Annotation-Agnostic Differential Expression and Binding Analyses

by

Leonardo Collado Torres

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

June, 2016

Abstract

My thesis work is centered around the development of `R` software packages for analyzing RNA sequencing (RNA-seq) and ChIP sequencing (ChIP-seq) high throughput genomic data. Chapter 2 describes the `derfinder` Bioconductor package which implements the DER Finder approach for identifying differentially expressed regions with RNA-seq data in an annotation-agnostic manner. Chapter 3 shows how `derfinder` can be applied to ChIP-seq data to identify differentially bounded regions. Chapter 4 describes the `regionReport` Bioconductor package for producing HTML or PDF reports from region-based genomic analyses, such as the `derfinder` analyses described in the previous chapters.

Thesis Committee

Primary Readers

Jeffrey Leek (Primary Advisor)
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Daniele Fallin (Committee Chair)
Professor
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health

Kasper Hansen
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Alexis Battle
Assistant Professor
Department of Computer Science
Johns Hopkins Whiting School of Engineering

Andrew Jaffe (5th non-voting member)
Assistant Professor
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health

Alternate Readers

Hongkai Ji
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Fernando Pineda
Associate Professor
Department of Molecular Microbiology and Immunology
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

There are many people I would like to thank for their support and guidance along my career, as well as more specifically during my Ph.D. I tried my best to include everyone.

This story would have not have started without Alejandra Medina convincing me to attend BioC2008, the annual Bioconductor meeting. At BioC2010 Rafael Irizarry and Ingo Ruczinski convinced me to apply to the Department of Biostatistics's Ph.D. program and I am grateful for their help during the admission process. I would also like to thank Daniel Sharfstein and Karen Bandeen-Roche for believing in me. Like my middle school teacher Véronique Georges once told me, you only need someone to give you one opportunity and then it's up to you to make the best of it.

Thank you Thomas Louis for helping me in my transition to the field of Biostatistics as my first year academic advisor, our meetings were always fun. Thank you Kasper Hansen for the valuable advice during my Ph.D., particularly regarding time management and cultural differences. Thank you Karen and Rafa for being there for me when I dislocated my left shoulder during the first year. Than you again Rafa for introducing me to the DER Finder project and my primary advisor Jeffrey Leek.

I specially thank Jeff for his complete support, positive feedback, guidance, energy, and incredible ideas, some of which I had the privilege to work on and contribute. I think that you are doing a terrific work as an advisor and feel proud to have had the opportunity to work with you. Also thank you for being a good sport whenever México beat the US in fútbol.

Andrew Jaffe, thank you very much for inviting me to work with you in the middle of my Ph.D. Had I not known, I would have not guessed that I was your first Ph.D. student. That speaks for itself about your professionalism and aptitude for advising students. It has been a pleasure working with you, learning how to get things done, and grasping the sweet spot between genomic data science and biostatistics. You are the most productive while traveling person that I know.

I am also grateful to other members of our weekly meetings: Alyssa Frazee who helped me get used to Enigma and working with Jeff, Prasad Patil who always had interesting topics to talk about, Abhinav Nellore who is the Slack master, Kai Kammers, Jack Fu, Claire Ruberman and Margaret Taub.

Thank you Marie Diener-West and Karen for helping me become a good teaching assistant, the opportunities you provided me to keep improving and your passion to teach.

Thank you Jeff, Andrew, Marie, Karen and Brian Caffo for supporting me when doctors scared me more than they should have in 2015.

Thank you Margaret, Jeff, Andrew, Rafa and Alexis Battle for helping me during my job hunt at the end of my Ph.D. The whole process was way more complicated than I had anticipated and it was interesting to learn of all the

different nuances involved as well as opinions on the best path to follow.

I would also like to thank my Oral Exam and Thesis Defense committee members Ben Langmead, Hongkai Ji, James Taylor, Jiou Wang, Ingo Ruczinski, Fernando Pineda, Daniele Fallin, Kasper Hansen, Alexis Battle, Andrew Jaffe, and Jeff Leek.

However, work is just one component of our lives. As such I want to thank the Department of Biostatistics (including all staff members!) for being my home this past few years and all the fun moments like tea time, the department retreats, get togethers at ENAR, among other events. Thank you Karen for supporting the cultural mixers I co-organized with Amanda Mejia, the Genomics for Students group I co-organized with Jean-Philippe Fortin and Kasper, and the Epi vs Biostat events. With your support I was able to implement some ideas which I believe improved the department.

I am grateful to my Baltimore friends, friends from México, and friends from my undergrad for their support, conversations, trips together, and for always believing in me. I look forward to hanging out with all of you in the future, sharing news and enjoying our lives regardless of the career paths that we decide to follow. There are too many of you to mention everyone, but you know who you are!

Finally, I am thankful to my family, even if we get a bit too nerdy at times with our *family symposiums*. My mom has always and will always cheer for me, my dad always pushes me to do better, and they both have provided me with everything that I ever needed. The person I am most grateful to is my brother who is the light that dispels all darkness. He is probably the most

ambitious person I know that is achieving all his goals and is a fundamental source of inspiration for me. You make me feel like I am there with you when you climb mountains. My family's love is stronger than I can describe it, just check https://youtu.be/UBTPPlLDsE4.

Funding

LCT was supported by Consejo Nacional de Ciencia y Tecnología México 351535.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis work is part of a larger collective effort to address the public health problem presented by neuropsychiatric disorders. This thesis work will lead to improving our understanding of the genomic data and will generate hypothesis, which will then be further analyzed by the team of scientists at the Lieber Institute for Brain Development as well as other institutions. This work is part of the team effort to improve the health and quality of life of individuals with neuropsychiatric disorders.

The goal of this thesis work is to develop statistical methods and software that enable researchers to differentiate the sources of variation observed in RNA-seq while minimizing the dependance on known annotation. This will allow researchers to correct for technological variation and study the biological variation driving their phenotype of interest. Then apply these methods to further our understanding of neuropsychiatric disorders using the Lieber Institute for Brain Development human brains collection (> 1000 samples).

To accomplish this goal, this thesis work was jointly supervised by Jeffrey

T. Leek from the Johns Hopkins Bloomberg School of Public Health Department of Biostatistics and Andrew E. Jaffe from the Lieber Institute for Brain Development. This work resulted in several R packages which are all part of the `Bioconductor` project [1], which means that the software is very well documented, regularly tested, and easy to install by R users.

1.1 `derfinder` applied to RNA-seq data

Differential expression analysis of RNA sequencing (RNA-seq) data typically relies on reconstructing transcripts or counting reads that overlap known gene structures. In under to better understand the human brain, transcriptome analysis provides fundamental insight into development and disease. However, this type of analysis typically relies on the existing annotation which might not be complete in some situations, particularly for less studied organisms.

As a complement to typical transcriptome analysis pipelines, the DER Finder statistical approach seeks to identify contiguous regions of the genome showing differential expression signal at single base-pair resolution [2]. DER Finder does not rely on existing annotation or potentially incomplete transcriptome, thus allowing researchers to further study tissues like the human brain. Chapter 2 describes the `derfinder R` package that implements the DER Finder approach [3] with visualizations created with the `derfinderPlot` [4] package. We used it with data generated by the Lieber Institute for Brain Development and determined that the human brain transcriptome annotation is incomplete [5].

1.2 derfinder applied to ChIP-seq data

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments identify regions of the genome with binding signal for a protein of interest. When multiple samples are collected for different conditions, treatments or other covariates, researchers will ask if there is differential binding between these conditions. The current strategies for answering this question rely on merging peaks from the different samples which can lead to un-wanted issues. These strategies do not take into account the variability across samples when merging peaks. Chapter 3 shows how `derfinder` can be used with ChIP-seq data to identify differentially bound regions. We illustrate this application using data from the *EpiMap* study [6] for histone marks H3K4me3 and H3K27ac from the anterior cingulate cortex and dorsolateral prefrontal complex of the human brain.

1.3 Interactive region based reports with regionReport

Many analyses of genomic data result in regions along the genome that associate with a covariate of interest. These genomic regions can result from identifying differentially bound peaks from ChIP-seq data, identifying differentially methylated regions (DMRs) from DNA methylation data, `derfinder` analyses, among other analysis pipelines. The genomic regions themselves are commonly stored in a `GRanges` object from `GenomicRanges` [7] and have some common properties such as p-values associated with each region. Chapter

4 describes the `regionReport` [8] R package for creating interactive HTML reports for region-based genomic analyses. These reports are useful for exploring results and can be shared with collaborators. `regionReport` can also be used to explore `DESeq2` [9] and `edgeR-robust` [10] results, which are among the most widely used differential expression software packages.

References

[1] R. C. Gentleman et al. "Bioconductor: Open software development for computational biology and bioinformatics". In: *Genome Biology* 5 (2004), R80.

[2] A. C. Frazee et al. "Differential expression analysis of RNA-seq data at single-base resolution". In: *Biostatistics* 15 (2014). PMID: 24398039, pp. 413–426. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxt053.

[3] L. Collado-Torres et al. "Flexible expressed region analysis for RNA-seq with derfinder". In: *bioRxiv* (2016), p. 015370. DOI: 10.1101/015370.

[4] L. Collado-Torres, A. E. Jaffe, and J. T. Leek. *derfinderPlot: Plotting functions for derfinder*. Version 1.6.0. 2015. URL: http://www.bioconductor.org/packages/derfinderPlot.

[5] A. E. Jaffe et al. "Developmental regulation of human cortex transcription and its clinical relevance at single base resolution". In: *Nature Neuroscience* 18.1 (2015). PMID: 25501035 PMCID: PMC4281298, pp. 154–161. ISSN: 1546-1726. DOI: 10.1038/nn.3898.

[6] S. Akbarian and P. Sklar. *Cis-Regulatory Epigenome Mappings in Schizophrenia*. U01MH103392. EpiMap. 2016.

[7] M. Lawrence et al. "Software for Computing and Annotating Genomic Ranges". In: *PLoS Computational Biology* 9 (8 2013). PMID: 23950696 PMCID: PMC3738458, e1003118. DOI: 10.1371/journal.pcbi.1003118.

[8] L. Collado-Torres, A. E. Jaffe, and J. T. Leek. "regionReport: Interactive reports for region-level and feature-level genomic analyses [version2; referees: 2 approved, 1 approved with reservations]". In: *F1000Research* 4 (2016), pp. 1–10. DOI: 10.12688/f1000research.6379.2.

[9] M. I. Love, W. Huber, and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". eng. In: *Genome Biology* 15.12 (2014). PMID: 25516281 PMCID: PMC4302049, p. 550. ISSN: 1465-6914. DOI: 10.1186/s13059-014-0550-8.

[10] X. Zhou, H. Lindsay, and M. D. Robinson. "Robustly detecting differential expression in RNA sequencing data using observation weights". eng. In: *Nucleic Acids Research* 42.11 (2014). PMID: 24753412 PMCID: PMC4066750, e91. ISSN: 1362-4962. DOI: 10.1093/nar/gku310.

Chapter 2

Flexible expressed region analysis for RNA-seq with `derfinder`

Leonardo Collado-Torres[1,2,3], Abhinav Nellore[1,2,4], Alyssa C. Frazee[1,2], Christopher Wilks[2,4], Michael I. Love[5,6], Ben Langmead[1,2,4], Rafael A. Irizarry[5,6], Jeffrey T. Leek[1,2,*], Andrew E. Jaffe[1,2,3,7,†].

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
2. Center for Computational Biology, Johns Hopkins University
3. Lieber Institute for Brain Development, Johns Hopkins Medical Campus
4. Department of Computer Science, Johns Hopkins University
5. Department of Biostatistics, Harvard T.H. Chan School of Public Health
6. Dana-Farber Cancer Institute, Harvard University
7. Department of Mental Health, Johns Hopkins University

* *corresponding author;* `jtleek@gmail.com`

† *corresponding author;* andrew.jaffe@libd.org

2.1 Abstract

Background Differential expression analysis of RNA sequencing (RNA-seq) data typically relies on reconstructing transcripts or counting reads that overlap known gene structures. We previously introduced an intermediate statistical approach called differentially expressed region (DER) finder that seeks to identify contiguous regions of the genome showing differential expression signal at single base resolution without relying on existing annotation or potentially inaccurate transcript assembly.

Results

We present the `derfinder` software that improves our annotation-agnostic approach to RNA-seq analysis by: (1) implementing a computationally efficient bump-hunting approach to identify DERs which permits genome-scale analyses in a large number of samples, (2) introducing a flexible statistical modeling framework, including multi-group and time-course analyses and (3) introducing a new set of data visualizations for expressed region analysis. We apply this approach to public RNA-seq data from the Genotype-Tissue Expression (GTEx) project and BrainSpan project to show that `derfinder` permits the analysis of hundreds of samples at base resolution in R, identifies expression outside of known gene boundaries and can be used to visualize expressed regions at base-resolution. In simulations our base resolution approaches enable discovery in the presence of incomplete annotation and is nearly as powerful as feature-level methods when the annotation is complete.

Conclusions `derfinder` analysis using expressed region-level and single base-level approaches provides a compromise between full transcript reconstruction and feature-level analysis. The package is available from `Bioconductor` at `www.bioconductor.org/packages/derfinder`.

Keywords RNA sequencing, differential expression analysis, coverage, gene annotation, gene expression.

2.2 Introduction

The increased flexibility of RNA sequencing (RNA-seq) has made it possible to characterize the transcriptomes of a diverse range of experimental systems, including human tissues [1, 2, 3], cell lines [4, 5] and model organisms [6, 7]. The goal of many experiments involves identifying differential expression with respect to disease, development, or treatment. In experiments using RNA-seq, RNA is sequenced to generate short "reads" (36-200+ base pairs). These reads are aligned to a reference genome, and this alignment information is used to quantify the transcriptional activity of both annotated (present in databases like Ensembl) and novel transcripts and genes.

The ability to quantitatively measure expression levels in regions not previously annotated in gene databases, particularly in tissues or cell types that are difficult to ascertain, is one key advantage of RNA-seq over hybridization-based assays like microarray technologies. As complicated transcript structures are difficult to completely characterize using short read sequencing technologies [8], the most mature statistical methods used for RNA-seq analysis rely on existing annotation for defining regions of interest - such as genes

or exons - and counting reads that overlap those regions [9]. These counts are then used as measures of gene expression abundance for downstream differential expression analysis [10, 11, 12, 13, 14, 15, 16, 17, 18]. Unfortunately, the gene annotation may be incorrect or incomplete, which can affect downstream modeling of the number of reads that cross these defined features.

We previously proposed an alternative statistical model for finding differentially expressed regions (DERs) that first identifies regions that show differential expression signal and then annotates these regions using previously annotated genomic features [19]. This analysis framework first proposed using coverage tracks (i.e. the number of reads aligned to each base in the genome) to identify differential expression signal at each individual base and merges adjacent bases with similar signal into candidate regions. However, the software for our first version was limited to small sample sizes, the ability to interrogate targeted genomic loci, and comparisons between only two groups.

Here we expand the DER finder framework to permit the analysis of larger sample sizes with more flexible statistical models across the genome. This paper introduces a comprehensive software package called `derfinder` built upon base-resolution analysis, which performs coverage calculation, preprocessing, statistical modeling, region annotation and data visualization. This software permits differential expression analysis at both the single base level, resulting in direct calculation of DERs [20], and a feature summarization we introduce here call "expressed region" (ER)-level analysis. We show that ER analysis allows us to perform base resolution analysis on larger scale RNA-seq

data sets using the BrainSpan project [21] http://developinghumanbrain.org and Genotype-Tissue Expression (GTEx) project data [3] to demonstrate that derfinder can identify differential expression signal in regions outside of known annotation without assembly. We use these DERs to illustrate the post-discovery annotation capabilities of derfinder and label each DER as exonic, intronic, intergenic or some combination of those labels. We show that some of these DERs we identify are outside of annotated protein coding regions and would not have been identified using gene or exon counting approaches.

In the GTEx data, we identify differentially expressed regions (DERs) that differentiate heart (left ventricle), testis and liver tissues for 8 subjects. There are many potential reasons for this observed intronic expression including intron retention, background levels of mis-transcription, or incomplete protein-coding annotation. A subset of these strictly intronic ERs are associated with tissue differences, even conditional on the expression of the nearest annotated protein-coding region. However, we point out that intronic expression may be artifactual and it our package permits visualization and discovery of potential expression artifacts not possible with other packages.

Finally, using simulated differentially expressed transcripts, we demonstrate that when transcript annotation is correct, derfinder is nearly as powerful as exon-count based approaches with statistical tests performed by limma [16] (or DESeq2 [14], edgeR-robust [13]) and ballgown [22] after summarizing the information using Rsubread [13] and StringTie [23] respectively. Finally, we also demonstrate that when annotation is incomplete, derfinder can be

substantially more powerful than methods that rely on a complete annotation.

2.3 Materials & Methods

2.3.1 Overview of R Implementation

We chose to implement `derfinder` entirely in the `R` statistical environment `www.R-project.org/`. Our software includes upstream pre-processing of BAM and/or BigWig files into base-resolution coverage. At this stage the user can choose to summarize the base resolution coverage into feature-level counts and apply popular feature-level RNA-seq differential expression analysis tools like `DESeq2` [14], `edgeR-robust` [13], `limma` [16, 15] and `voom` [17].

`derfinder` can be used to identify regions of differential expression agnostic to existing annotation (Figure 4.1). This can be done with either the expressed regions (ER)-level or single base-level approaches, described in detail in the following subsection and Supplementary Section 2.12.1. The resulting regions can then be visualized to identify novel regions and filter out potential artifacts.

After differential expression analysis, `derfinder` can plot DERs using base-resolution coverage data by accessing the raw reads within differentially expressed regions for posthoc analysis like clustering and sensitivity analyses. We have also created a lightweight annotation function for quickly annotating DERs based on existing transcriptome annotation, including the UCSC knownGene hg19, Ensembl p12, and Gencode v19 databases as well as newer versions.

Vignettes with detailed instructions and examples are available through the Bioconductor pages for `derfinder` and `derfinderPlot`. The main functions for the expressed region and single base-level approaches are further described in Supplementary Section 2.11.1.

2.3.2 Expressed region level analysis

In the expressed region approach, we compute the mean coverage for all base pairs from all the samples and filter out those below a user specified cutoff. Contiguous bases passing this filtering step are then considered a candidate region (Figure 2.2A). Then for each sample, we sum the base-level coverage for each such region in order to create an expression matrix with one row per region and one column per sample. This matrix can then be used with feature-level RNA-seq differential expression analysis tools.

2.3.3 Annotation and "Genomic State" Objects

We have implemented a "genomic state" framework to efficiently annotate and summarize resulting regions, which assigns each base in the genome to exactly one state: exonic, intronic, or intergenic, based on any existing or user-defined annotation (e.g. UCSC, Ensembl, Gencode). At each base, we prioritize exon $>$ intron $>$ unannotated across all annotated transcripts.

Overlapping exons of different lengths belonging to different transcripts are reduced into a single "exonic" region, while retaining merged transcript annotations. We have a second implementation that further defines promoters and divides exonic regions into coding and untranslated regions (UTRs)

which may be useful for the user to more specifically annotate regions - this implementation prioritizes coding exon > UTR > promoter > intron > unannotated.

2.3.4 Data Processing for Results in Main Manuscript

2.3.4.1 BrainSpan data

BigWig files for all 487 samples across 16 brain regions were downloaded from the *BrainSpan* website [21]. The samples for *HSB169.A1C*, *HSB168.V1C* and *HSB168.DFC* were dropped due to quality issues. Based on exploratory analyses the coverage was assumed to be reads-per-million mapped reads in this data set. We set the coverage filter to 0.25 for both the single base-level and ER-level `derfinder` approaches. Since the coverage is already adjusted to reads per million mapped reads we did not include a library size adjustment term in the single base-level `derfinder` analysis (see Supplementary Section 2.12.1 for details on this adjustment term). The details for the single base-level `derfinder` analysis are described further in Supplementary Section 2.12.2. For the ER-level approach we only considered regions longer than 5 base-pairs.

We sought to identify differences in expression across brain region (neocortical regions: DFC, VFC, MFC, OFC, M1C, S1C, IPC, A1C, STC, ITC, V1C and non-neocortical regions: HIP, AMY, STR, MD, and CBC) and developmental stage (fetal versus postnatal). We therefore fit the following region-by-stage interaction alternative model, which included main effects for fetal versus postnatal (binary) and categorical brain region variable (15 region indicators,

relative to A1C), and interaction terms for each brain region and developmental stage. This resulted in a total of 32 terms in the model (intercept; 16 main effects, 15 interaction terms). In equation (2.1), y_{ij} is the scaled $\log_2$ coverage for the expressed region i and sample j. That is, $y_{ij} = \log_2\left(\text{mean coverage}_{ij} + 1\right)$. The model is completed by an intercept term α_i, a indicator variable for fetal status β_i, m indicators variables γ for the brain region, and m interaction variables ζ between fetal status and brain region. The term ϵ_{ij} represents residual error.

$$y_{ij} = \alpha_i + \beta_i Fetal_j + \sum_{q=1}^{m} \gamma_{iq} Region_{jq} + \sum_{q=1}^{m} \zeta_{iq} Fetal_j * Region_{jq} + \epsilon_{ij} \qquad (2.1)$$

We compared the above model to an intercept-only model using the `lmFit` function from `limma` [16, 15]. The p-values for the ER-level DERs were adjusted via the Bonferroni method and those with adjusted p-values less than 0.05 were determined to be significant. We then calculated the mean coverage for each significant expressed region DERs in each sample, resulting in a mean coverage matrix (DERs by samples), and we performed principal component analysis (PCA) on this $\log_2$-transformed matrix (after adding an offset of 1).

Once the DERs were identified, we identified which of them overlap ENCODE blacklisted regions of the genome [4] using the file at `hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz`. For identifying which DERs overlap lincRNAs we used `EnsDb.Hsapiens.v75` [24], which can also be used for a variety of transcript types. We then performed the gene ontology

analysis for the DERs using `GOstats` [25] using as background all genes that are within 5 kb of an ER.

2.3.4.2 GTEx data

We selected samples from individuals that had data from heart (left ventricle), liver and testis tissues with RIN values greater than 7. 8 subjects matched this criteria and we selected only 1 sample if their tissue was analyzed more than once, leaving us with 24 samples. The data was aligned using `Rail-RNA` [26] version 0.2.1 with the code as described at `github.com/nellore/runs`. We created a normalized mean BigWig file for these 4 samples adjusted for library sizes of 40 million reads. We then identified the ERs using a cutoff of 5 using the function `railMatrix` from `derfinder` version 1.5.19.

For each expressed region greater than 9bp, we assigned its annotation status by using a genomic state object created with the Ensembl GRCh38.p5 database. We then performed principal component analysis (PCA) on the $\log_2$-transformed matrix (after adding an offset of 1) separately for strictly exonic and strictly intronic ERs. Using `limma` [16, 15] functions `lmFit`, `ebayes` we fit an intercept-only null model and an alternative model with coefficients for tissue differences. For each ER we calculated a F-statistic and determined whether it was differentially expressed by tissue using a Bonferroni adjusted p-value cutoff of 0.05.

