Functional Interpretation of High-Throughput Sequencing Data

by

Chee Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2016

Doctoral Committee:

Associate Professor Maureen A. Sartor, Chair
Professor Margit Burmeister
Assistant Professor Maria E. Figueroa
Assistant Professor Stephen C.J. Parker
Assistant Professor Indika Rajapakse

Dedication

For my friends and family, to whom I look to for my inspiration, motivation, and strength.

Acknowledgements

I would like to thank my advisor, Dr. Maureen Sartor, for her abiding and positive mentorship. She has taught me how to approach research problems from points of views that I have never considered, and provided guidance whenever I needed it. She always has an open door, and is extremely dedicated to the success of her students. Her style of mentorship has inspired a fun yet productive atmosphere in our lab. I have been very fortunate to find such a wonderful mentor for the last 5+ years.

Next I would like to thank my committee members, past and present, for sharing their expertise and enduring through long, often difficult to schedule, committee meetings. I did my first research rotation with Dr. Margit Burmeister, and years later chose her to be on my committee because her ability to quickly identify computational and biological connections. She did not disappoint in doing the same during all my meetings. Dr. Jim Cavalcoli was an expert in sequencing technologies and provided helpful advice for the sequencing problems that we tackled. Dr. Ken Figueroa was often the one in meetings who would anchor us computational biologist from diverging too far into the technical side. She provided a much needed perspective on the practicality and biological usefulness of my research. Dr. Stephen Parker was always excited about my researcher from the first day I met him during a student luncheon for his faculty interview. I am very grateful he was able to join my committee and share his experience with gene set enrichment testing and repetitive elements. Dr. Indika Rajapakse always has a smile on his face and the ability to think outside of the box. His expertise in genome organization has helped inspire many ideas in this dissertation.

I would also like to acknowledge additional bioinformatics faculty and staff members. Thank you to Dr. Brian Athey, Dr. Dan Burns, and additionally thanks to Dr. Margit Burmeister for their academic and personal advising. Their dedication to the success of the students in the department goes beyond what is asked of them. Thank you to Dr. Alla Karnovsky, with whom I did a research rotation with, for serving on my prelim committee and also providing valuable career guidance. I would also like to thank Dr. Tricia Wittkopp, and additional thanks to Dr. Dan Burns and Dr. Indika Rajapakse for also serving on my prelim committee. My prelim committee was critical to providing direction in the scientific process, which ultimately made me a better researcher. Thank you to Dr. Jeff de Wet for being a great teacher and introducing more programming languages to me. My success in the program will also not be possible without the help of all the bioinformatics staff. In particular I would like to thank Julia Eussen for handling all the paperwork, scheduling, and all things behind-the-scenes so seamlessly that we didn't even notice it was being done. The same goes to Alex Terzian, Aaron Bookvich, and Paul Trombley who has helped set up all of our technology needs from AV equipment to poster printing and designing graphics for our web apps. Also thanks to Terry Weymouth for helping us develop our websites for our various gene set enrichment programs.

I have been very fortunate to find a nourishing, fun, and positive research lab. Again I would like to thank Maureen, as well as all Sartor Lab members past and present for creating an environment where we helped each other and were involved in each's other successes. Thank you to Yanxiao Zhang, Raymond Calvalcante, Snehal Patil, Julie Kim, Yongseok Park, Yu-Hsuan Lin, Lada Koneva, Pelle Hall, Shama Virani, Chris Lee, and Yidan Liu for your friendships and research camaraderie.

My experience in graduate school has also been shaped by my classmates and my friends. From the bioinformatics and biostatistics departments, I want to thank Yanxiao Zhang, Ellen Schmidt, Tom Wolfe, Alex Schmidt, Ryan Welch, Paul Imbriano, and Lee Sam for working alongside me in class projects, helping me with coding when I needed it, and for being great research buddies. My friends whom I met in Michigan: Ariell, Kathryn, Esha, Elizabeth, Kaitlyn, and Purvi - we have kept each other sane all these years. My close friends in MN whom I miss dearly: Meredith, Rachel, Jenna, Brittany, Joe, Andy, Dave, and Chris. Thank you so much for your friendships. My companion, Jonathan Las Fargeas, thank you for being there for me, your infectious energy helps keep my spirits up.

Finally I would like to thank my family: my dad, Nao Bee, my mom, Na, and my sisters: Mai Tong, Dia, Mai Song, and my brothers: Touko, Muachi, Teng, and Muamong. You are huge inspirations to me and I love you all so much.

Table of Contents

List of Figures

Figures

Supplementary Figures

List of Tables

Tables

Supplementary Tables

List of Abbreviations

cFC: corrected fold change

DBP: DNA binding protein

DE: differential expression

DEG: differentially expressed gene

DEX: dexamethasone

DHT: dihydrotestosterone

EtOH: ethanol

FAIRE-seq: Formaldehyde-Assisted Isolation of Regulatory Elements

FDR: false discovery rate

FET: Fisher's exact test

GO: Gene Ontology

GREAT: Genomic Regions Enrichment of Annotations Tool

GRα: glucocorticoid receptor α

GSE: Gene set enrichment

GSEA: Gene Set Enrichment Analysis

GWAS: genome-wide association studies

H3K27me3: trimethylation of histone 3 lysine 27

HTS: high-throughput sequencing

KEGG: Kyoto Encyclopedia of Genes and Genomes

L1: LINE-1

LINEs: long interspersed nuclear elements

LTRs: long terminal repeats

MEME: Multiple EM for Motif Elicitation

NRSF: neuron-restrictive silencer factor

PAX5: paired box 5

QQ: quantile-quantile

RPKM: reads per kilobase per million tags sequenced

SINEs: short interspersed nuclear elements

SIX5: SIX homeobox 5

TES: transcription end site

TSS: transcription start site

Abstract

Functional interpretation of high-throughput sequencing (HTS) data provides insight into biological systems, including important pathways in the context under study. A common approach is gene set enrichment (GSE) testing. GSE emerged in the age of microarrays as a way to biologically interpret long lists of differentially expressed genes (DEGs). However, HTS data has characteristics not present in microarray data that can bias GSE results. My thesis is focused on identifying, characterizing, and accounting for biases to improve functional interpretation in HTS data.

In this thesis, I present GSE tests designed for ChIP-seq data and RNA-seq data. Our tests have applications beyond HTS data, which we show by using them to analyze genomic features, including mappability and repeat content. ChIP-Enrich is a GSE test for ChIP-seq data. It includes a database of locus definitions to annotate peaks to different gene loci (such as exons, introns, promoters, and other intergenic regions), which allows for biological discovery unique to different regions. ChIP-Enrich empirically adjusts for the observed bias due to the varying lengths of these gene loci in its enrichment test. RNA-Enrich is a GSE test for RNA-seq data. RNA-Enrich corrects for the selection bias often observed in RNA-seq data, where long and highly expressed genes are more likely to be identified as DEGs. Unlike other GSE tests for RNA-seq data, RNA-Enrich does not require permutations or a cut-off to define DEGs, and works well with small sample sizes. For both ChIP-Enrich and RNA-Enrich, we showed well-calibrated type I error compared to competing methods. Finally, we characterize sequence mappability, which is one potential bias in the interpretation of HTS data. We characterize properties of the main contributors of low mappability (transposons and segmental duplications), overall mappability, and their relationship with gene locus length and function. Across different transcribed and regulatory regions, certain gene functions showed unique signatures involving significantly more/fewer associated repeats, higher/lower mappability, and longer/shorter locus length. Our analyses provide insight into evolutionary selection pressures that maintain complexity of gene regulation. Overall, we demonstrate that considering characteristics of the human genome is essential to improving functional interpretation of HTS data.

Chapter 1 Introduction

1.1 Introduction

The era of high-throughput sequencing (HTS), also known as next-generation sequencing or massively parallel sequencing, has inspired progress in genomics that has produced an incredible amount of data. Along with generating the data, researchers have also developed various algorithms for each step of data processing. What starts as a multitude of short DNA sequences eventually undergoes quality control, genome alignment, gene assignment or quantitation, a myriad of statistical analyses, and then, finally, interpretation [1, 2]. Since the assembly of the human genome, we have expanded high-throughput sequencing from sequencing of full genomes to a wide variety of applications that can measure gene expression [3], gene regulation and epigenetic marks [4]. The human genome is complicated but not random. Studying it poses many challenges. The organization of the human genome (e.g. exon/intron structures, spatial organization, and sequence redundancy) [5, 6] can perpetuate as biases in downstream analyses of HTS studies, resulting in incorrect interpretations of results, and therefore also may lead researchers to draw erroneous conclusions. This dissertation is focused on identifying, characterizing, and accounting for such biases to improve functional interpretation in various HTS platforms, including RNA-seq and ChIP-seq. Biases due to gene length, sequencing selection, and sequence mappability (the ability to uniquely align short DNA sequences) will be explored. While the research only includes select sequencing platforms, the findings and the methodology may be applicable to many types of current and, perhaps, future iterations of high-throughput sequencing technologies.

1.2 Background

1.2.1 High-throughput technologies

In this section, I explore some basic designs of popular high-throughput technologies, and pros and cons of each technology as relevant to this dissertation, beginning with the technology that predates HTS: microarrays. Prior to HTS, microarrays were the tools of choice for measuring gene expression, copy number variation, DNA-binding (ChIP-chips), SNP genotyping, and more. They still remain popular in some areas of study such as DNA methylation and SNP genotyping. Gene expression microarrays make use of oligo hybridization and fluorescent labelling of probes to measure gene expression. Probes contain DNA sequence targets that are spread across the genome to target various parts of the gene body and/or intergenic regions. Probes occupy a small chip, and each sample cDNA (generated from sample RNA) can be PCR amplified to determine levels of expression based on the fluorescent signal. The Human MethylationEPIC BeadChip, for example, is the latest methylation microarray from Illumina and has over 850,000 probes that measure methylation using bisulfite-converted DNA across the genome. Many statistical applications for high-throughput data already existed for microarrays when next-generation sequencing was developed. Naively, approaches developed for microarrays were applied to sequencing data with little regard to whether underlying assumptions were correct. As I will further discuss in this dissertation, understanding the differences in data from microarrays and from sequencing is essential in developing the right tools, and for biological interpretation.

The incentive to complete the Human Genome Project inspired next-generation sequencing technologies, which in turn motivated the development of a variety of molecular methods to explore a wide range of biological phenomena. The basis of massively parallel sequencing requires library preparation from select fragmentation of DNA. Fragments are then ligated to common adaptor sequences, and optionally undergo multiple rounds of amplification to increase product input [7]. In RNA-seq, cDNA libraries can be prepared from RNA with specific features, for example those with a poly-A tail, a unique feature to mRNA [3]. Variations on RNA-seq to increase power to detect specific isoforms and increase coverage include creating libraries that are paired-end and/or strand-specific. RNA-seq achieves a much higher dynamic range than microarrays and is not subjected to the same selection bias that may occur due to probe placement. However, RNA-seq is not without its problems. Longer genes and more highly expressed genes are more likely to have more reads, and therefore are more likely to be called as significantly differentially expressed for many of the commonly used tests [8, 9]. In addition, sequenced fragments exhibit positional and sequence-specific preferences [10]. Several methods have been proposed to correct for such biases at the gene level [11-13], however if left uncorrected (or corrected for poorly), these biases can affect interpretation at the gene function level, e.g. gene set enrichment testing – a topic I explore in Chapter 3 of this dissertation.

While RNA-seq is the HTS equivalent to gene expression microarrays, ChIP-seq is the HTS equivalent of ChIP-chip. ChIP is chromatin immunoprecipitation, a procedure to study the interaction between proteins and DNA *in vivo*. ChIP-chip is ChIP combined with microarrays, whereas ChIP-seq is ChIP combined with massively parallel sequencing. In ChIP-seq, which is used to study genome-wide protein-DNA interactions (e.g. to identify transcription factor binding sites), libraries can be prepared from DNA bound by protein using an anti-body to target the particular protein of interest after crosslinking of protein and DNA, and then sample fragmentation. ChIP-seq has various modifications for applications beyond transcription factors. Histone modification ChIP-seq involves using antibodies that can detect specific histone tail modifications such as methylation or acetylation of one of the histone amino acids in the tetramer nucleosome protein complex. DNase-seq bypasses the antibody and performs fragmentation by targeting DNase hypersensitive sites with DNaseI digestion. FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements) does not use antibodies and therefore is not limited to particular DNA-binding proteins. Proteins are crosslinked, followed by sample fragmentation, sonication, and then phenol-chloroform extraction of DNA [4, 14]. Similar to RNA-seq, ChIP-seq also has a selection bias, in that longer genes and genes with more intergenic space around them are often more likely to have an associated peak. Technically speaking, peaks are areas of the genome where there is a significant number of consensus sequence read alignments; biologically speaking, they are the predicted protein-DNA binding sites. A study of the performance of a dozen popularly used peak calling algorithms [15] (including MACS, spp, and PeakSeq – three peak callers that have been used by the ENCODE consortium [16]) on ChIP-seq datasets for three transcription factors with different binding profiles, showed that the number of peaks identified can vary by tens of thousands among different peak callers. There are several newer peak callers [17-19] that have shown significant improvement in calling peaks with higher specificity – i.e. tested datasets produce peaks with high occurrence of binding motif and can be reproduced with ChIP and PCR. Unfortunately, it is often difficult to convince users to diverge from their usual protocol. Conceivably, a list of ChIP peaks that contains a substantial amount of false negatives may affect downstream analysis if certain categories of genes consistently have many peaks or few peaks.

A critical step to analysis of HTS data is alignment of reads. Longer read lengths are more likely to uniquely align to areas of the genome but shorter reads may align to multiple places, and therefore are often less mappable (i.e. have less unique sequence). Repeats pose a problem to sequence alignment. An estimated 45% of the human genome consists of repetitive elements called transposons [20, 21]. As we show in Chapter 4, they especially have a high occurrence in introns and intergenic regions. *Alu* elements, a type of short interspersed nuclear element, make up about 11% of the human genome and often occur near the transcription start site (TSS) [22]. Depending on how non-uniquely mappable reads are handled by the chosen sequence aligner, ChIP-seq peaks that occur near the TSS may be less likely to be detected if the peak region is not highly mappable. This also applies to any other region in the genome that is not uniquely mappable.

1.2.2 Gene set enrichment testing

Gene set enrichment (GSE) testing, also known as functional analysis of genes, is a way to identify important gene functions that differ between two different biological states. We have used it, for different genomic features like mappabiilty, repeat content, and gene length (Chapter 4). GSE can give insights into how a biological system works, and perhaps which pathways are important targets. GSE emerged in the age of microarrays as a way to interpret the biological relevance of long lists of differentially expressed genes (DEGs). Microarrays conquered the problem of obtaining gene expression profiles; however often the result was a list of hundreds or thousands of DEGs, which was simply too much information to absorb one gene at a time. The most common goal of GSE testing is to find pathways or biological processes that were affected by the conditions of an experiment. For example, in a microarray experiment that tested changes in gene expression before and after a drug treatment, GSE testing can enlighten researchers about which pathways were most affected by the treatment. In ChIP-seq data, one may be interested in discovering what biological processes are regulated by a transcription factor, or in what diseases it may play a role.

Gene sets may be constructed with various purposes in mind. Gene Ontology (GO), a commonly used gene set database, describes gene products in terms of biological processes, molecular functions, and cellular components [23]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways is a collection of manually drawn pathway maps representing molecular interactions and reaction networks [24]. Genes can also be classified into diseases (using MeSH terms or OMIM), spatially by open or condensed chromatin (cytoband), drug target lists, and many more. Typically GSE tests for overrepresentation, or enrichment, of gene sets. If the GSE test is two-sided, it may also test for underrepresentation, or depletion.

Several popular GSE methods exist for microarray data that are commonly applied to HTS data, three of which I will highlight here: Fisher’s exact test, GSEA, and random sets. Fisher’s exact test (FET) is a statistical test that analyzes contingency tables. In the case of GSE, the table is typically 2-by-2, where rows are gene set status (if the gene is in the gene set or not), and columns are gene significance status (for example, either differentially expressed or not, have a ChIP peak or not, etc). DAVID is perhaps the most widely used FET-based GSE test [25, 26]. DAVID modifies the FET by subtracting 1 from the table cell with the number of DEGs that are in the gene set. This modified FET is more conservative, it reduces the unpredictability of small gene sets, while having minimal effect on larger gene sets. Many implementations of FET exist besides DAVID, including GoMiner [27, 28] and HOMER [29] – which is designed for HTS data, allowing for association of peaks to genes as well as GSE testing. The underlying assumption of FET that is often violated with HTS data is that all genes are assumed to have equal power (similar type II error rate) and to be equally likely to be a false positive (similar type I error rate). As we show in chapters 2-4, there are several factors that make a gene more or less likely to be identified as differentially expressed, have a peak, or any differential status.

Another widely used GSE method is GSEA, abbreviated from “Gene Set Enrichment Analysis”. GSEA uses a weighted Kolmogorov-Smirnov test, a non-parametric test (i.e. it does not rely on a statistical distribution) [30, 31]. Input for GSEA is a ranked list of gene-level statistics, (for example, fold change ranked from most up-regulated to most down-regulated). A running-sum of the ordered gene-level statistics of genes in the gene set is calculated and compared to those not in the gene set to obtain an enrichment score. To calculate an associated p-value, a null distribution is generated by permuting phenotype labels, and the original enrichment score is compared to the distribution of permuted enrichment scores. There are several versions of GSEA that have been adapted for HTS data, including GSAASeqSP [32], which permutes read counts of genes, and SeqGSEA [33], which permutes the negative binomial statistics after using DESeq to test for differential gene expression. While GSEA, and some GSEA-like methods give the user the option to permute genes instead of phenotype labels, the GSEA authors recommend doing the latter, which “preserves gene-gene correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting genes” [31]. However, this cannot be done with small sample sized experiments, because a sufficient number of unique permutations is required to obtain p-values with reasonable accuracy. Both GSAASeqSP and SeqGSEA recommend sample sizes of at least 6-7 in each phenotype for accurate GSE results. Often in HTS experiments, it is not feasible to have this many samples in each condition.

Finally, I would like to describe the random sets method for GSE testing [34]. Random sets is, in a way, a hybrid of FET and GSEA. While FET is limited to a binary gene status, the random sets approach, like GSEA, allows gene level statistics to be any general quantitative expression score, s_g, for gene *g*. However, the test may not behave properly if the distribution of scores, s_g, is far from normally distributed. The unstandardized enrichment score for random sets is $\bar{X} = \frac{1}{m}\int_{g\in C} s_g$, where *m* is the number of genes in gene set *C*. Intuitively, the enrichment score of one gene set is compared to random sets of the same size *m* that are drawn from the population of *G* genes $\binom{G}{m}$, however this comparison is done with a simple theoretical formula rather than performed directly, resulting in a quick, reproducible result. This approach is equivalent to permuting gene-level statistics among the gene labels. In the case where s_g is binary, random sets reduces to FET. Newton et al (2007) showed the distribution of $\bar{X}$ is approximately normal, and used the method of moments to determine a test statistic and associated p-value for enrichment. While random sets does not require a binary gene status like FET, nor is it limited by small sample sizes like GSEA, it does continue to make the assumption that all genes are equally likely to be detected as significant (i.e. all genes have approximately the same power and same type 1 error). I expand on the performance of random sets in more detail in Chapter 3.

Other GSE tests mentioned in this dissertation belong to the class of model-based methods. LRpath, also developed for microarray data, uses a logistic regression model to test whether differential expression p-values, or any significance values of choice, for genes in a gene set are more or less significant than those for other genes [35, 36]. LRpath makes the same assumption that all genes are assumed to have equal power and equally likely to be a false positive. In Chapter 4, we use a different logistic regression GSE method, Broad-Enrich [37], that uses coverage proportion of gene (appropriate for HTS data like ChIP-seq of histone modifications) as the values of interest and corrects for gene locus length. In some cases, FET, GSEA, random sets, and LRpath can be applied to HTS data. However, as I will explain in this dissertation, the underlying assumptions of these tests are often not met with HTS data.

1.3 Overview of dissertation

Why is it important to account for bias in gene set enrichment testing? What are the origins of bias in HTS data? And what can we do to correct for these biases so that our conclusions are biologically sound? These are important questions that I seek to answer in this dissertation. Functional interpretation (e.g. GSE testing) bridges the gap between unwieldy, often hard to organize, high-throughput sequencing data and biological relevance. Improving our methods for analyzing HTS data begins with identification and characterization of the biases in the current state of HTS technology. It is important to note both technical and biological biases, as sometimes both are not mutually exclusive and often will affect the accuracy of conclusions that one draws from analyzing HTS data.

In Chapter 2, we present a gene set enrichment test, called ChIP-Enrich, that was designed for ChIP-seq data and other assays that produce a list of relatively narrow genomic regions (peaks). To perform GSE testing with ChIP-seq data, peaks have to be assigned to genes. The common practice is to assign peaks (defined by their midpoints) to the nearest TSS. As I have mentioned before, genes with longer loci are more likely to have a peak assigned to them. We address the bias of gene locus length in assignment of ChIP-seq peaks to genes, and developed a method that empirically adjusts for the observed bias. In doing so, we created a database of locus definitions to annotate peaks to different loci, such as introns, exons, nearest TSS, and 5kb around a TSS, therefore requiring different locus length adjustments and allowing for biological discovery that may be unique to different regulatory regions. We demonstrated that ChIP-Enrich is able to correct for all kinds of peak-to-locus-length relationships while maintaining a good type I error. We also introduce the significance of correcting for sequence mappability, which I expound upon in Chapter 4. In the ChIP-Enrich project, my contributions as a co-first author, in addition to helping to write the manuscript, were: (1) creating the locus definitions for the available genomes; (2) calculating mappability values for different *k*mer lengths and genomes; (3) applying ChIP-Enrich to 63 different ENCODE ChIP-seq datasets; (4) along with co-first author Ryan Welch, performing permutation testing on select ENCODE datasets to show how type I error rates of ChIP-Enrich compared to competing methods; (5) analyzing and interpreting a case study on a glucocorticoid receptor ChIP-seq dataset from ENCODE in respect to its regulatory activity in promoter and distal regions; and (6) assisting in creating the R Bioconductor package, *chipenrich* and *chipenrich.data* that are also employed on our website: http://chip-enrich.med.umich.edu/.

In Chapter 3, we present a gene set enrichment test, RNA-Enrich, to address a selection bias often observed in RNA-seq data where long and highly expressed genes are more likely to be identified as significantly differentially expressed (i.e. not all genes have the same power). We modify the random sets method to adjust for average read count per gene. To calculate a test statistic for gene set enrichment, we determine a weight based on the observed relationship between read count and significance values in the data. In datasets where there is no relationship, RNA-Enrich approximates random sets. In datasets where there is a relationship, correcting for the bias allows for improved identification of truly enriched or depleted gene sets. We compare RNA-Enrich to other GSE methods: random sets, DAVID, GOseq, GSAASeqSP and SeqGSEA. We also implement RNA-Enrich using significance values from different sources, including differential gene expression p-values from two different methods and a corrected fold change instead of p-values. We show that using average gene read count as a proxy for the selection bias greatly improves the type I error compared to other GSE tests for RNA-seq data.

Chapter 4 delves into sequence mappability, its relationship with repeat elements and gene length, and their correlation with gene function. We perform GSE testing using highly prevalent transposons in the human genome: L1 elements, which are a type of long intersperse nuclear element, and *Alu* elements, a type of short intersperse nuclear element. Together, these two repetitive elements make up an estimated 26% of the human genome. Segmental duplications, which are long duplications of DNA sequence that are 1-200kb in length and have >90% identity, make up 5% of the human genome. We show that across different regulatory regions, certain gene functions show unique enrichment signatures of *Alu* elements, L1 elements, segmental duplication, mappability, and gene length. That is, certain types of genes have significantly more/fewer associated repeat elements, higher/lower mappability, and longer/shorter gene locus length. While sequence mappability is a technical measurement that depends on sequence read length, we show that it can elucidate genomic architecture that relates to gene length and repeat elements. Our analyses gives insight into how evolutionary selection has been used to maintain the required complexity of gene regulations, and which types of genes have been most affected.

Chapter 2 ChIP-Enrich: Gene set enrichment testing for ChIP-seq data

2.1 Introduction

Genome-wide high-throughput experiments can assess transcription factor binding, epigenetic marks, differential gene expression or disease association, and often result in thousands of identified genomic regions or genes. Gene set enrichment testing is one way to determine how these lists of genomic regions or genes are related biologically, e.g. by assessment of Gene Ontology (GO)) terms [38-40]. For ChIP-seq experiments, oftentimes thousands of transcription factor binding sites or histone modification sites are identified. Enrichment testing of this data, or with a union or intersection of multiple ChIP-seq datasets, can identify key biological processes, functions, disease gene signatures, or other biological concepts regulated by the factor(s) under the given experimental conditions [41]. Conversely, ChIP-seq data can be used to create gene sets against which other experimental datasets can be tested for significant enrichment, including other ChIP-seq data [42, 43].

Gene set enrichment tests can generally be classified as competitive [36, 39, 44], self-contained [31], or a hybrid [31, 45], as discussed by Efron and Tibshirani in [46]. The hypothesis of competitive approaches is that there is a higher proportion of identified genes (or a higher level of significance overall) in the gene set of interest than in the remaining genes. In contrast, self-contained methods only use information about the genes in the gene set of interest, and test whether the significance level of the set is greater than expected given a null hypothesis. The enrichment testing methods used for sets of genomic regions (ChIP-seq data), including FET and binomial based tests, are

Chapter 2 is published as Welch RP*, **Lee C***, Imbriano PM, Patil S, Weymouth TE, Smith RA, Scott LJ, Sartor MA: ChIP-Enrich: gene set enrichment testing for ChIP-seq data. Nucleic Acids Research 2014, 42(13).

all competitive approaches [40, 47].

Fisher's exact test (FET), and slight variations on it, have traditionally been used for gene set enrichment in microarray gene expression data [39, 48-51]. FET makes the assumption that each gene has an equal probability of being identified as significant. Across gene sets, this means that each gene set is expected under the null hypothesis to have approximately the same proportion of significant genes as the overall proportion of significant genes. In contrast to microarray data, the data generated from ChIP-seq, RNA-Seq, and genome-wide association studies (GWAS), often show a positive correlation between the length of the relevant genomic region and detection of the gene [9, 52, 53]. In ChIP-seq data, the probability of a peak occurring within a gene or its surrounding non-coding sequence, which together we denote as the gene locus, is often positively correlated with the length of the locus [54]. Due to this correlation, genes with longer locus lengths contribute a disproportionate amount to the enrichment signal, and this bias introduced in the signal due to gene locus length violates the assumptions of FET. Furthermore, because many commonly tested gene sets contain genes with substantially longer (e.g., developmentally and nervous system related genes) or shorter (e.g., electron transport, rRNA processing) than average locus length [54], the gene sets with longer or shorter than average locus length are more or less likely, respectively to be detected as significantly enriched [53]. Therefore, lack of effective adjustment for gene locus length can lead to false positive findings.

Several approaches have been developed to adjust for locus length in ChIP-seq [47], RNA-Seq (for example, GOseq) [9], and GWAS data [52, 55]. For ChIP-seq data, a commonly used binomial-based test asks if the total number of peaks within the loci in a gene set is greater than expected, given the total locus length of the gene set, the total number of peaks and the corresponding length of the genome (implemented in Genomic Regions Enrichment of Annotations Tool (GREAT)) [47, 53]. In contrast to FET, the assumptions of the binomial test are met when the number of peaks in a locus is proportional to locus length, and the variability of peak counts among genes, given gene locus length, is consistent with that expected by the binomial distribution.

We examined the gene locus length-to-peak presence relationships in 63 ENCODE ChIP-seq GM12878 datasets and found they ranged from no relationship to strongly positively correlated. Given these observations, our goal was to develop a gene set enrichment method for ChIP-seq data (ChIP-Enrich) that empirically models and adjusts for the relationship between locus length and peak presence. ChIP-Enrich maintained the expected type I error rate (false positive rate) in all tested datasets, whereas FET and the binomial test did not. For each DNA binding protein (DBP), we asked if different (potential) regulatory region definitions would identify different enriched/disenriched gene sets. For the glucocorticoid receptor α (GRα), we examined the ability of ChIP-Enrich to detect known and potentially novel functions. Our method is freely available in the R Bioconductor package *chipenrich* and as a web-based program (http://chip-enrich.med.umich.edu).

2.2 Materials and Methods

2.2.1 Experimental ChIP-seq peak datasets

We used ENCODE ChIP-seq peak datasets from 63 DNA binding proteins for cell line GM12878 [56] (see http://chip-enrich.med.umich.edu/SummaryEncode.jsp) (Supplementary Table 2.1). We used the existing peak calls, which were called by the original authors using one or two of three peak calling methods (MACS, spp or Scripture [57-59]). For the subset of datasets that were called by two callers (MACS and spp), we use results from MACS, as we generally observed a larger number of called peaks for MACS than for spp.

2.2.2 Gene loci definitions and presence of peaks in a locus

We define a gene as the region between the furthest upstream transcription start site (TSS) and furthest downstream transcription end site (TES) for that gene. The positions of the TSSs and TESs for each gene were extracted from the UCSC knownGene table (human genome build hg19). We removed small nuclear RNAs as they are likely to have different regulatory mechanisms than other genes and often reside within the boundaries of other genes. For gene set enrichment testing we assign ChIP-seq peaks to genes (based on the peak midpoint) using three primary definitions of a gene's designated regulatory region (locus definitions). 1) *Nearest TSS:* the region between the upstream and downstream midpoints between a gene and the two adjacent genes' TSSs. This is equivalent to assigning each peak to the gene with the nearest TSS. 2) *Nearest gene*: the region from the midpoint between the TSS and the adjacent gene's TSS or TSE (whichever is closest) to the midpoint between the TES and the adjacent gene’s TSS or TES (whichever is closest). This is equivalent to assigning peaks to the nearest gene. 3) *≤1kb from TSS*: the region within 1kb of all TSSs in a gene. If TSSs from the adjacent gene(s) are less than 2kb away, we use the midpoint between the two TSSs as the boundary of the locus for each gene. Additionally we define *≤5kb from TSS*, using the same rules as we defined *≤1kb from TSS,* and we define *>10kb from TSS*, by subtracting the 10kb regions around the TSS from the *nearest TSS* locus definition. We define peak presence in a locus as ≥ 1 peak midpoint within the gene locus boundaries.

2.2.3 Gene Ontology terms

GO terms from GO molecular functions, GO cellular components, and GO biological processes were extracted from Bioconductor species specific annotation packages and the *GO.db* R package. We removed genes from each GO term that do not exist in our gene locus definitions as these genes could not have a peak assigned to them. For testing in the manuscript and in our tool, we exclude GO terms with <10 genes as they have more limited power to detect significant results, and as a rule of thumb logistic regression requires at least 10 events for each explaining variable [60]. In the manuscript, we also exclude reporting GO terms with >500 genes, as the categories become broader and less informative in interpreting the results. Q-values were calculated using all GO terms with 10-2000 genes (our tool’s defaults).

2.2.4 Overdispersion test of peak count (given locus length) in each gene set

Overdispersion is defined as higher variability in a data set than expected based on the distribution used to model it. The binomial test in GREAT uses a binomial distribution to model the combined number of peaks for all genes in a gene set, so if significant overdispersion in peak counts exists among genes, the binomial distribution assumption is not satisfied. We tested for overdispersion in the number of peaks per gene (given locus length) in each gene set using Tarone’s Z statistic [61]. Tarone’s Z allows better estimates of overdispersion when the binomial probabilities are close to 0 or 1 (the probabilities of having a peak for each basepair are very close to 0). We tested all gene sets with a minimum of 50 genes (as gene sets with fewer genes often do not have adequate power for this test) and a maximum of 500 genes (the maximum gene set size used throughout the paper). For each DBP, we reported the proportion of gene sets that had significantly higher variability than expected based on the binomial distribution (q-value≤0.05).

2.2.5 Mappability calculations

To estimate the mappable proportion of each gene locus for different read lengths, we first calculated base pair mappability for reads of lengths 24, 36, 40, 50, 75, and 100 base pairs using mappability data for *Homo sapiens* (build hg19) from the UCSC Genome Browser. The UCSC browser mappability tracks provide, for each base pair *i*, the reciprocal of the number of locations in the genome to which a read beginning at *i* and extending for read length *K* could map; a value of 1 indicates the read maps to one location in the genome, a smaller value indicates the read maps to two or more locations in the genome. We set reads with mappability <1 to 0 and calculated base pair mappability as the average read mappability of all possible reads of size *K* that include a specific base pair location, *i,*:

$$B_i = \left(\frac{1}{2K-1}\right) \sum_{j=i-K+1}^{i+(K-1)} M_j \qquad \text{(equation 1)}$$

where B_i is the mappability of base pair *i*, and M_j is the read mappability (from UCSC's mappability track) of a read of length *K* beginning at position *j*. We define gene locus mappability, *m*, as the average of all base pair mappability, B_i, values for a gene locus; each gene locus mappability score *m* represents the proportion of the gene locus that is uniquely mappable (given the read length of the data).

2.2.6 ChIP-Enrich method

We developed a logistic regression approach to test for gene set enrichment while adjusting for $\log_{10}$ mappable locus length for each gene. Suppose that for a given set of genomic regions (referred to as peaks), we have assigned each peak to a gene locus. The dependent variable is a binary vector defined as 1 if ≥1 peak is assigned to a gene's locus, and 0 if none are assigned to the gene's locus. For each gene set, the explanatory variable of interest is gene set membership, g, defined as 1 for genes in the gene set, and 0 for all other genes. Let L be the locus length, such that $m{\cdot}L$ is the mappable locus length. Let π be the probability that a gene with gene set membership g, and adjusted for mappable locus length, has ≥1 peak. Then $\pi/(1-\pi)$ are the corresponding odds that a gene, given g = 0 or 1 and mappable locus length $m{\cdot}L$, has ≥1 peak. If the log-odds differ by gene set membership adjusted for (mappable) locus length, then we conclude that peak presence is associated with the gene set. Our model is:

$$\log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 g + f\left(\log_{10}(mL+1)\right) \qquad \text{(equation 2)}$$

where β_0 is the intercept, β_1 is the coefficient of interest, and the function $f(log_{10}(m{\cdot}L+1))$ is a binomial cubic smoothing spline term that adjusts for $\log_{10}$ mappable locus length (or $\log_{10}$ locus length if m is omitted). We apply the $\log_{10}$ transformation to locus length as this improves the model fit (data not shown). The smoothing spline is estimated with a penalized spline using a cubic spline basis fit with 10 knots distributed evenly throughout the data. Placing a knot at each data point as in a true smoothing spline would not be computationally feasible. The model is fit using penalized likelihood maximization, where the smoothing penalty is the squared second derivative penalty, and generalized cross-validation is used to choose the optimal value for the smoothing parameter, λ [62, 63]. We use the *gam* function of the R package *mgcv* to fit the model [64], and the Wald statistic to test for significance of the gene set term, β_1,which is calculated as:

$$W = \left(\frac{\hat{\beta}_1}{s_{\hat{\beta}1}}\right)^2 \qquad \text{(equation 3)}$$

where $\hat{\beta}_1$ is the penalized maximum likelihood estimate for β_1, and $s_{\hat{\beta}1}$ is the standard error for $\hat{\beta}_1$. W is distributed as X^2 with one degree of freedom under the null hypothesis β_1=0, and p-values are calculated accordingly for the alternative hypothesis, $\beta_1 \neq 0$. P-values for the gene sets are corrected for multiple testing using the Benjamini-Hochberg false discovery rate approach [65]. To be included in the analysis, genes had to be annotated in GO and have a locus defined. For example, we have 19,051 human genes with the *nearest TSS* locus definition and 16,653 (87.4%) of these genes have ≥ 1 GO term annotation (with no restriction for GO term size).

2.2.7 R package and website

Our ChIP-Enrich gene set enrichment testing method is implemented in the *chipenrich* package for the R statistical software environment and available through Bioconductor, and as a web version at http://chip-enrich.med.umich.edu/. We also provide Fisher's exact test as an alternative enrichment method. In addition to Gene Ontology, we include 12 additional annotation sources containing over 20,000 total gene sets [35]. We currently support the human genome (hg19), mouse genome (mm9, mm10), and rat genome (rn4). Precomputed mappability is available for hg19 (for read lengths specified above) and for mm9 (read lengths 36, 40, 50, 75, and 100 base pairs). Users may either supply an R data frame (for the R package) or a BED format file containing the peak locations as input. Runtime is typically 10-14 minutes for testing all GO terms but varies depending on the dataset, number of cores, and choice of locus definition. In addition to the *nearest TSS, nearest gene, ≤1kb from TSS* and *≤5kb from TSS* locus definitions, described above, in ChIP-Enrich we also offer *Exons*: peaks are assigned to gene exons, ignoring all peaks outside of an exon. Users may also supply their own custom locus definition and/or mappability file. This enables users to study functional binding patterns relative to alternative gene features (e.g., 3'UTRs) or at different distances from TSSs, and to use different estimates of the observable region for each gene locus. Diagnostic plots are available to visualize the relationship between locus length and proportion of genes with a peak, and to examine the proportion of peaks binding proximal or distal to TSSs. We also offer an ENCODE ChIP-Enrich Results website (http://chip-enrich.med.umich.edu/summaryReport.jsp), where users can download enrichment testing results for individual DBPs or in bulk for the GM12878 and K562 cell lines.

2.2.8 Fisher's exact test for gene set enrichment testing of ChIP-seq data

For each GO term, we tested for association of peak presence and GO term membership using a two-sided Fisher's exact test. For inclusion in the analysis, genes had to be annotated in GO and have a locus defined.

2.2.9 Binomial test for GO term enrichment testing of ChIP-seq data

We used a slight modification of the one-sided binomial test for GO term enrichment described by Taher et al (2009) [53] and implemented in GREAT [47]. We calculate the one-sided probability of seeing greater than or equal to the number of peaks we observe for a GO term, π, with the following formula:

$$\sum_{i=k_\pi}^{n} \binom{n}{i} p_\pi^i (1 - p_\pi)^{n-i} \qquad \text{(equation 4)}$$

where n is the total number of peaks within gene loci present in any GO term, and k_π is the number of peaks annotated to GO term π. We define p_π as the expected proportion of peaks in GO term *π*, as the total non-gapped length of the gene loci in the GO term, divided by the total non-gapped length of loci with ≥1 GO term annotation. P-values are calculated as the probability of observing k_π or greater number of peaks in the GO term. Our implementation is consistent with other GO term enrichment programs which restrict the background gene set to those annotated in GO [66]. In contrast, GREAT uses the total non-gapped genome as the denominator for p_π and defines n as all observed peaks.

2.2.10 Permutations to create ENCODE ChIP-Seq data with no biological enrichment

We performed permutations to assess the behavior of each enrichment test under two null scenarios of no true enrichment. For both scenarios, we used three ENCODE ChIP-seq datasets from cell line GM12878: SIX5 (Figure 2.1a,d), PAX5 (Figure 2.1b,e), and H3K27me3 (Figure 2.1c,f). For each of the two permutation scenarios below, we perform 300 permutations and test each permuted dataset for GO term enrichment (5519 GO terms) using the three tests (ChIP-Enrich, Fisher's exact test, and the binomial test).

Under both scenarios, we do not expect to detect enrichment, as we have removed any association between gene membership in GO terms, and the count of peaks. To help visualize the two permutation scenarios, consider a data table, where each row represents a gene, and contains the following columns: count of peaks per gene, locus length of each gene, and one column for each GO term containing a (0,1) indicator variable for whether the gene belongs to that GO term. In the GO term permutations scenario, we randomly permute the count of peaks per gene and the locus length as a unit. This results in a dataset where genes (identified by their peak count and locus length) have been reassigned to new GO terms and the locus length bias inherent in GO terms has been removed, but the number of genes per GO term, correlations between GO terms, and the relationship between locus length and count of peaks have all been preserved. In the GO term permutation by locus length bin scenario, we first order the data by locus length and then randomly permute peak count and locus length as a unit, but restrict this permutation within successive bins of gene locus length (100 genes per bin). This is similar to the first scenario, but preserves the relationship between locus length and GO term membership.

2.2.11 GRα analysis

We applied ChIP-Enrich to ChIP-seq peaks for GRα data from the A549 cell line from Reddy et al (2009): ChIP-Seq peaks with FDR <0.02 (4,392 peaks). In Reddy et al (2009), sequence reads of 36mer length, were generated from Illumina GA1, aligned using ELAND, and peaks were called using MACS. Reddy et al. (2009) identified 209 genes as differentially expressed based on RNA-Seq data from A549 cells that were treated for 1hr with 100mM of Dexamethasone (DEX) or with 0.02% Ethanol control (EtOH). Briefly, in Reddy et al. (2009), gene expression levels were estimated using ERANGE to calculate reads per kilobase per million tags sequenced (RPKM) values, which were then adjusted for dependence of variance on expression level. A custom maximum likelihood approach was used to calculate p-values for the observed change in gene expression between DEX-treated and ethanol-treated cells. Finally, genes with FDR<0.05 were called significant [67]. Using the 209 reported differentially expressed genes, we tested for GO term enrichment (over-representation) with the R package goseq [9]. For Table 3, we pruned the list of top-ranked, enriched GO terms of closely related terms for presentation by removing terms whose parents, children, or siblings in the ontology tree were present at a higher rank in the list. We used the R package GO.db to determine relationships among GO terms.

