annotations. These gene annotations in conjunction with the pseudo-genome FASTA files were passed as arguments into the RSEM (v1.2.31) [80] command *rsem-prepare-reference* with default parameter specifications.

RNA-seq and data processing

Total RNA was isolated from flash-frozen tissue samples using a Qiagen miRNeasy Kit (Valencia, CA) according to the manufacturer's protocol. RNA purity and integrity were evaluated using a Thermo Scientific Nanodrop 2000 (Waltham, MA) and an Agilent 2100 Bioanalyzer (Santa Clara, CA), respectively. A minimum RNA integrity value of 7.0 was required for RNA samples to be used for library preparation and sequencing. Libraries for samples with a sufficient RNA integrity value were prepared using the Illumina TruSeq Total RNA Sample Prep Kit (Illumina, Inc., San Diego, USA) with ribosomal depletion. Single-end (50bp) sequencing was performed (Illumina HiSeq 2500). RNA sample preparations were performed by the Rusyn Lab and sequencing was done by the sequencing facility at TAMU and the UNC-CH High-throughput Sequencing Facility (HTSF).

Following sequencing, reads were filtered to retain only those with a quality score of 20 or greater for at least 90 percent of read positions. Additionally, reads with adapter contamination were removed using TagDust [81]. For each sequenced RNA sample, reads were mapped to the appropriate pseudo-transcriptome using the RSEM command *rsem-calculate-expression* with STAR (v2.5.3a) [82] as the specified aligner (parameter set: --star). RSEM utilizes STAR with alignment options that follow ENCODE3 RNA-seq read mapping guidelines (https://www.encodeproject.org/pipelines/ENCPL002LSE/).

Gene expression quantification and gene set finalization

The RSEM command *rsem-calculate-expression* used for the RNA-seq read mapping also performs gene expression quantification and produces a transcripts per million (TPM) value for each gene specified in the pseudo-transcriptome. Samples were grouped by a combination of tissue type (liver, lung, and kidney) and treatment status (control, 625 ppm, and 1500 ppm) to produce a total of 9 sample groups. TPM values for samples within a group were median ratio normalized using DESeq2 [83] to make the values more comparable across samples. A requirement that the normalized TPM value must exceed 1 for a gene in at least 5% of the samples within a group was applied to exclude genes with sparse expression across samples. As a final filtering step, genes on chrY and chrM were excluded.

ATAC-seq and data processing

Flash frozen tissue samples were pulverized in liquid nitrogen using the BioPulverizer (Biospec) to break open cells and allow even exposure of intact chromatin to Tn5 transposase. Pulverized material was thawed in glycerol containing nuclear isolation buffer to stabilize nuclear structure [84] and then filtered through Miracloth (Calbiochem) to remove large tissue debris. Nuclei were washed and directly used for treatment with Tn5 transposase. Tissue processing was performed by the Crawford Lab at Duke University. Single-end (50bp) sequencing was performed by UNC-CH HTSF (Illumina HiSeq 2500).

Following sequencing, reads were filtered to retain only those with a quality score of 20 or greater for at least 90 percent of read positions. Additionally, reads with adapter contamination were removed using TagDust, and a maximum of 5 read duplicates were allowed. Prior to read mapping, a GSNAP database for each pseudo-genome was built using GMAP and the pseudo-genome FASTA file (parameter set: -k 15, -q 1). For each sample, reads remaining after filtering were aligned to the appropriate pseudo-genome using GSNAP (parameter set: -k 15, -m 1, -i 5, --sampling=1, --trim-mismatch-score=0, --genome-unk-mismatch=1, --query-unk-mismatch=1) [85]. Any reads that mapped to more than 4 genomic locations were removed.

Satellite repetitive elements, regions with high sequence homology to mitochondrial DNA, rRNA, and regions on chrX with high sequence homology to chrY are prone to producing artifactual signals caused by experimental or technical biases. Consequently, it has been recommended that these regions be excluded from sequencing-based analyses [86]. The ENCODE Consortium [15] created "blacklists" containing the aforementioned problematic regions for the human genome, and blacklists were generated following a similar procedure for the mm9 mouse reference genome. In the same manner, pseudo-genome specific blacklists were created by combining RepeatMasker [87] annotations and BLAT [88] derived chrX/Y homologous segments and genomic regions in strong sequence homology to mitochondrial DNA. These pseudo-genome blacklists were used to remove problematic genomic regions from consideration in further analyses.

Using the CC strain MOD files, mapped reads for each ATAC-seq sample were converted to mm9 genomic coordinates to enable direct comparison of data between samples. To account for any differences between the pseudo-genome blacklists and the mm9 blacklist, converted reads that mapped to mm9 blacklist regions were removed. Following conversion, all reads aligning to the positive strand were offset +5 bp, and all reads aligning to the negative strand were offset by -5 bp. These read shiftings account for a previously characterized behavior in the integration of adaptors by Tn5 transposase upon DNA binding [89].

Chromatin accessibility quantification and windowing

Samples were grouped by a combination of tissue type (liver, lung, and kidney) and treatment group (control and 625 ppm) to produce a total of 6 sample groups. For each sample, genomic regions representing high chromatin accessibility, i.e. peaks, were determined using the peak-calling software F-seq [58] with default parameters. To define an initial common set of chromatin regions for between group comparisons, across all sample groups the union set of the top 50,000 peaks (ranked by F-seq score) from each sample was derived and overlapping peaks were merged, resulting in 310,620 chromatin regions. These peaks were subsequently divided into overlapping 300 bp windows as previously described [90]. Briefly, peaks smaller than 300 bp were expanded to 300 bp, and for any peak larger than 300 bp, the number of 300 bp windows to segment the peak and not exceed its boundaries was determined using an initial overlap constraint of 100 bp. If the windows spanned less than 90% of bases within the peak, an additional window was added and the overlap was adjusted to produce uniformly spaced windows that exactly spanned the peak region. This windowing protocol resulted in 1.8 million windows, and per sample read coverage of each window was calculated using BEDTools *coverageBed* [91].

Read count values for samples within a group were median ratio normalized using DESeq2 to make the values more comparable across samples. To exclude windows with sparse read counts across samples, a filtering procedure similar to that described previously [92] was used. Windows were retained if at least 20% of samples within a sample group had high chromatin accessibility. High chromatin accessibility was defined as being in the top 20th percentile of normalized read counts of all windows across all samples in a group. Further filtering was done to include only the top 5% of windows ranked by total normalized counts across samples in a group. Presumably, these windows would more likely contain active regulatory elements while also providing the highest power to detect associations. As a final filtering step, regions on chrY and chrM were excluded.

Principal component analysis cluster significance testing

In determining whether two groups of points (or vectors) in a given principal component analysis (PCA) transformed space form distinct clusters, I first calculate 1) the Euclidean distance between a given point and the centroid of the group for which the point belongs and 2) the Euclidean distance between a given point and the centroid of the group for which the point does not belong. These are referred to as within group and between group distances respectively. The distributions of within group and between group distances were compared using a Wilcoxon rank-sum test with $\alpha \leq 0.05$ to determine if there exists a statistically significant difference between the distances, indicating significant clustering.

Differential gene expression analysis

The number of differentially expressed genes between treatment groups was determined using DESeq2. The 625 ppm and 1500 ppm groups were combined to represent one BD exposure group. Both the control and BD exposed group were reduced to samples representing only the strains in common between both groups to mitigate inconsistencies in genetic background. Expected counts estimated by RSEM were used to generate a matrix of gene abundances. Genes with no expected counts across samples were removed prior to analysis by DESeq2. To mitigate the influence of technical variation, batch and sequencing center were included as covariates in the DESeq2 model.

Haplotype reconstruction and segmentation for eQTL and cQTL analysis

Diplotype probabilities for each CC strain in this study were obtained from the CC resource database (http://csbio.unc.edu/CCstatus/index.py?run=FounderProbs) for genome version B37. These haplotype reconstructions were previously calculated using an HMM model for genotype array data [93]. The haplotype reconstructions provide probabilities of a genomic region being inherited from each of the 8 founder strains represented as probabilities for 36 genotype calls (8 homozygous and 28 heterozygous founder strain calls) for each genotype array marker. These probabilities were converted into haplotype dosages, i.e., the expected number of haplotypes. Let G be a symmetric matrix of genotype call probabilities for a marker m and G_{ij} be the genotype call probability for genotype ij comprised of two founder strain haplotypes $i,j \in \{A,B,C,D,E,F,G,H\}$. Note that in this case G_{ij} and G_{ji} are considered equivalent. The haplotype dosage calculation for founder strain haplotype k and marker m is

$$\sum_{i,j} f(i,j,k)\, G_{ij}$$

$$f(i,j,k) = \begin{cases} 2, & i = j \wedge i = k \wedge j = k \\ 1, & i \neq j \wedge (i = k \vee j = k) \\ 0, & i \neq k \wedge j \neq k \end{cases}$$

To reduce the computational burden of testing for haplotype-phenotype associations at each genotype marker, segmentation analysis was performed [94]. Briefly, since array marker densities exceed the total density of recombinations in the CC population, haplotype segment boundaries were redefined based on transitions between highest dosage haplotype. For each CC strain, a segment breakpoint was defined at a genotype marker whenever the highest dosage haplotype differed from the previous marker. The union set of breakpoints across all 50 CC strains in this study was used to construct the final segment boundaries, resulting in 4,970 segments. Segments on chrY and chrM were excluded. The mean and median segment sizes were 0.48 and 0.22 Mb respectively with the largest segment size being 11.47 Mb. For each haplotype, dosages were averaged for all markers within a segment. Using a consolidated set of segments for QTL analysis has been shown to produce essentially identical results to using a full set of genotype markers while simultaneously providing improvements in computational efficiency [94].

eQTL and cQTL analysis workflow

For both eQTL and cQTL analysis, the general approach for detecting haplotype-phenotype associations is the same, but the analyses differ in the phenotypes being assessed. For clarity and simplicity, the methodology here is elaborated in terms of gene expression (eQTL) for a single phenotype (gene) and haplotype segment but applies to chromatin windows (cQTL) as well. Associations are identified using a modified regression on probabilities (ROP) framework [95,96]. Normalized TPM values transformed using a rank-based inverse normal transform were regressed on averaged haplotype dosages. Sequencing center and batch were included as covariates. An *F*-test was performed to assess model fit with the null model specified to exclude haplotype dosages. This procedure was applied to all genes for all segments, and statistical significance of p-values was determined at a 5% False Discovery Rate (FDR) using the R package *qvalue* [97].

RESULTS

Experimental Design

A major goal of this study was to characterize the impact of genetic variability and BD exposure on gene expression and chromatin accessibility. To do so, mice representing 50 Collaborative Cross (CC) mouse strains were assigned into three groups that we denote as “control”, “625 ppm”, and “1500 ppm”. Each group contained one mouse for a given strain. The control group mice were placed in exposure chambers circulating clean air for 6 hours a day, five days per week, for two weeks. The 625 ppm and 1500 ppm groups were placed in exposure chambers with approximately 625 ppm and 1500 ppm concentrations of BD respectively. RNA-seq was performed on lung, liver, and kidney tissue for mice in each group as well as ATAC-seq for mice in the control and 625 ppm groups (Figure 3.1). The experiments were designed to maximize the number of strains represented in each group, but due to limitations and experiment-specific factors, RNA-seq and ATAC-seq data were not available for all tissues across all strains and groups (Table 3.1). After data processing, 35 strains were shared for both assays across all tissues and groups, 37 strains across all tissues and groups for RNA-seq, 40 strains across all tissues and groups for ATAC-seq, 43 strains across tissues in the ATAC-seq control group, 47 strains across tissues in the ATAC-seq 625 ppm group, 49 strains across tissues in the RNA-seq control group, 43 strains across tissues in the RNA-seq 625 ppm group, and 44 strains across tissues in the RNA-seq 1500 ppm group.

Tissue-type strongly contributes to gene expression variation between samples

To identify variables that prominently contribute to overall gene expression variability, I performed principal components analysis (PCA) on the gene expression profiles, derived from RNA-seq, of each individual in our study across all treatment groups and tissues. Principal Components (PCs) 1 and 2 contributed to 35.7% and 17.9% of the total variance respectively with the remaining 46.4% of total variance distributed among the remaining 406 PCs (Figure 3.2). Visualization of PCs 1-4 individually as Gaussian kernel density estimates and pairwise as scatterplots showed a clear separation of samples by tissue type in PCs 1 and 2 (Figure 3.3). PCs 3 and 4 did not exhibit separation by tissue, but did portray partial clustering by BD exposure status (Figure 3.4). To assess whether the clustering by BD exposure status was statistically significant, I applied a Wilcoxon rank-sum test to compare distributions consisting of the between group and within group distances to the group centroids using PCs 3 and 4 (see Materials & Methods). The p-value obtained provides support for a significant separation between clusters on PCs 3 and 4. (p < 2.2e-16).

Within-tissue gene expression variation reveals BD exposure associated effects

Because between-tissue gene expression variability showed more prominent tissue associated variation than BD associated effects, I performed PCA of gene expression profiles from all treatment groups for each tissue independently to investigate whether BD exposure contributed to noticeable variation within a tissue. Relative to the PCA across all three tissues, PC 1 for each tissue contributed less to the total variance with percent contributions being 9.7%, 8.5%, and 8.6% for lung, liver, and kidney respectively (Figure 3.5). Visualization of PCs for each tissue in a similar manner as the across-tissue PCA revealed a distinct separation of samples by BD exposure status in lung and liver but not in kidney on PCs 1 and 2 (Figures 3.6, 3.7, and 3.8). For samples in kidney, clusters associated with BD exposure became apparent on PCs 4 and 5. Statistical significance of the perceived groupings within each tissue were evaluated by the Wilcoxon rank-sum test as with the across tissue PCA analysis. For the clustering assessment of lung samples, PCs 1 and 2 showed a statistically significant separation between groups (p < 2.2e-16), as did PC 1 for liver samples (p < 2.2e-16) and PCs 4 and 5 for kidney samples (p < 2.2e-16). The clustering of kidney samples by BD exposure status occurred at later PCs relative to the PCA of lung and liver samples, suggesting that gene expression changes in kidney tissue in response to BD are not as prominent as in lung and liver tissue. Differential expression analysis was conducted using DESeq2 to determine the number of differentially expressed genes in the three tissues between control and BD exposed mice. Based on observations from PCA, the expected outcome was that fewer genes would be differentially expressed in kidney tissue compared to lung and liver, and this was observed. In kidney 3,639 genes were significantly differentially expressed between control and BD exposed samples whereas lung and liver had 6,936 and 6,512 differentially expressed genes respectively (FDR 0.05, Figure 3.9). These results in conjunction with the PCA show that BD exposure associated effects on gene expression are more pronounced in lung and liver tissue than kidney.

Chromatin accessibility differences by tissue type are more pronounced than BD exposure associated variation.

From PCA of gene expression profiles, differences between tissues were noted as the strongest source of variation. To assess whether this applied to chromatin region accessibility as well, PCA was performed on chromatin accessibility profiles constructed from ATAC-seq data for each sample across all treatment groups and tissues. Initially, the top 5% of 300 bp chromatin windows ranked by accessibility (total read counts) were analyzed (see Materials & Methods). A substantial amount of the total variance was captured by PC 1 (52.5%) with a sharp decrease to 6.9% explained by PC 2 (Figure 3.10). Visualization of PCs 1-5 as previously done for the gene expression PCA showed a less distinct separation of samples by tissue (Figure 3.11). Additionally, no separation of samples by treatment group was visually apparent (Figure 3.12). Because the top 5% of chromatin windows may be capturing many sites commonly accessible across tissues, PCA was also applied to the top 50% most accessible chromatin windows. Using a broader set of chromatin windows reduced the variance contribution of PC 1 to 32%, but increased the variance explained by PC 2 to 15.2%. The remaining PCs each contributed less than 1% (Figure 3.10). Interestingly, PCs 1 and 2 produce a stark separation of samples by tissue type that was less pronounced using the top 5% of chromatin windows, but samples still do not cluster by treatment group within the top 5 PCs (Figures 3.13 and 3.14). The observations made through comparison of the top 5% and top 50% of chromatin windows suggest that the most accessible regions are more likely to be active across tissues. In both cases, BD exposure associated variation did not appear to produce sample clusters.

Within-tissue chromatin accessibility variation captures BD exposure associated effects in lung and liver

To investigate whether samples within each tissue segregate by BD exposure status, PCA of chromatin accessibility profiles was performed separately for each tissue. Both the top 5% and top 50% most accessible chromatin windows were analyzed. Similar to the across tissue PCA, PC 1 for the top 5% of windows in each tissue comprised more variance than in the top 50% of windows (34.7-45.6% vs. 9.4-13.3%; Figures 3.15 and 3.16). In lung, PCs 3 and 4 for both sets of windows showed potential clustering by treatment group (Figures 3.17 and 3.18), but evaluation of clusters by a Wilcoxon rank-sum test only showed significant grouping for the top 50% (top 5% set p = 0.054; top 50% set p = 1.83e-12). From individual and pairwise visualization of the top 5 PCs, liver and kidney samples did not clearly separate by treatment status for the top 5% most accessible chromatin windows (Figures 3.19 and 3.21), and evaluation of their sample to centroid distances as aforementioned confirmed a lack of statistically significant grouping (liver p = 0.41; kidney p = 0.07). However, the top 50% set for liver samples produced a significant p-value on PCs 1-5 (p = 1.58e-5; Figure 3.20), but the p-value for the kidney top 50% set still did not reach significance (p = 0.1; Figure 3.22). These observations of chromatin accessibility variation reflect the general pattern seen with gene expression profiles where lung and liver appear to be more strongly affected by BD exposure than kidney tissue.

Identification of local and distal gene expression QTLs in CC mice

The public availability of CC genotype data allowed for me to investigate how genetic variation influences variability in transcriptional output. Treating gene expression as a molecular quantitative trait and regressing on haplotype dosages of genomic segments, eQTL mapping was performed using RNA-seq data for each treatment group (control, 625 ppm BD exposure, 1500 ppm BD exposure) in lung, liver, and kidney tissue. Because the numbers of CC strains in each group were not consistent, I reduced the strains in each group to the universal set of 35 strains in common. The total number of control group eQTLs detected in lung, liver, and kidney were 400, 505, and 869 respectively (FDR 0.05; Table 3.2; Figures 3.24-3.26). When comparing the genomic locations of segments to the positions of their significantly associated genes (eGenes), in all three tissues most associations were local which we define as 10 Mb from the gene transcription start site (TSS; Figure 3.23). This observation is consistent with previous eQTL studies in both mice and human [94,98]. For lung, 34 of the 67 eGenes had local associations and 36 eGenes paired with a distal segment. Of the 168 eGenes in liver, I observed 101 local and 82 distal associations. Kidney possessed the highest number of eGenes identified at 213 of which 105 paired with local segments and 124 had distal associations.

In the 625 ppm group, 505 lung eQTLs were found with 97 eGenes (FDR 0.05). Of those eGenes 58 were from local eQTLs, and 44 were associated with distal segments. In liver 842 eQTLs were detected for 102 eGenes with 44 pairing with local and 65 pairing with distal segments. Lastly, for the 2,336 kidney eQTLs observed, 162 of the 283 eGenes had local associations and 136 were distal (Table 3.2; Figures 3.27-3.29). Compared to the 625ppm group, the total number of 1500 ppm group eQTLs identified increased for all tissues to 869 (lung), 1,803 (liver), and 3,523 (kidney). The number of eGenes was 114 (66 local; 55 distal), 216 (128 local; 104 distal), and 417 (255 local; 188 distal) in lung, liver, and kidney respectively (Table 3.2; Figures 3.30-3.32). Relative to the control group, the number of eQTLs identified increased for each tissue in both the 625ppm and 1500 ppm group with the exception of the liver 625 ppm group. In comparing the distances of associations, in all treatment groups and tissues most of the associations were within 10 Mb (Figure 3.23).

Within a treatment group, the fraction of eQTLs that overlapped between tissues ranges from 0.04 to 0.21 when considering all eQTLs (FDR 0.05; Figure 3.33). Across treatment groups for the same pairwise comparisons, the pattern of overlaps remained fairly consistent with the largest change being 0.07. In each treatment group, the fraction of lung and liver eQTLs overlapping kidney was the highest. When considering local and distal eQTL overlaps separately, local eQTL overlaps were more concordant than distal within a treatment group (Figure 3.34). Between treatment groups within a tissue, the fraction of concordant eQTLs ranged from 0.1 to 0.26 for lung, 0.13 to 0.34 for liver, and 0.18 to 0.39 for kidney (Figure 3.33). Similar to the overlap within a treatment group and between tissues, a breakdown into local and distal eQTLs showed that overall, local eQTLs more consistently appeared across treatment groups than distal eQTLs (Figure 3.35). Observations that local regulation of gene expression within a tissue is the most consistent despite BD exposure supports the notion that tissue effects of the local regulatory landscape are more prominent than treatment effects. However, the consistency across treatment groups varies by tissue with kidney being the most consistent and lung being the least, suggesting that treatment effects are relatively stronger in lung. I also note that the low level of similarity between tissues suggests more disparate regulatory landscapes between tissues. In all cases, distal eQTLs appear less stable than local eQTLs, which could be related to difficulties in distal eQTL detection due to the small sample size and smaller effect sizes relative to local eQTLs.

Identification of local and distal chromatin accessibility QTLs in CC mice

In addition to the eQTL mapping, cQTL mapping was performed for the control and 625 ppm BD exposure group for lung, liver, and kidney tissues. Accessible chromatin regions often represent nucleosome-depleted, active gene regulatory elements [41], thus characterizing how genetic variation impacts chromatin accessibility provides an additional layer of information to complement eQTLs. Again, using the universal set of 35 CC strains, the control group cQTL mapping resulted in 3,328 lung, 88 liver, and 5,700 kidney cQTLs (FDR 0.05; Table 3.3, Figures 3.37-3.39). When assessing the distance of cQTL associations, in all three tissues the majority of cQTLs were local (within 10 Mb; Figure 3.36). For lung, 72 of the 325 cQTL chromatin windows had local associations, and 256 chromatin windows paired with a distal segment. Of the 15 cQTL chromatin windows in liver, 7 had local and 11 had distal associations. In kidney, 353 cQTL chromatin windows were identified of which 105 paired with local segments and 257 had distal associations. In the 625 ppm group, 8,472 lung cQTLs were found with 594 chromatin windows (FDR 0.05). Of those windows 93 had local associations, and 510 were associated with distal segments. Surprisingly, no cQTLs were considered significant at FDR 0.05 for liver. The lack of associations may be due to a combination of small sample size, limited chromatin windows tested, and less genetic variation associated with chromatin accessibility for our mouse strains, ultimately resulting in limited power to discover cQTLs (see Discussion). The number of kidney cQTLs decreased to 2,532 compared to the 5,700 detected in the control group. I observed 87 of the 238 chromatin windows with local associations and 164 with distal associations (Table 3.3, Figures 3.40-3.42). Similar to the control group results, most cQTL associations were shorter than 10 Mb, despite more chromatin windows having distal associations than local associations (Figure 3.36).

When assessing the cQTL overlap within a treatment group and between tissues, for the control group, 41% of lung cQTLs overlapped with kidney cQTL (FDR 0.05; Figure 3.42), but the percentage decreased to 12% for the 625 ppm group. Conversely, the percentage of kidney cQTLs that overlap lung cQTLs changes from 24% to 39% between treatment groups. This inverse relationship of lung and kidney between treatment groups appears related to the relationship between total cQTLs discovered, where detected lung cQTLs increased 2.5 fold upon BD exposure, and kidney cQTLs decreased 56%. This suggests that increasing the number of cQTLs detected will generally decrease the overlap fraction. When looking more specifically at local versus distal cQTL overlap, the control group showed similar local and distal cQTL overlap between lung and kidney, but the 625 ppm group exhibited no clear patterns. Notably, distal kidney cQTLs overlapped lung cQTLs by 47% (Figure 3.43).

In the eQTL analysis, overlaps within a tissue and between treatment groups were more consistent than within a treatment group and between tissues. The relationship is less clear for the cQTL results, but the highest overlaps were still observed within a tissue and between treatment groups (Figure 3.42). Breakdown by local and distal cQTLs revealed that more lung control group cQTLs overlapped with the 625 ppm group than vice versa for both local and distal cQTLs. The opposite relationship was observed in kidney. This further supports my observation that increasing the number of cQTLs detected will decrease the overlap fraction and may be contributing to a less apparent pattern in the cQTL results compared to the eQTL results.

Lung cQTL hotspots show founder strain specific phenotype clustering

The genome-wide plots of lung cQTL associations in both treatment groups produced visibly pronounced “hotspots” where a genomic locus was enriched for associations with chromatin windows genome-wide (Figures 3.37 and 3.40). The control group hotspot fell on chr14, and the 625 ppm hotspot was observed on chr13. Although this has not been previously characterized for cQTLs in the CC population, past studies have reported the occurrence of eQTL hotspots [99,100]. To better understand if a particular CC founder strain is driving the associations seen within the lung control group hotspot, the control mice were ordered by level of lung chromatin accessibility for chromatin windows chr14:76244575-76244874 and chr9:24956096-24956395 and haplotype dosages were visualized for haplotype segment UNC24188333-UNC24192133 (chr14:69792831-70034137, Figure 3.45). These chromatin windows produced the most significant local and distal associations for UNC24188333-UNC24192133. These cQTLs exhibited clustering of lower chromatin accessibility values for founder haplotype NOD/ShiLtJ and higher values for NZO/HiLtJ. The same visualization was applied for segment UNC23481318-UNC23486670 (chr13:118864010-119236105) and 2 significantly associated chromatin windows for the 625 ppm group cQTL mapping (Figure 3.46). For both associations, the C57BL/6J, NOD/ShiLtJ, and NZO/HiLtJ haplotypes showed distinct phenotype values that spanned the spectrum of measured chromatin accessibility. Although more hotspot associations need to be evaluated, these initial observations suggest that these haplotypes would be good candidates for further investigations into which haplotypes are driving the hotspot associations.

As noted previously, a moderate number of distal 625ppm kidney cQTLs overlapped with lung cQTLs. Upon assessing the locations of these overlapping cQTLs, I identified another hotspot location on chr8 shared between these two tissues (Figure 3.47). For two chromatin windows, chr8:19981785-19982084 and chr8:197000103-19700402, chromatin accessibility portrayed clear clustering by founder strain haplotype at segment JAX00663067 (chr8:23562597). Specifically, CAST/EiJ and PWK/PhJ showed the highest phenotype values, and WSB/EiJ had the lowest values. The contrast between the three haplotypes gives evidence for their strong effect on chromatin accessibility in these chromatin regions.

DISCUSSION

In this study I sought to characterize variability in gene regulation and transcription in three mouse tissues. In the GTEx pilot analysis, samples largely grouped by tissue based on hierarchical clustering of gene expression [24]. Through multi-tissue PCA, I made the same conclusion. Tissue type had the largest effect on both gene expression and chromatin accessibility. Interestingly, tissue type clustering was less apparent in the ATAC-seq samples when considering the top 5% of chromatin windows compared to the top 50%. This observation suggests that many of the more accessible chromatin regions were common across tissues, and inclusion of more windows incorporated more distal, tissue-specific chromatin regions. When evaluating the impact of BD exposure on gene expression and chromatin, I witnessed different levels of response between tissues. Lung presented the largest number of differentially expressed genes between control and BD exposed mice followed by liver. Both lung and liver exhibited greater than 1.7 times the number of differentially expressed genes compared to kidney. Lung and liver also exhibited the more notable variation in chromatin accessibility by treatment status compared to kidney. These observations align with previous findings of significant epigenetic changes in lung and liver but not kidney [69]. Given the more dynamic response of lung and liver gene expression to BD, further investigation of these genes may elucidate important regulatory pathways in BD metabolism and BD-related carcinogenesis.

An important advantage of the experimental design for this study was the use of CC mice. Being that the CC population consists of recombinant inbred mouse strains, CC mice with the same genetic background could be studied in three different treatment environments. Simultaneously, the representation of multiple CC strains within a treatment group provided genetic diversity that was utilized for eQTL and cQTL mapping. In the eQTL analysis, eQTLs were detected in all three tissues with most of the associations being local. In comparing the eQTLs identified for each tissue and treatment group, tissue specificity was noted by the higher overall overlap of eQTLs within a tissue and across treatment groups than between tissues and within a treatment group. In line with the aforementioned observation that lung has the most dynamic gene expression response to BD, eQTLs in lung were the least consistent across treatment groups relative to liver and kidney eQTLs.

Although the reproducible genetic diversity of the CC was advantageous in characterizing eQTLs, the small number of samples used in the cQTL mapping decreased power to detect associations. The choice to use a set of 35 CC strains common across all tissue and treatment groups was motivated by concerns that additional genetic variation from a CC strain unique to a specific group would create less comparable results. Unfortunately, a consequence of this decision was the discovery of only a handful of liver cQTLs. To assess the impact of including the additional 13 CC strains, I performed cQTL mapping for the liver control group utilizing the 48 mice/strains available for that group. Using an FDR 0.05 significance threshold, 7,991 cQTLs were detected as opposed to the 88 identified with a sample size of 35. The drastic difference reflects a significant loss in power to detect genotype related chromatin accessibility variation by excluding the 13 CC strains in cQTL mapping. Nevertheless, I still identified thousands of lung and kidney cQTLs with the reduced set of CC strains. As with the eQTL analysis, most associations were local, but cQTL hotspots were observed in lung and kidney despite the expectation that distal cQTL detection would be underpowered. To my knowledge, these hotspots are the first to be characterized for cQTLs in mice. Observations of haplotype dosages for select associations in each of the hotspots revealed likely founder strain haplotypes driving the chromatin accessibility differences. However, to better infer haplotype effects, more rigorous follow-up analyses need to be done such as the application of Diploffect, a Bayesian model described in [96] that estimates haplotype effects while accounting for uncertainty in the diplotype probability estimates.

In summary, these results serve as an initial characterization of tissue specific, genetic, and BD exposure related variability to gene expression and chromatin accessibility. I observed that all three factors play a role in gene regulatory differences, providing a basis for any or all factors being further investigated more in-depth. Through this work, I also demonstrated the strengths and weaknesses of the experimental design that will be informative for future studies that take advantage of the Collaborative Cross. As a whole, these results contribute to the growing body of studies that seek to better understand the relationship between genetics, gene regulation, environment, and phenotype.

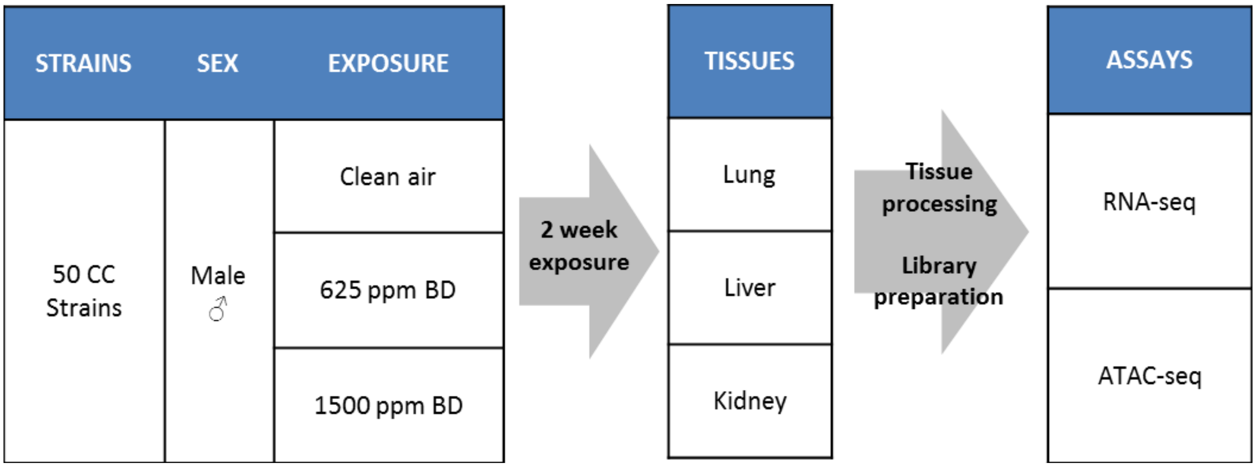

Figure 3.1. Experimental design overview. Male mice representing 50 CC strains were assigned to one of three exposure groups. After a two-week treatment period, lung, liver, and kidney tissue were obtained from each mouse which were processed for sequencing. ATAC-seq was not performed on tissue from the 1500 ppm group mice.