For the conditional expression analysis, we found the nearest exonic ER for each intronic ER using the `distanceToNearest` function from `GenomicRanges`

[27]. For each intronic ER we fitted two linear regression models for the $\log_2$-transformed coverage matrix (after adding an offset of 1). For the alternative model we used as covariates two tissue indicator variables (Heart as the reference) and the coverage from the nearest stricly exonic ER as shown in Equation (2.2) for ER i and sample j. For the null model we only used the coverage from the nearest exonic ER. We calculated an F-statistic using the `anova` function that tests whether β_{1i} or β_{2i} are equal to 0 and used a Bonferroni adjusted p-value cutoff of 0.05 to identify which intronic ERs had differential expression adjusting for the coverage at the nearest exonic ER.

$$y_{ij} = \alpha_i + \beta_{1i} Testis_j + \beta_{2i} Liver_j + \gamma_i ExonicCoverage_j + \epsilon_{ij} \tag{2.2}$$

2.3.4.3 Simulated data

We simulated 100 bp paired-end reads (250bp fragments, sd = 25) with `polyester` [28] for two groups with five samples per group from human chromosome 17 with uniform error rate of 0.005 and replicated this process three times. One sixth of the transcripts were set to have higher expression (2x) in group 2, a sixth to have lower expression in group 2 (1/2x) and the remaining two thirds to be equally expressed in both groups. Given a RNA-seq experiment with 40 million paired-end reads, assuming that all transcripts are equally expressed we would expect 1,989,247 of them to be from chromosome 17 based on the length of all exons using the known transcripts UCSC knownGene hg19 annotation. We used this information and the transcript length to assign the number of reads per transcript in chromosome

17 and generated the number of reads with the `NB` function from `polyester` with mean μ and size (see `stats::rnbinom` function in R) equal to $\frac{1}{3}\mu$. This resulted in an average of 2,073,682 paired-end reads per sample. For each simulation replicate, paired-end reads were aligned to to the hg19 reference genome using `HISAT` version 0.1.6-beta [29] and `Rail-RNA` version 0.2.2b [26]. We created a GTF file using all known transcripts from chromosome 17 as well as one with 20% of the transcripts missing (8.28% of exons missing). Using these two GTF files we performed transcript quantification with `StringTie` version 1.2.1 [23] as well as exon counting allowing multiple overlaps with the `featureCounts` function from `Rsubread` version 1.21.4 [13]. ERs were determined with `derfinder` version 1.5.19 functions `regionMatrix` and `railMatrix` respectively from the `HISAT` BAM and `Rail-RNA` BigWig output using a mean cutoff of 5 for libraries adjusted to 80 million single-end reads. Count matrices resulting from `featureCounts` and `derfinder` were analyzed with *limma* [16], *DESeq2* [14] and *edgeR*-robust [18] controlling the FDR at 5% and testing for differences between the two groups of samples. We used `ballgown` version 2.2.0 [22] to perform differential expression tests using coverage at the transcript and exon levels, controlling the FDR at 5%.

The 3900 transcripts from chromosome 17 are composed in total by 39,338 exons (15,033 unique). To avoid ambiguous truth assignments, we used only the 3,868 that overlap only 1 transcript and assigned the truth status based on whether that transcript was set to have a high or low expression on group 2 for the replication replicate under evaluation. We assessed the different pipelines by checking if these 3,868 exons overlapped at least one differentially

expressed unit: exons (`featureCounts` and `ballgown`), transcripts (`ballgown`), and ERs (`derfinder`) respectively. We then calculated the empirical power, false discovery rate and false positive rate.

2.4 Results

2.4.1 Overview of the `derfinder` package

The `derfinder` package includes functions for several stages in the analysis of data from an RNA-sequencing experiment (Figure 4.1).

First, `derfinder` includes functions for pre-processing coverage data from BAM files or bigWig coverage files. The base-level coverage data for multiple samples can be loaded and filtered since most bases will show zero or very low coverage across most samples. Then, the software allows for definition of contiguous regions that show average coverage levels above a certain threshold. These expressed regions are non-overlapping subsets of the genome that can then be counted to arrive at a matrix with an expression value for each region in each sample. Alternatively, the software provides options for counting exons or genes for use in more standard analysis pipelines.

Next, `derfinder` can be used to perform statistical tests on the region level expression matrix. These tests can be carried out using any standard package for differential expression of RNA-seq data including `edgeR` [10, 12], `DESeq` [11], `DESeq2` [14], or `limma`-voom [17].

`derfinder` can then be used to annotate the differentially expressed regions (DERs). We have developed functions that label each region according to

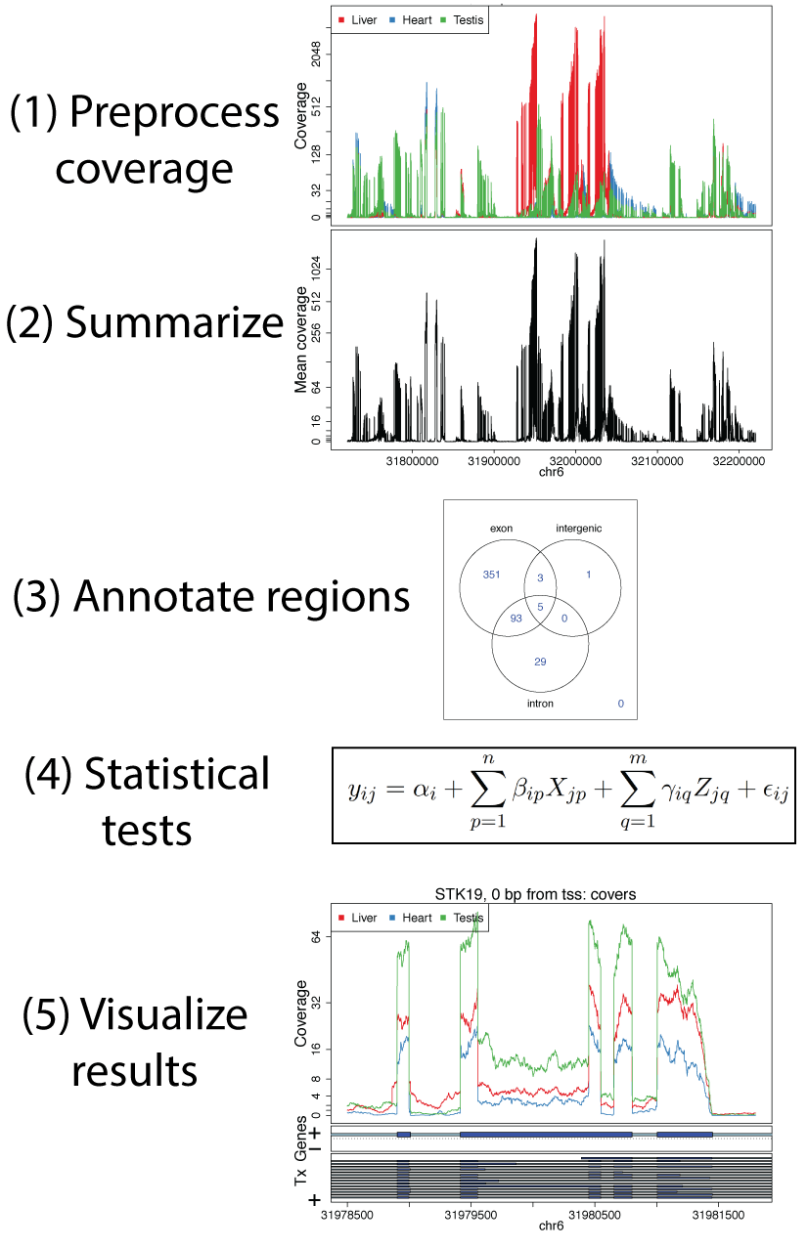


Figure 2.1: An overview of the `derfinder` suite The `derfinder` software package includes functions for processing and normalizing coverage per sample, performing statistical tests to identify differentially expressed regions, labeling those regions with known annotation, and visualizing the results across groups.

whether it falls entirely in a previously annotated protein coding exon (exonic), entirely inside a previously annotated intronic region (intronic), or outside of any previously annotated gene (intragenic). The software also will report any region that overlaps any combination of those types of regions.

Finally, data from an expressed region analysis can be visualized using different visualization approaches. While region-level summaries can be plotted versus known phenotypes, `derfinder` also provides functions to plot base resolution coverage tracks for multiple samples, labeled with color according to phenotype.

We now provide more detail on each of these steps.

2.4.2 Finding expressed regions

The first step in a `derfinder` analysis is to identify expressed regions. Reads should be aligned using any splicing aware alignment tool such as `TopHat2` [30], `HISAT` [29] or `Rail-RNA` [26].

Base resolution coverage information can be read directly from the BAM files that are produced by most alignment software [30, 29, 26]. This process can be parallelized across multiple cores to reduce computational time. An alternative is to read bigWig [31] coverage files. Recent alignment software such as `Rail-RNA` [26] produces these files directly, or they can be created using samtools [32] or produced using the `derfinder` package. Reading bigWig files can produce significant computational and memory advantages over reading from BAM files.

The coverage information represents the number of reads that covers each

genomic base in each sample. `derfinder` first filters out bases that show low levels of expression across all samples. Since most genomic bases are not expressed, this filtering step can reduce the number of bases that must be analyzed by up to 90%, reducing both CPU and memory usage. We originally proposed performing a statistical test for every base in the genome [19] and this approach is still supported by the `derfinder` package for backwards compatibility (Supplementary Section 2.11.3).

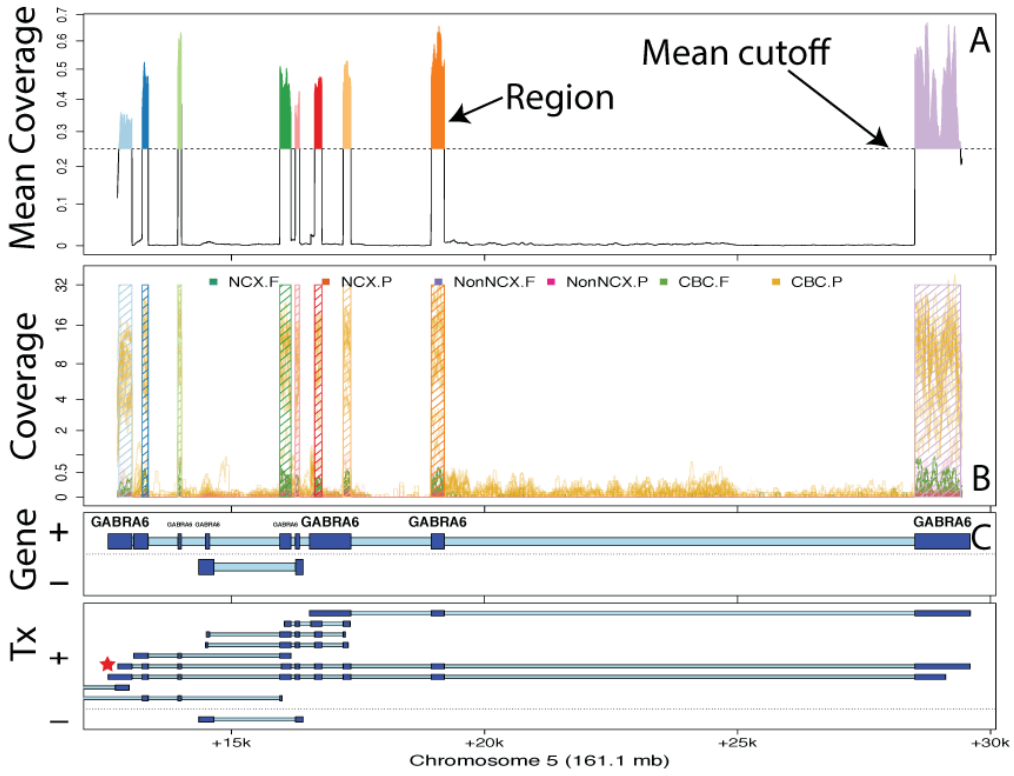


Figure 2.2: Finding regions via expressed region-level approach on chromosome 5 with *BrainSpan* data set. A Mean coverage with segments passing the mean cutoff (0.25) marked as regions. **B** Raw coverage curves superimposed with the candidate regions. Coverage curves are colored by brain region and developmental stage (NCX: Neocortex: Non-NCX: Non-neocortex, CBC: cerebellum, F: fetal, P: postnatal). **C** Known exons (dark blue) and introns (light blue) by strand for genes and subsequent transcripts in the locus. The DERs best support the *GABRA6* transcript with a red star, indicating the presence of a differentially expressed transcript.

Here we focus on a new approach based on the bump-hunting methodology for region level genomic analysis [33] (Figure 2.2). This approach first calculates expressed regions (ERs) across the set of observed samples. For each base, the average, potentially library size-adjusted, coverage is calculated across all samples in the data set. This generates a vector of (normalized) mean level expression measurements across the genome. Then an average-coverage cutoff is applied to this mean coverage vector to identify bases that show minimum levels of expression. An expressed region is any contiguous set of bases that has expression above the mean expression cutoff.

The next step is to count the number of reads (including fractions of reads) that overlap each expressed region. As we have pointed out previously [19] that counting expression in genes and exons is complicated by overlapping annotation. Expressed regions are non-overlapping, so this means that each read can be unambiguously assigned to the appropriate region.

2.4.3 Expressed region level statistical tests

The result of the expressed region (ER) step is a coverage matrix with each row corresponding to one ER and each column corresponding to one sample. This count matrix can then be analyzed using statistical models that have been developed for gene or exon counts such as `limma` [16, 15], `voom` [17], `edgeR-robust` [18], and `DESeq2` [14]. We emphasize that unlike other feature-level counting approaches, our approach is annotation-agnostic: ERs are defined empirically using the observed sample data and coverage threshold. So if there is sufficient expression in a region outside of previously annotated genes

it will be quantified and analyzed with our approach.

2.4.4 Visualizing differentially expressed regions

After statistical modeling, `derfinder` produces a set of DERs with summary statistics per region. They are stored as a `GRanges` object [27] and can be visualized with a range of packages from the Bioconductor suite. We have also developed several visualization tools specific to the `derfinder` approach.

These plots can be made at different levels of summarization. First, the `derfinder` and `derfinderPlot` packages provide a range of visualizations of coverage tracks at single base resolution. These plots can be used to identify coverage patterns that may diverge from annotated protein-coding regions. For example, using the GTEx example we can visualize genes that have consistently high intronic expression as shown in Figure 2.3. We show several examples of genes known to be functionally important in heart - *LBD3* and *MYOZ2* (Figure 2.3A,B) [34, 35], and liver - *HGD* and *UPB1* (Figure 2.3C,D) [36, 37]. The coverage profiles can provide additional insight into transcription, and well as potential technical artifacts, beyond the level of annotated genes, exons and transcripts, which we include in our base-resolution plots.

DERs can be grouped into larger regions by distance, which can be useful to identify potentially systematic artifacts such as coverage dips (Figure 2.4), perhaps due to sequence composition. Visualizing the base-level coverage for a set of nearby candidate DERs can reveal patterns that explain why one DER is sometimes fragmented into two or more shorter DERs. Coverage dips (Figure 2.4), spikes and data quality in general can affect the borders of the

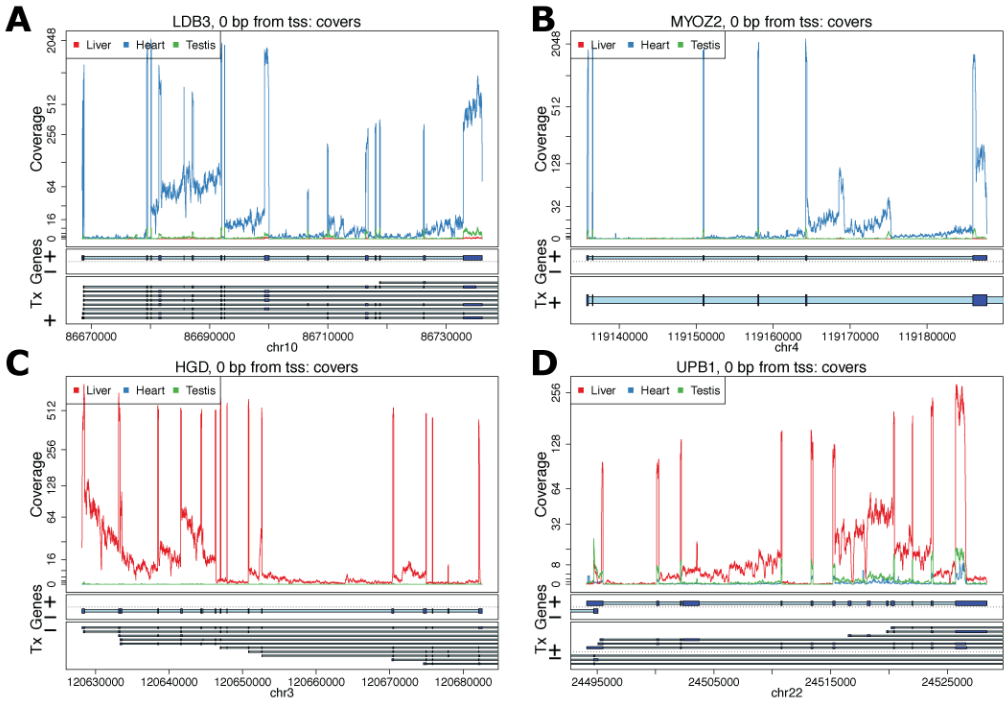


Figure 2.3: Coverage plots for the average coverage levels for the GTEx example. Average coverage profile for heart (blue), liver (red), and testis (green) from the GTEx example near genes: **A** *LDB3*, **B** *MYOZ2*, **C** *HGD*, and **D** *UPB1*.

candidate DERs. Some artifacts can be discarded, like candidate DERs inside repetitive regions. Base-pairs inside repetitive regions available in repeat masker tracks can be flagged and filtered out from the analysis. Other known potentially problematic regions of the genome, like those with extreme GC content or mappability issues can also be filtered out, either before identifying candidate DERs or post-hoc.

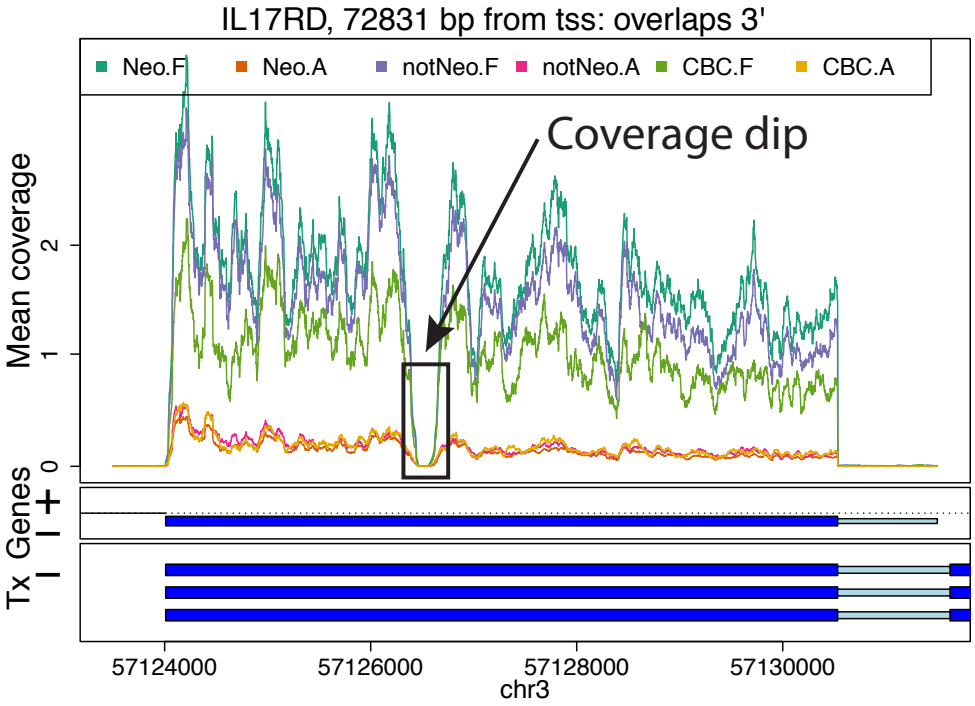


Figure 2.4: Example of a coverage dip. Mean coverage per group for the *BrainSpan* data set for a region that results in two DERs for a single exon due to a coverage dip. The genome segment shown corresponds to the DERs cluster ranked 15th in terms of overall signal by the single base-level approach applied to the *BrainSpan* data set.

2.4.5 Annotating differentially expressed regions

The DERs can be annotated to their nearest gene or known feature using `bumphunter` [33]. The basic approach is to overlap DERs genomic coordinates with the genomic coordinates of known genomic features. By default, `derfinder` labels each identified region as exonic, intronic, intragenic or some combination of those three labels.

A region may overlap multiple genomic features (say an exon and the adjacent intron). Using this information candidate DERs can further be compared to known gene annotation tables (Methods Section 2.3.3) to identify

potentially novel transcription events. Using this information, visualizations of specific loci for overlap with annotation can be made with `derfinderPlot`. The regions can be exported to CSV files or other file formats for followup and downstream analyses. We have also developed a complementary R package for creating reproducible reports incorporating the annotation and visualization steps of the `derfinder` pipeline called `regionReport` [38].

2.4.6 Application: large-scale expression analysis at base resolution

We used `derfinder` to detect regions that were differentially expressed across the lifespan in the human brain. We applied `derfinder` to the *BrainSpan* RNA-seq coverage data (Methods Section 2.3.4.1), a publicly available data set consisting of 484 postmortem samples across 16 brain regions from 40 unique individuals that collectively span the full course of human brain development [21]. We used the expressed region approach described above for this analysis. For comparison we applied the single-based resolution approach previously utilized on independent dorsolateral prefrontal cortex RNA-seq data [20] (Supplementary Section 2.11.4).

We identified 174,610 ERs across the 484 samples with mean across-sample normalized coverage > 0.25, which constituted 34.57 megabases of expressed sequence. The majority (81.7%) of these ERs were labeled as strictly exonic while only a small subset (5.4%) were strictly non-exonic by Ensembl annotation. These ERs largely distinguished the fetal and postnatal samples using PCA - the first principal component explained 40.6% of the variance of the mean coverage levels and separated these developmental stages across all

brain regions. This separation was consistent regardless of the annotation status of the DERs including in the strictly intronic regions (Figure 2.5 and Supplementary Figure 2.9). The separation between brain regions in intronic regions may be due to noisy or incorrect splicing [39] or may be due to missing annotation [19] or mistaken sequencing of pre-mRNA. The base resolution visualizations available as part of `derfinder` and `derfinderPlot` make it possible to explore to determine if it is biology or artifacts driving these expression differences.

The PCA plots also appear to show patterns consistent with potential artifacts such as batch effects [40] (Figure 2.5). Regardless, the new ER approach we present here provides options for analysts who wish to discover patterns of expression outside of known annotation on hundreds of samples - an analysis of this scope and scale was unfeasible with earlier versions of our single base resolution software [19].

Using statistical models where expression levels were associated with developmental stage (fetal versus postnatal) and/or brain region (Methods Section 2.3.4.1), we found that 129,278 ERs (74%) were differentially expressed by brain region and/or developmental stage at the ER-level controlling the family-wise error rate (FWER) at $< 5\%$ via Bonferroni correction. We controlled the FWER instead of the FDR due to the expected large effects between the developmental stages and/or brain regions. The 129,278 ER-level DERs overlapped a total of 17,525 Ensembl genes (13,016 with gene symbols), representing a large portion of the known transcriptome. Of the significant ER-level DERs, 93,355 (72.2%) overlapped at least 1 significant single base-level DER

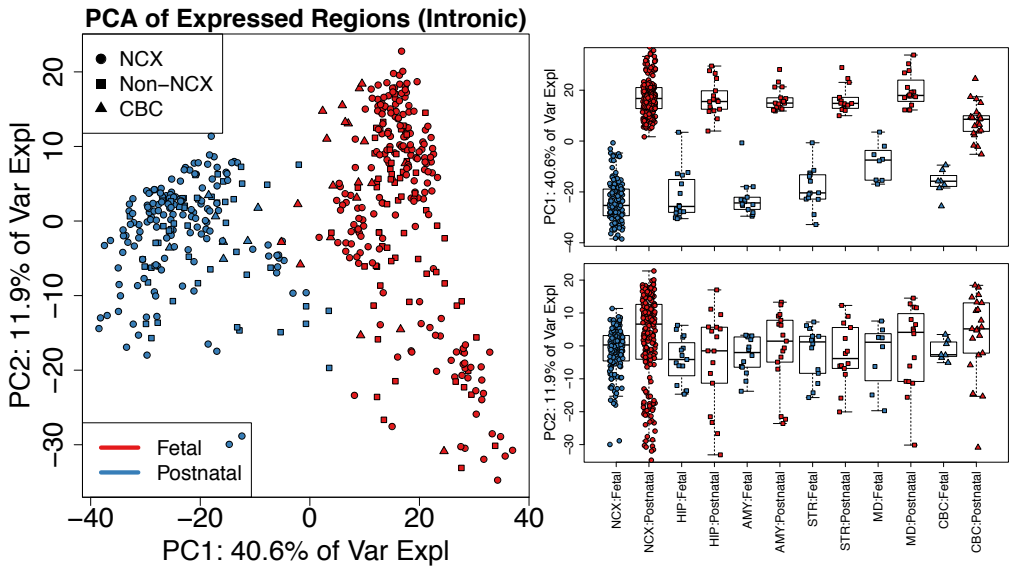


Figure 2.5: Principal components analysis reveals clusters of samples in the BrainSpan data set. (Left) First two principal components (PCs) with samples colored by sample type (F: Fetal or P: Postnatal) and shape given by brain region using only the strictly intronic expressed regions (ERs). Analysis of other subsets of ERs produce similar results (Supplementary Figure 2.9). (Right) Boxplots for PCs 1 and 2 by brain region (NCX: neocortex, HIP: hippocampus, AMY: amygdala, STR: striatum, MD: thalamus, CBC: cerebellum) and sample type with non-neocortex brain decomposed into its specific regions. Using the single base-level approach (Supplementary Figure 2.10) produces similar results as shown in Supplementary Figure 2.11.