2.3 Results

2.3.1 Observed relationship between gene locus length and presence of at least one peak in ENCODE ChIP-seq datasets

We first explore the relationship between gene locus length and the presence of a peak in 63 ENCODE ChIP-seq datasets from tier 1 cell line GM12878 [16, 56] using a binomial cubic smoothing spline to model the relationship (see Experimental ChIP-seq peak datasets and ChIP-seq method sections of Methods) [62, 63]. GM12878 is a lymphoblastoid cell line, transformed using Epstein-Barr Virus, and which has a normal karyotype. Lymphoblasts are immature cells that typically differentiate into lymphocytes, and serve as a good model for functional studies as they are known to express a wide range of metabolic pathways [68]. This exploration of ChIP-seq data is motivated by the opposing assumptions underlying FET and the binomial test: for FET that there is no association between locus length and presence of a peak, and for binomial-based tests, that the number of peaks per locus is proportional to locus length. In Figure 2.1, we assigned peaks to the gene with the nearest TSS (see Methods) and grouped the ENCODE datasets based on the total number of peaks (three equal sized groups). For datasets with the smallest number of peaks, we noticed that a large fraction of peaks were close to a TSS, and there was no or little relationship between locus length and probability of a peak (Figure 2.1a,d; n=21) which is consistent with the assumptions of FET. All were transcription factor datasets. In contrast, datasets with the largest number of peaks tended to have the smallest proportion of peaks within 1kb of a TSS and had a strong positive locus length-to-peak presence relationship (Figure 2.1,f; n=21), which is potentially consistent with the assumptions of the binomial test. Many of these datasets were histone modifications that tend to occur much more widely across the genome than TF binding. The locus length-to-peak presence patterns within datasets with intermediate numbers of peaks show larger variability and are often not consistent with either FET or the binomial test assumptions (Figure 2.1b,e).

The binomial test sums the peaks over all the genes/loci in a gene set. This summation assumes that the underlying probability of a peak per unit length is the same for each gene in the gene set (and the same for each gene not in the gene set), i.e. the variance of peak counts among genes in a gene set is no greater than expected based on the binomial distribution. We tested for variability greater than that of the binomial distribution, in GO terms containing between 50 and 500 genes. All DBPs showed a substantial proportion of GO terms with significantly (FDR<0.05) higher variability than expected, with many DBPs having over 99% of GO terms with significant extra variability (Supplementary Table 2.2) (see Overdispersion test in Methods). Thus, even DBPs that have a strong positive relationship between number of peaks and locus length (Figure 2.1f) do not satisfy the binomial test assumptions.

2.3.2 ChIP-Enrich method

Given the observed locus length-to-peak presence relationships, we sought to develop a gene set enrichment testing approach for ChIP-seq data that would empirically model this relationship (Figure 2.2). To capture different aspects of the underlying regulatory biology, we define loci based on one or more genomic features, and assign peaks in the defined genomic regions to genes (locus definitions). In this paper we use as primary locus definitions: 1) the region(s) within 1kb of every TSS of a gene (≤1kb from TSS), 2) the region between the upstream and downstream midpoints between a gene’s TSS and the adjacent genes' TSSs (nearest TSS), and 3) the gene and half the intergenic region between adjacent genes, defined by the closest TSS/TES of each gene (nearest gene) (See Gene loci definitions section of Methods for more details). Consistent with previous observations [54], genes with long locus lengths defined by the nearest TSS definition were significantly enriched for neuronal processes, development, and adhesion (Supplementary Table 2.3), while genes with short locus lengths were enriched for translation and chromatin-related processes (Supplementary Table 2.4),

We test for gene set enrichment using a logistic regression model, and adjust for the probability of a peak as a function of log10(observable locus length) using a binomial cubic smoothing spline (see ChIP-Enrich method section of Methods). Since a logistic regression model without the smoothing spline term approximately corresponds to Fisher's exact test, our model is motivated by FET while accounting for locus length. Because we observed that the average mappability of gene loci both differed substantially among genes and that many GO terms were enriched with highly or lowly mappable genes (Supplementary Text and Supplementary Figure 2.1), we also account for the average mappability of each gene locus. We calculate the proportion of each locus length that is uniquely mappable as the mappability score, and use locus length × mappability as an estimate of the observable locus length (see Mappability Methods section). Although mappability often improved the spline fit (Supplementary Figure 2.2), it had little effect on the results of these analyses. Our R package and web-based tool offer a number of additional options, including custom locus and mappability definitions (see R package and website Methods section). Thirteen gene annotation databases [35] are available for testing; for simplicity, we use GO terms to illustrate our method in our analyses below (see Gene Ontology terms Methods section).

2.3.3 Comparison of ChIP-Enrich, Fisher's exact test and the binomial test for permuted and non-permuted ENCODE datasets

To illustrate the performance of the different tests, we selected three ENCODE GM12878 DBPs with different locus length-to-peak presence relationships: SIX homeobox 5 (SIX5) (weak relationship, Figure 2.1d), paired box 5 (PAX5) (moderate positive relationship, Figure 2.1e), and trimethylation of histone 3 lysine 27 (H3K27me3) (strong positive relationship, Figure 2.1f) (Supplementary Figure 2.3). These DBPs have 75, 26, and 5% of peaks ≤1kb from a TSS (Figure 2.1a-c) and 4,442, 19,618 and 41,464 total peaks, respectively. We first tested for GO term enrichment with FET, the binomial test, and ChIP-Enrich in the original data (see Methods for implementation details of FET and the binomial test). The top ranked terms from the three tests were highly different for H3K27me3, moderately different for PAX5, and similar for SIX5 where several very strongly enriched GO terms were identified by all tests (Table 2.1 Comparison of top ranked GO terms for three DBPs from cell line GM12878 using ChIP-Enrich, FET, and the binomial test.). However, other datasets with total peaks counts similar to SIX5 (few peaks) (Figure 2.1a,d) had less agreement between the top ranked terms for ChIP-Enrich and the binomial test (data not shown).

Under the null hypothesis of no true gene set enrichment, the type I error rate for a dataset at a given threshold α is the proportion of gene sets with p-value less than α. A method with type I error rate higher than the expected α level will have an increased number of false positives. Therefore, we investigated the type I error rates for ChIP-Enrich, the binomial test, and FET. We assessed the type I error rate using two permutation scenarios that preserve the GO term correlation structure but under which no biological enrichment exists, and therefore none should be detected. In the first scenario, we grouped gene locus length and gene peak count and permuted them together across all genes, which removes any relationship between GO term membership and locus length (permutations across all genes). In the second scenario, we grouped locus length and gene peak count and permuted them together within bins of 100 genes ordered by locus length, which retains the GO term-locus length relationship (permutations within locus length bins) (see Permutations Methods section).

In the permutations across all genes, ChIP-Enrich and FET showed slightly conservative type I error for both permutation scenarios at α=0.05 and 0.001 (Table 2.2 and Supplementary text), with the slight deflation expected due to the discrete nature of the data [69]. The lack of inflation for FET was expected since this permutation breaks the GO term-locus length relationship. In contrast, the binomial test had very high type I error rates at all three tested alpha levels (Table 2.2).

For the permutations within locus length bins, ChIP-Enrich again had the expected type I error rate (Table 2.2). FET showed inflation of type I error rates for PAX5 and H3K27me3, but not for SIX5. SIX5 shows little relationship between locus length and peak presence, and therefore the assumptions for FET are approximately satisfied. As a check of the ChIP-Enrich method, we compared the –log10(p-values) in the original SIX5 data and found they were highly correlated between ChIP-Enrich and FET (Pearson’s r=0.97), illustrating that in this case ChIP-Enrich closely approximates Fisher’s exact test. The binomial test again had very high type I error rates for every DBP, with particularly high error for H3k27me3 (minimum permuted p-value = 1 × 10-57). Using the binomial test we observed 761 gene sets with p<0.001 in the original

H3k27me3 data, compared to a median of 618 for the permutated data, implying that most of the significant results for the original H3k27me3 data are false positives. For SIX5 using permutations within locus length bin, >75% of gene sets with short average locus lengths had p-values <0.05 with the binomial test, whereas nearly all the gene sets with long average locus lengths had p-values>0.9. The binomial model assumes that genes with longer locus length will have proportionally more peaks, which is not satisfied in the SIX5 data (Supplementary Figure 2.4a). We observed the same behavior using the GREAT program (Supplementary text and Supplementary Figure 2.5), but not for ChIP-Enrich (Supplementary Figure 2.4b). To see whether the bias in ranks based on locus length for the binomial test carried over from the permuted to the original unpermuted data, we asked if the ranks for original and permuted SIX5 datasets were correlated. We observed a high correlation for the binomial test (r =0.71) between the ranks of results from the original SIX5 data and the average ranks from permutations within locus length bins, but not for permutations across all genes (Supplementary Figure 2.6a,b), indicating that the correlation is due to locus length. With ChIP-Enrich, there was no correlation between ranks of the original and permuted data (r =-0.02) as expected (Supplementary Figure 2.6c,d).

To complement our permutation study, we also simulated ChIP-seq peak datasets with no true biological enrichment under various scenarios and tested for enrichment with ChIP-Enrich, the binomial test, and FET. In these simulations, the binomial test had an inflated type I error rate when peak counts were not proportional to locus length or when extra-variability (overdispersion) was added to gene peak counts. Only ChIP-Enrich showed the expected type I error rate in all simulations (Supplementary text and Supplementary Figure 2.7, Supplementary Figure 2.8).

2.3.4 Influence of locus definition on detection of gene set enrichment

For each of the 63 GM12878 ChIP-seq datasets, we asked if dissimilar sets of biologically-related genes were detected using different locus definitions, as a way to identify DBPs that regulate distinct biological functions from different regulatory regions. Comparing ChIP-Enrich results for peaks assigned to the *nearest TSS* to those of the *nearest gene*, we found moderate to high correlations in the enrichment results

(Pearson's r=0.62-0.99 for $-\log_{10}$ p-values) and p-values of similar magnitude, indicating that the two definitions are capturing similar information.

We observed much greater variability in comparisons between the *≤1kb from TSS* and *nearest TSS* locus definitions, with four distinct patterns emerging (Figure 2.3 and Supplementary Figure 2.9). 1) We found similar results for *≤1kb from TSS* and *nearest TSS* for DBPs that tend to bind near TSSs, such as SIX5 (Figure 2.3a), and for a subset of other DBPs (Supplementary Figure 2.9). 2) We identified distinct GO terms for *≤1kb from TSS* and *nearest TSS* for JunD and a small number of other DBPs (Figure 2.3b). JunD showed strong enrichment for calcium ion-related terms only within 1kb of a TSS and enrichment for the JNK and MAPK cascades only using *nearest TSS* (not shown). JunD regulates varied physiological processes [70]; these results suggest it may regulate different processes from near versus far TSSs. 3) We identified much stronger enrichment using *nearest TSS* than *≤1kb from TSS* for H3K36me3 (Figure 2.3c), H3k79me2 and H4k20me1 (Supplementary Figure 2.9) which bind along gene bodies [71]. 4) Finally, we saw much stronger GO term enrichment using *≤1kb from TSS* compared to *nearest TSS* for CTCF (Figure 2.3d), WHIP, and a subset of DBPs with a small percent of peaks ≤1kb from the TSS (Supplementary Figure 2.9).

Although CTCF is a well-known insulator in intergenic regions, both CTCF-binding and housekeeping genes are enriched in the boundary regions of genomic topological domains [72], and we see many of the same strongly enriched GO terms for CTCF binding ≤1kb from a TSS (RNA processing, mitochondrion, and cell cycle) as for genes identified at the boundary regions. WHIP binds to damaged DNA and in that capacity is not expected to bind within or near genes with specific functions [73, 74]. The most highly enriched gene sets for WHIP using the *≤1kb from TSS* definition included DNA repair ($p=1.1\times10^{-17}$), chromatin organization ($p=3.6\times10^{-15}$) and cell cycle regulation suggesting transcriptional roles of WHIP related to its direct function in DNA repair. Other DBPs with relatively small percentages of peaks near a TSS also showed stronger ≤1kb to TSS enrichment results; these have known transcriptional functions and/or involvement in DNA repair (ZNF143, CHD2) [75, 76], chromatin structure (EBF1) [77], or centromere formation (SMC3) [78], which may explain the lack of biological enrichment from more distal peaks (Supplementary Figure 2.9).

2.3.5 ChIP-Enrich analysis of the glucocorticoid receptor α (GRα)

We asked whether ChIP-Enrich could identify known and potential new biology of a well-characterized transcription factor, the glucocorticoid receptor α (GRα) [79]. Previous analysis identified 4,392 peaks in A549 cells treated with 100nM DEX (dexamethasone stimulates GR activity); only 4.7% of the peaks were within 1kb of a TSS (Figure 2.4a). GO term enrichment testing yielded largely distinct subsets of significant (FDR≤0.05) terms for nearest TSS (195 terms) and ≤1kb from TSS (72 terms) with only 16 overlapping terms (Figure 2.4b,d; Supplementary Table 2.5). The most significant terms (after collapsing similar terms) are shown in Table 2.3. Terms significant using one or both locus definitions include “epithelial cell differentiation” (q-values: nearest TSS=1.8×10-6; ≤1kb from TSS=1.0) and “negative regulation of blood coagulation” (q-values: nearest TSS=0.077, ≤1kb from TSS=3.19×10-7, with the related term “regulation of wound healing” (q-values: nearest TSS=0.0064, ≤1kb from TSS=0.0029). In addition, we observed “response to glucocorticoid stimulus” (q-values: nearest TSS=0.0035; ≤1kb from TSS=0.55) and “regulation of lipid metabolic process” (q-values: nearest TSS=0.0062, ≤1kb from TSS=0.74). GRα is known to be involved in the response to steroids and the activation of lipolysis [80, 81], although knowledge of the transcriptional role of GRα in wound healing and blood coagulation is more limited. We also tested for enrichment using non-overlapping locus definitions for regions closer to a TSS (≤5kb from TSS; 14.5% of peaks) and further from a TSS (>10kb from TSS; 75.6% of peaks) and again identified largely distinct gene sets (Supplementary Figure 2.10).

We also compared the enrichment results (using nearest TSS) from ChIP-Enrich with those using the binomial test and FET. Due to inflated type I error rates for the binomial test and FET for nearest TSS, the specific p-values and number of terms with FDR<0.05 cannot be used. Instead, we compared the top ranked terms among the methods, using the number of top ranked terms with FDR <.05 for ChIP-Enrich (195). There was substantial overlap, with 57 (29%) GO terms identified by all three methods and 150 (77%) identified by at least two (Figure 2.4c). Both FET and the binomial test had higher overlap with ChIP-Enrich than with each other, consistent with the fact that the locus length-to peak presence relationship modeled by ChIP-Enrich is intermediate between the assumptions of FET and the binomial test.

To evaluate the biological relevance of our results, we compared the ChIP-seq enrichment results from ChIP-Enrich with RNA-Seq enrichment results based on differential expression between control and 100nm DEX treated A549 cells [18] (See GRα analysis section of Methods). Of 4,544 GO terms tested for enrichment based on RNA-Seq differential expression, 458 (10%) were significant at FDR≤0.05. "Vascular development", the most significant GO term based on differential expression, was also significantly enriched for GRα binding using the nearest TSS analysis (q-value=0.0047) but not using ≤1kb from TSS (q-value=0.97). Eighty-six (29%) of the significant terms from RNA-Seq were significant with one or both of the locus definitions in ChIP-seq data (Figure 2.4d). From the ChIP-seq perspective, many of the most highly significant terms using nearest TSS and <1kb from TSS were significant for RNA-Seq (Table 2.3, Figure 2.4e,f). Seventy-two (37%) of the significant GO terms for nearest TSS were significant for RNA-Seq, whereas only 20 (28%) of the significant GO terms for ≤1kb from TSS were significant for RNA-Seq, indicating potentially stronger correspondence of the gene expression data with the GRα peaks captured by the nearest TSS definition than only those peaks ≤1kb from a TSS. Correlations with RNA-Seq results using a custom >10kb from TSS locus definition (see Gene loci definitions Methods section) were similar to nearest TSS and those for ≤5kb from TSS were similar to ≤1kb from TSS (not shown). GO terms enriched only in RNA-seq may be regulated by genes downstream of those directly regulated by GRα or be GRα-independent DEX effects. GO terms enriched only in ChIP-seq data may indicate pathways that are poised to be regulated, either from proximal promoter or more distal enhancer regions.

2.4 Discussion

We developed a gene set enrichment testing method for ChIP-seq data, ChIP-Enrich, that empirically models and adjusts for the effect of gene locus length. In contrast to Fisher's exact and the binomial test, ChIP-Enrich maintains the correct type I error rate for datasets with a wide range of locus length-to-peak presence relationships. FET and the binomial test make assumptions that are inconsistent with the observed relationships, which can lead to inflated type I error rates (false positive results). Strikingly, the binomial test often has significantly more false positives than FET.

ChIP-Enrich uses a binomial smoothing spline to empirically model the relationship with gene locus length; an approach similar to that employed by GOseq, which was developed for RNA-seq data [9]. Whereas GOseq uses either a resampling approach or the approximate Wallenius method to calculate GO term enrichment p-values, ChIP-Enrich incorporates the smoothing spline in a logistic regression model, allowing more precise p-value calculations and in less time than a resampling approach requires. Compared to the Wallenius approximation, ChIP-Enrich has greater power, as determined by finding more significantly enriched GO terms (36% more on average) across the 63 GM12878 ChIP-seq datasets (data not shown).

For many DBPs, particularly those with more binding near TSSs, testing for enrichment using the nearest TSS and ≤1kb from TSS locus definitions identifies largely overlapping gene sets, suggesting the two definitions often capture similar regulatory information. However, for a subset of DBPs, these two locus definitions detect very different enriched gene sets. JunD, for example, may be regulating different biological processes nearer to and further from the TSS, possibly with different cooperating factors. For datasets with a small proportion of peaks ≤1kb of a TSS, but stronger levels of enrichment detected with those peaks (examples WHIP and CTCF), it is possible that DBP binding >1kb from the TSS may not be properly assigned to the regulated gene(s), or that some of the widespread DBP binding may not regulate genes in any specific biological processes. Thus for DBPs with unknown function, comparisons of patterns of gene set enrichment could help predict an alternative role for the DBP, such as DNA repair and/or chromatin remodeling or looping.

To further explore the biological relevance of our results, we compared the gene sets enriched for differential expression of mRNA following activation of GRα to the gene sets enriched for GRα binding [79]. For GRα a subset of gene sets, many of which were not detected using the ≤1kb from TSS locus definition and including vasculature development, showed substantial enrichment for both differential expression and GRα binding. GRα has been reported to play a limited role in vasculature development, mainly through non-transcription factor activities; the extent to which it directly regulates vasculature development genes as a group was thus far unknown [82-84]. This suggests that GRα regulates many genes and functions via binding further from TSSs, consistent with the observations of Reddy et al (2009) [79], and this regulation would be missed if only peaks within 1kb were examined (such as could be tested without bias using FET).

Unlike the binomial test, ChIP-Enrich results are not influenced by a single gene or few genes with a large number of peaks. However, because higher numbers of a bound DBP in a gene locus may exert stronger biological effects (49), the use of a model based on peak counts, that accounts for extra-variability and diverse locus length-to-peak count relationships, could be considered. For example, a negative binomial or beta binomial model may be able to account for the extra variability among genes in the peak count data. However, it is unclear whether these models can fully account for both the extra variability and the observed negative correlation between peak occurrence rate and locus length, or how best to empirically adjust for locus length.

In conclusion, we developed a gene set enrichment testing method, ChIP-Enrich, which allows enrichment analysis of ChIP-seq data with any locus length-to-peak presence relationship with the expected type I error rate. This is in contrast to currently available methods, which often exhibit highly elevated type I error and/or gene set ranking biased towards genes with long or short locus length, leading to false positive results. Based on our observations, we recommend testing each set of genomic regions for enrichment with both a locus definition representing promoter regions (e.g., ≤1kb from TSS or ≤5kb from TSS) and a locus definition representing all regions or regions more distal to TSSs (e.g., nearest TSS, nearest gene, or >10kb from TSS). ChIP-Enrich can be used to further assess and refine regulatory region definitions, based on empirical exploration, and to identify biological functions of regions exhibiting various complex patterns of histone marks or protein binding using the wealth of biological data from ENCODE, the Roadmap Epigenomics Program and other public and non-public sources. With the option for user-defined locus definitions and/or mappability tracks, this framework can also be used with other genome-wide sequencing data such as RNA-

Seq (with potential bias from transcript length and/or read depth) or bisulfite sequencing data (with potential bias from number of measured CpG sites).

2.5 Supplementary Methods

2.5.1 Testing for enriched GO terms with genes of longer (or shorter) than average locus length

We used DAVID [85] to test for GO term enrichment in the top 500 genes with longest (or shortest) locus lengths. For both tests, the complete set of genes in our locus definition file was used as the background gene list. Results were limited to GO terms with ≤2,000 genes and FDR≤0.05 in order to report more specific categories.

2.5.2 Testing for enriched GO terms with genes having higher or lower than average mappability

We tested for GO terms enriched with genes having higher or lower than expected mappability scores using a logistic regression model with GO term membership as the outcome and average gene locus mappability scores as the predictor (LRpath [86]; lrpath.ncibi.org). Because LRpath typically accepts p-values as input which are then log-transformed, we exponentiated mappabililty values before input to preserve the original mappability scale. Results were limited to GO terms with ≤2,000 genes and FDR≤0.05.

2.5.3 Simulation and enrichment testing of data under the null hypothesis of no GO term enrichment

We simulated ChIP-seq peaks under the null hypothesis of no association with any GO term. As an alternative to simulating peak locations, we randomly sampled genes with replacement and set the number of times the gene was selected to the count of peaks occurring within the locus of a gene. Genes were sampled in two ways: 1) randomly (not in proportion to locus length), and 2) randomly in proportion to locus length. The first simulates peaks occurring within genes with no dependence on locus length (FET assumption). The second method simulates peaks being assigned to genes with probability in proportion to locus length (binomial test assumption). We simulated datasets of 10,000 peaks with varying percentages (0, 50 and 100%) sampled in proportion to locus length. For FET and ChIP-Enrich, a gene is labeled as having a peak if the count of peaks is ≥ 1. Each GO term was tested for enrichment using FET, the binomial test, and ChIP-Enrich. We repeated this process 1,000 times for each test and percentage of genes sampled by locus length, and calculated the median of the 1,000 simulation p-values at each quantile of the 2,565 GO term p-values to create the plots for the bottom row of Supplementary Figure 2.7.

To examine the effect of overdispersion (added variation in peak count among genes) on each of the three tests, we simulated data with 100% of peaks sampled by locus length (satisfying the binomial test assumption) but with additional overdispersion in number of peaks per gene. In the simulations above without overdispersion, we sampled genes in proportion to their locus lengths to represent ChIP-seq peaks occurring in gene loci, by assigning each gene a weight proportional to its locus length. Here, we sample genes in proportion to random deviates of their locus lengths using a gamma distribution with mean equal to the gene's locus length and variance set to one of four different levels (0, 0.01, 0.1, 0.5) to simulate increasing overdispersion. For each simulation, a weight is drawn for each gene and then 10,000 draws of genes are made based on the weights to represent 10,000 peaks. For each gamma variance level, we performed 1,000 simulations. The simulated data was tested for enriched GO terms using FET, ChIP-Enrich and the binomial test, and results were presented as median quantile p-values as above (Supplementary Figure 2.8).

Code for all simulations is available in Supplementary_code.zip.

2.5.4 GREAT testing on permuted ChIP-seq datasets

To confirm that our results in Supplementary Figure 2.4 were not restricted to our implementation of the binomial test, we repeated the analysis of GO term permutations by locus length bins data (permuted ENCODE datasets for SIX5, PAX5, and H3k27me3) with the GREAT website (Supplementary Figure 2.5). For each of the three experimental datasets, we used GREAT with the "single gene" setting, where "each gene is assigned a regulatory domain that extends in both directions to the nearest gene's TSS but no more than the maximum extension in one direction." The "maximum extension" was set to 100,000 kb in order to insure each peak is assigned to a regulatory domain (i.e. gene), which is most equivalent to our *nearest TSS* locus definition.

2.6 Supplementary Results

2.6.1 Effect of locus length and read mappability on gene set enrichment tests

Gene set enrichment testing with Fisher's exact test (FET) can be confounded when there is a positive relationship between locus length and the presence of a peak, since many gene sets contain genes with substantially longer or shorter locus lengths than average [54]. Taher and Ovcharenko (2009) identified Gene Ontology (GO) terms with much longer or shorter than expected gene loci (defining a locus as the gene and half the intergenic region between adjacent genes) [53]. Similarly, when we assigned peaks to the gene with the nearest TSS, we found that GO terms related to nucleosome, protein-DNA complexes, and translation have genes with shorter than average locus lengths (Supplementary Table 3) and nervous system development, cell adhesion, and transcription have genes with longer than average locus lengths (Supplementary Table 4).

The probability of calling a ChIP-seq peak can depend on the mappability of the reads in the binding region [87, 88]. To account for mappability (which similar to locus length, varied significantly by gene set (Supplementary Figure 2.1)), we use locus length × mappability as the observable locus length in our analyses; this can improve the spline fit (Supplementary Figure 2.2), although it had little effect on the final results of the presented enrichment testing analyses (*data not shown*). To assess the ability of mappability to confound the relationship between the presence of a peak and gene set membership, we estimated the average mappability of each gene locus based on base pair mappabilities for 50bp reads (see Supplementary Figure 2.1a for comparison of mappability at different read lengths). Genes with less mappable loci are significantly more likely to be present in sensory, xenobiotic response and oxygen related terms, whereas genes with highly mappable loci are more likely to be involved in nervous system or development terms (q-value < $3.0\text{x}10^{-16}$) (Supplementary Figure 2.1b,c). We observed similar results at other read lengths (*data not shown*). Several GO terms (e.g., central nervous system development) had both longer locus lengths and higher mappability, increasing the possibility for confounding.

In addition to mappability, GC content has also been noted to influence sequencability and thus detection of ChIP-seq peaks. To examine a potential bias due to GC content, we downloaded the UCSC Genome Browser's GC content track for hg19, which provides GC content for every 5bp. We calculated average GC content for four definitions of gene loci (*nearest TSS*, *exons*, *1kb* and *5kb*), and observed very little spread in the distribution of GC content. Testing for GO terms enriched with low or high GC content genes (using the *nearest TSS* definition and the same LRpath approach as used for testing low or high mappability genes), we found only 14 significant terms as compared to 717 significant terms for mappability (FDR<0.05). Given the tight distribution of GC content among gene loci and the small number of significantly associated GO terms, we conclude that it is unlikely that GC content is a confounding variable or significantly biases the enrichment testing results.

2.6.2 Comparison of ChIP-Enrich, Fisher's exact test, and the binomial test under the null hypothesis of no enrichment using simulated data

To examine the sensitivity of each test to varying mixtures of peak distributions that meet the FET or binomial test assumptions, we simulated datasets of 10,000 peaks with 0%, 50%, and 100% of the peaks simulated in proportion to locus length relative to those simulated irrespective of locus length. As the percentage of peaks simulated in proportion to locus length increases from 0 to 100%, the relationship between the probability of a gene having at least one peak and locus length changes from flat (Supplementary Figure 2.7a) to increasingly correlated (Supplementary Figure 2.7b,c).

Using our simulated datasets, we tested for GO term enrichment and plotted the observed $-\log_{10}$(p-values) versus the expected $-\log_{10}$(p-values) under the null hypothesis of no enrichment in quantile-quantile (QQ) plots (Supplementary Figure 2.7d-f.) For all three scenarios, ChIP-Enrich shows no inflation of significance levels from the expected distribution but has a slight deflation of the most significant p-values. When all peaks are simulated with each gene having equal probability of having a peak

(0% proportional to locus length), Fisher's exact test shows the expected distribution of p-values (observed = expected) also with a slight deflation of the most significant p-values similar to ChIP-Enrich, expected due to the discrete nature of the data [89] . With an increasing percentage of peaks sampled in proportion to locus length, FET becomes increasingly anti-conservative (Supplementary Figure 2.7d → e → f), such that p-values as low as 10^{-10} are observed in the absence of any true enrichment. The binomial test shows the opposite behavior; when peaks are sampled in proportion to locus length (100% proportional to locus length) and without any additional variability among genes in a gene set, the binomial test has the expected p-value distribution (again with a slight deflation as for FET when 0% random) but when peaks are sampled independent of locus length (0% proportional to locus length) the test becomes increasingly anti-conservative (Supplementary Figure 2.7f → e → d), with even lower p-values than observed for Fisher's exact test.

2.6.3 Test behaviors in the presence of overdispersion of peak counts among genes, given locus length

To better understand the difference in binomial test behavior between 1) the H3K27me3 dataset GO term permutation by locus length bin (which shows a strong inflation of significance levels despite peaks occurring approximately in proportion to locus length) and 2) simulations in which 100% of peaks were simulated in proportion to locus length (Supplementary Figure 2.8f; which shows no inflation of significance levels when peaks occur in proportion to locus length), we performed additional simulations with 100% of the peaks simulated in proportion to locus length. In these simulations, we added increasing levels of extra variability (overdispersion) in peak counts among genes (gamma variance levels of 0.01, 0.1, and 0.5). The overdispersion did not visually change the observed spline fit (*not shown*). Again, we tested for GO term enrichment and plotted the observed $-\log_{10}$(p-values) versus the expected $-\log_{10}$(p-values) in QQ plots (Supplementary Figure 2.8). As before, ChIP-Enrich shows no inflation of significance levels. The binomial test, however, shows increasing inflation of significance levels with increasing overdispersion. FET shows decreasing levels of inflation with increasing overdispersion, but remained biased for each overdispersion scenario.

2.6.4 Slight deflation in p-values compared to what is expected under the null

When the assumptions of Fisher's exact test are met, e.g. for the transcription factor SIX5, Fisher's exact test shows a slight deflation of the most significant p-values compared to what is expected if we assume a uniform distribution of p-values under the null hypothesis. This trend is expected due to the discrete nature of the data [89]. We observe the same slight deflation for both ChIP-Enrich (Figure 2.3c,g,k, and Supplementary Figure 2.7d,e,f, and 8b) and the binomial test (Supplementary Figure 2.7f and Supplementary Figure 2.8c) when the assumptions of the test are satisfied.

2.6.5 Sensitivity analysis for GR

As a sensitivity analysis we also repeated the GRα analyses with a larger set of peaks identified using a less stringent cutoff. This set contains 15,837 peaks with p-value ≤ 1 x10^{-9}, equivalent to an FDR < 0.23 (Supplementary dataset 1 from Reddy et al.) [79]. Results using *nearest TSS* with the 15,837 peaks were similar to those from the more stringent peak calling (r =0.61 for $-\log_{10}$(p-value) comparison) (see Supplementary text), with 81/216 (38%) of the significant GO terms also significant in the RNA-seq enrichment analysis. However, the *≤1kb from TSS* analysis results from the less stringent peak calling identified only 26 GO terms compared to 72 from the more stringent peak calling with only 8 GO termsin common. Only five (19%) of the 26 GO terms were also significant in the RNA-seq GOseq enrichment analysis, consistent with our finding in the main text that there is potentially stronger correspondence of gene expression data with GRα binding captured by *nearest TSS* (mainly distal regions) than only those peaks ≤1kb from a TSS.

2.7 Figures

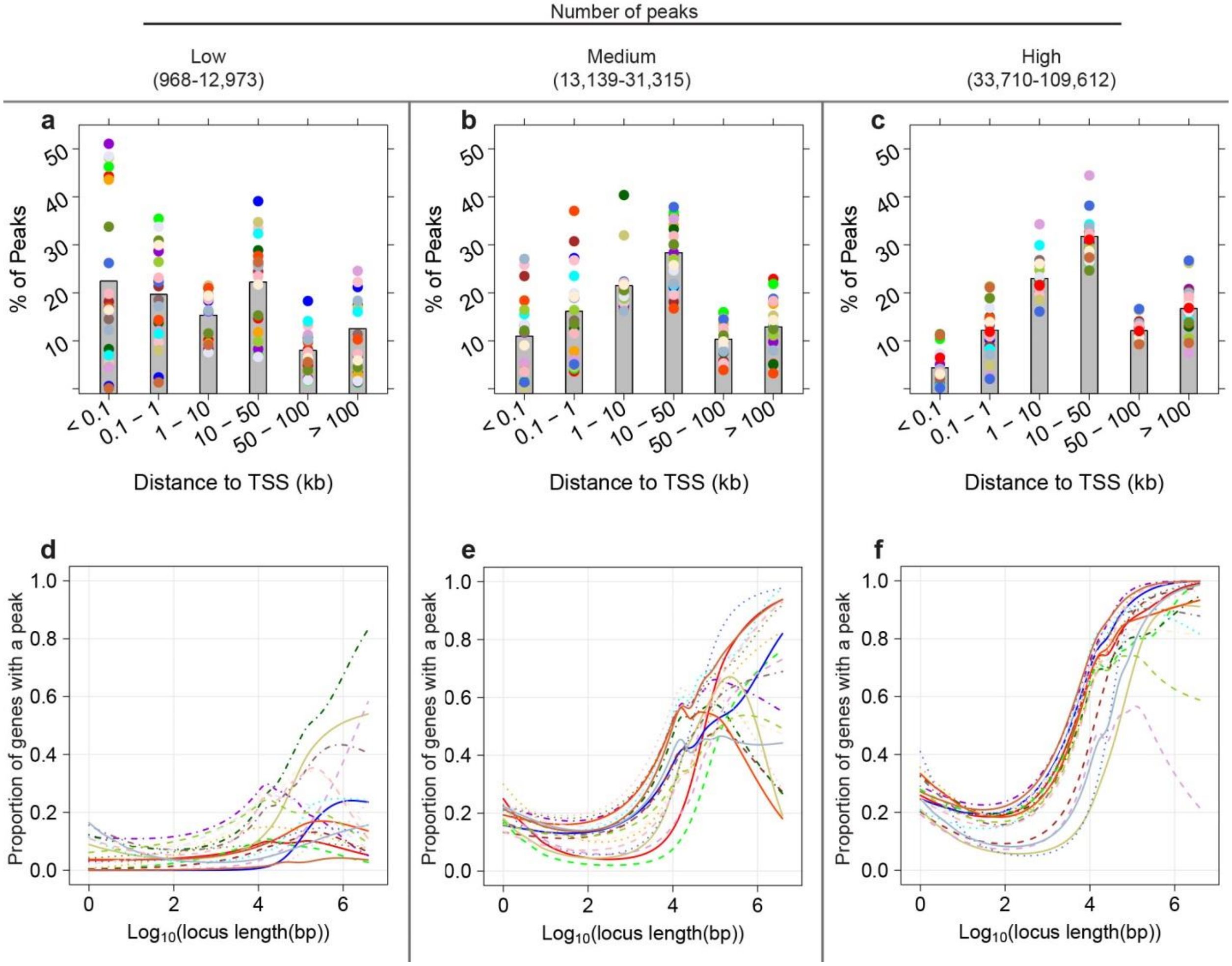
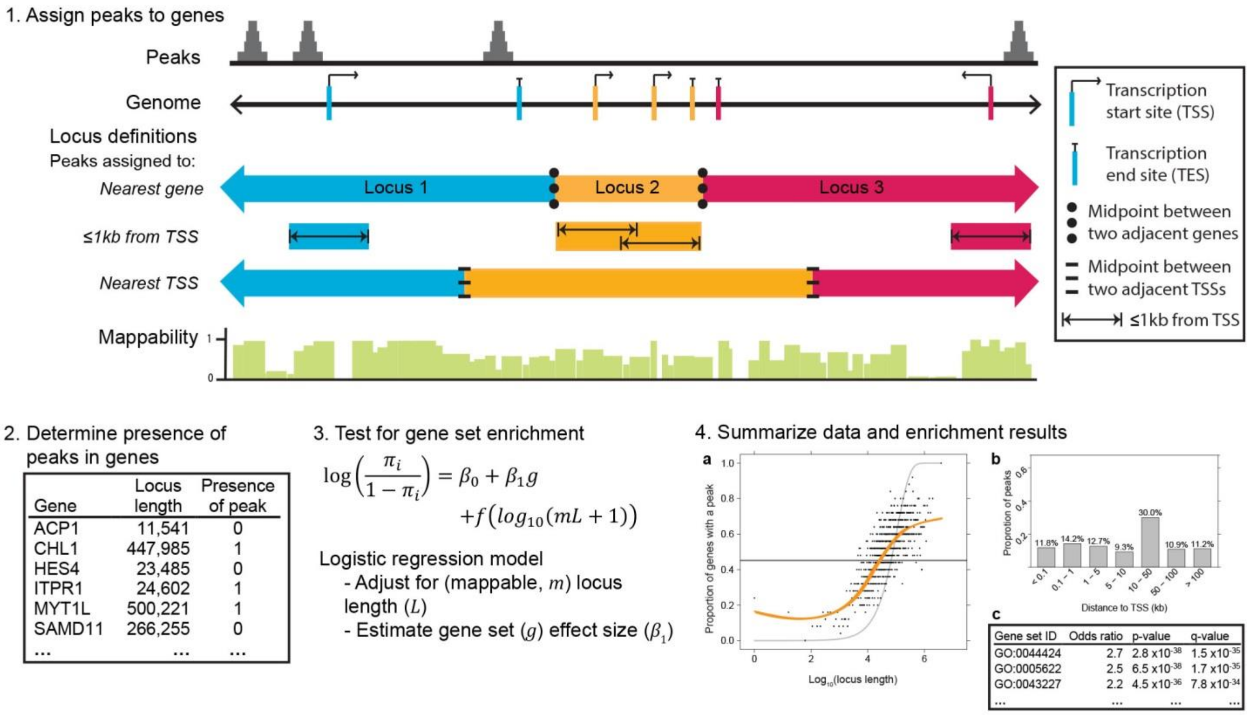


Figure 2.1 Gene locus length-to-peak presence relationship becomes stronger as total number of peaks increases. The relationship between gene locus length and proportion of genes with ≥1 peak in a gene locus varies widely in 63 ENCODE ChIP-seq datasets, from no relationship to strongly positive. DNA binding proteins (DBPs) from the GM12878 cell line were categorized into three groups of 21 DBPs by the total number of peaks. For each DBP, the relationship between log10 locus length and proportion of genes with a peak was modeled using a binomial cubic smoothing spline (see Methods). (a-c) Barplots show the average proportion of peaks present within the specified distance from the TSS (kb) (gray bar) and the proportions for individual DBPs (colored dots, same color as line in the corresponding plot). DBPs with fewer peaks tend to have a higher proportion of binding close to TSSs. (d) The locus length-to peak presence relationship tends to be weak for datasets with few peaks. (e-f) The relationship becomes strongest when the number of peaks is highest (f). None of the DBPs in (d), two of the DBPs in (e) and 10 of the DBPs in (f) are histone modifications.



Gene	Locus length	Presence of peak
ACP1	11,541	0
CHL1	447,985	1
HES4	23,485	0
ITPR1	24,602	1
MYT1L	500,221	1
SAMD11	266,255	0
...	...	...

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 g + f\left(log_{10}(mL+1)\right)$$

Logistic regression model
- Adjust for (mappable, m) locus length (L)
- Estimate gene set (g) effect size (β_1)

Gene set ID	Odds ratio	p-value	q-value
GO:0044424	2.7	2.8×10^{-38}	1.5×10^{-35}
GO:0005622	2.5	6.5×10^{-38}	1.7×10^{-35}
GO:0043227	2.2	4.5×10^{-36}	7.8×10^{-34}
...	...	...	...

Figure 2.2. Overview of ChIP-Enrich. We describe ChIP-Enrich in four steps. (1) ChIP-seq peaks are assigned to genes using a chosen gene locus definition. Definitions include: nearest gene, ≤1kb from TSS and nearest TSS. (2) It is determined whether ≥ 1 peak is present in each gene locus. (3) Gene set enrichment is performed for each gene set using a logistic regression model, adjusting for locus length with a binomial cubic smoothing spline term (represented as f in the model equation.) (4) Data and results are summarized. a) Plot of observed spline fit for log10 locus length versus proportion of genes with a peak (orange). Expected line if no relationship between log10 locus length and proportion of genes with a peak (dark gray, satisfies Fisher's exact test assumptions). Expected line if number of peaks observed is proportional to locus length (light gray, binomial test assumption). For visualization only, each point is the proportion of genes assigned a peak within sequential bins of 25 genes; b) Barplot of proportion of peaks found at various distances from the TSS; c) Abbreviated ChIP-Enrich output.