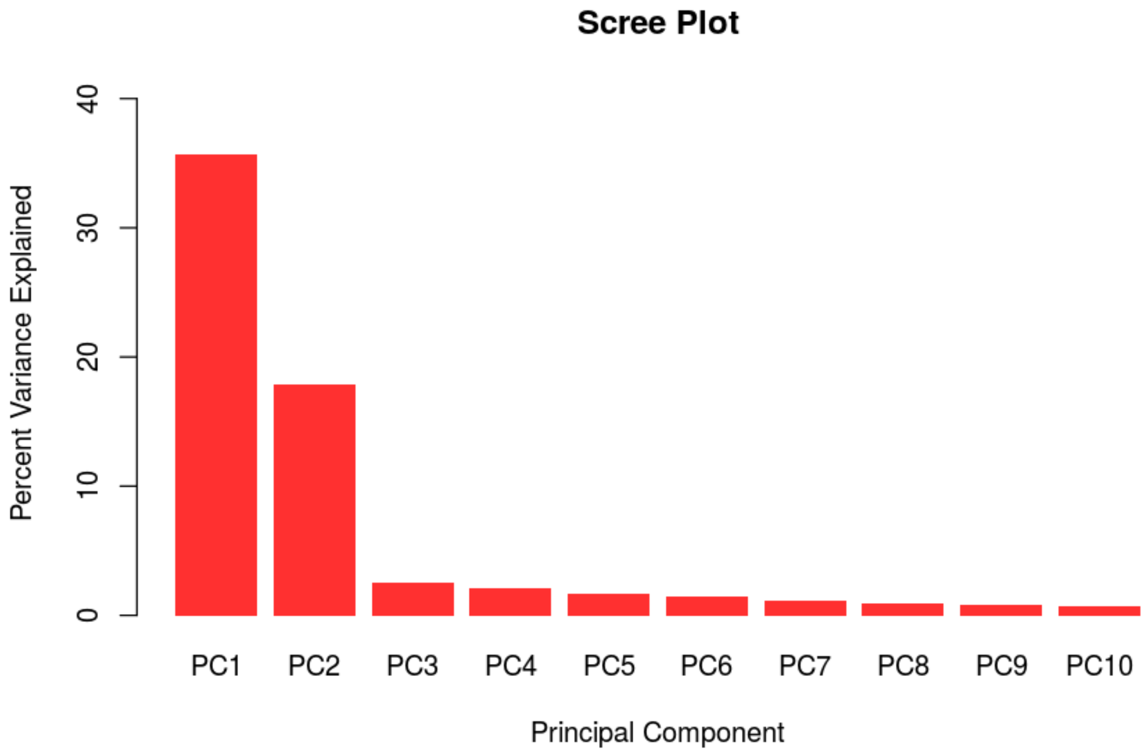

Figure 3.2. Multi-tissue gene expression PCA scree plot. Percent of variance explained by the top 10 PCs for PCA of gene expression profiles for all samples.

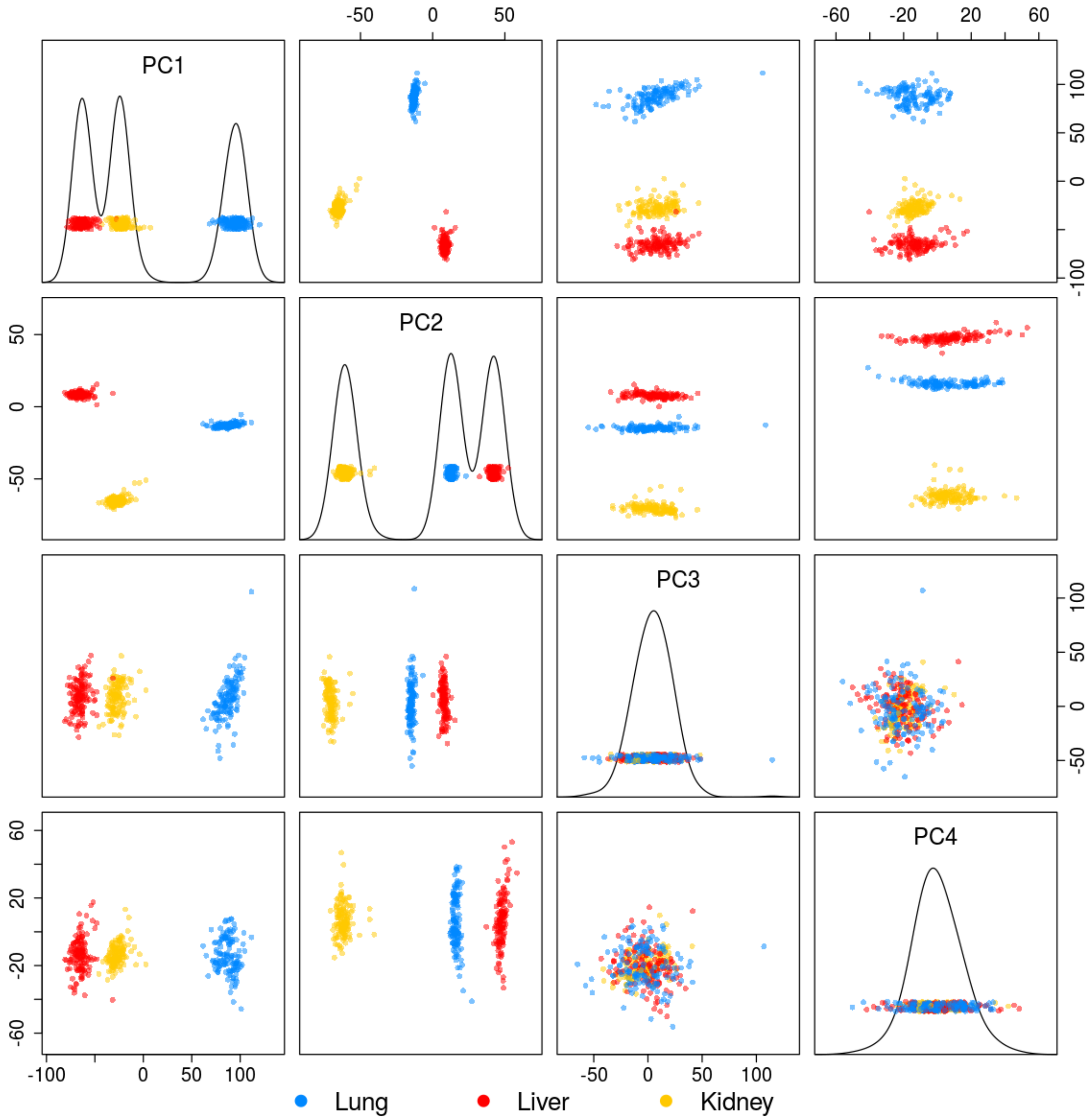

Figure 3.3. Multi-tissue gene expression PCA plot colored by tissue type. PCA plot for the top 4 PCs from PCA of gene expression profiles for all samples. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-4.

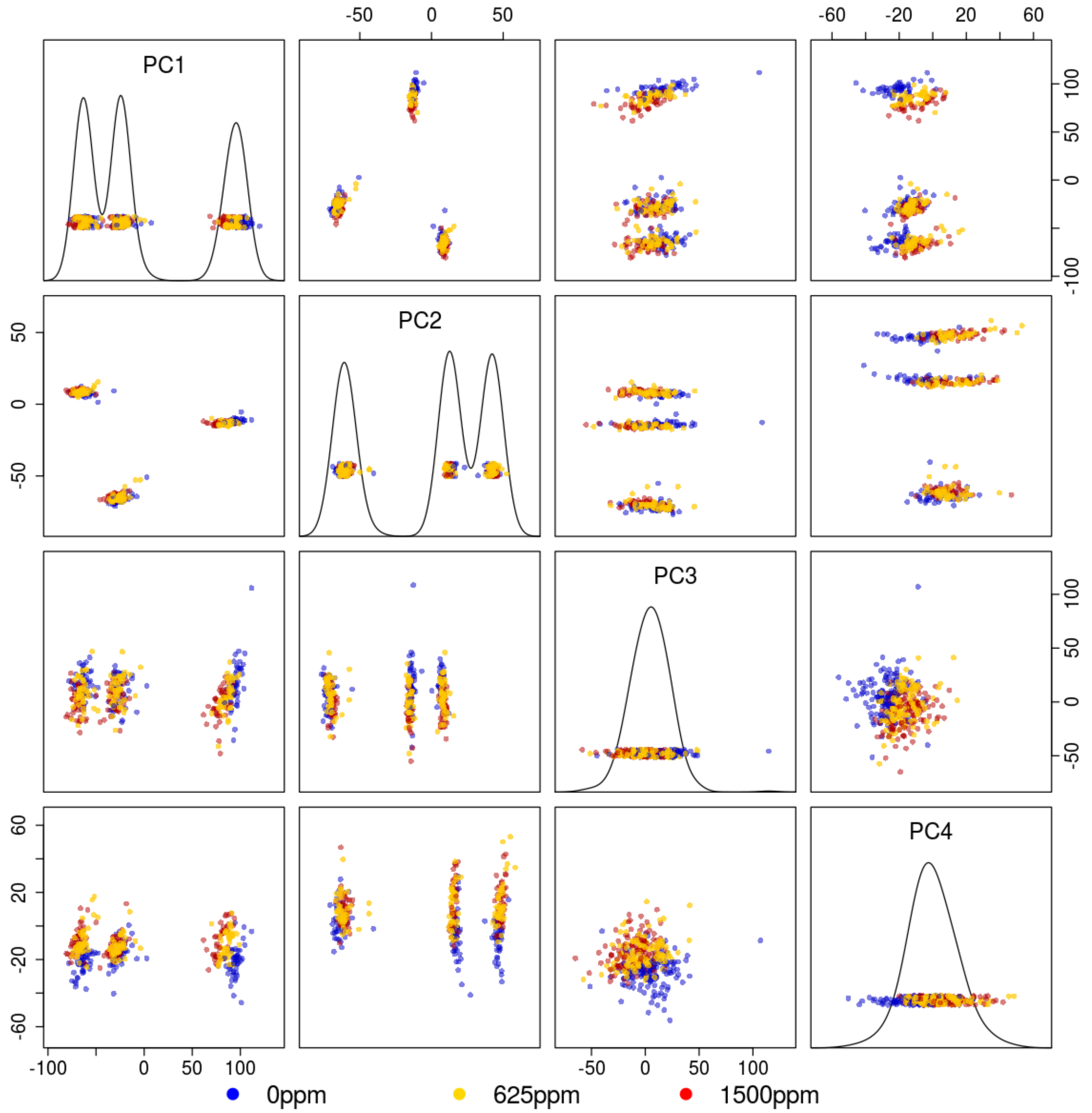

Figure 3.4. Multi-tissue gene expression PCA plot colored by treatment status. PCA plot for the top 4 PCs from PCA of gene expression profiles for all samples. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-4.

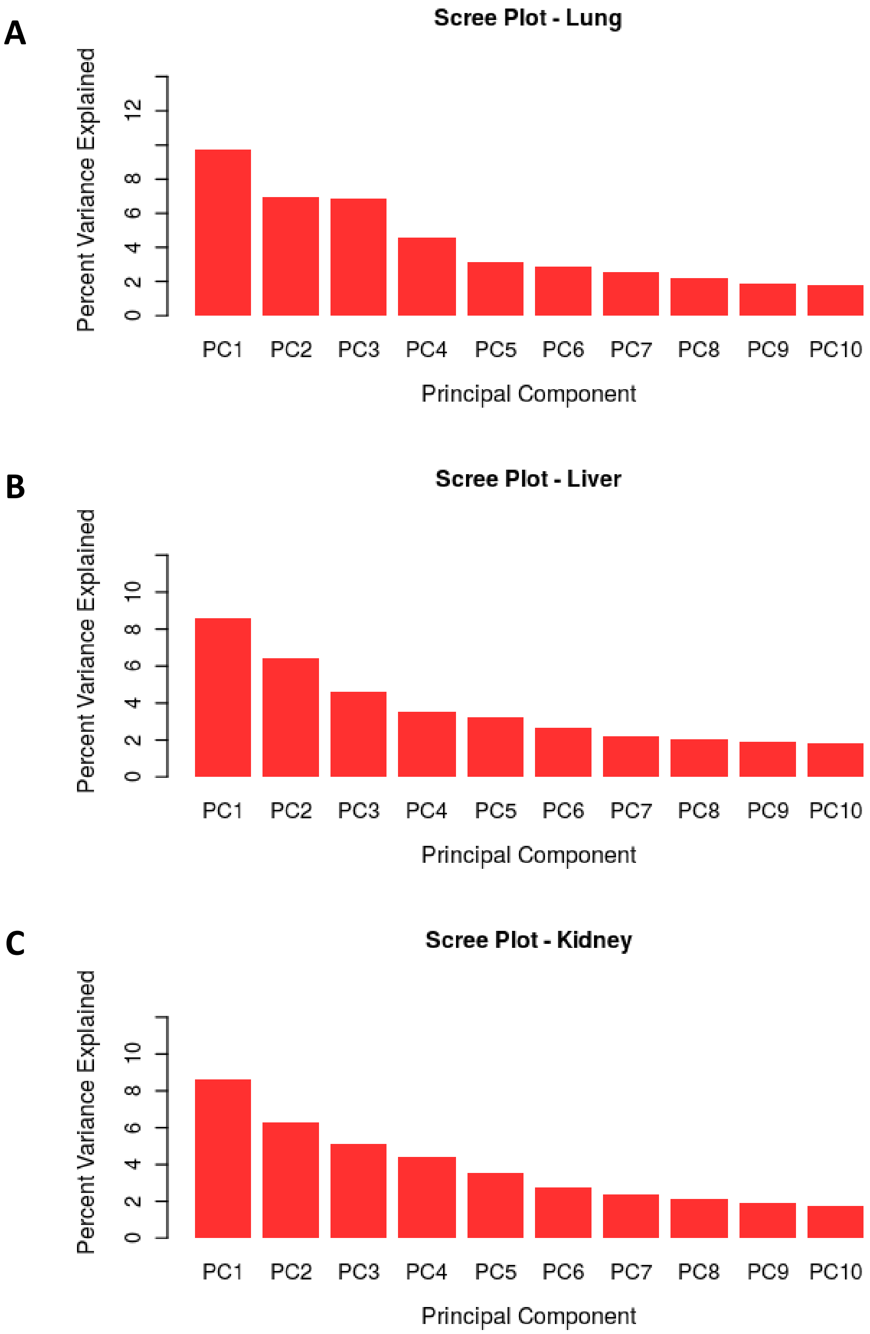

Figure 3.5. Per-tissue gene expression PCA scree plots. Percent of variance explained by the top 10 PCs for PCA of gene expression profiles for A) Lung, B) Liver, and C) Kidney samples.

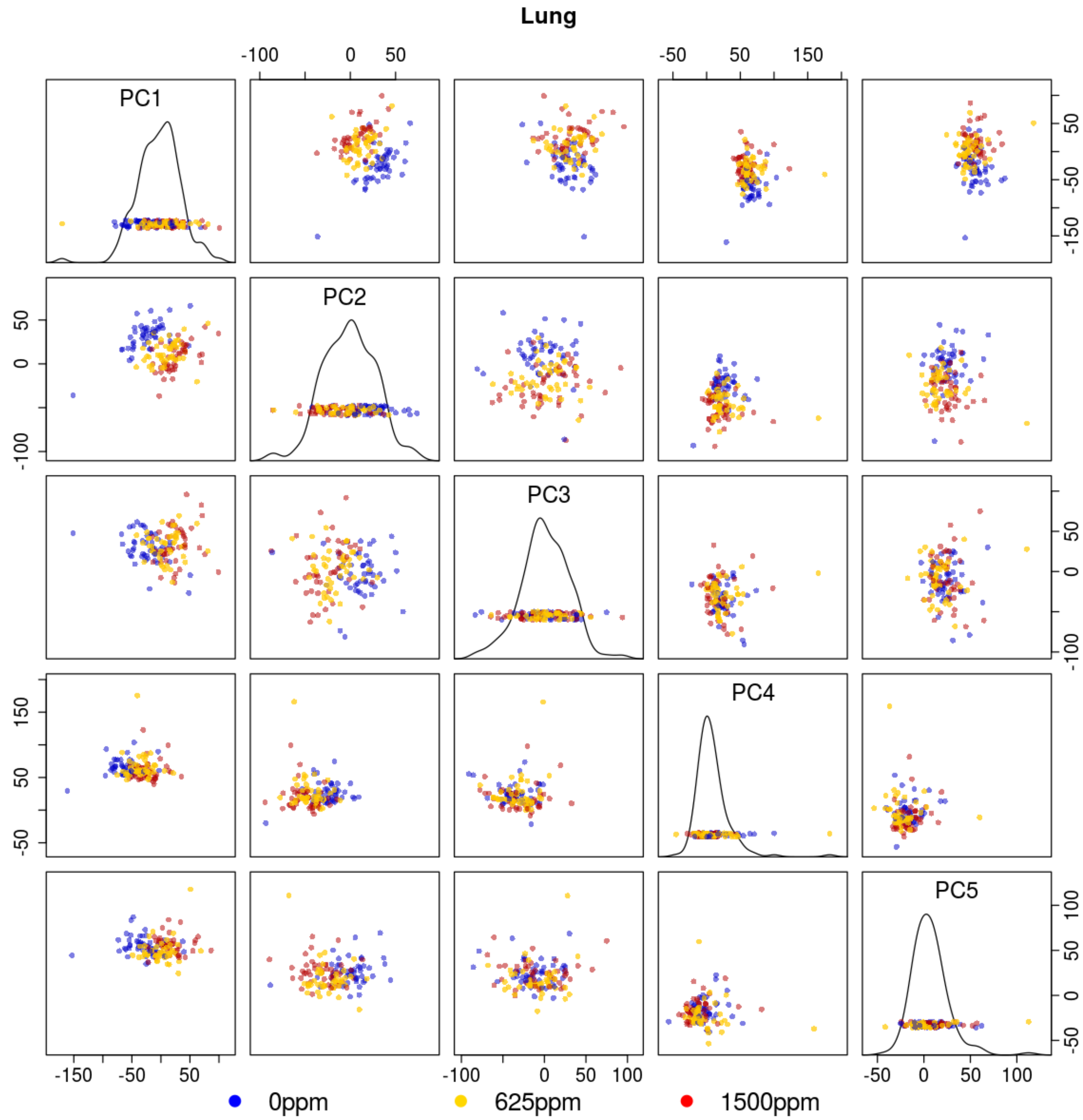

Figure 3.6. Lung gene expression PCA plot. PCA plot for the top 5 PCs from PCA of gene expression profiles for lung samples. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

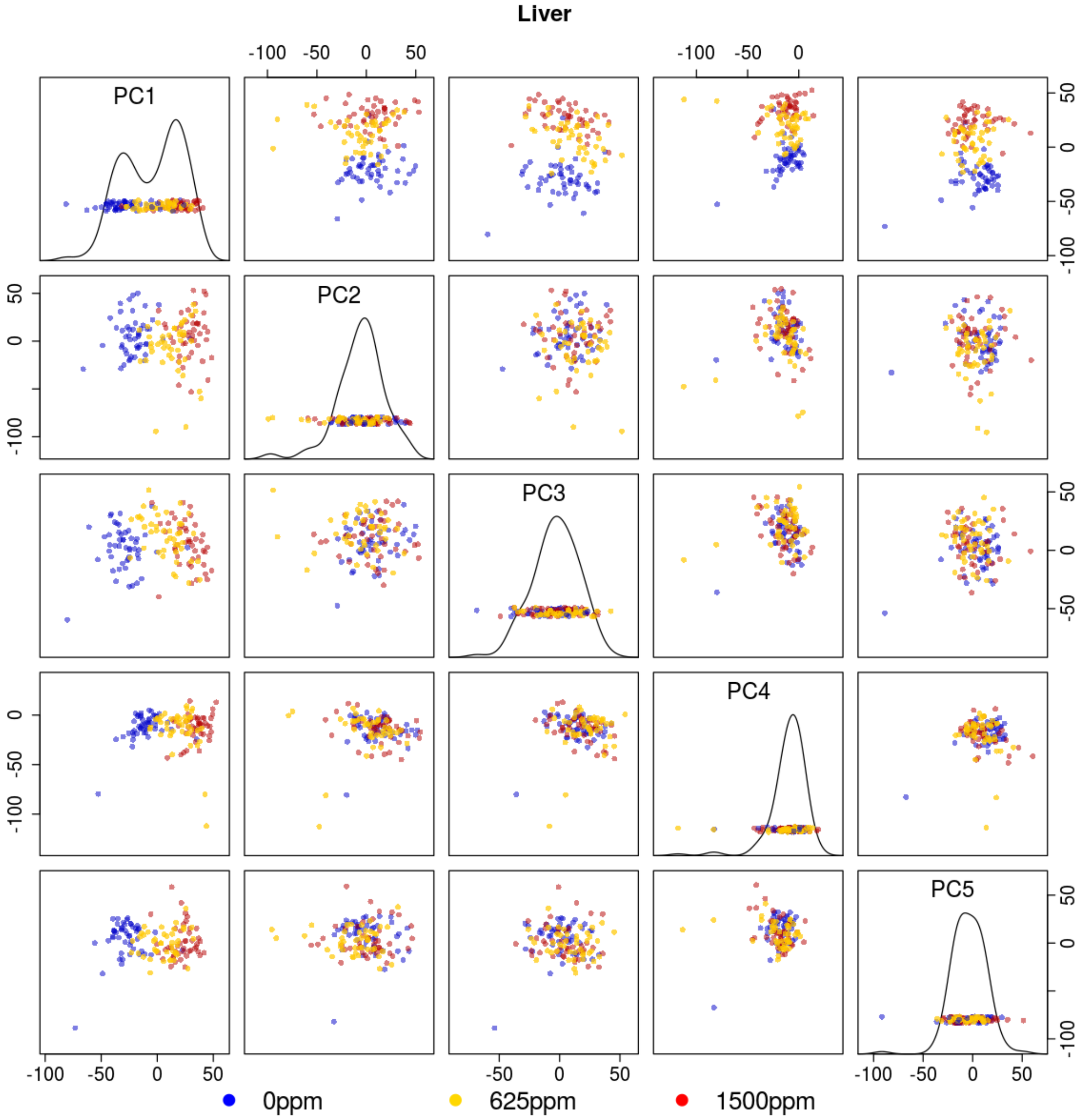

Figure 3.7. Liver gene expression PCA plot. PCA plot for the top 5 PCs from PCA of gene expression profiles for liver samples. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

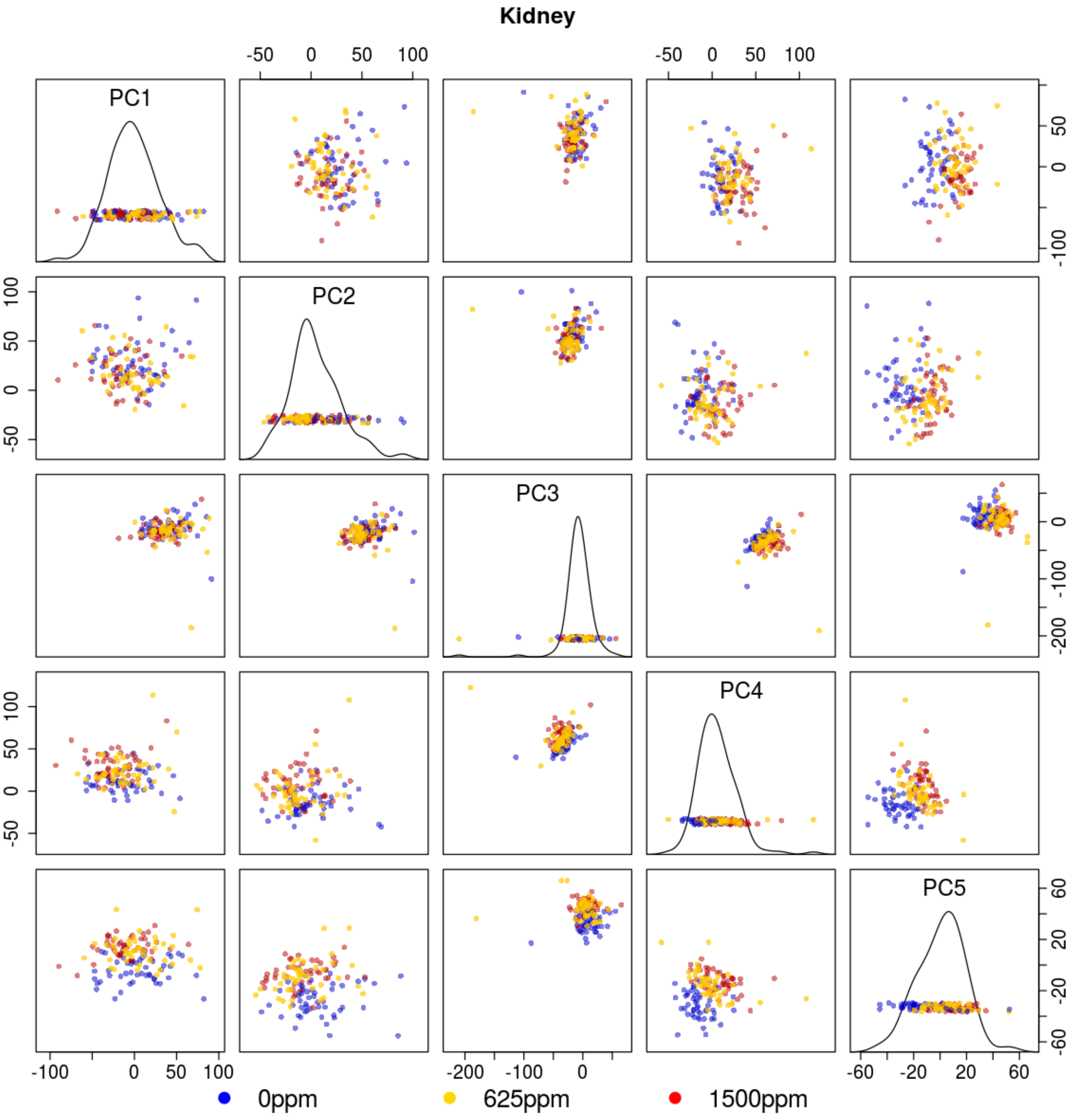

Figure 3.8. Kidney gene expression PCA plot. PCA plot for the top 5 PCs from PCA of gene expression profiles for kidney samples. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

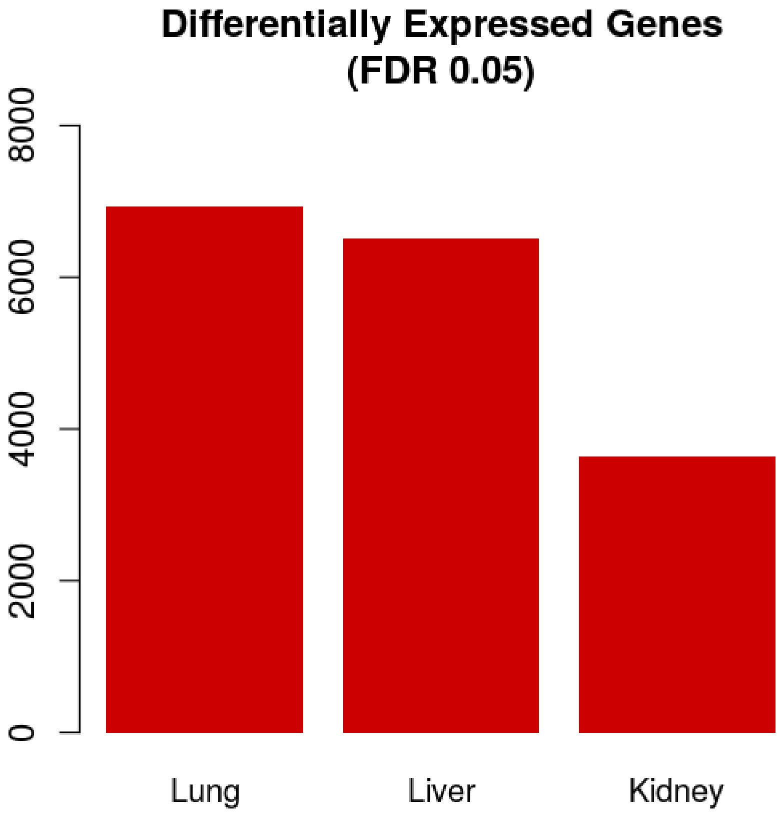

Figure 3.9. Differentially expressed genes by tissue. Number of differentially expressed genes between control and BD exposed treatment groups detected by DESeq2 at FDR 0.05 for lung, liver, and kidney tissue samples.

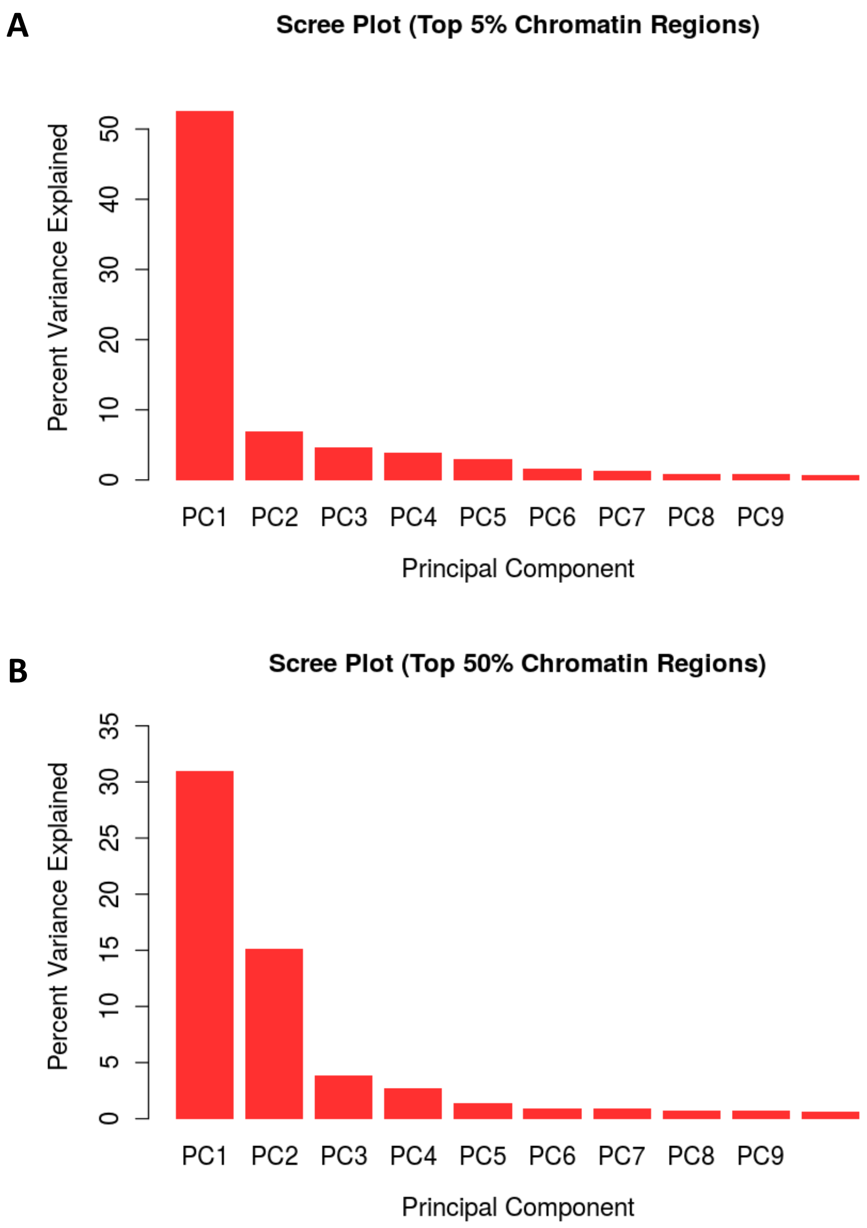

Figure 3.10. Multi-tissue chromatin accessibility PCA scree plots. Percent of variance explained by the top 10 PCs for PCA of chromatin accessibility profiles for all samples using A) the top 50% of chromatin windows ranked by chromatin accessibility and B) the top 5% of chromatin windows ranked by chromatin accessibility.