(Supplementary Section 2.11.4). Lack of overlap results from almost half (45.2%) of single base-level DERs having an average coverage lower than the expression cutoff determining ERs (0.25). For example, there was high expression only in the samples from a few brain regions, or only one development period. Decreasing the cutoff that defines the ERs from 0.25 to 0.1 results in a larger number of regions (217,085) that have a higher proportion of non-exonic sequence (12.1%), suggesting that the choice of this expression cutoff requires some initial exploratory data analysis as shown in Supplementary Section 2.11.5.

We highlight the utility of the ER-level analysis (using the original 0.25 cutoff) to identify regions differentially expressed within subsets of the data by analyzing brain regions within a single developmental period. We identified 1,170 ERs that were differentially expressed comparing striatum versus hippocampus samples in the fetal developmental stage. These DERs mapped to 293 unique genes. Genes more highly expressed in the striatum include *ARPP-21*, previously shown to localize in the basal ganglia [41], and dopamine receptor genes *DRD1* and *DRD2* [42]. Genes more highly expressed in the hippocampus in fetal life were strongly enriched for neurodevelopmental genes including *FZD7* [43], *ZBTB18* [44], and *NEUROD1* [45]. The ER-level analysis therefore permits subgroup analysis without the need to rerun the full `derfinder` single base-level pipeline - another improvement over previous versions of single base resolution analysis software [19].

DERs are non-standard in the sense that they don't necessarily match with known exons. Depending on the application, you might interested in filtering out DERs that overlap problematic regions of the genome. This can be done prior to defining the ERs or once the candidate DERs have been identified. In the *BrainSpan* application, only 0.086% of the 129,278 DERs overlap ENCODE blacklisted regions [4] and 1.58% overlap lincRNAs. Similarly one can check if the DERs overlap other known features of interest. The genes overlapped by the DERs are enriched for gene ontology terms such as *neuron differentiation* (GO:0030182, p-value 4.13e-15), *neurogenesis* (GO:0022008, p-value 4.62e-14) and *neuron projection development* (GO:0031175, p-value 1.4e-12) among other terms associated to neuronal development.

2.4.7 Identification of expressed regions that differentiate tissues using a subset of the GTEx data

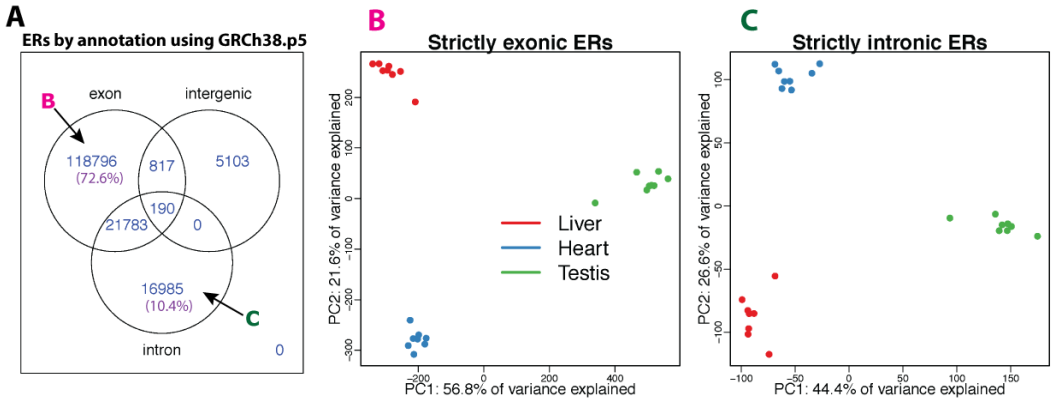


Figure 2.6: GTEx expressed regions analysis using 24 samples from the heart (left ventricle), liver and testis for 8 subjects. A expressed regions (longer than 9 bp) overlapping known annotation based on GRGh38.p5 (hg38). 72.6% of the ERs only overlap known exons (strictly exonic) while 10.4% only overlap known introns (strictly intronic). **B** First two principal components (PCs) with samples colored by sample type (red: liver, blue: heart, green: testis) using only the strictly exonic ERs. **C** First two PCs with samples colored by sample type using only only the strictly intronic ERs. The sign change of the second principal component is simply a rotation and the results are consistent between the strictly exonic and strictly intronic ERs.

We selected a subset of subjects from the GTEx project [3] that had RNA-seq data from heart (left ventricle), liver and testis, specifically the eight subjects with samples that had RNA Integrity Numbers (RINs) greater than 7, given RIN's impact on transcript quantification [46]. Using only one sequencing library from each subject aligned with `Rail-RNA` [26], we applied the ER-level `derfinder` approach with a cutoff of 5 normalized reads (after normalizing coverage to libraries of 40 million reads). We found a total of 163,674 ERs with lengths greater than 9 base-pairs. Figure 2.6A shows that 118,795 (72.6%) of the ERs only overlapped known exonic regions of the genome using the Ensembl GRCh38.p5 database [47].

we performed PCA on the $\log_2$ adjusted coverage matrix using just the 118,795 strictly exonic ERs (Figure 2.6B). Here the first two PCs explain 56.8% and 21.6% of the variance respectively and show three distinct clusters of samples that correspond to the tissue of the sample. We found that the 16,985 (10.4%) ERs (Figure 2.6A) that only overlap annotated introns can also differentiate tissues using PCA, as shown in Figure 2.6C. The total percent of variance explained by the first two principal components is slightly lower (44.4% + 26.6% = 71% versus 56.8% + 21.6% = 78.4%) when using only the strictly intronic ERs versus the strictly exonic ERs. This may represent a different biological signal and/or potentially noisy splicing (as in Figure 2.3B). but we use this example to illustrate the potential to use `derfinder` to explore regions outside of known annotation.

Using `limma` [16, 15] to test for differential expression between tissues (Methods Section 2.3.4.2) we found that 42,880 (36.1%) of the strictly exonic ERs and 4,401 (25.9%) of the strictly intronic ERs were differentially expressed (FWER of 5% via Bonferroni correction). Overall 59,776 (36.5%) of the ERs were differentially expressed between tissues. Given the similar global patterns of expression between annotated and unannotated ERs, we considered the scenario that the strictly intronic ERs were differentially expressed between tissues in the same pattern as the nearest exonic ERs due to possible run-off transcription events. To assess this scenario we fitted a conditional regression for each strictly intronic ER adjusting for the coverage of the nearest strictly exonic ER. 749 (4.4%) of the strictly intronic ERs differentiate tissues while adjusting for the coverage at the nearest exonic ER at a FWER of 5%. Figure

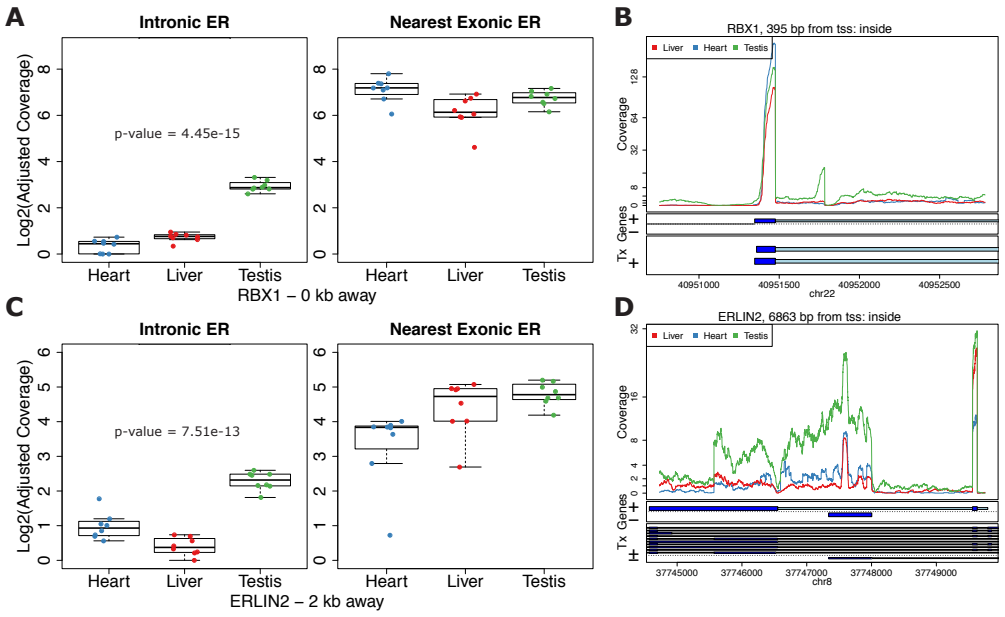


Figure 2.7: Differential expression on strictly intronic expressed regions adjusting for expression on the nearest strictly exonic ER. Boxplots (**A** and **C**) and region coverage plots (**B** and **D**) for two strictly intronic ERs showing differential expression signal adjusting for the nearest exonic ER. Boxplots show the $\log_2$ adjusted coverage for the strictly intronic ERs by tissue with the corresponding boxplot for the nearest strictly exonic ERs. The p-value shown is for the differential expression between tissues on the intronic ERs conditional on the expression values for the nearest exonic ERs. The distance to the nearest strictly exonic ER and the gene symbol are shown below. The region coverage plots are centered at the strictly intronic ER with the neighboring 2kb and 5kb for **C** and **D** respectively. **A,B** Expression on the exonic ER is fairly similar between the groups but different on the intronic ER. **C,D** Expression on the exonic ER has an increasing pattern from heart to liver to testis but has a different pattern on the intronic ER.

2.7A,B shows an example where the expression is similar between tissues in the nearest exonic ER but there is a clear tissue difference in the intronic ER with testis having higher expression than the other two tissues. Figure 2.7C,D shows different patterns between the intronic and exonic ERs where in the exonic ER the expression is lowest in the heart, higher in liver and slightly higher at the testis. However in the intronic ER, liver is the tissue that has the lowest

expression. These results suggest that expression at unannotated sequence could have biological relevance beyond local annotated exonic sequence.

2.4.8 Simulation results

We lastly performed a simulation study to evaluate the statistical properties of `derfinder` with and without complete annotation. To compare `derfinder` against feature-level alternatives, we simulated reads for 2 groups, 10 samples in total (5 per group) with $\frac{1}{6}$ of the transcripts having higher and $\frac{1}{6}$ lower expression in group 2 versus group 1 at fold changes of 2x and $\frac{1}{2}$x respectively. Reads were simulated from chromosome 17 using `polyester` [28] with the total number of reads matching the expected number given paired-end library with 40 million reads (Methods Section 2.3.4.3). We used `HISAT` [29] to align the simulated reads and summarized them using either `featureCounts` from the `Rsubread` package [13] or `StringTie` [23] and performed the statistical tests on the resulting coverage matrices using `limma` and `ballgown` [22] respectively. We performed the `ballgown` statistical test at the exon-level as well as the transcript-level. We performed the feature-level analyses using the complete annotation and with an annotation set missing 20% randomly selected transcripts (8.28% unique exons missing). We then used `derfinder` to find the ERs from the same `HISAT` alignments as well as from `Rail-RNA` [26] output and performed the statistical test with `limma`. For all statistical tests we controlled the FDR at 5% and we repeated the simulation three times.

Table 2.1 shows the range of the empirical power, false positive rate (FPR)

Table 2.1: Minimum and maximum empirical power, false positive rate (FPR) and false discovery rate (FDR) observed from the three simulation replicates for each analysis pipeline. `ballgown` analyses were done at either the exon or transcript levels. Pipelines that rely on annotation were run with the full annotation or with 20% of the transcripts missing (8.28% exons missing). Count matrices were analyzed with `limma`, `DESeq2` and `edgeR-robust` (Supplementary Table 2.3). FDR of 5% was targeted.

Power	FPR	FDR	Annotation complete	Aligner	Summary method	Statistical method
(93.6-94.2)	(6.4-9.3)	(12.8-16.5)		HISAT	derfinder	limma
(93.7-94.2)	(6.5-9.1)	(12.5-16.1)		Rail-RNA	derfinder	limma
(69-77.6)	(2.5-3.3)	(6-7.7)	No	HISAT	featureCounts	limma
(94.4-95.1)	(3.1-4.5)	(6.5-7.5)	Yes	HISAT	featureCounts	limma
(68.4-77)	(2.8-3)	(5.5-8.3)	No	HISAT	StringTie	ballgown-exon
(93.7-94.6)	(3.6-4)	(5.9-7.8)	Yes	HISAT	StringTie	ballgown-exon
(53.2-60)	(0.6-2.2)	(1.4-8.1)	No	HISAT	StringTie	ballgown-trans
(67.2-71.9)	(0.6-1.1)	(1.4-3.2)	Yes	HISAT	StringTie	ballgown-trans

and false discovery rate (FDR) for all these methods based on the three simulation replicates. `derfinder`'s expressed region approach resulted in overlapping empirical power ranges to the exon-level methods that are supplied the complete annotation. The exon-level methods had a 18% to 27% loss in power when using the incomplete annotation set compared to the complete set even though only 8.28% of the unique exons were missing. `derfinder`, being annotation-agnostic, does not rely on having the complete annotation but did show increased FPR and FDR compared to the exon-level methods. We recommend performing sensitivity analyses of the cutoff parameter used for defining ERs or the FDR control in the statistical method used to determine which ERs are differentially expressed (i.e. DERs). Transcript-level analyses had the lowest FPR and FDR but also the lowest power. Note that we only performed transcript expression quantification with `StringTie` and did not use the data to determine new transcripts. Doing so resulted in a much larger transcript set than originally present in the data: 3,900 in the original set versus 15,920 (average for the three replicates using the complete annotation).

Supplementary Section 2.11.6.1 shows the results when using `DEseq2` or `edgeR`-robust for performing the statistical tests. Figure 2.8 shows the mean empirical power against the observed FDR for the different combinations of methods when controlling the FDR at 1%, 5%, 10%, 15% and 20%. Results with `derfinder` are among the set with the highest empirical power, at the cost of a higher observed FDR than what was controlled for.

Identifying ERs uses computational resources and runs in similar time to summarization steps required for the exon-level pipelines used in this

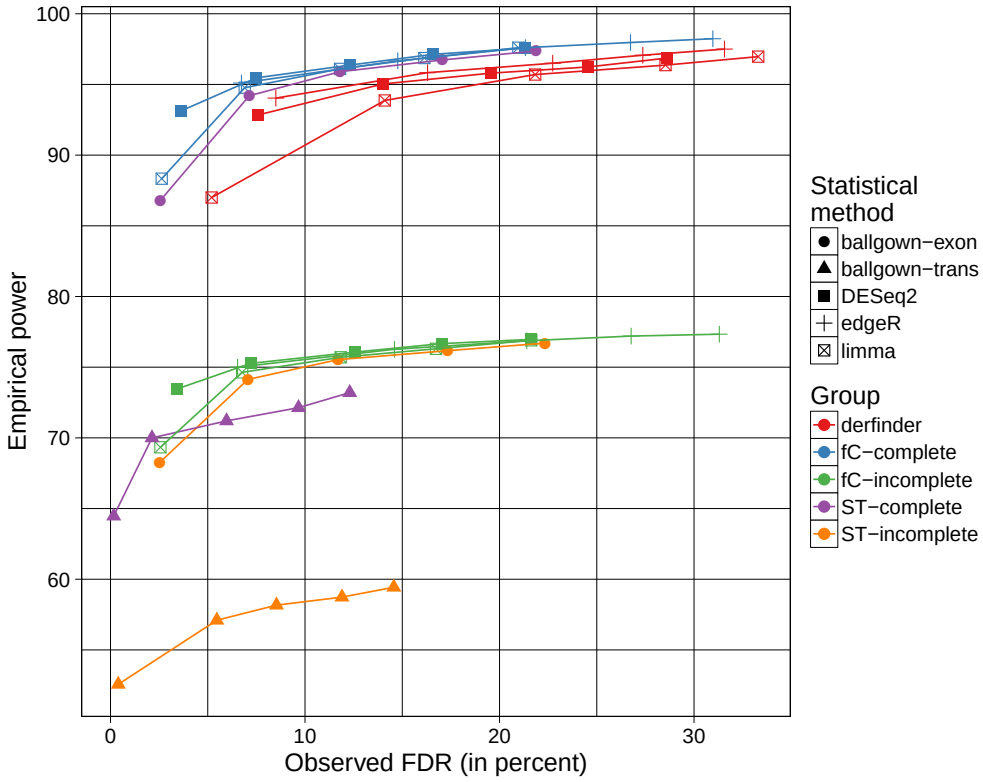


Figure 2.8: Mean empirical power versus observed False Discovery Rate (FDR) across the 3 simulation replicates for a combination of statistical and summary methods. For FDR cutoffs of 1, 5, 10, 15 and 20% the mean empirical power and FDR across the 3 simulation replicates is displayed for the combination of statistical method (`ballgown` at exon or transcript level, `limma`, `DESeq2`, `edgeR`-robust) the summary method (`derfinder`, `featureCounts` (fC), `StringTie` (sT)) and whether the annotation used was complete or not (complete, incomplete).

simulation (Supplementary Section 2.11.6.2) and is the fastest when using BigWig files such as those produced by `Rail-RNA`. These results suggest that the `derfinder` approach performs well when differentially expressed features overlap known annotation and appear in unannotated regions of the genome. If you are only interested in studying known regions, other methods have better FDR control than `derfinder` as shown in Figure 2.8.

2.5 Discussion

Here we introduced the `derfinder` statistical software for performing genome-scale annotation-agnostic RNA-seq differential expression analysis. This approach utilizes coverage-level information to identify differentially expression regions (DERs) at the expressed region or single base-levels, and then generates useful summary statistics, visualizations and reports to further inspect and validate candidate regions.

The reduced dependence on the transcriptome annotation permits the discovery of novel regulated transcriptional activity, such as the expression of intronic or intergenic sequences, which we highlight in publicly available RNA-seq data and our previous `derfinder` application [20]. As shown with a subset of GTEx, strictly intronic ERs can differentiate tissues when adjusting for the expression from the nearest exonic expressed region, suggesting that some intronic DERs may represent signal beyond run-off transcription. Furthermore, the structure of DERs across a given gene can permit the direct identification of differentially expressed transcripts (e.g. Figure 2.2C), providing useful information for biologists running validation experiments. Lastly, this software and statistical approach may be useful for RNA-seq studies on less well-studies species, where transcript annotation is especially likely to be incomplete.

The software pipeline, starting with BAM or BigWig files, and ending with lists of DERs, reports, and visualizations, runs at comparable speeds to existing RNA-seq analysis software. Given the appropriate computing resources, `derfinder` can scale to analyze studies with several hundred samples. For

such large studies, it will be important to correct for batch effects and potentially expand `derfinder`'s statistical model for base-level covariates. This approach provides a powerful intermediate analysis approach that combines the benefits of feature counting and transcript assembly to identify differential expression without relying on existing gene annotation.

2.6 Competing interests

The authors declare that they have no competing interests.

2.7 Funding

JTL was supported by NIH Grant 1R01GM105705, LCT was supported by Consejo Nacional de Ciencia y Tecnología México 351535, and AEJ was supported by 1R21MH109956.

2.8 Author's contributions

AEJ, JTL, RAI conceived the software. LCT wrote the software under the supervision of JTL and AEJ. LCT analyzed the data with the supervision of JTL and AEJ. AN, CW and BL helped with the GTEx data analysis. All authors contributed to writing the paper.

2.9 Acknowledgments

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI/SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The raw data (sequencing reads and phenotype data) used for the analyses described in this manuscript were obtained from SRA accession number phs000424.v6.p1 on 10/07/2015.

2.10 Additional Files

The `derfinder` vignettes detail how to use the software and its infrastructure. The latest versions are available at `www.bioconductor.org/packages/derfinder`.

The Supplementary Methods and Results describe in more detail the R implementation, the single base-level approach, and the analysis of the *BrainSpan* data set with the single base-level approach. Supplementary file 1 contains the

identified candidate single base-level DERs in CSV format (gzip compressed) for the *BrainSpan* data set.

The code and log files detailing the versions of the software used for all the analyses described in this paper is available at the Supplementary Website: leekgroup.github.io/derSupplement.

Supplementary Methods and Results

This document describes R implementation details of `derfinder`, the single base-level approach, and results from applying the single base-level approach to the *BrainSpan* data set. It also includes the simulation results when performing the statistical tests using `edgeR-robust` [13] or `DESeq2` [14] instead of `limma` [16].

2.11 Supplementary Results

2.11.1 R implementation

The `derfinder` package can be used for different types of analyses such as DER finding (single base-level and ER-level approaches) as well as creating a feature counts matrix. The overall relationship between these functions is shown in section *Flow charts* subsection *DER analysis flow chart* of the *derfinder users guide* vignette available at `www.bioconductor.org/packages/derfinder`.

For the single base-level approach, the main function is `analyzeChr()` which makes it easier for users to run this type of analysis. This function is a wrapper for other functions available in `derfinder`, as can be seen section *Flow charts* subsection *analyzeChr() flow chart* of the *derfinder users guide* vignette. It splits the data, calculates the F-statistics, identifies the null regions, and annotates them.

The expressed regions (ERs) approach is described in section *Flow charts* subsection *regionMatrix() flow chart* of the *derfinder users guide* vignette. This type of analysis requires fewer functions, as the user only needs to load the

data and then identify the ERs with the `regionMatrix()` function. The *regionMatrix() flow chart* shows which other functions are internally used by `regionMatrix()` that filter the coverage by using a mean cutoff, identify the regions, and produce the region-level count matrix. The function `railMatrix()` is optimized for identifying ERs from BigWig files, specially those created with `Rail-RNA` (DOI: 10.1101/019067).

2.11.2 Differential expression in the developing human brain via expressed region-level analysis

Figure 2.9 complements Figure 2.5 with the results of performing principal component analysis of ERs found in the *BrainSpan* data set given the known annotated elements they overlap with. The results are consistent regardless of the type of ERs under study.

2.11.3 Single base-level statistical test

A single base-level resolution analysis in `derfinder` starts with read alignment and coverage calculation as done in the ER-level approach. Next, a standard differential expression analysis is performed at each base by comparing nested null and alternative linear models using an F-statistic. The statistical models may include adjustments for confounders such as library size [48], demographic variables, and batch effects [40].