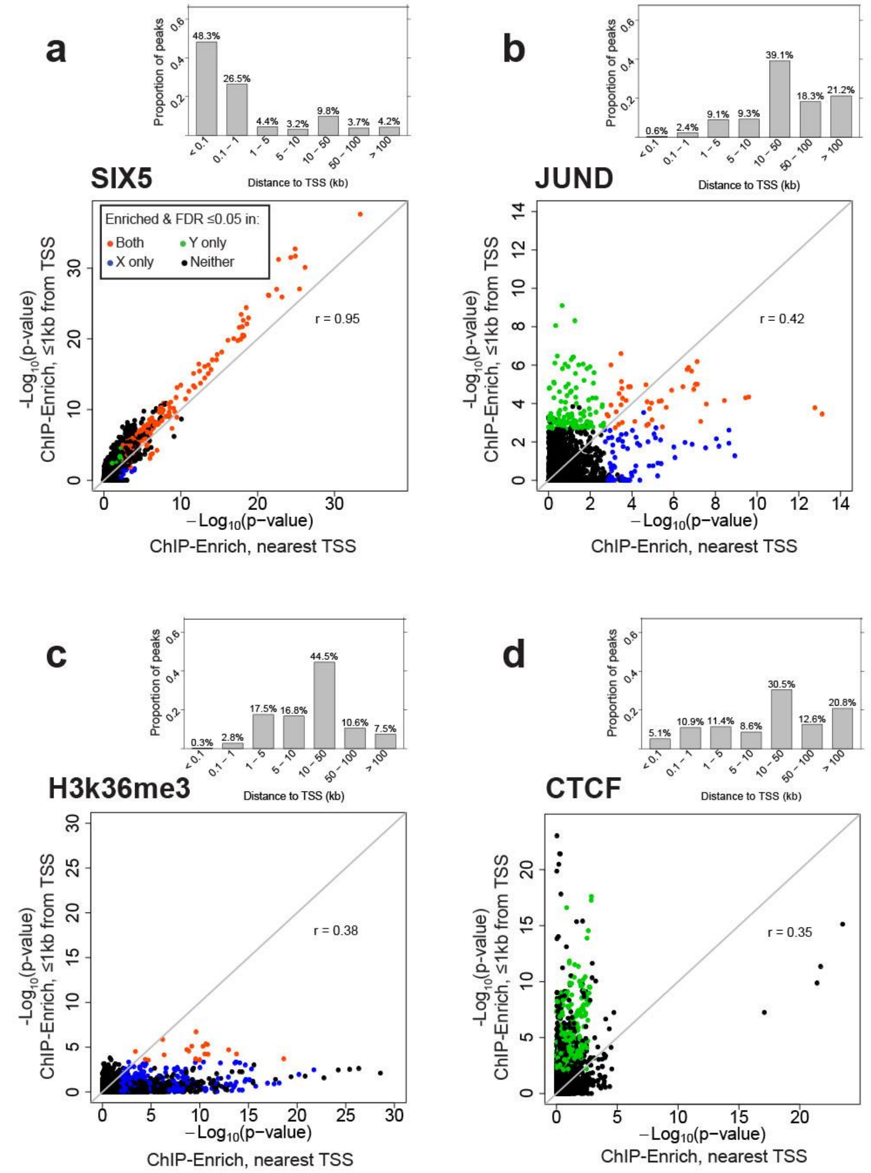


Figure 2.3. Representative plots of the 4 patterns of enrichment comparing the ≤1kb from TSS and nearest TSS locus definitions. Gene set enrichment testing using the ≤1kb from TSS and nearest TSS locus definitions may identify similar (a) or different (b) sets of significant GO terms for the same DBP. Alternatively, most of the enrichment signal may come from nearest TSS which uses all peaks (c) or ≤1kb from TSS which ignores peaks >1kb from a TSS (d). (a-d) Upper plot: Barplot of proportion of peaks at different distances from the TSS. Lower plot: Comparison of –log10(p-values) from ChIP-Enrich GO term enrichment testing using ≤ 1kb from TSS versus nearest TSS locus definitions in ENCODE data for the GM12878 cell line. GO terms enriched with FDR ≤0.05 for: ≤1kb from TSS only (green); nearest TSS only (blue); ≤1kb from TSS and nearest TSS (orange); neither analysis (black). r, Pearson correlation coefficient. These patterns are representative of patterns present in 63 ENCODE DBPs from the GM12878 cell line.

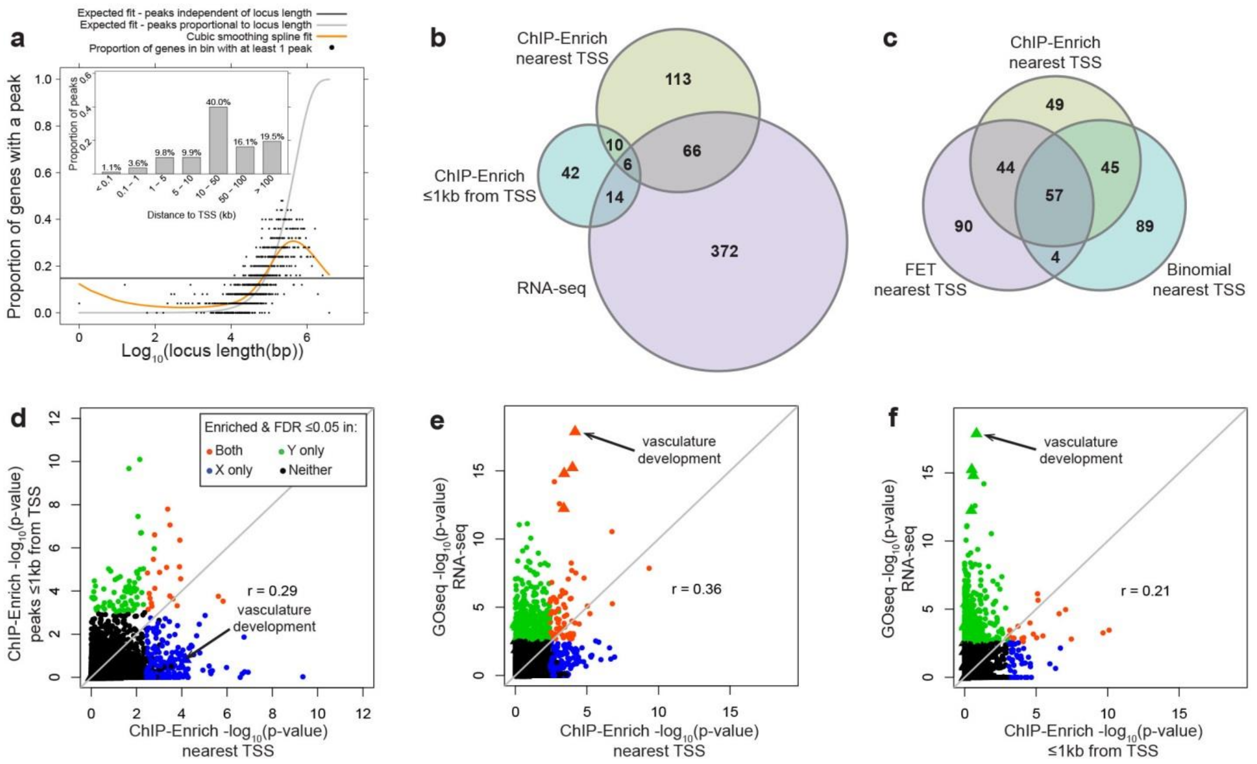


Figure 2.4 Comparison of GRα enrichment results for ChIP-seq (using two locus definitions) and RNA-seq data from A549 cells. Enriched GO terms for differentially expressed transcripts and GRα binding following 100nM DEX treatment show stronger overlap using the *nearest TSS* locus definition than using the *≤1kb from TSS* definition*.* (a) Observed spline fit for GRα fits neither FET nor the binomial test assumption (orange); barplot of proportion of peaks at different distances from the TSS. See Fig 2.4.a and b for further details. (b) Using the *nearest TSS* locus definition with GRα results in more overlapping terms with RNA-seq results than using *≤1kb from TSS* (c) Using the top 195 ranked terms for each test, FET and the binomial test have more overlap with ChIP-Enrich than with each other. (d-f) Comparison of –log$_{10}$(p-values) for GO term enrichment tests based on ChIP-seq data (ChIP-Enrich) and/or RNA-seq (GOseq) data. (f) Many enriched RNA-seq terms would have been missed in the ChIP-seq data if only peaks in promoter regions were considered. GO terms enriched and FDR ≤0.05: for Y-axis test only (green); for X-axis test only (blue); for X and Y-axis tests (orange); for neither (black). Vasculature development and related GO terms (triangles). The majority of GO terms that overlap between *≤1kb from TSS* and *nearest TSS* are related to fatty acid metabolism, reactive oxygen species and unfolded proteins, or blood coagulation.

2.8 Tables

Table 2.1 Comparison of top ranked GO terms for three DBPs from cell line GM12878 using ChIP-Enrich, FET, and the binomial test. (a) H3K27me3, (b) PAX5, and (c) SIX5. The most extreme differences are observed for H3K27me3, which also had the highest type I error rate for the binomial test. Differences among the tests are more moderate for PAX5. SIX5 had several extremely significant GO terms with ChIP-Enrich, which were also easily detected by the other two methods. All tests were performed using the nearest TSS locus definition. CE=ChIP-Enrich; Binom=binomial test; FET = Fisher's exact test.

a H3k27me3

CE rank	Binom Rank	FET rank	GO term	CE q-value	Binom q-value	FET q-value	GO term avg locus length %ile*
1	898	**1**	extracellular matrix	1.5x10^{-9}	0.013	2.2x10^{-20}	69.6
2	14	**4**	regulation of hormone levels	3.3x10^{-7}	4.4x10^{-16}	3.9x10^{-13}	58
3	1633	**3**	proteinaceous extracellular matrix	3.8x10^{-7}	0.15	9.2x10^{-17}	70.4
4	648	311	cytokine activity	2.7x10^{-6}	2.6x10^{-3}	1.2x10^{-3}	20.9
5	1137	122	anchored to membrane	2.9x10^{-6}	0.036	1.6x10^{-5}	88.2
691	**1**	1066	3',5'-cyclic-GMP phosphodiesterase activity	0.28	9.8x10^{-32}	0.089	52.1
986	**2**	3715	IgG binding	0.41	1.9x10^{-26}	0.77	1.8
256	**3**	2696	pancreatic ribonuclease activity	0.095	1.10x10^{-24}	0.77	0.1
3537	**4**	3186	cytoplasmic dynein complex	0.87	9.28x10^{-23}	1	37.4
2842	**5**	3049	localization within membrane	0.99	2.6x10^{-21}	0.92	42
14	4946	**2**	synapse	1.7x10^{-4}	1.0	3.6x10^{-17}	91.6
21	1250	**5**	sensory organ development	8.7x10^{-4}	0.053	4.6x10^{-13}	77.8

*Average locus length percentile for the top 20 terms for: ChIP-Enrich- 59.1; binomial test- 41.6; FET- 82.2.

b PAX5

CE rank	Binom rank	FET Rank	GO term	CE q-value	Binom q-value	FET q-value	GO term avg locus length %ile*
1	6	**2**	immune response-regulating signaling pathway	1.4x10^{-7}	1.1x10^{-53}	4.5x10^{-10}	39.6
2	**4**	**1**	immune response-activating signal transduction	1.5x10^{-7}	3.0x10^{-54}	4.5x10^{-10}	39.2
3	111	13	protein localization to organelle	2.8x10^{-7}	9.0x10^{-17}	4.8x10^{-6}	27
4	13	66	viral reproduction	3.2x10^{-7}	1.4x10^{-41}	5.8x10^{-4}	9.3
5	**3**	3	leukocyte activation	5.8x10^{-7}	1.3x10^{-54}	5.0x10^{-9}	48.9
20	**1**	39	regulation of immune response	8.6x10^{-5}	5.2x10^{-74}	1.1x10^{-4}	28.5
170	**2**	405	innate immune response	0.024	4.2x10^{-61}	0.10	20.9
49	**5**	31	induction of apoptosis	5.2x10^{-4}	3.3x10^{-54}	8.6x10^{-5}	30.9
6	11	**4**	lymphocyte activation	1.3x10^{-6}	5.5x10^{-44}	9.8x10^{-9}	52
8	19	**5**	immune response-activating cell surface receptor signaling pathway	7.1x10^{-6}	8.7x10^{-37}	4.8x10^{-8}	46.7

*Average locus length percentile for the top 20 terms for ChIP-Enrich: 25.9, binomial test: 33.3, and FET: 48.6

C SIX5

CE rank	Binom rank	FET Rank	GO term	CE q-value	Binom q-value	FET q-value	GO term avg locus length %ile*
1	**1**	**1**	Ribosome	4.4x10^{-32}	1.5x10^{-60}	1.9x10^{-34}	3.4
2	**4**	**2**	structural constituent of ribosome	9.8x10^{-25}	4.2x10^{-49}	1.7x10^{-27}	3.1
3	6	**4**	establishment of protein localization to organelle	2.6x10^{-23}	1.1x10^{-43}	8.4x10^{-24}	8.6
4	28	6	mRNA processing	6.2x10^{-23}	5.8x10^{-26}	2.3x10^{-22}	22.7
5	**3**	**5**	ncRNA metabolic process	1.0x10^{-22}	1.0x10^{-49}	8.8x10^{-23}	6.6
6	**2**	6	viral reproduction	1.1x10^{-22}	1.0x10^{-52}	2.3x10^{-22}	9.3
7	**5**	**3**	ribosomal subunit	2.5x10^{-22}	1.0x10^{-46}	2.9x10^{-24}	2.6

*Average locus length percentile for the top 20 terms for ChIP-Enrich: 8.4, binomial test: 5.1, and FET: 8.0

Table 2.2. Fisher's exact test and the binomial test, but not ChIP-Enrich, show strongly inflated type I error rates. ChIP-Enrich shows the expected type I error rate in permuted ENCODE GM12878 ChIP-seq data; Fisher's exact test and the binomial test can show substantial inflation of type I error rate. Values represent the proportion of tests with p-value less than the given □ level. For both permutation scenarios (Permuted overall and permuted in locus length bins), a well-calibrated test should have type I error rate approximately equal to the □ level. The total number of tests was 300 permutations × 5519 GO terms = 1,655,700 tests. CE=ChIP-Enrich; Binom=binomial test; FET = Fisher's exact test.

		α level = 0.05			α level = 0.001			α level = 10^{-5}		
		CE	Binom	FET	CE	Binom	FET	CE	Binom	FET
Permuted across all gene	SIX5	0.038	0.11	0.038	6.2x10^{-4}	0.012	5.9x10^{-4}	6.5x10^{-5}	.0033	6.4x10^{-5}
	PAX5	0.043	0.25	0.040	4.3x10^{-4}	0.093	7.8x10^{-4}	2.8x10^{-5}	0.054	6.9x10^{-5}
	H3k27me3	0.045	0.30	0.040	4.7x10^{-4}	0.14	7.4x10^{-4}	3.9x10^{-5}	0.096	5.1x10^{-5}
Permuted within locus length bins	SIX5	0.038	0.13	0.039	7.4x10^{-4}	0.034	7.3x10^{-4}	6.7x10^{-5}	0.019	6.2x10^{-5}
	PAX5	0.043	0.25	0.073	3.9x10^{-4}	0.11	0.0046	3.4x10^{-5}	0.073	0.0011
	H3k27me3	0.044	0.32	0.18	4.2x10^{-4}	0.17	0.044	3.1x10^{-5}	0.12	0.024

Table 2.3. Most significant Gene Ontology terms from GRα ChIP-Enrich analysis using *nearest TSS* and *≤1kb from TSS* locus definitions show a large degree of overlap with significant GO terms from RNA-seq data from the same cell line. Most highly significant GO terms (after collapsing related terms; q-value ≤0.05) detected using ChIP-Enrich with the a) *nearest TSS* and b) *≤1kb from TSS* locus definitions. The highest ranked GO term from each related set of GO terms is displayed. Bold rows designate GO terms with q-value ≤0.05 in GOseq analysis of RNA-Seq data. In total, 458 GO terms (with ≤500 genes) were significantly enriched for the RNA-seq results.

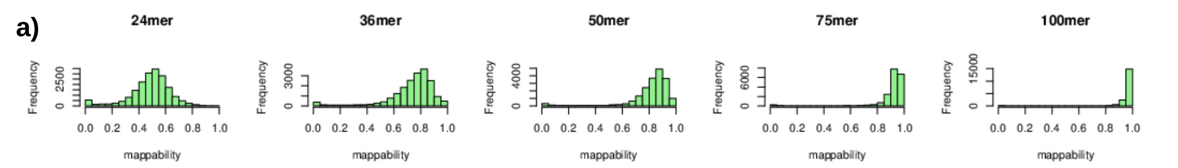
a

ChIP-Enrich rank *nearest TSS*	GOseq rank RNA-seq data	GO Term	ChIP-Enrich q-value		GOseq q-value
			nearest TSS	*≤1kb from TSS*	
1	**22**	**epithelial cell differentiation**	**1.8x10^{-6}**	**1.0**	**1.2x10^{-6}**
2	936	adherens junction	5.3x10^{-5}	1.0	0.39
4	**85**	**negative regulation of sequence-specific DNA binding transcription factor activity**	**5.5x10^{-5}**	**1.0**	**3.0x10^{-4}**
5	**9**	**anti-apoptosis**	**5.5x10^{-5}**	**0.34**	**3.2x10^{-9}**
7	1040	basolateral plasma membrane	1.7x10^{-4}	1.0	0.52
8	501	unsaturated fatty acid metabolic process	3.2x10^{-4}	0.028	0.063
10	872	focal adhesion	4.5x10^{-4}	1.0	0.32
13	**132**	**regulation of small GTPase mediated signal transduction**	**8.6x10^{-4}**	**1.0**	**1.3x10^{-3}**
14	**95**	**response to inorganic substance**	**1.2x10^{-3}**	**0.075**	**4.3x10^{-4}**
15	1616	response to growth hormone stimulus	1.4x10^{-3}	1.0	1.0

b

ChIP-Enrich rank *nearest TSS*	GOseq rank RNA-seq data	GO Term	ChIP-Enrich q-value		GOseq q-value
			≤1kb from TSS	*nearest TSS*	
1	**267**	**negative regulation of blood coagulation**	**3.2x10^{-7}**	**0.077**	**0.010**
7	1143	intrinsic to external side of plasma membrane	1.8x10^{-4}	0.062	0.68
8	1648	leukotriene metabolic process	2.2x10^{-4}	6.4x10^{-3}	1.0
10	4193	anchored to plasma membrane	2.1x10^{-3}	0.39	1.0
14	**323**	**positive regulation of leukocyte chemotaxis**	**3.5x10^{-3}**	**0.092**	**0.017**
15	1091	platelet alpha granule lumen	4.7x10^{-3}	0.25	0.61
18	1099	ameboidal cell migration	5.2x10^{-3}	0.31	0.94
19	1108	regulation of nuclease activity	5.2x10^{-3}	0.083	0.66
20	**192**	**cellular response to biotic stimulus**	**5.2x10^{-3}**	**6.1x10^{-3}**	**3.7x10^{-3}**
22	**876**	**nucleotide-binding domain, leucine rich repeat containing receptor signaling pathway**	**6.1x10^{-3}**	**0.15**	**0.010**

2.9 Supplementary Figures



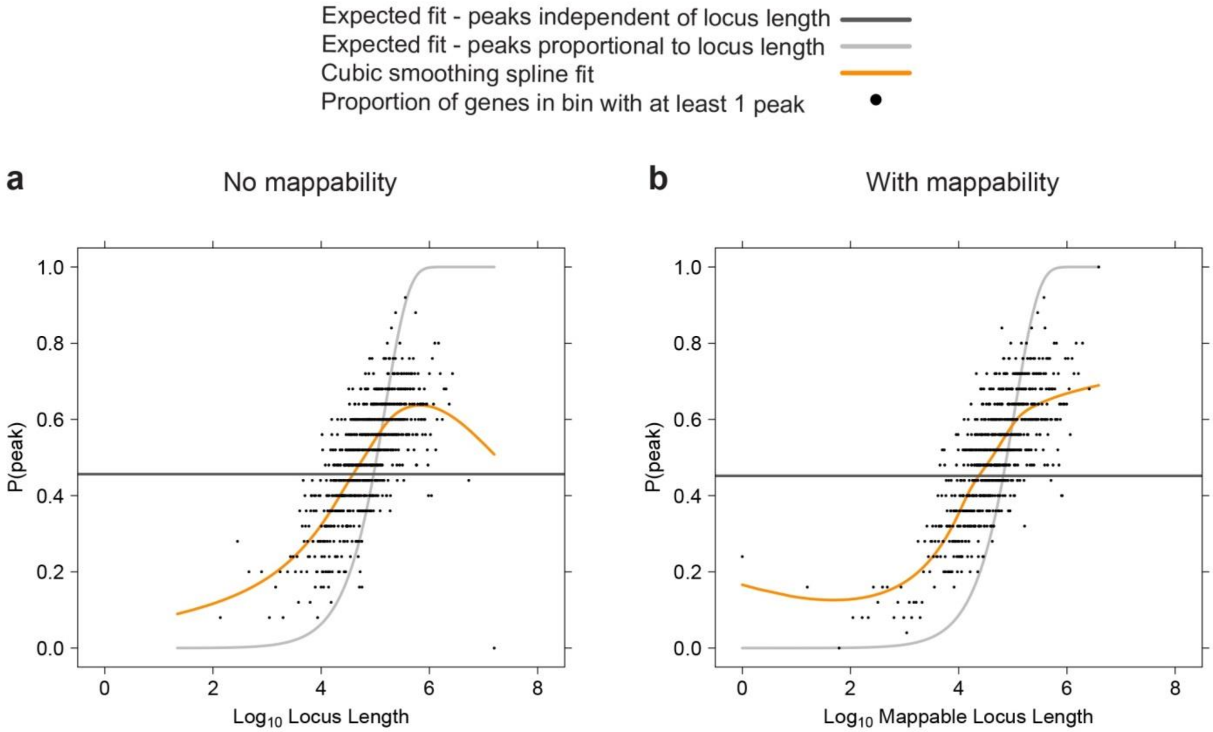
b) GO Terms whose Genes' Loci Have Higher Mappability

GO Term	# Genes	P-value	Q-value
Organ morphogenesis	642	2.6E-22	5.5E-19
Central nervous system development	454	2.9E-19	3.0E-16
Neurogenesis	634	1.4E-18	9.8E-16
Neuron differentiation	534	2.7E-18	1.4E-15
Cell development	786	5.5E-18	2.3E-15
Generation of neurons	589	1.6E-17	5.6E-15
Skeletal system development	272	2.6E-16	7.8E-14
Regionalization	217	1.9E-15	4.9E-13

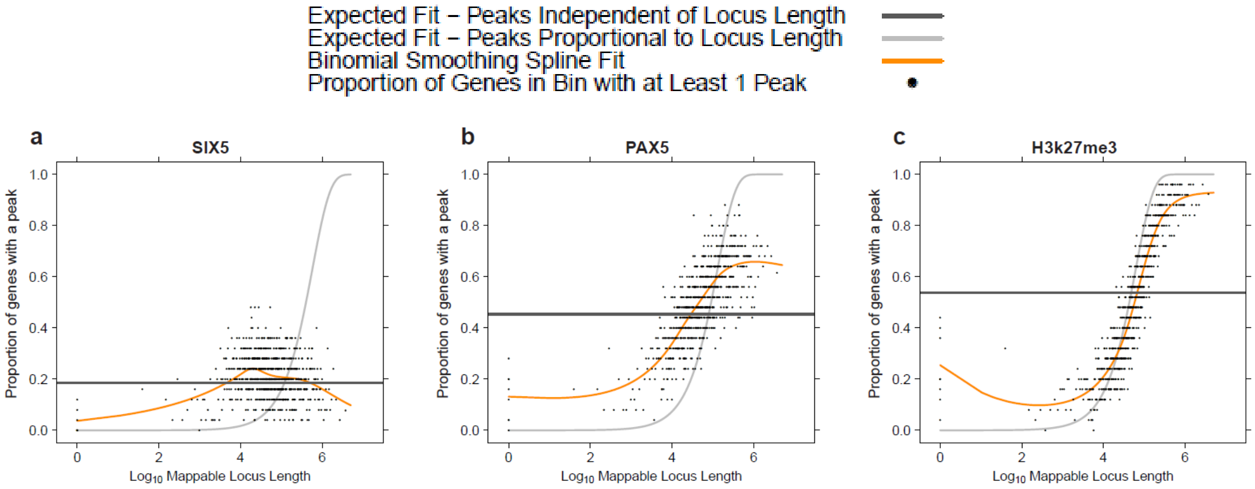
c) GO Terms whose Genes' Loci Have Lower Mappability

GO Term	# Genes	P-value	Q-value
Olfactory receptor activity	114	1.6E-11	7.0E-09
Sensory perception of smell	131	1.3E-09	6.3E-08
Cellular defense response	60	3.0E-08	9.0E-07
Sensory perception of chemical stimulus	167	3.7E-08	1.1E-06
Oxygen binding	44	7.7E-08	8.7E-06
Cellular response to xenobiotic stimulus	35	2.2E-07	5.1E-06
Xenobiotic metabolic process	35	2.2E-07	5.1E-06
Electron carrier activity	156	4.9E-07	3.0E-05

Supplementary Figure 2.1. Gene loci with high (or low) average mappability are enriched for specific Gene Ontology terms. (a) Distribution of human (hg19) mappability scores (calculated as the average mappability for each gene locus using the *nearest TSS* locus definition) for five different sequencing read lengths. (b) Most significantly enriched GO terms associated with high mappability using 50mer reads (c) Most significantly enriched GO terms associated with low mappability using 50mer reads. GO biological processes and molecular functions were tested using the LRpath gene set enrichment program [36].



Supplementary Figure 2.2. Using the mappable locus length (locus length x mappability) tends to improve the fit of the binomial cubic smoothing spline in the model, illustrated here with PAX5. Adjusting for mappability often shifts up the spline fit for the longest locus lengths; the dip in the fit without mappability is due to outlier points with long locus length and very low mappability.

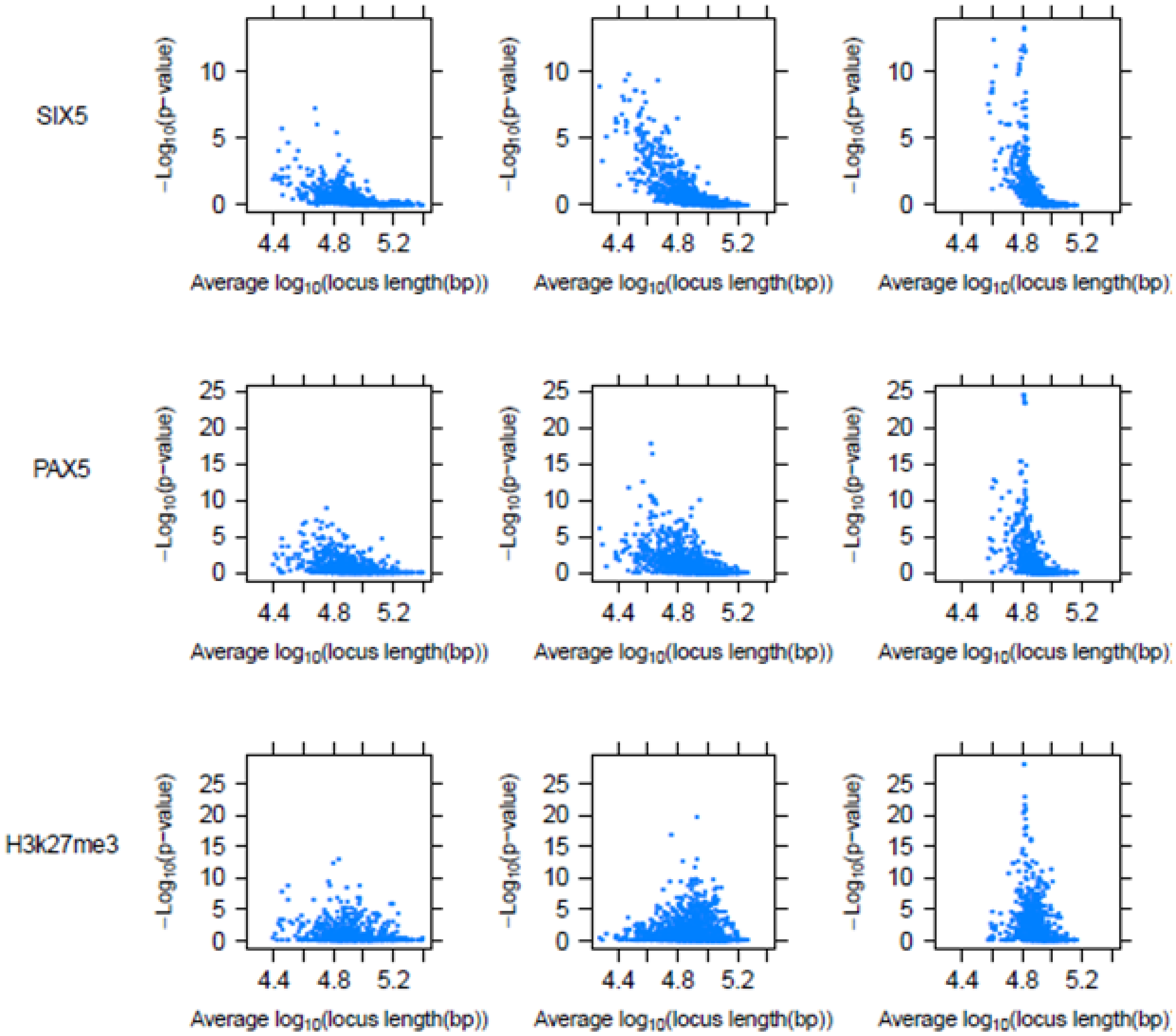


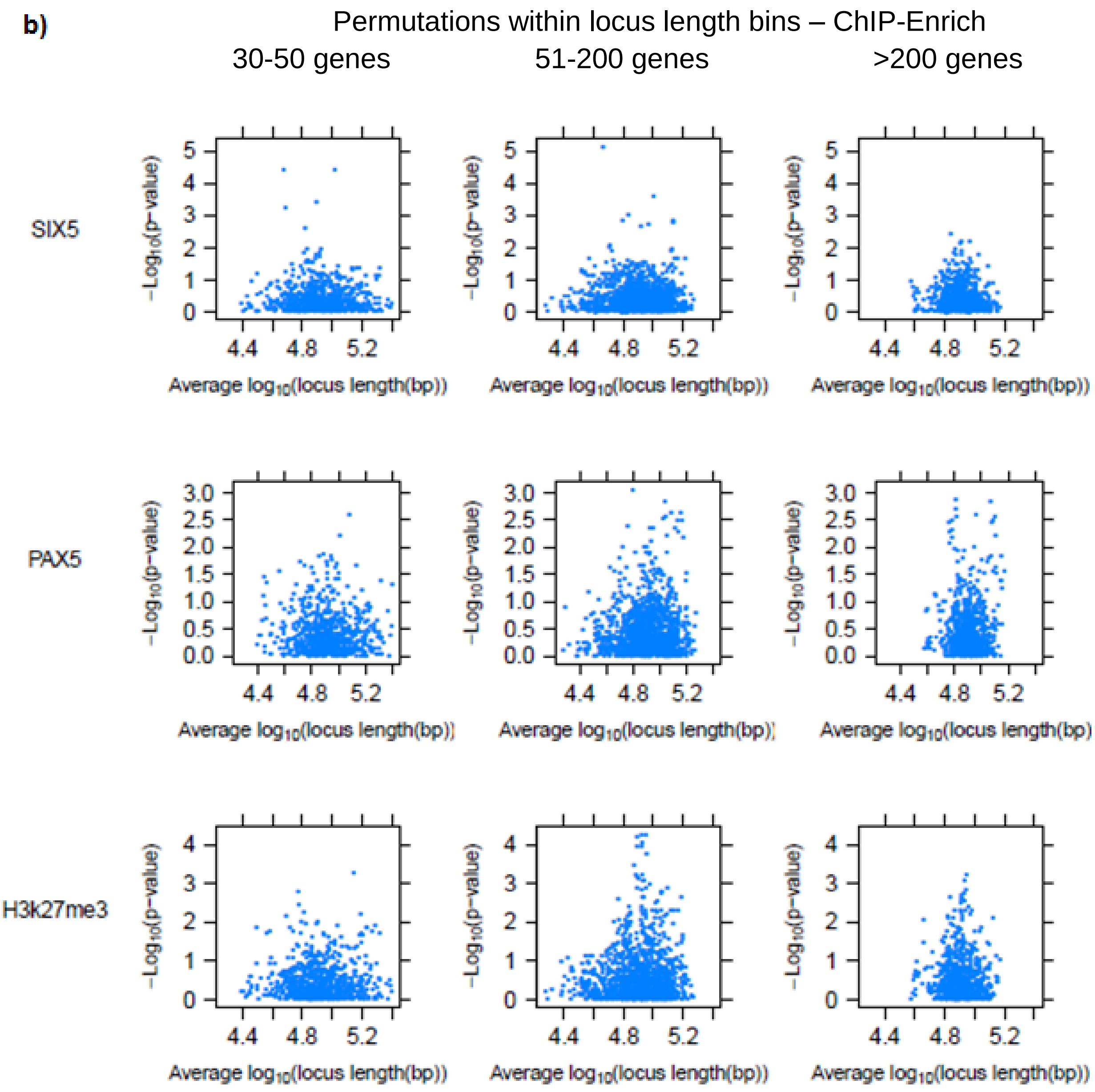
Supplementary Figure 2.3. SIX5, PAX5, and H3K27me3 have different gene locus-length-to-peak presence relationships. SIX5 (also in Figure 2.1a) has a weak relationship; PAX5 (also in Figure 2.1b) has a mid-level relationship; H3K27me3 (also in Figure 2.1c) has a strong relationship. The relationships between locus length and proportion of genes with a peak were estimated using a binomial cubic smoothing spline (orange line). Expected line if no relationship between presence of ≥ 1 peak and log10 locus length (dark gray, satisfies Fisher's exact test assumptions). Expected line if number of peaks observed is proportional to locus length (light gray, binomial test assumption). For visualization only, each point is the proportion of genes assigned a peak within sequential bins of 25 genes.

a)

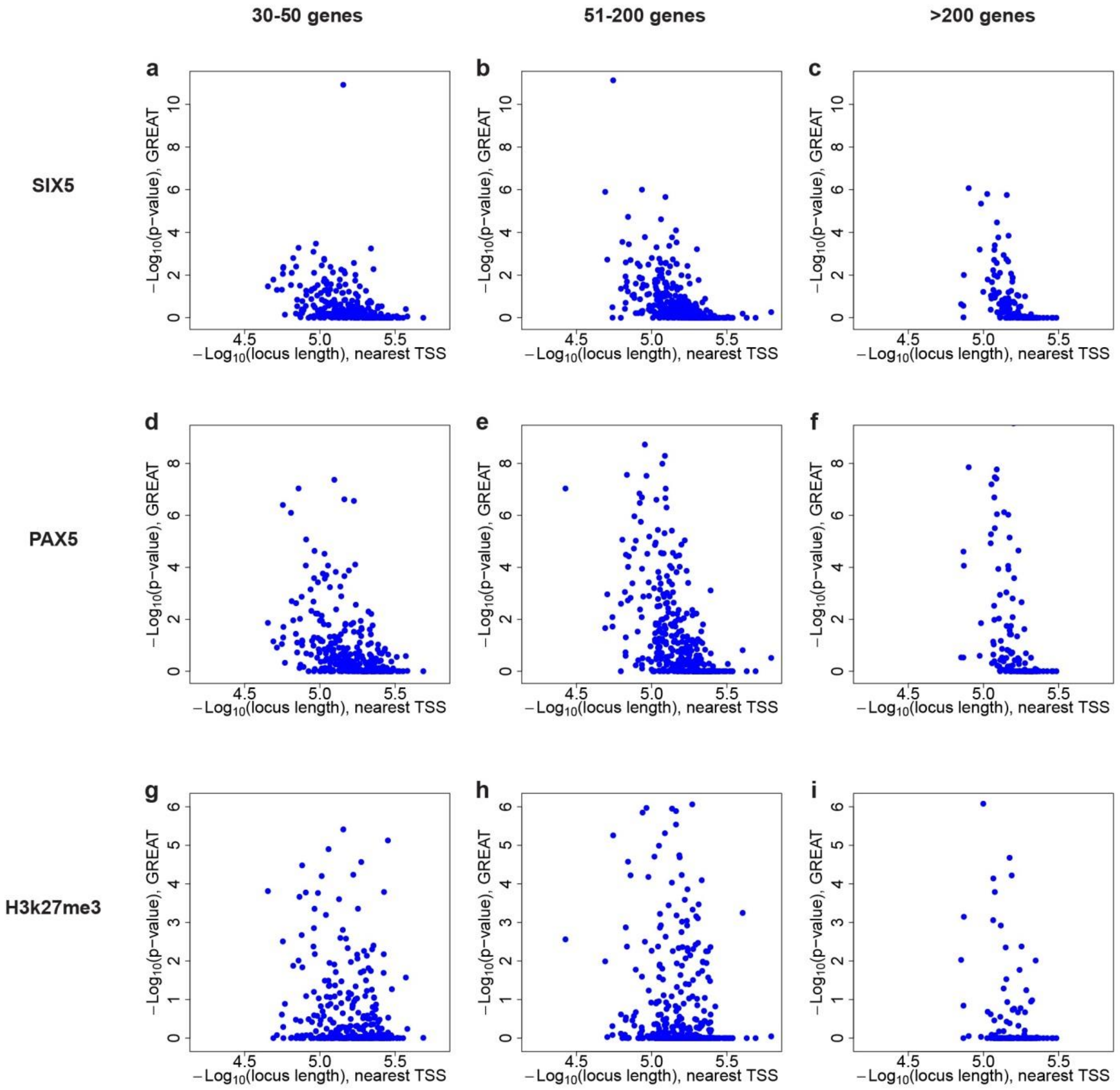
Permutations within locus length bins – Binomial test

30-50 genes 51-200 genes >200 genes

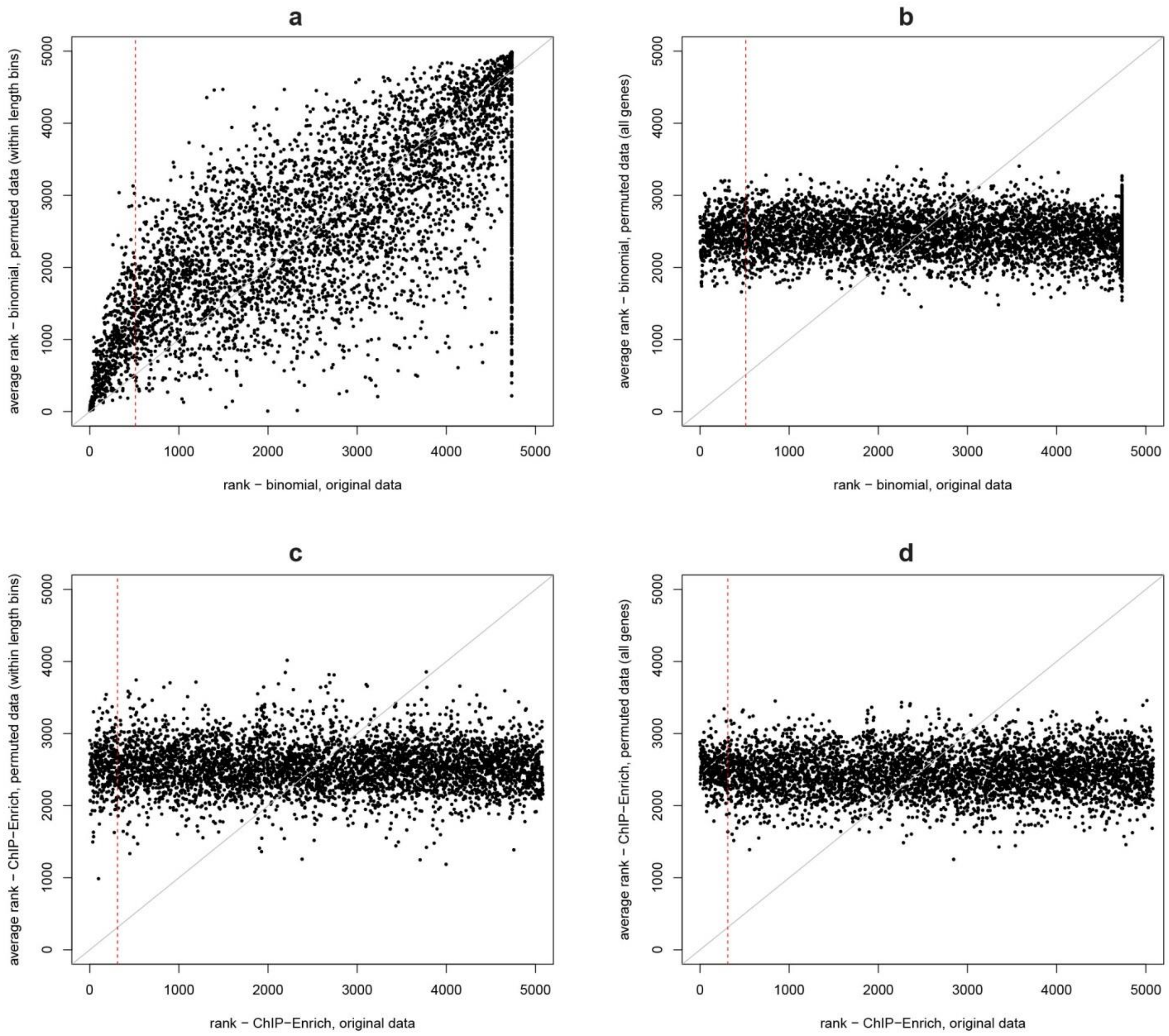




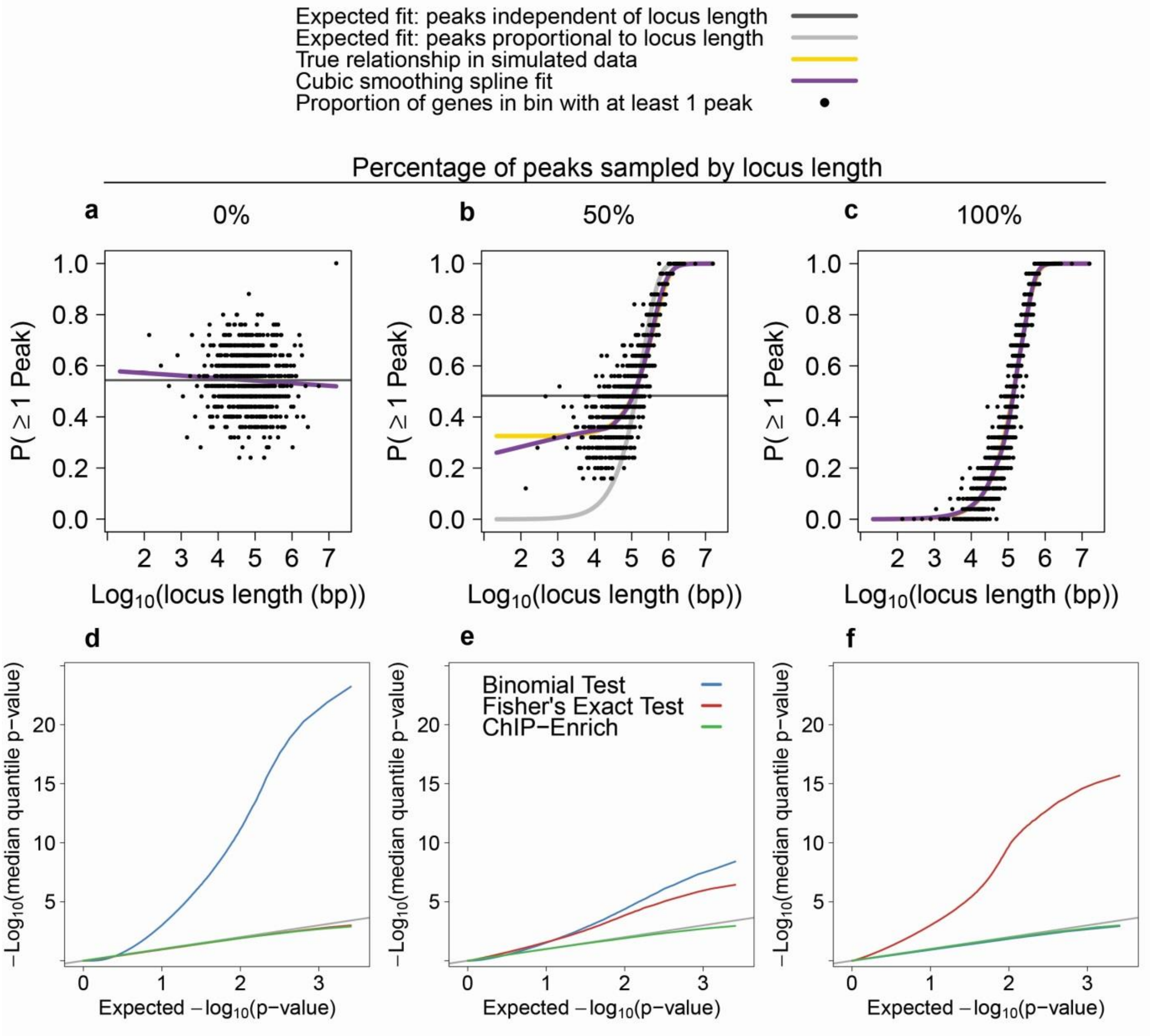
Supplementary Figure 2.4. The binomial test tends to identify gene sets with short locus length as significant (p < 0.05), especially for SIX5. Panel (a) shows the –log10 p-values from the binomial test versus the average log10 locus length of each gene set tested. Each row shows results from a permutation of the DBP dataset, where the original DBP dataset has been permuted by shuffling genes within bins of locus length. Each column subdivides all gene sets by their number of genes: the first column has gene sets with 30-50 genes, the next 51-200 genes, and the last > 200 genes. Panel (b) shows plots as in (a) for ChIP-Enrich. The binomial test shows a trend of much larger –log10 p-values for gene sets with low average log10 locus length, and this trend is most pronounced for sets of genes with fewer than 200 genes (first and second columns.) ChIP-Enrich does not show this trend for any of the three datasets tested.