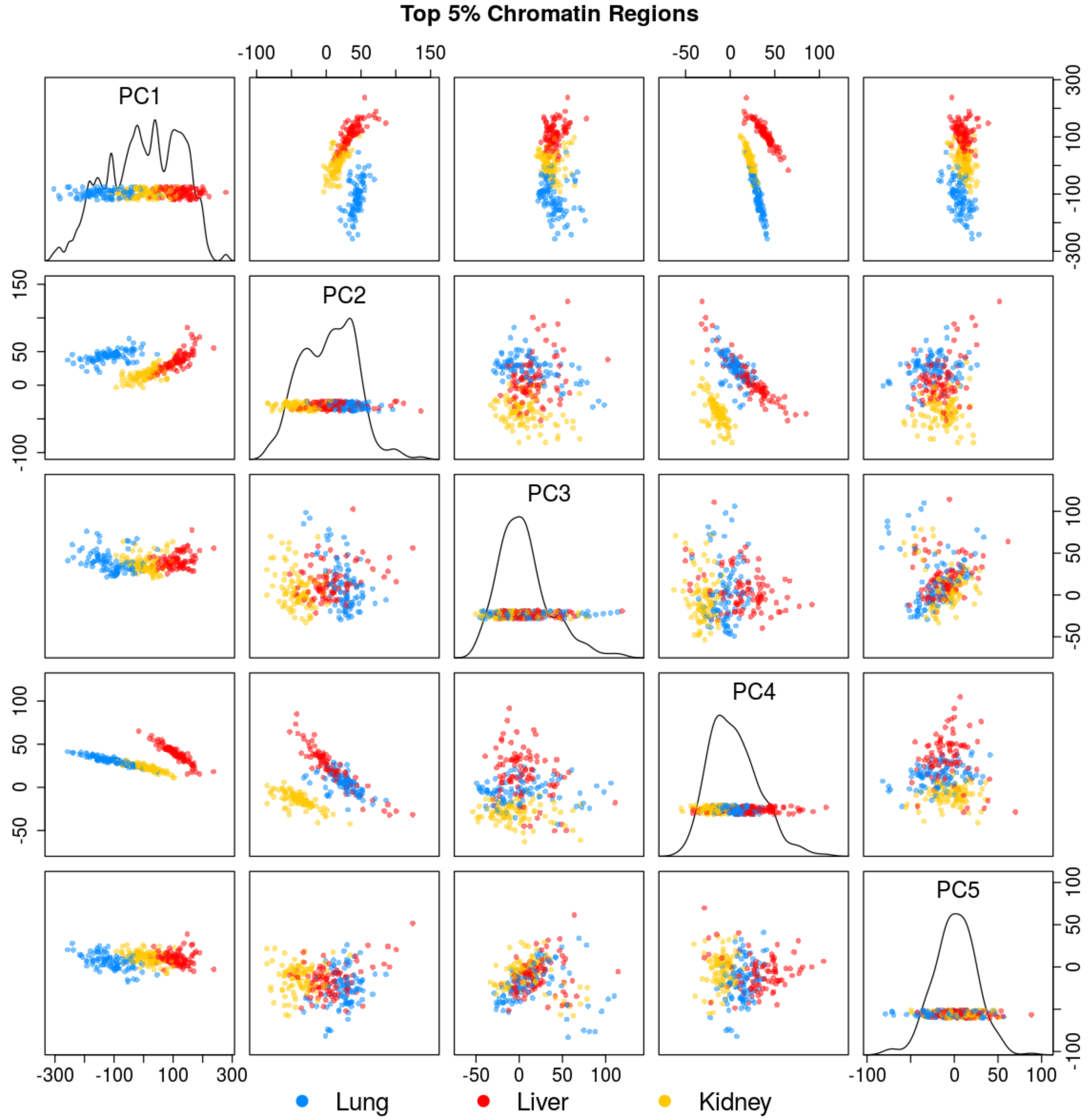

Figure 3.11. Multi-tissue top 5% chromatin windows PCA plot colored by tissue type. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for all samples using the top 5% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

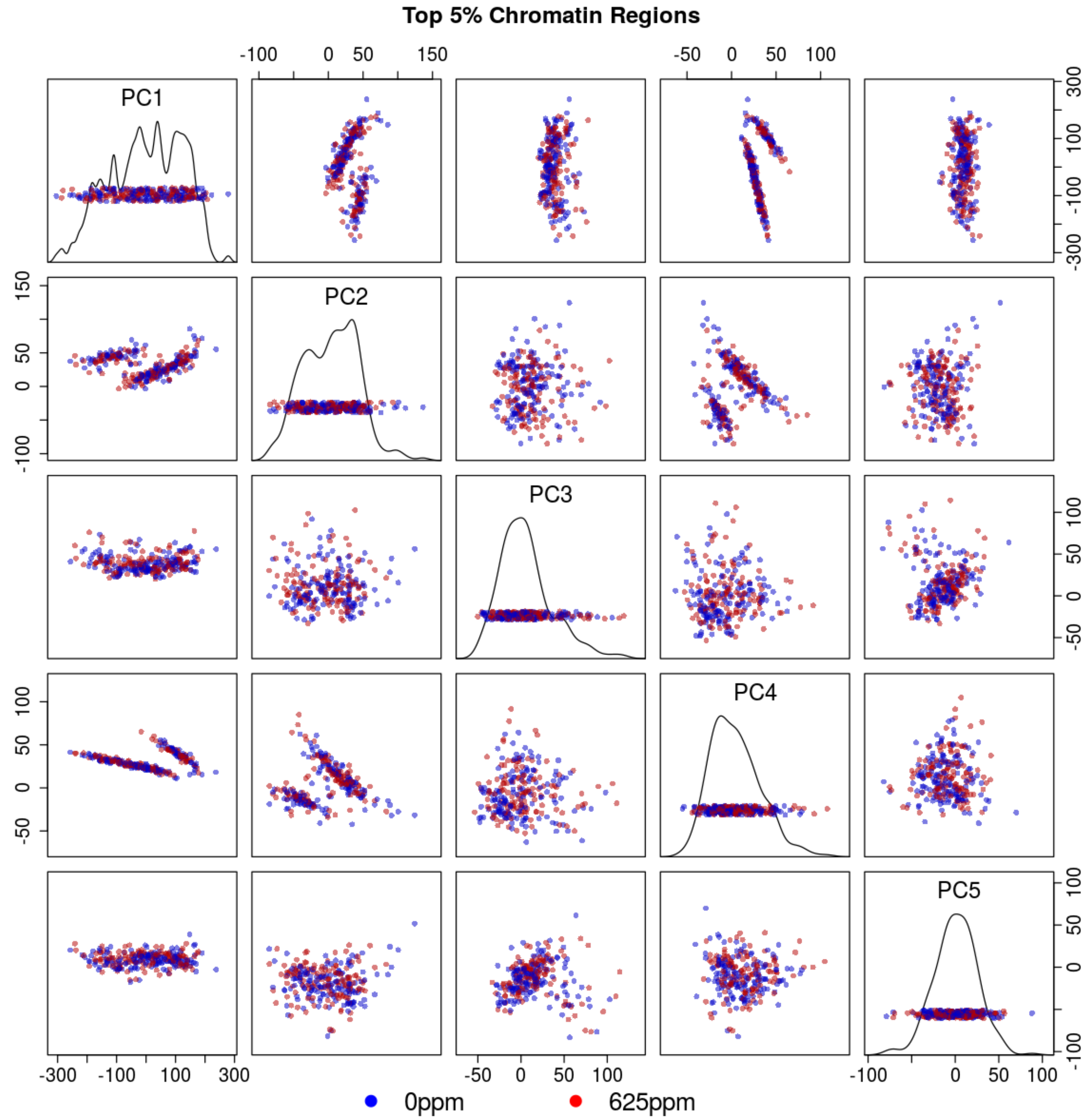

Figure 3.12. Multi-tissue top 5% chromatin windows PCA plot colored by treatment status. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for all samples using the top 5% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

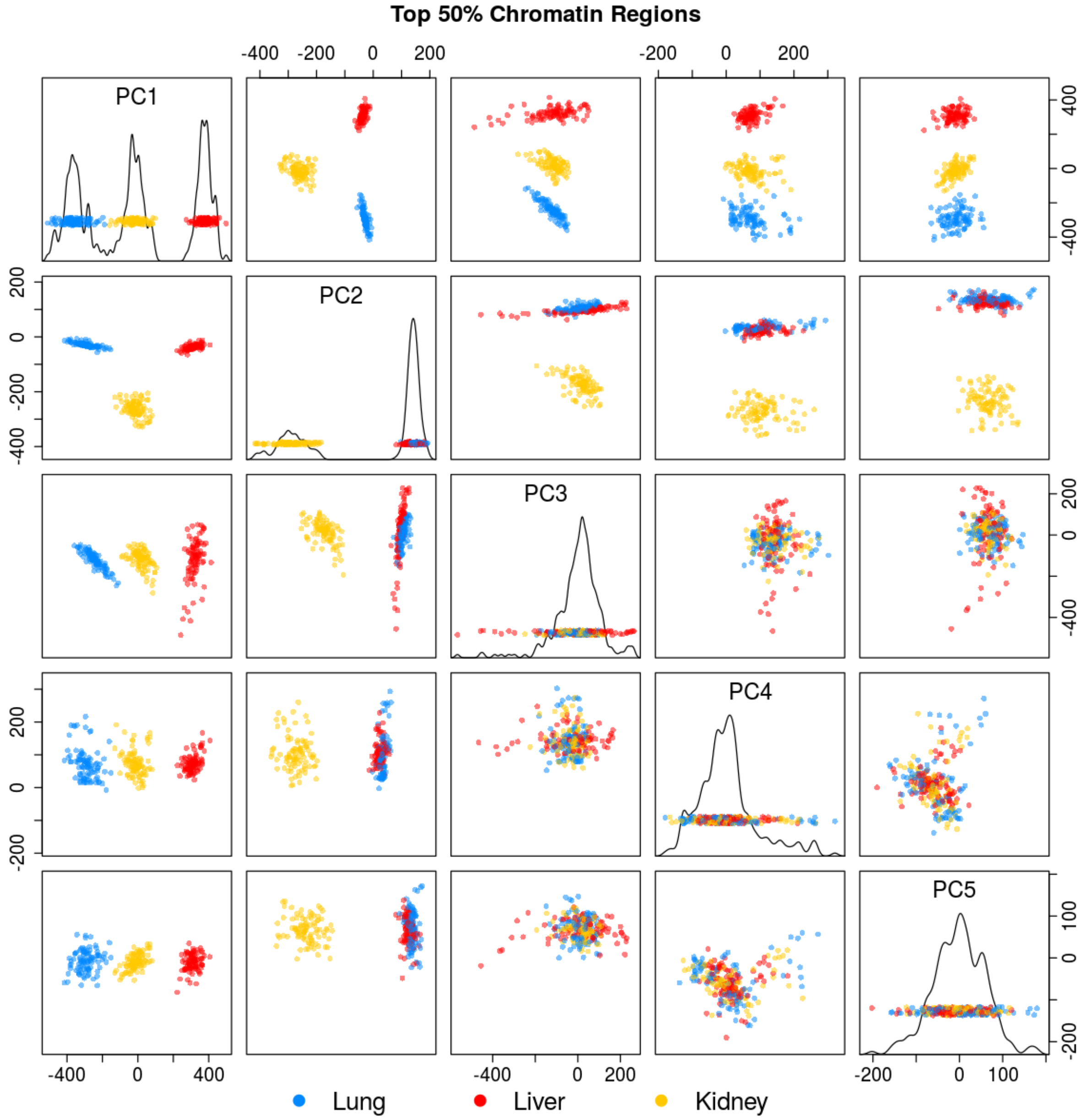

Figure 3.13. Multi-tissue top 50% chromatin windows PCA plot colored by tissue type. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for all samples using the top 50% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

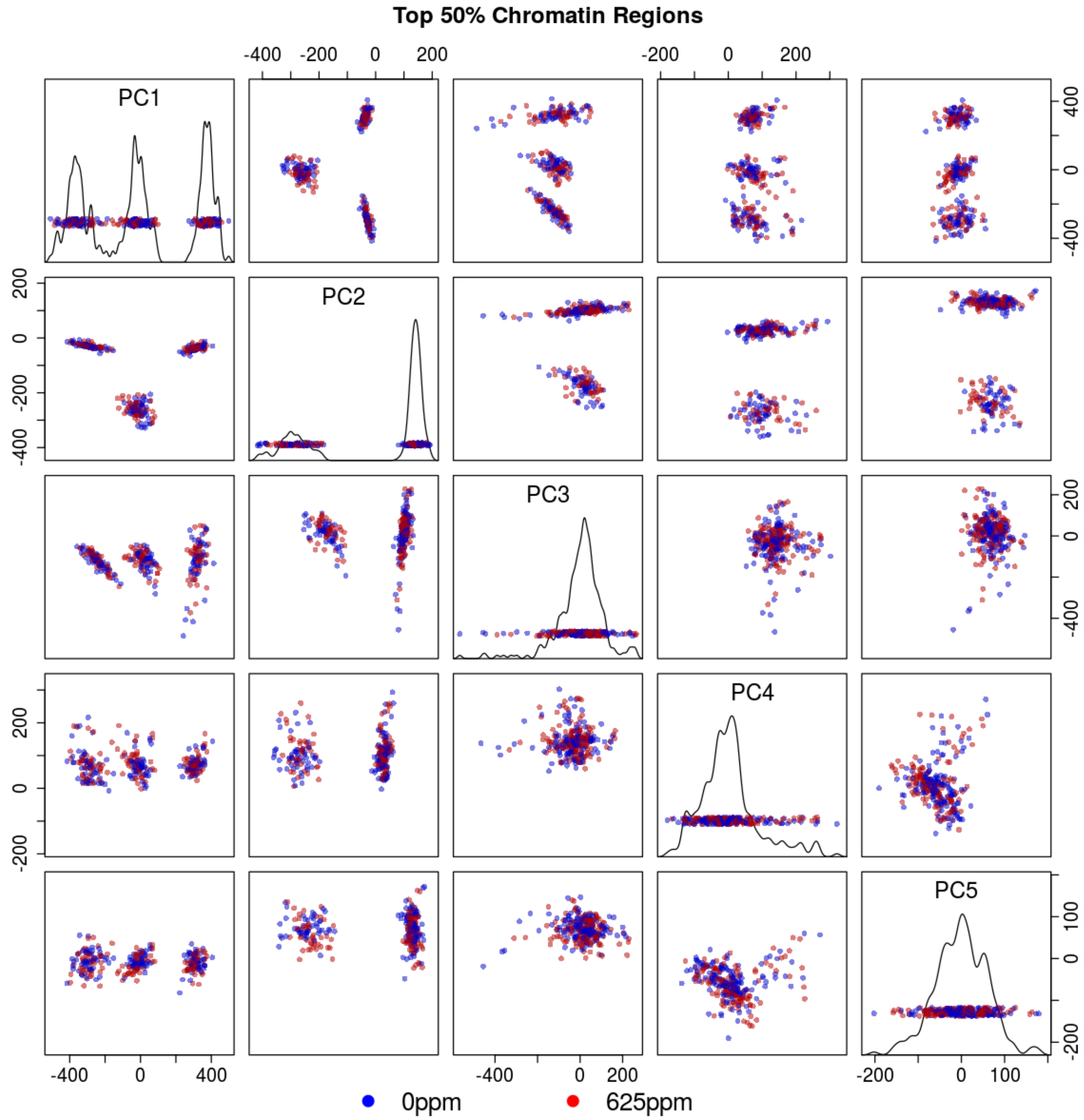

Figure 3.14. Multi-tissue top 50% chromatin windows PCA plot colored by treatment status. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for all samples using the top 50% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

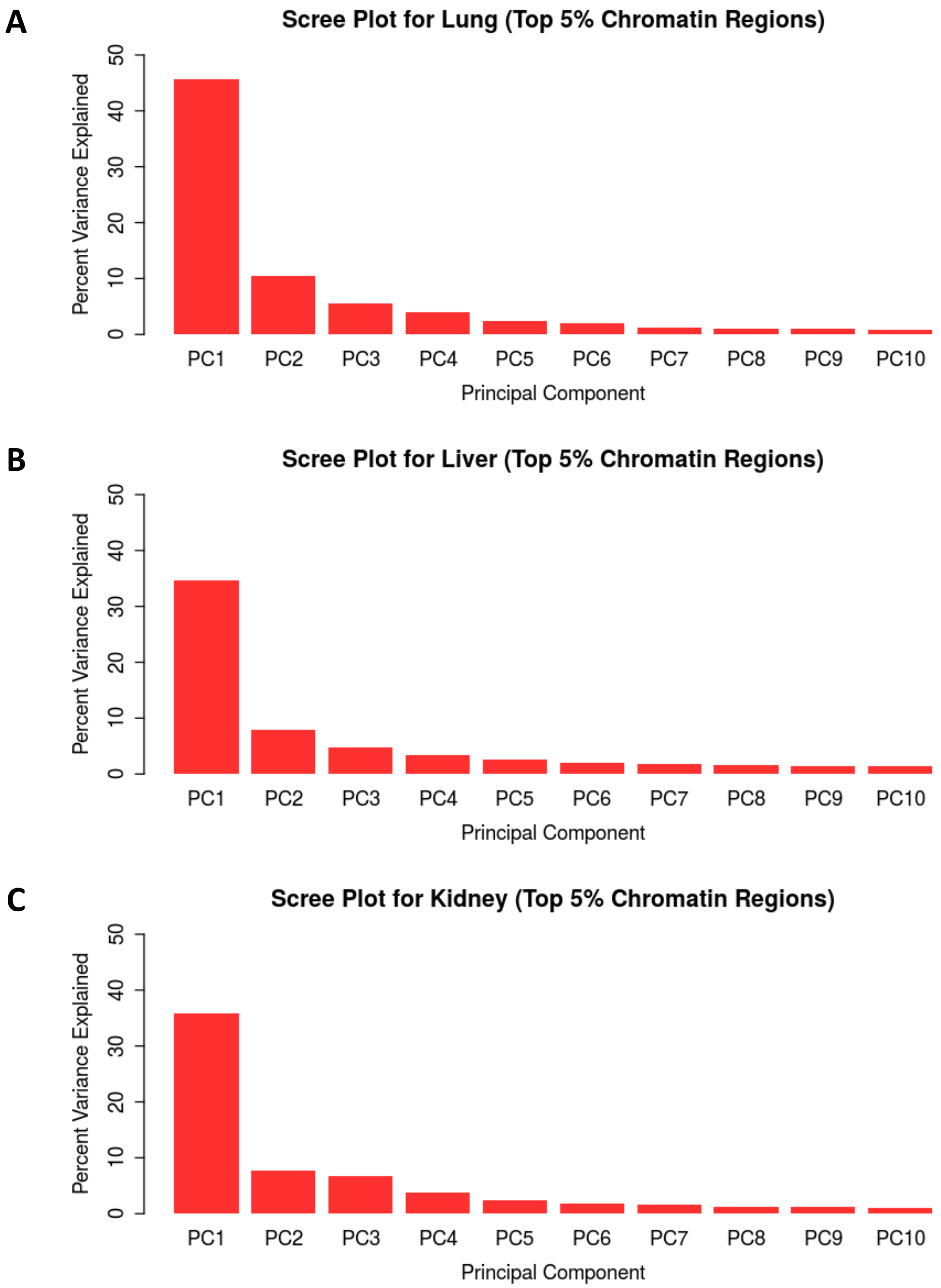

Figure 3.15. Per-tissue top 5% chromatin windows PCA scree plots. Percent of variance explained by the top 10 PCs for PCA of chromatin accessibility profiles for all samples using the top 5% of chromatin windows ranked by chromatin accessibility in A) lung, B) liver, and C) kidney.

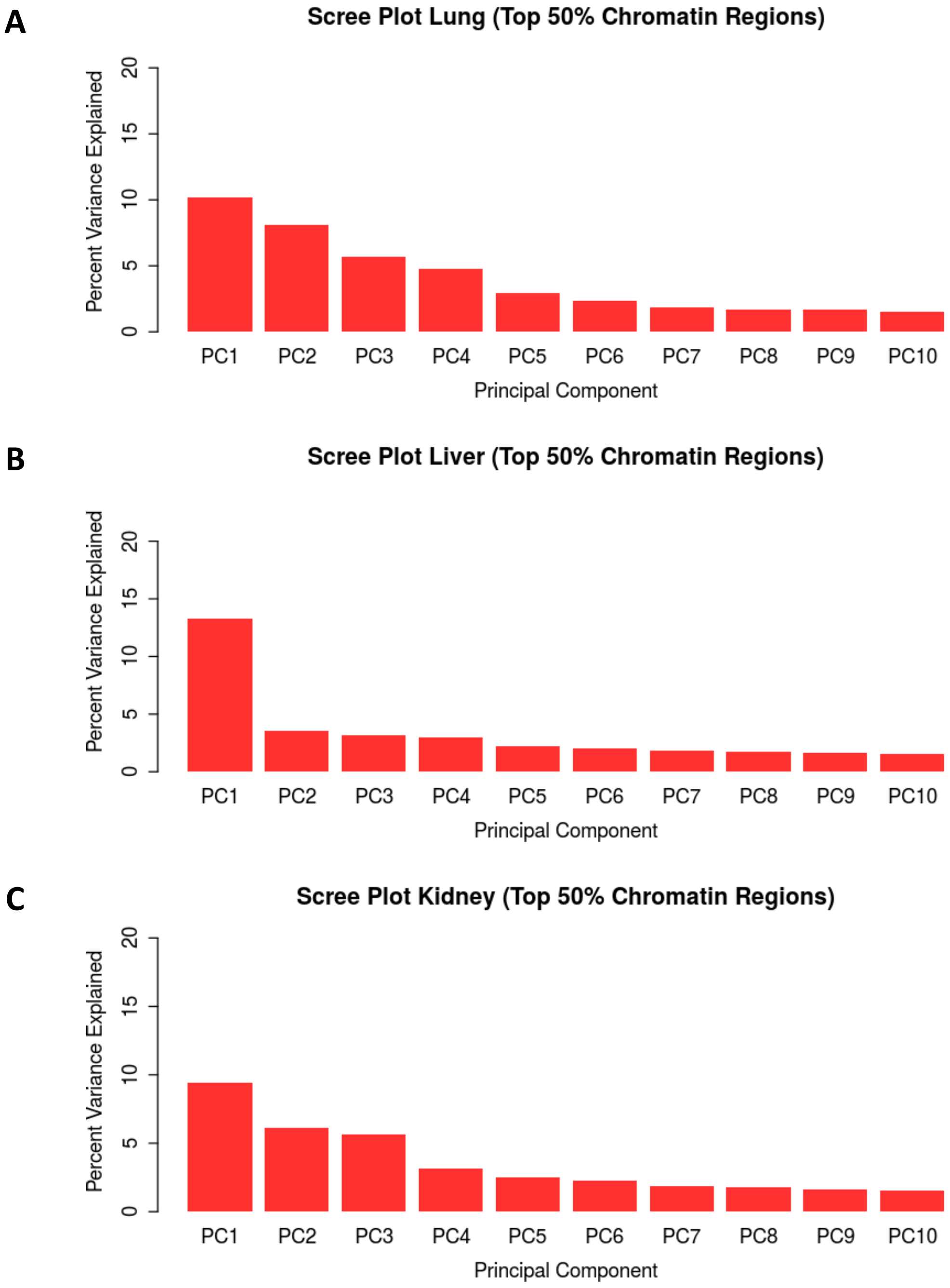

Figure 3.16. Per-tissue top 50% chromatin windows PCA scree plots. Percent of variance explained by the top 10 PCs for PCA of chromatin accessibility profiles for all samples using the top 50% of chromatin windows ranked by chromatin accessibility in A) lung, B) liver, and C) kidney.

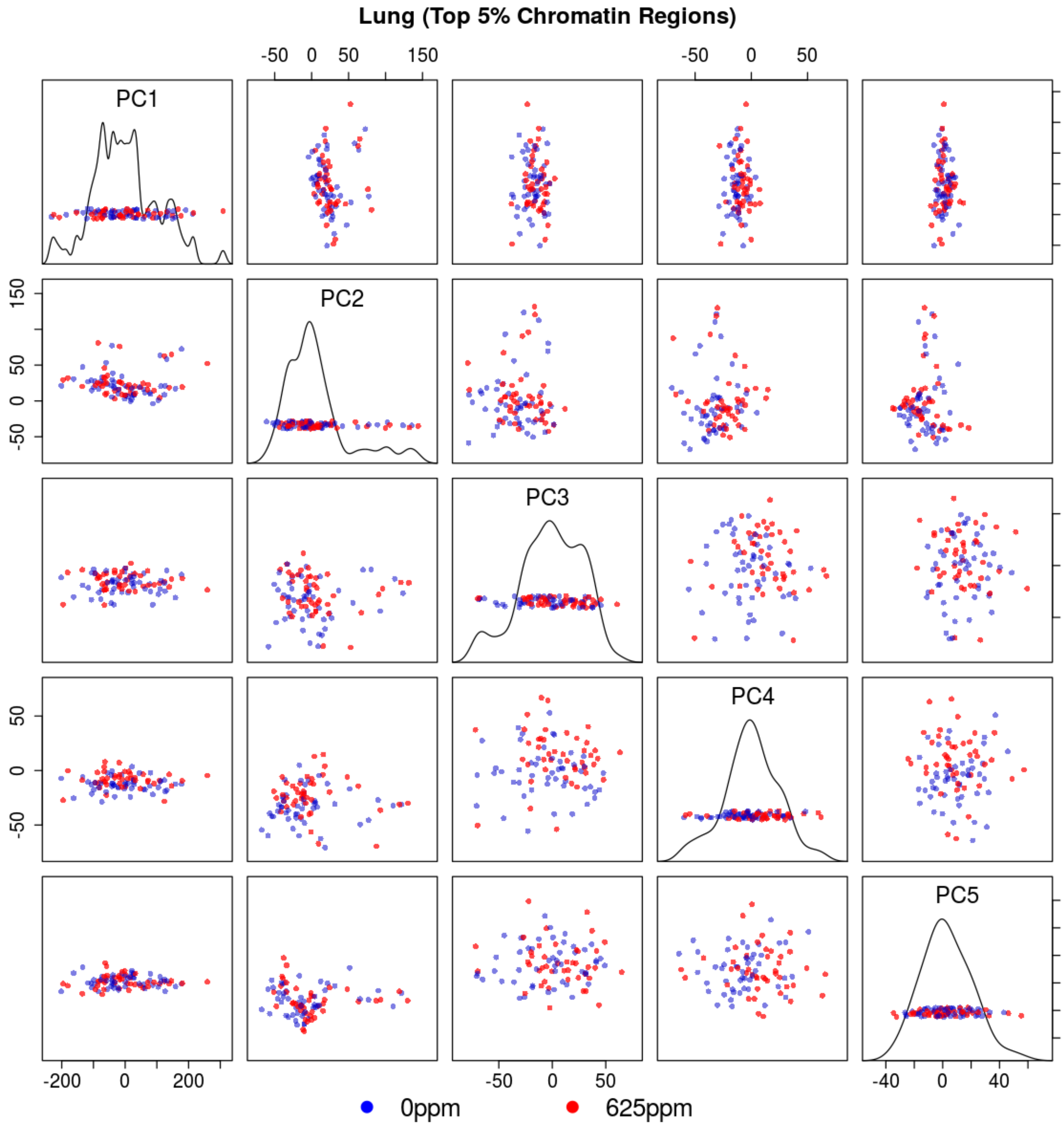

Figure 3.17. Lung top 5% chromatin windows PCA plot. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for lung samples using the top 5% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

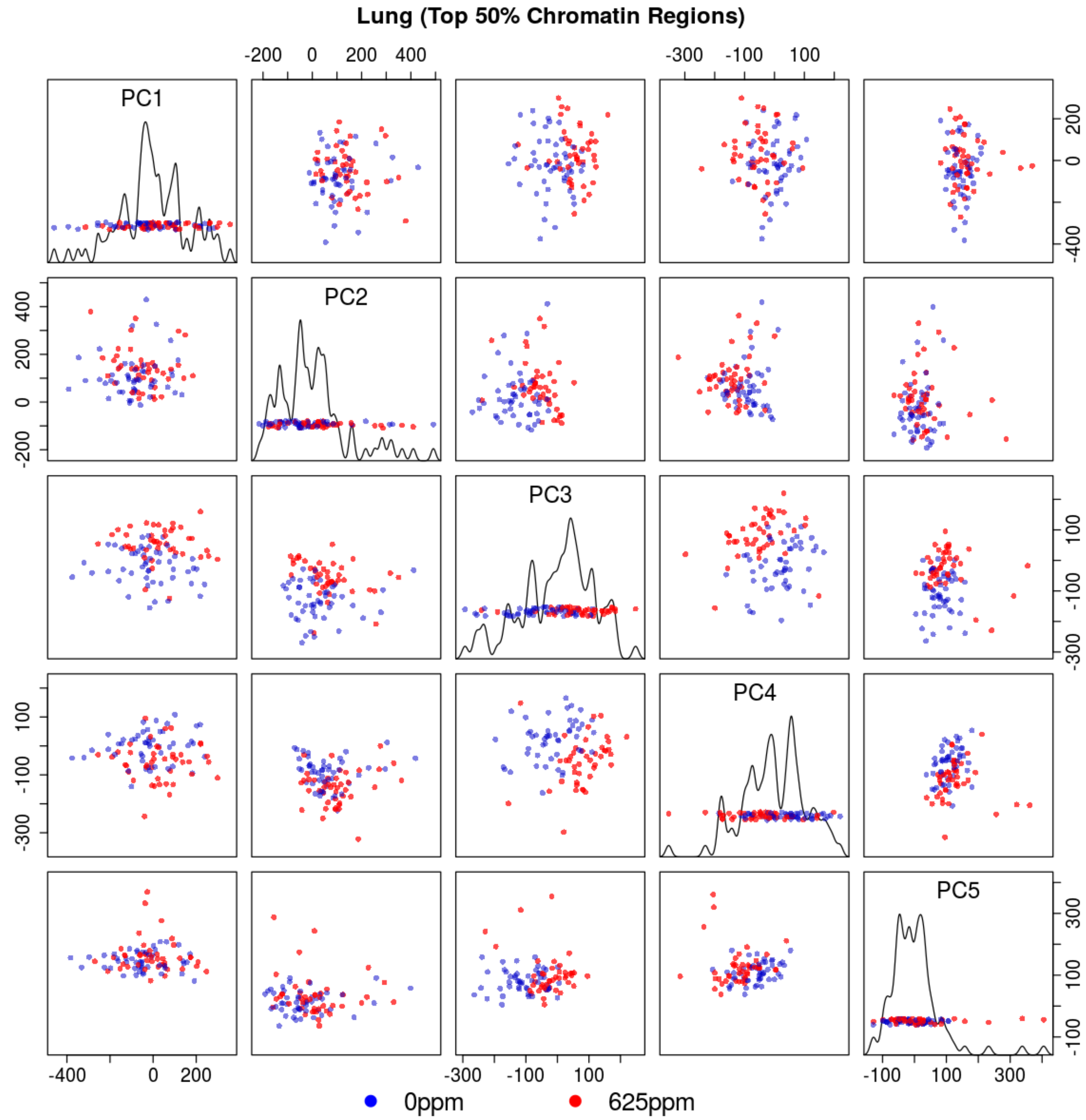

Figure 3.18. Lung top 50% chromatin windows PCA plot. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for lung samples using the top 50% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