Once an F-statistic is calculated at each base, we identify differentially expressed regions (DERs) using a "bump hunting" approach [33]. First we

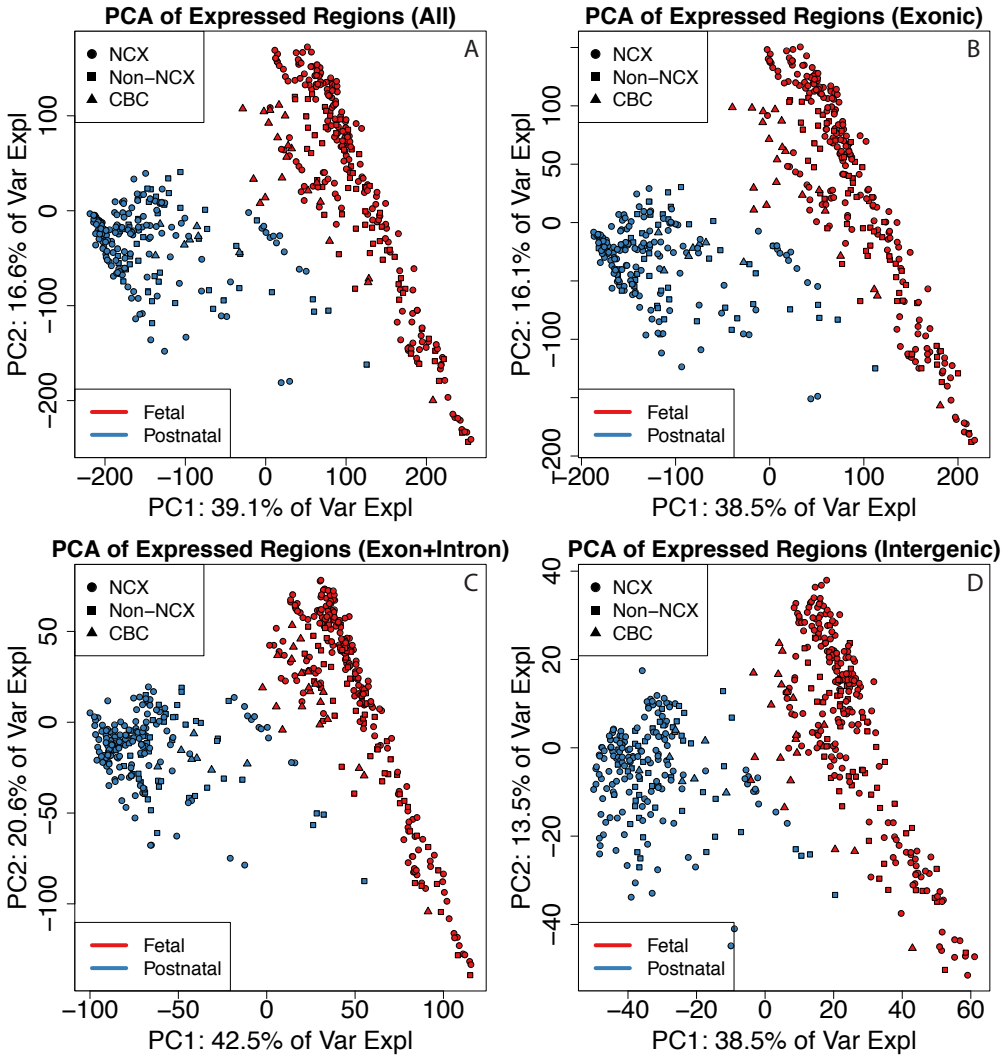


Figure 2.9: Principal components analysis reveals clusters of samples in the BrainSpan data set. First two principal components (PCs) with samples colored by sample type (F: Fetal or P: Postnatal) and shape given by brain region using all ERs (top left), strictly exonic ERs (top right), ERs overlapping exons and introns (bottom left) and strictly intergenic ERs (bottom right).

find candidate DERs by identifying regions of the genome where the base-level F-statistics pass a genome-wide threshold (Figure 2.10 with *BrainSpan* data set, see Supplementary Section 2.12.1). We then calculate a summary statistic for each candidate region based on the length of the region and the

size of the statistics within the region. To evaluate the statistical significance of these candidate regions, we permute the sample labels and recompute candidate regions and summary statistics. The result is a region-level p-value, which can be adjusted to control the family-wise error rate. Alternatively, the region-level p-values can be adjusted for multiple testing using standard false discovery rate techniques [49, 50].

2.11.4 Differential expression in the developing human brain via single base-level analysis

At the single base-level, we identified 113,691 genome-wide significant DERs (FWER $< 5\%$) with the same statistical models used with the ER-level analysis described in the main text. These resulting single base-level DERs largely distinguished the fetal and postnatal samples representing the first principal component and 49.4% of the variance of the mean coverage levels within the DERs (Figure 2.11). The most significant DERs map to genes previously implicated in development, and contained many of the DERs we previously identified in the frontal cortex in 36 independent subjects [20]. For example, 59% of our previously published 50,650 developmental DERs (and 72.6% in the 10,000 most significant) in the frontal cortex overlapped these DERs identified in the *BrainSpan* data set. The potential lack of overlap may be explained by unmodeled artifacts as there appear to be clusters in the principal components calculated on the base resolution data (Figure 2.11, left panel).

While the majority (68.1%) of single base-level DERs overlap exclusively exonic sequence using Ensembl database v75, we find that a fraction (22.2%)

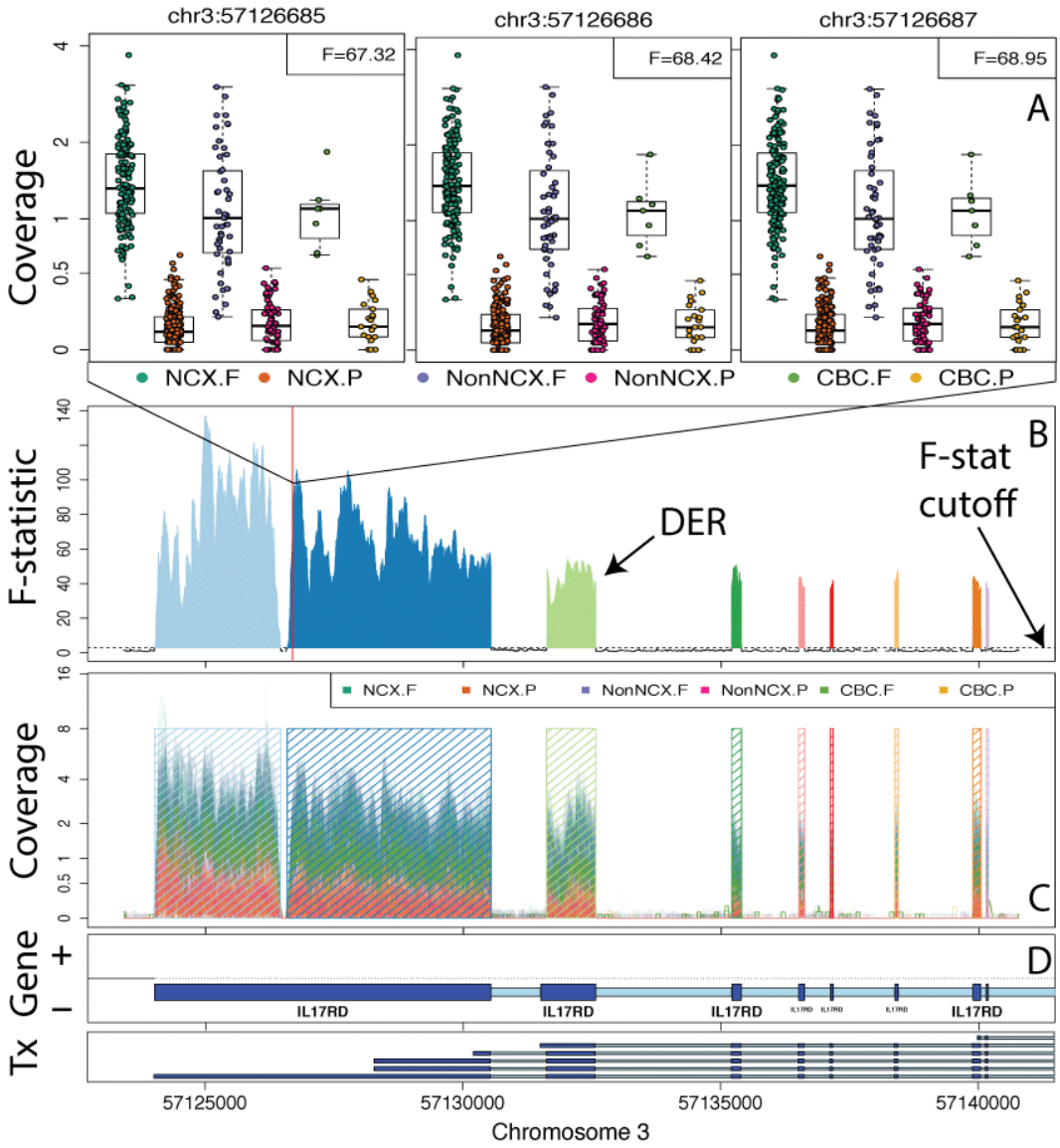


Figure 2.10: Finding DERs on chromosome 3 with *BrainSpan* data set using six groups: Neocortical regions (NCX: DFC, VFC, MFC, OFC, M1C, S1C, IPC, A1C, STC, ITC, V1C), Non-neocortical regions (NonNCX: HIP, AMY, STR, MD), and cerebellum (CBC) split by whether the sample is from a fetal (F) or postnatal (P) subject. **A** Boxplots for three specific bases. **B** F-statistics curve with regions passing the F-stat cutoff marked as candidate DERs. **C** Raw coverage curves superimposed with the candidate DERs. **D** Known exons (dark blue) and introns (light blue) by strand. The third DER matches the shorter version of the second exon shown in the *Tx* track.

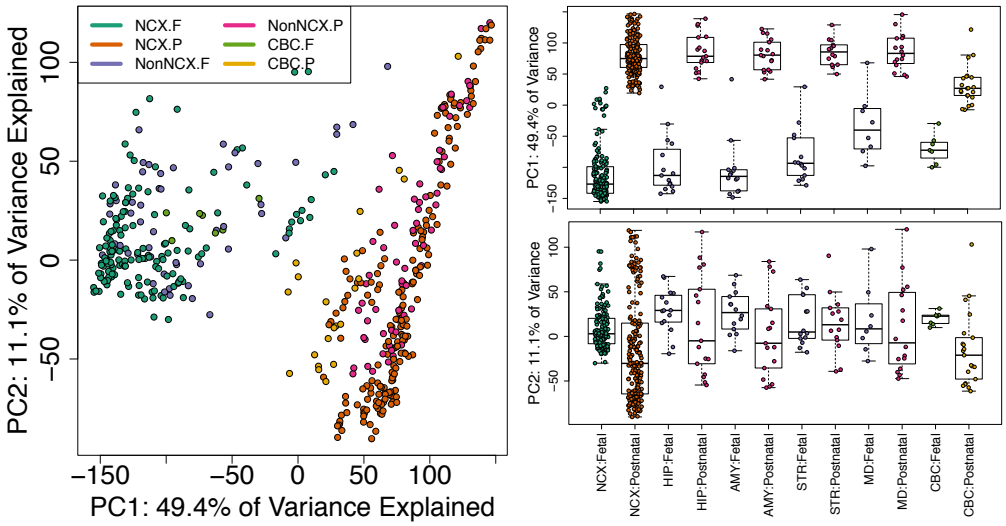


Figure 2.11: Principal components analysis reveals clusters of samples in the BrainSpan data set. (Left) First two principal components (PCs) with samples colored by sample type (F: Fetal or P: Postnatal) and shape given by brain region. (Right) Boxplots for PCs 1 and 2 by brain region (NCX: neocortex, HIP: hippocampus, AMY: amygdala, STR: striatum, MD: thalamus, CBC: cerebellum) and sample type with non-neocortex brain decomposed into its specific regions.

of the single base-level DERs map to sequence previously annotated as non-exonic (e.g. solely intronic or intergenic). The proportion of exonic sequence is higher than our previous analyses in the frontal cortex [20]. When the single base-level DERs are stratified by brain region and developmental period with the highest expression levels (Table 2.2), we find the highest degree of unannotated regulation in the cerebellum, the brain region with the largest degree of region-specific genes in a previous analyses [51]. The majority of DERs, regardless of their annotation, are most highly expressed in fetal life, particularly within the neocortex, hippocampus, and amygdala. Non-exonic expression might be due to incomplete transcript annotation in reference databases, background expression, or previously undetected artifacts.

Table 2.2: Classification of single base-level DERs in the *BrainSpan* project. For each statistically significant DER, we identified the developmental period and region with the highest average expression levels, stratified by annotation relative to the Ensembl gene database. NCX: neocortex, HIP: hippocampus, AMY: amygdala, STR: striatum, MD: thalamus, CBC: cerebellum. Region assignment is prioritized by exon > intron > intergenic.

Group		Exonic	Intergenic	Intronic	Total
NCX	Fetal	15583	1946	1196	18725
	Postnatal	2750	882	415	4047
HIP	Fetal	12511	889	523	13923
	Postnatal	1021	237	144	1402
AMY	Fetal	14705	1178	727	16610
	Postnatal	1193	229	167	1589
STR	Fetal	6952	1706	1199	9857
	Postnatal	4734	1060	905	6699
MD	Fetal	4671	890	431	5992
	Postnatal	2922	425	348	3695
CBC	Fetal	9984	1815	1118	12917
	Postnatal	11382	2932	3921	18235

2.11.5 Exploratory analysis of the cutoff used for the expressed regions-level analysis in the developing human brain

The cutoff used in the expressed regions-level `derfinder` analysis impacts how many ERs are found (Figure 2.12A), their length in base pairs (width, Figure 2.12B). It can also affect the percent of the known annotation that at least overlaps one ER (Figure 2.12C) and conversely the percent of ERs that overlap at least one known exon (Figure 2.12D). Figure 2.12 shows the effect of the cutoff used with the *BrainSpan* data set for a range of cutoffs from 0.025 to 0.5 in increments of 0.025. Note that this data set was already normalized to a library size of 1 million reads. We recommend choosing a cutoff in the

elbow of these curves. In Section 2.4.6 we present the results for cutoffs 0.1 and 0.25 which are at the beginning and the end of the elbow, respectively.

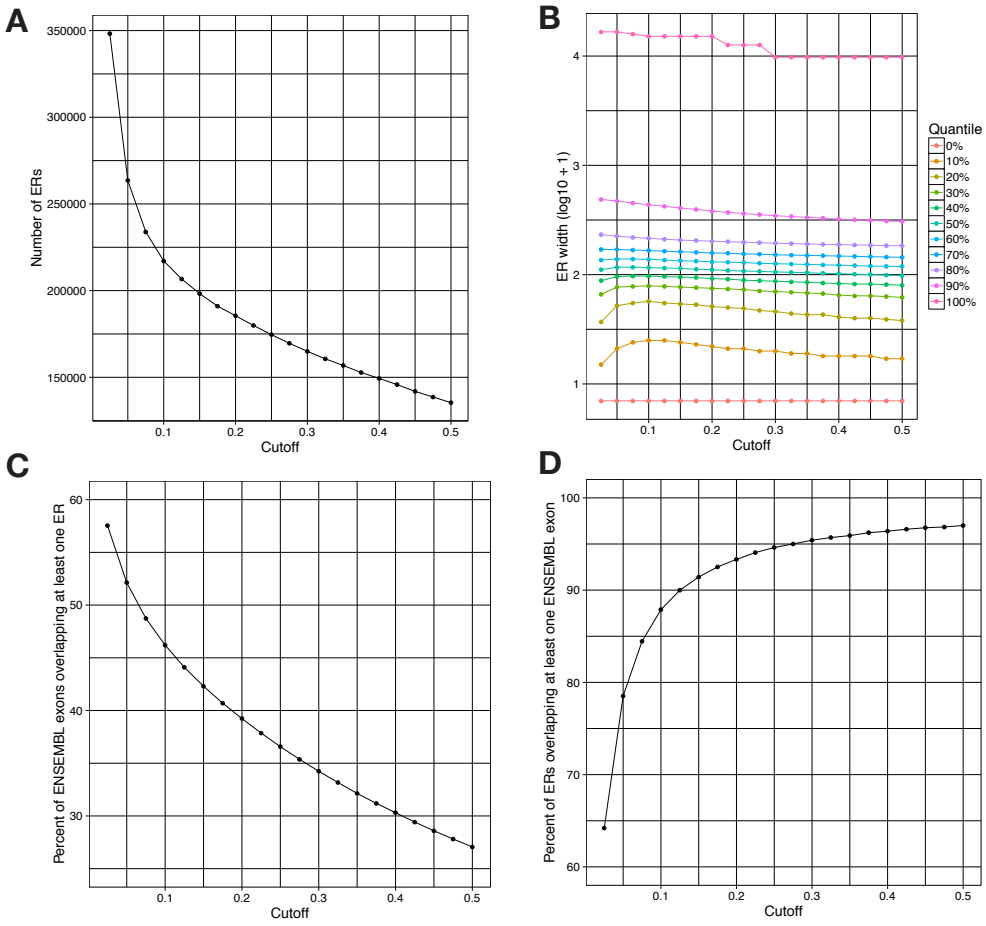


Figure 2.12: Exploratory analysis of the expressed regions cutoff used for the *BrainSpan* data set. A Relationship between number of ERs of at least 6 base-pairs in length against the cutoff used in Figure 2.2A. **B** Distribution of the width of the ERs for each cutoff summarized by quantiles in 10% increments and $\log_1 0$ transformed. **C** Percent of ENSEMBL v75 exons overlapping at least one ER by cutoff. **D** Percent of ERs overlapping at least one ENSEMBL v75 exon by cutoff.

2.11.6 Simulation analysis

2.11.6.1 Simulation results with DESeq2 or edgeR-robust

Table 2.3 shows the empirical power, false positive rate (FPR) and false discovery rate (FDR) for the different analysis pipelines that result in a count matrix which we analyzed with DESeq2 [14] or edgeR-robust [18] while controlling the FDR to 5%. The observed power for edgeR-robust is slightly higher than the corresponding results using DESeq2 [14]. The observed FPR and FDR with edgeR-robust are higher than in the DESeq2 results, with overlapping ranges for the derfinder analyses and non-overlapping ones when summarizing the data with featureCounts [13].

2.11.6.2 Timing and computational resources used

Table 2.4 shows a summary of the computational resources used for the different pipelines used in the simulation as well as the time for running them. In general, the maximum memory per core is low (most are below 3.2 GB) regardless of the analysis step. The exception is alignment with Rail-RNA because of how our computing cluster measures memory usage: it artificially increases when processes spawn shared-many sub-processes by counting more than once the memory used by shared objects. Time-wise all analysis steps except for alignment take only 11 minutes at most. Notably, the ER-level approach is much faster with Rail-RNA output than with HISAT output. This is because derfinder can load the data much faster from BigWig files than from BAM alignment files and the railMatrix has been optimized for the BigWig files that Rail-RNA produces. In this particular simulation, Rail-RNA is slower

Table 2.3: Simulation results for pipelines that used `DESeq2` or `edgeR-robust` for the statistical tests. Minimum and maximum empirical power, false positive rate (FPR) and false discovery rate (FDR) observed from the three simulation replicates for each analysis pipeline that resulted in a count matrix analyzed with `DESeq2` or `edgeR-robust`. `ballgown` analyses were done at either the exon or transcript levels. Pipelines that rely on annotation were run with the full annotation or with 20% of the transcripts missing (8.28% exons missing).

Power	FPR	FDR	Annotation complete	Aligner	Summary method	Statistical method
(94.3-95.5)	(6.8-9)	(12.2-16.1)		HISAT	derfinder	DESeq2
(94.1-95.1)	(6.5-8.3)	(11.1-15.7)		Rail-RNA	derfinder	DESeq2
(95.6-96.1)	(8.4-10.5)	(13.9-18.9)		HISAT	derfinder	edgeR
(94.7-95.9)	(8.3-10.2)	(12.8-18.5)		Rail-RNA	derfinder	edgeR
(70-78.3)	(2.1-3.6)	(5.6-9.8)	No	HISAT	featureCounts	DESeq2
(95.1-95.9)	(2.9-4.8)	(6.3-9.7)	Yes	HISAT	featureCounts	DESeq2
(71.2-79.2)	(5-8.1)	(12-19.5)	No	HISAT	featureCounts	edgeR
(96.4-97.2)	(6.8-9.9)	(12.7-18.1)	Yes	HISAT	featureCounts	edgeR

than HISAT for aligning reads, but this is expected since Rail-RNA is better suited at analyzing larger data sets in the cloud and decreasing false positives when determining new splice junctions. This is reflected on Table 2.1 and 2.3 with slightly reduced FPR and FDR when using Rail-RNA compared to HISAT. The timing results for each computing job are available in the Supplementary Website.

Table 2.4: Summary of computing resources required for each analysis step for the different simulation pipelines. This table shows the maximum memory (GB) per core, the time in minutes to run the analysis with all jobs running sequentially and the maximum number of cores used in any step of the simulation analysis for the different pipelines. Note that the ERs (H), the feature-level counts and ballgown pipelines rely on HISAT alignments. Rail-RNA is abbreviated as (R).

Max memory by core (GB)	Time (minutes)	Peak cores	Pipeline	Analysis step
(2.8-3.1)	(2.1-3.2)	10	ER-level (R)	Align prep
(32.8-39.1)	(137.6-218.4)	10	ER-level (R)	Align
(3.2-3.2)	(47.2-72.1)	40	HISAT	Align
(1.4-1.4)	(1.5-1.5)	1	ER-level (R)	Summarize
(0.6-0.6)	(1.3-1.9)	1	ER-level (R)	Statistical tests
(0.8-0.8)	(5-7)	4	ER-level (H)	Summarize
(0.6-0.6)	(1.5-1.9)	1	ER-level (H)	Statistical tests
(2.2-2.2)	(1.6-1.6)	8	Feature counts	Summarize
(0.6-0.6)	(3.7-5.3)	2	Feature counts	Statistical tests
(2.1-2.1)	(8.7-11)	80	StringTie Ballgown	Summarize
(0.7-0.7)	(0.7-0.8)	2	StringTie Ballgown	Statistical tests

2.12 Supplementary Methods

2.12.1 single base-level `derfinder`

The single base-level approach implemented in `derfinder` requires two models. The alternative model (3.2) contains an intercept, the primary covariate of interest, and optionally adjustment variables. The primary variable can be as simple as a case-control variable or a more complicated model including smoothing functions (e.g. splines) over time. The adjustment variables can include a library size normalization factor for raw data and optionally other potential confounders like age, sex, and batch variables. There are different library size normalization factors you can consider using and `derfinder` implements a version in the `sampleDepth` function based on Paulson et. al [52].

$$y_{ij} = \alpha_i + \sum_{p=1}^{n} \beta_{ip} X_{jp} + \sum_{q=1}^{m} \gamma_{iq} Z_{jq} + \epsilon_{ij} \tag{2.3}$$

In both models y_{ij} is the scaled $\log_2$ base-level coverage for genomic position i and sample j. That is, $y_{ij} = \log_2 \Big(\text{coverage}_{ij} + \text{scaling factor}\Big)$. The model is completed by the n group effects $\mathbf{fi_i}$, m adjustment variable effects $\mathbf{fl_i}$ and potentially correlated measurement error ϵ. The null model (3.1) is nested within model (3.2) and contains only the intercept and adjustment variables.

$$y_{ij} = \alpha_i + \sum_{q=1}^{m} \gamma_{iq} Z_{jq} + \epsilon_{ij} \tag{2.4}$$

`derfinder` uses a fixed design matrix, testing the same hypothesis at every

base. This permits fast vectorized differential expression analysis. At each base we compute a moderated F-statistic [16] of the form in equation (2.5), where RSS0_i and RSS1_i are the residual sum of squares of the null and alternative models for base i. Furthermore, df_0 and df_1 are the degrees of freedom for the null (3.1) and alternative (3.2) models respectively, n is the number of samples, and an offset can be used for smaller experiments to shrink large F-statistics that may be driven by few biological replicates that cluster tightly.

$$F_i = \frac{(\text{RSS0}_i - \text{RSS1}_i)/(\text{df}_1 - \text{df}_0)}{\text{offset} + (\text{RSS1}_i/(n - \text{df}_1))} \tag{2.5}$$

We then perform "bump hunting" adapted to `Rle` objects in order to identify candidate DERs, R_k. Candidate DERs are defined as contiguous sets of bases where $F_i > T$ for a fixed threshold T. We then calculate an "area" statistic for each candidate DER which is the sum of the F-statistics above the threshold within the region: $S_k = \sum_{j \in R_k} F_j$ (Figure 2.10B). We have previously applied this approach to identify local differentially and variably methylated regions and more long range changes in methylation [33, 53, 54]. One key difference compared to previous implementations in DNA methylation data is that we do not explicitly smooth the F-statistics, allowing for precise discovery of intron-exon boundaries in the data (Figure 2.10C).