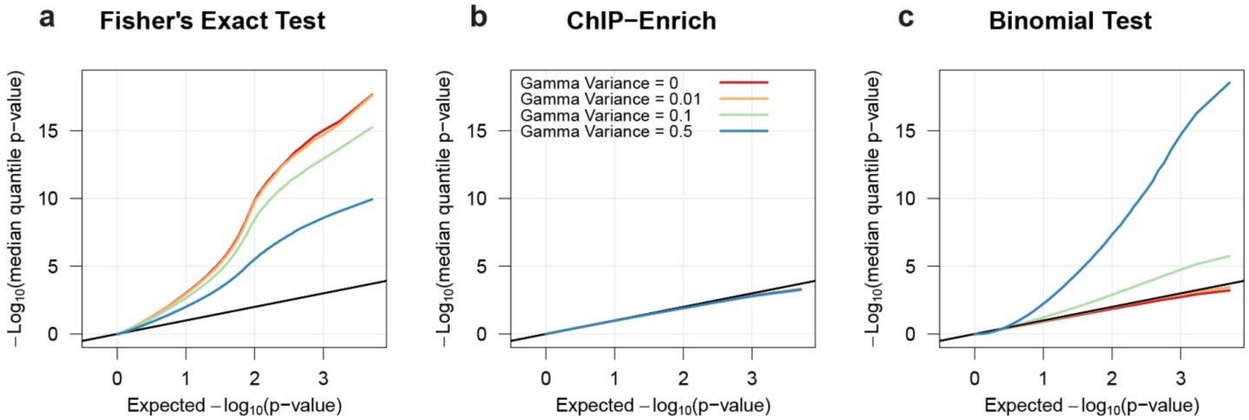
Supplementary Figure 2.5. The GREAT website test tends to detect gene sets with shorter than average locus length, especially for SIX5. Plots show the –log$_{10}$ p-values from the GREAT test versus the average log$_{10}$ locus length of each gene set tested. Each row show results from a permutation of the DBP dataset, where the original DBP dataset has been permuted by shuffling genes within bins of locus length. Each column subdivides all gene sets by their number of genes. Gene set enrichment testing using GREAT on each set of permuted peaks from SIX5, PAX5, and H3k27me3 found significantly enriched GO terms (FDR≤0.05), when none should have been detected. The trend with locus length was again greatest for SIX5 and least for H3k27me3.



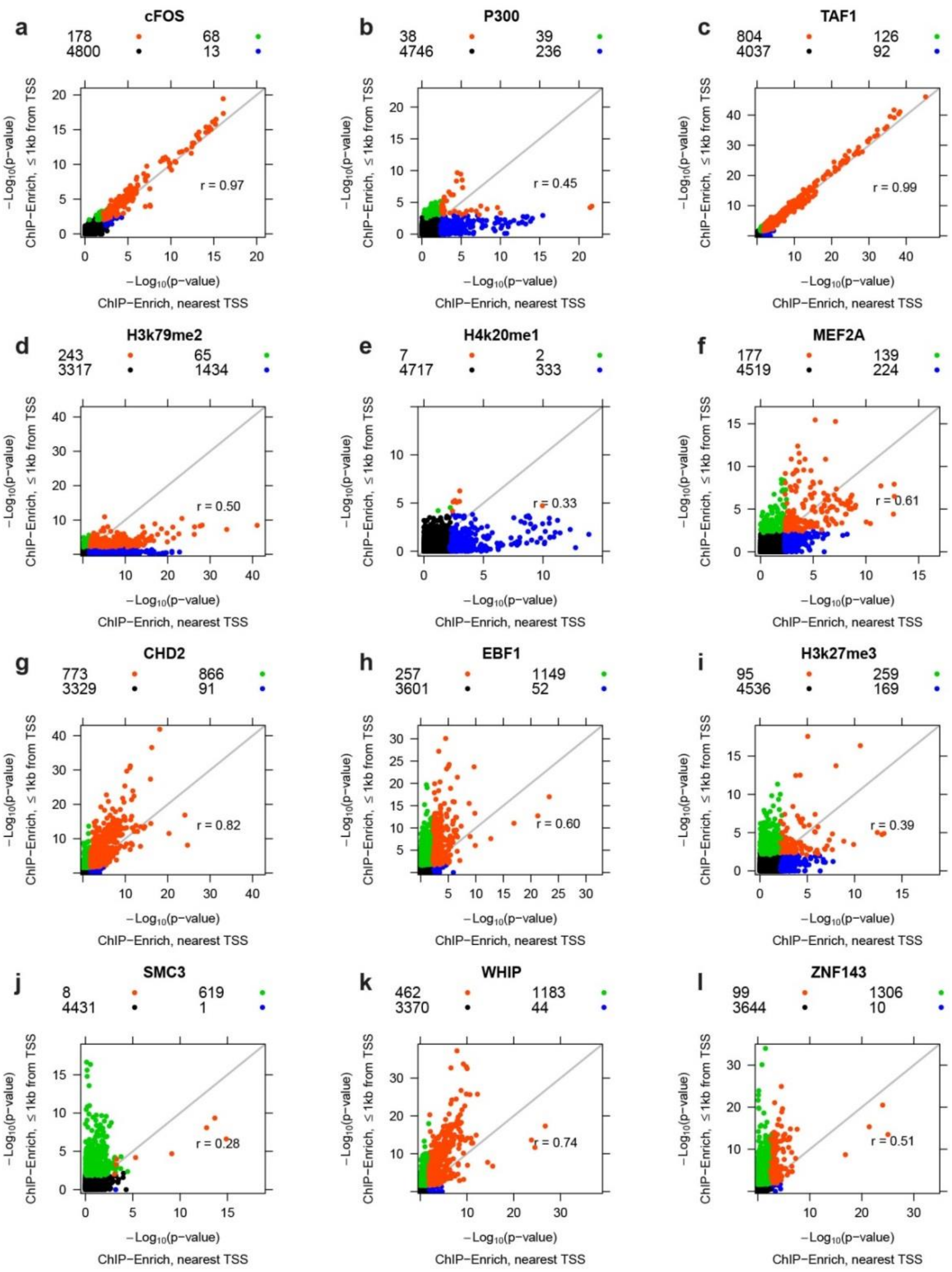
Supplementary Figure 2.6. For SIX5 the ranks of the binomial test results from the original data are highly correlated the average ranks from 25 permutations (within locus length bins) of the original data (Spearman r = 0.71). For ChIP-Enrich the rank of test results from original data and the average rank permuted data are not correlated. The fact that this correlation is not observed in (b, permutations across all genes), implies that the correlation is due to the locus length bias. Each plot compares the average ranks of results from 25 permutations to rank in the original data, and the red dash line indicates the highest rank where FDR≤0.05 for the original data. (a) Binomial test results for permutations within locus length bins. Of the 509 significantly enriched (FDR≤0.05) GO terms (with ≤500 genes) using the original, non-permuted data, 413 (81.1%) were also significantly enriched in at least one of the permuted data sets. Of the 4,325 not significantly enriched GO terms, only 583 (13.5%) were enriched in at least one of the permuted data sets. (b) Binomial test results from permutations across all genes. The average ranks of the binomial test results are not correlated with the ranks of the original data (Spearman r=-0.06), indicating that the correlation in (a) is due to the confounding by locus length. ChIP-Enrich test results from (c) permutations within locus length bins and (d) permutations across all genes, respectively. In both permutation scenarios, the ranks of ChIP-Enrich results from permuted and the original data were not correlated (Spearman r=-0.02 and -0.005, respectively).



Supplementary Figure 2.7. Relationship in simulated datasets between locus length and presence of at least one peak (a-c), and QQ-plots showing the type 1 error rate of Fisher's exact test, the binomial test, and ChIP-Enrich under these relationships (d-f). Simulated datasets of 10,000 peaks and 0% (a, d), 50% (b, e), or 100% (c, f) of peaks sampled in proportion to locus length. Top row (a-c) - For visualization, each point is a bin of 25 genes, plotted as the average proportion of genes having a peak within the bin against the average log10 locus length. The dark grey horizontal line represents the model where peaks occur within genes with no relationship to their locus length. The light grey line represents the probability of a locus having ≥1 peak if peaks are randomly distributed across the genome (binomial test assumption). The purple line is a binomial smoothing spline fit to the underlying data (the 0/1 vector denoting whether a peak was assigned to a gene vs. the log10 locus length of each gene). The yellow line represents the known relationship that exists in the simulated data. Bottom row (d-f) – QQ plots showing Fisher's exact and the binomial test represent two extreme assumptions for enrichment testing for ChIP-seq data, while ChIP-Enrich empirically estimates the correct balance between these two extremes. Incorrect assumptions at either end leads to biased significance levels. Median p-values (solid lines) are shown for 1000 simulations of Fisher's exact test, ChIP-Enrich, and the binomial test.

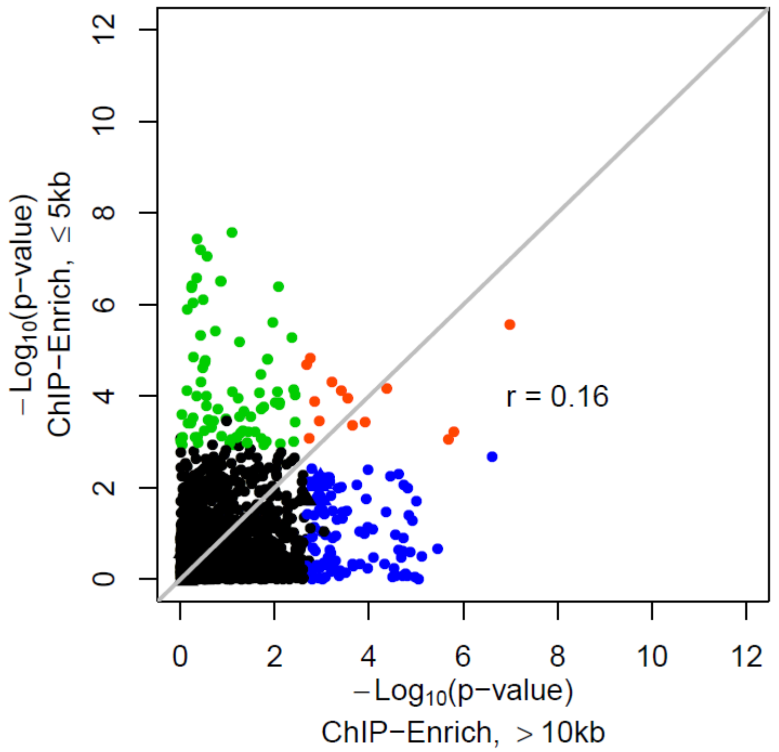


Supplementary Figure 2.8. Increasing overdispersion in peak counts among genes increases the type 1 error rate of the binomial test, decreases type 1 error for Fisher's exact test, and has no effect on ChIP-Enrich. QQ plots of expected versus observed –$\log_{10}$(p-values) for (a) Fisher's exact test, (b) ChIP-Enrich, and (c) the binomial test. Increasing levels of overdispersion, modeled using a gamma distribution, were assessed ranging from no overdispersion (red) to a gamma distribution with variance = 0.5 (blue). Red line (no overdispersion) represents the same simulation as that in Supplementary Figure 2.7c,f. Simulated datasets of 10,000 peaks were used, and median p-values for 1000 simulations are shown.



Supplementary Figure 2.9. Gene set enrichment testing using ≤ *1kb from TSS* and *nearest TSS* locus definitions often identifies very different sets of significant GO terms for the same DBP.

Comparison of –log$_{10}$(p-values) from testing GO terms with ChIP-Enrich using ≤ *1kb from TSS* versus *nearest TSS* locus definitions in ENCODE data for the GM12878 cell line. GO terms with: FDR ≤.05 for ≤ *1kb from TSS* only (green); FDR ≤.05 for *nearest TSS* only (blue); FDR ≤.05 for ≤ *1kb from TSS* and *nearest TSS* (orange); FDR >.05 in both analyses (black). r: Pearson correlation coefficient. DBPs are arranged by groupings in Figure 2.1a, (a-c) are DBPs with low number of peaks, (d-f) are DBPs with medium number of peaks, and (g-l) are DBPs with high number of peaks.



Supplementary Figure 2.10. A comparison of ChIP-Enrich GO term enrichment results for GR using peaks ≤5kb from the TSS and peaks >10kb from the TSS. Only 14 gene sets were significantly enriched (q≤0.05) in both tests. Vasculature development (shown as the blue triangle) was only significant using peaks *>10kb from the TSS*. GO terms with: FDR ≤.05 for ≤ *5kb from TSS* only (green); FDR ≤.05 for *>10kb from TSS* only (blue); FDR ≤.05 for ≤ *5kb from TSS* and ≥*10kb from TSS* (orange); FDR >.05 in both analyses (black). r: Pearson correlation coefficient.

2.10 Supplementary Tables

Supplementary Table 2.1. List of DBPs from Figure 2.1 with their total peak counts and associated peak caller. All DBPs are from ENCODE cell line GM12878.

DNA binding protein	Peak caller	Number of peaks	% of peaks ≤ 1kb from TSS
ATF3	spp	1884	66.9
BATF	spp	24600	4.9
BCL11A	spp	13256	5.7
BCL3	MACS	22503	13.1
BCLAF1	MACS	29162	30.8
BHLHE40	MACS	57698	21.7
BRCA1	MACS	15431	40.6
C-Fos	spp	1744	81.8
CHD2	MACS	42652	29.4
CTCF	MACS	44056	16.1
Ctcf	Scripture	61525	12.8
EBF1	MACS	98976	14.2
EGR1	spp	13662	54.3
ELF1	spp	20528	52.7
ETS1	spp	2879	72.6
Ezh2	Scripture	64277	9.9
GABP	spp	5095	79.9
H2az	Scripture	95358	15.0
H3k27ac	Scripture	56069	18.5
H3k27me3	Scripture	41464	5.4
H3k36me3	Scripture	33710	3.0
H3k4me1	Scripture	109612	8.9
H3k4me2	Scripture	79675	15.3
H3k4me3	Scripture	57476	17.6
H3k79me2	Scripture	28302	13.7
H3k9ac	Scripture	41266	25.5
H3k9me3	Scripture	74515	2.3
H4k20me1	Scripture	23943	5.0
JunD	spp	1715	3.0
MAX	spp	2087	39.7
MEF2A	spp	16694	11.5
MEF2C	MACS	968	15.1
NF-E2	MACS	12973	24.2
NFKB	spp	10073	22.1
NRF1	spp	5042	12.7
NRSF	spp	2541	39.2
P300	spp	3687	16.5

PAX5	spp	19618	18.5
PBX3	spp	7431	26.1
POL2	MACS	14989	33.1
POL3	MACS	112	55.5
POU2F2	spp	14441	32.9
PU.1	spp	35821	11.3
RAD21	MACS	23947	6.5
RFX5	MACS	26336	33.4
RXRA	spp	2965	31.4
SIX5	spp	4442	74.8
SMC3	MACS	64597	14.4
SP1	spp	13139	46.1
SRF	spp	2412	48.7
STAT3	MACS	24257	15.1
TAF1	spp	5169	82.4
TBP	MACS	31315	30.4
TCF12	spp	15028	25.8
TR4	MACS	1530	29.5
USF1	spp	7074	43.0
USF2	MACS	30248	28.3
WHIP	MACS	88803	17.0
YY1	MACS	42162	32.3
ZBTB33	spp	1934	64.8
ZEB1	spp	8304	46.5
ZNF143	MACS	81743	18.4
ZNF274	MACS	1483	1.5

Supplementary Table 2.2. Significant overdispersion in peak count among genes is observed for a substantial number of GO terms in all 63 ENCODE ChIP-seq datasets from cell line GM12878. Number and percentage of GO terms (with 50-500 genes) that contain significant overdispersion in peak counts among the genes (q≤0.05). DBPs from Figure 2.1 panels c,f have more peaks than DBPs from panels a, d and thus higher power to detect significant overdispersion.

DBP	Figure 1 panel	# over-dispersed GO terms	% GO terms over-dispersed		DBP	Figure 1 panel	# over-dispersed GO terms	% GO terms over-dispersed
ATF3	a, d	630	35		PAX5	b, e	1831	100
cFOS	a, d	758	42		POL2	b, e	1828	100
ETS1	a, d	1253	68		POU2F2	b, e	1827	100
GABP	a, d	1620	88		RAD21	b, e	1755	96
JunD	a, d	692	38		RFX5	b, e	1829	100
MAX	a, d	705	39		SP1	b, e	1829	100
MEF2C	a, d	561	31		STAT3	b, e	1830	100
NF-E2	a, d	1800	98		TBP	b, e	1831	100
NFKB	a, d	1803	98		TCF12	b, e	1828	100
NRSF	a, d	493	27		USF2	b, e	1830	100
P300	a, d	1413	77		BHLHE40	c, f	1831	100
PBX3	a, d	1751	96		CHD2	c, f	1832	100
RXRA	a, d	1319	72		CTCF	c, f	1831	100
SIX5	a, d	1448	79		Ctcf	c, f	1831	100
SRF	a, d	902	49		EBF1	c, f	1832	100
TAF1	a, d	1614	88		Ezh2	c, f	1830	100
TR4	a, d	577	32		H2az	c, f	1832	100
USF1	a, d	1763	96		H3k27ac	c, f	1832	100
ZBTB33	a, d	769	42		H3k27me3	c, f	1831	100
ZEB1	a, d	1807	99		H3k36me3	c, f	1832	100
ZNF274	a, d	832	59		H3k4me1	c, f	1832	100
BATF	b, e	1821	99		H3k4me2	c, f	1832	100
BCL11A	b, e	1813	99		H3k4me3	c, f	1832	100
BCL3	b, e	1826	100		H3k9ac	c, f	1832	100
BCLAF1	b, e	1831	100		H3k9me3	c, f	1832	100
BRCA1	b, e	1824	100		MXI1	c, f	1832	100
EGR1	b, e	1825	100		PU1	c, f	1828	100
ELF1	b, e	1831	100		SMC3	c, f	1831	100
H3k79me2	b, e	1832	100		WHIP	c, f	1832	100
H4k20me1	b, e	1830	100		YY1	c, f	1832	100
MEF2A	b, e	1815	99		ZNF143	c, f	1832	100
NRF1	b, e	1829	100					

Supplementary Table 2.3. GO terms most strongly associated with short locus length.
The 500 genes with shortest locus lengths (ranging from 23bp to 5,066bp) were tested for GO term enrichment relative to all remaining genes (having a computed locus length) using DAVID [39].

GO Term	# genes	total genes in term	fold enrich	p-value	q-value
nucleosome	13	58	11.64	$8.0x10^{-10}$	$2.5x10^{-7}$
protein-DNA complex	14	81	8.97	$4.3x10^{-9}$	$6.6x10^{-7}$
translation	25	314	4.10	$9.7x10^{-9}$	$1.3x10^{-5}$
DNA packaging	14	105	6.87	$1.1x10^{-7}$	$7.2x10^{-5}$
nucleosome assembly	12	74	8.35	$1.6x10^{-7}$	$7.0x10^{-5}$
chromatin assembly	12	77	8.03	$2.5x10^{-7}$	$8.0x10^{-5}$
ribosome	18	201	4.65	$3.4x10^{-7}$	$2.6x10^{-5}$
protein-DNA complex assembly	12	81	7.63	$4.2x10^{-7}$	$1.1x10^{-4}$
cellular macromolecular complex assembly	22	304	3.73	$4.6x10^{-7}$	$1x10^{-4}$
nucleosome organization	12	83	7.45	$5.4x10^{-7}$	$1x10^{-4}$

Supplementary Table 2.4. GO terms most strongly associated with long locus length.
The 500 genes with the longest locus lengths (ranging from 879 kb to 15,8 Mb) were tested for GO term enrichment relative to all remaining genes (having a computed locus length) using DAVID [39].

GO Term	# genes	total genes in term	fold enrich	p-value	q-value
homophilic cell adhesion	29	130	8.65	$1.9x10^{-18}$	$3.9x10^{-15}$
nervous system development	78	1066	2.84	$2.2x10^{-17}$	$2.3x10^{-14}$
cell adhesion	60	686	3.39	$9.9x10^{-17}$	$7.6x10^{-14}$
biological adhesion	60	687	3.39	$1.0x10^{-16}$	$5.7x10^{-14}$
cell-cell adhesion	37	271	5.30	$4.3x10^{-16}$	$1.8x10^{-13}$
generation of neurons	43	549	3.04	$2.1x10^{-10}$	$4.0x10^{-8}$
calcium ion binding	58	896	2.49	$2.4x10^{-10}$	$1.4x10^{-7}$
neurogenesis	44	591	2.89	$6.1x10^{-10}$	$1.1x10^{-7}$
neuron differentiation	34	429	3.07	$1.9x10^{-8}$	$3.0 x10^{-6}$
axonogenesis	21	191	4.26	$1.0x10^{-7}$	$1.5x10^{-5}$

Supplementary Table 2.5a. Top enriched GO terms (not collapsed, with ≤500 genes) for GR ChIP-seq data (4,392 peaks) that were significantly enriched (q ≤0.05) using the *nearest TSS* locus definition.
Bolded terms are significantly enriched in both ChIP-Enrich and GOseq results. The complete list of enriched GO terms for GR using the *nearest TSS* and *≤1kb from TSS* locus definition is included as a supplemental excel file, "Supplementary_table_5expanded.csv."

a	Rank	GO term	*nearest TSS* q-value	*≤1kb from TSS* q-value	GOseq q-value
	1	**epithelial cell differentiation**	**1.8x10^{-6}**	**1.0**	**1.2x10^{-6}**
	2	adherens junction	5.3x10^{-5}	1.0	0.39
	3	anchoring junction	5.3x10^{-5}	1.0	0.49
	4	**negative regulation of sequence-specific DNA binding transcription factor activity**	**5.5x10^{-5}**	**1.0**	**3.0x10^{-4}**
	5	**anti-apoptosis**	**5.5x10^{-5}**	**0.34**	**3.2x10^{-9}**
	6	regulation of epithelial cell differentiation	7.6x10^{-5}	1.0	0.14
	7	basolateral plasma membrane	1.7x10^{-4}	01.0	0.52
	8	unsaturated fatty acid metabolic process	2.9x10^{-4}	3.5x10^{-3}	0.52
	9	**icosanoid metabolic process**	**1.8x10^{-4}**	**2.9x10^{-3}**	**4.1x10^{-4}**
	10	focal adhesion	4.5x10^{-4}	1.0	0.32
	11	cell-substrate junction	4.5x10^{-4}	1.0	0.39
	12	cell-substrate adherens junction	4.5x10^{-4}	1.0	0.35
	13	**regulation of small GTPase mediated signal transduction**	**8.7x10^{-4}**	**1.0**	**1.3x10^{-3}**
	14	**response to inorganic substance**	**1.2x10^{-3}**	**0.075**	**4.3x10^{-4}**
	15	response to growth hormone stimulus	1.4x10^{-3}	1.0	1.0
	16	**regulation of cellular component movement**	**1.8x10^{-3}**	**0.74**	**5.7x10^{-6}**
	17	monocarboxylic acid metabolic process	1.9x10^{-3}	0.15	0.66
	18	response to calcium ion	0.0024	0.33	0.14
	19	regulation of anti-apoptosis	2.9x10^{-3}	0.76	0.63
	20	**negative regulation of protein metabolic process**	**3.2 x10^{-3}**	**0.57**	**5.5 x10^{-3}**
	21	response to glucocorticoid stimulus	3.5 x10^{-3}	0.56	0.39
	22	positive regulation of anti-apoptosis	3.7 x10^{-3}	0.62	0.49
	23	response to corticosteroid stimulus	3.9 x10^{-3}	0.63	0.18
	24	regulation of epidermal cell differentiation	4.1 x10+	1.0	0.84
	25	Ras protein signal transduction	4.1 x10^{-3}	1.0	0.26
	26	energy reserve metabolic process	4.2 x10^{-3}	1.0	0.60
	27	**negative regulation of transcription from RNA polymerase II promoter**	**4.4 x10^{-3}**	**1.0**	**2.5x10^{-6}**
	28	actin cytoskeleton organization	4.4 x10^{-3}	1.0	0.15
	29	**vasculature development**	**4.7 x10^{-3}**	**0.97**	**7.4x10^{-16}**
	30	**small GTPase mediated signal transduction**	**4.7 x10^{-3}**	**1.0**	**4.9 x10^{-3}**

Supplementary Table 2.5b. Top enriched GO terms (not collapsed, with ≤500 genes) for GR ChIP-seq data (4,392 peaks) that were significantly enriched (q ≤0.05) using the *≤1kb from TSS* locus definition. Bolded terms are significantly enriched in both ChIP-Enrich and GOseq results. The complete list of enriched GO terms for GR using the nearest TSS and ≤1kb from TSS locus definition is included as a supplemental excel file, "Supplementary_table_5expanded.csv."

b

Rank	GO term	*≤1kb from TSS* q-value	*nearest TSS* q-value	GOseq q-value
1	**negative regulation of blood coagulation**	**3.2x10^{-7}**	**0.077**	**0.010**
2	**negative regulation of coagulation**	**4.3x10^{-7}**	**0.15**	**0.015**
3	**fibrinolysis**	**3.5x10^{-5}**	**0.088**	**0.033**
4	**regulation of blood coagulation**	**7.0x10^{-5}**	**0.011**	**5.4x10^{-4}**
5	regulation of fibrinolysis	1.4x10^{-4}	0.073	0.11
6	**regulation of coagulation**	**1.4x10^{-4}**	**0.029**	**9.9x10^{-4}**
7	intrinsic to external side of plasma membrane	1.8x10^{-4}	0.062	0.68
8	leukotriene metabolic process	2.2x10^{-4}	6.3x10^{-3}	1.0
9	**positive regulation of coagulation**	**3.1x10^{-3}**	**0.061**	**0.028**
10	anchored to plasma membrane	2.1x10^{-3}	0.39	1.0
11	**regulation of wound healing**	**2.9x10^{-3}**	**6.4x10^{-3}**	**1.4X10^{-4}**
12	**regulation of response to external stimulus**	**2.9x10^{-3}**	**0.014**	**5.0X10^{-5}**
13	**positive regulation of coagulation**	**3.0x10^{-3}**	**0.061**	**0.028**
14	**positive regulation of leukocyte chemotaxis**	**3.5 x10^{-3}**	**0.092**	**0.017**
15	platelet alpha granule lumen	4.7 x10^{-3}	0.25	0.61
16	secretory granule lumen	4.7 x10^{-3}	0.29	0.63
17	cytoplasmic membrane-bounded vesicle lumen	5.1 x10^{-3}	0.25	0.64
18	ameboidal cell migration	5.2 x10^{-3}	0.31	0.94
19	regulation of nuclease activity	5.2 x10^{-3}	0.083	0.66
20	**cellular response to biotic stimulus**	**5.2x10^{-3}**	**6.1x10^{-3}**	**3.7x10^{-3}**
21	vesicle lumen	5.5 x10^{-3}	0.20	0.68
22	nucleotide-binding domain, leucine rich repeat containing receptor signaling pathway	6.1 x10^{-3}	0.15	0.32
23	peptidyl-glutamic acid carboxylation	6.9 x10^{-3}	0.11	1.0
24	**regulation of leukocyte chemotaxis**	**0.011**	**0.18**	**0.028**
25	long-chain fatty acid transport	0.011	0.028	0.12
26	**second-messenger-mediated signaling**	**0.012**	**0.66**	**0.047**
27	**positive regulation of leukocyte migration**	**0.014**	**0.20**	**0.031**
28	carboxylic acid transport	0.014	0.18	0.39
29	external side of plasma membrane	0.015	0.63	1.0
30	endoplasmic reticulum unfolded protein response	0.069	0.015	0.29

Chapter 3 RNA-Enrich: A cut-off free functional enrichment testing method for RNA-seq with improved detection power

3.1 Introduction

Functional enrichment testing is one of the most common downstream analyses for transcriptomics experiments, facilitating a deeper interpretation of results. Examples of gene set databases used for testing are Gene Ontology (GO) which includes biological processes, cellular components, and molecular functions, and the Kyoto Encyclopedia of Genes and Genomes (KEGG) which places genes in metabolic and other pathways. Most current gene set enrichment (GSE) methods, such as DAVID [66], were developed for microarray data. These methods often only make use of differential expression (DE) p-values or ranks, or simply a list of significant genes. With RNA-seq, which uses whole transcriptome sequencing to quantify gene expression, tests for DE often exhibit a relationship between read count and likelihood of detecting DE. For example, when power is greater to detect longer and/or higher expressed genes, gene sets that have long genes or that are highly expressed are more likely to be detected as significant, violating common test assumptions. Thus, accounting for read count per gene may improve standard GSE methods, which may otherwise not be appropriate for RNA-seq data.

RNA-seq achieves a very high dynamic range, with gene read counts often varying across six or more orders of magnitude. Read-count based methods such as those using a negative binomial model (e.g., edgeR and DEseq2) can be more likely to identify longer and highly-expressed transcripts as significant. Two methods that can account for this bias in GSE testing are GOseq [9] which requires a p-value cut-off, and

Chapter 3 is published as **Lee C**, Patil S, Sartor MA: RNA-Enrich: a cut-off free functional enrichment testing method for RNA-seq with improved detection power. Bioinformatics 2015.

GSAASeqSP [32] and SeqGSEA[33], which require permutations and moderate to large sample sizes to obtain a sufficient number of unique permutations of phenotype labels. We have developed RNA-Enrich, a GSE method that empirically adjusts for average read count per gene, and does not require a cut-off to define differentially expressed genes (DEGs), time consuming permutations, or regression models. Similar cut-off free methods for microarray data have shown improved ability to detect gene sets enriched with either a few very strong DEGs or many only moderate DEGs [35, 36].

3.2 Methods

3.2.1 Model

RNA-Enrich models the relationship between $\log_{10}$(average read count) per gene and -$\log_{10}$(significance score) using a binomial cubic smoothing spline. The significance scores, usually p-values, and read counts are input by the user. Per gene weights (w_g) are calculated from the spline fit as the ratio between mean –$\log_{10}$(p-value) and fitted values, and then normalized to have a mean of 1. A modified version of the random sets method, as proposed by [34], is used. We calculate the test statistic $\bar{x}$ for genes in a gene set:

$$\bar{x} = mean(w_g * s_g) \qquad (1)$$

where s_g is the –$\log_{10}$(p-value) from a differential gene expression test such as edgeR or DESeq2. The distribution of the statistic to test whether $\bar{x}$ is significantly different from what is expected by chance is intractable. Instead, we use the first and second moments of the distribution to define approximate z-scores which are then used to calculate p-values of enrichment [34]. Adjusted p-values (q-values) are calculated to correct for multiple testing.

The use of weights ensures that if a relationship exists between read count and DE p-values genes, it will be adjusted for properly. The original random sets method does not include the w_g terms, i.e. all genes are equally weighted; the method for calculating p-values using approximate z-scores was the same. Our website supports 16 different annotation databases plus custom gene sets, seven organisms, and clustering of results.

3.2.2 Datasets

Prostate cancer LNCaP cells treated with an androgen hormone

Li, et al [90] collected RNA-seq data from LNCaP cells, a prostate cancer model cell line, treated with an androgen hormone, which is associated with survival in prostate cancer. There were 7 total samples: 3 replicates treated with 100μM of dihydrotestosterone (DHT) to stimulate androgen production, and 4 controls treated with an inactive compound. Expression data was downloaded from http://yeolab.ucsd.edu/yeolab/Papers; and differential gene expression testing was performed using edgeR accounting for tag-wise overdispersion using R code provided by [91] in the edgeR user's guide on Bioconductor, and with DESeq2 ([92]).

A549 cells treated with dexamethasone

The ENCODE dataset, wgEncodeHaibRnaSeqA549Dex100nm, consists of 4 total samples from A549 cells, adenocarcinomic human alveolar basal epithelial cells: 2 replicates treated with 100μM of dexamethasone (DEX) and 2 controls treated with 0.02% ethanol solution. We performed differential gene expression testing using edgeR accounting for tag-wise overdispersion and with DESeq2.

Tunicamycin-treated mice embryonic fibroblasts

Embryonic fibroblasts from transgenic mice [93] were treated with tunicamycin for 10 hours. Two treated samples were compared to two controls. Expression data was downloaded from GEO (ascension number GSE35681). We performed differential gene expression testing using edgeR accounting for tag-wise overdispersion and with DESeq2.

3.2.3 Description of Permutations

We performed two sets of permutations: "permuted within bins" and "permuted overall." Permuted data was then tested for enrichment of Gene Ontology (GO) terms. When data was permuted within bins, the input data, which includes three columns: gene ID, differential gene expression p-value and the read count, is ordered by read count and divided into bins of 100 genes. Per each group of 100 genes, p-value and read count is randomized. This permutation scenario preserves any relationship between read count and p-values but removes any association between p-values and

GO term membership. In the second permutation scenario, p-values and read counts were randomized over all genes, which removes their relationship as well as any association between p-values and GO term membership. The difference between the results of these two permutations is due to the effect of the relationship between read counts and significance values. For example, in the barplots, if a method is conservative for “permuted overall” but has an excess of p-values in the left-most bar for “permuted within bins”, then that method is not adequately adjusting for the relationship between p-values and read count.

For each dataset, we created 100 permuted datasets for each permutation scenario. The median p-value of each GO term was calculated across all permutations in a permutation scenario for each dataset. We tested each dataset with RNA-Enrich, random sets, GOseq, and DAVID.

3.2.4 Performance Comparison

To assess the type I error rate for RNA-Enrich, we created permuted datasets from two experiments. The first, prostate cancer LNCaP cells treated with dihydrotestosterone (DHT), an androgen hormone [90], showed increasing read counts with increasing significance (Figure 3.1a). The second, A549 cells treated with dexamethasone (ENCODE dataset wgEncodeHaibRnaSeqA549Dex100nm) showed steady read counts with increasing significance (Figure 3.1d).

The original datasets were sorted by read counts, and then within each bin of 100 genes, GO term membership, DE p-value and average read count were permuted as a group. This scenario preserved the association between p-values and read count but removed functional enrichment significance from the data, allowing us to assess type I error under the null hypothesis and given the observed relationship with read count. We also tested the use of corrected fold changes instead of p-values; in this case the relationship with read count differed by dataset, but still existed. For both the LNCaP dataset and the A549 dataset, 100 permutations were performed. Each original dataset and permutation was tested using RNA-Enrich, the random sets method, GOseq and DAVID for all GO terms containing 10 – 500 genes. The median p-value of each GO term was calculated across all permutations for each dataset. We also provide results for a simpler type of permutations, where data was permuted over all genes; this does not preserve the association between DE p-values and read count, and is thus an estimate of what type I error would be if no relationship with read count existed.

3.2.5 Other comparisons

We used GOseq version 1.18.0 on R version 3.1.1 for enrichment testing on the original datasets and the permuted datasets. Average normalized read counts were provided as the bias data. We tested the original and the permuted datasets using both the sampling method and Wallenius approximation [9].

We reimplemented DAVID [39], which uses a modified Fisher's exact test, to use in R in order to test the same GO database that was tested with RNA-Enrich, the random sets method, and GOseq.

We calculated a corrected fold change (cFC) = log2((X+C) / (Y+C)) where C = 10th percentile of read counts, X is the average normalized read count per gene for treatment cases, and Y is the average normalized read count per gene for control cases. Each sample's read counts were normalized for library size by dividing by total number of reads of sample, and then multiplied by average read count across all samples. cFC was used in place of $–\log_{10}$(p-value) in RNA-Enrich and random sets.

3.3 Results & Discussion

3.3.1 Method performance with permutated data

Using permuted datasets we compared the type I error of RNA-Enrich to the random sets method (does not account for any bias in the data), GOseq (can adjust for read counts, but using a cut-off based method), and DAVID (does not adjust for read counts, and uses a cut-off based method). We show that when there is a relationship between read count and $–\log_{10}$(p-values), adjusting for read count improves the type I error rate compared to random sets (Figure 3.1a-c and Figure 3.2). Without adjusting for read count, 37 GO terms were enriched in the permuted data for random sets but only 3 for RNA-Enrich (q-value≤0.05). When the relationship does not exist, as is observed in the A549 dataset, RNA-Enrich and random sets have nearly identical type I error rates (Figure 3.1d-f). DAVID had 0 GO terms enriched in the permuted data for the LNCap experiment, but its type I error was overly conservative in cases where no relationship exists (Figure 3.3). RNA-Enrich provides a diagnostic plot for the user to determine if a relationship between read count and $–\log_{10}$(p-value) exists in their data (Figure 3.1a,d). If a relationship does exist, we recommend using RNA-Enrich to provide more biologically relevant results. RNA-Enrich also has favorable type I error rate compared to GOseq and DAVID (Figure 3.3, Figure 3.4, Figure 3.5). The performance of RNA-Enrich with p-values from DESeq2 instead of edgeR resulted in the same conclusions (Figure 3.6). Use of corrected fold change instead of p-values as input showed a different relationship exists, but similarly resulted in a benefit for RNA-Enrich compared to random sets (Figure 3.7).

3.3.2 Method performance with experimental results

Using RNA-Enrich with the LNCaP cells treated with DHT we found 192 enriched GO terms (q-value ≤0.05) (Table 3.1 and Supplementary Table 3.1). In comparison, the random sets, GOseq, and DAVID methods identified 35, 8, and 30 enriched GO terms, respectively. We tested a second dataset, mice embryonic fibroblasts treated with tunicamycin, that also revealed a relationship between read counts and significance levels, and resulted in conclusions similar to the LNCaP dataset (Figure 3.8, Figure 3.9, Figure 3.10). Again, RNA-Enrich detected more GO terms than the alternatives (Figure 3.8).

In the A549 dataset, we did not expect an advantage to RNA-Enrich over random sets, since there was no observed relationship between read count and significance levels. RNA-Enrich found 367 enriched GO terms including *negative regulation of transcription*, *vasculature development* and *fat cell differentiation* – all top ranked enriched GO terms also found by random sets and GOseq. Random sets and GOseq identified 347 and 363 GO terms, respectively. Based on Figure 3.1e,f and our overall findings, RNA-Enrich has the desirable property of reducing to the random sets method when no relationship with read count exists.