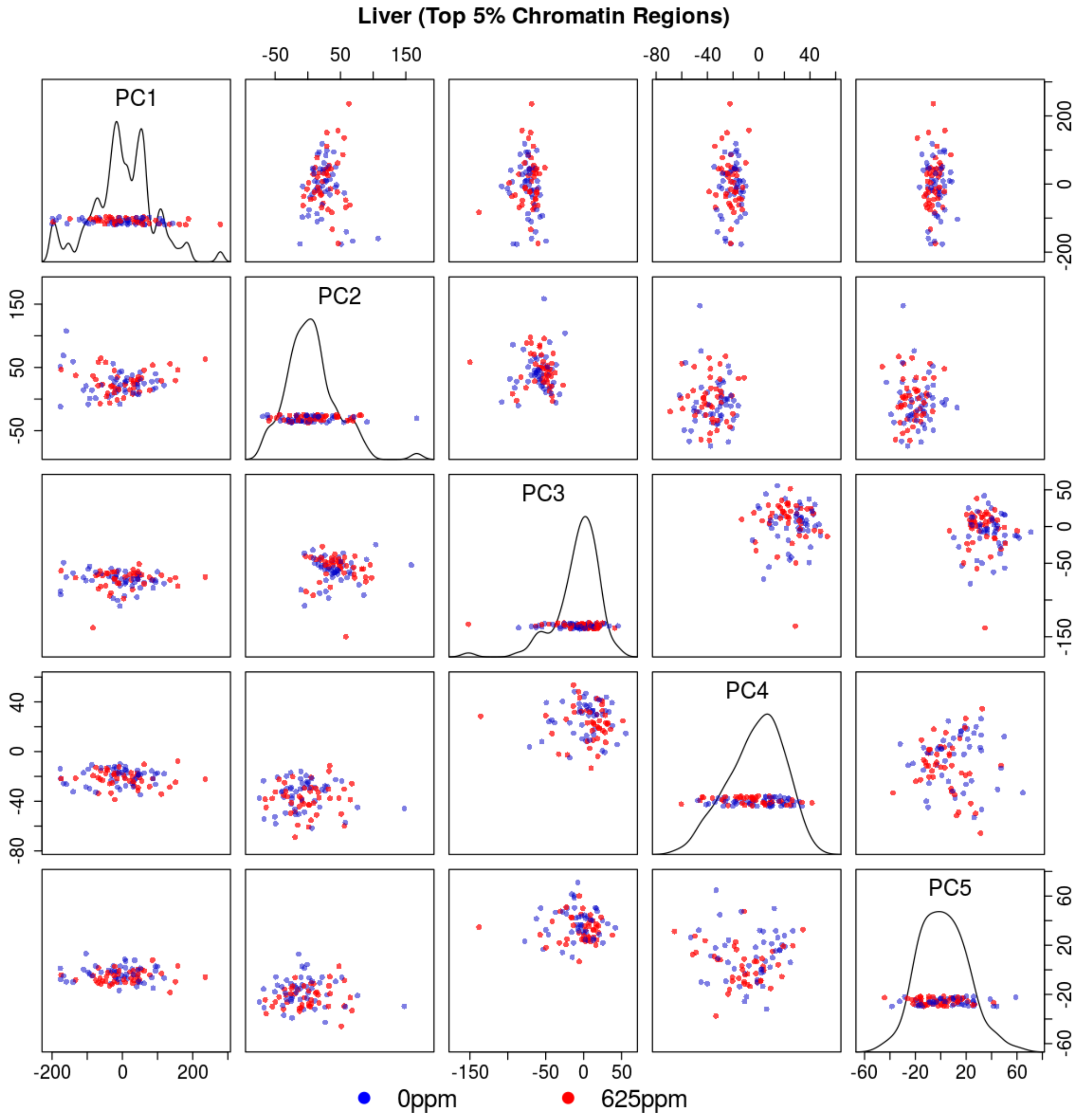

Figure 3.19. Liver top 5% chromatin windows PCA plot. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for liver samples using the top 5% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

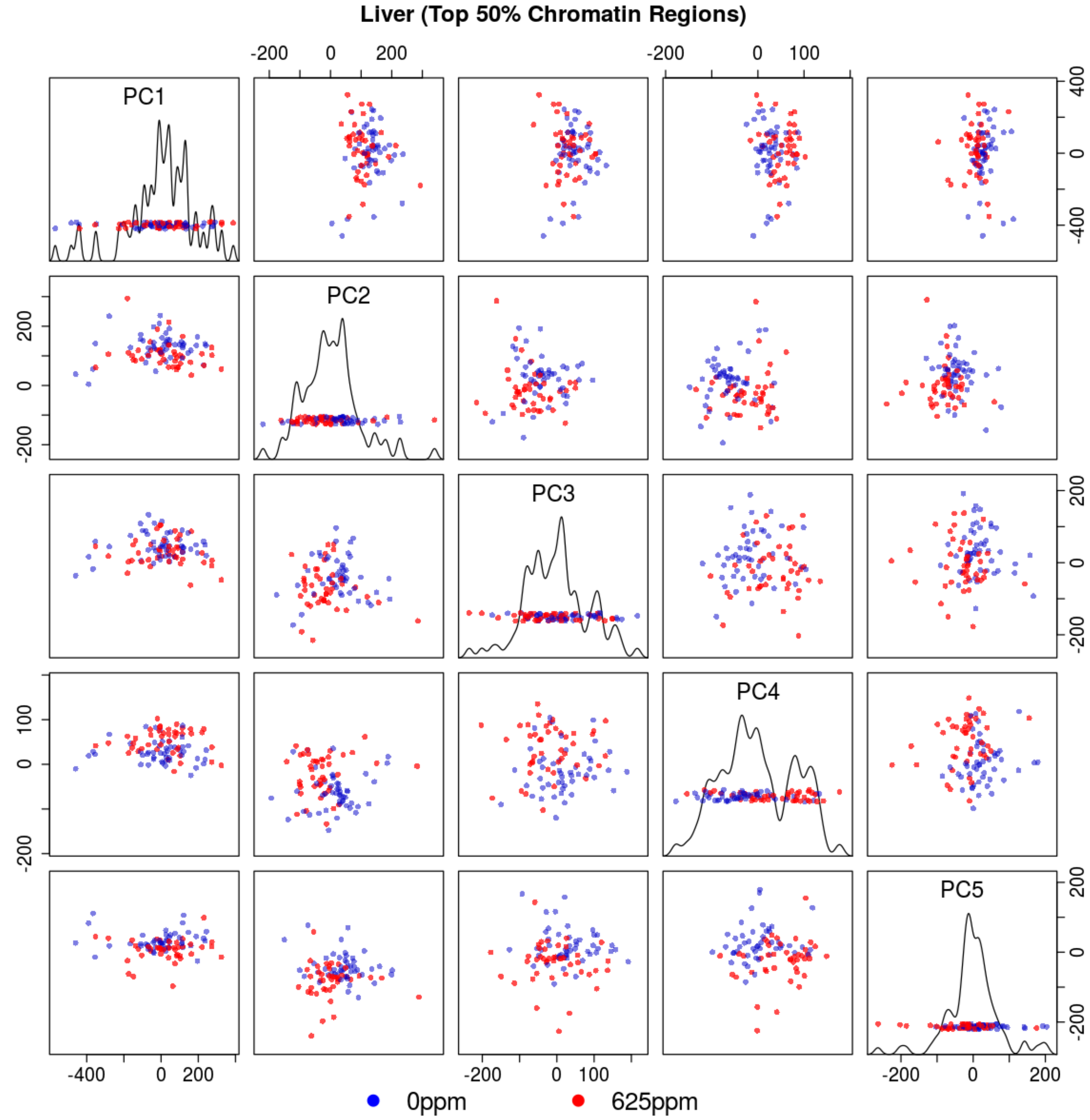

Figure 3.20. Liver top 50% chromatin windows PCA plot. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for liver samples using the top 50% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5

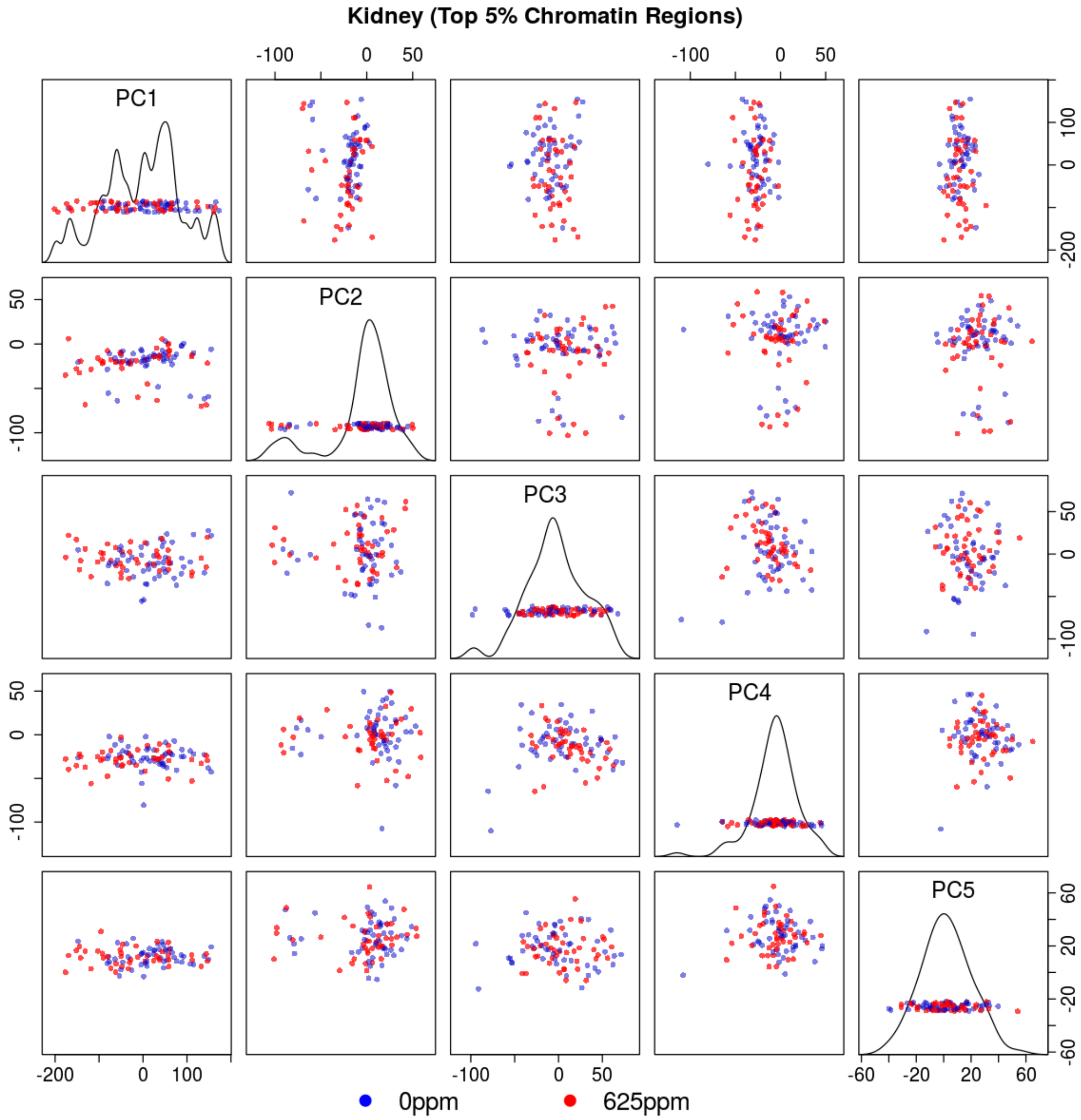

Figure 3.21. Kidney top 5% chromatin windows PCA plot. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for kidney samples using the top 5% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

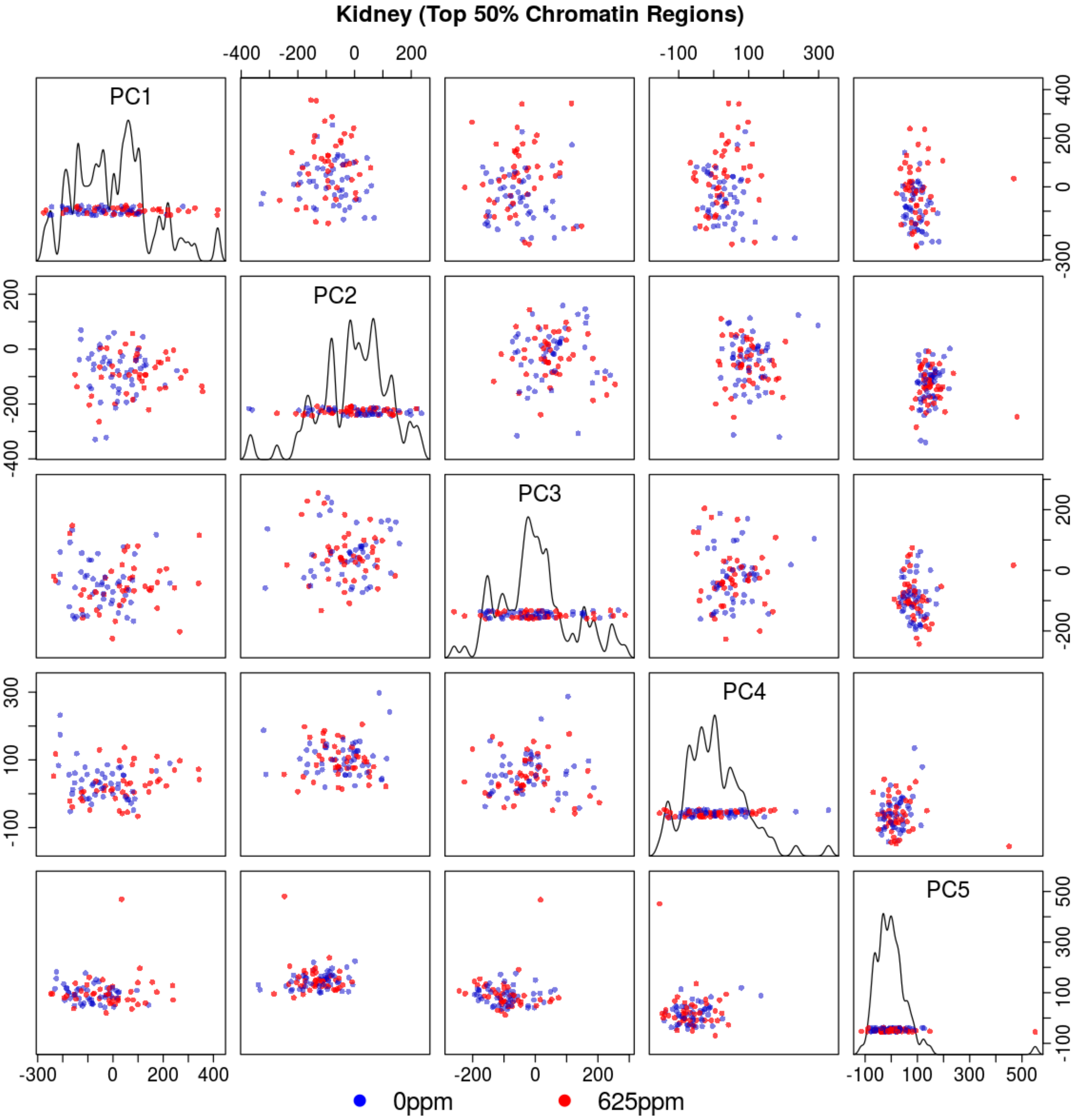

Figure 3.22. Kidney top 50% chromatin windows PCA plot. PCA plot for the top 5 PCs from PCA of chromatin accessibility profiles for kidney samples using the top 50% of chromatin windows ranked by accessibility. Diagonal subplots are kernel density estimates of sample values for a given PC. Off diagonal subplots are scatterplots of all pairwise comparisons of PCs 1-5.

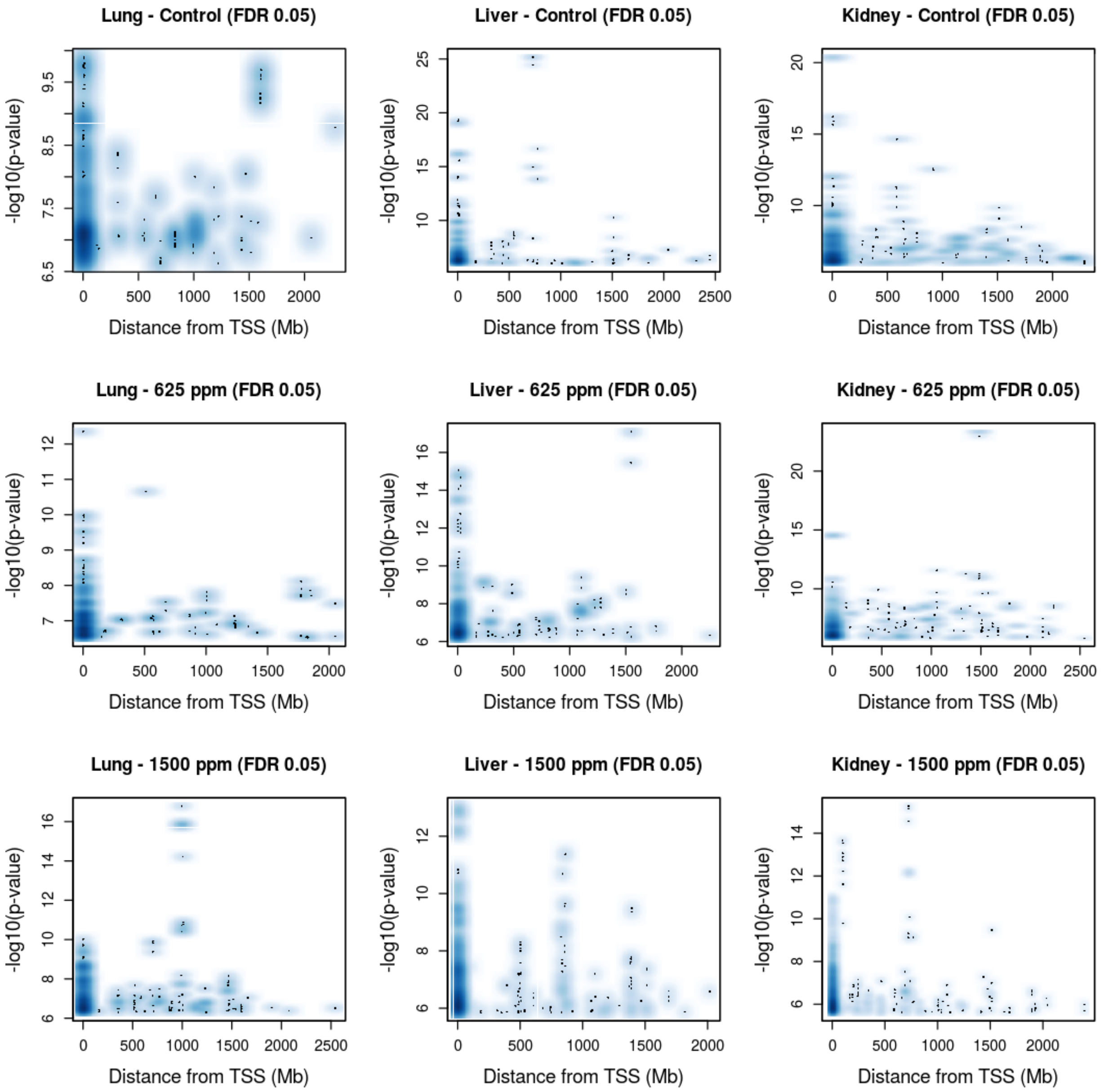

Figure 3.23. Distance of eQTL associations. Genomic distance between a gene TSS and its significantly associated haplotype segment for each detected eQTL (FDR 0.05) compared to the eQTL *p*-value. Each subplot is for eQTLs identified in a specific treatment group and tissue.

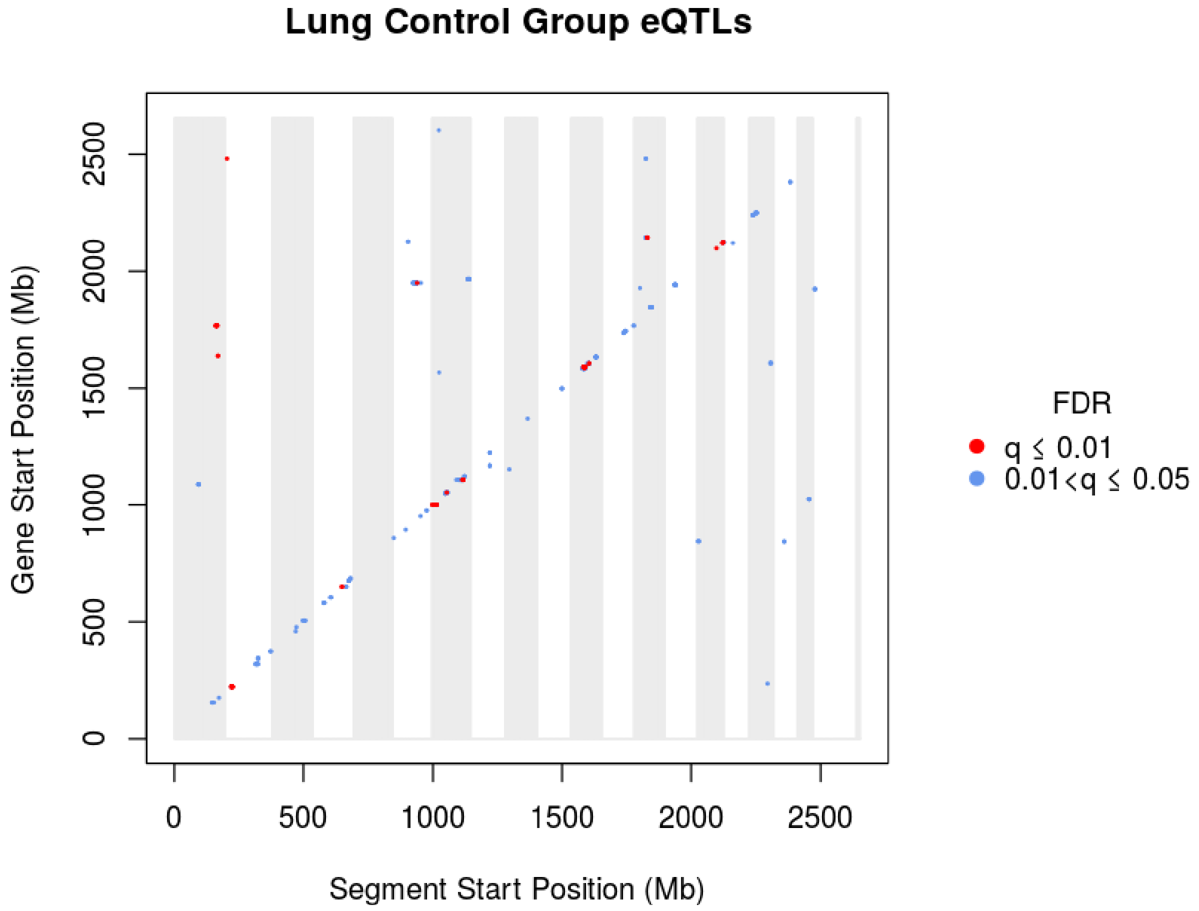

Figure 3.24. Lung control group eQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated gene for identified eQTLs (FDR 0.05)

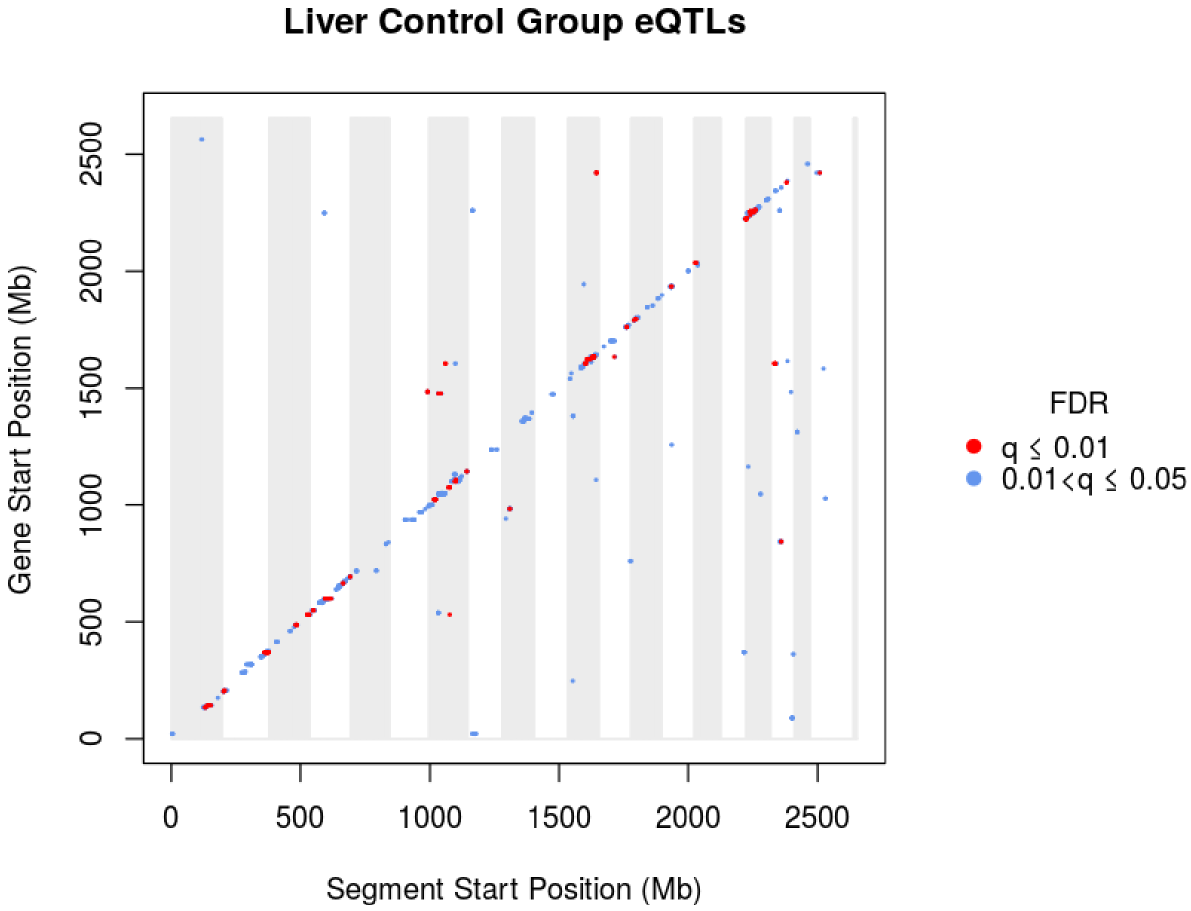

Figure 3.25. Liver control group eQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated gene for identified eQTLs (FDR 0.05)

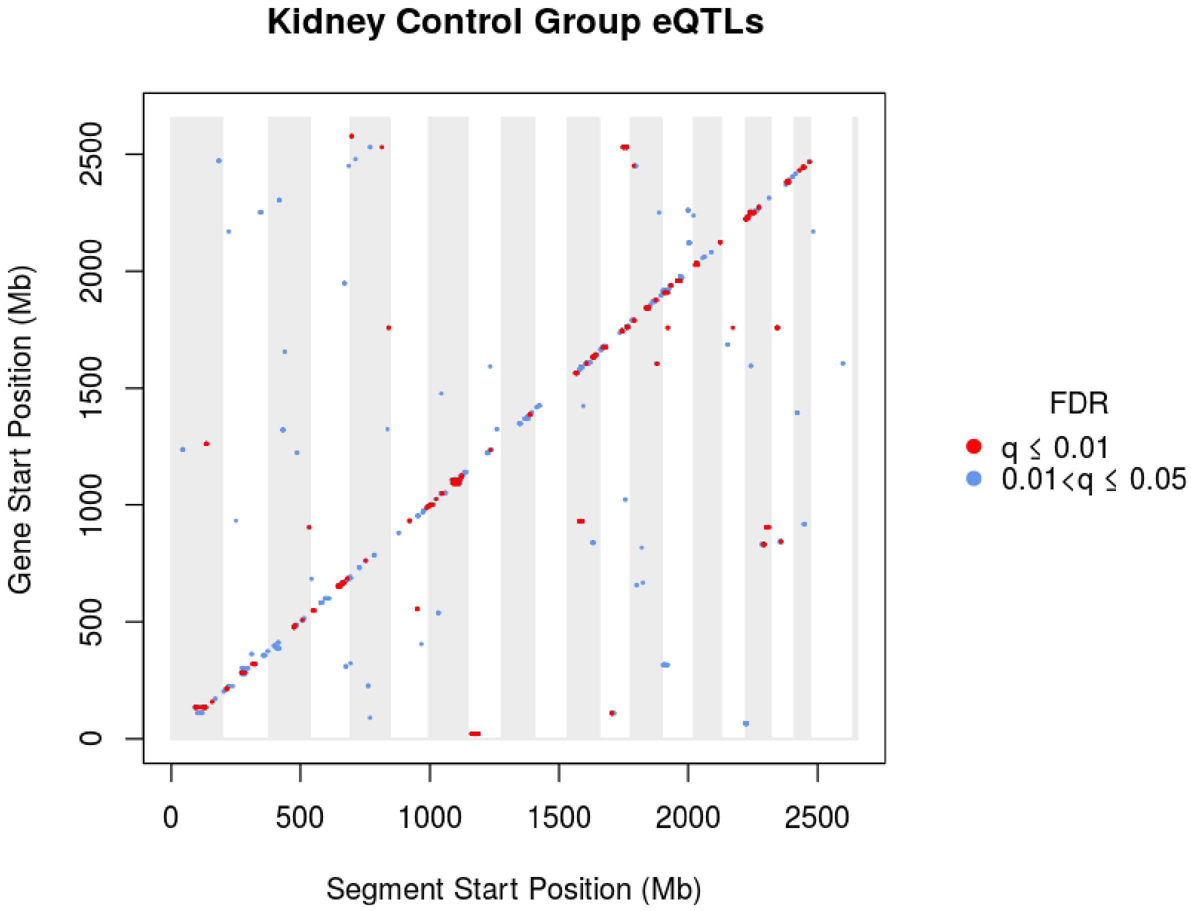

Figure 3.26. Kidney control group eQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated gene for identified eQTLs (FDR 0.05)