Permutation analysis generates statistical significance for each of these candidate DERs by permuting the sample labels, re-calculating the F-statistics, identifying null candidate regions and region-level statistics in this permuted data set, and then calculating empirical p-values and/or directly estimating the family-wise error rate (FWER) [33]. Alternatively, the empirical p-values

can be adjusted to control the false discovery rate (FDR) via `qvalue` [49].

2.12.2 Data Processing: BrainSpan data

For the single base-level analysis, we used a scaling factor of 1 and chose the F-statistic cutoff T such that $P(F > T) = 10^{-6}$. We used the same alternative model described for the expressed region analysis in the main text. We compared the alternative model to an intercept-only model, and identified DERs using the single base-level analysis. We then calculated the mean coverage for each significant single base-level DERs in each sample, resulting in a mean coverage matrix (DERs by samples), and we performed principal component analysis (PCA) on this $\log_2$-transformed matrix (after adding an offset of 1), which were subsequently plotted in Figure 2.11.

References

[1] C. M. Farrell et al. “Current status and new features of the Consensus Coding Sequence database”. In: *Nucleic acids research* 42.Database issue (2014). PMID: 24217909 PMCID: PMC3965069, pp. D865–872. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1059.

[2] GTEx Consortium. “The Genotype-Tissue Expression (GTEx) project”. In: *Nature genetics* 45.6 (2013). PMID: 23715323, pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653.

[3] GTEx Consortium. “Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans”. In: *Science* 348.6235 (2015). PMID: 25954001 PMCID: PMC4547484, pp. 648–660. ISSN: 1095-9203. DOI: 10.1126/science.1262110.

[4] ENCODE Project Consortium et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012). PMID: 22955616 PMCID: PMC3439153, pp. 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247.

[5] T. Lappalainen et al. “Transcriptome and genome sequencing uncovers functional variation in humans”. In: *Nature* 501.7468 (2013). PMID: 24037378 PMCID: PMC3918453, pp. 506–511. ISSN: 1476-4687. DOI: 10.1038/nature12531.

[6] A. A. Dillman et al. “mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex”. In: *Nature neuroscience* 16.4 (2013). PMID: 23416452 PMCID: PMC3609882, pp. 499–506. ISSN: 1546-1726. DOI: 10.1038/nn.3332.

[7] B. Daines et al. “The Drosophila melanogaster transcriptome by paired-end RNA sequencing”. In: *Genome research* 21.2 (2011). PMID: 21177959 PMCID: PMC3032934, pp. 315–324. ISSN: 1549-5469. DOI: 10.1101/gr.107854.110.

[8] T. Steijger et al. “Assessment of transcript reconstruction methods for RNA-seq”. In: *Nature methods* 10.12 (2013). PMID: 24185837 PMCID: PMC3851240, pp. 1177–1184. ISSN: 1548-7105. DOI: 10.1038/nmeth.2714.

[9] S. Anders, P. T. Pyl, and W. Huber. “HTSeq—a Python framework to work with high-throughput sequencing data”. In: *Bioinformatics* 31.2 (2015). PMID: 25260700, pp. 166–169. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btu638.

[10] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010). PMID: 19910308 PMCID: PMC2796818, pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616.

[11] S. Anders and W. Huber. “Differential expression analysis for sequence count data”. In: *Genome biology* 11.10 (2010). PMID: 20979621 PMCID: PMC3218662, R106. ISSN: 1465-6914. DOI: 10.1186/gb-2010-11-10-r106.

[12] McCarthy et al. “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation”. In: *Nucleic Acids Research* 40.10 (2012). PMID: 22287627 PMCID: PMC3378882, pp. 4288–4297. DOI: 10.1093/nar/gks042.

[13] Y. Liao, G. K. Smyth, and W. Shi. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7 (2014). PMID: 24227677, pp. 923–930. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btt656.

[14] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (2014). PMID: 25516281 PMCID: PMC4302049, p. 550. ISSN: 1465-6914. DOI: 10.1186/s13059-014-0550-8.

[15] M. E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* 43.7 (2015). PMID: 25605792 PMCID: PMC4402510, e47.

[16] G. K. Smyth. “Linear models and empirical Bayes methods for assessing differential expression in microarray experiments”. In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004). PMID: 16646809, pp. 1–25.

[17] C. W. Law et al. “Voom: precision weights unlock linear model analysis tools for RNA-seq read counts”. In: *Genome biology* 15.2 (2014). PMID: 24485249, R29. ISSN: 1465-6914. DOI: 10.1186/gb-2014-15-2-r29.

[18] X. Zhou, H. Lindsay, and M. D. Robinson. “Robustly detecting differential expression in RNA sequencing data using observation weights”. In: *Nucleic Acids Research* 42.11 (2014). PMID: 24753412 PMCID: PMC4066750, e91. ISSN: 1362-4962. DOI: 10.1093/nar/gku310.

[19] A. C. Frazee et al. “Differential expression analysis of RNA-seq data at single-base resolution”. In: *Biostatistics* 15 (2014). PMID: 24398039, pp. 413–426. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxt053.

[20] A. E. Jaffe et al. “Developmental regulation of human cortex transcription and its clinical relevance at single base resolution”. In: *Nature Neuroscience* 18.1 (2015). PMID: 25501035 PMCID: PMC4281298, pp. 154–161. ISSN: 1546-1726. DOI: 10.1038/nn.3898.

[21] BrainSpan. *Atlas of the Developing Human Brain*. Funded by ARRA Awards 1RC2MH089921-01, 1RC2MH090047-01, and 1RC2MH089929-01. 2011. URL: http://developinghumanbrain.org.

[22] A. C. Frazee et al. “Ballgown bridges the gap between transcriptome assembly and expression analysis”. In: *Nature Biotechnology* 33.3 (2015). PMID: 25748911, pp. 243–246. ISSN: 1546-1696. DOI: 10.1038/nbt.3172.

[23] M. Pertea et al. “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads”. In: *Nature Biotechnology* 33.3 (2015). PMID: 25690850 PMCID: PMC4643835, pp. 290–295. ISSN: 1546-1696. DOI: 10.1038/nbt.3122.

[24] J Rainer. *EnsDb.Hsapiens.v75: Ensembl based annotation package*. R package version 0.99.12. 2015.

[25] S Falcon and R Gentleman. “Using GOstats to test gene lists for GO term association.” In: *Bioinformatics* 23.2 (2007), pp. 257–8.

[26] A. Nellore et al. “Rail-dbGaP: analyzing dbGaP-protected data in the cloud with Amazon Elastic MapReduce”. In: *Bioinformatics* btw177v2 (2016). PMID: 27153614, pp. 1–3. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw177.

[27] M. Lawrence et al. “Software for Computing and Annotating Genomic Ranges”. In: *PLoS Computational Biology* 9 (8 2013). PMID: 23950696 PMCID: PMC3738458, e1003118. DOI: 10.1371/journal.pcbi.1003118.

[28] A. C. Frazee et al. “Polyester: simulating RNA-seq datasets with differential transcript expression”. In: *Bioinformatics* 31.17 (2015). PMID: 25926345 PMCID: PMC4635655, pp. 2778–2784. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv272.

[29] D. Kim, B. Langmead, and S. L. Salzberg. “HISAT: a fast spliced aligner with low memory requirements”. In: *Nature Methods* 12.4 (2015). PMID: 25751142 PMCID: PMC4655817, pp. 357–360. ISSN: 1548-7105. DOI: 10.1038/nmeth.3317.

[30] D. Kim et al. “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”. In: *Genome biology* 14.4 (2013). PMID: 23618408, R36. ISSN: 1465-6914. DOI: 10.1186/gb-2013-14-4-r36.

[31] W. J. Kent et al. “BigWig and BigBed: enabling browsing of large distributed datasets”. In: *Bioinformatics* 26.17 (2010), pp. 2204–2207.

[32] H. Li et al. “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[33] A. E. Jaffe et al. “Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies”. In: *International journal of epidemiology* 41.1 (2012). PMID: 22422453 PMCID: PMC3304533, pp. 200–209. ISSN: 1464-3685. DOI: 10.1093/ije/dyr238.

[34] D. Selcen and A. G. Engel. “Mutations in ZASP define a novel form of muscular dystrophy in humans”. In: *Annals of Neurology* 57.2 (2005). PMID: 15668942, pp. 269–276. ISSN: 0364-5134. DOI: 10.1002/ana.20376.

[35] A. Osio et al. “Myozenin 2 is a novel gene for human hypertrophic cardiomyopathy”. In: *Circulation Research* 100.6 (2007). PMID: 17347475 PMCID: PMC2775141, pp. 766–768. ISSN: 1524-4571. DOI: 10.1161/01.RES.0000263008.66799.aa.

[36] A. W. Duncan et al. “Aneuploidy as a mechanism for stress-induced liver adaptation”. In: *The Journal of Clinical Investigation* 122.9 (2012). PMID: 22863619 PMCID: PMC3428097, pp. 3307–3315. ISSN: 1558-8238. DOI: 10.1172/JCI64026.

[37] T. Sakamoto et al. “Expression and properties of human liver beta-ureidopropionase”. In: *Journal of Nutritional Science and Vitaminology* 47.2 (2001). PMID: 11508704, pp. 132–138. ISSN: 0301–4800.

[38] L. Collado-Torres, A. E. Jaffe, and J. T. Leek. "regionReport: Interactive reports for region-level and feature-level genomic analyses [version2; referees: 2 approved, 1 approved with reservations]". In: *F1000Research* 4 (2016), pp. 1–10. DOI: 10.12688/f1000research.6379.2.

[39] J. K. Pickrell et al. "Noisy splicing drives mRNA isoform diversity in human cells". In: *PLoS Genet* 6.12 (2010), e1001236.

[40] J. T. Leek et al. "Tackling the widespread and critical impact of batch effects in high-throughput data". In: *Nature Reviews Genetics* 11.10 (2010). PMID: 20838408 PMCID: PMC3880143, pp. 733–739. DOI: 10.1038/nrg2825.

[41] C. C. Ouimet, H. C. Hemmings, and P. Greengard. "ARPP-21, a cyclic AMP-regulated phosphoprotein enriched in dopamine-innervated brain regions. II. Immunocytochemical localization in rat brain". In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 9.3 (1989). PMID: 2538585, pp. 865–875. ISSN: 0270-6474.

[42] R. Cachope and J. F. Cheer. "Local control of striatal dopamine release". In: *Frontiers in Behavioral Neuroscience* 8 (2014). PMID: 24904339 PMCID: PMC4033078, p. 188. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2014.00188.

[43] K. Melchior et al. "The WNT receptor FZD7 contributes to self-renewal signaling of human embryonic stem cells". In: *Biological Chemistry* 389.7 (2008). PMID: 18681827, pp. 897–903. ISSN: 1431-6730. DOI: 10.1515/BC.2008.108.

[44] V. M. Tatard et al. "ZNF238 is expressed in postmitotic brain cells and inhibits brain tumor growth". In: *Cancer Research* 70.3 (2010). PMID: 20103640, pp. 1236–1246. ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-09-2249.

[45] G. Poulin, B. Turgeon, and J. Drouin. "NeuroD1/beta2 contributes to cell-specific transcription of the proopiomelanocortin gene". In: *Molecular and Cellular Biology* 17.11 (1997). PMID: 9343431 PMCID: PMC232521, pp. 6673–6682. ISSN: 0270–7306.

[46] I. Gallego Romero et al. "RNA-seq: impact of RNA degradation on transcript quantification". In: *BMC biology* 12 (2014). PMID: 24885439 PMCID: PMC4071332, p. 42. ISSN: 1741-7007. DOI: 10.1186/1741-7007-12-42.

[47] F. Cunningham et al. “Ensembl 2015”. In: *Nucleic Acids Research* 43.Database issue (2015). PMID: 25352552 PMCID: PMC4383879, pp. D662–669. ISSN: 1362-4962. DOI: 10.1093/nar/gku1010.

[48] A. Mortazavi et al. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature methods* 5.7 (2008). PMID: 18516045, pp. 621–628. DOI: 10.1038/nmeth.1226.

[49] A. Dabney and J. D. Storey. *qvalue: Q-value estimation for false discovery rate control*. R package version 1.40.0. 2014. URL: http://www.bioconductor.org/packages/qvalue.

[50] J. D. Storey and R. Tibshirani. “Statistical significance for genomewide studies”. In: *Proceedings of the National Academy of Sciences* 100.16 (2003). PMID: 12883005 PMCID: PMC170937, pp. 9440–9445. DOI: 10.1073/pnas.1530509100.

[51] H. J. Kang et al. “Spatio-temporal transcriptome of the human brain”. In: *Nature* 478.7370 (2011). PMID: 22031440 PMCID: PMC3566780, pp. 483–489. ISSN: 1476-4687. DOI: 10.1038/nature10523.

[52] J. N. Paulson et al. “Differential abundance analysis for microbial marker-gene surveys”. In: *Nature methods* (2013). PMID: 24076764 PMCID: PMC4010126. ISSN: 1548-7105. DOI: 10.1038/nmeth.2658.

[53] A. E. Jaffe et al. “Significance analysis and statistical dissection of variably methylated regions”. In: *Biostatistics* 13.1 (2012). PMID: 21685414 PMCID: PMC3276267, pp. 166–178. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxr013.

[54] K. D. Hansen et al. “Increased methylation variation in epigenetic domains across cancer types”. In: *Nature genetics* 43.8 (2011). PMID: 21706001 PMCID: PMC3145050, pp. 768–775. ISSN: 1546-1718. DOI: 10.1038/ng.865.

Chapter 3

Differential binding analysis with `derfinder`

Leonardo Collado-Torres[1,2,3], Jeffrey T. Leek[1,2,*], Andrew E. Jaffe[1,2,3,4,†].

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
2. Center for Computational Biology, Johns Hopkins University
3. Lieber Institute for Brain Development, Johns Hopkins Medical Campus
4. Department of Mental Health, Johns Hopkins University

* *corresponding author;* `jtleek@gmail.com`

† *corresponding author;* `andrew.jaffe@libd.org`

3.1 Abstract

Background

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments identify regions of the genome with binding signal for a protein of interest. When multiple samples are collected for different conditions, treatments or other covariates, researchers will ask if there is differential binding between these conditions. The current strategies for answering this question rely on merging peaks from the different samples which can lead to un-wanted issues. These strategies do not take into account the variability across samples when merging peaks.

Results

Here we show that the Bioconductor package `derfinder` can be used to identify differentially bound peaks using ChIP-seq data, bypassing the peak calling step. The software is flexible, annotation-agnostic and takes into account the variability across all samples in determining differentially bound peaks. We illustrate the approach using ChIP-seq data from the *EpiMap* study for histone marks H3K4me3 and H3K27ac from the human brain. We identify differentially bound peaks associated with cell type, brain region and/or age at time of death. We show that most of the differentially bound peaks are associated with cell type, although some are also associated with technical covariates. We compare our approach to results from `DiffBind`, one of the most widely used software for differential binding analysis.

Conclusions

`derfinder` can be successfully used to identify differentially bound peaks using ChIP-seq data. This approach solves the merging peaks problem where you have to choose between analyzing wide peaks or peak summits.

The package is available at www.bioconductor.org/packages/derfinder.

Keywords ChIP sequencing, differential binding analysis, ChIP-seq, ATAC-seq, ChIP-exo, DNase-seq.

3.2 Introduction

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is the main assay used to identify regions of the genome where a given protein of interest binds to the genome. For example ChIP-seq can be used with transcription factors to identify transcription factor binding sites. ChIP-seq experiments nowadays are widely used as the technology has benefited from decreased sequencing costs. These experiments typically produce high-throughput short sequence reads for samples of interest as well as control input samples. The control input samples can be used by peak caller software to adjust for potential noise from the immunoprecipitation step. Model-based Analysis for ChIP-seq (MACS) [1] is one of the most commonly used peak callers and is among the best as evaluated with different metrics [2].

ChIP-seq experiments can be used to identify differential binding peaks between two conditions or more complicated designs. The analysis pipeline for determining differential binding peaks typically begins by using a peak-caller such as MACS [1] for each sample (Figure 3.1A). Once the peaks for each sample have been identified, the next step is to merge them to build a consensus peak set (Figure 3.1B) using custom scripts or software such as DiffBind [3, 4], diffReps [5], among others [6]. A count matrix based on this consensus peak set is then constructed in a similar process to how RNA-seq

count matrices are created. Then this count matrix is analyzed with software for differential expression analysis such as `DESeq2` [7]. Alternatively, window based analyses are possible with software such as `csaw` [8].

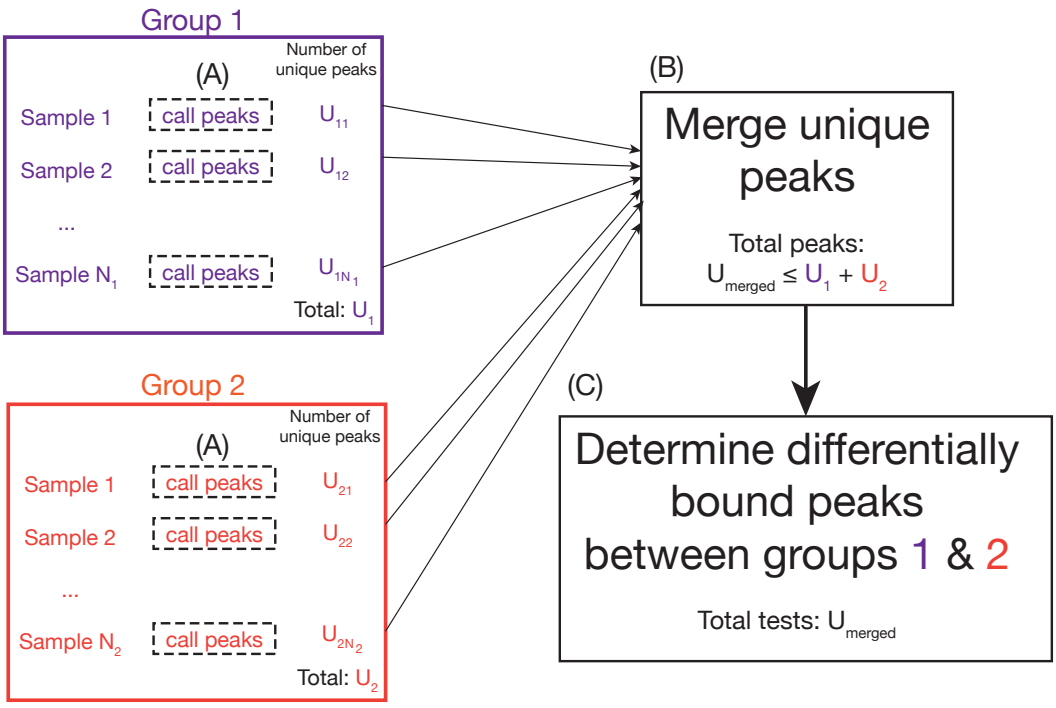


Figure 3.1: Current strategy for identifying differentially bound peaks between two conditions. (A) A peak-caller is used independently for each sample from both groups to identify peaks. **(B)** The peaks from all samples are merged to determine a common set across all samples. **(C)** For each merged peak, a statistical test is performed to determine whether the peak is differentially bound between the conditions.

The current strategy for identifying differentially bound peaks does not take into account the variability across samples for determining peaks. This strategy also ignores the variability among samples when merging. The merging step is performed by sequentially identifying which unique peaks overlap with each other (Figure 3.2A). These can lead to wide consensus peaks as shown in Figure 3.2B for the consensus peak with the highest fold

change that increased over time for time-course ChIP-seq experiment [9]. This widening effect can be limited by identifying the base-pair with the highest coverage among all samples called summit in `DiffBind` [3] and using only a fixed window size surrounding this peak.

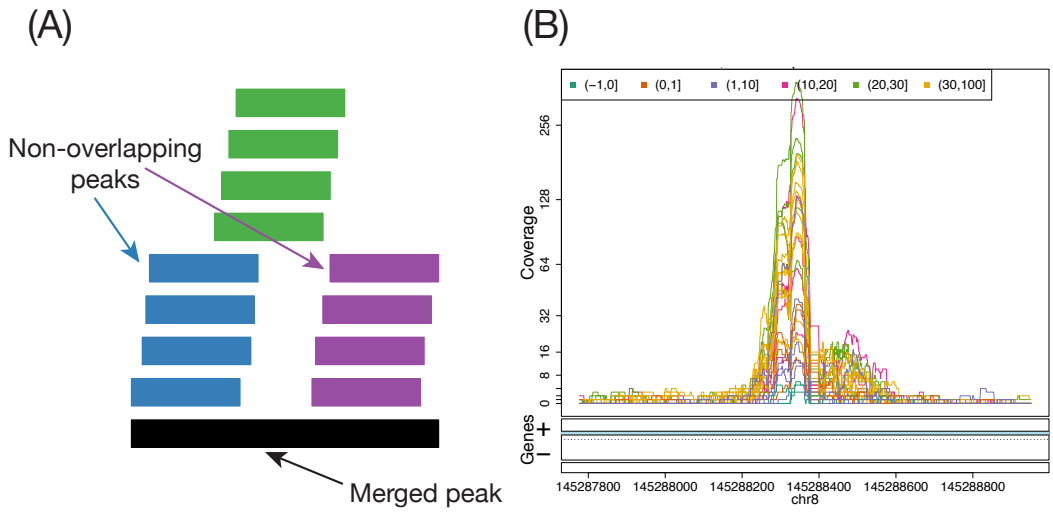


Figure 3.2: Merging peaks by overlaps can lead to wide peaks. **(A)** Peaks are merged sequentially by finding other peaks than overlap them, which can lead to two non-overlapping peaks being merged into the same merged peak. **(B)** Example of a wide merged peak with strong coverage support in the middle region of the peak and low support on the ends. Samples are colored by age group.

Here we show that `derfinder` [10] can be used to directly identify the differentially bound peaks with minor alterations to the pipeline used for RNA-seq data as shown in Figure 3.3. We illustrate our proposed strategy using ChIP-seq samples for histone marks H3K4me4 and H3K27ac from the human brain. Using `derfinder` for identifying differentially bound peaks skips the traditional peak calling step.

The differentially bound peaks (dbPeaks) identified with `derfinder` have differential binding signal support for all bases, which is not necessarily true

for current strategies (Figure 3.2B). Furthermore, by calculating F-statistics `derfinder` does take into account the sample variability across each base-pair when determining the dbPeaks. The proposed strategy does not rely on input samples since it does not identify peaks per sample. Strategies that rely on input samples can be noisy since many input samples show anomalous signal in some regions. This problem can be addressed using software such as `GreyListChIP` [11]. While `derfinder` is not a peak caller, given several replicate input runs `derfinder` could be used to identify differential binding between samples and input for peak calling.

3.3 Results

3.3.1 Finding differentially bound peaks with `derfinder`

The DER Finder methodology [12] was initially designed to identify differentially expressed regions with RNA-seq data without relying on annotation. `derfinder` [10] implements this method in two approaches. The single base-level approach is based on calculating F-statistics at every single base-pair of the genome where enough input signal is available. It aims to find sharp boundaries which is a desired feature given the nature of RNA-seq data. We modified the single base-level approach for ChIP-seq data, and in general any genomic data type where the regions of interest do not have sharp boundaries.

First, we define two models: an alternative and a null model. Using these models we calculate F-statistics at every base as shown in Figure 3.3A which take into account the sample variability. This results in a F-statistic curve along the genome as shown in Figure 3.3B. With RNA-seq data, we

would then determine differentially expressed regions at that point using a global cutoff based on the distribution of the F-statistics. For ChIP-seq data, we perform a smoothing step to the F-statistic curve as shown in Figure 3.3C. We then determine the differentially bound peaks using a global cutoff. Once the differentially bound peaks (dbPeaks) have been identified, we can identify where the dbPeaks are located in the genome and visualize them using `derfinderPlot` [13] as shown in Figure 3.3D-F.

`derfinder` is very flexible and can identify differentially bound peaks using different models. It can be used for time-course ChIP-seq experiments and does not rely on input samples for determining the differentially bound peaks. Technical details are described in Methods Section 3.5.1.