3.4 Figures and Tables

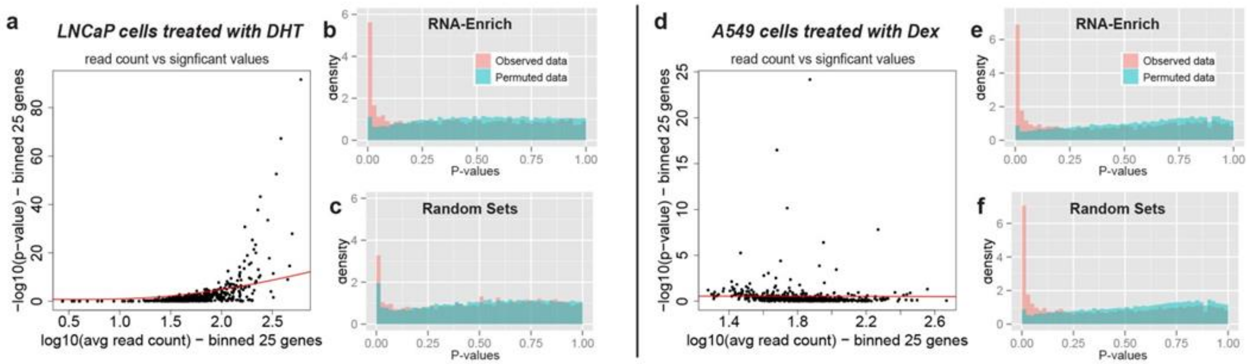


Figure 3.1. Comparison of RNA-Enrich with random sets with two datasets exhibiting two different p-value to average read count trends. Permuted data is “permuted within bins.” (a) RNA-seq data from LNCaP cells treated with DHT compared to a control showed a relationship between average gene read count and –log10(p-values) from DE tests. (b-c) Histogram of permutation p-values (teal color) should be uniformly distributed for acceptable type I error rate. For RNA-Enrich, the type I error rate is approximately uniform (b), but for the random sets approach for which there is no correction, more p-values are significant than expected (c). With the original data, RNA-Enrich identifies more significant GO terms than the random sets method (pink color). (d) RNA-seq data from A549 cells treated with Dex compared to ethanol showed no relationship between read count and –log10(p-values). (e-f) With or without the read count bias correction, type I error rate is approximately uniform, indicating that no correction is needed and either test is valid.

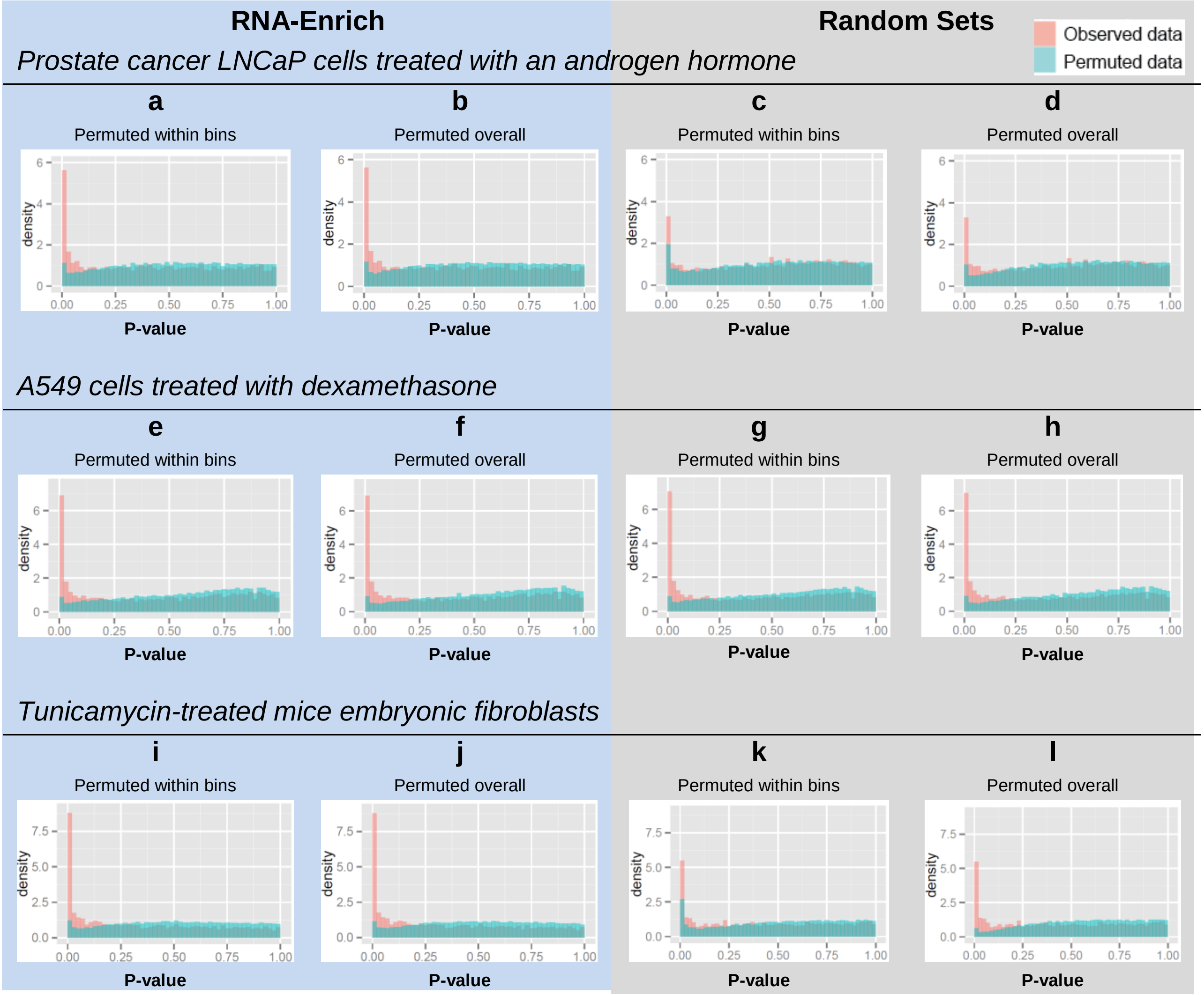


Figure 3.2. RNA-Enrich and random sets performance on permutated data. P-value distributions of RNA-Enrich versus the random sets method on datasets that were "permuted within bins" and "permuted overall." Histograms of permutation p-values (teal color) should be uniformly distributed for acceptable type I error rate. Both the prostate cancer LNCaP dataset (a-d) and a third dataset of tunicamycin-treated mice embryonic fibroblasts (i-l) exhibited a positive relationship between read count and DE p-values. When enrichment testing was applied to the permuted datasets that were permuted within bins, which preserves that relationship, RNA-Enrich had a better type I error rate than the random sets method, which shows an excess of low p-values when data is permuted within bins (c,k). When data is permuted overall (b,d,f,h,j,l), the type I error of RNA-Enrich and random sets is similar, which implies that the difference between methods observed for the "permuted within bins" scenario is due to random sets' inability to adjust for the effect of read counts. RNA-Enrich calls more significantly enriched gene sets than random sets for this dataset. When RNA-Enrich and random sets were applied to the A549 cells dataset, type I error was very similar between the two methods and both methods called a similar number and list of enriched gene sets.

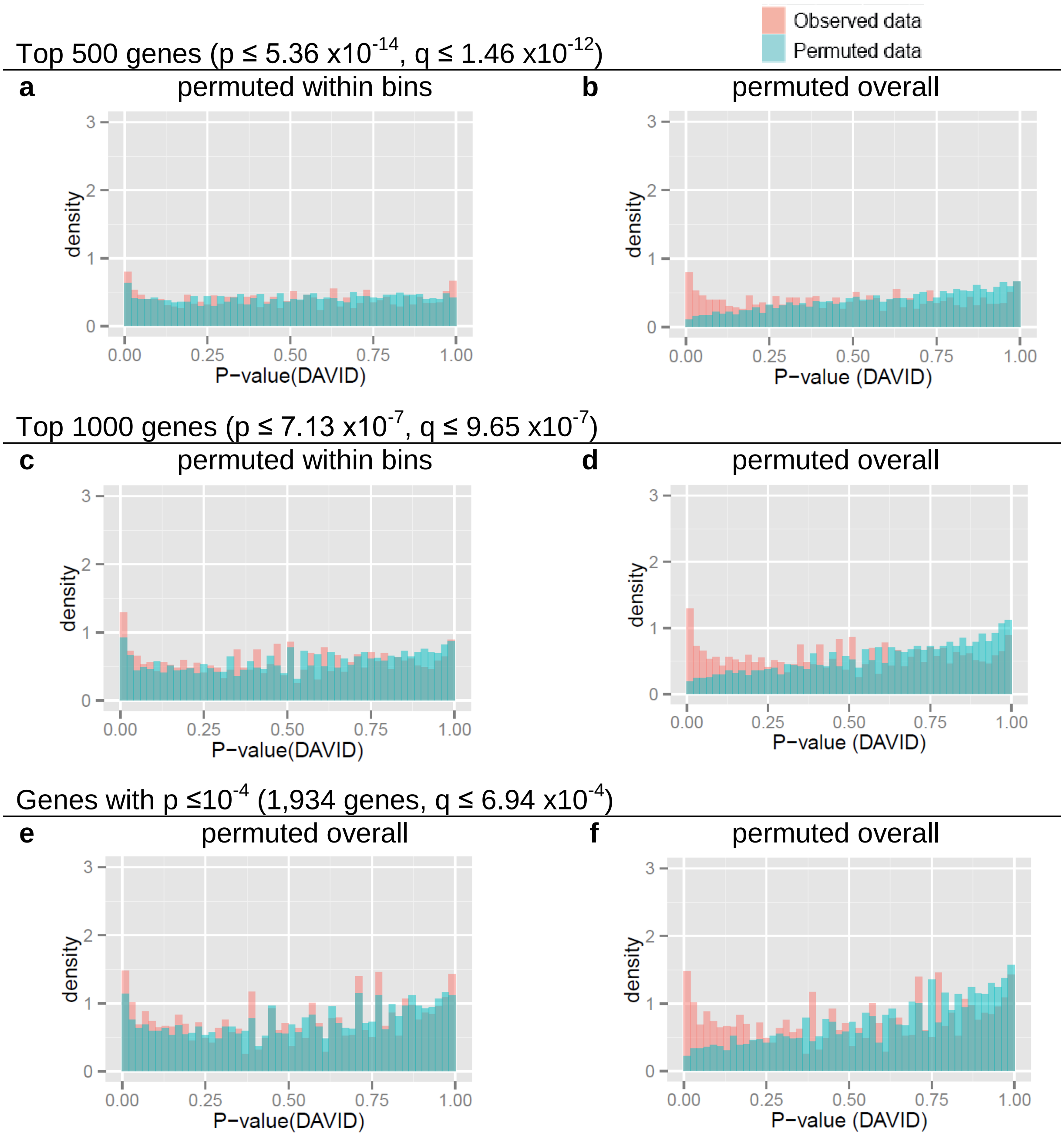


Figure 3.3. DAVID results: Prostate cancer LNCaP cells treated with an androgen hormone. P-value distributions from DAVID with the LNCaP dataset. Histograms of permutation p-values (teal color) should be uniformly distributed for acceptable type I error rate. (a, c, e) When the permuted data retains the relationship between read counts and DE p-values, DAVID calls more significantly enriched GO terms when the cutoff for differential expression includes more genes. (b, d, f) When the permuted data no longer has a relationship between read counts and DE p-values, the type I error rate of DAVID is conservative.

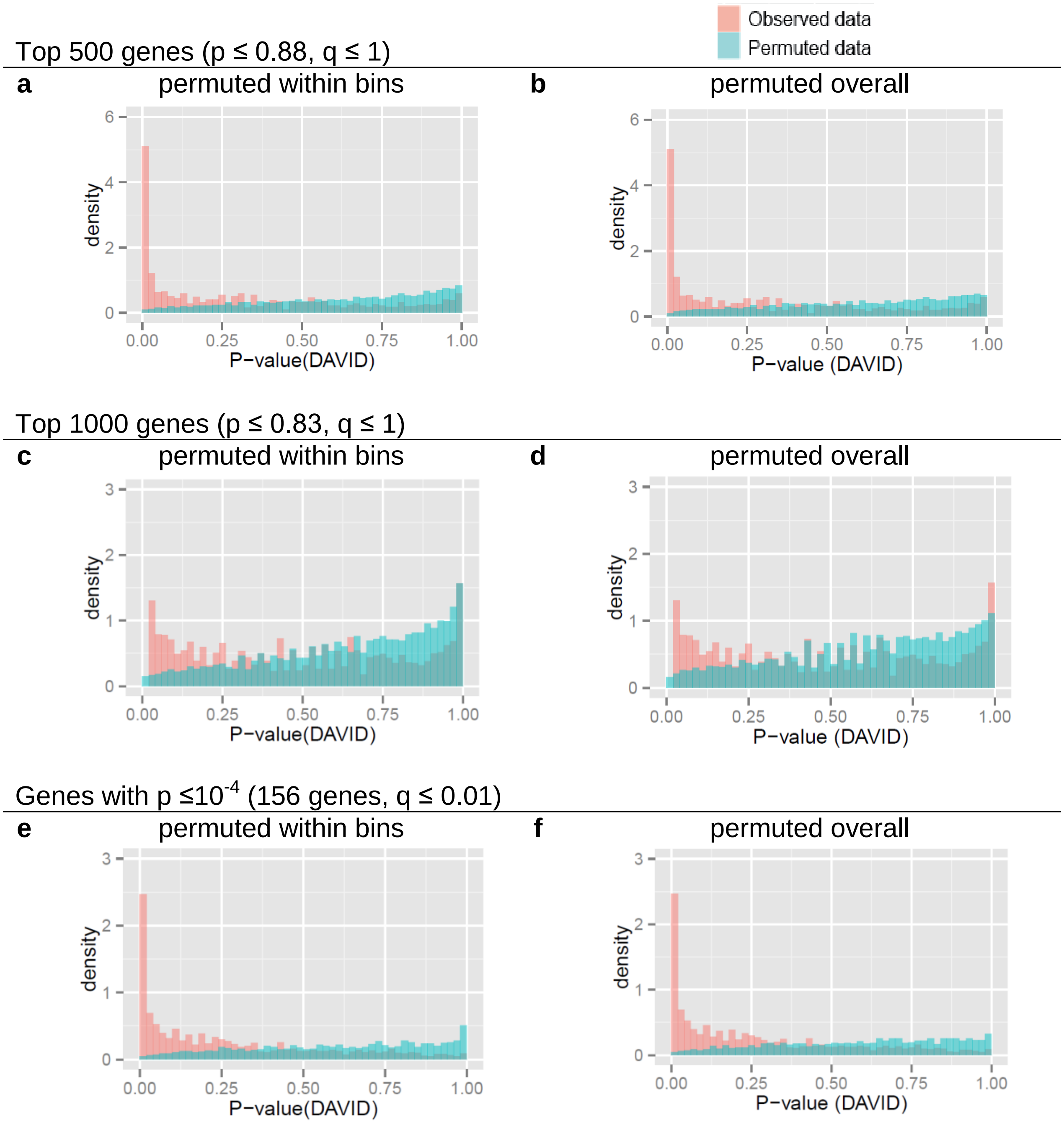


Figure 3.4. DAVID results: A549 cells treated with dexamethasone. P-value distributions from DAVID with the A549 cells dataset. Histograms of permutation p-values (teal color) should be uniformly distributed for acceptable type I error rate. (a-f) The type 1 error rate for DAVID is similar when testing on data permuted within bins (retains relationship between read count and DE p-values) and permuted overall (relationship is not retained), because this data set does not have a relationship between read count and differential gene expression p-values. The type I error rate of DAVID is somewhat conservative.

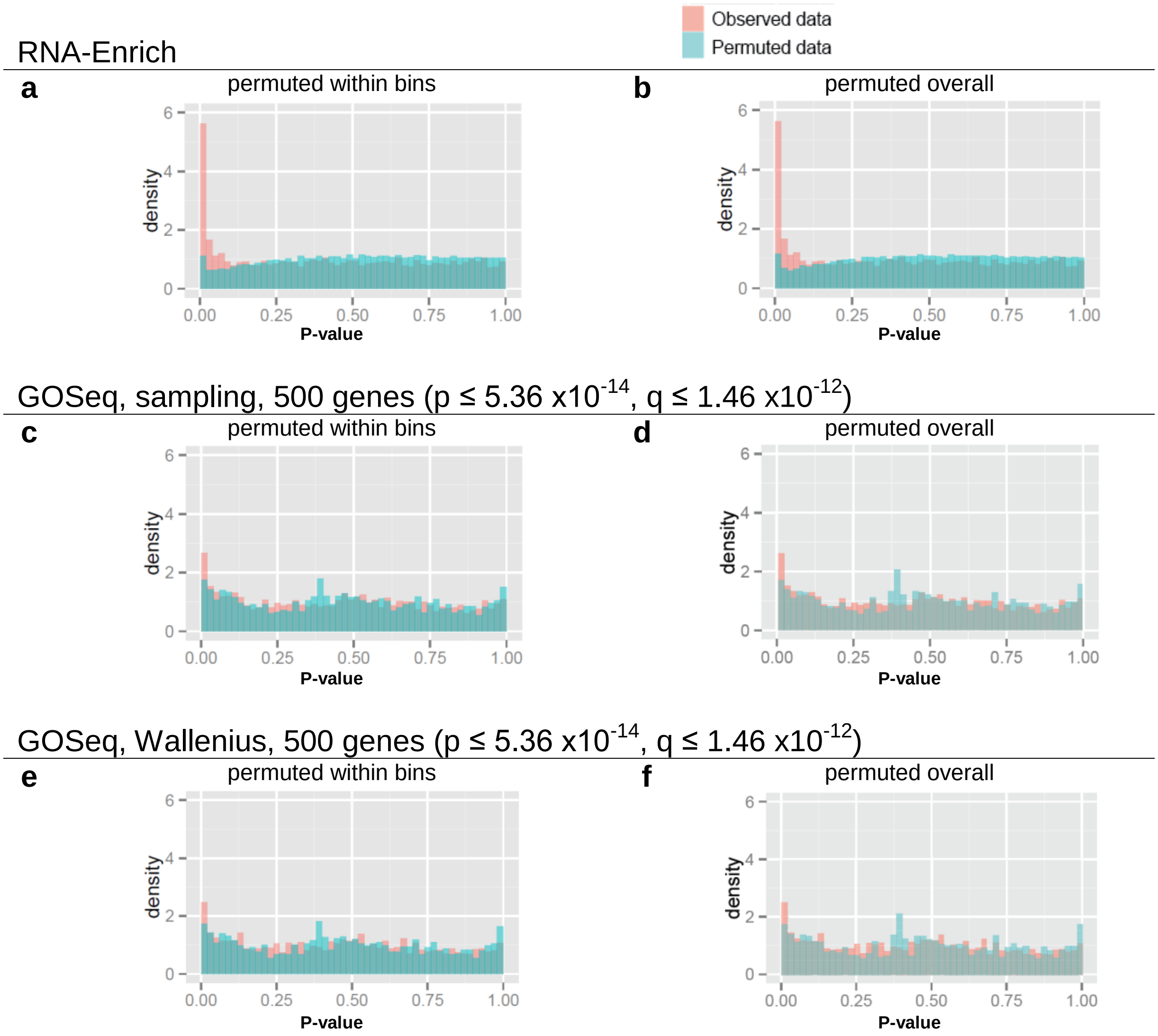


Figure continues on next page…

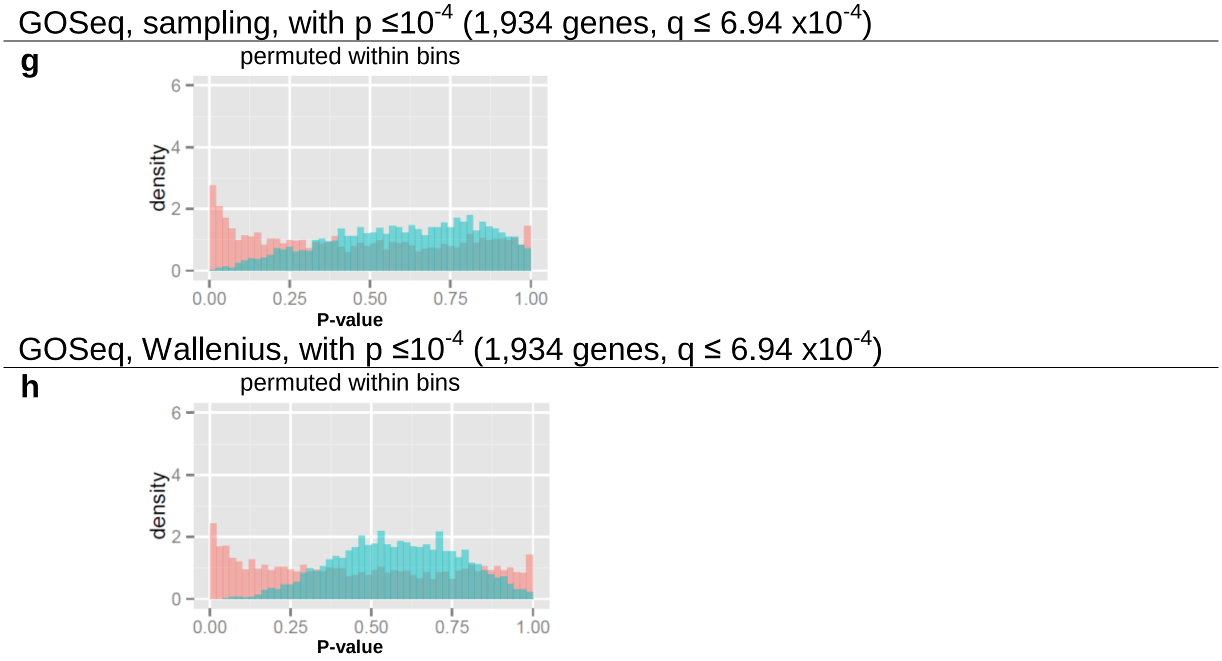


Figure 3.5. RNA-Enrich vs GOseq: Prostate cancer LNCaP cells treated with an androgen hormone. RNA-Enrich versus GOseq for LNCaP dataset. Histograms of permutation p-values (teal color) should be uniformly distributed for acceptable type I error rate. For the GOseq analysis, we used two different cut-offs to define differentially expressed genes: the top 500 significant genes, and genes with p-value ≤10^{-4}. We tested GOseq using both the default sampling method as well as the Wallenius approximation method. (a-b) For RNA-Enrich, the type I error rate is approximately uniform. (c-d) GOseq using the sampling method and a cutoff of the top 500 genes to define differentially expressed genes, results in an approximately uniform distribution. (a,b) With the original data, RNA-Enrich identifies more significant GO terms than GOseq (pink color). (e,f) GOseq using the Wallenius approximation and a cutoff of the top 500 genes to define differentially expressed genes. Again, the type I error rate is acceptable. (g) GOseq using the sampling method and genes with p- value ≤10^{-4}. (h) GOseq using the Wallenius approximation and genes with p- value ≤10^{-4}. (d,f) Type I error of RNA-Enrich and GOseq using permuted data that was permuted overall, which removes any relationship between read count and differential gene expression p-values, was similar to that using the data that was permuted within bins. This suggests both methods do account for that bias. However, all cutoffs we tried with GOseq resulted in less significant GO terms than with RNA-Enrich.

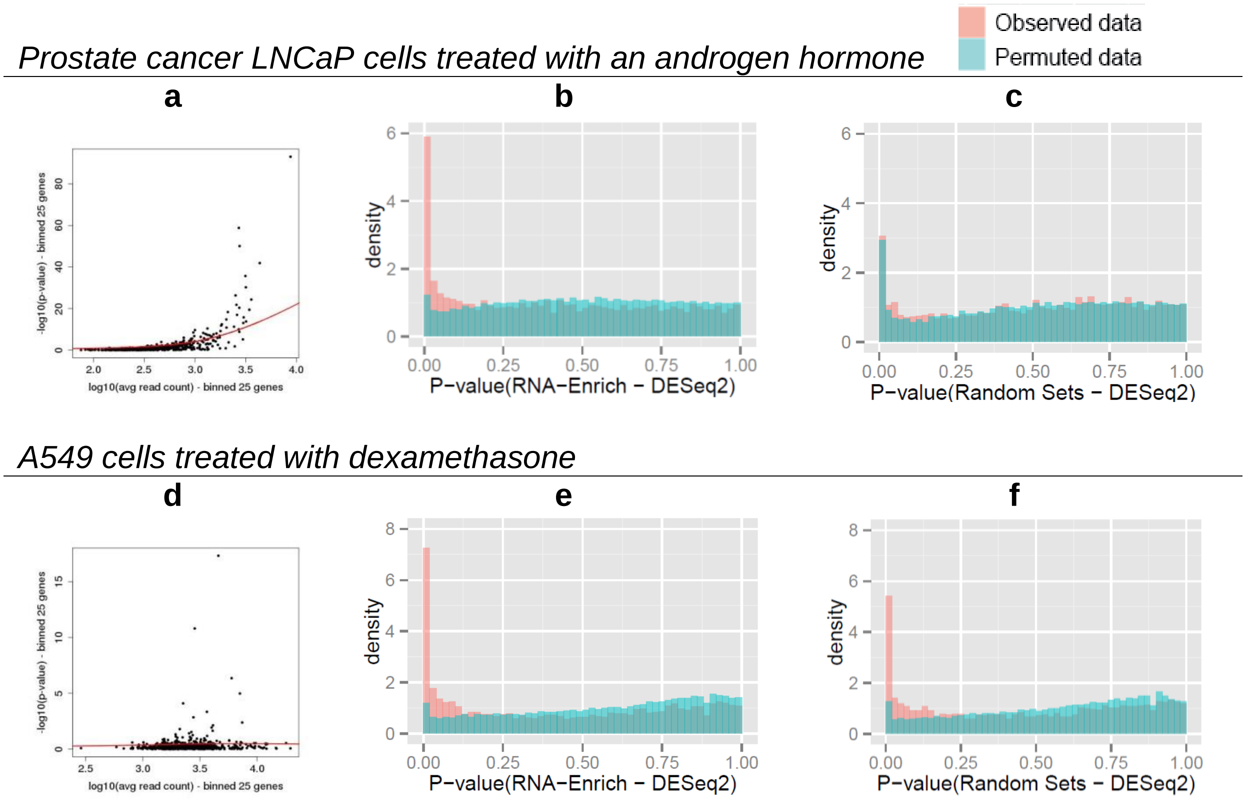


Figure 3.6. Comparison of RNA-Enrich and Random Sets using DESeq2 differential gene expression significance values. Relationships between read counts and significance values, and p-value distributions for RNA-Enrich and random sets using p-values from DESeq2 instead of edgeR. Results were similar between the two differential gene expression tests. (a,d) Relationship between read count and differential gene expression p-values remain the same for each dataset as they appeared for edgeR. (b,e; teal color) Type I error of RNA-Enrich remains improved over (c,f; teal color) random sets when there exists a relationship between read count and DE p-value (b,c) and is the same when there is no relationship (e,f). RNA-Enrich still appears to have improved detection power (b,c,e,f; pink color).

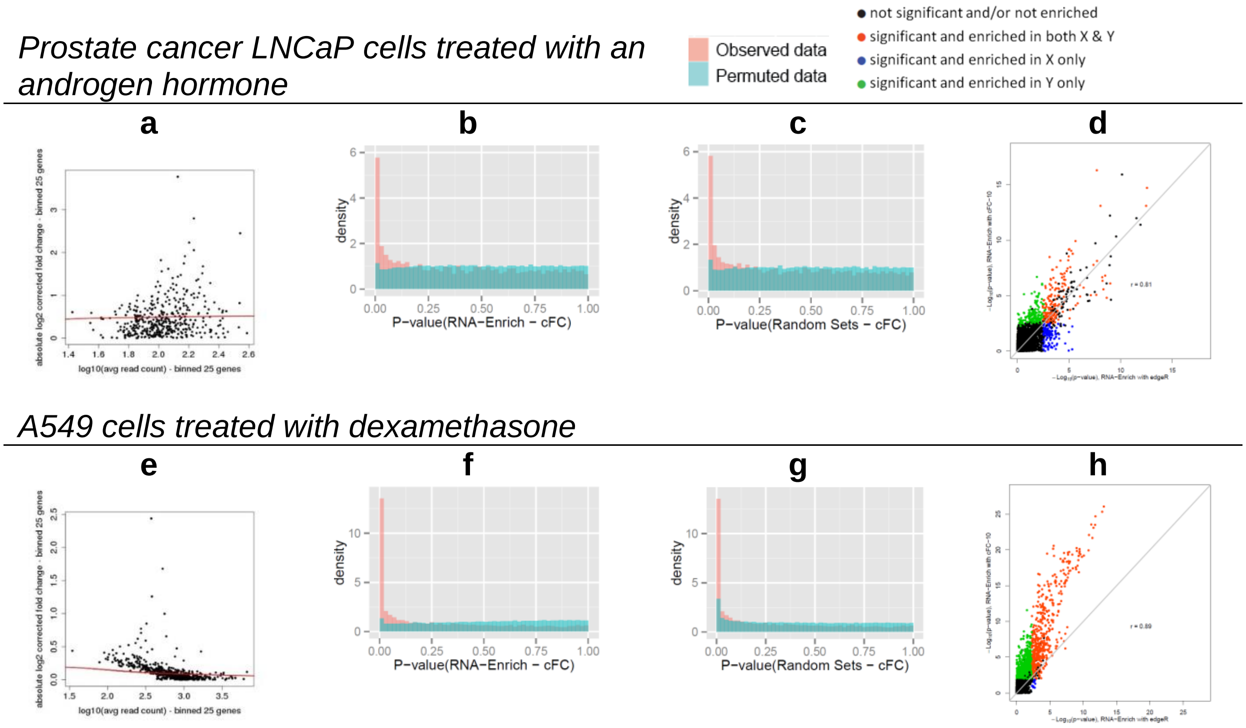


Figure 3.7. Comparison of RNA-Enrich and Random Sets using corrected fold change as significance values. Results for RNA-Enrich and random sets using corrected fold change (cFC) = log((X+C) / (Y+C)) where C = 10th percentile of read counts, X is average normalized read count per gene for treatment cases, and Y is average normalized read count per gene for control cases. (a,e) Relationship between read counts and cFC exists for A549 dataset, but not for the LNCaP dataset, which is opposite from what was observed based on p-values. Permutation p-value distributions for RNA-Enrich (b,f; teal color) compared to random sets (c,g; teal color) show that random sets has more lower p-values than expected for the A549 dataset, as evident in the higher bar at p=0 – 0.02. (d,h) GSE –$\log_{10}$(p-values) were highly correlated between RNA-Enrich using cFC and RNA-Enrich using p-values from edgeR.

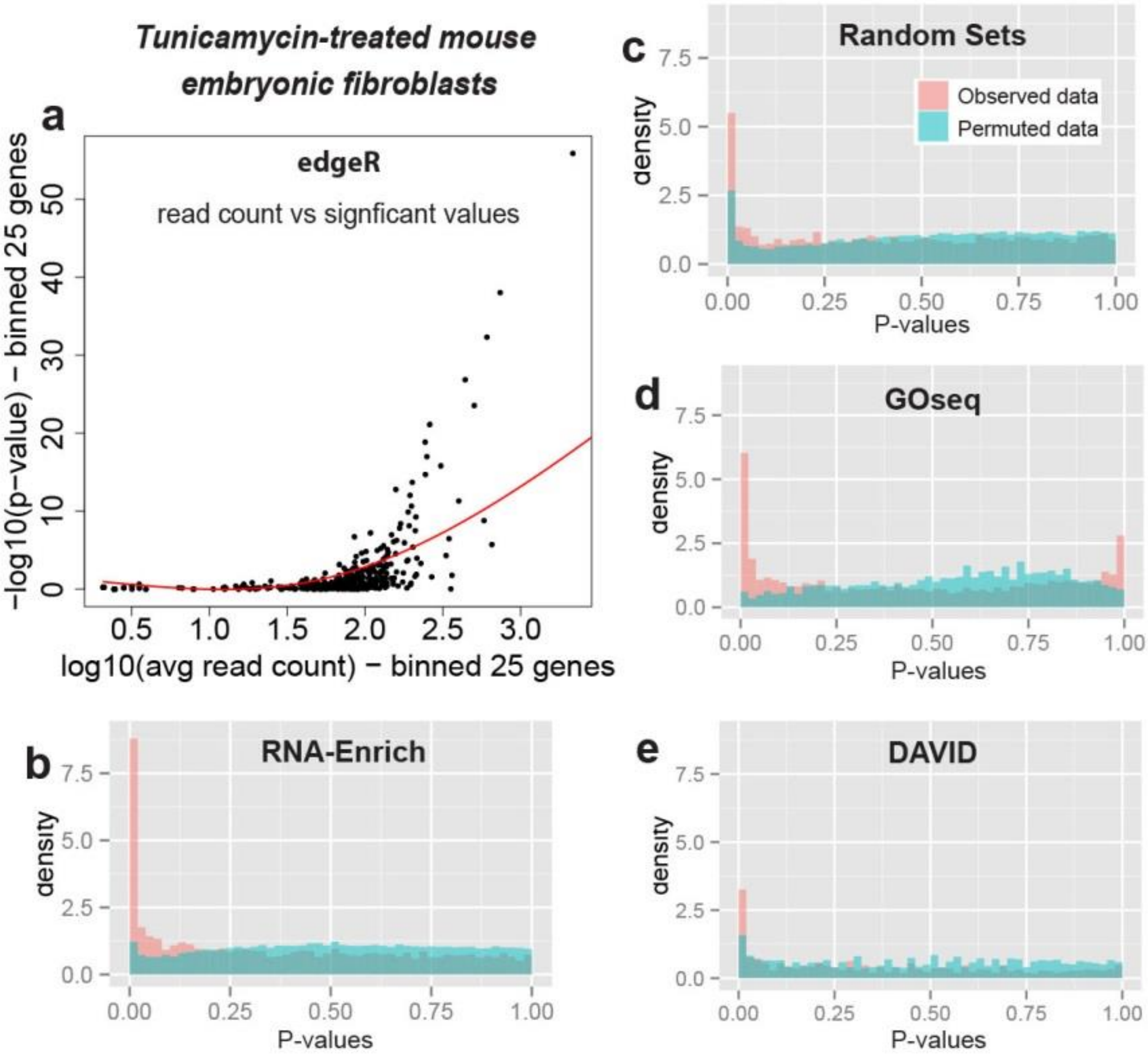


Figure 3.8. RNA-Enrich vs random sets in tunicamycin-treated mouse embryonic fibroblasts dataset using edgeR. (a) RNA-seq data from tunicamycin-treated mouse embryonic fibroblasts also showed a relationship between average gene read count and $–\log_{10}$(p-values) from differential expression tests using edgeR. (b-e) Histograms of permutation p-values (teal color) should be uniformly distributed for acceptable type I error rate. (b) For RNA-Enrich, the type I error rate is again approximately uniform. (c) For the random sets approach for which there is no correction, more p-values are significant than expected. With the original data, RNA-Enrich identified more significant GO terms (548) than the random sets method (310) (pink color). (d) Using GOseq (sampling method) and a cutoff of $q≤10^{-4}$ (795 genes) to define differentially expressed genes, 158 GO terms were enriched. (e) Using DAVID with the top 1,000 genes (p ≤ 4.44 x10-5, q ≤ 7.80 x10-4), 274 GO terms were enriched.

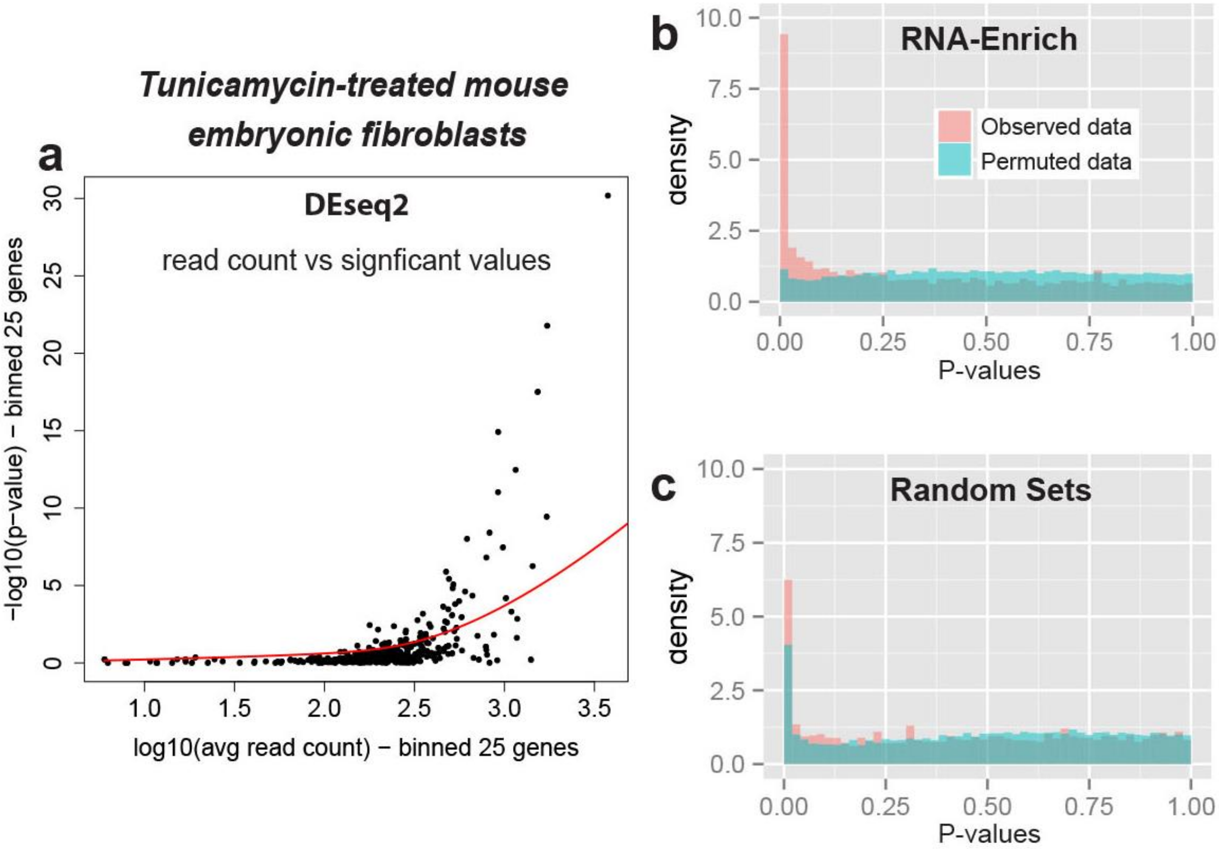


Figure 3.9. RNA-Enrich vs random sets in tunicamycin-treated mouse embryonic fibroblasts dataset using DEseq2. (a) RNA-seq data from tunicamycin-treated mouse embryonic fibroblasts also showed a relationship between average gene read count and –log10(p-values) from differential expression tests using DEseq2. (b-c) Histograms of permutation p-values (teal color) should be uniformly distributed for acceptable type I error rate. (b) For RNA-Enrich, the type I error rate is again approximately uniform. (c) For the random sets approach for which there is no correction, more p-values are significant than expected. With the original data, RNA-Enrich identified more significant GO terms (771) than the random sets method (415) (pink color).

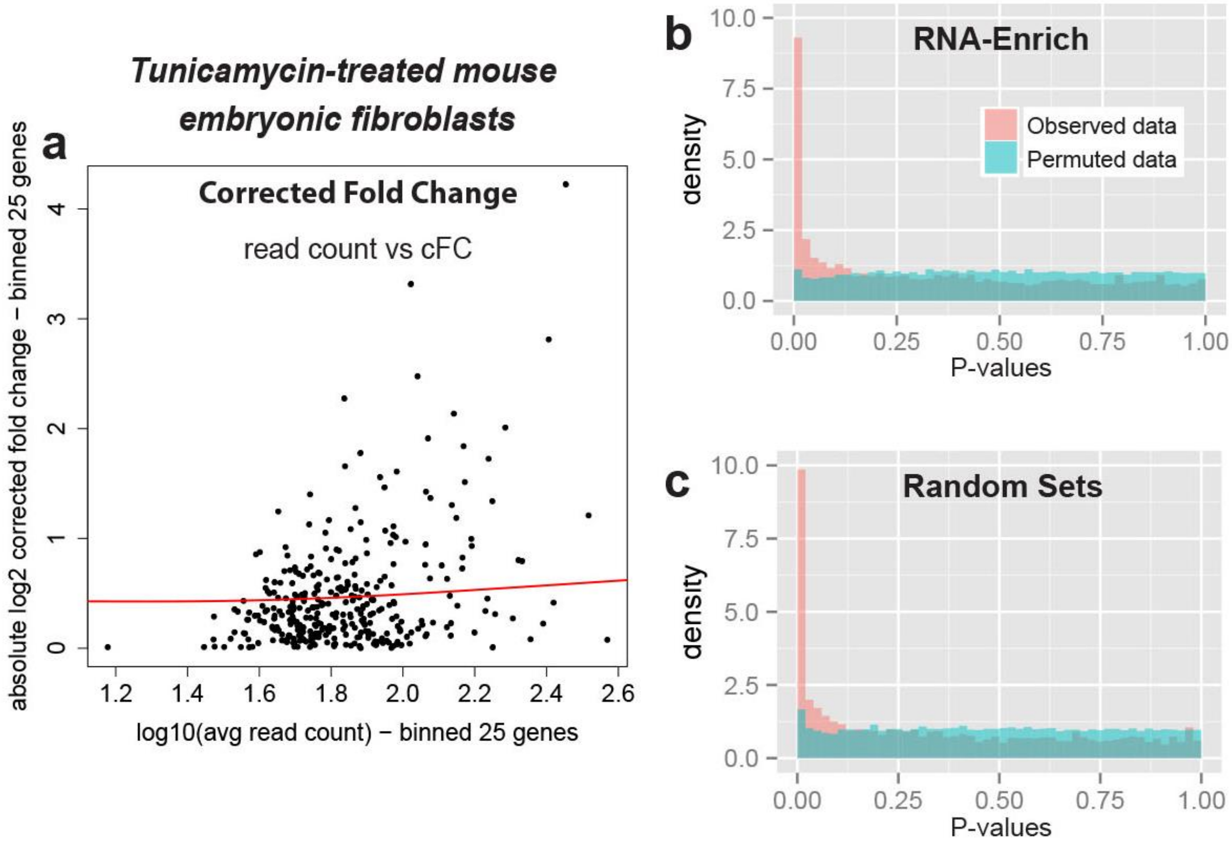


Figure 3.10. RNA-Enrich vs random sets in tunicamycin-treated mouse embryonic fibroblasts dataset using corrected fold change. (a) RNA-seq data from tunicamycin-treated mouse embryonic fibroblasts showed only a very slight relationship between average gene read count and cFC. (b-c) Histograms of permutation p-values (teal color) should be uniformly distributed for acceptable type I error rate. (b) For RNA-Enrich, the type I error rate is again approximately uniform. (c) For the random sets approach for which there is no correction, only slightly more p-values are significant than expected. In this case with the original data, RNA-Enrich identified 758 significant GO terms and the random sets method identified 842 GO terms as significant (pink color).