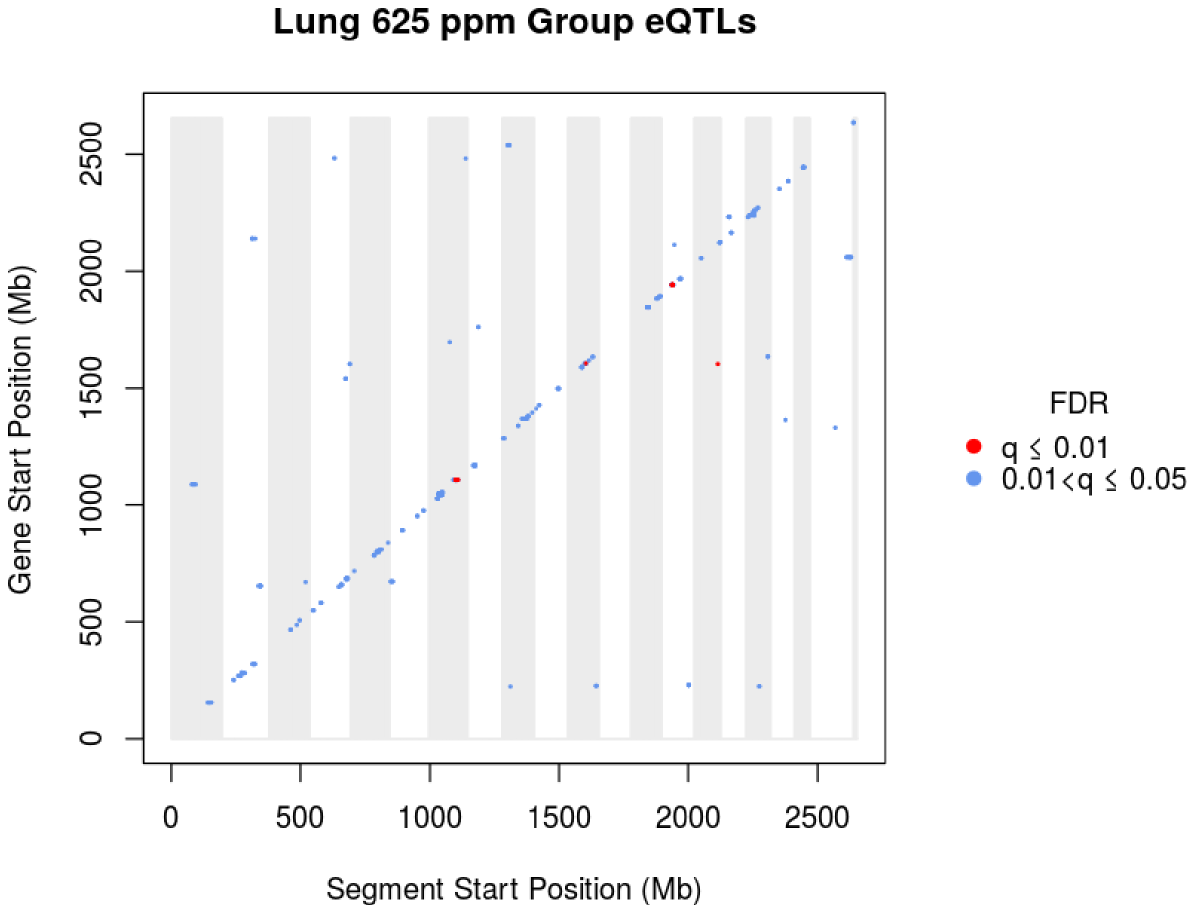

Figure 3.27. Lung 625 ppm group eQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated gene for identified eQTLs (FDR 0.05)

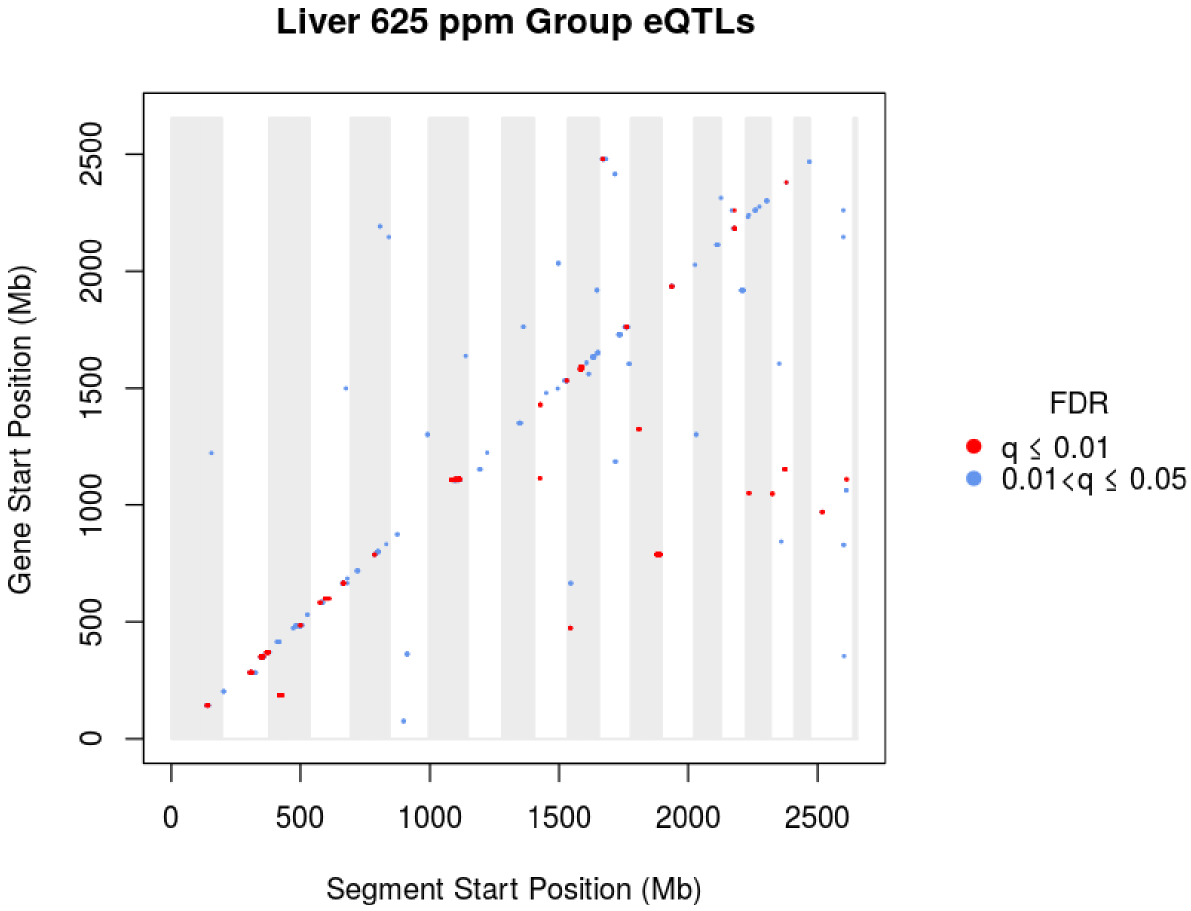

Figure 3.28. Liver 625 ppm group eQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated gene for identified eQTLs (FDR 0.05)

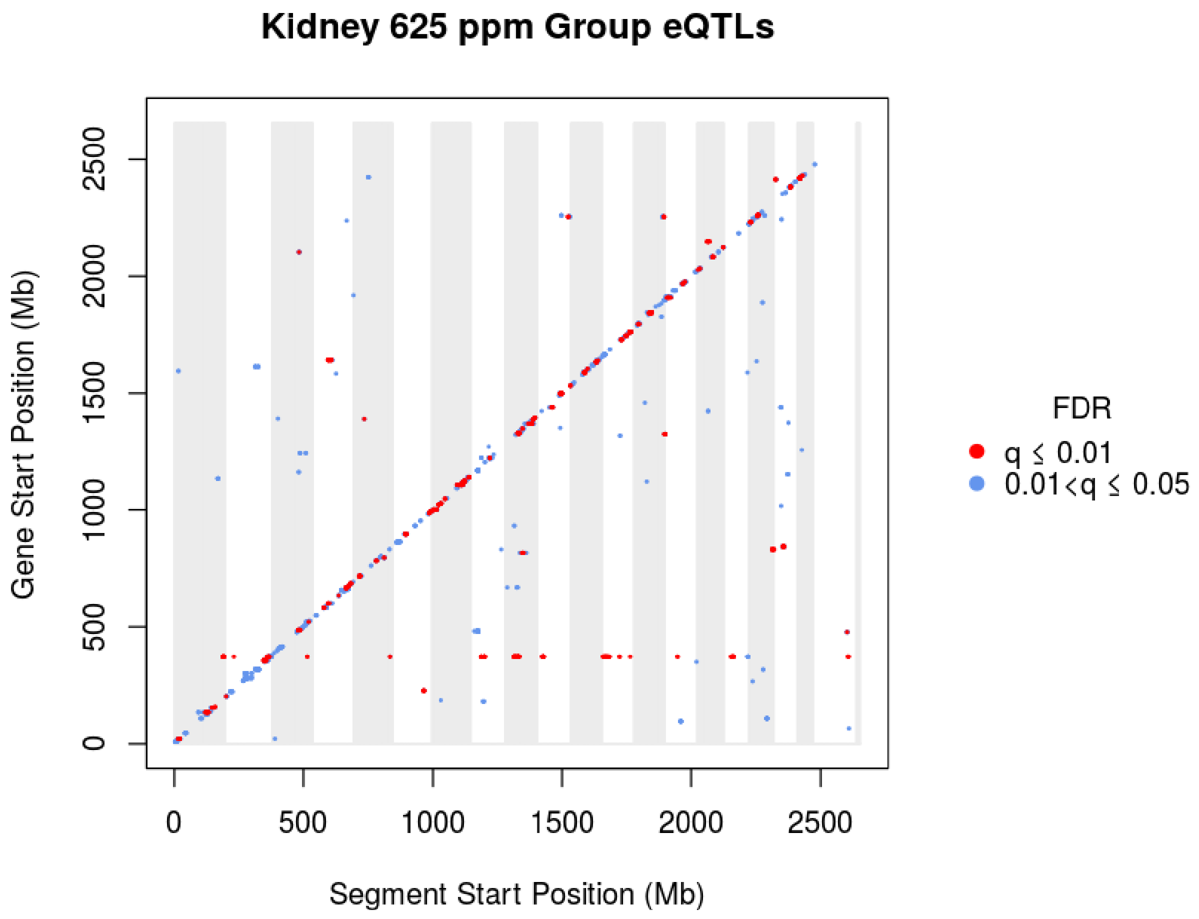

Figure 3.29. Kidney 625 ppm group eQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated gene for identified eQTLs (FDR 0.05)

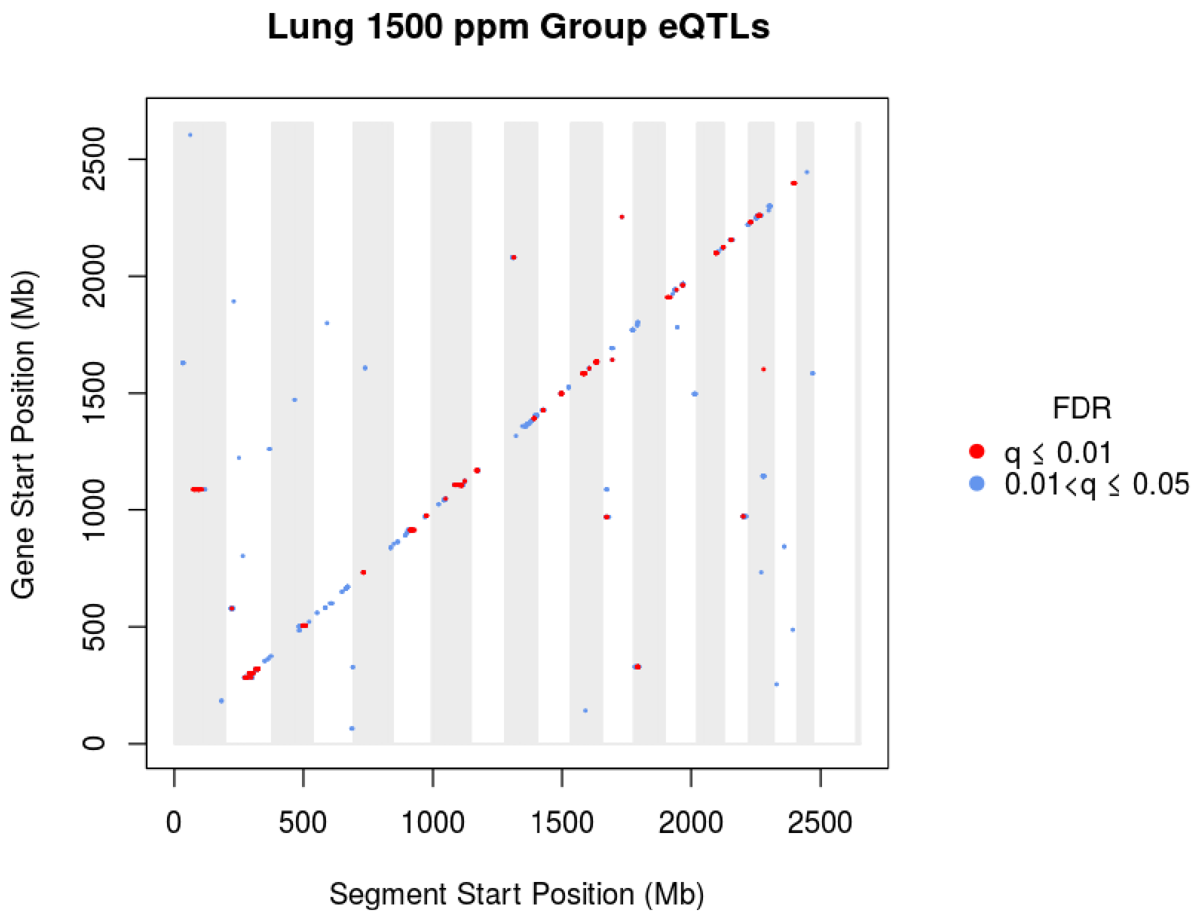

Figure 3.30. Lung 1500 ppm group eQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated gene for identified eQTLs (FDR 0.05)

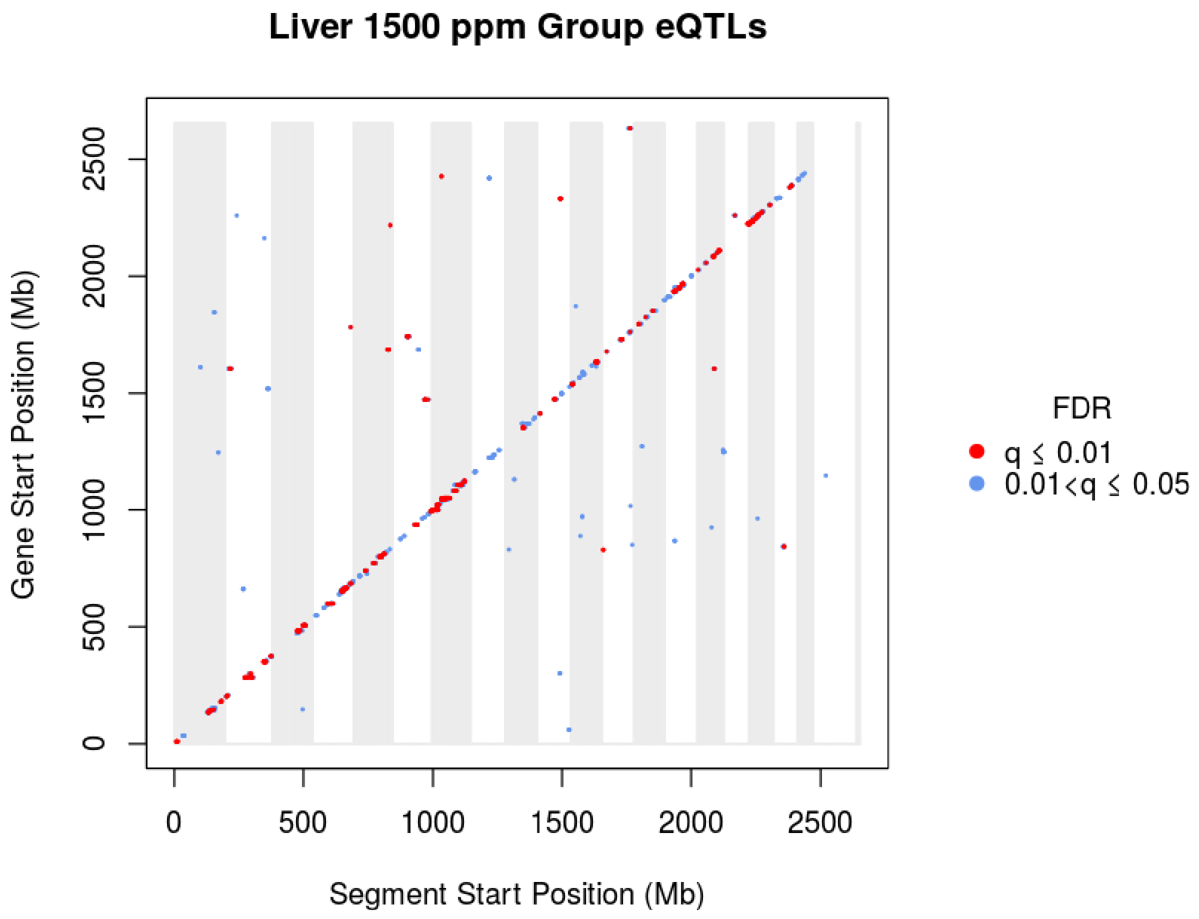

Figure 3.31. Liver 1500 ppm group eQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated gene for identified eQTLs (FDR 0.05)

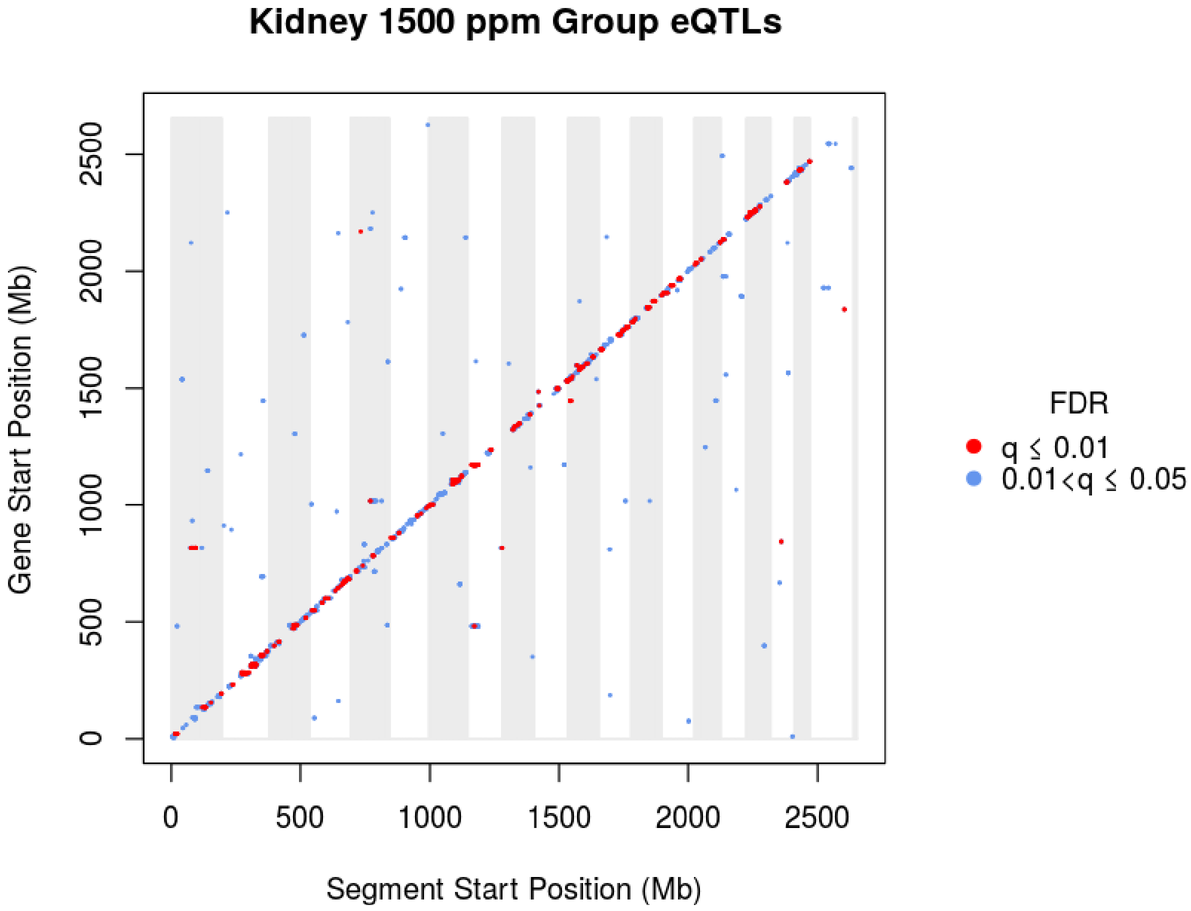

Figure 3.32. Kidney 1500 ppm group eQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated gene for identified eQTLs (FDR 0.05)

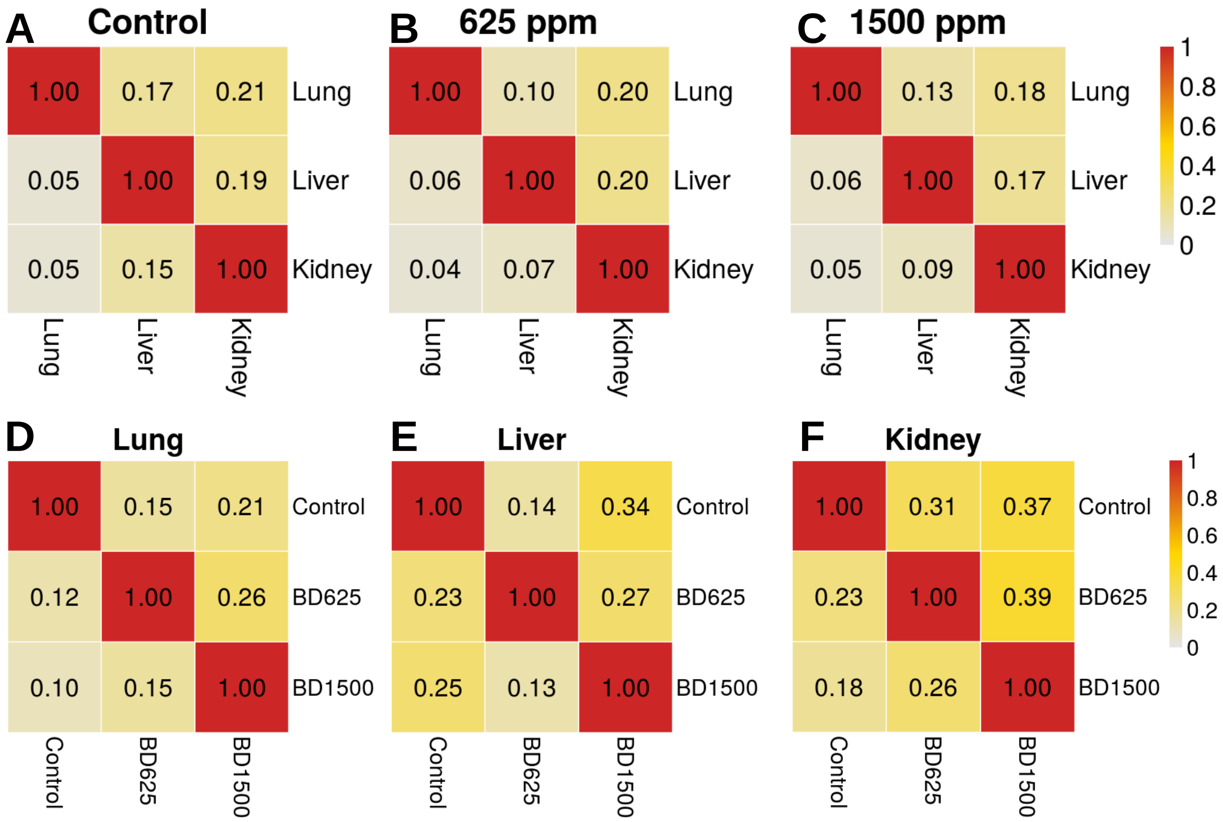

Figure 3.33. Total fraction of overlapping eQTLs. Pairwise comparisons of the fraction of significant eQTLs (FDR 0.05) that overlap A-C) between tissues within a treatment group and D-F) between treatment groups within a tissue. The row label of each matrix signifies the group used as the denominator of a given fraction in a comparison. Colors denote degree of overlap.

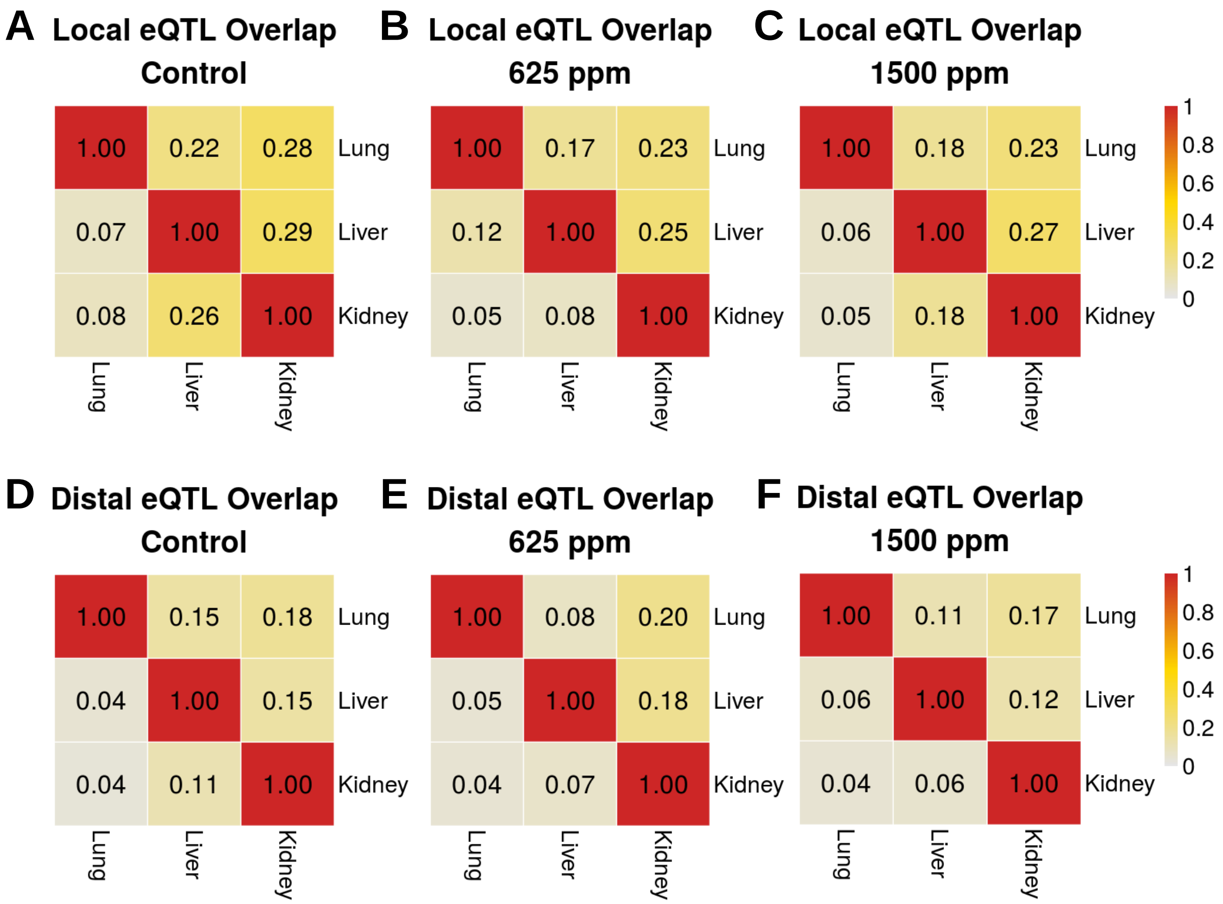

Figure 3.34. Fraction of overlapping local and distal eQTLs across tissues. Pairwise comparisons of the fraction of significant eQTLs (FDR 0.05) that overlap between tissues within a treatment group for A-C) local and D-F) distal eQTLs. Local eQTL is defined as an association less than 10 Mb in length. The row label of each matrix signifies the group used as the denominator of a given fraction in a comparison. Colors denote degree of overlap.

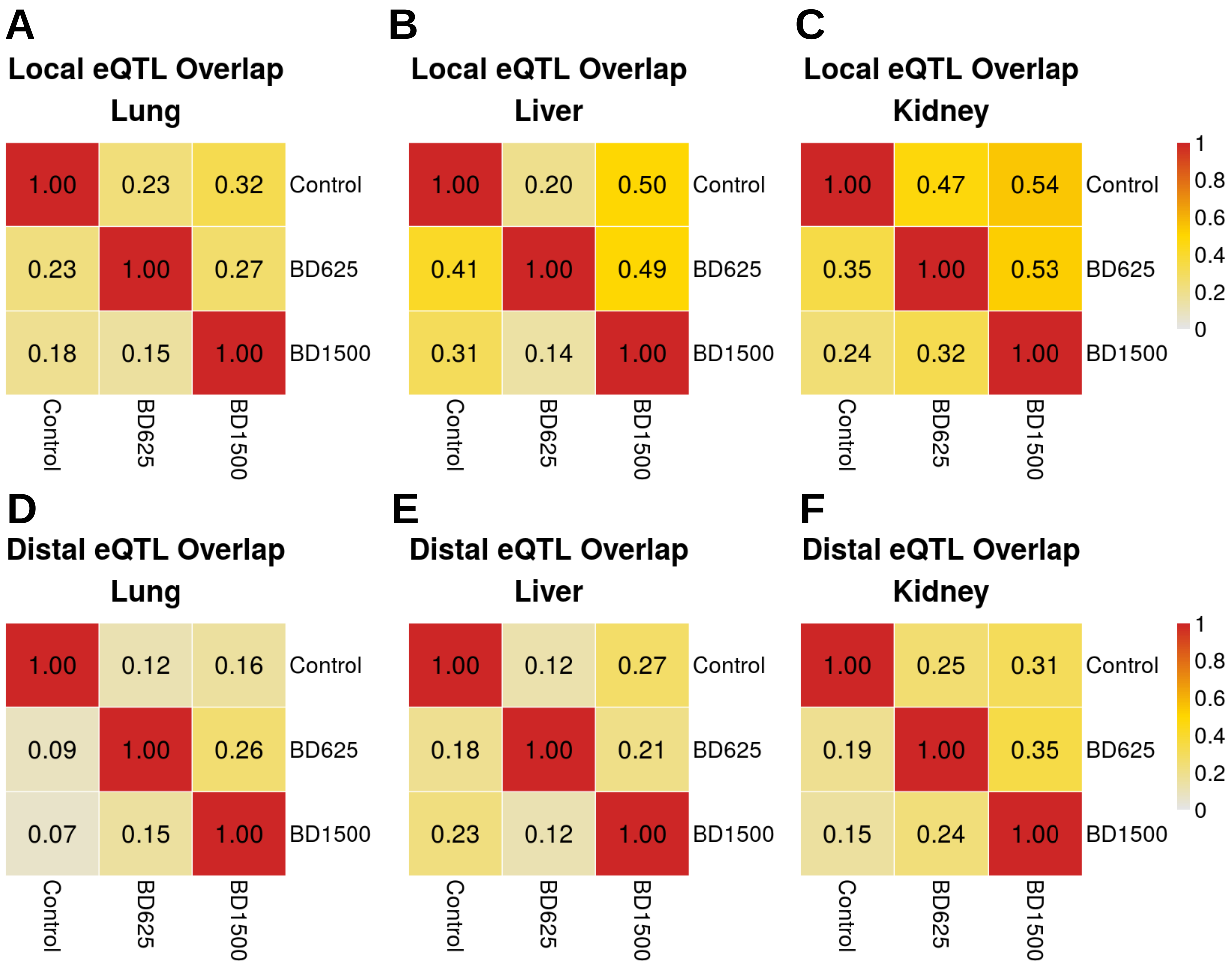

Figure 3.35. Fraction of overlapping local and distal eQTLs across treatment groups. Pairwise comparisons of the fraction of significant eQTLs (FDR 0.05) that overlap between treatment groups within a tissue for A-C) local and D-F) distal eQTLs. Local eQTL is defined as an association less than 10 Mb in length. The row label of each matrix signifies the group used as the denominator of a given fraction in a comparison. Colors denote degree of overlap.