3.3.2 Differentially bound peaks for histone marks H3K4me3 and H3K27ac in the human brain

To illustrate our approach for identifying differentially bound peaks with `derfinder`, we used 62 and 57 ChIP-seq samples from the *EpiMap* study [14] for histone marks H3K4me3 and H3K27ac, respectively. Using fluorescence-activated cell sorting, 31 and 28 of the samples were determined to be negative for the NeuN antibody, respectively for H3K4me3 and H3K27ac. For both histone marks, 29 of the samples are from anterior cingulate cortex (ACC) with the remaining samples extracted from the dorsolateral prefrontal complex (DLPFC). Details are shown in Table 3.1 and Supplementary Methods 3.10.

For each histone mark, we used `derfinder`'s single base-level approach to identify differentially bound peaks (dbPeaks) for the 8 groups given by

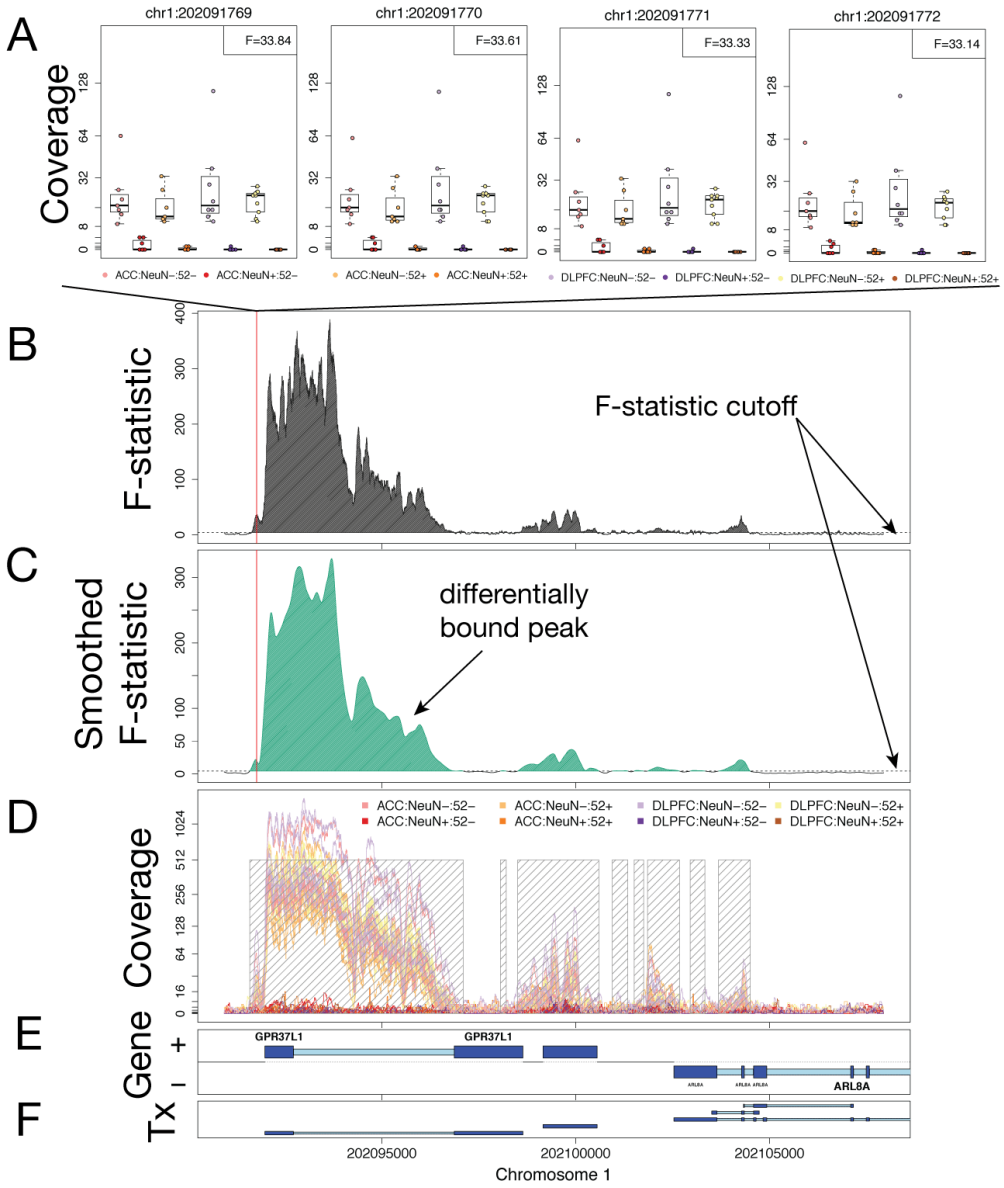


Figure 3.3: Identification of differentially bound peaks with `derfinder`. **A** Boxplots of the coverage for 4 consecutive bases with the F-statistic for difference between the 8 groups. **B** F-statistic curve across a window of chromosome 1. **C** Smoothed F-statistic curve across the same window. Regions above the cutoff are labeled as candidate differentially bound peaks (dbPeaks). **D** Raw sample coverage plots with candidate dbPeaks highlighted. **E** Known gene and **F** transcripts for this window of chromosome 1. Known exons (dark blue) and introns (light blue) are shown by strand. The data is from the H3K4me3 histone mark.

Table 3.1: Sample information from the *EpiMap* study for histone marks H3K4me3 and H3K27ac. Further information about these samples is available at Supplementary Methods 3.10.

Histone mark	Brain region	NeuN antibody	Number of samples
H3K4me3	ACC	negative	14
		positive	15
	DLPFC	negative	17
		positive	16
H3K27ac	ACC	negative	14
		positive	15
	DLPFC	negative	14
		positive	14

cell type, brain region, and age at time of death. We smoothed the F-statistics using a window of 300 base-pairs. We then identified the set of candidate dbPeaks with a family-wise error rate (FWER) adjusted p-value less than 0.05. This resulted in 29,939 and 204,026 dbPeaks for H3K4me3 and H3K27ac histone marks with median lengths of 1,292 and 1,135 base-pairs, respectively. The minimum and maximum lengths in base-pairs are 182 and 24,598 for H3K4me3 and 141 and 65,971 for H3K27ac.

The dbPeaks for these histone marks overlap in different ways Ensembl v75 [15]'s features as shown in Figure 3.4. For the histone mark H3K4me3, the dbPeaks overlapping known exons and introns are the most frequent group (33%, Figure 3.4A) and span 20.5 mega base-pairs (mb) of the genome representing 38.95% of all bases spanned by these dbPeaks. For H3K27ac, most of the dbPeaks overlap exclusively known intronic regions (51.1%, Figure 3.4B) and in total the dbPeaks cover a much larger portion of the genome: 347.9 versus 52.6 mb. For both histone marks, dbPeaks that overlap all types

of features show an increase in percent of the total by length of the genome spanned with respect to the total number of peaks (1.42 and 2.67 fold change). This is expected since the peaks have to be long to cover all three types of features.

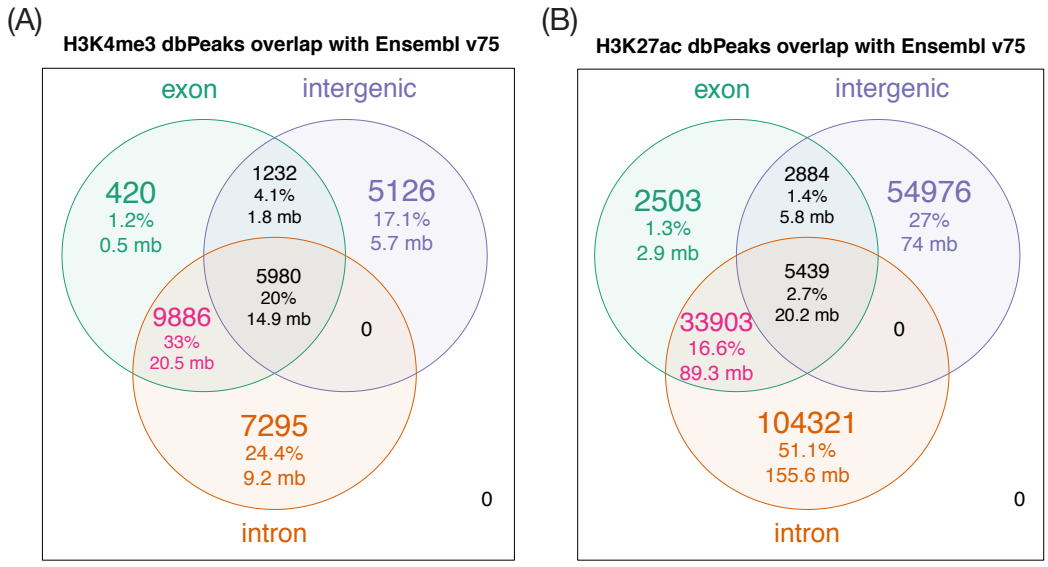


Figure 3.4: Overlap between differentially bound peaks for H3K4me3 and H3K27ac marks and Ensembl v75 features. Overlaps are shown in venn diagrams for **(A)** H3K4me3 and **(B)** H3K27ac differentially bound peaks (dbPeaks) by cell type, brain region or age at death. Percent of dbPeaks and total mega base-pairs spanned are shown below the number of dbPeaks.

3.3.2.1 Characterization of differentially bound peaks by modeled covariates

For each of histone marks we calculated a count matrix summarizing the base-level information as described in Methods Section 3.5.2. Using this $\log_2$-transformed matrix, we fit a linear model to using an intercept term and one of the main covariates: brain region, cell type, or age at death (continuous). For each covariate we identified the dbPeaks that are significantly associated

(FWER < 0.05) with the covariate. 89.2% and 94.5% of the dbPeaks are only associated with cell type for H3K4me3 and H3K27ac (Figure 3.9).

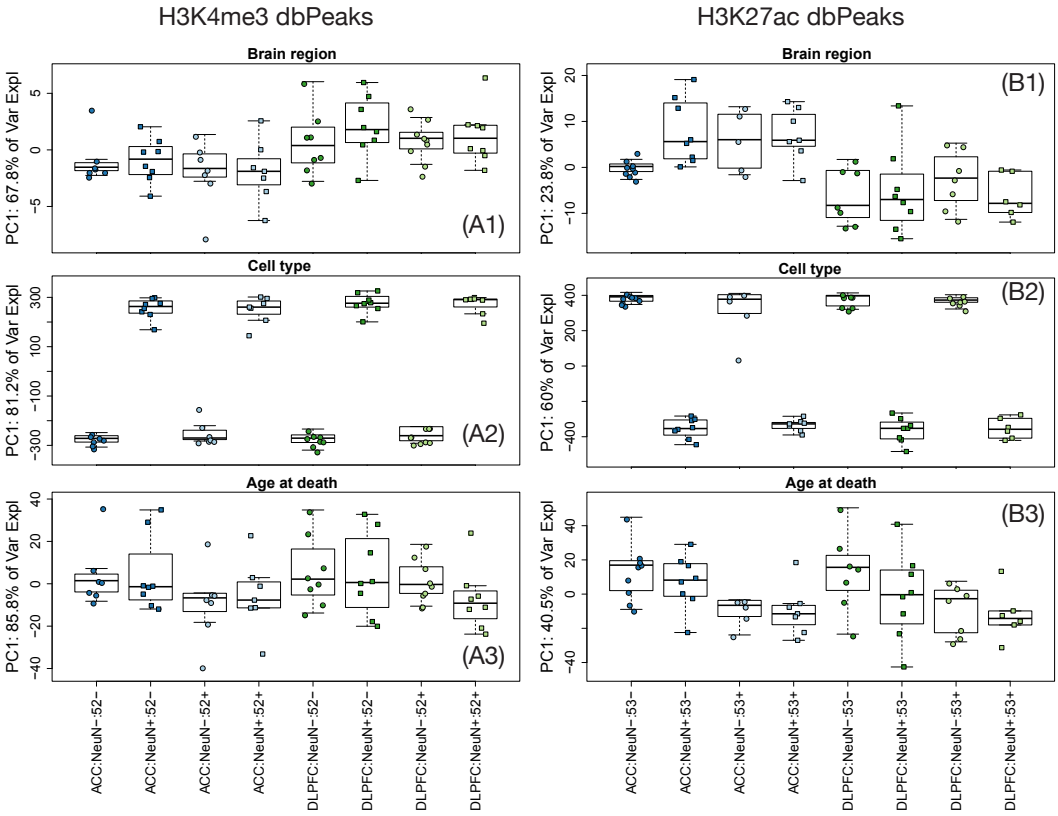


Figure 3.5: Boxplots showing the first principal component for dbPeaks with significant associations with a modeled covariate. (A) H3K4me and **(B)** H3K27ac dbPeaks associated with brain region (**1**), cell type (**2**) and age at death (**3**). ACC samples are shown in blue, DLPFC samples in green. NeuN- samples are shown with circles, NeuN+ samples with squares. Darker colors and lighter colors are used for samples below and above the median age at time of death, respectively. The number of dbPeaks for each principal component analysis is given in Figure 3.9.

For the dbPeaks associated with each of the modeled covariates (brain region, cell type, age at death), we performed a principal component analysis using the $\log_2$-transformed matrix. For H3K4me3, the first principal component (PC) explains at least 67.8% of the variance (Figure 3.5A) while the equivalent analysis for H3K27ac showed that the first PC explains at most

60% of the variance (Figure 3.5B). For age at death, this drop can be explained by a decrease in the percent of dbPeaks strictly associated with age at death between H3K4me3 and H3K27ac (Figure 3.9). When performing the principal component analysis with all dbPeaks, the first PC explains 77.9% and 58.8% of the variance for H3K4me3 and H3K27ac dbPeaks and is markedly associated with cell type.

3.3.2.2 Example differentially bound peaks highlight problems with the current strategy for merging peaks

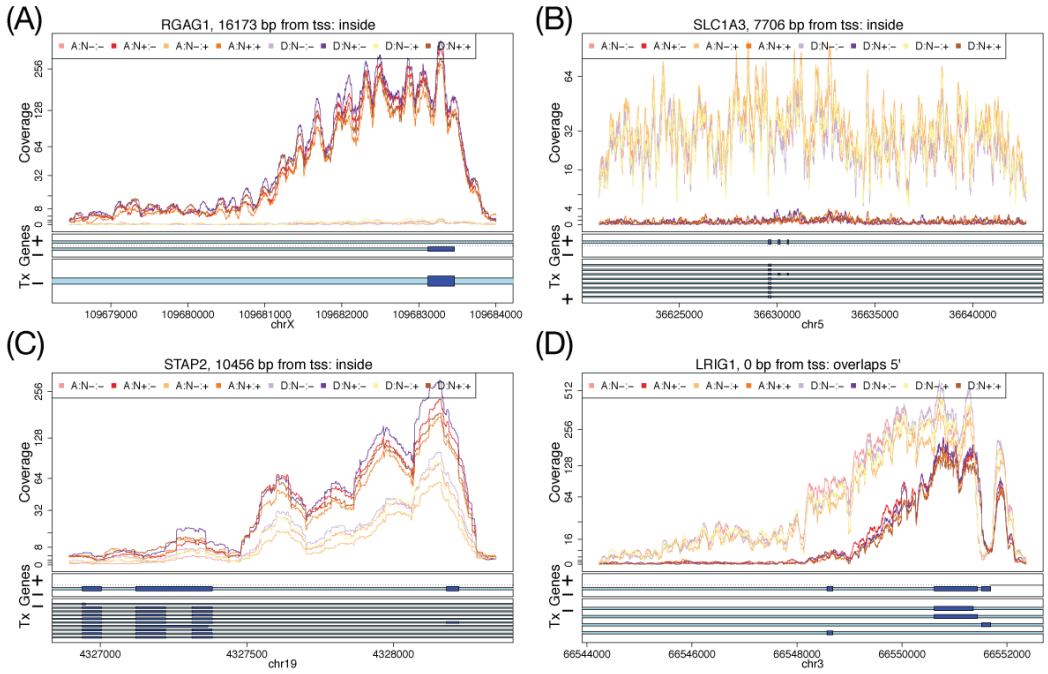


Figure 3.6: Coverage plots for average coverage levels for differentially bound peaks. (A) The sixth strongest H3K4me3 dbPeak and **(B)** second strongest H3K27ac dbPeak associated with cell type. **(C)** and **(D)** are two other H3K4me3 dbPeaks that show differences by cell type. Lines are colored by group with NeuN- samples shown in lighter colors. ACC and DLPFC abbreviated as *A* and *D*, NeuN- and NeuN+ as *N-* and *N+*, and below or above the median age at time of death as - and +, respectively.

The strong association with cell type is remarkable when visually exploring

the resulting dbPeaks as shown in Figure 3.6. For example, the dbPeak with the sixth strongest association with cell type for H3K4me3 shown in Figure 3.6A and it overlaps *RGAG1*, a gene that has been associated to non-syndromic X-linked intellectual disability [16]. The H3K27ac dbPeak shown in Figure 3.6B overlaps *SLC1A3*, a gene that encodes a high affinity glutamate transporter family known to be important for astrocytes [17].

In Figure 3.6 the mean coverage for all groups by cell type is very similar. In Figure 3.6A the NeuN+ samples compose a coverage curve with multiple summits and would likely be broken into different consensus peaks when using `DiffBind` with the summits argument to control the widening effect of the merging step. This could be a problem and it shows that with the current strategy for identify differentially bound peaks an arbitrary choice leads to either wide consensus peaks (Figure 3.2B), or to splitting peaks.

3.3.2.3 Variation in the differentially bound peaks

We produced a similar figure to Figure 3C from *Geuvadis* RNA-seq analysis [18] and reproduced in Figure 3 of the `Rail-RNA` software paper [19]. For each histone mark, we used the $\log_2$-transformed count matrix and fit a model for each dbPeak with the 3 modeled covariates (brain region, cell type, age at death) as well as 12 other covariates, some of which are biological (BMI, sex) and some of which are technical such as flowcell batch and total mapped reads. Figure 3.7 shows the percent of variance explained for all the dbPeaks by each of these 15 covariates and the residual variation.

In contrast to the *Geuvadis* RNA-seq data [18, 19], residual variation is

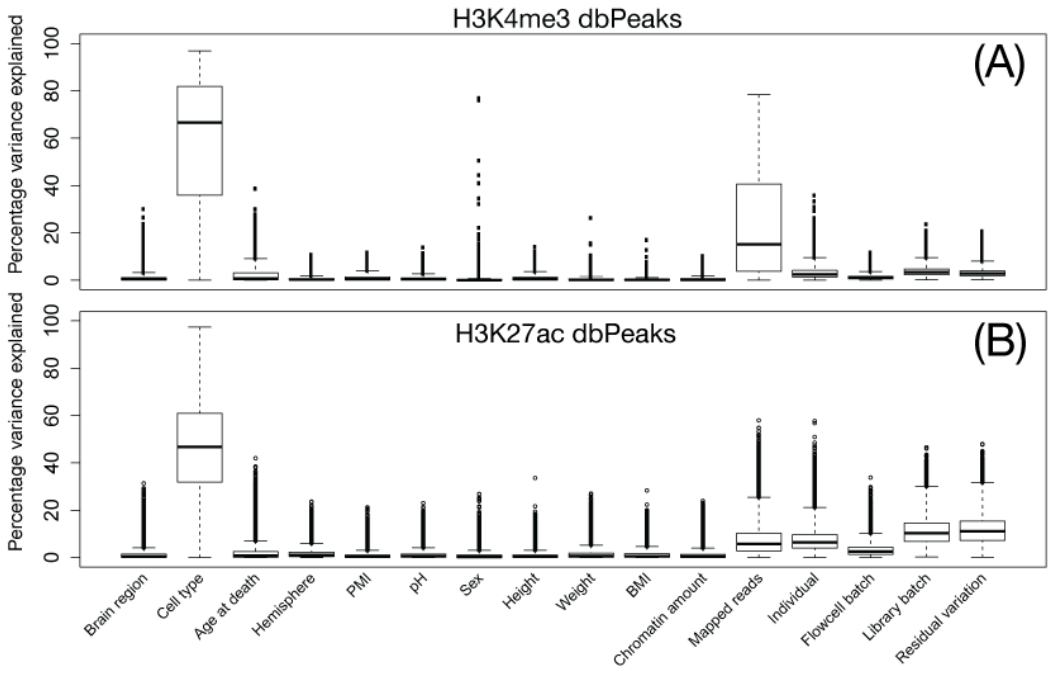


Figure 3.7: Boxplots of percentage of variation explained by the 3 modeled covariates, 12 other covariates, and residual variation. Boxplots for **(A)** H3K4me3 and **(B)** H3K27ac dbPeaks.

not the most important factor. This is expected since we are only analyzing regions of the genome that showed differential binding by brain region, cell type and/or age at death. For both histone marks, cell type explains most of the variance and it does so more strongly for H3K4me3. Notably, the total number of mapped reads explains more of the variability than brain region and age at death. This effect is weaker for H3K27ac, whose samples were prepared such that around 80 million uniquely mapping reads would be generated versus 40 million H3K4me3. However, the effect of the individual, flowcell batch and library batch covariates is stronger for H3K27ac than H3K4me3. Overall, the technical covariates are more closely clustered in H3K27ac than in H3K4me (Figure 3.11).

In the `Rail-RNA` re-analysis, the figure did not show differences by annotation features. Figure 3.10 shows the equivalent annotation breakdown, and while most the picture is similar across annotation features there is a difference in the percent of variance explained by the total mapped reads. This difference is most marked when focusing on the dbPeaks overlapping only known exons and introns.

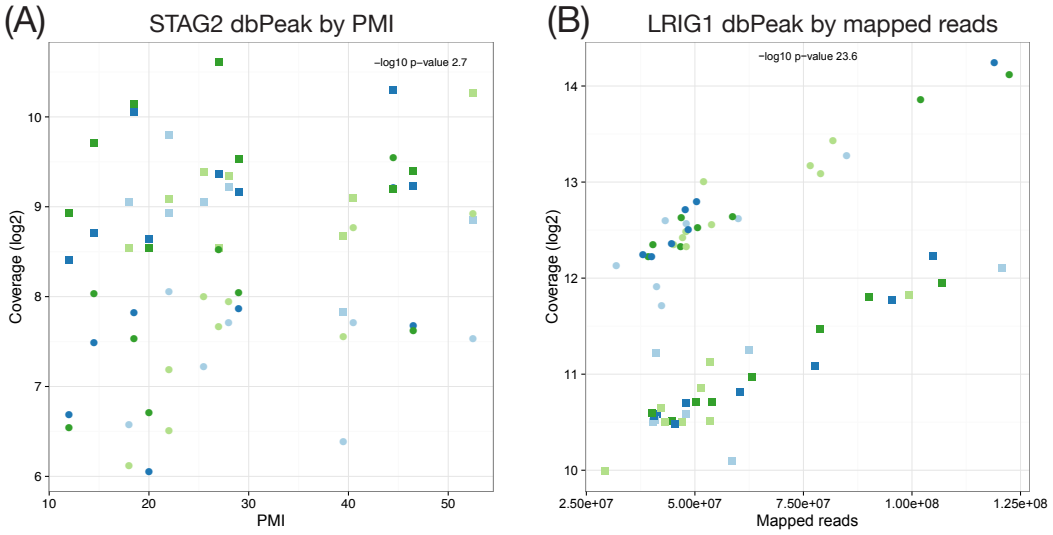


Figure 3.8: Scatterplots between the total coverage ($\log_2$) and un-modeled covariates. (A) H3K4me3 dbPeak from Figure 3.6C has the third strongest association with post-mortem interval (PMI). **(B)** H3K4me3 dbPeak from Figure 3.6D has the fifth strongest association with total mapped reads. The -$\log_1 0$ Bonferroni adjusted p-value for adding PMI or total mapped reads to a model accounting for brain region, cell type and age at death is shown. Colors and shapes are as described in Figure 3.5.