Table 3.1. Top ranked GO terms from RNA-Enrich for LNCaP cell line treated with DHT. Results shown are limited to the top unrelated GO terms.

Rank	GO term	P-value	FDR
1	extracellular space	2.3 x 10^{-8}	1.6 x 10^{-6}
2	vasculature development	4.6 x 10^{-9}	3.4 x 10^{-6}
7	signaling receptor activity	3.9 x 10^{-7}	1.3 x 10^{-4}
9	epithelial cell differentiation	2.3 x 10^{-6}	5.0 x 10^{-4}
10	cellular biogenic amine metabolic process	2.4 x 10^{-6}	5.0 x 10^{-4}
12	response to endoplasmic reticulum stress	8.9 x 10^{-6}	1.3 x 10^{-3}

Supplementary Table 3.1. Extended version of Table 3.1.

Chapter 4 Transposons, segmental duplications, sequence mappability, and gene length: Deciphering their relationship with gene function

4.1 Introduction

The ability to uniquely map short DNA read sequences to a reference genome (referred to as mappability) varies significantly across the mammalian genome [88, 94], with repetitive and duplicated regions generally reducing mappability. Mappability tracks are available on the UCSC Genome Browser for the human and mouse genomes [88] and some peak finders for the discovery of DNA binding sites in ChIP-seq data, such as PeakSeq [88], MOSAICS [18] and MUSIC [95], can adjust for mappability when calling peaks. Others have noted read bias in copy number variation calls due to mappability and have proposed methods to adjust for this [96]. Although mappability is a technical factor, it can serve as a measure of sequence uniqueness for any genomic region of interest, and may help to reveal the underlying regulatory architecture of the functional genome.

While mappability has not yet been studied in conjunction with gene function, there are several studies that suggest genes requiring more complex and tissue-specific regulation have at least some regions with high mappability, and vice versa. The creators of the CRG mappability tracks, which we utilize in this study, found that the approximately 900 olfactory receptor genes annotated in GENCODE were 10% less mappable than the average of protein-coding genes [94]. Olfactory receptor genes are highly paralogous. The majority of them in humans are no longer functional but are pseudogenes, and many have multiple copies [97], resulting in lower mappability. On the other end of the mappability spectrum, transposon-free regions are long stretches of sequence that are devoid of transposons (a class of repetitive elements), and therefore are more likely to be highly mappable. Though exons have a higher proportion of transposon-free regions, the majority of transposon-free regions occur in intronic and intergenic regions. The longest transposon-free regions are enriched in or near genes involved in DNA binding, regulation of transcription and development [21]. The distribution of transposon-free regions and their functional annotation suggest that certain genes need more finely tuned regulation, while other genes, such as olfactory receptor genes, do not need as finely tuned regulation and are more tolerant of repetitive element insertion, mutations, and may gain/lose copies over time.

Evolutionarily, the cell can fine-tune gene regulation by (1) having more potential regulatory space around the gene exons (i.e. longer intergenic distances and/or longer introns), (2) having more unique sequence space (e.g. fewer repetitive elements), or (3) both. Nelson *et al.* [98] have shown these two properties are related to regulatory complexity in *Drosophila melanogaster* and *Caenorhabditis elegans.* We hypothesize that in mammals, mappability, along with gene and intergenic length, can be used as an indirect measure of the complexity of a gene's regulation. Although mappability has been shown to bias read coverage and the results of multiple sequencing applications, little has been published on the relationship between mappability and gene function in mammals. To study this relationship, one needs to consider the main factors that contribute to mappability: transposons and segmental duplications.

Repetitive elements, such as transposons and segmental duplications, pose a major problem to sequence alignment. Transposons such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) , long terminal repeats (LTRs), and DNA transposons (transposons that do not require an RNA intermediate) together constitute an estimated 45% of the human genome [20, 21] (and possibly up to 66% [99]) . These regions have a high level of short sequence repeats and tend to be poorly mappable, although their mappability is correlated with age since older transposons have had more time to accumulate mutations. LINEs and SINEs are classes of autonomous mobile DNA sequences. Both LINEs and SINEs primarily move in a "copy and paste" manner using reverse transcriptase to insert DNA copies into the genome, however SINEs do not encode the reverse transcription machinery and rely upon reverse transcriptase produced from intact LINEs. LINEs and SINEs have been evolutionarily selected against in coding regions such as exons, but make up a large proportion of non-coding DNA, which include promoters, introns, and intergenic regions [100]. In the human genome, the most abundant transposons in the LINE and SINE classes are LINE-1 (L1) elements and *Alu* elements, respectively. L1 elements make up an estimated 15% of the human genome, while *Alu* elements make up an estimated 11% and are twice as abundant as L1 elements [22, 100, 101]. The *Alu* sequence is only 300bp, while the canonical L1 sequence is 6kb in length, however the majority of L1 elements in the genome are fragments of the original. *Alu* elements are more often observed in gene-rich regions, while L1s are enriched in gene-poor regions. Consequently, *Alu* elements are also enriched in R-bands (euchromatin) and CpG rich regions. However, *Alu* elements are more likely to be evolutionarily selected against over time compared to L1 elements [22].

Segmental duplications are long duplications of DNA sequence, inter- and intra-chromosomal, that are 1-200kb in length and have >90% identity. They make up about 5% of the human genome [102-105] and are enriched near centromeres and telomeres, as well as specific focal regions within euchromatin. It is estimated that 10.6% of highly identical (>98% identity) segmental duplications are paralogs, genes related by duplication that may evolve to have new functions [102]. Indeed segmental duplications have been implicated as the source of evolution of novel genes [106, 107], including genes that contributed to the divergence between humans and apes [108, 109].

While the distribution of repetitive elements and segmental duplications across the human genome is well-characterized, studies of their relationship with gene function have been limited. *Alu* elements have been shown to have a preference for (or are tolerated in) promoter regions of housekeeping genes over tissue-specific genes [110], and canresult in new binding sites for multiple transcription factors with single nucleotide mutations [111, 112]. An analysis of *Alu* distribution in chromosomes 21 and 22showed that *Alu* elements on these chromosomes were enriched in or near transport, metabolism, and signaling genes [113]. A recent study of L1 elements in coding regions found that these genes produced proteins such as transcriptional factors, and topoisomerase, and were involved in histone modification, RNA elongation, signal transduction, membrane receptors and extracellular growth factors [114]. L1 elements in intergenic regions have been suggested to help recruit the protein *Xist* RNA for X-inactivation [115].

Despite the above specific studies, to our knowledge, a genome-wide analysis of repetitive elements and segmental duplications and their relationship with gene function has not been conducted. One reason why this may not yet have been accomplished is that until recently, there was no functional enrichment testing method developed to handle data with these characteristics that have extremely high prevalence across the genome in both genic and intergenic regions. However, given an appropriate enrichment testing method, genome-wide functional enrichment of repetitive element families could elucidate evolutionary selection pressures, and identify which gene functions may benefit from maintaining or deleting repeats in the various regulatory regions of the associated genes.

Although little is known about the relationship between repetitive element families and gene function, the relationship between gene function and the length of a gene or the amount of intergenic distance surrounding a gene has been well-studied [38-40]. For example, genes involved in various developmental programs, nervous system related processes, and regulation of transcription (as determined by Gene Ontology (GO)) tend to be longer and/or have longer intergenic distances surrounding them. Conversely, genes involved in electron transport, chromatin assembly and organization, and the ribosome and rRNA processing tend to be shorter and/or have shorter intergenic distances.

Since repetitive elements and segmental duplications are the main contributors to mappability, we reasoned that if they are all independently depleted in the same gene functions as intergenic distance, it would be strong evidence that natural selection pressure drives these same gene functions to have overall higher levels of unique, potential regulatory sequence. We hypothesized that some or all of the gene functions that require complex regulation, also require a high level of uniqueness in the surrounding sequences. This would result in higher than average mappability levels and maintenance of long gene locus length that would provide protection from degradation of essential DNA sequence or distances for DNA loops. This leads to the hypothesis that genes with fewer transposons and segmental duplications in their surrounding regions, and with longer intronic/intergenic regions, require complex regulation. Conversely, we hypothesize that genes containing more transposons and segmental duplications require simpler regulation and/or may have adapted the repetitive elements for their own regulatory purposes.

Here we examine the levels of L1 elements, *Alu* elements, segmental duplications, and overall mappability in different genic and intergenic regions, as well as gene length, to discern significant patterns of enrichment or depletion across gene functions. We first show how average mappabiilty differs among gene locus regions (e.g. exons, introns, promoters). Secondly, we characterize the contribution of L1, *Alu*, and other repetitive elements, and segmental duplications to mappability across various regions in the human genome. We illustrate a strong relationship between repetitive elements and gene function and compare that with those of gene length and mappability. Our results suggest mappability could have important implications for interpretation of deep sequencing applications, and the evolutionary mechanisms used to achieve proper regulation.

4.2 Methods

4.2.1 Locus regions

We define the locus regions similarly to those in Welch and Lee, *et al* [116]. The *TSS extended* locus region is defined as the genomic region between the upstream and downstream midpoints between a gene and the TSS's of the two adjacent genes. This region represents an estimate of the entire region that is part of, or regulates, the gene. *≤5kb from TSS* is defined as the region within 5kb upstream and downstream of all TSSs in a gene. If TSSs from the adjacent gene(s) are less than 10kb away, we use the midpoint between the two TSSs as the boundary of the locus for each gene. We define *>5kb upstream from TSS* as the region between 5kb upstream of a TSS to the midpoint of the adjacent upstream TSS. We also defined two additional locus regions: *exons,* the exonic regions of each gene (in the case where exons from multiple transcripts of the same gene overlap, the region was reduced to one non-overlapping region), and *introns*, the intronic regions of each gene, that is the region between two non-overlapping exons of transcripts belonging to the same gene.

4.2.2 Mappability calculations

We calculated base pair mappability for reads of lengths 24, 36, 50, 75, and 100 base pairs derived from mappability data for *Homo sapiens* (build hg19) from the CRG mappability tracks on the UCSC Genome Browser [94]. An illustration of our mappability calculations is shown in Figure 4.1. Let B_i be the average read mappability of all possible reads of size K that encompass a specific base pair location, i, based on the mappability tracks from the UCSC browser. The values from the UCSC mappability tracks are the reciprocal of the number of mapped locations in the genome to which a read beginning at position j and extending for length K maps uniquely. A value of 1 indicates the read maps to one location in the genome. A value of 1/n indicates the read maps to n distinct places in the genome. We converted the values from the mappability track to binary (either uniquely mappable or not). For each base pair position, and then determined the proportion of reads overlapping that position that are uniquely mappable. We defined the mappability of a locus as the average of all base pair mappability values in the defined gene locus region. We chose to primarily use 50mer read length throughout the chapter due to the common choice of this read length, especially for ChIP-seq experiments.

4.2.3 Repetitive elements and segmental duplications

We defined repetitive elements using the repeat masker (“rmsk”) and segmental duplications (“superdupe”) tables from the UCSC genome browser. For *Alu* and L1 elements, we used the subset of the repeat masker table defined by the respective transposon family. Mappability of each non-overlapping region was calculated in the same way as locus definitions; that is, we averaged base pair mappability over the repetitive element or segmental duplication.

4.2.4 Contribution of repetitive elements and segmental duplications to mappability

To quantitate the contribution of each repeat type to mappability, we created a contribution score for each repeat type, *C*, which we define as 1 minus the mappability of the region where the repeat type overlapped with the loci, multiplied by the percent of the loci covered by the repeat type. We performed simple linear regression using the model:

$$mappability = \beta_1 C_{Alu} + \beta_2 C_{L1} + \beta_3 C_{Other} + \beta_4 C_{SegD}$$

and calculated the R^2 values for each factor and the overall model, which we report as the percent of mappability explained by repetitive elements and segmental duplications.

4.2.5 Gene set enrichment testing

We used RNA-Enrich [117] (http://lrpath.ncibi.org) to test for Gene Ontology (GO) terms enriched with genes with high or low mappability, which corrected for any effect of gene length on average mappability. For GO term enrichment testing of genes with long or short locus length, we use LRpath [36] in order to use a continuous measure for gene length rather than a method that requires a cut-off to define short and long genes. For GO term enrichment testing of the *Alu* and L1 elements, and segmental duplications, we use Broad-Enrich [37]. Broad-Enrich tests for enrichment/depletion of percent coverage of repeats across gene loci, while correcting for locus length. This allows us to discover gene functions with genes whose loci consist majorly of repeats (enrichment) or rarely contain repeats (depletion). Since repeats occurred in almost all genes, using a method that reduces the repeat coverage to a binary measure would not have been appropriate. In all cases, we corrected p-values for multiple testing using the False Discover Rate (FDR) approach. We report results for GO terms with ≤500 genes to avoid overly broad terms.

4.2.6 Clustering

Gene sets that had significantly high or low mappability (q-value≤0.001) in at least two locus regions in at least 1 type of repeat (*Alu* elements*,* L1 elements, or segmental duplications), were included in the clustering. P-values from gene set enrichment testing were $\log_{10}$ transformed and then multiplied by -1 if the gene set was enriched (resulting in a positive number), thereby creating “signed” $-\log_{10}$(p-values). Hierarchical clustering was performed using uncentered correlation and average linkage. Clustering was performed with gene set enrichment results of GO terms across the five locus regions in the following order: *exons*, *introns*, *TSS extended*, *≤5kb from TSS,* and *>5kb upstream from TSS*. The order was chosen to show trends using different partitions of the genome, from coding (*exons*) to non-coding (*introns)* regions, overall regions that included the previous and intergenic regions, and then the potential regulatory regions surrounding a gene.

4.2.7 Motif discovery and mappability

We used MEME (Multiple EM for Motif Elicitation) to perform an unsupervised search for transcription factor motifs (with a maximum 21bp width) in the ENCODE ChIP-seq experiment for neuron-restrictive silencer factor (NRSF) in cell line K562, with 6,016 peaks called using the peak finder PePr. MEME output includes a logo of the motif as well as a position-specific probability matrix. Sequences for each NRSF peak were extracted using the UCSC genome browser. The position-specific probability matrix was used as a position weight matrix in the Bioconductor R package, *Biostrings*, to calculate all areas in the human genome where the motif matched ≥80%. Genomic coordinates for each instance where the motif was found was extended upstream and downstream by 100bp to simulate ChIP-seq peaks. Simulated peaks were overlapped with the NRSF peaks to determine potential peaks that were not identified by the peak finder. Mappability was calculated for each simulated peak that had a motif and compared between simulated peaks that overlapped and did not overlap with the NRSF peaks.

4.3 Results

4.3.1 Mappability varies by read length and region of gene

We first describe the mappability of each gene’s overall locus (defined as the gene body and its surrounding genomic region) and the mappability of specific regions for each gene, such as exons, introns, and upstream regions. We calculated the mappability at each base pair for read lengths ranging from 24 to 100mer for human

(Figure 4.1a; see *methods* for details). The overall percent mappability in the human genome ranged from 48.6% (for 24mers) up to 93.6% (100mers). As expected, longer reads had consistently higher average mappability (Figure 4.1). We then calculated the average mappability of the following locus regions for each gene: (1) *TSS extended* - the region between the midpoints of a gene's TSS and the upstream and downstream TSSs of adjacent genes, (2) *exons*, (3) *introns*, (4) *≤5kb from TSS*- the 5kb region upstream and downstream of a TSS, and (5) *>5kb upstream from TSS* – the region from 5kb upstream to the midpoint between the gene's TSS and the neighboring gene TSS (Figure 4.1b). The last four locus regions are encompassed in *TSS extended*.

In human, using the *TSS extended* regions, the percent of gene loci having an average mappability ≥ 90% for 24, 36, 50, 75, and 100mers was 0.21, 7.41, 24.2, 76.2, and 90.1 %, respectively. Although the results for 100mers was very high, still only 2.7% of gene loci had 100% unique mappability. For all read lengths, *exons* were the most mappable locus region; this was particularly pronounced for the shorter read lengths (Figure 4.2). Focusing on the widely used read length of 50nt, regions *>5kb upstream of TSS* tend to be least mappable, although the distributions of different locus regions across all genes are very similar (Figure 4.3a).

4.3.2 Mappability of repetitive elements and their distribution across genic regions

We next sought to understand how much various factors contribute to the mappability of different gene locus regions, and how they are distributed across genes. To accomplish this, we examined different locus regions across genes and calculated (1) the mappability of different repetitive elements and (2) the proportion of each region of each gene covered by these repetitive elements (Figure 4.3b-f). As expected, most repetitive elements were not highly mappable, however L1 elements were relatively highly mappable with an average of 86% mappabiilty (Figure 4.3d). Segmental duplications and *Alu* elements were least mappable with an average of 41% and 46% mappability, respectively. Other repetitive elements were relatively highly mappable with an average of 92% mappability. Overall, mappability of *Alu* and L1 elements were positively correlated with age as measured by percent divergence from consensus sequence (Figure 4.4).

Repetitive elements occurred in almost all *TSS extended* loci, with *Alu* elements in 98% of genes, L1 elements in 90% of genes, and other repetitive elements in 99% of genes. Segmental duplications were associated with only 27% of genes. As expected, repetitive elements were least prominent in *exons*, while segmental duplications were most prominent in *exons* and had a similar distribution of coverage for the other four locus regions (*introns*, *≤5kb from TSS*, and *>5kb upstream*, and *TSS extended*) (Figure 4.3). *Alu* elements on average covered 16-19% of a non-exonic locus region, while L1 elements covered 6-16% (Figure 4.3b-c). Compared to *Alu* elements, L1 elements had a more significant depletion in coverage in *introns* and *≤5kb of the TSS*. Both *Alu* elements and L1 elements were most prominent in the *>5kb upstream from TSS* loci, which contain distal regulatory regions.

Overall, repetitive element coverage negatively correlated with mappability. Loci with segmental duplications were among those with lower mappability (Figure 4.5). To quantitate the contribution of each repeat type to mappability, we created a contribution score that measures how much the *unmappability* (1-mappability) of a locus is due to coverage by a repeat type (see methods for further details). Contribution scores of repetitive elements and segmental duplications were highly correlated with loci mappability (Figure 4.6). We performed linear regression to determine separately and altogether how much repetitive elements and segmental duplications contribute to the mappability of different locus regions. In order from explaining most to least, the regions were *≤5kb of the TSS* (95.5%), *introns* (94.5%), *TSS extended* (90.3%), *>5kb upstream from TSS* (89.3%)*,* and *exons* (79.9%) (Figure 4.7). Thus, these repetitive elements account for the great majority of the unmappability of regions. The contribution to exon mappability may be lowest due to the low coverage of most repetitive elements in these regions and because exons may contain many duplicated regions too small to be defined as a segmental duplication.

Also interesting was the range of coverage proportions observed across genes both for *Alu* and L1 elements. For example, although on average only 15% of a gene's promoter region is covered by *Alu* elements, some genes had as high as >50% *Alu* element coverage in their promoter while others had <5%. For L1 elements, we observed that a certain group of genes had near 0% L1 coverage across their entire range, while the rest had a fairly spread out distribution, resulting in a bimodal distribution. This led us to wonder whether the types of genes at either end of these distributions were random, or whether they tend to belong to certain functions and processes. And if so, were they similar or different functions and processes for the different contributors to low mappability (*Alu*s, L1s, segmental duplications, etc)?

4.3.3 Repetitive elements and overall mappability in both genic and regulatory regions are associated with gene function and locus length

To assess gene functions that are significantly enriched with or depleted of different repetitive element types, we performed gene set enrichment tests using GO terms. For comparison, we also performed similar tests for gene functions that have significantly higher/lower than average overall mappability, and significantly longer/shorter than average locus length. We found gene length to be a potentially confounding factor for both mappability and repeat coverage, as the average mappability and coverages were correlated with locus length in various ways (Figure 4.8). Therefore, we used recently-developed tests that empirically adjust for locus length and were appropriate for testing mappability and all repeat types. For mappability testing, we used a method that automatically adjusts for any relationship with gene locus length [117]; for repetitive element testing, we used a method (Broad-Enrich [37]) that adjusts for locus length and models the proportion of each gene locus covered by a genomic region of interest. Both methods were developed by our group, and allowed us to carry out these analyses in an unbiased way.

For each type of repetitive element we found strong enrichments and depletions, with some overlapping GO terms across the different locus regions. For example, we found that *Alu* elements were most strongly depleted in transporter activity, transcription factor binding, and brain development; and enriched in immune related functions, and olfactory and xenobiotic processes. For L1 elements, we found depletion in early and nervous system development functions; and enrichment in olfactory processes, immune related functions, cellular organization, chromatin modifications, and RNA processes. Segmental duplications were not as strongly depleted/enriched among gene functions as *Alu* and L1 elements (i.e. gene set enrichment p-values were not as significant). Segmental duplications were depleted in early and nervous system development functions and processes involved in transcription regulation, and were enriched in cellular organization processes, hormone metabolism, cytokine activity, and xenobiotic processes. Other top enriched and depleted terms are shown in Table 4.1, and results of all GO terms in Supplementary Table 4.1. For locus length, our results are in strong agreement with previous reports [54].

Given the large number of tests performed (5 locus regions x 3 repeat types, plus overall mappability and locus length), we wished to visualize the results all together. We clustered significance values, using signed –log10(p-values) (negative for depleted terms, positive for enriched terms) (Figure 4.9). One of the most notable observations is that the overall mappability and locus length results are highly correlated with each other and negatively correlated with repetitive element coverages. That is, gene functions that have high mappability also tend to have long locus lengths; conversely, gene sets with low mappability tend to have shorter locus lengths, albeit with a lesser degree of agreement (Figure 4.10). Another notable observation is that while there are many gene functions consistently depleted in all three repeat types (*Alu* elements, L1 elements, and segmental duplications), and that had high mappability and long locus length, there were no gene functions that were consistently enriched in all three.

Overall, we identified 8 distinct clusters, each with a unique enrichment signature. Clusters 1-4 included GO terms that were lowly mappable and had shorter locus lengths (using *TSS extended*). Clusters 1 and 2 were generally depleted in segmental duplications but enriched with L1 elements. Cluster 1 was different than Cluster 2 in that Cluster 1 was highly enriched with L1 elements in exons and had shorter locus lengths. Cluster 1 included terms that were related to RNA (28.6%) - such as *RNA catabolic process*, *ribosome*, and *mRNA transport,* mitochondria - such as *mitochondrial matrix* and *mitochondrial respiratory chain,* and terms related to *transcription* and *translation*. Cluster 2 included terms that were related to cellular organization and division such as *microtubule*, *spindle assembly* and *kinetochore*, protein modifications such as *protein targeting* and *histone modification*, and chromatin such as *centrosome* and *chromosomal part*. Cluster 3 differed from Clusters 1 and 2 in that it was enriched with segmental duplications and depleted of L1 elements except in *TSS extended*, suggesting that L1 elements were enriched in regions downstream of the TSS. Cluster 3 was dominated by terms related to immune response–r*esponse to bacterium* and *innate immune response*, cell signaling such as *cytokine activity* and *steroid metabolic process*, and protein secretion. Cluster 4 was unlike cluster 3 in that it exhibited very strong enrichment of *Alu* elements. It included terms related to immunoglobulin binding, as well as GO terms like *mismatch repair, olfactory receptor activity,* and *heme binding*. GO terms in clusters 1-4 where those whose genes allowed the most repetitive elements of various types in their surrounding locus regions, suggesting they may be under strong positive selection.

The other half of the heatmap, Clusters 5-8, included GO terms that were highly mappable and had longer locus lengths. Cluster 5, which was depleted in *Alu* elements and segmental duplications, and mixed for L1s (enriched with L1 elements only in the region ≥*5kb upstream from TSS)* was dominated by terms related to kinase activity, transcription factor binding, actin, and regulation of GTPase. Cluster 6 was strongly depleted of L1 elements and somewhat enriched with segmental duplications; it included many terms related to ion channel activity and ion transportation, and some developmental terms such as *head development, vasculature development,* and *kidney development*. Cluster 7, the largest cluster, was depleted in *Alu* elements, L1 elements, and segmental duplications. The majority of this cluster is terms related to development, as 72% of the GO terms in this cluster included the word “development,” “formation,” “differentiation,” or the suffix “-genesis.” The development terms were related to organ formation, tissue development, embryo development and nervous system development, including terms such as *brain development*, *axon, dendrite*, and *synapse*. Non-developmental terms included *transcription factor complex* and *transcription regulatory region sequence-specific DNA binding*, processes where mutations or repetitive element insertion could cause widely detrimental *trans*-regulation effects. Cluster 8 was the smallest cluster, distinct in that it was enriched only with *Alu* elements. This cluster included specific terms such as *cranial nerve development, regulation of kidney* *development*, and *developmental pigmentation*. The full list of GO terms associated with each cluster is shown in Supplementary Table 4.2.

4.3.4 Comparison of mouse and human mappability and locus length

To determine the extent to which our above findings are limited to human, versus generalize to other mammals, we performed a simplified analysis in mouse testing for gene functions enriched with high/low mappability and long/short locus length. We found that GO terms with high mappability in human were also generally highly mappable in mouse (Figure 4.11a, r =0.35) such as terms related to nervous system development, early development, and transcription factor binding. Over 500 GO terms had significantly high mappability in both species using the *TSS extended* regions; these included terms related to nervous system development like *axon guidance, regulation of neurogenesis,* and *synaptic membrane*, as well as *pattern specification process, regionalization,* and *sequence-specific DNA binding RNA polymerase II transcription factor activity*. There was less agreement with lowly mappable terms, however there were still seventeen terms overlapping between the two species, including *olfactory receptor activity* and *defense response to bacterium* (Figure 4.11a). This divergence for lowly mappable gene functions suggests the possibility that the different repetitive element types in mouse compared to human have been adapted to help regulate different types of processes, and/or that different types of genes are more likely to have a segmental duplication. Overall, there was higher concordance between human and mouse for GO terms enriched with long genes or short genes than there was for mappability (Figure 4.11, r=0.35 for mappability, and r=0.59 for locus length). There were 445 GO terms that were both highly mappable and had long locus lengths in both human and mouse (q ≤0.05 and ≤500 genes); of those, 187 (42%) included the word “development”, “formation, “differentiation” or the suffix “-genesis.” The top ranked GO terms with high mappability and long locus length in both species were *regulation of neuron differentiation* and *morphogenesis of an epithelium.* Mouse enrichment results are provided in Supplementary Table 4.3.

4.3.5 Effect of mappability on ChIP-seq peak detection

Finally, we asked how variations in mappability might affect the ability of ChIP-seq to detect true DNA protein binding. Successful peak detection (detection of protein-DNA binding sites) requires the ability to uniquely align sequencing reads to the binding region. Binding sites in regulatory regions with lower mappability are less likely to be detected, because fewer reads can be mapped to those regions. To investigate this, we calculated the mappability of the region surrounding a DNA binding protein motif and asked if motifs under ChIP-seq peaks for that DNA binding protein had higher mappability than motifs not in a peak. Specifically, we used Motif Em for Motif Elicitation (MEME[118]) to scan for DNA binding motifs in 6,016 peaks (called by PePr [119]) from the ENCODE ChIP-seq experiment for neuron-restrictive silencer factor (NRSF) in cell line K562, an immortalised myelogenous leukemia cell line. As expected, the most prominent motif was for NRSF (Figure 4.12a). Using the resulting position weight matrix (PWM), we identified 17,021 predicted NRSF motif sites in the genome (hg19). Of these computationally predicted motif sites, 2,607 (15.3%) occurred in a ChIP-seq peak; these motif sites tended to occur near the center of the peaks (Figure 4.12b). We compared the mappability of motifs in ChIP-seq peaks to those outside of peaks, calculating the mappability of the regions 100bp up- and down-stream of each motif site. Motif regions in ChIP-seq peaks had significantly greater mappability than motif regions not in a peak ($p = 8.65 \times 10^{-6}$) (Figure 4.12c). A likely explanation is that many of the NRSF bound regions with lower mappability were not detected by the peak calling algorithm. If true, given the results of the previous sections, undetected binding sites may be in the loci of genes with low mappability; and thus the detected binding sites may be enriched for subsets of genes (and GO terms) with high mappability.

4.4 Discussion

The mappability of genomic loci is directly related to sequence uniqueness. Factors that affect the mappability of genomic regions include read length and presence of repetitive elements such as transposons and segmental duplications. Our partitioning of the genome into five different locus regions: (1) *TSS extended,* (2) *exons*, (3) *introns*, (4) ≤*5kb from TSS*, and (5) *>5kb upstream from TSS* showed different patterns of mappability and coverage of repetitive elements and segmental duplications. Repetitive elements and segmental duplications explained >90% of unmappable sites in all locus regions except exons. As expected, *exons* were generally most highly mappable and had the lowest amount of repetitive elements and segmental duplications. Any unmappability in *exons* not explained by repetitive elements and segmental duplications may be due to other factors, such as duplications not long enough to be categorized as segmental duplications, but are replicated to code for the same protein domain in various proteins. One example is tandemly duplicated exons, estimated to occur in about 10% of annotated genes in *Homo sapiens*, that are present in expressed sequence tags and cDNAs, and therefore are likely functional [120].

Loci with *Alu* and L1 elements and segmental duplications depleted or enriched with gene functions. Across the different repetitive elements and segmental duplications we examined, we showed that certain gene functions exhibit specific repeat enrichment signatures and many of these coincide with the enrichment signatures of overall mappability and locus length. Furthermore, results for repeat elements varied across different locus regions. We observed depletion, but never strong enrichment, of all three repeat types across all locus regions, implying that certain regions tend to tolerate a specific type of repeat, but do not tolerate repetitive elements in general.

The strong associations of *Alu* elements, L1 elements, and segmental duplications with gene function show that incorporation and proliferation of repeats is not random, and suggest locus-specific tolerance. Our results also suggest that the selection for one type of repeat over another may be because that repeat type is beneficial to regulation of genes involved in that function, and perhaps that repeat type has even been adapted for use by the genes, a process called exaptation. For example, cluster 3 of the heatmap (Figure 4.9) showed a signature highly enriched for *Alu* elements in genes involved with immune response. It has been previously shown that an *Alu* element had been exapted as a highly conserved binding site for the *CAMP* gene, which is regulated in the vitamin D pathway, a pathway involved in innate immune response in humans and primates [121].

The top enrichment results for segmental duplications were less significant than for *Alu* and L1 elements. Segmental duplications occur at a larger scale than transposons, often resulting in large blocks of duplications, and inter- and intra-chromosomal rearrangements; therefore, survival of segmental duplications may be less dependent on gene regulation mechanisms. Rather genes in gene functions enriched with segmental duplications may have one or a combination of favorable features: (1) tolerant of having duplications, (2) located in regions of chromosomes (pericentromeric, subtelomeric, and “duplication cores” [122]) that are more likely to be duplicated and re-arranged , and/or (3) have simpler regulation not requiring many unique sequences. For the gene functions in cluster 3 (Figure 4.9) that are enriched with segmental duplications and *Alu* elements, it is possible that these are the result of a positive relationship between segmental duplications and *Alu* elements. It has been shown that a significant amount of pericentromeric and interstitial (occurring in euchromatin) segmental duplications are enriched with *AluY* and *AluS,* younger subfamilies of *Alu* elements, at the boundaries of segmental duplication events [123, 124].

We consistently found developmental genes, especially those involved in nervous system development and early development, to be highly mappable, have longer locus length, and to be strongly depleted of *Alu* and L1 elements, and segmental duplications in multiple locus regions. While this is expected for coding regions like exons, depletion in surrounding intergenic regions suggest these functions require more unique sets of sequences and potentially preserved distances between them,to maintain proper regulation from promoter and enhancer regions. Genes in these processes tend to be consistent across everything tested, having both longer and more highly mappable intergenic space around them. These findings agree with and expand on those of Simons, et al. [21], i.e. that transposon-free regions are highly enriched in developmental genes. Interestingly, these genes may also be resistant to acquiring mutations. For example, ultra-conserved elements have been found to be enriched in “gene deserts” that are 10-100 kilobases (kb) away from known genes, where the closest flanking genes are associated with early development functions [125, 126]. Also consistent with the idea that developmental genes require complex regulation and therefore more sequence uniqueness, Lawson and Zhang [127] found that the 5’-UTRs of tissue-specific genes, which often require complex regulatory control, have significantly fewer simple sequence repeats than the 5'-UTRs of housekeeping genes. Tissue-specific regulation requires the cooperation of multiple transcription factors, and thus multiple binding locations [128], and genes with complex expression patterns often require long-range cis-regulatory elements. This was first revealed by the fact that intergenic chromosomal breaks in disease disrupted these genes [129].

The relationships we have described among transposons, segmental duplications, mappability, and gene locus length with gene function, suggest that mappability may need to be taken into account when looking at the biological relevance of genomic regions from genome-wide sequencing results. In particular, genes involved in biological functions tending to have high mappability will be more likely identified than genes involved in functions tending to have low mappability, simply due to higher sequence coverage. Our example of how mappability affects ChIP-seq peak detection with the NRSF dataset shows that even though many of our computationally predicted binding sites were likely false positives, we were still able to detect a significant shift towards higher mappability in the sites covered by a detected peak. These results are in agreement with those of Rozowsky, et al. [88] who found that mappability had a significant effect on modifying the overall ChIP-seq signal.

Our analyses were made possible by gene set enrichment tests [37, 117] that are able to account for confounding factors and to measure the proportion of loci covered by a repeat. It is well known that gene locus length can bias gene set enrichment tests [54], and it should therefore be accounted for. A previous study by Tsirigos, et al [130], who used a permutation-based approach to find gene functions enriched with *Alu* elements in the human genome did not take into account gene locus length. We compared our results to theirs and found little agreement. In the case of highly occurring genomic features, such as repetitive elements and segmental duplications, gene set enrichment tests that reduce the genomic features to a binary value or even the number of features, would not be as informative since, or almost all genes, would have that genomic feature and they can be various lengths. The method we used, Broad-Enrich [37], uses percentage of loci covered by the repeat type for gene set enrichment testing, and therefore is well-suited for *Alu* and L1 elements, and segmental duplications.

There are multiple limitations of our study and/or future directions for further analysis. Mappability may play an even larger role in bisulfite sequencing studies, where unmethylated cytosines are converted to uracil (and read as thymine on sequencers). In these studies, the genome is essentially reduced to a three letter alphabet, significantly reducing the unique information content of short read sequences. Although we identified many significant associations, our study was limited by our method of using nearest genes for defining gene regulatory regions. Future studies may take into account insulators (e.g., CTCF sites) and identified DNA loops between enhancer and proximal promoter regions, which could refine enrichment results. In our study, we focused on read lengths of 50bp. Although sequencers are now capable of longer reads, these short read results remain relevant due to the large number of publicly-available sequencing experiments available and still being extensively used that were performed with shorter reads. For example, nearly all of the ENCODE ChIP-seq data were performed with 35-40nt read length (ENCODE TFBS metadata, column Z). Finally, other classes and subclasses of LINEs and SINEs, besides *Alu* and L1 elements, may be associated with different gene functions. Liang, *et al.* [131] found that tandem repetitive elements, unlike transposable elements, were more highly enriched in genes involved with development and regulation functions, especially in 5' UTR regions. However, their analysis did not take into account gene locus length, and only looked at the presence (while we look at the proportion) of a repetitive element in 5'-, 3'-UTR, and set regions upstream and downstream of the transcription start/end sites. Our approach in this study can further be applied to other repeat types as well. Also of particular interest would be an analysis of subfamilies of *Alu* and L1 elements, as these subfamilies are of different ages and sequence similarity. Segmental duplications that are >10kb and >95% identity are more prone to duplication-mediated rearrangements and non-allelic homologous recombination [103, 132, 133]. Paralogous segmental duplications or segmental duplications of different lengths and identity may be associated with distinct gene functions, not captured in this analysis. Also of interest could be a comparison of pericentromeric, subtelomeric, and interstitial segmental duplications. There is much more to discover, but here we have described and applied the most comprehensive study of depletion and enrichment of highly occurring repetitive elements in gene functions across different gene locus regions in the human genome.

4.5 Figures and Tables

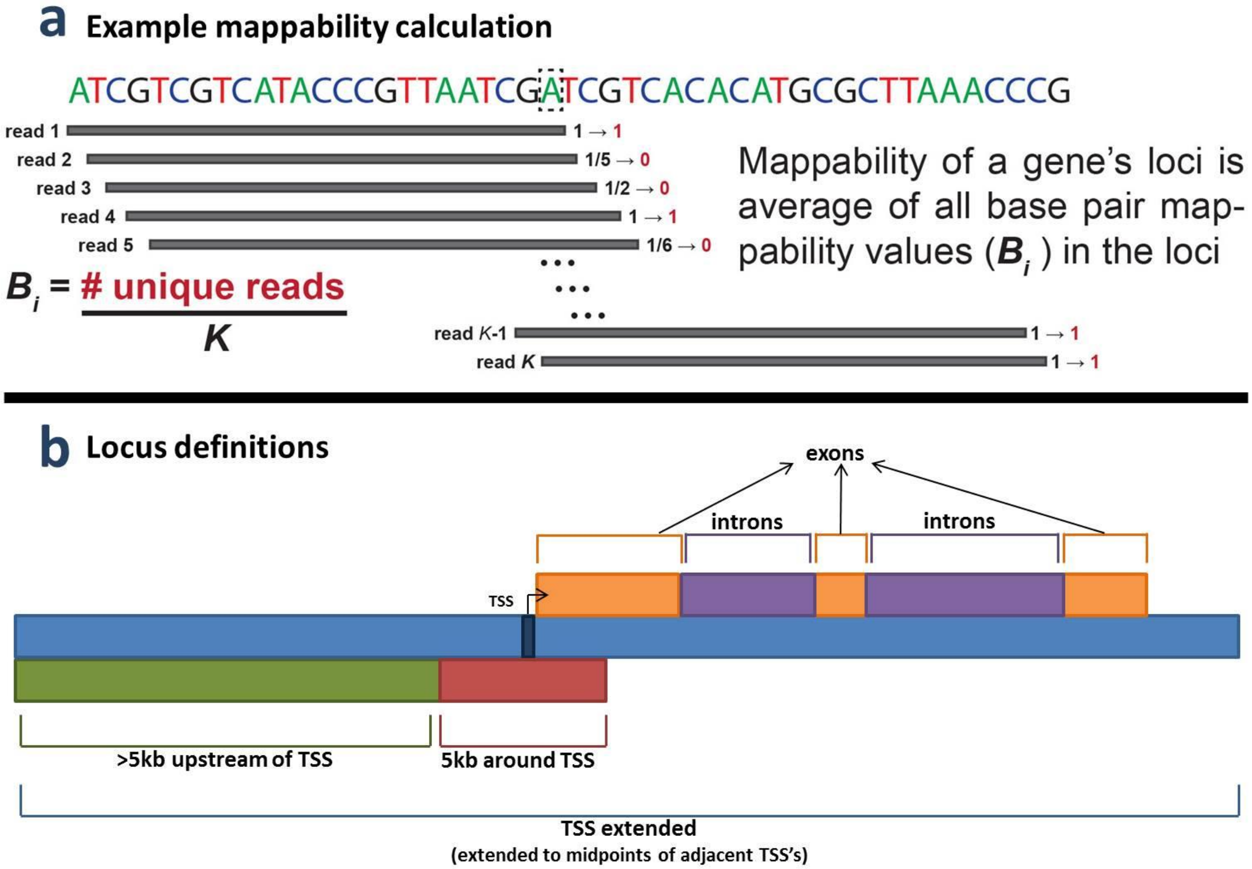

a
Example mappability calculation
ATCGTCGTCATACCCGTTAATCGATCGTCACACATGCGCTTAAACCCG
read 1
1 → 1
read 2
1/5 → 0
read 3
1/2 → 0
read 4
1 → 1
read 5
1/6 → 0
read K-1
1 → 1
read K
1 → 1
Mappability of a gene's loci is average of all base pair mappability values (B_i) in the loci
B_i = # unique reads / K
b
Locus definitions
exons
introns
introns
TSS
>5kb upstream of TSS
5kb around TSS
TSS extended
(extended to midpoints of adjacent TSS's)

Figure 4.1. Illustration of locus definitions and example of mappability calculation.
(a) An example our mappabiilty calculation using a *K*mer read. Original values from UCSC CRG mappability tracks are 1/number of locations to which the read aligns. We convert UCSC mappability to 1 if the read sequence is unique, otherwise it is assigned a value of 0. Each base pair in the genome is given a base pair mappabiilty (B_i), which is the average number of uniquely mappable reads that span over base pair *i.* Mappability of a gene's loci is the average of all B_i in the loci. (b) We created five different locus definitions: *TSS extended* - the region between the midpoints of a gene's TSS and the upstream and downstream TSSs of adjacent genes, *exons* – the exons of a transcript of a gene, *introns* – the regions between two exons of a transcript of a gene, *≤5kb from TSS* - the 5kb region surrounding a TSS, and *>5kb upstream from TSS*- the region >5kb upstream of a TSS.