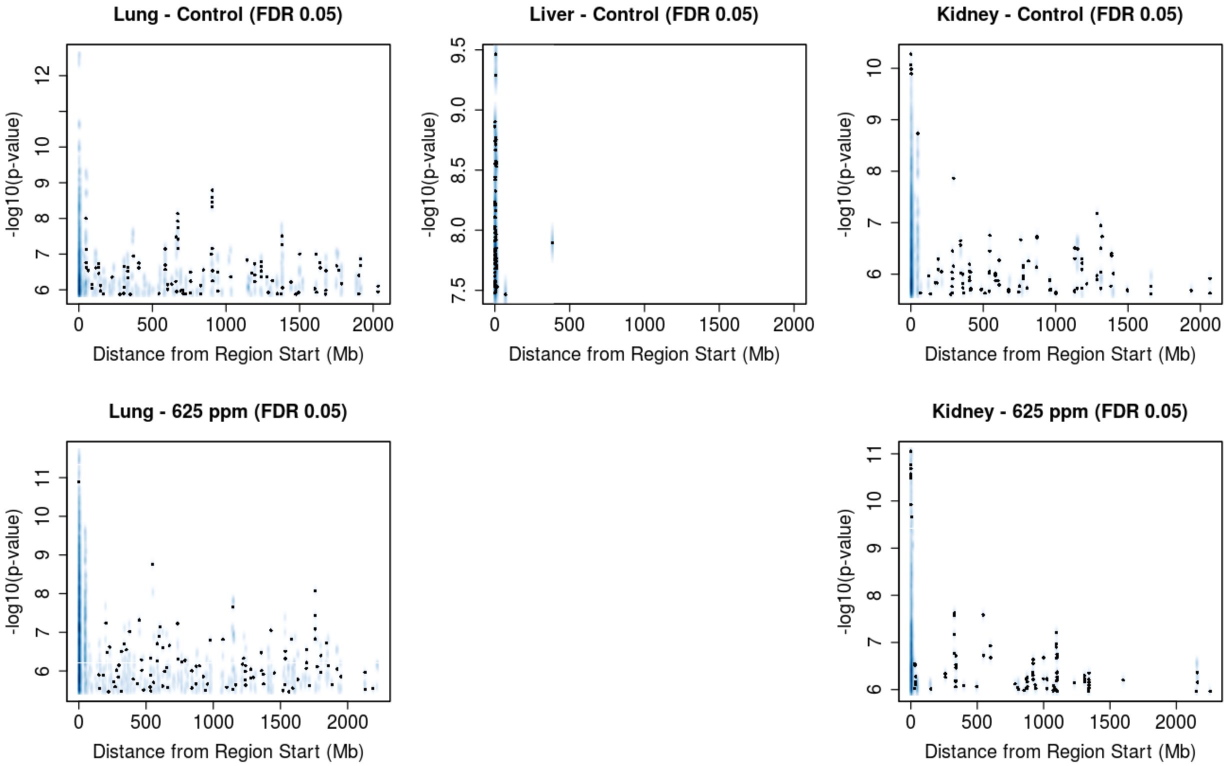

Figure 3.36. Distance of cQTL associations. Genomic distance between a chromatin window start position and its significantly associated haplotype segment for each detected cQTL (FDR 0.05) compared to the cQTL *p*-value. Each subplot is for cQTLs identified in that treatment group and tissue. The liver 625 ppm group is not shown due to no significant cQTLs discovered.

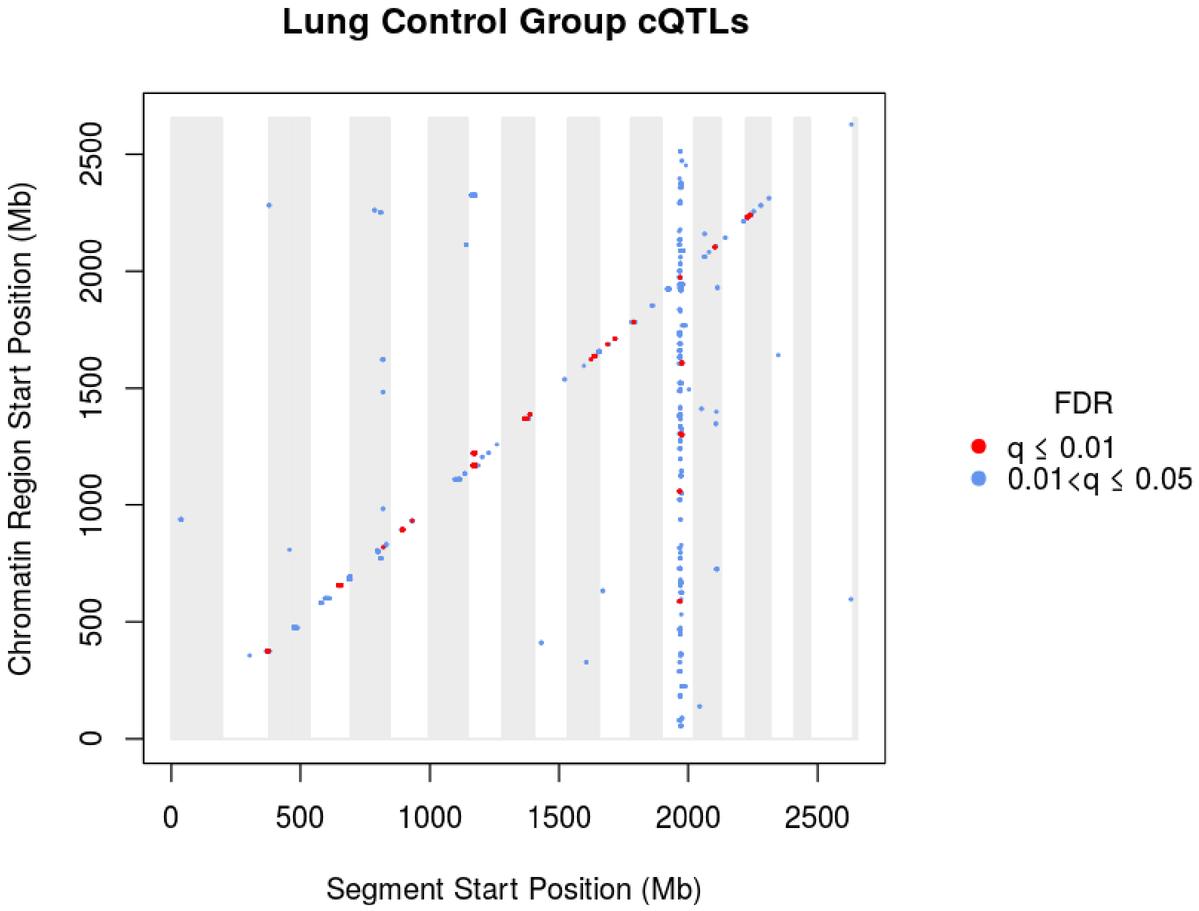

Figure 3.37. Lung control group cQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated chromatin window for identified cQTLs (FDR 0.05).

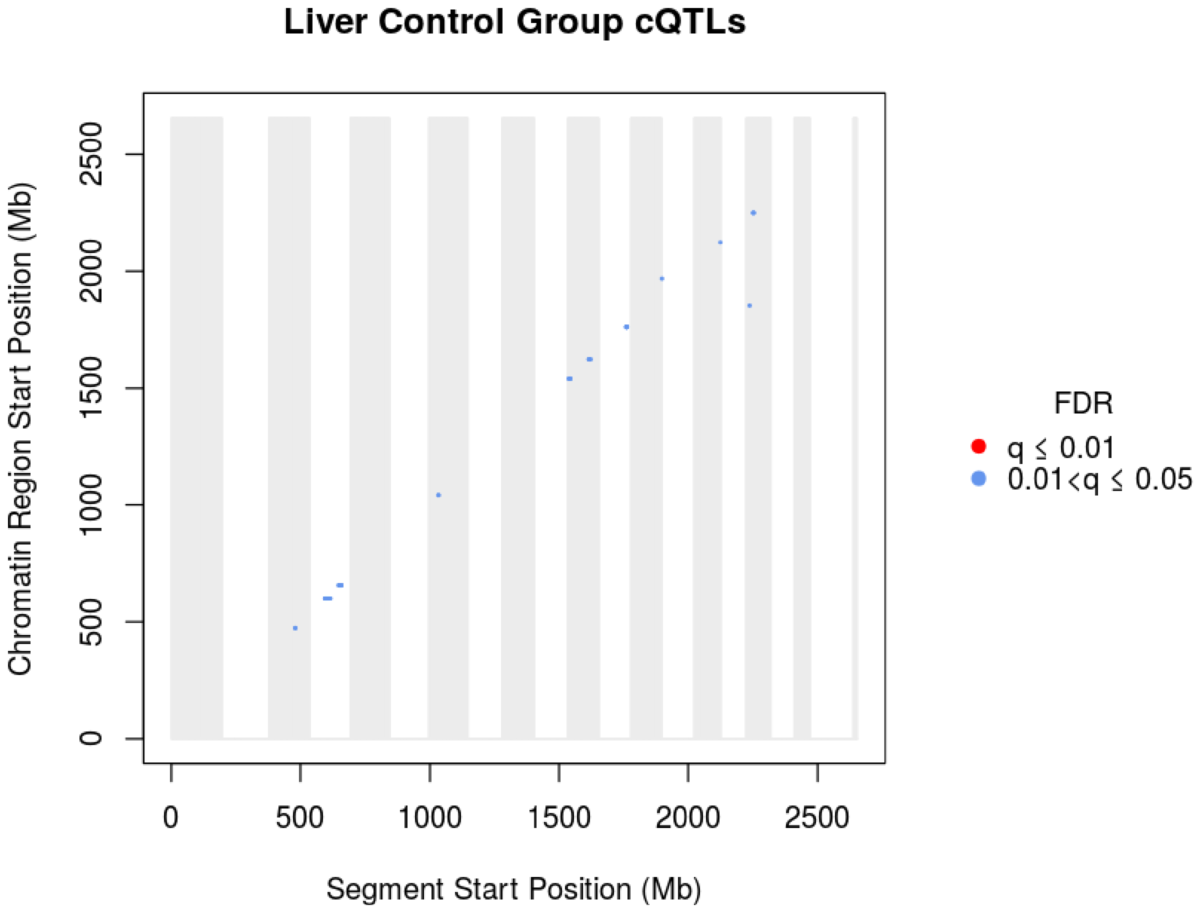

Figure 3.38. Liver control group cQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated chromatin window for identified cQTLs (FDR 0.05).

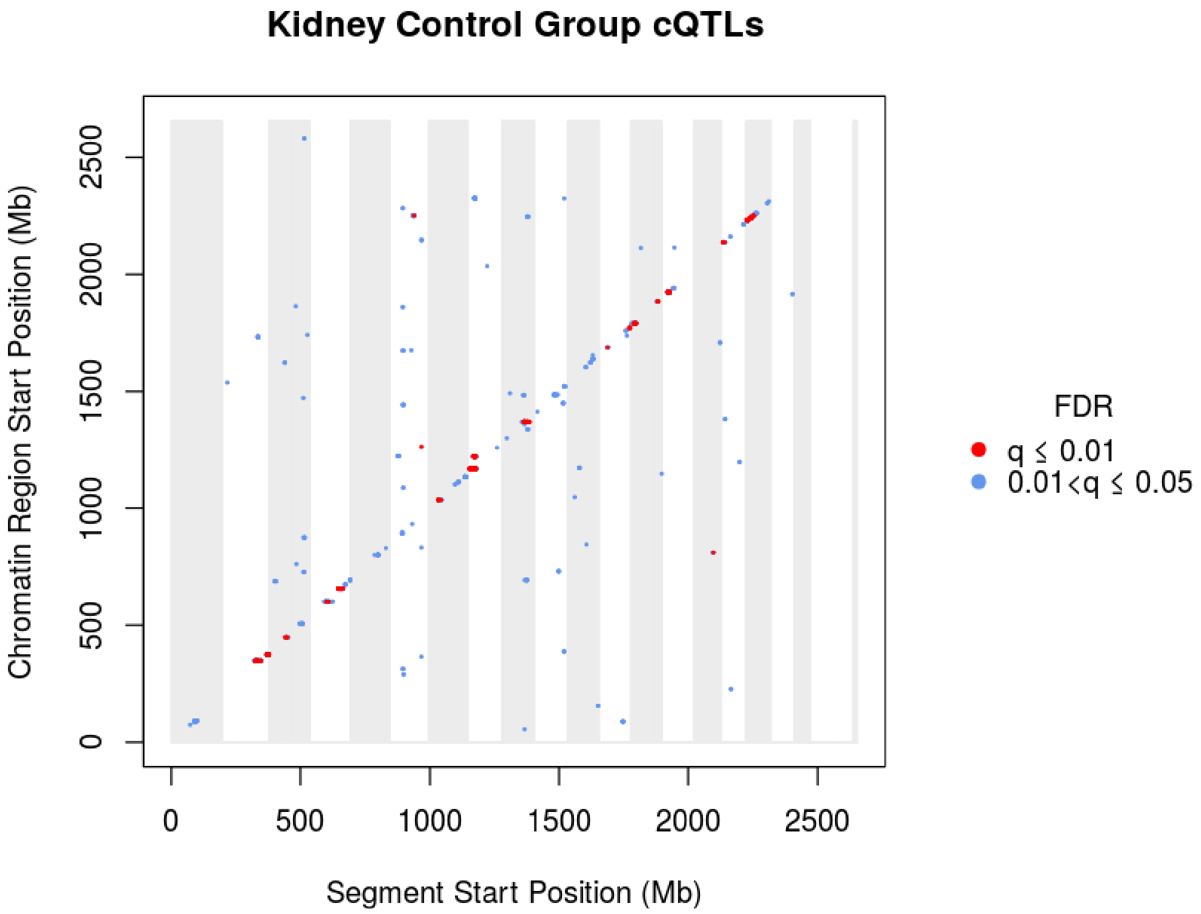

Figure 3.39. Kidney control group cQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated chromatin window for identified cQTLs (FDR 0.05).

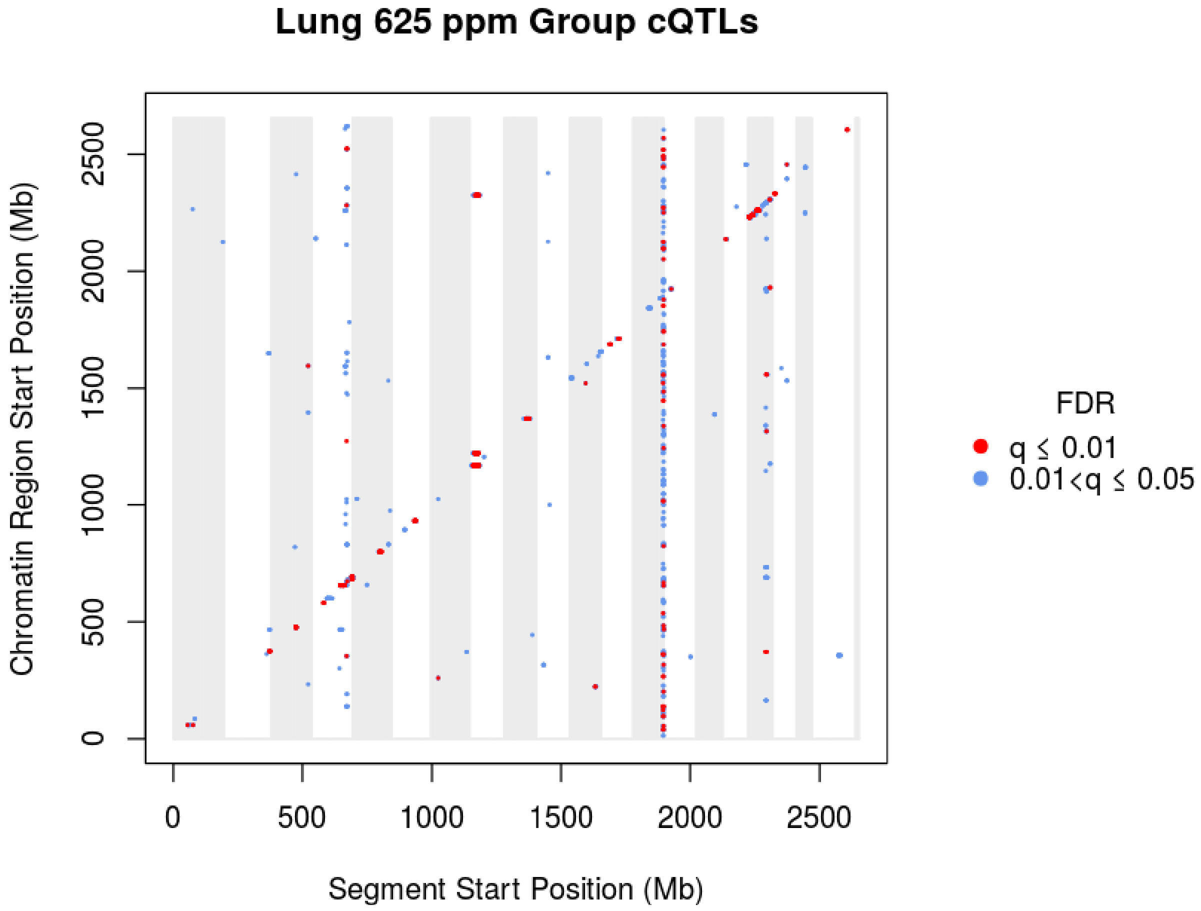

Figure 3.40. Lung 625 ppm group cQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated chromatin window for identified cQTLs (FDR 0.05).

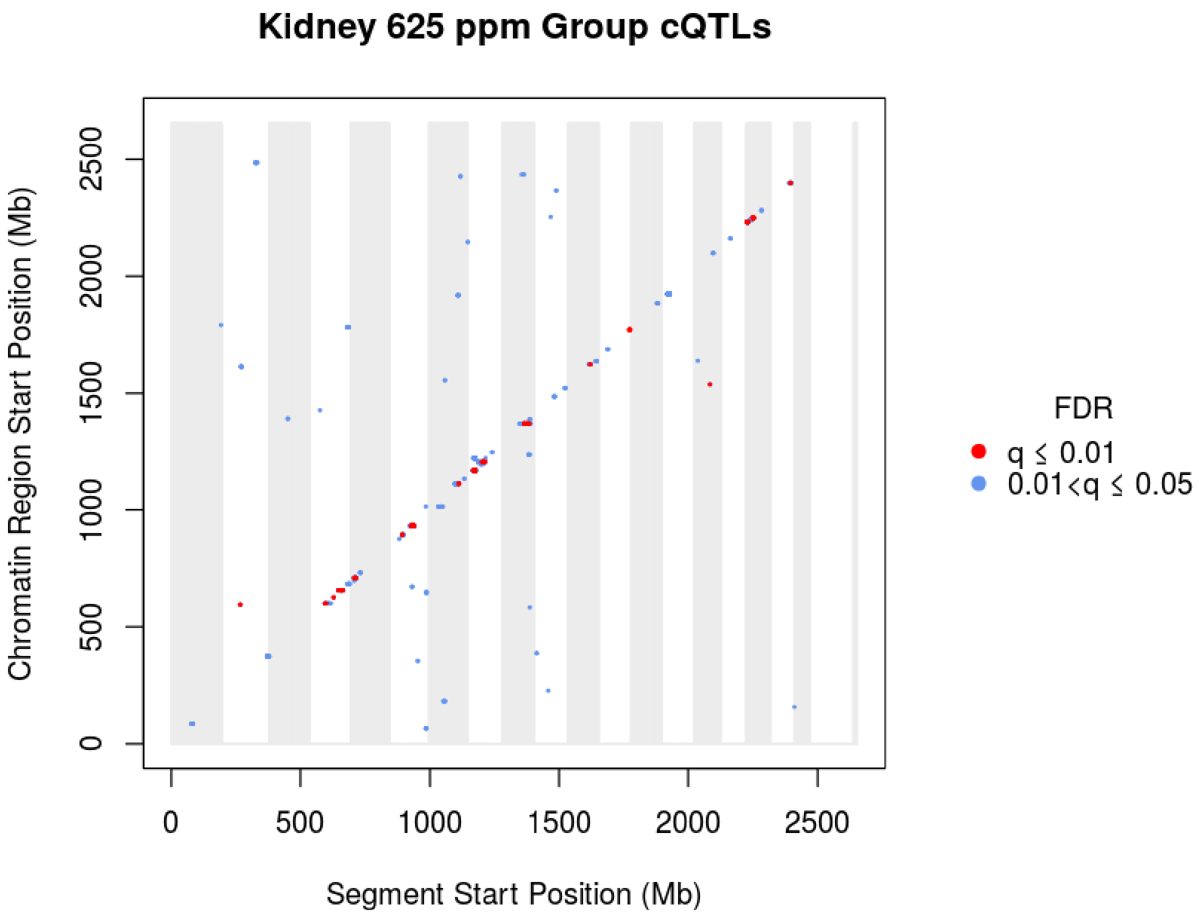

Figure 3.41. Kidney 625 ppm group cQTL map. Genome-wide scatterplot comparing the location of a segment to the position of its significantly associated chromatin window for identified cQTLs (FDR 0.05).

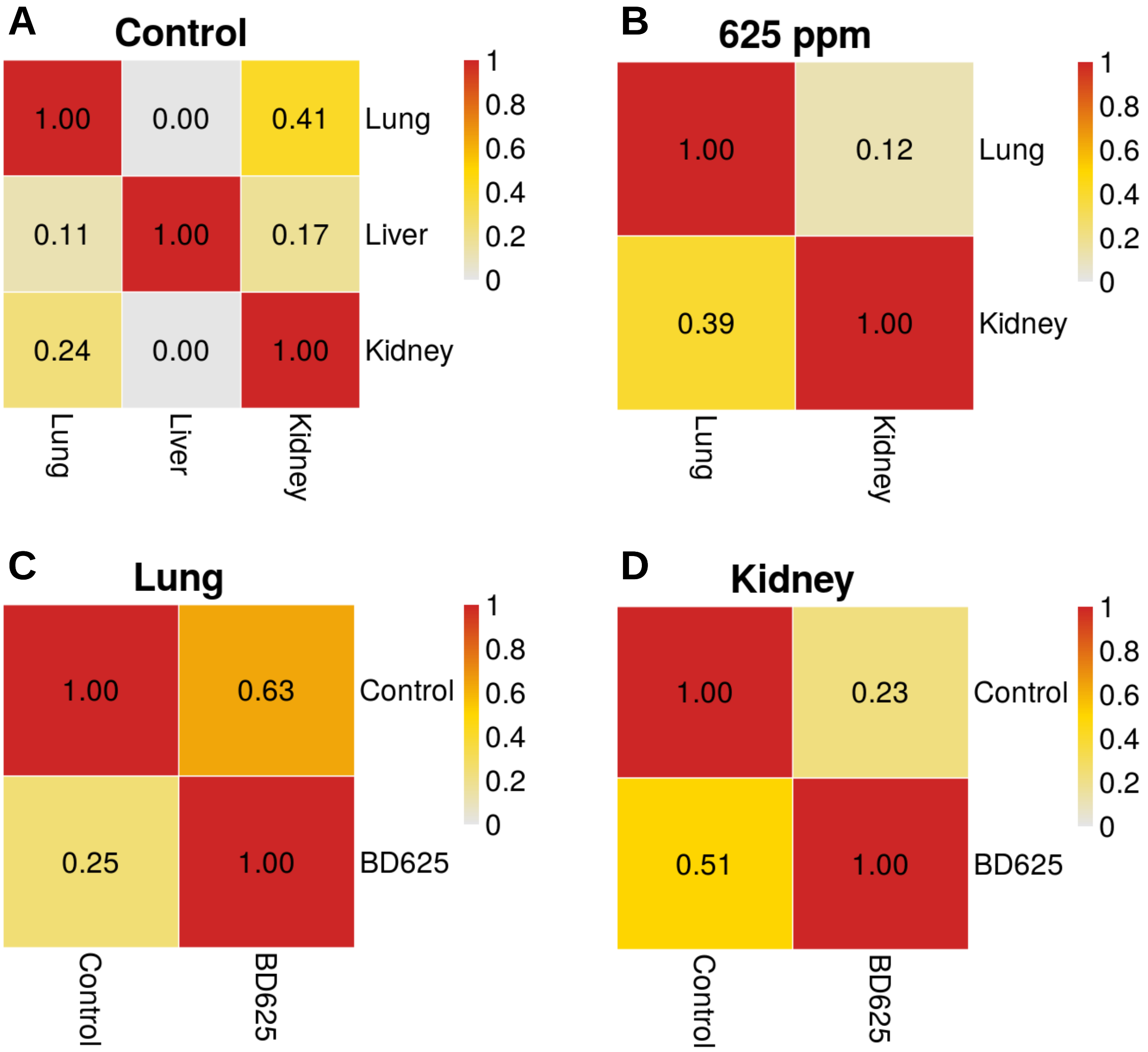

Figure 3.42. Total fraction of overlapping cQTLs. Pairwise comparisons of the fraction of significant cQTLs (FDR 0.05) that overlap A,B) between tissues within a treatment group and C,D) between treatment groups within a tissue. The row label of each matrix signifies the group used as the denominator of a given fraction in a comparison. Colors denote degree of overlap.

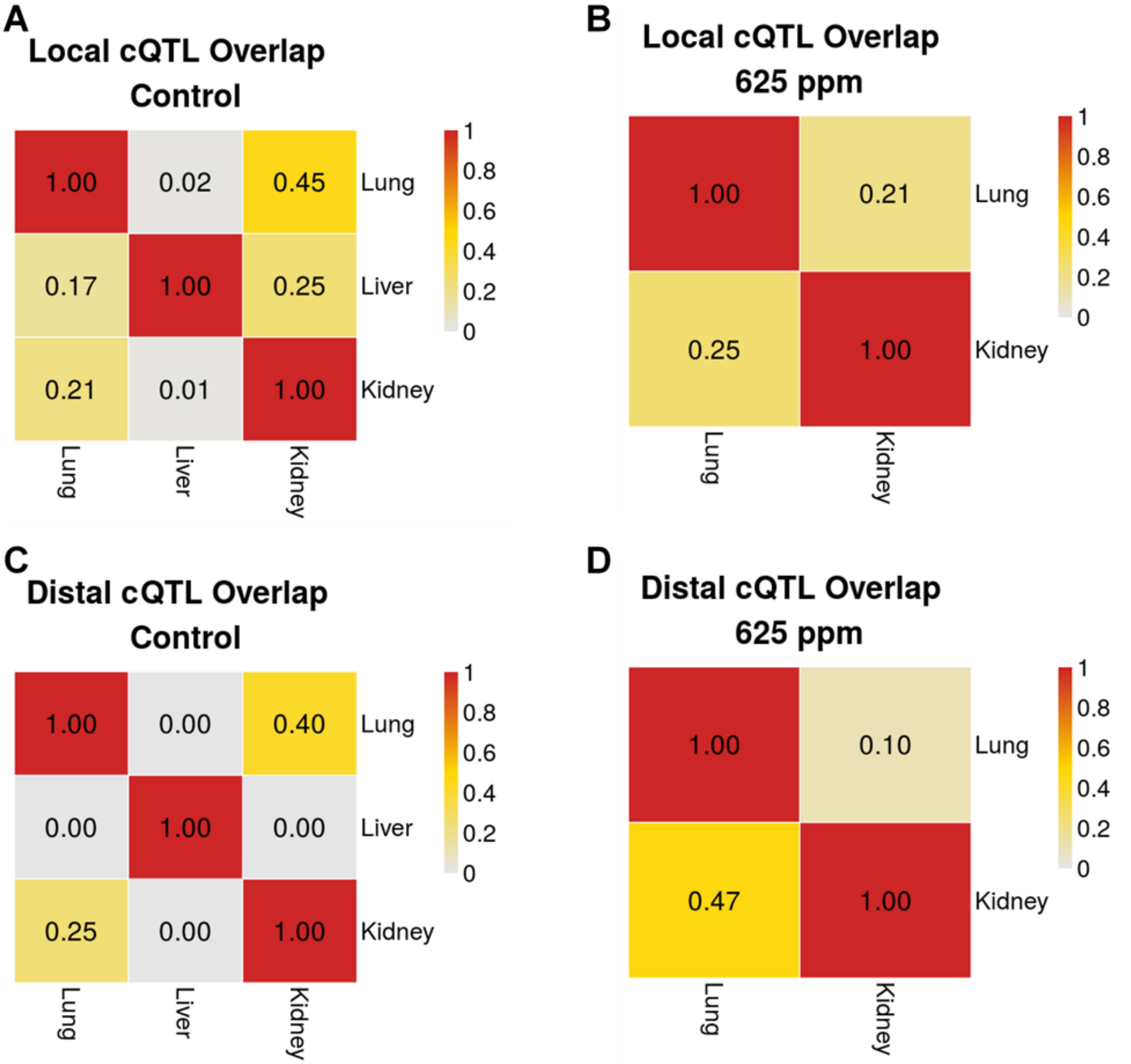

Figure 3.43. Fraction of overlapping local and distal cQTLs across tissues. Pairwise comparisons of the fraction of significant cQTLs (FDR 0.05) that overlap between tissues within a treatment group for A,B) local and C,D) distal associations. Local associations are defined as being less than 10 Mb in length. The row label of each matrix signifies the group used as the denominator of a given fraction in a comparison. Colors denote degree of overlap.

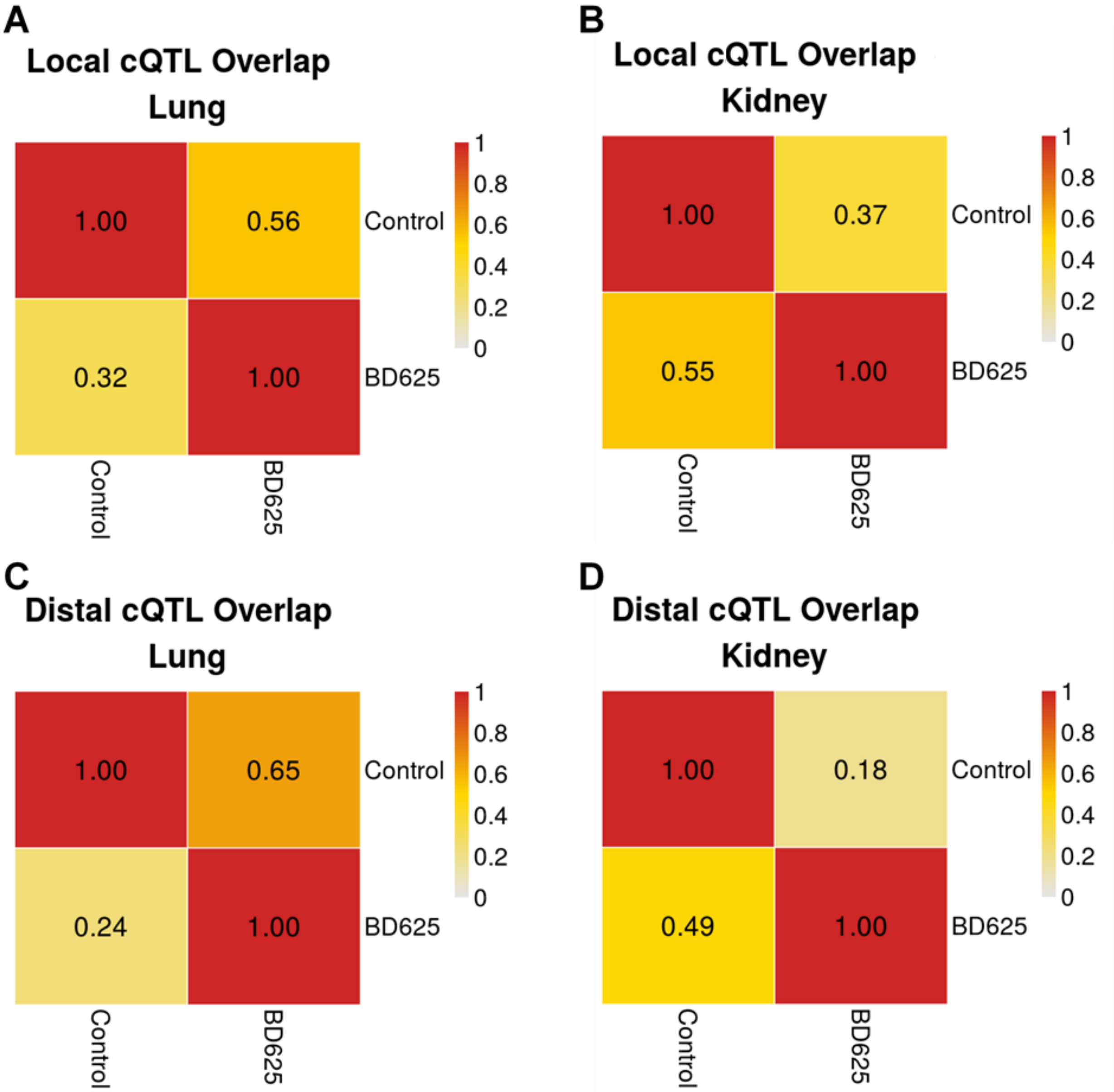

Figure 3.44. Fraction of overlapping local and distal cQTLs across treatment groups. Pairwise comparisons of the fraction of significant cQTLs (FDR 0.05) that overlap between treatment groups within a tissue for A,B) local and C,D) distal associations. Local associations are defined as being less than 10 Mb in length. The row label of each matrix signifies the group used as the denominator of a given fraction in a comparison. Colors denote degree of overlap.