For each dbPeak and each of the 12 un-modeled covariates, we sequentially fit a model testing the contribution of the un-modeled covariate against a null model with brain region, cell type and age at death. We FWER adjusted the resulting p-values and for each of the un-modeled covariates we identified the set of dbPeaks significantly associated with said covariate (FWER < 0.05)

controlling for brain region, cell type and age at death. Figure 3.8 shows two H3K4me3 dbPeaks that are associated with post-mortem interval (PMI, Figure 3.8A) and the total number of mapped reads (Figure 3.8B); the -$\log_1 0$ p-value is shown in each panel. These two dbPeaks overlap *STAP2* associated with T-cells [20] and *LRIG1* that has been linked to inhibition of cancer cell growth [21]. These dbPeaks are associated with cell type as shown in Figure 3.6C and D, respectively. Notably, the cell type association is clear in Figure 3.8B with no apparent interaction with the total number of mapped reads.

3.3.3 Comparison with `DiffBind` derived differentially bound peaks

For each histone mark, we identified peaks using `MACS` [1] and performed the differential binding analysis using `DiffBind` while controlling the FWER at 5% as described in Methods Section 3.5.4. We focused on `DiffBind` since it is the most widely currently used software for differential binding analysis and is one of the best performers for studies with biological replicates [6]. For H3K4me3, 100,326 (70.6%) peaks were present in at least 2 samples (including the input sample) out of 142,102 peaks. The corresponding number of peaks for H3K27ac are 402,932 (71.2%) out of 565,787.

We ran the `DiffBind` analysis twice: once without restricting the analysis to peak summits, and the second with 500 base-pair summits. The summit windows significantly associated with cell type (FWER < 0.05) overlap 91.7% and 92.95% of the time the differentially bound peaks as identified without restricting `DiffBind`'s analysis to the peak summits for H3K4me3 and H3K27ac respectively. The inverse is similar (92.16% and 95.17%), thus the `DiffBind`

results are highly consistent regardless of how consensus peaks are derived.

We explored whether the candidate dbPeaks derived from `derfinder` and `DiffBind` overlapped or not, alternating which set was used as the query (reference) and which one was used as the target as shown in Table 3.2. For H3K4me3 90.88% of the `DiffBind` consensus peaks derived without using summits (wide peaks) overlap at least one candidate dbPeak from `derfinder` (94.17% for H3K27ac). Thus most `DiffBind` consensus peaks are tested for differential binding in both methods. However, the inverse is not true as only 47.15% of the candidate H3K4me3 `derfinder` dbPeaks overlap a peak tested with `DiffBind` (42.12% for H3K27ac). These percents drop of when comparing against the peak summits as expected by the fact that the peak summits are mostly 500 base-pairs long.

When considering the significant dbPeaks (FWER < 0.05) from both methods, only 30.33% of `DiffBind`'s wide significant dbPeaks for H3K4me3 overlap `derfinder`'s significant dbPeaks (37.56% for H3k27ac). This can be explained by the fact that `DiffBind` labels as significant 69.4% and 76.8% of the consensus wide peaks while `derfinder` determined that 5.7% and 4.6% of the candidate peaks were significantly differentially bound for H3K4me3 and H3K27ac. Thus `derfinder` is more conservative in which candidate peaks are identified as differentially bound. The reciprocal comparison shows that 76.53% and 75.23% of `derfinder`'s dbPeaks overlap a `DiffBind` wide dbPeak for these histone marks. These percents drop when comparing against `DiffBind`'s summit dbPeaks.

We created 1 kilo base-pairs (kb) non-overlapping windows along the

Table 3.2: Percent overlap between dbPeaks identified with `derfinder` and `DiffBind`. Percent of dbPeaks that overlap betweeen strategies using all candidate dbPeaks or only significant dbPeaks at FWER $< 5\%$. The query set determines which list of dbPeaks is the reference.

Histone mark	Query	Target	Significant	`DiffBind` no summits	`DiffBind` with summits
H3K4me3	`DiffBind`	`derfinder`	No	90.88	88.71
			Yes	30.33	31.26
	`derfinder`	`DiffBind`	No	47.15	33.53
			Yes	76.53	71.18
H3K27ac	`DiffBind`	`derfinder`	No	94.17	91.46
			Yes	37.56	39.89
	`derfinder`	`DiffBind`	No	42.12	23.61
			Yes	75.23	58.22

Table 3.3: Percent of 1 kb windows of the genome overlapping differentially bound peaks. Percent of genome windows (1 kb each) that overlap at least one dbPeak using `derfinder` or `DiffBind`. All candidates dbPeaks are shown first, then the dbPeaks that are significant at FWER $< 5\%$.

Histone mark	Significant	`derfinder`	`DiffBind` no summits	`DiffBind` with summits
H3K4me3	No	8.95	6.6	4.57
	Yes	2.56	4.68	3.25
H3K27ac	No	39.89	25.98	16.47
	Yes	15.6	21.42	13

genome and calculated the percent of them that overlap a candidate dbPeak as well as dbPeaks from all three methods as shown in Table 3.3. The results show that `derfinder` considers candidate peaks in a larger percent of these 1 kb windows than `DiffBind`. When considering only the significant dbPeaks, the percent of 1 kb windows overlapping `derfinder` dbPeaks is comparable to the percent of windows with `DiffBind`'s summit dbPeaks and smaller than `DiffBind`'s wide peaks. This is result is consistent with the widening effect from merging peaks and the property of `derfinder`'s dbPeaks that all base-pairs have differential binding signal.

3.4 Conclusions

Here we illustrated how `derfinder` [10] can be used to identify differentially bound peaks (dbPeaks) using ChIP-seq data. This strategy for identifying dbPeaks resolves the widening effect produced by merging peaks and the limitations of focusing the analysis on peak summits. The dbPeaks identified with `derfinder` have signal in all base-pairs which might not be the case for

peaks derived from the currently available methods. This could lead to a reduction in false positives in downstream analyses by avoiding regions of the genome with low differential binding signal.

Using data from the *EpiMap* study [14] we showed how to identify dbPeaks for more than 2 groups. We identified dbPeaks for histone marks H3K4me3 and H3K27ac based on 8 groups given by brain region, cell type and age at time of death. These dbPeaks are mostly associated with cell type (NeuN negative or positive) as it explains most of the variation in these peaks followed by technical covariates. The overall residual variation is smaller than in the *Geuvadis* RNA-seq experiment [18] due to how the dbPeaks are selected versus using all genes. In a comparison with `DiffBind` [3] we showed that `derfinder` is more conservative yet `derfinder` leads to similar percent of the genome overlapping dbPeaks when compared against the merging strategy that focuses on peak summits.

`derfinder` is flexible, annotation-agnostic and can be used for a wide variety of models, including time-course analyses and for more than 2 conditions. `derfinder` can be used for both sharp and wide ChIP enrichment and defines the regions based on the data at hand, but could be used for pre-defined regions thus making it more versatile than most available methods for differential binding analysis [6]. The core `derfinder` functionality is similar for RNA-seq and ChIP-seq analyses, which could be an advantage for users analyzing both types of data sources. With the changes we made to `derfinder` for ChIP-seq data we believe that it can be used for determining regions that have differential signal between two or more conditions with genomic assays

such as ATAC-seq, ChIP-exo, DNase-seq [22], among others. Note that no input samples are needed for `derfinder` and that it can be used with `bigWig` coverage files instead of `BAM` files.

3.5 Methods

3.5.1 Changes in `derfinder` for ChIP-seq data

In order for `derfinder` [10] to allow smoothing of the F-statistics, we changed the `findRegions()` code. The new version of this function has additional parameters to control how to perform the smoothing. By default, when smoothing is used in `derfinder`, the smoothing is performed with the function `locfitByCluster()` from the `bumphunter` package [23]. We recommend setting the `minNum` and `minInSpan` arguments to the read length and the `bpSpan` argument to the expected peak size. Note that smoothing is disabled by default for backward compatibility with RNA-seq users.

3.5.2 Identification of dbPeaks from the *EpiMap* study with `derfinder`

We downloaded the BAM files and sample phenotype data provided by *EpiMap* [14]. We adjusted the coverage of each sample based on the total number of mapped reads to libraries of 80 million base-pairs. For each histone mark, we calculated F-statistics for the base-pairs that had at least one sample with coverage greater or equal to 10 reads. The F-statistics are derived from an intercept-only null model (3.1) for where y_{ij} corresponds to the $\log_2$ adjusted

coverage for base-pair i of sample j with an offset factor of 32.

$$y_{ij} = \alpha_i + \epsilon_{ij} \tag{3.1}$$

For the alternative model (3.2), we used a model with covariates for the brain region (reference: ACC), cell type (reference: NeuN-), and age at time of death.

$$y_{ij} = \alpha_i + \beta_{i1}\text{BrainRegion}_j + \beta_{i2}\text{CellType}_j + \beta_{i3}\text{AgeAtDeath}_j + \epsilon_{ij} \tag{3.2}$$

The F-statistics were smoothed with `locfitByCluster()` with arguments `minNum` = 100, `minInSpan` = 100 and `bpSpan` = 300. The global F-statistic cutoff used corresponds to a p-value of 0.01, candidate peaks were clustered using `maxClusterGap` = 3000, and a total of 100 permutations were used to determine family-wise error rate (FWER) adjusted p-values. A cutoff of 0.05 was used to determine the differentially bound peaks. Results were first explored with reports created with `regionReport` [24].

3.5.3 Analysis of dbPeaks identified with `derfinder`

We determined the overlap between significant dbPeaks for each histone mark and Ensembl v75 [15] features using `derfinder` [10] as shown in Figure 3.4.

For each histone mark, we calculated the total coverage divided by the read length for each significant dbPeak in each sample, resulting in a count matrix (dbPeaks by samples). We $\log_2$-transformed this matrix (after adding

an offset of 1) and selected only the dbPeaks with a FWER adjusted p-value less than 0.05.

For each significant dbPeak we calculated the Bonferroni adjusted p-value from adding brain region, cell type, or age at death (continuous) as a covariate to an intercept-only model using the $\log_2$-transformed matrix for the corresponding histone mark. The dbPeaks that associated (FWER < 0.05) with these three covariates are shown in Figure 3.9 and were used in three separate principal component analysis (PCA) as shown in Figure 3.5. Venn diagrams were created by modifying code from `limma` [25].

Similarly, for each significant dbPeak we calculated the Bonferroni adjusted p-value from adding one of the 12 other covariates shown in Figure 3.7 to a model with the three main covariates. The resulting Bonferroni adjusted p-values were used for clustering these 12 un-modeled covariates (Figure 3.11). We made coverage plots with `derfinderPlot` [13] and scatterplots with `ggplot2` [26] for the top 50 dbPeaks that are associated (FWER < 0.05) with each of the 15 covariates. Some of them are highlighted in Figures 3.6 and 3.8.

We also performed a joint model with all 15 covariates using the $\log_2$-transformed matrix for each significant dbPeak. We calculated the percent of variance explained by each covariate and summarized the results in boxplots as displayed in Figure 3.7.

3.5.4 Identification of dbPeaks with `DiffBind`

Peaks were called with `MACS` version 2.1.0 [1] for each sample using the corresponding input sample for the cell type analyzed. `MACS` was used with

arguments `--tsize` = 100, `-bw` = 230. For each histone mark we merged the peaks using `DiffBind` [3] with argument `minOverlap` = 2 and using the input sample for the corresponding cell type. Differential binding between cell types was determined using `DESeq2` [7] as implemented in `DiffBind`. All consensus peaks were retrieved using the argument `th` = 1 in the `dba.report()` function. The resulting p-values were Bonferroni adjusted to control the FWER and a cutoff of 0.05 was used to determine the differentially bound peaks. For the `DiffBind` analysis that controlled the widening effect produced by merging peaks, we used the argument `summits` = 250 in the function `dba.count()`.

Genome tiles were created using using `GenomicRanges` [27] and this same package was used to overlap `derfinder` and `DiffBind` results.

3.6 Competing interests

The authors declare that they have no competing interests.

3.7 Funding

JTL was supported by NIH Grant 1R01GM105705, LCT was supported by Consejo Nacional de Ciencia y Tecnología México 351535.

3.8 Author's contributions

LCT wrote the software under the supervision of JTL and AEJ. LCT analyzed the data with the supervision of JTL and AEJ. All authors contributed to writing the paper.

3.9 Supplementary Results

Figure 3.9 shows whether the significant dbPeaks for each histone mark are associated (FWER < 0.05) with brain region, cell type, or age at death. Notably, 89.2% and 94.5% of the dbPeaks are only associated with cell type for H3K4me3 and H3K27ac.

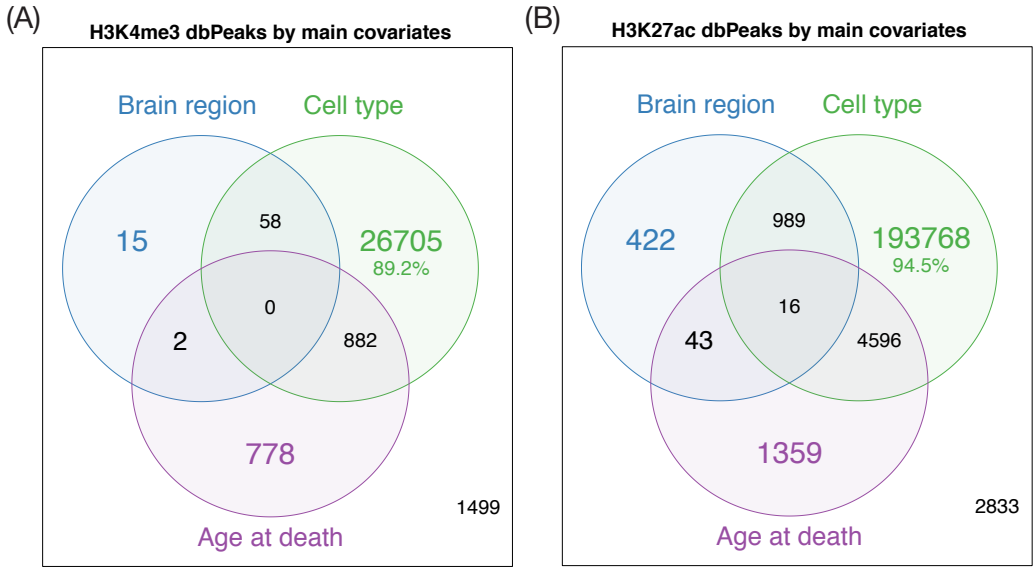


Figure 3.9: Differentially bound peaks for H3K4me3 and H3K27ac marks classified by the modeled covariates. (A) H3K4me3 and **(B)** H3K27ac dbPeaks. All dbPeaks by covariate were used in Figure 3.5.

Figure 3.10 shows the decomposition of Figure 3.7 by Ensembl v75 features shown on Figure 3.4. Interestingly, the dbPeaks that only overlap exonic and intronic sequences (Figure 3.10, row 4) show an increase percent of variation explained by the total mapped reads.

Figure 3.11 shows the relationship between the 12 un-modeled covariates

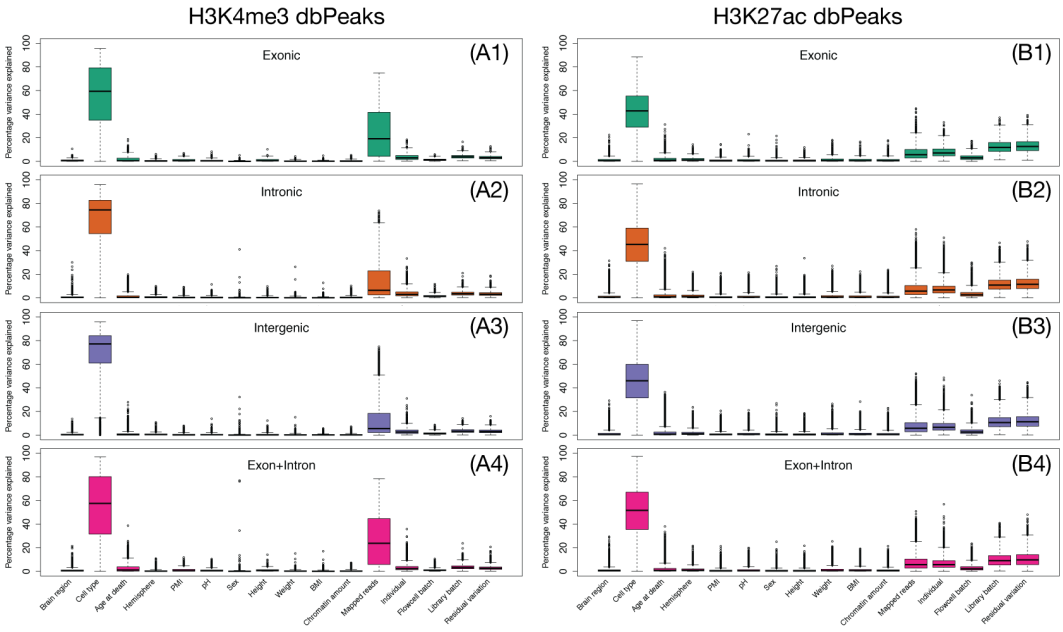


Figure 3.10: Boxplots of percentage of variation explained by the 3 modeled covariates, 12 other covariates, and residual variation by annotation. (A) H3K4me3 and **(B)** H3K27ac dbPeaks overlapping Ensembl v75 strictly exonic (**1**), strictly intronic (**2**), strictly intergenic (**3**), or exonic and intronic (**4**) features. The number of dbPeaks per annotation feature are as given in Figure 3.4.

when controlling for brain region, cell type and age at death as clustered by correlation between the -$\log_1 0$ Bonferroni adjusted p-values. For both marks, height, weight, sex and BMI are clustered together as well as individual id and library batch. The technical covariates are overall more closely related in H3K27ac than in H3K4me3.

3.10 Supplementary Methods

The following paragraph summarizes the sample information from the *EpiMap* study [14].

Samples for the *EpiMap* study come from the National Institute of Mental

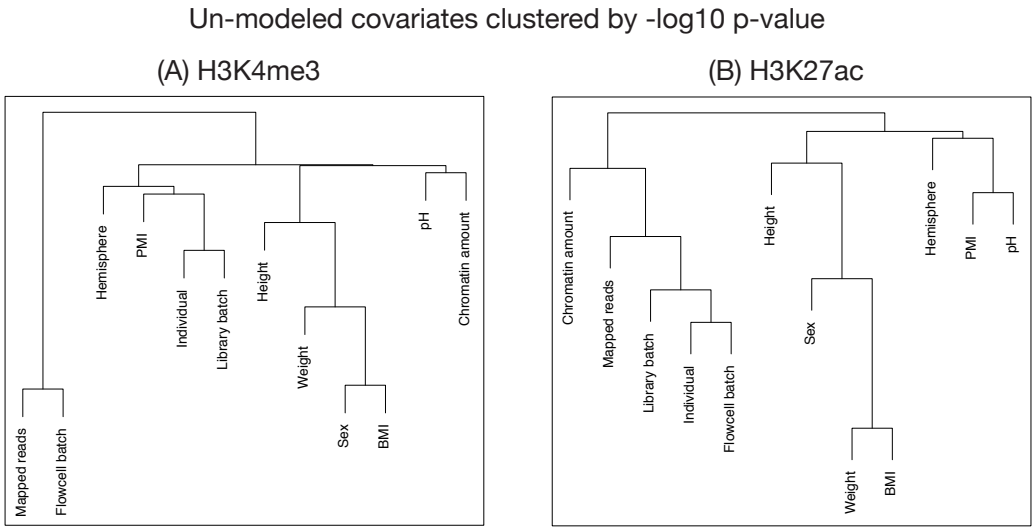


Figure 3.11: Hierarchical clustering of the 12 un-modeled covariates. Clustering of the -$\log_1 0$ Bonferroni adjusted p-values for the 12 un-modeled covariates compared sequentially to a model with brain region, cell type and age shown for **(A)** H3K4me3 and **(B)** H3K27ac dbPeaks.

Health (NIMH) Human Brain Collection Core (HBCC) http://www.nimh.nih.gov/labs-at-nimh/research-areas/research-support-services/hbcc/human-brain-collection-core-hbcc.shtml. The human brain specimens were collected in the Section on Neuropathology of the Clinical Brain Disorders Branch at NIMH. Samples were dissected at the HBCC and shipped to the Ichan School of Medicine - Mt Sinai for sample preparation. Samples for the *EpiMap* study were dissected from a combination of right and left hemisphere of fresh frozen coronal slabs cut at autopsy from the dorsolateral prefrontal cortex (DLPFC) and the anterior cingulate cortex (ACC) from Brodmann areas 9_10 and 24_32 respectively. For nuclei isolation the mouse monoclonal antibody (clone A60) against neuronal marker NeuN (Millipore, MAB377X) was used. Immuno-tagging with NeuN antibody conjugated to

AlexaFluor 488 allowed for sorting of the nuclei into 2 fractions: NeuN+ (neuronal) nuclei and NeuN- (non-neuronal) nuclei, using fluorescence-activated cell sorting (FACS). Chromatin immunoprecipitation (ChIP) assays for histone marks H3K4me3 and H3K27ac were carried out using Native ChIP. Micrococal Nuclease (MNase) (Sigma, N3755) treatment was used to digest chromatin into mononucleosomes. The following antibodies were used for chromatin pull-down: anti-H3K4me3 (Cell Signaling, Cat # 9751BC, lot 7) and anti-H3K27ac (Active Motif, Cat# 39133, Lot # 01613007). Agilent Bioanalyzer, Qubit concentration measurement and pQCR were used to quality control the ChIP results. For sequencing, libraries were prepared in batches of 8 samples using KAPA Hyper Prep Kit and BIOO Scientific Adapters. After each step, DNA was purified using AMPure beads (SPRI select) and final library size selection (200-350 bp) was performed using Pippin Prep. Libraries were barcoded based on the sequencing randomization scheme to allow for multiplexing. The presence of the main library product (275 bp) and the absence of adapter dimer (125 bp) was confirmed using Agilent Bioanalyzer as a quality control step. Libraries were sequenced with the goal that 40 millions of uniquely mapped paired end reads for H3K4me3 and 80 millions of uniquely mapped paired end reads for H3K27ac. Therefore, samples were sequenced in batches of 8 (for H3K4me3) or 4 (for H3K27ac) per lane of the Illumina flow cell. A pool of 4 or 8 barcoded libraries were layered on a random selection of one of the eight lanes of the Illumina flow cell. One-hundred base pair paired-end reads were obtained on a HiSeq 2500 in the Mount Sinai Genomics core facility. Paired FASTQ files were aligned to the Human Genome (HG19) with BWA mem (version 0.7.8). Picard (version 1.112) MarkDups was used

to mark duplicates in the bam fies and multi-mapped reads and improperly paired reads were filtered out with `samtools view -f 2 -F 2828 -q 1` (version 1.1).

References

[1] Y. Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome Biology* 9.9 (2008). PMID: 18798982 PMCID: PMC2592715, R137. ISSN: 1465-6914. DOI: 10.1186/gb-2008-9-9-r137.

[2] R. Thomas et al. “Features that define the best ChIP-seq peak calling algorithms”. In: *Briefings in Bioinformatics* (2016). PMID: 27169896, bbw035. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbw035.

[3] R. Stark and G. Brown. *DiffBind: differential binding analysis of ChIP-Seq peak data*. Version 2.0.0. 2011. URL: http://bioconductor.org/packages/DiffBind.

[4] C. S. Ross-Innes et al. “Differential oestrogen receptor binding is associated with clinical outcome in breast cancer”. In: *Nature* 481.7381 (2012). PMID: 22217937 PMCID: PMC3272464, pp. 389–393. ISSN: 1476-4687. DOI: 10.1038/nature10730.

[5] L. Shen et al. “diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates”. In: *PloS One* 8.6 (2013). PMID: 23762400 PMCID: PMC3677880, e65598. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0065598.

[6] S. Steinhauser et al. “A comprehensive comparison of tools for differential ChIP-seq analysis”. In: *Briefings in Bioinformatics* (2016). PMID: 26764273. ISSN: 1477-4054. DOI: 10.1093/bib/bbv110.

[7] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. eng. In: *Genome Biology* 15.12 (2014). PMID: 25516281 PMCID: PMC4302049, p. 550. ISSN: 1465-6914. DOI: 10.1186/s13059-014-0550-8.