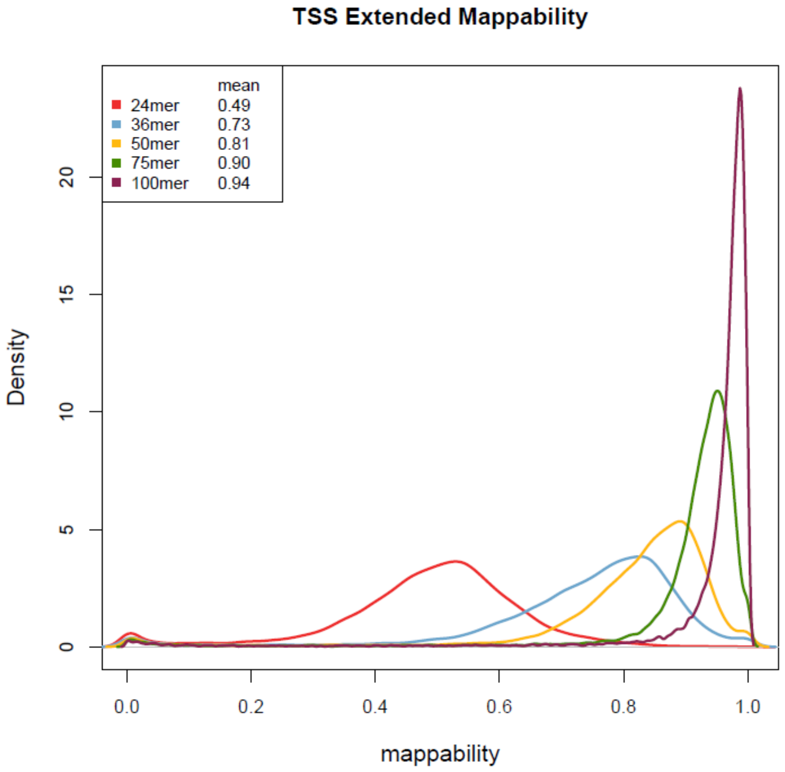


Figure 4.2. Density curves of mappability of *TSS extended* loci show increased mappability as read length increases.

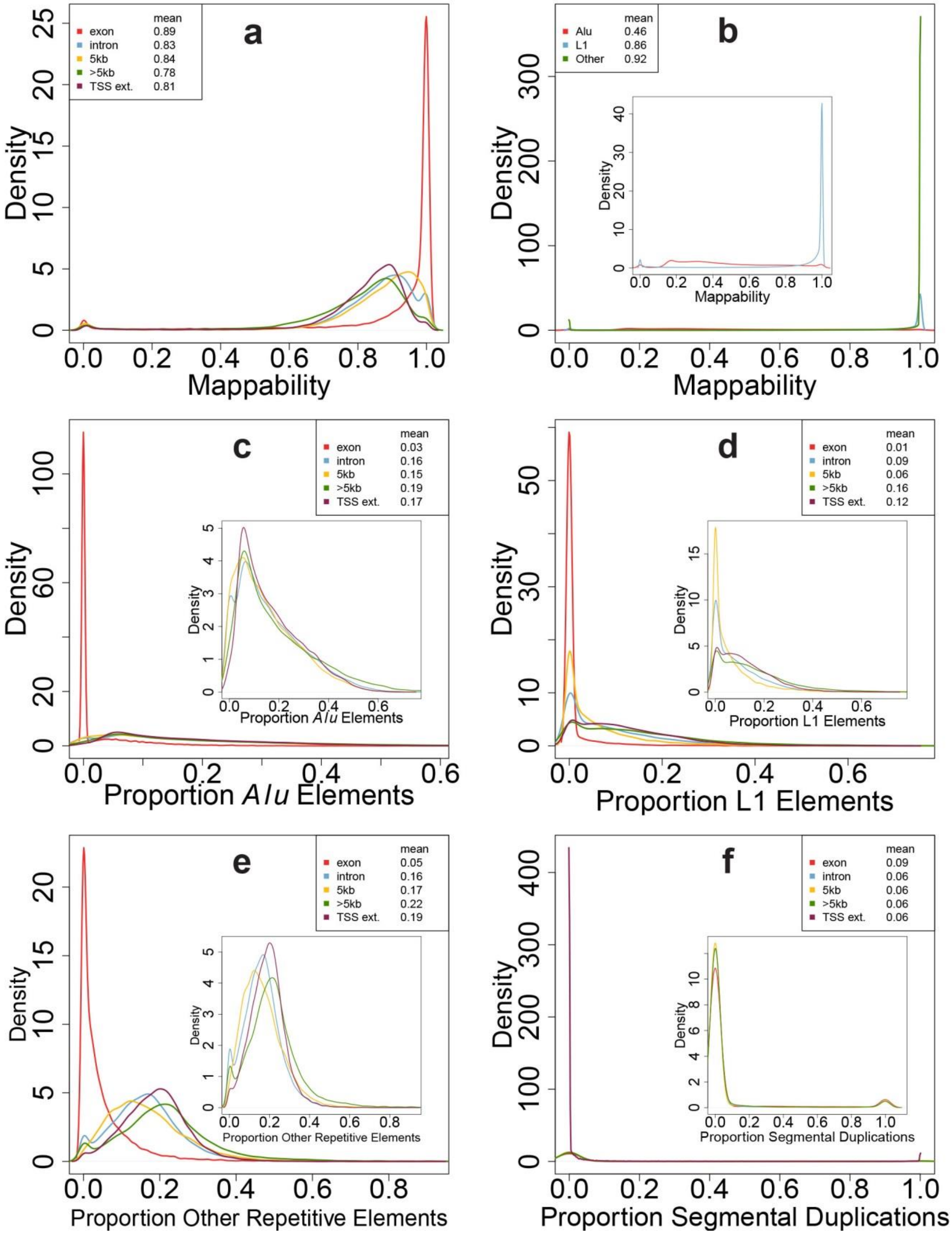


Figure 4.3. Distribution of mappability and repetitive elements across different loci. (a) Distribution of average mappability (50mer) across genes for different locus regions and (b) different repetitive elements: *Alu*, L1, and other repetitive elements. (c-f) Density curves across genes showing proportion of loci consisting of (c) *Alu* elements, (d) L1 elements, (e) other repetitive elements, and (f) segmental duplications. Panels b-f have inset plots excluding the most dominating density curve.

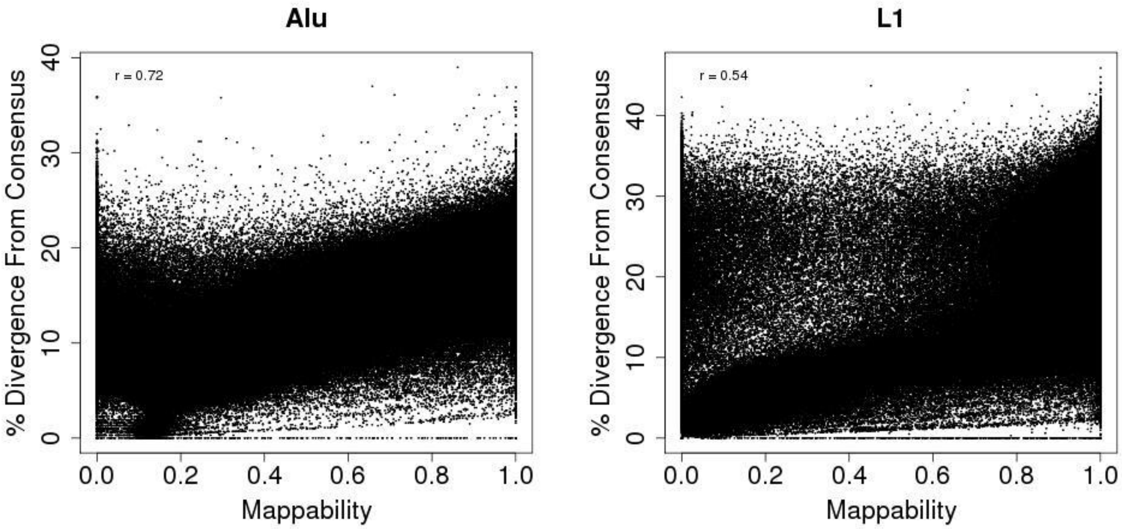
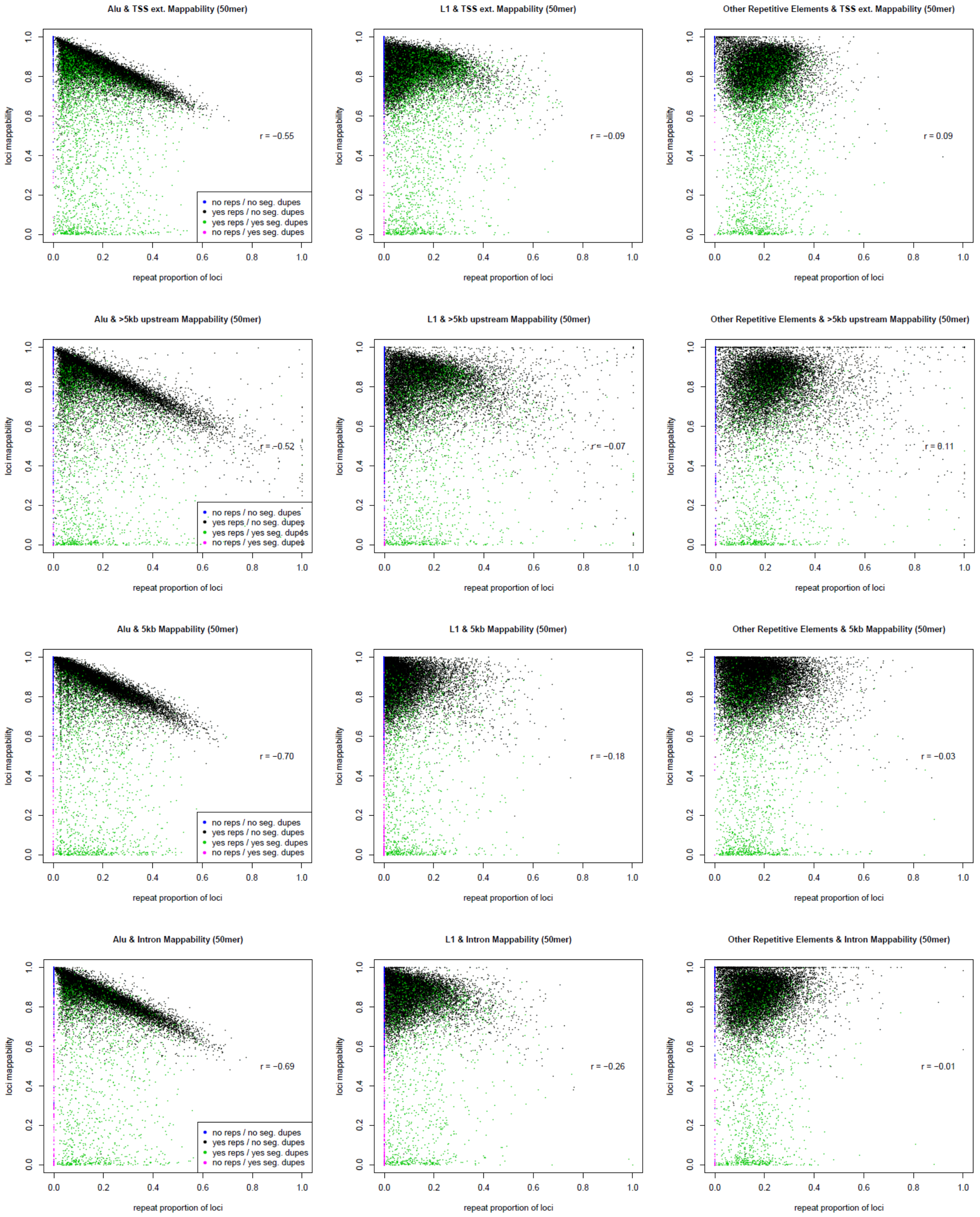


Figure 4.4. Mappability of a *Alu* and L1 elements are positively correlated with age as measured by percent divergence from consensus sequence. Spearman's correlation is 0.72 for *Alu* elements and mappability, and 0.54 for L1 elements and mappability.


Alu & TSS ext. Mappability (50mer)
r = −0.55
L1 & TSS ext. Mappability (50mer)
r = −0.09
Other Repetitive Elements & TSS ext. Mappability (50mer)
r = 0.09
Alu & >5kb upstream Mappability (50mer)
r = −0.52
L1 & >5kb upstream Mappability (50mer)
r = −0.07
Other Repetitive Elements & >5kb upstream Mappability (50mer)
r = 0.11
Alu & 5kb Mappability (50mer)
r = −0.70
L1 & 5kb Mappability (50mer)
r = −0.18
Other Repetitive Elements & 5kb Mappability (50mer)
r = −0.03
Alu & Intron Mappability (50mer)
r = −0.69
L1 & Intron Mappability (50mer)
r = −0.26
Other Repetitive Elements & Intron Mappability (50mer)
r = −0.01
loci mappability
repeat proportion of loci
0.0
0.2
0.4
0.6
0.8
1.0
no reps / no seg. dupes
yes reps / no seg. dupes
yes reps / yes seg. dupes
no reps / yes seg. dupes

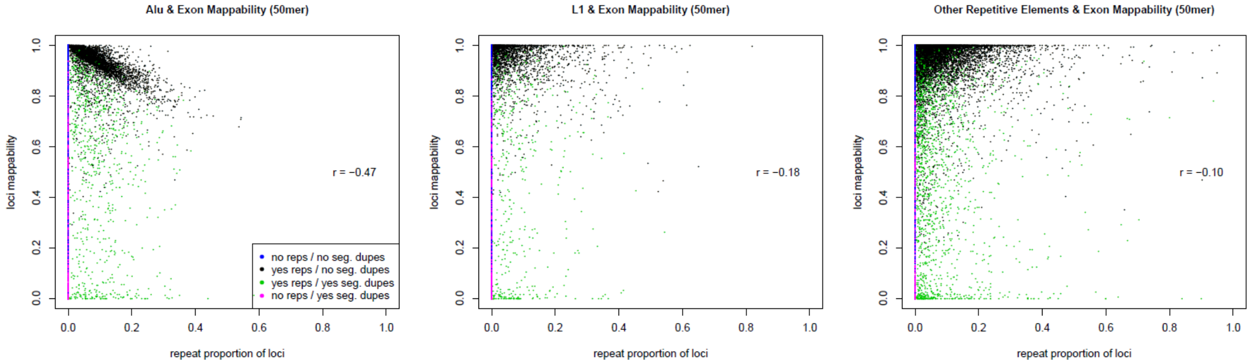


Figure 4.5. Scatterplots of mappability using 50mer reads and proportion of loci that are *Alu* elements, L1 elements, and other repetitive elements show generally that gene loci that contain larger proportion of a repetitive element have lower mappability and vice versa. *Alu* elements affect mappability more so than other repetitive elements in the all the different loci. Each point (i.e. gene) is colored blue if that gene locus has no respective repetitive element and no segmental duplications. Black points are gene loci with the respective repetitive element but do not have any segmental duplications. Blue points are gene loci with no respective repetitive element and no segmental duplications. Green points are gene loci that have both the respective repetitive element and at least one segmental duplication. Pink points are gene locus that have no respective repetitive element but do have at least one segmental duplication. Having at least one segmental duplication also decreases mappability. Gene loci that have at least one segmental duplication can have much lower mappability that those that do not, even when there are no repetitive elements, or are repetitive elements such as L1 elements and other repetitive elements that tend to be more highly mappable than *Alu* elements.

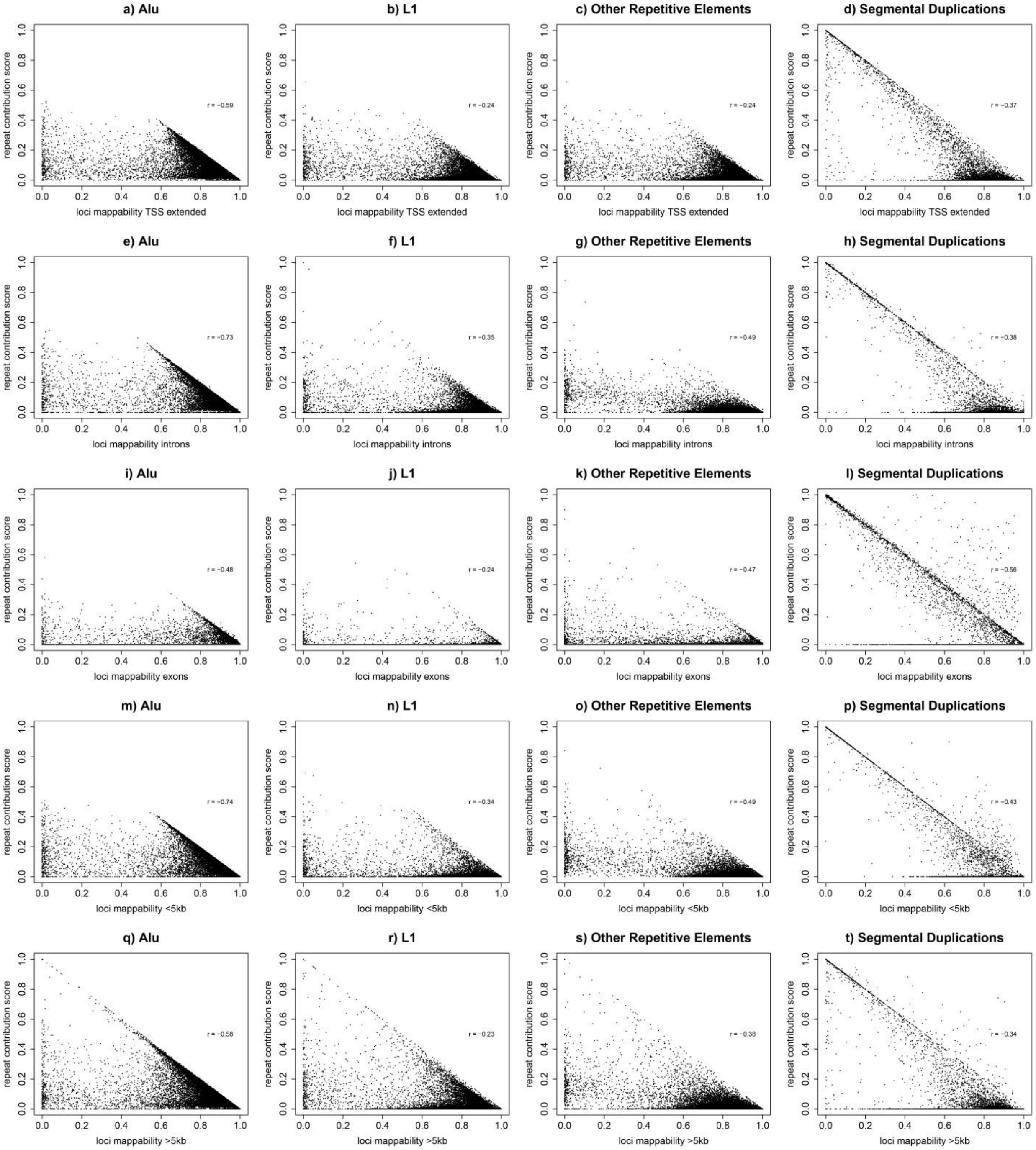


Figure 4.6. Contribution scores of repetitive elements and segmental duplications are correlated with mappability.

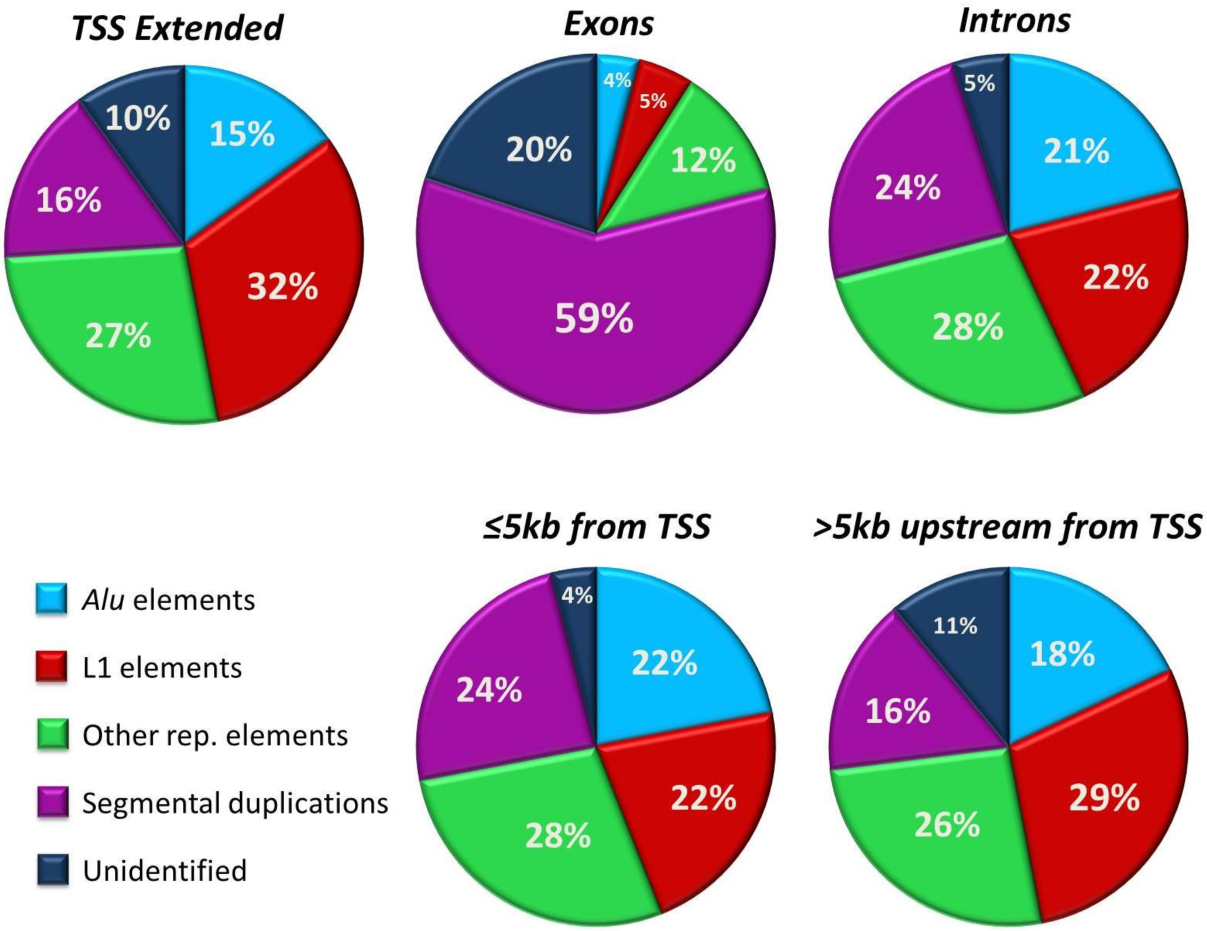


Figure 4.7. Individual contribution of repetitive elements and segmental duplications to mappability. In order from explaining mappabiilty most to least, the regions were *≤5kb of the TSS* (95.5%), *introns* (94.5%), *TSS extended* (90.3%), *>5kb upstream from TSS* (89.3%), and *exons* (79.9%).

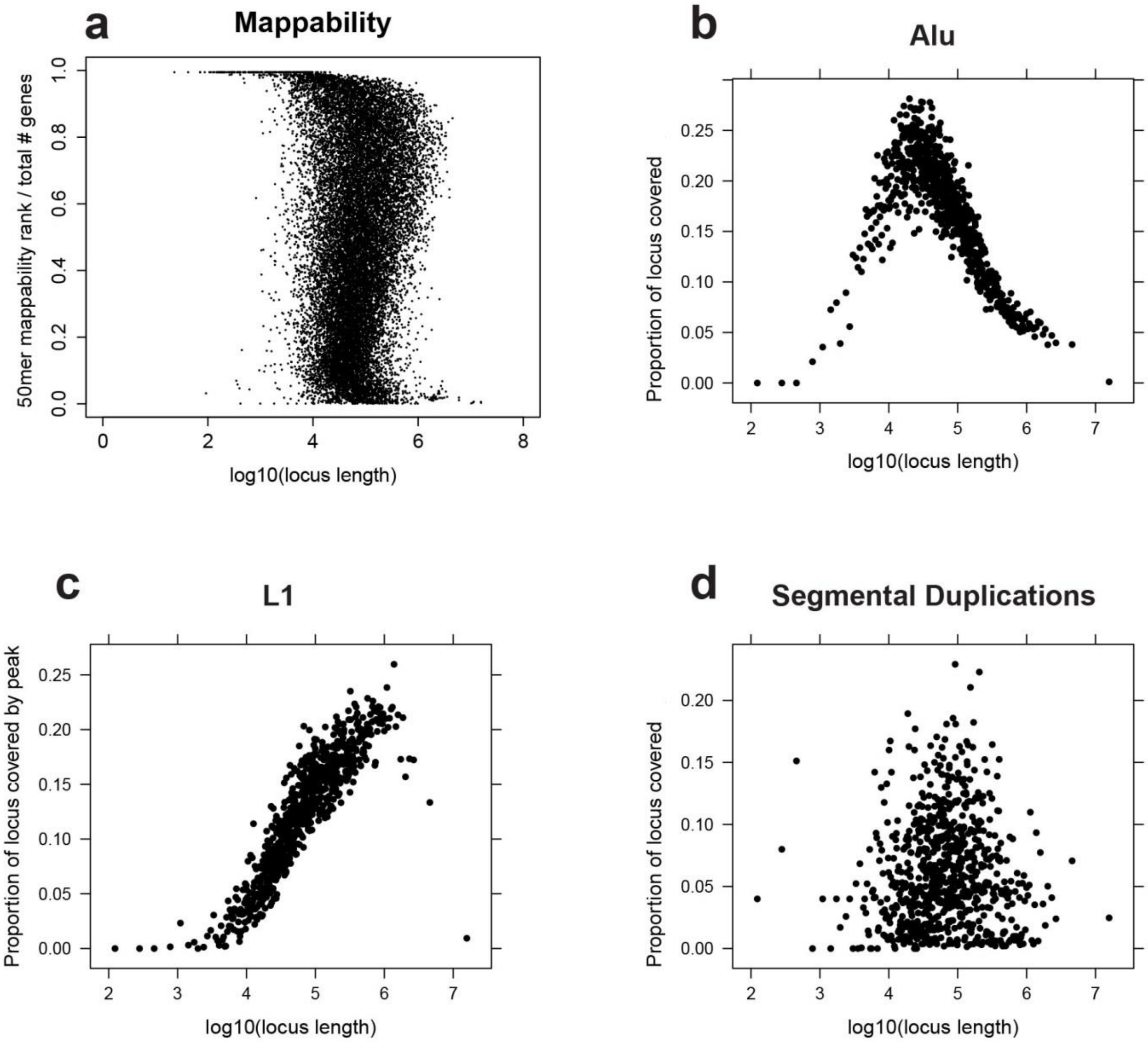


Figure 4.8. Relationship of (a) mappability, (b) *Alu* elements, (c) L1 elements, and (d) segmental duplications with gene length. (b-d) are output plots from Broad-Enrich, genes are grouped into bins of 25 genes ordered by locus length. Locus length here is calculated from *TSS extended* regions.

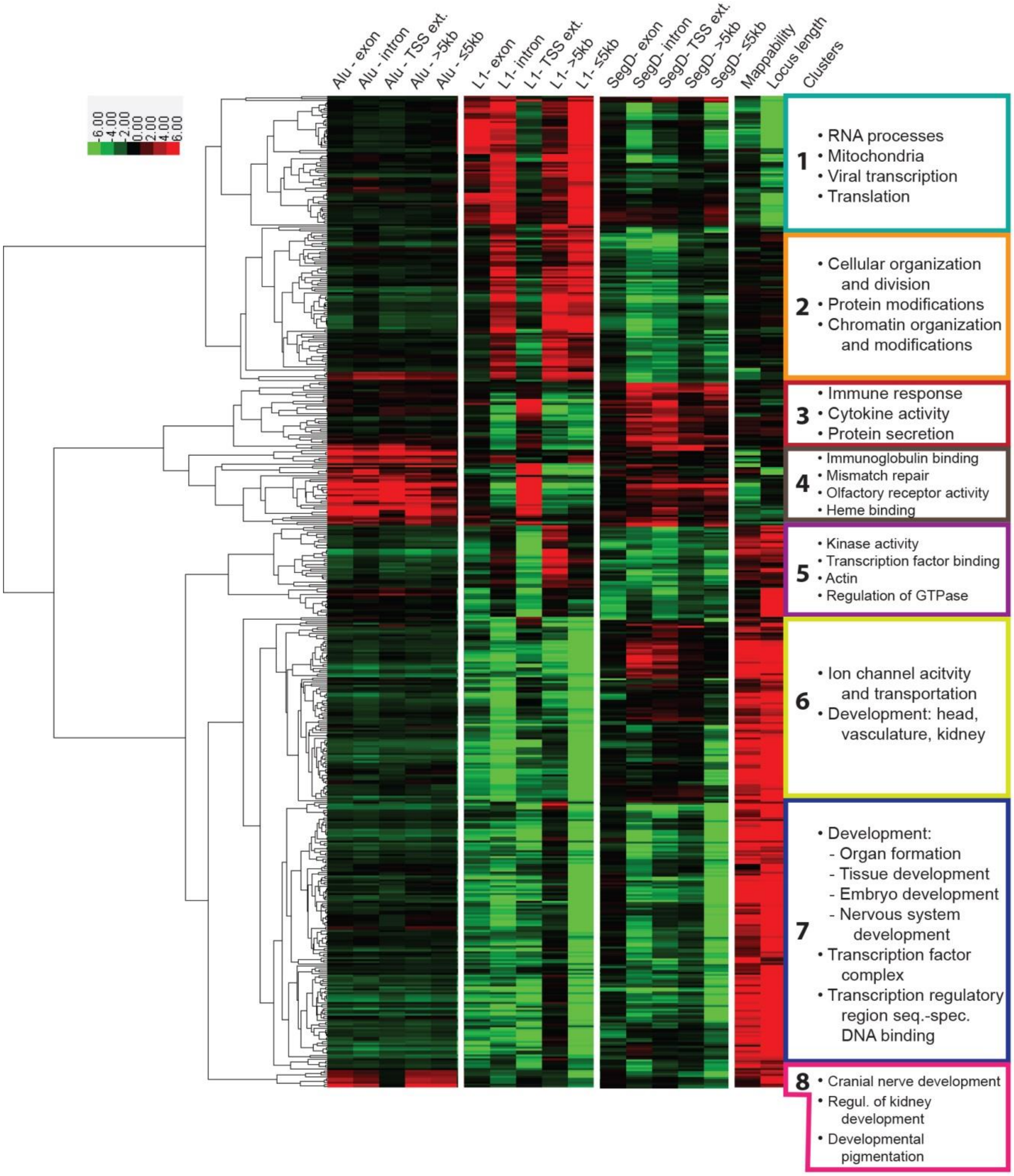


Figure 4.9. Heatmap showing signed –log10(p-values) (negative for depleted terms, positive for enriched terms) of GO gene set enrichment results. The first five columns are enrichment results for the different loci using proportion of *Alu* elements, the second set of five columns are enrichment results for the different loci using proportion of L1 elements, the third set of five columns are enrichment results for the different loci using proportion of segmental duplications, and the last two columns are enrichment results for highly/lowly mappable GO terms and short/long locus lengths using the *TSS extended* loci. We identified 8 distinct clusters of GO terms that show similar enrichment patterns. GO terms included were limited to those that had less than 500 genes and FDR≤0.001 in at least two locus definitions in at least one type of repetitive element or segmental duplications, which resulted in 510 gene sets.

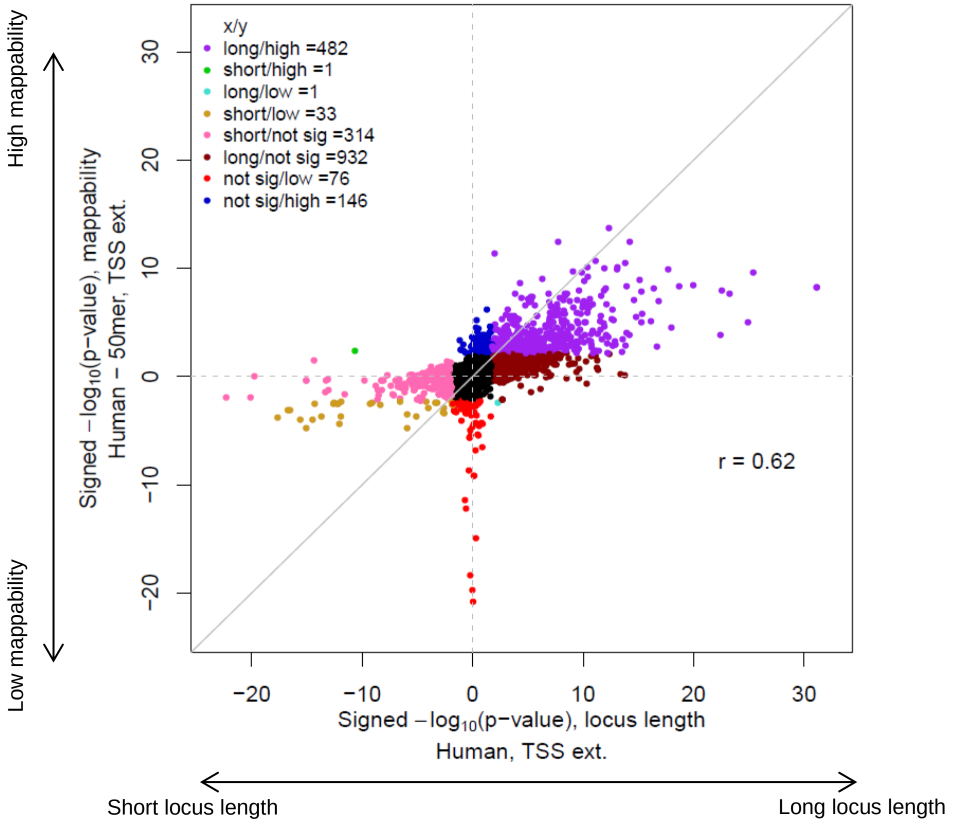


Figure 4.10. Mappability and locus length in human have similarly enriched gene sets. Each point represents a GO term. The axes are signed –$\log_{10}$(p-value), i.e. +$\log_{10}$(p-value) if long locus length/high mappability, or -$\log_{10}$(p-value) if short locus length/low mappability using *TSS extended* locus regions and 50mer mappability. Purple indicates terms that have long locus length/high mappability. Green indicates terms with short locus length/low mappability. Turquoise indicates terms with long locus length/high mappability. Gold indicates terms with short locus length/low mappability. Pink indicates terms that have only significantly short locus length. Dark red indicates terms that have only significantly long locus length. Bright red indicates terms that have only significantly low mappability. Dark blue indicates terms that have only significantly high mappability. The remaining terms are indicated in black. Reported GO terms are limited to those with ≤500 genes

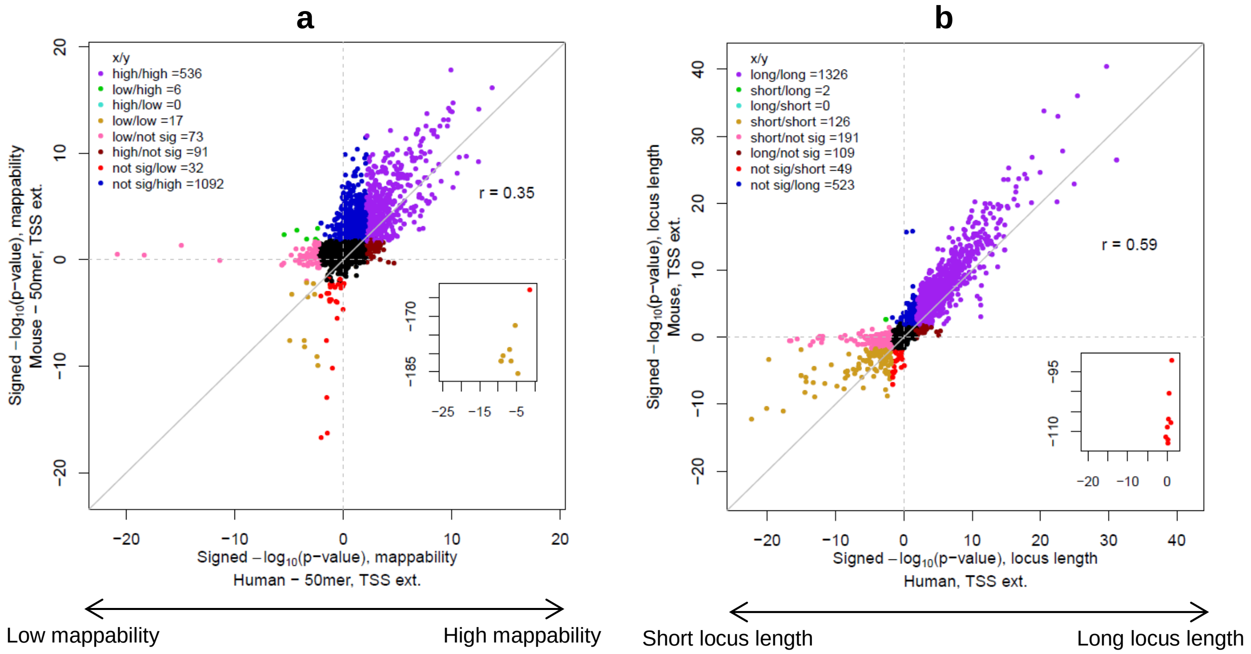


Figure 4.11. Human and mouse have similarly enriched gene sets for (a) high and low mappability and (b) long and short locus lengths. Each point represents a GO term. The axes are signed –log$_{10}$(p-value), i.e. +log$_{10}$(p-value) if highly mappable/long locus length, or -log$_{10}$(p-value) if lowly mappable/short locus length in mouse or human using 50mer mappability from *TSS extended* locus regions. Purple indicates terms that have high mappability/long locus length in both human and mouse. Green indicates terms with low mappability/short locus length in human and high mappability/long locus length in mouse. Turquoise indicates terms with high mappability/long locus length in human and low mappability/short locus lengths in mouse. Gold indicates terms with low mappability/short locus length in both human and mouse. Pink indicates terms that are only significantly lowly mappable/short locus length in human. Dark red indicates terms that are only significantly highly mappable/long locus length in human. Bright red indicates terms that are only significantly lowly mappable/short locus length in mouse. Dark blue indicates terms that are only significantly highly mappable/long locus length in mouse. The remaining terms are indicated in black. Reported GO terms are limited to those with ≤500 genes.

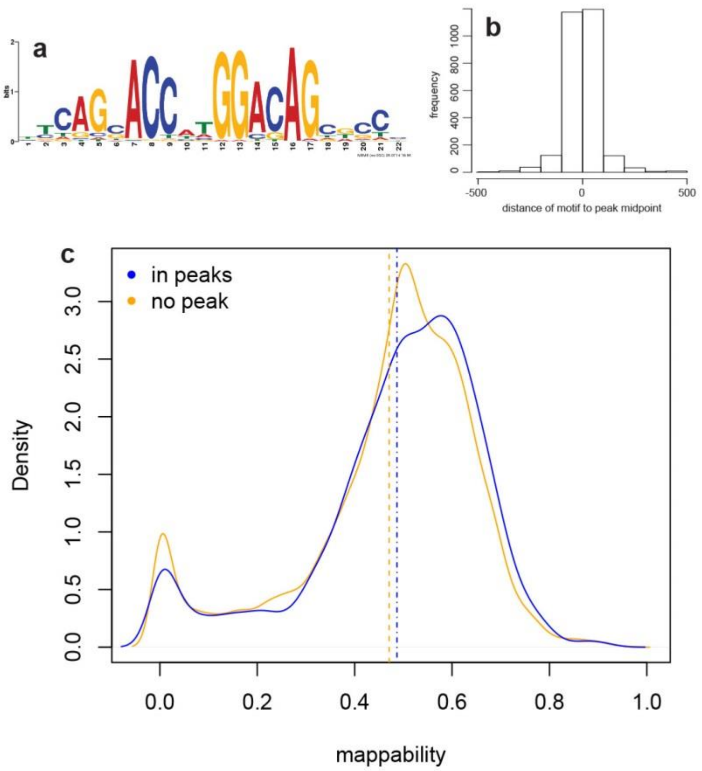


Figure 4.12. NRSF ChIP-seq results. (a) Motif logo identified by MEME using peaks called by PePr; (b) histogram showing that motifs had a strong tendency to occur in the middle of a peak; (c) histograms showing that the predicted NRSF binding sites in peaks (blue) tend to have higher mappability than the computationally predicted NRSF binding sites not overlapping a peak (orange) (p =8.65x10^{-6}).