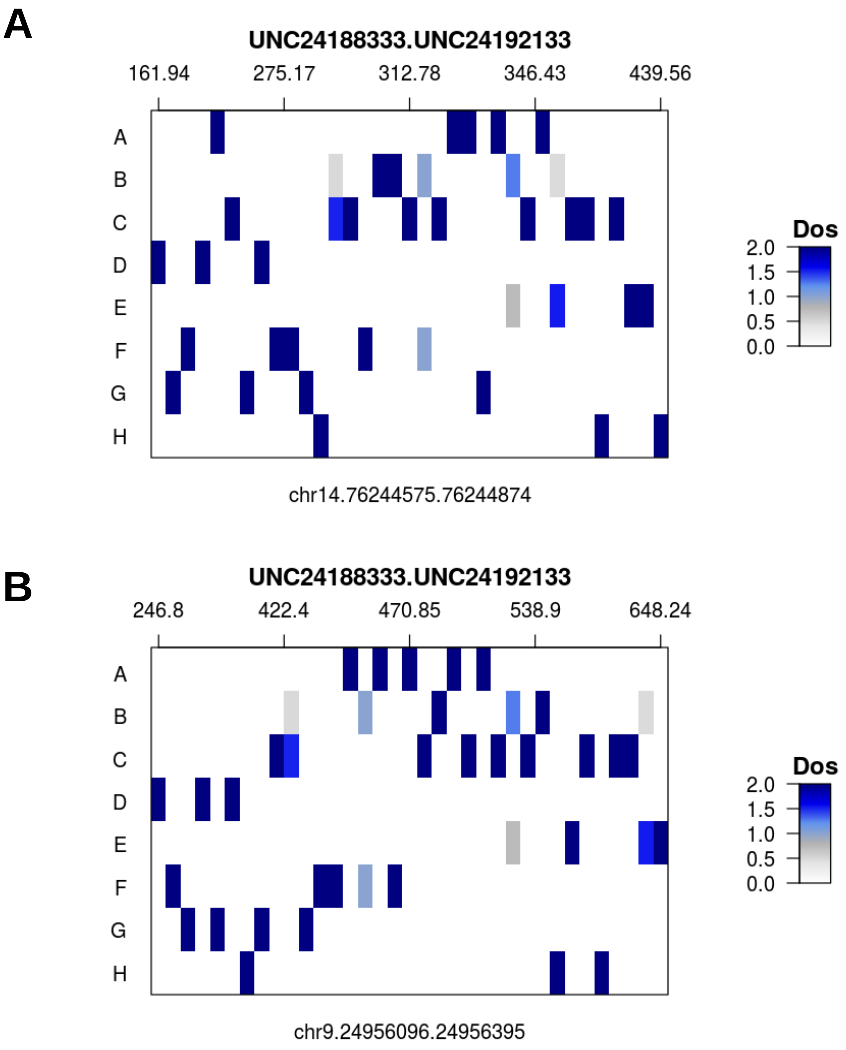

Figure 3.45. Lung control group hotspot haplotype dosages. Haplotype dosages at segment UNC24188333.UNC24192133 (chr14:69792831-70034137) for each mouse arranged by chromatin accessibility phenotype values at chromatin windows A) chr14:76244575-76244874 and B) chr9:24956096-24956395. Each column corresponds to the phenotype value for a mouse, and each row represents the founder strain dosage for a given mouse. Residual heterozygosity is observed in some mice for this haplotype segment. Haplotype codes correspond to A/J (A), C57BL/6J (B), 129S1/SvImJ (C), NOD/LtJ (D), NZO/HILtJ (E), CAST/EiJ (F), PWK/PhJ (G), and WSB/EiJ (H).

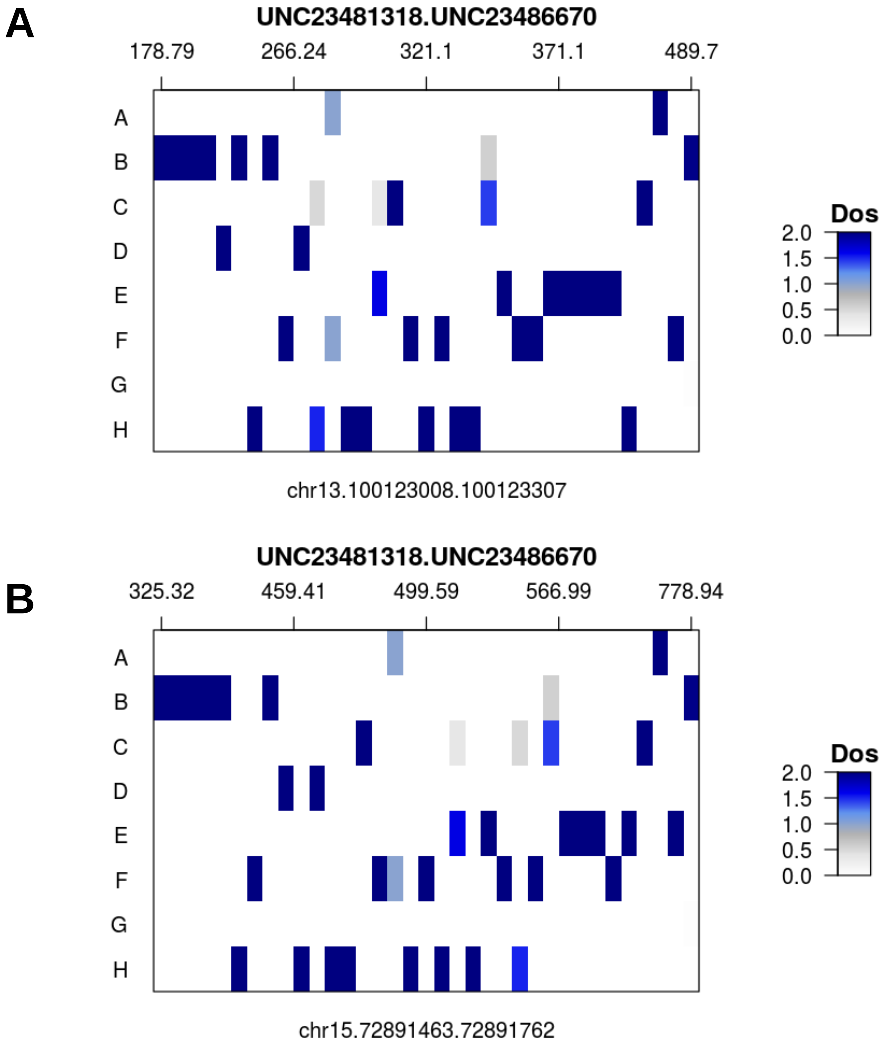

Figure 3.46. Lung 625 ppm group hotspot haplotype dosages. Haplotype dosages at segment UNC23481318.UNC23486670 (chr13:118864010-119236105) for each mouse arranged by chromatin accessibility phenotype values at chromatin windows A) chr13:100123008-100123307 and B) chr15:72891463-72891762. Each column corresponds to the phenotype value for a mouse, and each row represents the founder strain dosage for a given mouse. Residual heterozygosity is observed in some mice for this haplotype segment. Haplotype codes correspond to A/J (A), C57BL/6J (B), 129S1/SvImJ (C), NOD/LtJ (D), NZO/HILtJ (E), CAST/EiJ (F), PWK/PhJ (G), and WSB/EiJ (H).

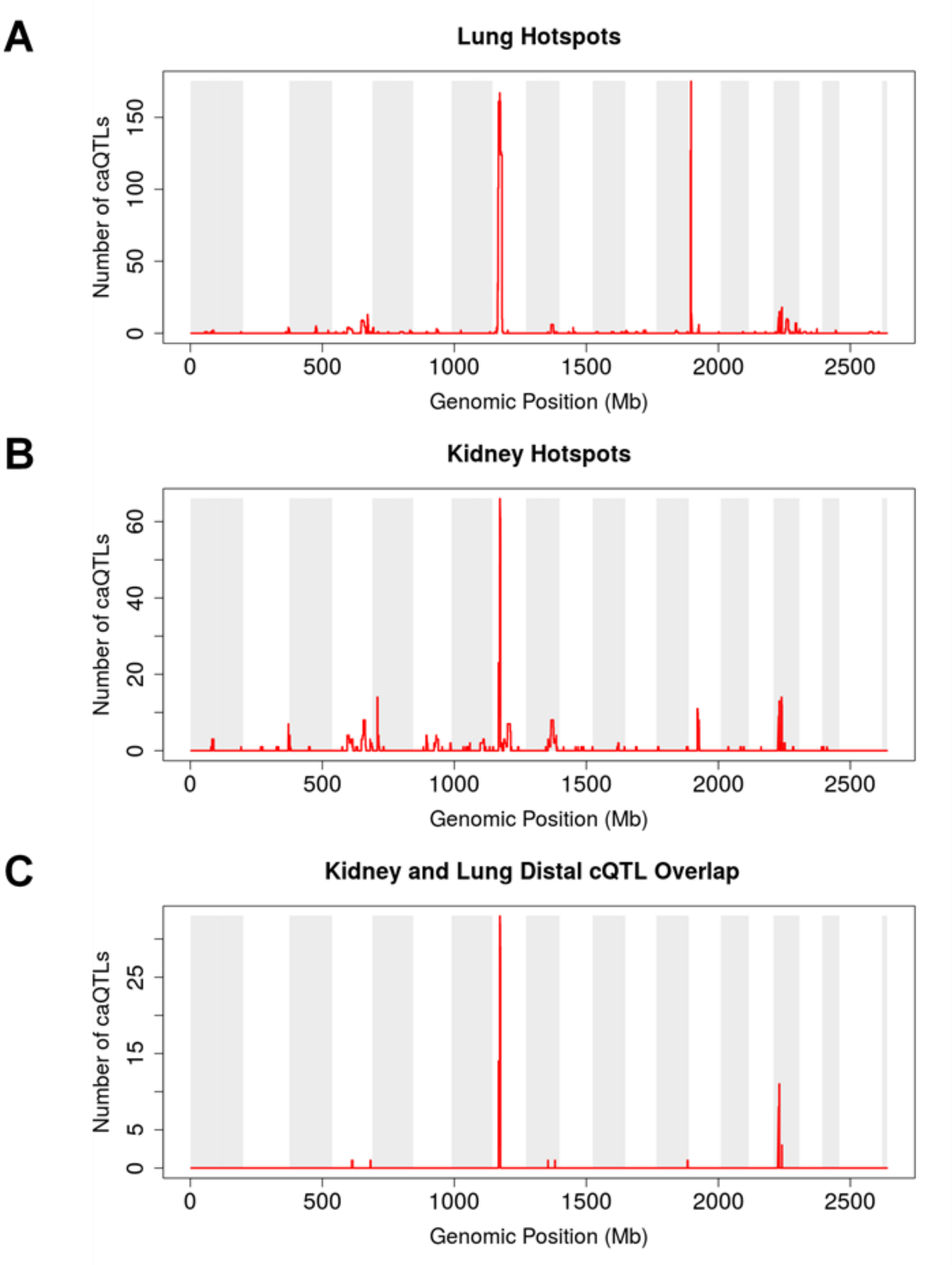

Figure 3.47. Genome-wide 625 ppm lung and kidney cQTL frequencies. Genome-wide view of the number of significant cQTLs (FDR 0.05) at each haplotype segment for A) lung and B) kidney. C) The number of significant distal cQTLs found in both lung and kidney in the 625 ppm group (FDR 0.05) at each haplotype segment.

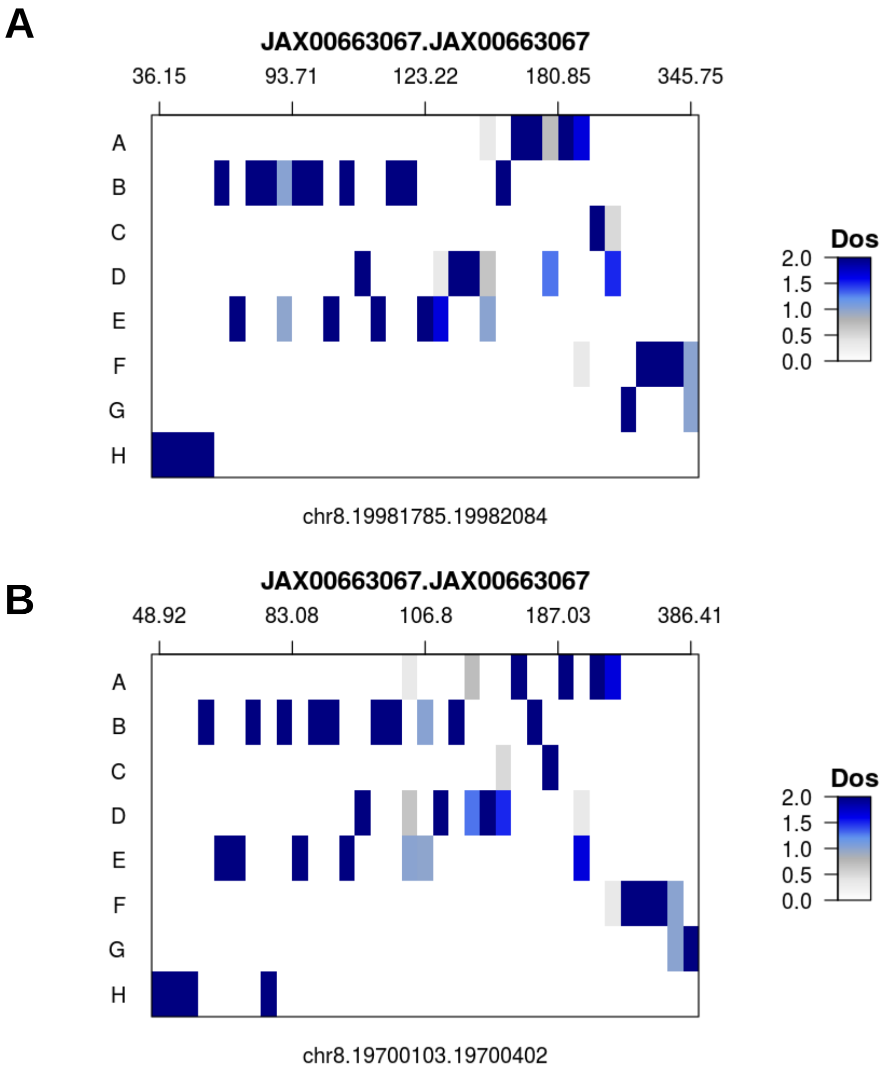

Figure 3.48. Lung chr8 hotspot haplotype dosages. Haplotype dosages at segment JAX00663067 (chr8:23562597) for each mouse arranged by lung 625 ppm chromatin accessibility phenotype values at chromatin windows A) chr8:19981785-19982084 and B) chr8:197000103-19700402. Each column corresponds to the phenotype value for a mouse, and each row represents the founder strain dosage for a given mouse. Residual heterozygosity is observed in some mice for this haplotype segment. Haplotype codes correspond to A/J (A), C57BL/6J (B), 129S1/SvImJ (C), NOD/LtJ (D), NZO/HILtJ (E), CAST/EiJ (F), PWK/PhJ (G), and WSB/EiJ (H).

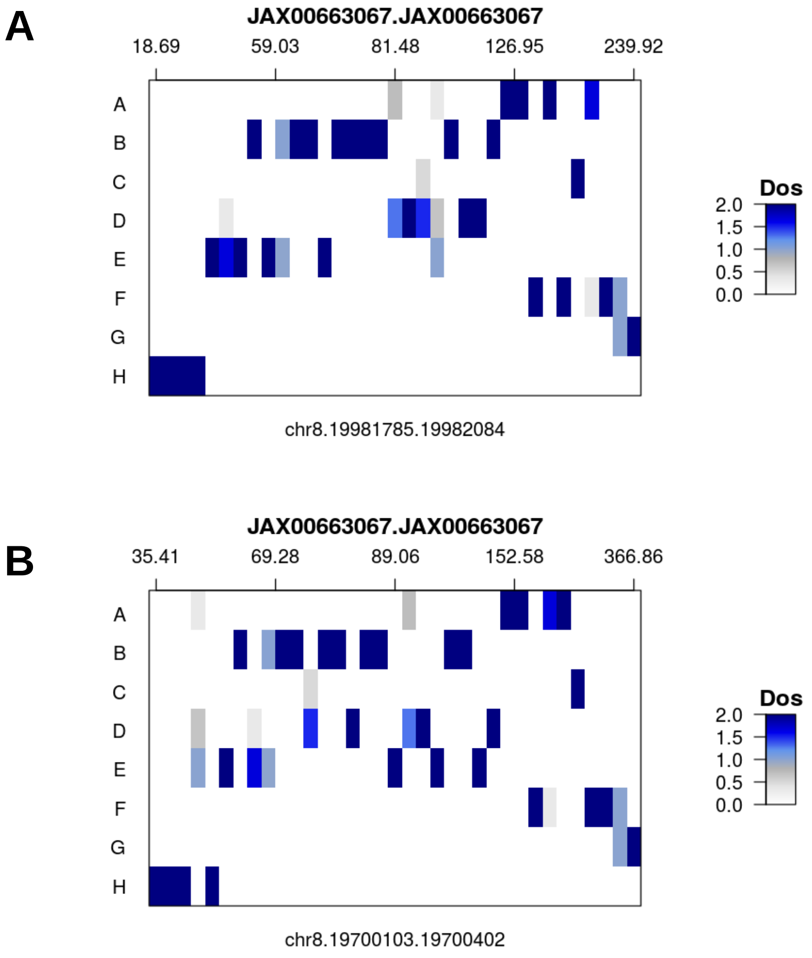

Figure 3.49. Kidney chr8 hotspot haplotype dosages. Haplotype dosages at segment JAX00663067 (chr8:23562597) for each mouse arranged by kidney 625 ppm chromatin accessibility phenotype values at chromatin windows A) chr8:19981785-19982084 and B) chr8:197000103-19700402. Each column corresponds to the phenotype value for a mouse, and each row represents the founder strain dosage for a given mouse. Residual heterozygosity is observed in some mice for this haplotype segment. Haplotype codes correspond to A/J (A), C57BL/6J (B), 129S1/SvImJ (C), NOD/LtJ (D), NZO/HILtJ (E), CAST/EiJ (F), PWK/PhJ (G), and WSB/EiJ (H).

CC Strain	Treatment Group	Tissue	RNA-seq	ATAC-seq
CC001	0 ppm	Kidney	yes	yes
CC001	0 ppm	Liver	yes	yes
CC001	0 ppm	Lung	yes	yes
CC001	1500 ppm	Kidney	yes	no
CC001	1500 ppm	Liver	yes	no
CC001	1500 ppm	Lung	yes	no
CC001	625 ppm	Kidney	yes	yes
CC001	625 ppm	Liver	yes	yes
CC001	625 ppm	Lung	yes	yes
CC002	0 ppm	Kidney	yes	yes
CC002	0 ppm	Liver	yes	yes
CC002	0 ppm	Lung	yes	yes
CC002	1500 ppm	Kidney	yes	no
CC002	1500 ppm	Liver	yes	no
CC002	1500 ppm	Lung	yes	no
CC002	625 ppm	Kidney	yes	yes
CC002	625 ppm	Liver	yes	yes
CC002	625 ppm	Lung	yes	yes
CC003	0 ppm	Kidney	yes	yes
CC003	0 ppm	Liver	yes	yes
CC003	0 ppm	Lung	yes	yes
CC003	1500 ppm	Kidney	yes	no
CC003	1500 ppm	Liver	yes	no
CC003	1500 ppm	Lung	yes	no
CC003	625 ppm	Kidney	yes	yes
CC003	625 ppm	Liver	yes	yes
CC003	625 ppm	Lung	yes	yes
CC004	0 ppm	Kidney	yes	yes
CC004	0 ppm	Liver	yes	yes
CC004	0 ppm	Lung	yes	yes
CC004	1500 ppm	Kidney	yes	no
CC004	1500 ppm	Liver	yes	no
CC004	1500 ppm	Lung	yes	no
CC004	625 ppm	Kidney	yes	yes
CC004	625 ppm	Liver	yes	yes
CC004	625 ppm	Lung	yes	yes
CC005	0 ppm	Kidney	yes	yes
CC005	0 ppm	Liver	yes	yes
CC005	0 ppm	Lung	yes	yes
CC005	1500 ppm	Kidney	yes	no
CC005	1500 ppm	Liver	yes	no
CC005	1500 ppm	Lung	yes	no
CC006	0 ppm	Kidney	yes	yes
CC006	0 ppm	Liver	yes	yes
CC006	0 ppm	Lung	yes	yes
CC006	625 ppm	Kidney	yes	yes

CC006	625 ppm	Liver	yes	yes
CC006	625 ppm	Lung	yes	yes
CC007	0 ppm	Kidney	yes	yes
CC007	0 ppm	Liver	yes	yes
CC007	0 ppm	Lung	yes	yes
CC007	1500 ppm	Kidney	yes	no
CC007	1500 ppm	Liver	yes	no
CC007	1500 ppm	Lung	yes	no
CC010	0 ppm	Kidney	yes	yes
CC010	0 ppm	Liver	yes	yes
CC010	0 ppm	Lung	yes	yes
CC010	1500 ppm	Kidney	yes	no
CC010	1500 ppm	Liver	yes	no
CC010	1500 ppm	Lung	yes	no
CC010	625 ppm	Kidney	yes	yes
CC010	625 ppm	Liver	yes	yes
CC010	625 ppm	Lung	yes	yes
CC011	0 ppm	Kidney	yes	yes
CC011	0 ppm	Liver	yes	yes
CC011	0 ppm	Lung	yes	yes
CC011	1500 ppm	Kidney	yes	no
CC011	1500 ppm	Liver	yes	no
CC011	1500 ppm	Lung	yes	no
CC012	0 ppm	Kidney	yes	yes
CC012	0 ppm	Liver	yes	yes
CC012	0 ppm	Lung	yes	yes
CC012	1500 ppm	Kidney	yes	no
CC012	1500 ppm	Liver	yes	no
CC012	1500 ppm	Lung	yes	no
CC012	625 ppm	Kidney	yes	yes
CC012	625 ppm	Liver	yes	yes
CC012	625 ppm	Lung	yes	yes
CC013	0 ppm	Kidney	yes	yes
CC013	0 ppm	Liver	yes	yes
CC013	0 ppm	Lung	yes	yes
CC013	1500 ppm	Kidney	yes	no
CC013	1500 ppm	Liver	yes	no
CC013	1500 ppm	Lung	yes	no
CC013	625 ppm	Kidney	yes	yes
CC013	625 ppm	Liver	yes	yes
CC013	625 ppm	Lung	yes	yes
CC015	0 ppm	Kidney	yes	yes
CC015	0 ppm	Liver	yes	yes
CC015	0 ppm	Lung	yes	yes
CC015	1500 ppm	Kidney	yes	no
CC015	1500 ppm	Liver	yes	no
CC015	1500 ppm	Lung	yes	no
CC015	625 ppm	Kidney	yes	yes
CC015	625 ppm	Liver	yes	yes
CC015	625 ppm	Lung	yes	yes

CC016	0 ppm	Kidney	yes	yes
CC016	0 ppm	Liver	yes	yes
CC016	0 ppm	Lung	yes	yes
CC016	1500 ppm	Kidney	yes	no
CC016	1500 ppm	Liver	yes	no
CC016	1500 ppm	Lung	yes	no
CC016	625 ppm	Kidney	yes	yes
CC016	625 ppm	Liver	yes	yes
CC016	625 ppm	Lung	yes	yes
CC017	0 ppm	Kidney	yes	yes
CC017	0 ppm	Liver	yes	yes
CC017	0 ppm	Lung	yes	yes
CC017	1500 ppm	Kidney	yes	no
CC017	1500 ppm	Liver	yes	no
CC017	1500 ppm	Lung	yes	no
CC017	625 ppm	Kidney	yes	yes
CC017	625 ppm	Liver	yes	yes
CC017	625 ppm	Lung	yes	yes
CC018	0 ppm	Kidney	yes	yes
CC018	0 ppm	Liver	no	no
CC018	0 ppm	Lung	yes	yes
CC018	1500 ppm	Kidney	yes	no
CC018	1500 ppm	Liver	yes	no
CC018	1500 ppm	Lung	yes	no
CC018	625 ppm	Kidney	yes	yes
CC018	625 ppm	Liver	yes	yes
CC018	625 ppm	Lung	yes	yes
CC019	0 ppm	Kidney	yes	yes
CC019	0 ppm	Liver	yes	yes
CC019	0 ppm	Lung	yes	yes
CC019	1500 ppm	Kidney	yes	no
CC019	1500 ppm	Liver	yes	no
CC019	1500 ppm	Lung	yes	no
CC019	625 ppm	Kidney	yes	yes
CC019	625 ppm	Liver	yes	yes
CC019	625 ppm	Lung	yes	yes
CC020	0 ppm	Kidney	yes	yes
CC020	0 ppm	Liver	yes	yes
CC020	0 ppm	Lung	yes	yes
CC020	1500 ppm	Kidney	yes	no
CC020	1500 ppm	Liver	yes	no
CC020	1500 ppm	Lung	yes	no
CC020	625 ppm	Kidney	yes	yes
CC020	625 ppm	Liver	yes	yes
CC020	625 ppm	Lung	yes	yes
CC021	0 ppm	Kidney	yes	yes
CC021	0 ppm	Liver	yes	yes
CC021	0 ppm	Lung	yes	yes
CC021	625 ppm	Kidney	yes	yes
CC021	625 ppm	Liver	yes	yes

CC021	625 ppm	Lung	yes	yes
CC023	0 ppm	Kidney	yes	yes
CC023	0 ppm	Liver	yes	yes
CC023	0 ppm	Lung	yes	yes
CC023	1500 ppm	Kidney	yes	no
CC023	1500 ppm	Liver	yes	no
CC023	1500 ppm	Lung	yes	no
CC023	625 ppm	Kidney	yes	yes
CC023	625 ppm	Liver	yes	yes
CC023	625 ppm	Lung	yes	yes
CC024	0 ppm	Kidney	yes	yes
CC024	0 ppm	Liver	yes	yes
CC024	0 ppm	Lung	yes	yes
CC024	1500 ppm	Kidney	yes	no
CC024	1500 ppm	Liver	yes	no
CC024	1500 ppm	Lung	yes	no
CC024	625 ppm	Kidney	yes	yes
CC024	625 ppm	Liver	yes	yes
CC024	625 ppm	Lung	yes	yes
CC025	0 ppm	Kidney	yes	yes
CC025	0 ppm	Liver	yes	yes
CC025	0 ppm	Lung	yes	yes
CC025	625 ppm	Kidney	yes	yes
CC025	625 ppm	Liver	yes	yes
CC025	625 ppm	Lung	yes	yes
CC027	0 ppm	Kidney	yes	yes
CC027	0 ppm	Liver	yes	yes
CC027	0 ppm	Lung	yes	yes
CC027	1500 ppm	Kidney	yes	no
CC027	1500 ppm	Liver	yes	no
CC027	1500 ppm	Lung	yes	no
CC027	625 ppm	Kidney	yes	yes
CC027	625 ppm	Liver	yes	yes
CC027	625 ppm	Lung	yes	yes
CC028	0 ppm	Kidney	yes	yes
CC028	0 ppm	Liver	yes	yes
CC028	0 ppm	Lung	yes	yes
CC028	1500 ppm	Kidney	yes	no
CC028	1500 ppm	Liver	yes	no
CC028	1500 ppm	Lung	yes	no
CC028	625 ppm	Kidney	yes	yes
CC028	625 ppm	Liver	yes	yes
CC028	625 ppm	Lung	yes	yes
CC029	0 ppm	Kidney	yes	yes
CC029	0 ppm	Liver	yes	yes
CC029	0 ppm	Lung	yes	yes
CC029	1500 ppm	Kidney	yes	no
CC029	1500 ppm	Liver	yes	no
CC029	1500 ppm	Lung	yes	no
CC029	625 ppm	Kidney	yes	yes

CC029	625 ppm	Liver	yes	yes
CC029	625 ppm	Lung	yes	yes
CC030	0 ppm	Kidney	yes	yes
CC030	0 ppm	Liver	yes	yes
CC030	0 ppm	Lung	yes	yes
CC030	1500 ppm	Kidney	yes	no
CC030	1500 ppm	Liver	yes	no
CC030	1500 ppm	Lung	yes	no
CC030	625 ppm	Kidney	yes	yes
CC030	625 ppm	Liver	yes	yes
CC030	625 ppm	Lung	yes	yes
CC031	0 ppm	Kidney	yes	yes
CC031	0 ppm	Liver	yes	yes
CC031	0 ppm	Lung	yes	yes
CC031	1500 ppm	Kidney	yes	no
CC031	1500 ppm	Liver	yes	no
CC031	1500 ppm	Lung	yes	no
CC031	625 ppm	Kidney	yes	yes
CC031	625 ppm	Liver	yes	yes
CC031	625 ppm	Lung	yes	yes
CC032	0 ppm	Kidney	yes	yes
CC032	0 ppm	Liver	yes	yes
CC032	0 ppm	Lung	yes	yes
CC032	1500 ppm	Kidney	yes	no
CC032	1500 ppm	Liver	yes	no
CC032	1500 ppm	Lung	yes	no
CC033	0 ppm	Kidney	yes	yes
CC033	0 ppm	Liver	yes	yes
CC033	0 ppm	Lung	yes	yes
CC033	1500 ppm	Kidney	yes	no
CC033	1500 ppm	Liver	yes	no
CC033	1500 ppm	Lung	yes	no
CC033	625 ppm	Kidney	yes	yes
CC033	625 ppm	Liver	yes	yes
CC033	625 ppm	Lung	yes	yes
CC035	0 ppm	Kidney	yes	yes
CC035	0 ppm	Liver	yes	yes
CC035	0 ppm	Lung	yes	yes
CC035	1500 ppm	Kidney	yes	no
CC035	1500 ppm	Liver	yes	no
CC035	1500 ppm	Lung	yes	no
CC035	625 ppm	Kidney	yes	yes
CC035	625 ppm	Liver	yes	yes
CC035	625 ppm	Lung	yes	yes
CC036	0 ppm	Kidney	yes	yes
CC036	0 ppm	Liver	yes	yes
CC036	0 ppm	Lung	yes	yes
CC036	1500 ppm	Kidney	yes	no
CC036	1500 ppm	Liver	yes	no
CC036	1500 ppm	Lung	yes	no

CC036	625 ppm	Kidney	yes	yes
CC036	625 ppm	Liver	yes	yes
CC036	625 ppm	Lung	yes	yes
CC037	0 ppm	Kidney	yes	yes
CC037	0 ppm	Liver	yes	yes
CC037	0 ppm	Lung	yes	yes
CC037	1500 ppm	Kidney	yes	no
CC037	1500 ppm	Liver	yes	no
CC037	1500 ppm	Lung	yes	no
CC037	625 ppm	Kidney	yes	yes
CC037	625 ppm	Liver	yes	yes
CC037	625 ppm	Lung	yes	yes
CC038	0 ppm	Kidney	yes	yes
CC038	0 ppm	Liver	yes	yes
CC038	0 ppm	Lung	yes	yes
CC038	1500 ppm	Kidney	yes	no
CC038	1500 ppm	Liver	yes	no
CC038	1500 ppm	Lung	yes	no
CC038	625 ppm	Kidney	yes	yes
CC038	625 ppm	Liver	yes	yes
CC038	625 ppm	Lung	yes	yes
CC039	0 ppm	Kidney	yes	yes
CC039	0 ppm	Liver	yes	yes
CC039	0 ppm	Lung	yes	yes
CC039	1500 ppm	Kidney	yes	no
CC039	1500 ppm	Liver	yes	no
CC039	1500 ppm	Lung	yes	no
CC039	625 ppm	Kidney	yes	yes
CC039	625 ppm	Liver	yes	yes
CC039	625 ppm	Lung	yes	yes
CC040	0 ppm	Kidney	yes	yes
CC040	0 ppm	Liver	yes	yes
CC040	0 ppm	Lung	yes	yes
CC040	1500 ppm	Kidney	yes	no
CC040	1500 ppm	Liver	yes	no
CC040	1500 ppm	Lung	yes	no
CC040	625 ppm	Kidney	yes	yes
CC040	625 ppm	Liver	yes	yes
CC040	625 ppm	Lung	yes	yes
CC041	0 ppm	Kidney	yes	yes
CC041	0 ppm	Liver	yes	yes
CC041	0 ppm	Lung	yes	yes
CC041	1500 ppm	Kidney	yes	no
CC041	1500 ppm	Liver	yes	no
CC041	1500 ppm	Lung	yes	no
CC042	0 ppm	Kidney	yes	yes
CC042	0 ppm	Liver	yes	yes
CC042	0 ppm	Lung	yes	yes
CC042	1500 ppm	Kidney	yes	no
CC042	1500 ppm	Liver	yes	no

CC042	1500 ppm	Lung	yes	no
CC042	625 ppm	Kidney	yes	yes
CC042	625 ppm	Liver	yes	yes
CC042	625 ppm	Lung	yes	yes
CC043	0 ppm	Kidney	yes	yes
CC043	0 ppm	Liver	yes	yes
CC043	0 ppm	Lung	yes	yes
CC043	625 ppm	Kidney	yes	yes
CC043	625 ppm	Liver	yes	yes
CC043	625 ppm	Lung	yes	yes
CC044	0 ppm	Kidney	yes	yes
CC044	0 ppm	Liver	yes	yes
CC044	0 ppm	Lung	yes	yes
CC044	1500 ppm	Kidney	yes	no
CC044	1500 ppm	Liver	yes	no
CC044	1500 ppm	Lung	yes	no
CC044	625 ppm	Kidney	yes	yes
CC044	625 ppm	Liver	yes	yes
CC044	625 ppm	Lung	yes	yes
CC045	0 ppm	Kidney	yes	yes
CC045	0 ppm	Liver	yes	yes
CC045	0 ppm	Lung	yes	yes
CC045	625 ppm	Kidney	yes	yes
CC045	625 ppm	Liver	yes	yes
CC045	625 ppm	Lung	yes	yes
CC046	0 ppm	Kidney	yes	yes
CC046	0 ppm	Liver	yes	yes
CC046	0 ppm	Lung	yes	yes
CC046	1500 ppm	Kidney	yes	no
CC046	1500 ppm	Liver	yes	no
CC046	1500 ppm	Lung	yes	no
CC049	0 ppm	Kidney	yes	yes
CC049	0 ppm	Liver	yes	yes
CC049	0 ppm	Lung	yes	yes
CC049	1500 ppm	Kidney	yes	no
CC049	1500 ppm	Liver	yes	no
CC049	1500 ppm	Lung	yes	no
CC049	625 ppm	Kidney	yes	yes
CC049	625 ppm	Liver	yes	yes
CC049	625 ppm	Lung	yes	yes
CC051	0 ppm	Kidney	no	no
CC051	0 ppm	Liver	yes	yes
CC051	0 ppm	Lung	yes	yes
CC051	1500 ppm	Kidney	yes	no
CC051	1500 ppm	Liver	yes	no
CC051	1500 ppm	Lung	yes	no
CC051	625 ppm	Kidney	yes	yes
CC051	625 ppm	Liver	yes	yes
CC051	625 ppm	Lung	yes	yes
CC053	0 ppm	Kidney	yes	yes

CC053	0 ppm	Liver	yes	yes
CC053	0 ppm	Lung	yes	yes
CC053	1500 ppm	Kidney	yes	no
CC053	1500 ppm	Liver	yes	no
CC053	1500 ppm	Lung	yes	no
CC053	625 ppm	Kidney	yes	yes
CC053	625 ppm	Liver	yes	yes
CC053	625 ppm	Lung	yes	yes
CC055	0 ppm	Kidney	yes	yes
CC055	0 ppm	Liver	yes	yes
CC055	0 ppm	Lung	yes	yes
CC055	1500 ppm	Kidney	yes	no
CC055	1500 ppm	Liver	yes	no
CC055	1500 ppm	Lung	yes	no
CC055	625 ppm	Kidney	yes	yes
CC055	625 ppm	Liver	yes	yes
CC055	625 ppm	Lung	yes	yes
CC057	0 ppm	Kidney	yes	yes
CC057	0 ppm	Liver	yes	yes
CC057	0 ppm	Lung	yes	yes
CC057	1500 ppm	Kidney	yes	no
CC057	1500 ppm	Liver	yes	no
CC057	1500 ppm	Lung	yes	no
CC057	625 ppm	Kidney	yes	yes
CC057	625 ppm	Liver	yes	yes
CC057	625 ppm	Lung	yes	yes
CC059	0 ppm	Kidney	yes	yes
CC059	0 ppm	Liver	yes	yes
CC059	0 ppm	Lung	yes	yes
CC060	1500 ppm	Kidney	yes	no
CC060	1500 ppm	Liver	yes	no
CC060	1500 ppm	Lung	yes	no
CC060	625 ppm	Kidney	yes	yes
CC060	625 ppm	Liver	yes	yes
CC060	625 ppm	Lung	yes	yes
CC061	0 ppm	Kidney	yes	yes
CC061	0 ppm	Liver	yes	yes
CC061	0 ppm	Lung	yes	yes
CC061	1500 ppm	Kidney	yes	no
CC061	1500 ppm	Liver	yes	no
CC061	1500 ppm	Lung	yes	no
CC061	625 ppm	Kidney	yes	yes
CC061	625 ppm	Liver	yes	yes
CC061	625 ppm	Lung	yes	yes
CC062	0 ppm	Kidney	yes	yes
CC062	0 ppm	Liver	yes	yes
CC062	0 ppm	Lung	yes	yes
CC062	1500 ppm	Kidney	yes	no
CC062	1500 ppm	Liver	yes	no
CC062	1500 ppm	Lung	yes	no

CC062	625 ppm	Kidney	yes	yes
CC062	625 ppm	Liver	yes	yes
CC062	625 ppm	Lung	yes	yes
CC068	0 ppm	Kidney	yes	yes
CC068	0 ppm	Liver	yes	yes
CC068	0 ppm	Lung	yes	yes
CC068	1500 ppm	Kidney	yes	no
CC068	1500 ppm	Liver	yes	no
CC068	1500 ppm	Lung	yes	no
CC068	625 ppm	Kidney	yes	yes
CC068	625 ppm	Liver	yes	yes
CC068	625 ppm	Lung	yes	yes

Table 3.1. Inventory of CC mouse samples. CC strains represented in each tissue and treatment group and whether RNA-seq and ATAC-seq data were processed for these strains.

Tissue and Treatment Group	Total Genes Tested	Total eQTLs (FDR 0.05)	eGenes Detected (FDR 0.05)	eGenes with Local Associations (FDR 0.05)	eGenes with Distal Associations (FDR 0.05)
Lung – Control	17,675	400	67	34	36
Lung – 625 ppm	17,536	505	97	58	44
Lung – 1500 ppm	17,376	869	114	66	55
Liver – Control	13,629	1,368	168	101	82
Liver – 625 ppm	13,355	842	102	44	65
Liver – 1500 ppm	13,299	1,803	216	128	104
Kidney – Control	15,894	1,718	213	105	124
Kidney – 625 ppm	15,625	2,336	283	162	136
Kidney – 1500 ppm	15,408	3,523	417	255	188

Table 3.2. eQTL mapping results overview. Summary of eQTL mapping results in each tissue and treatment group. The term “eGenes” denotes an eQTL gene with at least one significantly associated segment.

Tissue and Treatment Group	Total Chromatin Windows Tested	Total cQTLs (FDR 0.05)	Unique cQTL Chromatin Windows (FDR 0.05)	cQTL Chromatin Windows with Local Associations (FDR 0.05)	cQTL Chromatin Windows with Distal Associations (FDR 0.05)
Lung – Control	24,949	3,328	325	72	256
Lung – 625 ppm	24,666	8,472	594	93	510
Liver – Control	25,762	88	15	7	11
Liver – 625 ppm	25,785	0	0	0	0
Kidney – Control	25,081	5,700	353	105	257
Kidney – 625 ppm	25,280	2,532	238	87	164

Table 3.3. cQTL mapping results overview. Summary of cQTL mapping results in each tissue and treatment group.

CHAPTER IV

Discussion

Current technologies allow for researchers to employ a broad array of high-throughput methods to produce data related to many facets of gene regulation such as chromatin organization and interaction, DNA methylation, histone modifications, TF occupancy, microRNA expression, and gene expression. Studies have integrated these data types to elucidate the details of gene regulation, and in doing so provided insight into the variability in how these regulatory components present and interact between different cell types, tissues, and conditions. Furthermore, studies assessing the contributions of genetic variation on these gene regulatory properties have demonstrated regulatory variability due to genetic differences between individuals. With such an elaborate picture of the factors contributing to understanding context-specific gene regulation, much still needs to be learned about the biology as well as how to devise analyses that best take advantage of the data.

In chapter II, I provided an overview of footprinting and assessed currently held assumptions about TF footprints. Using ENCODE ChIP-seq and DNase-seq data in conjunction with TF motif site predictions, I showed that DNase-seq signals at active and inactive motif sites are more heterogeneous than previously assumed, violating assumptions many current footprinters use when identifying TFBSs. To address this heterogeneity, I introduced DeFCoM, a novel machine learning framework for predicting TFBSs using DNase-seq data. DeFCoM applied a supervised learning approach to classification in order to learn the characteristics of footprints as opposed to enforcing assumptions about their structure. Through a comprehensive comparison with 9 other footprinters using 71 test sets for 18 TFs, I showed that DeFCoM performed the best overall.

Furthermore, by assessing footprintability at varying sequencing depths and using data sets of different signal quality, I observed that footprintability varied drastically by TF. Intuitively, sequencing depth and data set signal quality should improve footprintability, but the degree of benefit for both are also TF dependent. In addition, I applied DeFCoM to ATAC-seq GM12878 data and noted similar but slightly decreased performance relative to using DNase-seq GM12878 data, though this may be attributable to differences in sequencing depth and signal-to-noise.

As an area of research, genomic footprinting is still maturing. The first genome-wide footprinting paper was published in 2009, and it reported the detection of footprints in the *Saccharomyces cerevisiae* genome [44]. Despite the lack of a comprehensive understanding of footprint characteristics, papers have been published that include extensive analyses of TF dynamics and networks for numerous cell types and across species based solely on computational footprint predictions [101,102] using a method that performed poorly in my footprinter comparisons. The conclusions drawn solely from footprinting raised concerns that were voiced in [103]. To appropriately make use of footprinting, DNase-seq and ATAC-seq signal at TFBSs need to be better characterized. My work in chapter II contributes to the need for better footprint characterization by highlighting the degree of footprint heterogeneity and showing the impact of sequencing depth and signal quality on footprintability. This research further expands the field by demonstrating a motif-centric approach to assessing footprint predication accuracy and using a single, unified framework to evaluate most currently existing footprinters.

Looking towards future research in genomic footprinting, a priority needs to be placed on further characterizing the biological and technical factors impacting TF footprint profiles. Currently, annotating TFBSs for footprinting studies relies predominately on ChIP-seq data, but the literature remains scarce in regards to how an imperfect concordance between ChIP-seq based annotations and TF footprints impacts footprint characterization and prediction. For elucidating properties of footprint profiles, more refined and accurate TFBS annotations would produce more robust characterizations. An improved understanding of footprint signals in chromatin accessibility data would allow for more refined and appropriate statistical models and machine learning methods to be implemented for footprint classification. As footprint characterization improves and quantitative models become more accurate, footprinting offers promising new avenues for investigating TF binding. In the context of differential chromatin accessibility studies, DNase-seq and ATAC-seq could be utilized more effectively to identify differential TF binding through evaluating changes in chromatin accessibility signal at TFBSs. Accurate identification of differential TF binding with chromatin accessibility assays would offer an alternative option to ChIP-seq for genome-wide studies. ATAC-seq and DNase-seq would capture binding events for many TFs within a single experiment as opposed to ChIP-seq which assays one TF per experiment. Another exciting avenue for genomic footprinting research lies in pairing genotype information with accurate quantification of footprint changes. In genomics studies where genotype and chromatin accessibility data are available, footprint variability could be defined as a quantitative trait and tested for associations with genetic variability using QTL mapping approaches. These analyses would help clarify TF binding and the gene regulatory mechanisms influenced by genetic variability within a given context.

In chapter III, I used liver, lung, and kidney ATAC-seq and RNA-seq data for CC mice from three different BD treatment groups (control, 625 ppm exposure, and 1500 ppm exposre) to assess the impact of BD exposure on gene expression and chromatin accessibility. From PCA of the samples, I observed that tissue differences accounted for more variability than BD exposure in both gene expression and chromatin accessibility. Further analyses revealed that in both gene expression and chromatin accessibility, lung and liver exhibited more pronounced changes than kidney. This result complements a previous study in which DNA methylation and histone modifications were shown to significantly change at the global level for lung and liver but not kidney tissue in C57BL6/J mice [69]. My analyses build upon these observations by showing a similar affect for chromatin accessibility and gene expression using mice with diverse genetic backgrounds. Since the ATAC-seq and RNA-seq data provide genome-wide information, these data can be further examined to identify specific regulatory elements and genes that are driving the tissue-specific differences to BD exposure.

In addition to the global variation assessments, I also provide a characterization of eQTLs and cQTLs in each of the tissues and treatment groups. In general, I observed eQTLs and cQTLs to be more consistent within a tissue and across treatment groups than within a treatment group and across tissues. Additionally, of the 3 tissues, I observed lung eQTLs to be the least concordant across treatment groups suggesting that the BD response in lung is causing unique gene expression changes that are also associated with genetic variation. For the cQTL analysis, I discovered previously uncharacterized cQTL hotspots in lung and kidney and identified potential causal founder haplotypes driving these hotspots. From these observations, the dynamic changes in lung and its significance in BD metabolism suggest that it should be prioritized in future analyses. To my knowledge, this cQTL characterization study is the first to be done for a mouse population, but both the eQTL and cQTL analyses leave many questions to be answered in regards to better understanding the genetic underpinnings of BD response. In [35], CC mice liver eQTLs were mapped using a "delta" phenotype model. Because CC strains are inbred, mice from the same strain can be subject to two different conditions, and the difference or ratio between measurements of gene expression can be used as the phenotype. With the CC population the genetic diversity between strains allows for this delta phenotype to be incorporated into eQTL mapping. For future work, this approach can be taken with the BD data to infer more direct relationships between genetics, chromatin accessibility, gene expression, and BD exposure. The ability to take advantage of such a model demonstrates the utility of the CC, in conjunction with sequencing-based assays of gene regulation and gene expression, in elucidating the gene regulatory architecture underlying toxic chemical exposure.

As genomic footprinting improves, TF binding dynamics can be incorporated into studies akin to the work described in chapter III. This experimental design would be able to integrate genotype, gene expression, chromatin accessibility, TF occupancy, and an environmental perturbation to paint a more comprehensive picture of context-specific gene regulation. Incorporating additional data types such as microRNA expression and 3D chromatin interactions among others would offer even greater explanatory power. With our current understanding of the complexity of gene regulation, such integrative approaches will be necessary to deduce the underlying mechanistic links between genetics and complex traits.

REFERENCES

1. Levine M. Transcriptional enhancers in animal development and evolution. Current Biology. 2010. doi:10.1016/j.cub.2010.06.070

2. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961;3: 318–56. doi:10.1016/S0022-2836(61)80072-7

3. Levo M, Segal E. In pursuit of design principles of regulatory sequences. Nat Rev Genet. 2014;15: 453–468. doi:10.1038/nrg3684

4. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012;337: 1190–1195. doi:10.1126/science.1222794

5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921. doi:10.1038/35057062

6. ENCODE Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518: 317–330. doi:10.1038/nature14248

7. Feingold E, Good P, Guyer M, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306: 636–40. doi:10.1126/science.1105136

8. Pai AA, Pritchard JK, Gilad Y. The Genetic and Mechanistic Basis for Variation in Gene Regulation. PLoS Genet. 2015;11. doi:10.1371/journal.pgen.1004857

9. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011;470: 198–203. doi:10.1038/nature09796

10. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. Analytical Chemistry. 2011. pp. 4327–4341. doi:10.1021/ac2010857

11. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316: 1497–1502. doi:1141319 [pii]\r10.1126/science.1141319

12. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132: 311–22. doi:10.1016/j.cell.2007.12.014

13. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, et al. Direct RNA sequencing. Nature. 2009;461: 814–818. doi:10.1038/nature08390

14. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014;515: 355–364. doi:10.1038/nature13992

15. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

16. Gilad Y, Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data. F1000 Research. 2015;121: 1–20. doi:10.12688/f1000research.6536.1

17. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Epigenomics: Roadmap for regulation. Nature. 2015;518: 314–316. doi:10.1038/518314a

18. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

19. The International HapMap Consortium. The International HapMap Project. Nature. 2003;426: 789–796. doi:10.1038/nature02168

20. Visscher PM, Wray NR, Zhang Q, Sklar P, Mccarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101: 5–22. doi:10.1016/j.ajhg.2017.06.005

21. Brem RB. Genetic Dissection of Transcriptional Regulation in Budding Yeast. Science. 2002;296: 752–755. doi:10.1126/science.1069516

22. Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W, Ayroles JF, et al. Genomic Variation and Its Impact on Gene Expression in Drosophila melanogaster. PLoS Genet. 2012;8. doi:10.1371/journal.pgen.1003055

23. Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA, Castellani LW, et al. Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. Nat Genet. 2005;37: 1224–1233. doi:10.1038/ng1619

24. Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348: 648–660. doi:10.1126/science.1262110

25. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of genetic variants that affect histone modifications in human cells. Science. 2013;342: 747–9. doi:10.1126/science.1242429

26. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–8. doi:10.1038/nmeth.2688

27. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet. 2015;48: 206–213. doi:10.1038/ng.3467

28. Briggs JP. The zebrafish: a new model organism for integrative physiology. Am J Physiol Regul Integr Comp Physiol. 2002;282: R3–R9. doi:10.1152/ajpregu.00589.2001

29. Jeibmann A, Paulus W. Drosophila melanogaster as a model organism of brain diseases. International Journal of Molecular Sciences. 2009. pp. 407–440. doi:10.3390/ijms10020407

30. Cheon D-J, Orsulic S. Mouse models of cancer. Annu Rev Pathol. 2011;6: 95–119. doi:10.1146/annurev.pathol.3.121806.154244

31. Threadgill DW, Hunter KW, Williams RW. Genetic dissection of complex and quantitative traits: From fantasy to reality via a community effort. Mamm Genome. 2002;13: 175–178. doi:10.1007/s00335-001-4001-y

32. Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet. 2004;36: 1133–1137. doi:10.1038/ng1104-1133

33. Iraqi F a., Mahajne M, Salaymah Y, Sandovski H, Tayem H, Vered K, et al. The genome architecture of the collaborative cross mouse genetic reference population. Genetics. 2012;190: 389–401. doi:10.1534/genetics.111.132639

34. Kelada SNP, Aylor DL, Peck BCE, Ryan JF, Tavarez U, Buus RJ, et al. Genetic Analysis of Hematological Parameters in Incipient Lines of the Collaborative Cross. G3: Genes, Genomes, Genetics. 2012;2: 157–165. doi:10.1534/g3.111.001776

35. Mosedale M, Kim Y, Brock WJ, Roth SE, Wiltshire T, Eaddy JS, et al. Candidate risk factors and mechanisms for tolvaptan-induced liver injury are identified using a collaborative cross approach. Toxicol Sci. 2017;156: 438–454. doi:10.1093/toxsci/kfw269

36. Furey TS, Sethupathy P. Genetics Driving Epigenetics. Science. 2013;342: 705–706. doi:10.1126/science.1246755

37. Vockley CM, Barrera A, Reddy TE. Decoding the role of regulatory element polymorphisms in complex disease. Current Opinion in Genetics and Development. 2017. pp. 38–45. doi:10.1016/j.gde.2016.10.007

38. Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011;21: 456–64. doi:10.1101/gr.112656.110

39. Quach B, Furey TS. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. Bioinformatics. 2016;33: btw740. doi:10.1093/bioinformatics/btw740

40. Thomas S, Li X-Y, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, et al. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. Genome Biol. 2011;12: R43. doi:10.1186/gb-2011-12-5-r43

41. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee B-K, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011;21: 1757–67. doi:10.1101/gr.121541.111

42. Nag R, Smerdon MJ. Altering the chromatin landscape for nucleotide excision repair. Mutat Res - Rev Mutat Res. 2009;682: 13–20. doi:10.1016/j.mrrev.2009.01.002

43. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316: 1497–1502. doi:1141319 [pii]\r10.1126/science.1141319

44. Hesselberth J, Chen X, Zhang Z. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 2009;6: 283–289. doi:10.1038/NMETH.1313

45. Chen X, Hoffman MM, Bilmes JA, Hesselberth JR, Noble WS. A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. Bioinformatics. 2010;26: 334–342. doi:10.1093/bioinformatics/btq175

46. Kahara J, Lahdesmaki H. BinDNase : a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. 2015;31: 2852–2859. doi:10.1093/bioinformatics/btv294

47. Luo K, Hartemink AJ. Using DNase digestion data to accurately identify transcription factor binding sites. Pac Symp Biocomput. 2013; 80–91. doi:10.1142/9789814447973_0009

48. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. Nature Publishing Group; 2012;489: 83–90. doi:10.1038/nature11212

49. Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res. 2013;41. doi:10.1093/nar/gkt850

50. Pique-Regi R, Degner JF, Pai A a, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011;21: 447–55. doi:10.1101/gr.112623.110

51. Raj A, Shim H, Gilad Y, Pritchard JK, Stephens M. msCentipede: Modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding. PLoS One. 2015;10: 1–15. doi:10.1371/journal.pone.0138030

52. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal A a, van Hoff JP, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol. 2014;32: 171–178. doi:10.1038/nbt.2798

53. Sung M-HH, Guertin MJJ, Baek S, Hager GLL. DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence. Mol Cell. 2014;56: 1–11. doi:10.1016/j.molcel.2014.08.016

54. Siggers T, Gordan R. Protein-DNA binding: Complexities and multi-protein codes. Nucleic Acids Res. 2014;42: 2099–2111. doi:10.1093/nar/gkt1112

55. Gusmão EG, Dieterich C, Zenke M, Costa IG. Detection of Active Transcription Factor Binding Sites with the Combination of DNase Hypersensitivity and Histone Modifications. Bioinformatics. 2014; 1–8. doi:10.1093/bioinformatics/btu519

56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2012;12: 2825–2830. doi:10.1007/s13398-014-0173-7.2

57. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing classifier performance in R. Bioinformatics. 2005;21: 3940–3941. doi:10.1093/bioinformatics/bti623

58. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics. 2008;24: 2537–8. doi:10.1093/bioinformatics/btn480

59. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2014;42: 2976–87. doi:10.1093/nar/gkt1249

60. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27: 1017–8. doi:10.1093/bioinformatics/btr064

61. Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. Proc 5th Annu ACM Work Comput Learn Theory. 1992; 144–152. doi:10.1.1.21.3818

62. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22: 1813–31. doi:10.1101/gr.136184.111

63. He HH, Meyer C a, Hu SS, Chen M-W, Zang C, Liu Y, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Methods. 2013;11: 73–8. doi:10.1038/nmeth.2762

64. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla N V., Herrera F, Alaiz-Rodriguez R, et al. A unifying view on dataset shift in classification. Pattern Recognit. 2012;45: 521–530. doi:10.1016/j.patcog.2011.06.019

65. Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL. Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics. 2012;28: 56–62. doi:10.1093/bioinformatics/btr614

66. Gusmao EG, Allhoff M, Zenke M, Costa IG. Analysis of computational footprinting methods for DNase sequencing experiments. Nat Methods. Nature Publishing Group; 2016;13: 303–9. doi:10.1038/nmeth.3772

67. Yardimci GG, Frank CL, Crawford GE, Ohler U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Res. 2014;42: 11865–11878. doi:10.1093/nar/gku810

68. Gusmao EG, Dieterich C, Zenke M, Costa IG. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics. 2014;30: 3143–3151. doi:10.1093/bioinformatics/btu519

69. Chappell G, Kobets T, O'Brien B, Tretyakova N, Sangaraju D, Kosyk O, et al. Epigenetic events determine tissue-specific toxicity of inhalational exposure to the genotoxic chemical 1,3-butadiene in male C57BL/6J mice. Toxicol Sci. 2014;142: 375–384. doi:10.1093/toxsci/kfu191

70. Efferth T, Volm M. Pharmacogenetics for individualized cancer chemotherapy. Pharmacology and Therapeutics. 2005. pp. 155–176. doi:10.1016/j.pharmthera.2005.02.005

71. Costa PM, Fadeel B. Emerging systems biology approaches in nanotoxicology: Towards a mechanism-based understanding of nanomaterial hazard and risk. Toxicol Appl Pharmacol. 2016;299: 101–111. doi:10.1016/j.taap.2015.12.014

72. White WC. Butadiene production process overview. Chem Biol Interact. 2007;166: 10–14. doi:10.1016/j.cbi.2007.01.009

73. Boldry EJ, Patel YM, Kotapati S, Esades A, Park SL, Tiirikainen M, et al. Genetic Determinants of 1,3-Butadiene Metabolism and Detoxification in Three Populations of Smokers with Different Risks of Lung Cancer. Cancer Epidemiol Prev Biomarkers. 2017;26. Available: http://cebp.aacrjournals.org.libproxy.lib.unc.edu/content/26/7/1034.long

74. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Occupational exposures to mists and vapours from strong inorganic acids, and other industrial chemicals. [Internet]. Available: http://publications.iarc.fr/Book-And-Report-Series/Iarc-Monographs-On-The-Evaluation-Of-Carcinogenic-Risks-To-Humans/Occupational-Exposures-To-Mists-And-Vapours-From-Strong-Inorganic-Acids-And-Other-Industrial-Chemicals-1992

75. Filser JG, Hutzler C, Meischner V, Veereshwarayya V, Csanády GA. Metabolism of 1,3-butadiene to toxicologically relevant metabolites in single-exposed mice and rats. Chem Biol Interact. 2007;166: 93–103. doi:10.1016/j.cbi.2006.03.002

76. Goggin M, Swenberg JA, Walker VE, Tretyakova N. Molecular dosimetry of 1,2,3,4-diepoxybutane induced DNA-DNA cross-links in B6C3F1 mice and F344 rats exposed to 1,3-butadiene by inhalation. Cancer Res. 2009;69: 2479–2486. doi:10.1158/0008-5472.CAN-08-4152

77. Melnick RL, Huff JE. 1,3-Butadiene induces cancer in experimental animals at all concentrations from 6.25 to 8000 parts per million. IARC Sci Publ. 1993; 309–22. Available: http://www.ncbi.nlm.nih.gov/pubmed/8070878

78. Owen PE, Glaister JR, Gaunt IF, Pullinger DH. Inhalation toxicity studies with 1,3-butadiene. 3. Two year toxicity/carcinogenicity study in rats. Am Ind Hyg Assoc J. 1987;48: 407–413. doi:10.1080/15298668791384959

79. Koturbash I, Scherhag A, Sorrentino J, Sexton K, Bodnar W, Swenberg J a., et al. Epigenetic mechanisms of mouse interstrain variability in genotoxicity of the environmental toxicant 1,3-butadiene. Toxicol Sci. 2011;122: 448–456. doi:10.1093/toxsci/kfr133

80. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2009;26: 493–500. doi:10.1093/bioinformatics/btp692

81. Lassmann T, Hayashizaki Y, Daub CO. TagDust - A program to eliminate artifacts from next generation sequencing data. Bioinformatics. 2009;25: 2839–2840. doi:10.1093/bioinformatics/btp527

82. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

83. Love MI, Anders S, Huber W. Differential analysis of count data - the DESeq2 package [Internet]. Genome Biology. 2014. doi:110.1186/s13059-014-0550-8

84. Masuda K, Takahashi S, Nomura K, Inoue M, Widholm JM. A simple procedure for the isolation of pure nuclei from carrot embryos in synchronized cultures. Plant Cell Rep. 1991;10: 329–333. Available: https://link.springer.com/content/pdf/10.1007%2FBF00193152.pdf

85. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. Methods in Molecular Biology. 2016. pp. 283–334. doi:10.1007/978-1-4939-3578-9_15

86. Carroll TS, Liang Z, Salama R, Stark R, de Santiago I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Front Genet. 2014;5. doi:10.3389/fgene.2014.00075

87. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. http://www.repeatmasker.org. 2013. Available: http://repeatmasker.org

88. Kent WJ. BLAT - The BLAST-like alignment tool. Genome Res. 2002;12: 656–664. doi:10.1101/gr.229202. Article published online before March 2002

89. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. BioMed Central Ltd; 2010;11: R119. doi:10.1186/gb-2010-11-12-r119

90. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, et al. Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection. PLoS Genet. 2012;8: e1002789. doi:10.1371/journal.pgen.1002789

91. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

92. Degner JF, Pai A a, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012;482: 390–4. doi:10.1038/nature10808

93. Fu C-P, Welsh CE, de Villena FP-M, McMillan L. Inferring ancestry in admixed populations using microarray probe intensities. Proc ACM Conf Bioinformatics, Comput Biol Biomed - BCB '12. 2012; 105–112. doi:10.1145/2382936.2382950

94. Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA, Baric RS, et al. Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. 2011;21: 1213–1222. doi:10.1101/gr.111310.110

95. Haley CS, Knott SA. A simple regression method for mapping quantitative trait loci in line crossess using flanking markers. Heredity (Edinb). 1992;69: 315–324.

96. Zhang Z, Wang W, Valdar W. Bayesian modeling of haplotype effects in multiparent populations. Genetics. 2014;198: 139–156. doi:10.1534/genetics.114.166249

97. Storey J. A Direct Approach to False Discovery Rates on JSTOR. Wiley Online Libr. 2002;64: 479–498. doi:10.1111/1467-9868.00346

98. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, et al. Heritability and genomics of gene expression in peripheral blood. Nat Genet. 2014;46: 430–7. doi:10.1038/ng.2951

99. Weiser M, Mukherjee S, Furey TS. Novel distal eQTL analysis demonstrates effect of population genetic architecture on detecting and interpreting associations. Genetics. 2014;198: 879–93. doi:10.1534/genetics.114.167791

100. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in Genetics. 2008. pp. 408–415. doi:10.1016/j.tig.2008.06.001

101. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos J. Circuitry and dynamics of human transcription factor regulatory networks. Cell. 2012;150: 1274–86. doi:10.1016/j.cell.2012.04.040

102. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature. 2014;515: 365–370. doi:10.1038/nature13972

103. Sung M, Baek S, Hager GL. Genome-wide footprinting : ready for prime time? Nat Methods. 2016;13. doi:10.1038/nmeth.3766