Table 4.1. Select significantly enriched/depleted GO terms. Top 5 non-related terms for each repeat type and locus regions, limited to GO terms with ≤500 genes. The full list is in Supplementary Table 4.1.

a) Top GO terms significantly *enriched* with *Alu* elements

GO term	Branch	Locus Regions (q-value ≤0.05)	Average Mappability	Locus Length %tile
dichotomous subdivision of an epithelial terminal unit	BP	intron, exon, ≤5kb, >5kb up.	0.62	88
enteric nervous system development	BP	intron, exon, ≤5kb, >5kb up.	0.66	61
glucuronosyltransferase activity	MF	All	0.76	43
heme binding	MF	TSS ext., intron, exon, >5kb up.	0.77	16
IgG binding	MF	All	0.25	1.7
immunoglobulin binding	MF	All	0.51	2.3
keratin filament	CC	intron, exon, >5kb up.	0.81	0.019
MHC class I protein complex	CC	All	0.7	0.42
monooxygenase activity	MF	TSS ext., intron, exon, >5kb up.	0.76	22
olfactory receptor activity	MF	intron, exon, ≤5kb, >5kb up.	0.73	26
oxidoreductase activity, acting on paired donors,…	MF	TSS ext., intron, exon, >5kb up.	0.69	3.9
positive regulation of kidney development	BP	intron, exon, ≤5kb, >5kb up.	0.64	92
response to xenobiotic stimulus	BP	All	0.76	16

b) Top GO terms significantly *depleted* with *Alu* elements

GO term	Branch	Locus Regions (q-value ≤0.05)	Average Mappability	Locus Length %tile
brain development	BP	intron, exon, >5kb up.	0.87	88
cell division	BP	exon	0.83	41
inorganic cation transmembrane transporter activity	MF	intron, exon, >5kb up.	0.85	64
metal ion transport	BP	intron	0.86	59
protein binding transcription factor activity	MF	TSS ext., exon, >5kb up.	0.85	63
regulation of cell development	BP	intron, exon, >5kb up.	0.87	88
regulation of nervous system development	BP	intron, exon, >5kb up.	0.87	90
sensory organ development	BP	intron	0.86	78
synapse	CC	TSS ext., intron, exon, >5kb up.	0.86	91

c) Top GO terms significantly *enriched* with L1 elements

GO term	Branch	Locus Regions (q-value ≤0.05)	Average Mappability	Locus Length %tile
chromatin modification	BP	intron, ≤5kb, >5kb up.	0.83	45
chromosomal part	CC	intron, ≤5kb, >5kb up.	0.82	32
keratin filament	CC	TSS ext.	0.81	0.019
keratinization	BP	TSS ext.	0.84	1.9
mitochondrial membrane	CC	intron, exon, ≤5kb, >5kb up.	0.81	15
mitochondrial membrane part	CC	intron, exon, 5kb	0.79	11
monooxygenase activity	MF	TSS ext.	0.76	22
mRNA processing	BP	intron, ≤5kb, >5kb up.	0.81	22

olfactory receptor activity	MF	TSS ext.	0.73	26
protein folding	BP	intron, 5kb	0.81	31
ribosome	CC	intron, exon, 5kb	0.8	3.3
RNA catabolic process	BP	intron, exon, ≤5kb, >5kb up.	0.8	5.2
RNA splicing	BP	intron, ≤5kb, >5kb up.	0.81	24
SRP-dependent cotranslational protein targeting to membrane	BP	intron, exon, 5kb	0.79	2.5
structural constituent of ribosome	MF	intron, exon, 5kb	0.8	3.1
transcription cofactor activity	MF	>5kb up.	0.85	63
transcription factor binding	MF	>5kb up.	0.83	65
translational termination	BP	intron, exon, 5kb	0.77	1.7
unfolded protein binding	MF	intron, 5kb	0.81	7.6
viral genome expression	BP	intron, exon, ≤5kb, >5kb up.	0.8	3
viral reproduction	BP	intron, exon, ≤5kb, >5kb up.	0.81	9.1
xenobiotic metabolic process	BP	TSS ext.	0.76	16

d) Top GO terms significantly depleted with L1 elements

GO term	Branch	Locus Regions (q-value ≤0.05)	Average Mappability	Locus Length %tile
actin filament-based process	BP	exon, TSS ext.	0.84	64
axon	CC	intron, exon, ≤5kb, TSS ext.	0.86	95
axonogenesis	BP	intron, exon, ≤5kb, TSS ext.	0.86	93
behavior	BP	intron, exon, ≤5kb, >5kb up.	0.86	82
brain development	BP	intron, exon, ≤5kb, TSS ext.	0.87	88
dendrite	CC	intron, exon, ≤5kb, TSS ext.	0.86	94
embryonic morphogenesis	BP	intron, exon, ≤5kb, TSS ext.	0.86	78
extracellular matrix	CC	intron, exon, ≤5kb, >5kb up.	0.85	69
intermediate filament	CC	intron, exon, ≤5kb, >5kb up.	0.84	1.8
monooxygenase activity	MF	intron, ≤5kb, >5kb up.	0.76	22
olfactory receptor activity	MF	intron, ≤5kb, >5kb up.	0.73	26
pattern specification process	BP	intron, exon, ≤5kb, TSS ext.	0.86	75
phospholipid binding	MF	exon, TSS ext.	0.83	69
regulation of cell development	BP	intron, exon, ≤5kb, TSS ext.	0.87	88
regulation of nervous system development	BP	intron, exon, ≤5kb, TSS ext.	0.87	90
regulation of system process	BP	All	0.87	76
sensory perception	BP	intron, ≤5kb, >5kb up.	0.85	57
skeletal system development	BP	All	0.86	77
synapse	CC	intron, exon, ≤5kb, TSS ext.	0.86	91
tissue morphogenesis	BP	intron, exon, ≤5kb, TSS	0.86	81

		ext.		
tube development	BP	intron, exon, ≤5kb, TSS ext.	0.85	87
vasculature development	BP	intron, exon, ≤5kb, TSS ext.	0.86	79
xenobiotic metabolic process	BP	intron, ≤5kb, >5kb up.	0.76	16

e) Top GO terms significantly *enriched* with segmental duplications

GO term	**Branch**	**Locus Regions (q-value ≤0.05)**	**Average Mappability**	**Locus Length %tile**
branched-chain amino acid catabolic process	BP	exon	0.81	28
channel activity	MF	intron, TSS ext.	0.85	73
CoA hydrolase activity	MF	exon, 5kb	0.75	3.9
cytokine activity	MF	intron, TSS ext.	0.84	21
cytokine receptor activity	MF	intron, ≤5kb, TSS ext.	0.82	28
defense response to bacterium	BP	intron, ≤5kb, TSS ext.	0.76	14
high-density lipoprotein particle	CC	exon	0.81	0.19
hormone metabolic process	BP	>5kb up., TSS ext.	0.82	41
immune effector process	BP	intron, TSS ext.	0.82	28
inflammatory response	BP	intron, ≤5kb, TSS ext.	0.82	34
innate immune response	BP	intron, ≤5kb, TSS ext.	0.81	21
ion channel activity	MF	intron, TSS ext.	0.86	76
lipid catabolic process	BP	intron, ≤5kb, >5kb up., TSS ext.	0.83	28
mitochondrial matrix	CC	5kb, TSS ext.	0.81	13
oxidoreductase activity, acting on CH-OH group of donors	MF	5kb, >5kb up.	0.81	22
regulation of defense response to virus by host	BP	exon	0.77	15
response to bacterium	BP	intron, TSS ext.	0.81	35
response to xenobiotic stimulus	BP	intron, ≤5kb, >5kb up., TSS ext.	0.76	16
small molecule catabolic process	BP	5kb, TSS ext.	0.82	30
steroid metabolic process	BP	intron, ≤5kb, >5kb up., TSS ext.	0.82	27
transferase activity, transferring hexosyl groups	MF	intron, ≤5kb, >5kb up., TSS ext.	0.82	75
transferase activity, transferring one-carbon groups	MF	exon	0.8	16
xenobiotic metabolic process	BP	intron, ≤5kb, >5kb up., TSS ext.	0.76	16

f) Top GO terms significantly *depleted* with segmental duplications

GO term	**Branch**	**Locus Regions (q-value ≤0.05)**	**Average Mappability**	**Locus Length %tile**
actin cytoskeleton	CC	intron, exon, TSS ext.	0.84	41
actin filament-based process	BP	exon, TSS ext.	0.84	64
axon guidance	BP	All	0.85	92
axonogenesis	BP	intron, exon, ≤5kb, TSS ext.	0.86	93
cell-cell adhesion	BP	exon, 5kb	0.85	96
chordate embryonic development	BP	intron, ≤5kb, >5kb up., TSS ext.	0.86	72
chromatin binding	MF	intron, ≤5kb, TSS ext.	0.84	67

chromatin organization	BP	intron, ≤5kb, TSS ext.	0.83	37
embryonic morphogenesis	BP	intron, ≤5kb, TSS ext.	0.86	78
homophilic cell adhesion	BP	exon, 5kb	0.86	99
Microtubule	CC	intron, exon, TSS ext.	0.82	44
mRNA processing	BP	intron, TSS ext.	0.81	22
muscle structure development	BP	intron, ≤5kb, TSS ext.	0.87	79
negative regulation of transcription from RNA polymerase II promoter	BP	intron, ≤5kb, >5kb up., TSS ext.	0.85	75
pattern specification process	BP	intron, ≤5kb, >5kb up., TSS ext.	0.86	75
posttranscriptional regulation of gene expression	BP	intron, 5kb	0.83	31
regionalization	BP	intron, ≤5kb, TSS ext.	0.88	72
regulatory region DNA binding	MF	intron, ≤5kb, TSS ext.	0.86	73
tissue morphogenesis	BP	intron, ≤5kb, TSS ext.	0.86	81
tubulin binding	MF	intron, >5kb up., TSS ext.	0.82	44

Supplementary Table 4.1. Enriched and depleted GO terms for *Alu* and L1 elements, and segmental duplications for all locus regions (extended version of Table 4.1).

Supplementary Table 4.2. Full list of GO terms associated with clusters from Figure 4.9.

Supplementary Table 4.3. GO term enrichment results for mappability and locus length, comparing mouse and human.

Chapter 5 Conclusions & Future Directions

5.1 Conclusions

In this dissertation, I contributed to the tools available for functional interpretation of high-throughput sequencing data such as ChIP-seq and RNA-seq data, as well as characterizing the main contributing factors to read mappability at the pathway level. Overall, the research I present has and will further our understanding of how gene length, read count, and read mappability can affect statistical tests for HTS data..

In Chapter 2, we introduced ChIP-Enrich, a gene set enrichment test for ChIP-seq data that adjusts for gene locus length in peak-to-gene assignments. ChIP-Enrich consists of two main parts: (1) locus definitions that allow peak-to-gene assignments for studying various genic and regulatory regions, and (2) a gene set enrichment test that empirically adjusts for the observed relationship between locus length and probability of having at least one peak in the gene loci. We showed through permutation testing that unlike other existing GSE tests, Fisher's exact test and the binomial test, ChIP-Enrich maintains an acceptable type I error rate even when there exists a relationship between locus length and probability of having at least one peak. We applied ChIP-Enrich to 63 ENCODE datasets that included transcription factors and histone modifications, and were generated using different peak callers. The datasets varied widely by binding patterns as well as number of peaks. We showed that ChIP-Enrich was able to account for all types of relationships between locus length and peak presence. FET was only appropriate for data sets where there was no relationship between locus length and peak presence. The other test we compared ChIP-Enrich to was the binomial test, which assumes number of peaks is proportional to locus length. However, for datasets with high number of peaks, which were the same datasets that often had peaks proportional to locus length, the binomial test underestimates variance because of over-dispersion of peaks among genes. This resulted in incorrect p-values and inflated type I error. In addition, we showed that limiting peaks to a restrictive locus definition, *≤1kb from TS*S, compared to an all-inclusive locus definition, *nearest TSS*, resulted in different discovered biology. We further examined the effect of the locus definition by applying ChIP-Enrich to a ChIP-seq dataset of glucocorticoid receptor activity in A549 cells treated with Dex, and showed that using two locus definitions, *≤1kb from TSS* and *nearest TSS*, elucidated different regulatory activity of GR in proximal and distal regulatory regions.

In Chapter 3, we introduced RNA-Enrich, a gene set enrichment test for RNA-seq data that corrects for any selection bias due to varying read counts. To correct for this relationship, weights for genes were created using a spline fitted to average gene read count and significance values like differential gene expression p-values. Therefore if genes that were more likely to have higher read counts, like highly expressed, long genes, also had more significant p-values, then those genes were weighed less and were less influential in enrichment of a gene set. Unlike other GSE tests for RNA-seq data such as DAVID (which uses a modified Fisher's exact test) and GOseq (which can also adjust for read count), RNA-Enrich does not require a cut-off to define differentially expressed genes. And unlike permutation-based tests like GSEA, GSAASeqSP, and SeqGSEA, RNA-Erinch does not require large sample sizes for power and accuracy, or a long run-time. RNA-Enrich also had substantially improved type I error compared to DAVID and GOseq, and maintained similar performance when we used p-values from DEseq instead of *edgeR* or a correct log fold change.

With ChIP-Enrich and RNA-Enrich, we showed that gene length and read count, respectively, can affect functional interpretation of ChIP-seq and RNA-seq data. However our GSE tests are not limited to these two types of HTS data. Here are some examples: ChIP-Enrich may also be applied to GWAS data, especially if genes with longer locus length are more likely to have significant SNP variants. In Chapter 4, we used RNA-Enrich to perform GSE testing of mappability, in which case we used mappability values as the significance values and locus length instead of read count. Bisulfite sequencing data may also be a good candidate for RNA-Enrich. Some bisulfite sequencing platforms arebias toward CpG islands, thus genes with more CpGs may be more likely to be identified as differentially methylated. Genes with more intergenic distance may also be more likely to contain CpGs.

We first introduced the concept of mappability in Chapter 2 as an option for ChIP-Enrich. However we did not further expand on how much effect mappability has on gene set enrichment testing after accounting for gene length. We showed in Chapter 4 that certain gene functions do tend to have higher or lower mappability, and therefore mappability should be considered in GSE tests. We also showed that mappability of potential DNA protein binding sites can affect which peaks would be detected in ChIP-seq. Thus we explored how certain gene functions had more or less mappable genes, and how that concurred with gene locus length. Our analysis of mappability and gene length showed that gene sets that tend to have shorter genes were also lowly mappable, while gene sets that tend to have longer genes were also highly mappable. We also found similar highly/lowly mappable and longer/shorter locus length gene sets in mouse, suggesting some conservation of gene functions across different species.

We suggested that in addition to co-varying with gene length, mappability can also be indicative of the complexity of gene regulation. Genes that need more complex regulation would need more unique surrounding sequence and intergenic distance. We expanded our analyses of mappability and gene length to include repetitive genomic features like transposons and segmental duplications. We showed that mappability is strongly affected by transposons and segmental duplications, which are both lowly mappable. We further analyzed the most prominent LINE elements, L1 elements, and the most prominent SINE elements, *Alu* elements, in the human genome, as well as segmental duplications. We discovered distinct enrichment and depletion signatures of *Alu* and L1 elements, segmental duplications, mappability and gene length exhibited by certain gene functions. For example, genes involved with development were strongly depleted of repeat elements, were highly mappable, and had longer genes - all of which suggest evolutionary pressure to maintain unique sequence and long intergenic distances for complex regulation of developmental genes. Gene sets enriched with *Alu* elements, L1 elements, and/or segmental duplications suggest evolutionary selection for the repeat element, and perhaps adaptation of the element for the gene’s own benefit. Overall, we showed that mappability is indicative of genomic architect and complexity of gene regulation. We found certain gene functions are more highly or lowly mappable, and therefore can bias GSE of ChIP-seq data. Fortunately our GSE test, ChIP-Enrich has the option to correct for mappability and therefore is robust to this effect.

Throughout this dissertation, we demonstrated that considering characteristics of the human genome is essential to improving functional interpretation of HTS data. The GSE tests we have developed have enabled us to perform functional interpretation of HTS data, repetitive elements and segmental duplications, and mappability for the first time taking into account locus length.

5.2 Future Directions

5.2.1 Chapter 2

In developing ChIP-Enrich, we generated various locus definitions for assigning peaks to genes, each resulting in different locus lengths. The regulatory regions we assigned to each gene were based on a linear and continuous organization of the genome. However, our understanding of the genome has evolved with studies of histone marks and chromatin conformation. Regulatory regions are not necessarily adjacent to gene TSS's. Studies of topologically associated domains, or TADs, show that the human genome is organized in various loops and compartments [72]. Furthermore, current ChIP-Enrich locus definitions do not allow overlap of loci among genes, which makes the assumption of a one-to-one locus to gene regulatory system. However, it is known that enhancers can regulate multiple genes, and some genes have multiple enhancers [134]. We are currently developing more biologically realistic locus definitions that make use of enhancer databases to better define promoter and enhancer regions of genes.

In our analysis of 63 ENCODE datasets, we observed that histone modification ChIP-seq experiments tend to call more peaks and broader peaks. ChIP-Enrich reduces peak information to a binary indicator, i.e. either a gene has no peaks or at least one peak, and peaks are only defined by their midpoints. This is not the best approach for histone modification data as peaks can occur in almost all genes and may even span multiple genes. As a follow-up to ChIP-Enrich, we developed Broad-Enrich [37], which uses the locus proportion covered by a peak instead of a binary peak indicator. In doing so, Broad-Enrich has improved power for histone data compared to ChIP-Enrich. Broad-Enrich is not limited to histone data as we have showed in Chapter 4. It can also be used with experiments that result in many genomic regions, especially occurring frequently throughout the genome (like transposons), and/or broad genomic regions (like segmental duplications).

Our analyses of ChIP-seq data has thus far been limited to comparing gene sets to one another. However, across different cell lines and different DNA-binding proteins, genes in the same gene sets may be regulated similarly. Some genes may be regulated by the proximal promoter, others more so from distal regions. These patterns of regulation may be observed on a global scale, i.e. not just limited to one cell line or one DNA-binding protein. This approach can give insight into more detailed regulatory patterns of genes and gene functions.

5.2.2 Chapter 3

We showed improved type I error in RNA-Enrich when we correct for the relationship between significance values and average read count, however we acknowledge that the results from our permutations are still not perfectly uniform as there are more gene sets with $p \leq 0.05$ than expected. There is evidence that sequences with high GC content more easily amplify compared to those that are not [8, 135, 136]. If GC content perpetuates at the gene set level, it may affect enrichment results. Other factors that may affect read alignment is mappability. Though we show in Chapter 4 that exons are most mappable, alternative splicing may result in a transcript that includes typically non-coding regions, which tend to be less mappable. Overall read coverage can also affect differential expression testing, which is not mutually exclusive of GC content bias. The dataset of A549 cells treated with Dex had higher coverage than the other two datasets we used, suggesting that the relationships we observed between differential expression significance values and average read count per gene may become less prominent with deeper sequencing. It would be of interest to apply RNA-Enrich to more datasets of varying read coverage.

5.2.3 Chapter 4

Transposons are a rich field of study. We have only performed our analyses on *Alu* and L1 elements, however there are 42 other repeat families in the UCSC genome browser “rmsk” table. For example, human endogenous retroviruses (HERVs), a type of long terminal repeat (LTR), comprise 5-8% of the human genome and have been implicated in cancer and as an inducer of genetic instability, methylation, transactivation and RNA interference. HERVs have been shown to act as promoter, enhancer, and transcriptional factor-binding site to potentially regulate neighboring genes [137]. Associations of HERVs with gene function may provide some insight into their evolution in the genome. Also of interest is examining whether the age of L1 elements is associated with gene function. Younger L1 elements tend to be located closer to genes than older elements [138] , whereas full length L1 elements are more abundant on sex chromosomes [139] and have been implicated in X-inactivation [115].

Thus far, we have only performed enrichment of repeat elements in human. We have performed gene set enrichment of mappability and gene length in mouse but will also perform the same analysis for repeat elements. Of particular interest is the relationship between gene function and transposons. In the evolutionary timeline, primate-rodent split of 7SL RNA derived elements (the origin of *Alu* elements in human, and B1 elements in mouse) diverged about 80 million years ago. The result was independent amplification, duplication and mutation accumulation in copies of *Alu* and B1 elements [140, 141]. While the sequences of L1 elements are similar between human and mouse, in mouse B1 elements are the most prominent SINEs. We hypothesize that we will find similar gene functions enriched with B1 in mouse as *Alu*s in human, as it has been shown that *Alu* and B1 elements have similar distributions in genomic features, for example, both are more prominent in upstream promoter regions of genes [130]. If our hypothesis holds true, this would be evidence for conserved evolutionary selection pressures, and identify which gene functions, despite independent evolution and accumulation of of *Alu* and B1 in human and mouse, may benefit from maintaining or deleting repeats in the various regulatory regions of the associated genes.

Also of interest in comparing mouse and human is the relationship of segmental duplications and gene function. Bailey, et al [142] and, more recently, She, et al [143] found that while the distribution of segmental duplications in humans are interspersed over large genomic distances, in mouse, segmental duplications are more locally clustered, taking the form of tandem duplications. Many of the identified mouse segmental duplications were copy number polymorphisms of immune response genes; similarly we found segmental duplications were also enriched in immune response genes in human. However, it is worth investigating if the differences in genomic architecture of segmental duplications in mouse and human perpetuate to differences in gene functions.

Further exploration of mappability in our GSE methods and different HTS data will show to what degree mappability improves performance and biological findings. For example, mappability may play an even larger role in bisulfite sequencing data, where unmethylated cytosines have been converted to uracil (and read as thymine on sequencers). The genome is essentially reduced to a three letter alphabet, significantly reducing the unique information content of short read sequences.

Bibliography

1. Trapnell C, Salzberg SL: **How to map billions of short reads onto genomes**. *Nature Biotechnology* 2009, **27**(5):455-457.
2. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies**. *Nat Methods* 2009, **6**(11 Suppl):S22-32.
3. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**(1):57-63.
4. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology**. *Nat Rev Genet* 2009, **10**(10):669-680.
5. Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome**. *Nat Rev Genet* 2006, **7**(2):85-97.
6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.
7. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat Biotechnol* 2008, **26**(10):1135-1145.
8. Oshlack A, Wakefield MJ: **Transcript length bias in RNA-seq data confounds systems biology**. *Biol Direct* 2009, **4**:14.
9. Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-seq: accounting for selection bias**. *Genome Biol* 2010, **11**(2):R14.
10. Hansen KD, Brenner SE, Dudoit S: **Biases in Illumina transcriptome sequencing caused by random hexamer priming**. *Nucleic Acids Res* 2010, **38**(12):e131.
11. Bohnert R, Ratsch G: **rQuant.web: a tool for RNA-Seq-based transcript quantitation**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W348-351.
12. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: **Improving RNA-Seq expression estimates by correcting for fragment bias**. *Genome Biol* 2011, **12**(3):R22.
13. Li J, Jiang H, Wong WH: **Modeling non-uniformity in short-read rates in RNA-Seq data**. *Genome Biol* 2010, **11**(5):R50.
14. Furey TS: **ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions**. *Nat Rev Genet* 2012, **13**(12):840-852.
15. Wilbanks EG, Facciotti MT: **Evaluation of algorithm performance in ChIP-seq peak detection**. *PLoS One* 2010, **5**(7):e11471.
16. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P *et al*: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**. *Genome Res* 2012, **22**(9):1813-1831.
17. Zhang Y, Lin YH, Johnson TD, Rozek LS, Sartor MA: **PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data**. *Bioinformatics* 2014.
18. Sun GN, Chung DH, Liang K, Keles S: **Statistical Analysis of ChIP-seq Data with MOSAiCS**. *Methods Mol Biol* 2013, **1038**:193-212.
19. Bansal M, Mendiratta G, Anand S, Kushwaha R, Kim R, Kustagi M, Iyer A, Chaganti RS, Califano A, Sumazin P: **Direct ChIP-Seq significance analysis improves target prediction**. *BMC Genomics* 2015, **16 Suppl 5**:S4.

20. Mills RE, Bennett EA, Iskow RC, Devine SE: **Which transposable elements are active in the human genome?** *Trends in Genetics* 2007, **23**(4):183-191.
21. Simons C, Pheasant M, Makunin IV, Mattick JS: **Transposon-free regions in mammalian genomes**. *Genome Res* 2006, **16**(2):164-172.
22. Deininger P: **Alu elements: know the SINEs**. *Genome Biol* 2011, **12**(12):236.
23. **Gene Ontology Consortium: going forward**. *Nucleic Acids Res* 2015, **43**(Database issue):D1049-1056.
24. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation**. *Nucleic Acids Res* 2016, **44**(D1):D457-D462.
25. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.
26. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 2009, **37**(1):1-13.
27. Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, Reimers M, Stephens RM, Bryant D, Burt SK *et al*: **High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID)**. *BMC Bioinformatics* 2005, **6**:168.
28. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S *et al*: **GoMiner: a resource for biological interpretation of genomic and proteomic data**. *Genome Biol* 2003, **4**(4):R28.
29. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities**. *Mol Cell* 2010, **38**(4):576-589.
30. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E *et al*: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**(3):267-273.
31. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.
32. Xiong Q, Mukherjee S, Furey TS: **GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data**. *Sci Rep* 2014, **4**:6347.
33. Wang X, Cairns MJ: **SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing**. *Bioinformatics* 2014, **30**(12):1777-1779.
34. Newton MA, Quintana FA, Den Boon JA, Sengupta S, Ahlquist P: **Random-Set Methods Identify Distinct Aspects of the Enrichment Signal in Gene-Set Analysis**. *Ann Appl Stat* 2007, **1**(1):85-106.
35. Kim JH, Karnovsky A, Mahavisno V, Weymouth T, Pande M, Dolinoy DC, Rozek LS, Sartor MA: **LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types**. *BMC Genomics* 2012, **13**:526.
36. Sartor MA, Leikauf GD, Medvedovic M: **LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data**. *Bioinformatics* 2009, **25**(2):211-217.

37. Cavalcante RG, Lee C, Welch RP, Patil S, Weymouth T, Scott LJ, Sartor MA: **Broad-Enrich: functional interpretation of large sets of broad genomic regions**. *Bioinformatics* 2014, **30**(17):I393-I400.
38. Curtis RK, Oresic M, Vidal-Puig A: **Pathways to the analysis of microarray data**. *Trends Biotechnol* 2005, **23**(8):429-435.
39. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.
40. Rivals I, Personnaz L, Taing L, Potier MC: **Enrichment or depletion of a GO category within a class of genes: which test?** *Bioinformatics* 2007, **23**(4):401-407.
41. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.
42. Joshi A, Hannah R, Diamanti E, Gottgens B: **Gene set control analysis predicts hematopoietic control mechanisms from genome-wide transcription factor binding data**. *Experimental hematology* 2013, **41**(4):354-366 e314.
43. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A: **ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments**. *Bioinformatics* 2010, **26**(19):2438-2444.
44. Wu D, Smyth GK: **Camera: a competitive gene set test accounting for inter-gene correlation**. *Nucleic Acids Res* 2012, **40**(17):e133.
45. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK: **ROAST: rotation gene set tests for complex microarray experiments**. *Bioinformatics* 2010, **26**(17):2176-2182.
46. Efron B, Tibshirani R: **On Testing the Significance of Sets of Genes**. *Ann Appl Stat* 2007, **1**(1):107-129.
47. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions**. *Nat Biotechnol* 2010, **28**(5):495-501.
48. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R: **A systems biology approach for pathway level analysis**. *Genome Res* 2007, **17**(10):1537-1545.
49. Khatri P, Voichita C, Kattan K, Ansari N, Khatri A, Georgescu C, Tarca AL, Draghici S: **Onto-Tools: new additions and improvements in 2006**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W206-211.
50. Sartor MA, Mahavisno V, Keshamouni VG, Cavalcoli J, Wright Z, Karnovsky A, Kuick R, Jagadish HV, Mirel B, Weymouth T *et al*: **ConceptGen: a gene set enrichment and gene set relation mapping tool**. *Bioinformatics* 2010, **26**(4):456-463.
51. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture**. *Nat Genet* 1999, **22**(3):281-285.
52. Segre AV, Consortium D, investigators M, Groop L, Mootha VK, Daly MJ, Altshuler D: **Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits**. *PLoS Genet* 2010, **6**(8).
53. Taher L, Ovcharenko I: **Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements**. *Bioinformatics* 2009, **25**(5):578-584.
54. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L: **Evolution and functional classification of vertebrate gene deserts**. *Genome Res* 2005, **15**(1):137-145.
55. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Wellcome Trust Case-Control C, Owen MJ, O'Donovan MC, Craddock N: **Gene ontology analysis of**

GWA study data sets provides insights into the biology of bipolar disorder. *American journal of human genetics* 2009, **85**(1):13-24.

56. Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D *et al*: **Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium**. *Nucleic Acids Res* 2013, **41**(Database issue):D171-176.
57. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C *et al*: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs**. *Nat Biotechnol* 2010, **28**(5):503-510.
58. Kharchenko PV, Tolstorukov MY, Park PJ: **Design and analysis of ChIP-seq experiments for DNA-binding proteins**. *Nat Biotechnol* 2008, **26**(12):1351-1359.
59. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al*: **Model-based analysis of ChIP-Seq (MACS)**. *Genome Biol* 2008, **9**(9):R137.
60. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: **A simulation study of the number of events per variable in logistic regression analysis**. *Journal of clinical epidemiology* 1996, **49**(12):1373-1379.
61. Tarone RE: **Testing the goodness of fit of the binomial distribution.** . *Biometrika* 1979, **66**(3):585-590.
62. Wood SN: **Generalized additive models : an introduction with R**: Chapman & Hall/CRC; 2006.
63. Wood SN: **Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011, **73**(1):3-36.
64. Wood SN: **mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL**. *R package version* 2010:1.6-2.
65. Benjamini Y, Hochberg Y: **CONTROLLING THE FALSE DISCOVERY RATE - A PRACTICAL AND POWERFUL APPROACH TO MULTIPLE TESTING**. *J R Stat Soc Ser B-Methodol* 1995, **57**(1):289-300.
66. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2009, **4**(1):44-57.
67. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci U S A* 2003, **100**(16):9440-9445.
68. Amoli MM, Carthy D, Platt H, Ollier WE: **EBV Immortalization of human B lymphocytes separated from small volumes of cryo-preserved whole blood**. *International journal of epidemiology* 2008, **37 Suppl 1**:i41-45.
69. Upton GJG: **Fisher's exact test**. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1992, **155**(3):395-402.
70. Hernandez JM, Floyd DH, Weilbaecher KN, Green PL, Boris-Lawrie K: **Multiple facets of junD gene expression are atypical among AP-1 family members**. *Oncogene* 2008, **27**(35):4757-4767.
71. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome**. *Cell* 2007, **129**(4):823-837.
72. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions**. *Nature* 2012, **485**(7398):376-380.
73. Comai L, Li B: **The Werner syndrome protein at the crossroads of DNA repair and apoptosis**. *Mechanisms of ageing and development* 2004, **125**(8):521-528.

74. Hayashi T, Seki M, Inoue E, Yoshimura A, Kusa Y, Tada S, Enomoto T: **Vertebrate WRNIP1 and BLM are required for efficient maintenance of genome stability**. *Genes & genetic systems* 2008, **83**(1):95-100.

75. Rajagopalan S, Nepa J, Venkatachalam S: **Chromodomain helicase DNA-binding protein 2 affects the repair of X-ray and UV-induced DNA damage**. *Environmental and molecular mutagenesis* 2012, **53**(1):44-50.

76. Wakasugi T, Izumi H, Uchiumi T, Suzuki H, Arao T, Nishio K, Kohno K: **ZNF143 interacts with p73 and is involved in cisplatin resistance through the transcriptional regulation of DNA repair genes**. *Oncogene* 2007, **26**(36):5194-5203.

77. Ramirez J, Lukin K, Hagman J: **From hematopoietic progenitors to B cells: mechanisms of lineage restriction and commitment**. *Current opinion in immunology* 2010, **22**(2):177-184.

78. Canudas S, Smith S: **Differential regulation of telomere and centromere cohesion by the Scc3 homologues SA1 and SA2, respectively, in human cells**. *The Journal of cell biology* 2009, **187**(2):165-173.

79. Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, Myers RM: **Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation**. *Genome Res* 2009, **19**(12):2163-2171.

80. Yu CY, Mayba O, Lee JV, Tran J, Harris C, Speed TP, Wang JC: **Genome-wide analysis of glucocorticoid receptor binding regions in adipocytes reveal gene network involved in triglyceride homeostasis**. *PloS one* 2010, **5**(12):e15188.

81. Xu C, He J, Jiang H, Zu L, Zhai W, Pu S, Xu G: **Direct effect of glucocorticoids on lipolysis in adipocytes**. *Mol Endocrinol* 2009, **23**(8):1161-1170.

82. Ayalasomayajula SP, Ashton P, Kompella UB: **Fluocinolone inhibits VEGF expression via glucocorticoid receptor in human retinal pigment epithelial (ARPE-19) cells and TNF-alpha-induced angiogenesis in chick chorioallantoic membrane (CAM)**. *Journal of ocular pharmacology and therapeutics : the official journal of the Association for Ocular Pharmacology and Therapeutics* 2009, **25**(2):97-103.

83. Dostert A, Heinzel T: **Negative glucocorticoid receptor response elements and their role in glucocorticoid action**. *Current pharmaceutical design* 2004, **10**(23):2807-2816.

84. Small GR, Hadoke PW, Sharif I, Dover AR, Armour D, Kenyon CJ, Gray GA, Walker BR: **Preventing local regeneration of glucocorticoids by 11beta-hydroxysteroid dehydrogenase type 1 enhances angiogenesis**. *Proc Natl Acad Sci U S A* 2005, **102**(34):12165-12170.

85. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):3.

86. Kim JH, Karnovsky A, Mahavisno V, Weymouth T, Pande M, Dolinoy DC, Rozek LS, Sartor MA: **LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types**. *BMC genomics* 2012, **13**(1):526.

87. Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD: **ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions**. *Genome Biol* 2011, **12**(7):R67.

88. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls**. *Nat Biotechnol* 2009, **27**(1):66-75.

89. Upton GJG: **Fisher's Exact Test**. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1992, **155**(3):395-402.

90. Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW: **Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model**. *P Natl Acad Sci USA* 2008, **105**(51):20179-20184.

91. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.

92. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.

93. Han J, Back SH, Hur J, Lin YH, Gildersleeve R, Shan J, Yuan CL, Krokowski D, Wang S, Hatzoglou M *et al*: **ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death**. *Nat Cell Biol* 2013, **15**(5):481-490.

94. Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, Ribeca P: **Fast computation and applications of genome mappability**. *PLoS One* 2012, **7**(1):e30377.

95. Harmanci A, Rozowsky J, Gerstein M: **MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability-Corrected Multiscale Signal Processing Framework**. *Genome Biol* 2014, **15**(10):474.

96. Janevski A, Varadan V, Kamalakaran S, Banerjee N, Dimitrova N: **Effective normalization for copy number variation detection from whole genome sequencing**. *BMC Genomics* 2012, **13 Suppl 6**:S16.

97. Niimura Y, Nei M: **Evolution of olfactory receptor genes in the human genome**. *Proc Natl Acad Sci U S A* 2003, **100**(21):12235-12240.

98. Nelson CE, Hersh BM, Carroll SB: **The regulatory content of intergenic DNA shapes genome architecture**. *Genome Biol* 2004, **5**(4):R25.

99. de Koning APJ, Gu WJ, Castoe TA, Batzer MA, Pollock DD: **Repetitive Elements May Comprise Over Two-Thirds of the Human Genome**. *Plos Genet* 2011, **7**(12).

100. Kazazian HH, Jr., Moran JV: **The impact of L1 retrotransposons on the human genome**. *Nat Genet* 1998, **19**(1):19-24.

101. Lander ES, Consortium IHGS, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

102. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE: **Segmental duplications: Organization and impact within the current Human Genome Project assembly**. *Genome Research* 2001, **11**(6):1005-1017.

103. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome**. *Science* 2002, **297**(5583):1003-1007.

104. Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui LC, Scherer SW: **Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence**. *Genome Biology* 2003, **4**(4).

105. Zhang L, Lu HH, Chung WY, Yang J, Li WH: **Patterns of segmental duplication in the human genome**. *Mol Biol Evol* 2005, **22**(1):135-141.

106. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H *et al*: **Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication**. *Cell* 2012, **149**(4).

107. Avramopoulos D, Wang RH, Valle D, Fallin MD, Bassett SS: **A novel gene derived from a segmental duplication shows perturbed expression in Alzheimer's disease**. *Neurogenetics* 2007, **8**(2):111-120.

108. Stankiewicz P, Park SS, Inoue K, Lupski JR: **The evolutionary chromosome translocation 4;19 in Gorilla gorilla is associated with microduplication of the chromosome fragment syntenic to sequences surrounding the human proximal CMT1A-REP**. *Genome Res* 2001, **11**(7):1205-1210.

109. Edelmann L, Stankiewicz P, Spiteri E, Pandita RK, Shaffer L, Lupski J, Morrow BE: **Two functional copies of the DGCR6 gene are present on human chromosome 22q11**

due to a duplication of an ancestral locus (vol 11, pg 208, 2001). *Genome Research* 2001, **11**(3):503-503.

110. Ganapathi M, Srivastava P, Das Sutar SK, Kumar K, Dasgupta D, Pal Singh G, Brahmachari V, Brahmachari SK: **Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes**. *BMC Bioinformatics* 2005, **6**:126.
111. Polak P, Domany E: **Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes**. *BMC Genomics* 2006, **7**:133.
112. Zemojtel T, Kielbasa SM, Arndt PF, Behrens S, Bourque G, Vingron M: **CpG deamination creates transcription factor-binding sites with high efficiency**. *Genome Biol Evol* 2011, **3**:1304-1311.
113. Grover D, Majumder PP, C BR, Brahmachari SK, Mukerji M: **Nonrandom distribution of alu elements in genes of various functional categories: insight from analysis of human chromosomes 21 and 22**. *Mol Biol Evol* 2003, **20**(9):1420-1424.
114. Wanichnopparat W, Suwanwongse K, Pin-On P, Aporntewan C, Mutirangura A: **Genes associated with the cis-regulatory functions of intragenic LINE-1 elements**. *BMC Genomics* 2013, **14**:205.
115. Bailey JA, Carrel L, Chakravarti A, Eichler EE: **Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis**. *Proc Natl Acad Sci U S A* 2000, **97**(12):6634-6639.
116. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, Scott LJ, Sartor MA: **ChIP-Enrich: gene set enrichment testing for ChIP-seq data**. *Nucleic Acids Res* 2014, **42**(13):e105.
117. Lee C, Patil S, Sartor MA: **RNA-Enrich: a cut-off free functional enrichment testing method for RNA-seq with improved detection power**. *Bioinformatics* 2015.
118. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W202-208.
119. Zhang Y, Lin YH, Johnson TD, Rozek LS, Sartor MA: **PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data**. *Bioinformatics* 2014, **30**(18):2568-2575.
120. Letunic I, Copley RR, Bork P: **Common exon duplication in animals and its role in alternative splicing**. *Hum Mol Genet* 2002, **11**(13):1561-1567.
121. Gombart AF, Saito T, Koeffler HP: **Exaptation of an ancient Alu short interspersed element provides a highly conserved vitamin D-mediated innate immune response in humans and primates**. *BMC Genomics* 2009, **10**:321.
122. Bailey JA, Eichler EE: **Primate segmental duplications: crucibles of evolution, diversity and disease**. *Nat Rev Genet* 2006, **7**(7):552-564.
123. Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV: **Duplication, coclustering, and selection of human Alu retrotransposons**. *Proc Natl Acad Sci U S A* 2004, **101**(5):1268-1272.
124. Zhou Y, Mishra B: **Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling**. *Proc Natl Acad Sci U S A* 2005, **102**(11):4051-4056.
125. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D: **Human genome ultraconserved elements are ultraselected**. *Science* 2007, **317**(5840):915.
126. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome**. *Science* 2004, **304**(5675):1321-1325.

127. Lawson MJ, Zhang L: **Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5'-UTR region**. *Gene* 2008, **407**(1-2):54-62.

128. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I: **Predicting tissue-specific enhancers in the human genome**. *Genome Res* 2007, **17**(2):201-211.

129. Kleinjan DA, van Heyningen V: **Long-range control of gene expression: emerging mechanisms and disruption in disease**. *American journal of human genetics* 2005, **76**(1):8-32.

130. Tsirigos A, Rigoutsos I: **Alu and B1 Repeats Have Been Selectively Retained in the Upstream and Intronic Regions of Genes of Specific Functional Classes**. *Plos Computational Biology* 2009, **5**(12).

131. Liang KC, Tseng JT, Tsai SJ, Sun HS: **Characterization and distribution of repetitive elements in association with genes in the human genome**. *Comput Biol Chem* 2015, **57**:29-38.

132. Stankiewicz P, Inoue K, Bi W, Walz K, Park SS, Kurotaki N, Shaw CJ, Fonseca P, Yan J, Lee JA *et al*: **Genomic disorders: genome architecture results in susceptibility to DNA rearrangements causing common human traits**. *Cold Spring Harb Symp Quant Biol* 2003, **68**:445-454.

133. Stankiewicz P, Lupski JR: **Genome architecture, rearrangements and genomic disorders**. *Trends Genet* 2002, **18**(2):74-82.

134. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T *et al*: **An atlas of active enhancers across human cell types and tissues**. *Nature* 2014, **507**(7493):455-461.

135. Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, Barillot E: **Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization**. *Bioinformatics* 2011, **27**(2):268-269.

136. Risso D, Schwartz K, Sherlock G, Dudoit S: **GC-content normalization for RNA-Seq data**. *BMC Bioinformatics* 2011, **12**:480.

137. Yu HL, Zhao ZK, Zhu F: **The role of human endogenous retroviral long terminal repeat sequences in human cancer (Review)**. *Int J Mol Med* 2013, **32**(4):755-762.

138. Medstrand P, van de Lagemaat LN, Mager DL: **Retroelement distributions in the human genome: variations associated with age and proximity to genes**. *Genome Res* 2002, **12**(10):1483-1495.

139. Boissinot S, Entezam A, Furano AV: **Selection against deleterious LINE-1-containing loci in the human lineage**. *Mol Biol Evol* 2001, **18**(6):926-935.

140. Quentin Y: **Emergence of master sequences in families of retroposons derived from 7sl RNA**. *Genetica* 1994, **93**(1-3):203-215.

141. Batzer MA, Deininger PL: **Alu repeats and human genomic diversity**. *Nat Rev Genet* 2002, **3**(5):370-379.

142. Bailey JA, Church DM, Ventura M, Rocchi M, Eichler EE: **Analysis of segmental duplications and genome assembly in the mouse**. *Genome Res* 2004, **14**(5):789-801.

143. She X, Cheng Z, Zollner S, Church DM, Eichler EE: **Mouse segmental duplication and copy number variation**. *Nat Genet* 2008, **40**(7):909-914.