[8] A. T. L. Lun and G. K. Smyth. "csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows". In: *Nucleic Acids Research* 44.5 (2016). PMID: 26578583 PMCID: PMC4797262, e45. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1191.

[9] H. P. Shulha et al. "Coordinated cell type-specific epigenetic remodeling in prefrontal cortex begins before birth and continues into early adulthood". In: *PLoS genetics* 9.4 (2013). PMID: 23593028 PMCID: PMC3623761, e1003433. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003433.

[10] L. Collado-Torres et al. "Flexible expressed region analysis for RNA-seq with derfinder". In: *bioRxiv* (2016), p. 015370. DOI: 10.1101/015370.

[11] G. Brown. *GreyListChIP: Grey Lists – Mask Artefact Regions Based on ChIP Inputs*. Version 1.4.0. 2015. URL: http://bioconductor.org/packages/GreyListChIP.

[12] A. C. Frazee et al. "Differential expression analysis of RNA-seq data at single-base resolution". In: *Biostatistics* 15 (2014). PMID: 24398039, pp. 413–426. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxt053.

[13] L. Collado-Torres, A. E. Jaffe, and J. T. Leek. *derfinderPlot: Plotting functions for derfinder*. Version 1.6.0. 2015. URL: http://www.bioconductor.org/packages/derfinderPlot.

[14] S. Akbarian and P. Sklar. *Cis-Regulatory Epigenome Mappings in Schizophrenia*. U01MH103392. EpiMap. 2016.

[15] F. Cunningham et al. "Ensembl 2015". In: *Nucleic Acids Research* 43.Database issue (2015). PMID: 25352552 PMCID: PMC4383879, pp. D662–669. ISSN: 1362-4962. DOI: 10.1093/nar/gku1010.

[16] A. Gazou et al. "Xq22.3-q23 deletion including ACSL4 in a patient with intellectual disability". In: *American Journal of Medical Genetics. Part A* 161A.4 (2013). PMID: 23520119, pp. 860–864. ISSN: 1552-4833. DOI: 10.1002/ajmg.a.35778.

[17] Y.-P. Liu et al. "Ca(2+)-dependent reduction of glutamate aspartate transporter GLAST expression in astrocytes by P2X(7) receptor-mediated phosphoinositide 3-kinase signaling". In: *Journal of Neurochemistry* 113.1 (2010). PMID: 20070863, pp. 213–227. ISSN: 1471-4159. DOI: 10.1111/j.1471-4159.2010.06589.x.

[18] P. A. C. 't Hoen et al. "Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories". In: *Nature Biotechnology* 31.11 (2013). PMID: 24037425, pp. 1015–1022. ISSN: 1546-1696. DOI: 10.1038/nbt.2702.

[19] A. Nellore et al. "Rail-RNA: Scalable analysis of RNA-seq splicing and coverage". In: *bioRxiv* (2015), p. 019067. DOI: 10.1101/019067.

[20] Y. Sekine et al. "Signal-transducing adaptor protein-2 regulates stromal cell-derived factor-1 alpha-induced chemotaxis in T cells". In: *Journal of Immunology (Baltimore, Md.: 1950)* 183.12 (2009). PMID: 19933863, pp. 7966–7974. ISSN: 1550-6606. DOI: 10.4049/jimmunol.0902096.

[21] S. Goldoni et al. "A soluble ectodomain of LRIG1 inhibits cancer cell growth by attenuating basal and ligand-dependent EGFR activity". In: *Oncogene* 26.3 (2007). PMID: 16847455, pp. 368–381. ISSN: 0950-9232. DOI: 10.1038/sj.onc.1209803.

[22] S. Mahony and B. F. Pugh. "Protein-DNA binding in high-resolution". In: *Critical Reviews in Biochemistry and Molecular Biology* 50.4 (2015). PMID: 26038153 PMCID: PMC4580520, pp. 269–283. ISSN: 1549-7798. DOI: 10.3109/10409238.2015.1051505.

[23] A. E. Jaffe et al. "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies". In: *International journal of epidemiology* 41.1 (2012). PMID: 22422453 PMCID: PMC3304533, pp. 200–209. DOI: 10.1093/ije/dyr238.

[24] L. Collado-Torres, A. E. Jaffe, and J. T. Leek. "regionReport: Interactive reports for region-level and feature-level genomic analyses [version2; referees: 2 approved, 1 approved with reservations]". In: *F1000Research* 4 (2016), pp. 1–10. DOI: 10.12688/f1000research.6379.2.

[25] M. E. Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015). PMID: 25605792 PMCID: PMC4402510, e47.

[26] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN: 978-0-387-98140-6. URL: http://ggplot2.org.

[27] M. Lawrence et al. "Software for Computing and Annotating Genomic Ranges". In: *PLoS Computational Biology* 9 (8 2013). PMID: 23950696 PMCID: PMC3738458, e1003118. DOI: 10.1371/journal.pcbi.1003118.

Chapter 4

`regionReport`: Interactive reports for region-level and feature-level genomic analyses

Leonardo Collado-Torres[1,2,3], Andrew E. Jaffe[1,2,3,4], Jeffrey T. Leek[1,2,†].

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
2. Center for Computational Biology, Johns Hopkins University
3. Lieber Institute for Brain Development, Johns Hopkins Medical Campus
4. Department of Mental Health, Johns Hopkins University

† *corresponding author;* `jtleek@gmail.com`

4.1 Abstract

`regionReport` is an `R` package for generating detailed interactive reports from region-level genomic analyses as well as feature-level RNA-seq results. The

reports include quality-control checks, an overview of the results, an interactive table of the genomic regions or features of interest and reproducibility information. `regionReport` provides specialized reports for exploring `DESeq2`, `edgeR` or `derfinder` differential expression analyses results. `regionReport` is also flexible and can easily be expanded with report templates for other analysis pipelines.

Keywords: Report, Interactive, Reproducibility, Genomics, Sequencing, ChIP-seq, RNA-seq, Methylation, Software.

4.2 Introduction

Many analyses of genomic data result in regions along the genome that associate with a covariate of interest. These genomic regions can result from identifying differentially bound peaks from ChIP-seq data [1], identifying differentially methylated regions (DMRs) from DNA methylation data [2], performing base-resolution differential expression analyses using RNA sequencing data [3, 4], among other analysis pipelines. The genomic regions themselves are commonly stored in a `GRanges` object from `GenomicRanges` [5] when working with R or the BED file format on the UCSC Genome Browser [6]. Other information on these regions, for example summary statistics on the magnitude of effects and statistical significance, also provide useful information and can be stored as metadata in `GRanges` objects. The usage of `R` in genomics is increasingly common due to the usefulness and popularity of the Bioconductor project [7], and in the latest release (version 3.3) 300 unique packages use `GenomicRanges` for many workflows, demonstrating the widespread

utility of identifying and summarizing characteristics of genomic regions.

Bioconductor is particularly strong for differential expression analyses, with 206 packages using the `Differential Expression` BiocView. RNA-seq data is commonly used to perform feature-level analyses at either the transcript, gene or exon levels with Bioconductor packages `DESeq2` [8] and `edgeR` [9, 10, 11], among others. The features can also be expressed regions identified in an annotation-agnostic procedure by `derfinder` [3]. In an exploratory data analysis of `DESeq2` or `edgeR` results it is common to create a set of plots in order to identify potentially problematic samples or features. For example, in such a exploratory analysis it is common to use a dimension reduction technique such as principal component analysis to determine if samples are clustering by group or another variable of interest. This type of plot is useful for detecting artifacts, such as mislabeling of samples.

Here we introduce `regionReport` which allows users to explore genomic regions of interest, `derfinder`, `DESeq2`, and `edgeR` results through interactive stand-alone HTML reports that can be shared with collaborators. These reports are flexible enough to display plots and quality control checks within a given experiment, but can easily be expanded to include custom visualizations or text describing the main conclusions of the exploratory analysis. The resulting HTML report emphasizes reproducibility of analyses [12] by including all the `R` code without obstructing the resulting plots and tables. Alternatively, static PDF reports can be generated and easily shared among collaborators. We envision `regionReport` will provide a useful tool for exploring and sharing genomic region-based, `DESeq2` and `edgeR` results from high

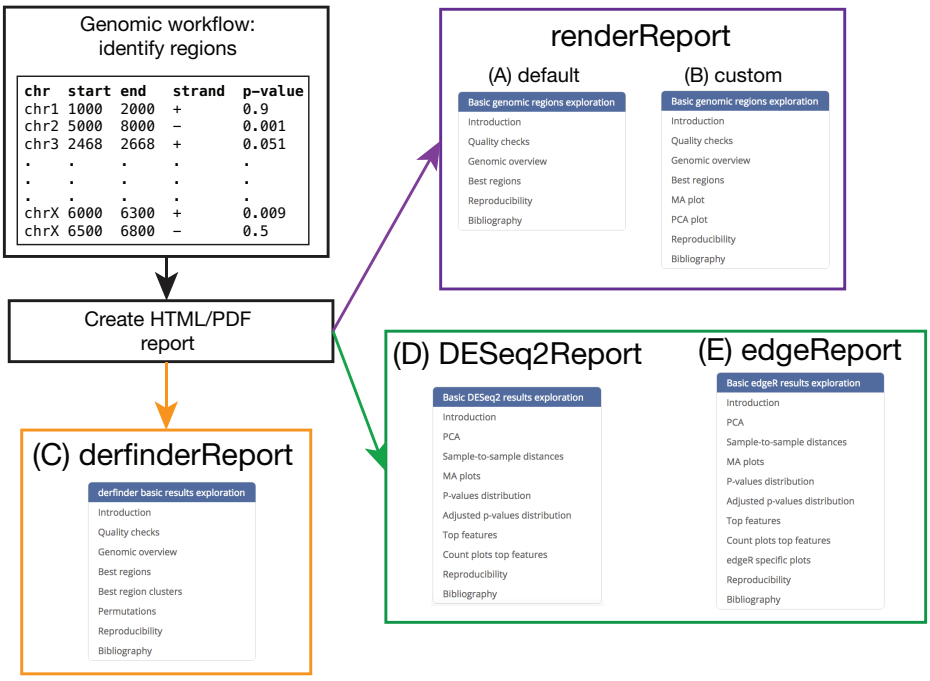


Figure 4.1: `regionReport` overview. Example region input, the appropriate `regionReport` function to use, and menu of the resulting report for: (**A**) the general use case, (**B**) a customized report, (**C**) `derfinder` results, (**D**) `DESeq2` results and (**E**) `edgeR` results.

throughput genomics experiments.

4.3 Methods

4.3.1 Implementation

The package includes R Markdown templates which are processed using `rmarkdown` [13] and `knitr` [14] to produce HTML or PDF reports. HTML reports can be styled using `knitrBootstrap` [15] or with `rmarkdown` templates that include interactive features. The `regionReport` package generates a report that includes a series of plots for checking the quality of the results and an interactive table of with the best regions or features. Each element of the report

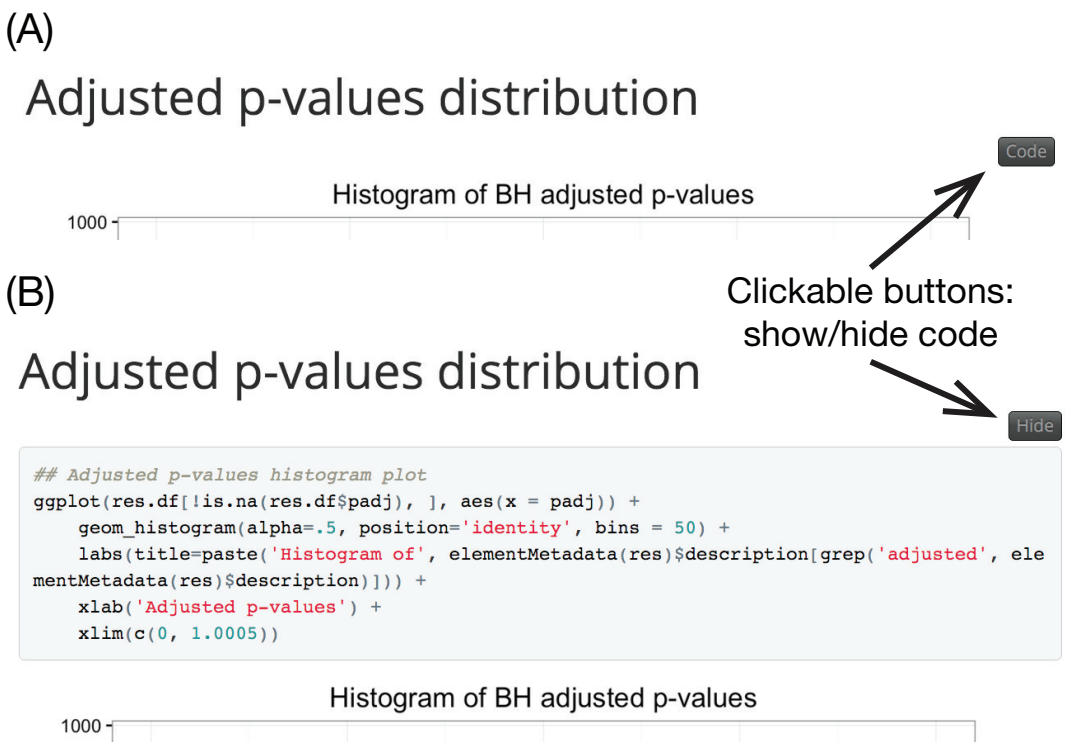


Figure 4.2: Interactively display the code for each table/figure in the report. (**A**) View by default and (**B**) after clicking on the "code" toggle for a section in the report. The HTML reports include a toggle to hide/show all the R code.

has a brief explanation, although actual interpretation of the results is dataset- and workflow-dependent. To facilitate navigation a menu is included, which is useful for users interested in a particular section of the report. Figure 4.1A shows the menu of the general report for a set of regions with associated p-values. The code for each plot or table is hidden by default and can be shown by clicking on the "code" button as shown in Figure 4.2. Further customization of the reports can be done by providing custom code, changing the default plots, or by modifying the R Markdown templates included in `regionReport`.

4.3.2 General region report

4.3.2.1 Quality checks

This section of the report includes a variety of quality control steps which help the user determine whether the results are sensible. The quality control steps explore:

- P-values, Q-values, and FWER adjusted p-values
- Region width
- Region area: sum of single-base level statistics (if available)
- Mean coverage or other score variables (if available)

A combination of density plots and numerical summaries are used in these quality checks. If there are statistically significant regions, the distributions are compared between all regions and the significant ones. For example, the distribution region widths might have a high density of small values for the global results, but shifted towards higher values for the subset of significant regions as shown in Figure 4.3.

4.3.2.2 Genomic overview

The report includes plots to visualize the location of all the regions as well as the significant ones. Differences between them can reveal location biases. The nearest known annotation feature for each region is summarized and visually inspected in the report. This type of plot can be useful to quickly check whether significant regions are concentrated in a chromosome or in an

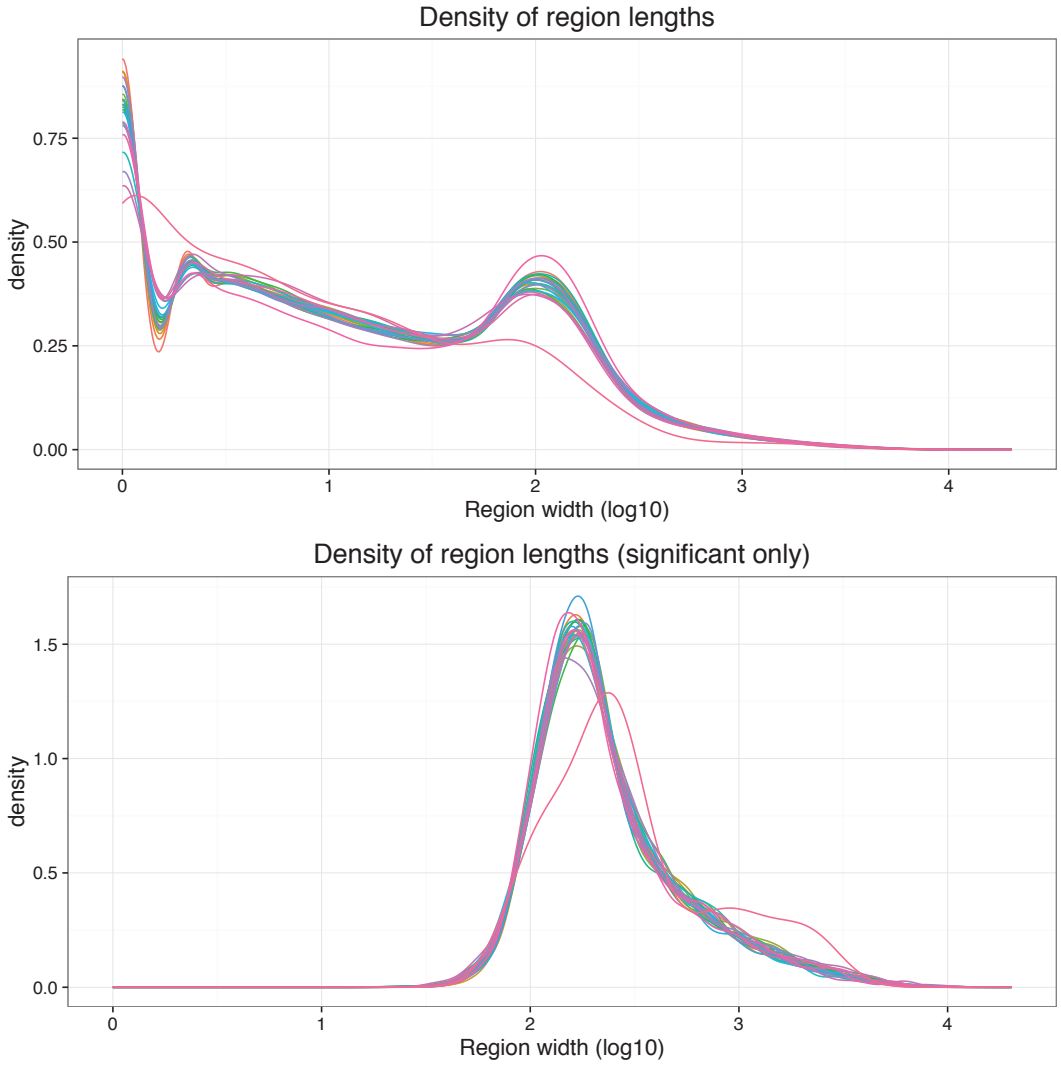


Figure 4.3: Distribution of region widths for all regions in the `derfinder` use case example with the *BrainSpan* dataset. The top figure shows the region width distribution for all regions while the bottom one shows it only for the significant regions. One line is is shown per chromosome in each of the plots.

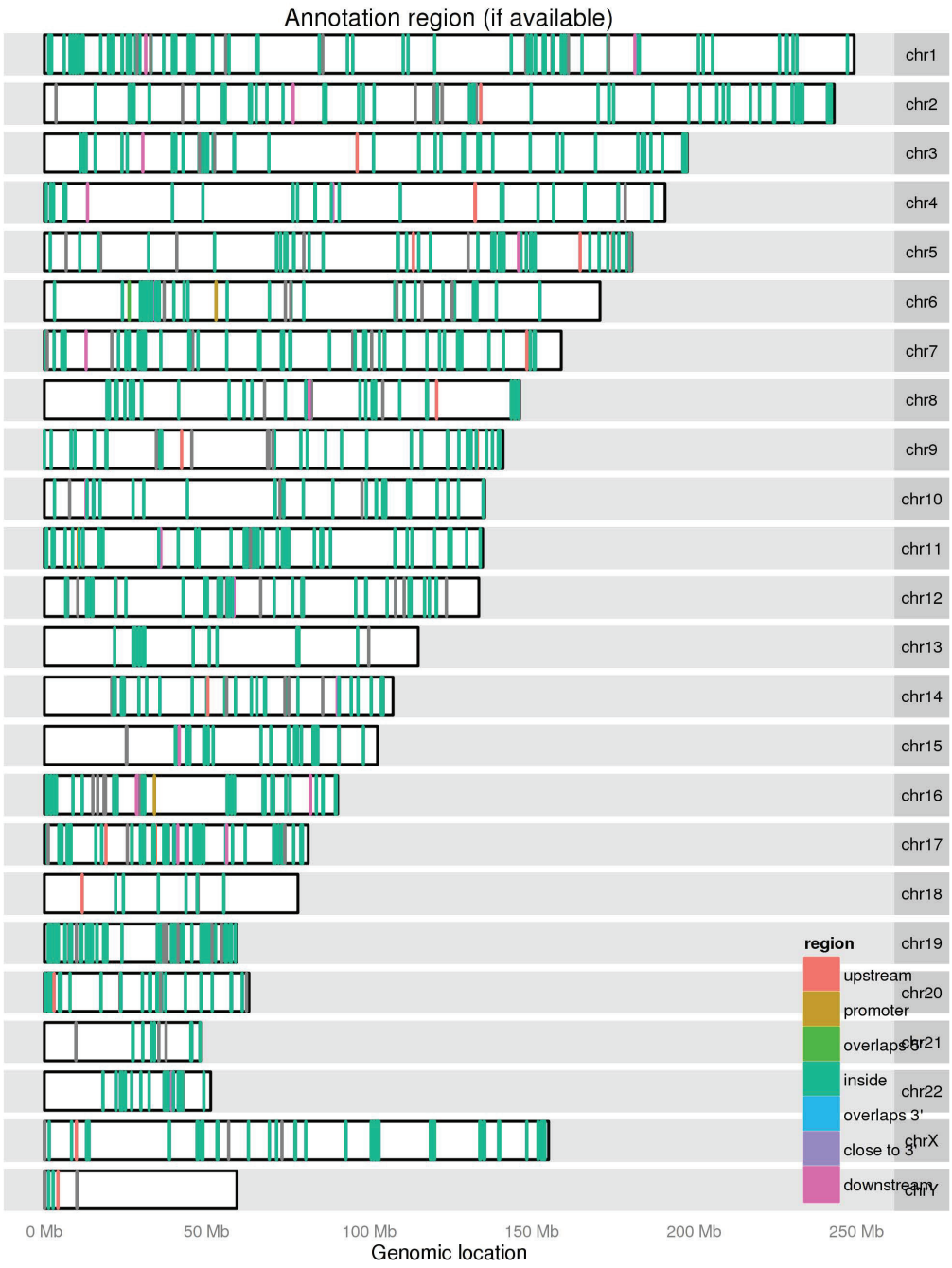


Figure 4.4: Genomic overview of the annotation type for the significant regions in the `derfinder` use case example with the *Hippo* dataset.

annotation type. For example, Figure 4.4 shows the annotation information for the significant regions with most regions contained inside genes, which is expected with RNA-seq data.

4.3.2.3 Best regions

An interactive table with the top regions (500 by default) is included in this section as shown in Figure 4.5A. This allows the user to sort the region information according to their preferred ranking option. For example, lowest p-value, longest width, chromosome, nearest annotation feature, etc. The table also allows the user to search and subset it interactively as shown in Figure 4.5B. A common use case is when the user wants to check if any of the regions are near a known gene of their interest.

4.3.2.4 Reproducibility

At the end of the report, detailed information is provided on how the analysis was performed. This includes the actual function call to generate the report, the path where the report was generated, time spent, and the detailed `R` session information including package versions of all the dependencies. An example is shown in Figure 4.6 with the `R` package information truncated.

The `R` code for generating the plots and tables in the report is included in the report itself, thus allowing users to manually reproduce any section of the report, customize them, or simply change the graphical parameters to their liking.

(A)

Show 10 entries Search:

seqnames	start	end	width	strand	area	value	cluster	L	clusterL	nar
chr1	2316	2631	316	*	-1.58	15.81	2	10	29.00	DDX1
chr2	451	551	101	*	1.59	4.77	3	3	20.00	FAM1
chr1	456	526	71	*	1.07	3.20	1	3	29.00	DDX1
chr1	2176	2211	36	*	0.78	1.57	2	2	29.00	DDX1
chr1	2841	2841	1	*	1.20	1.20	2	1	29.00	DDX1

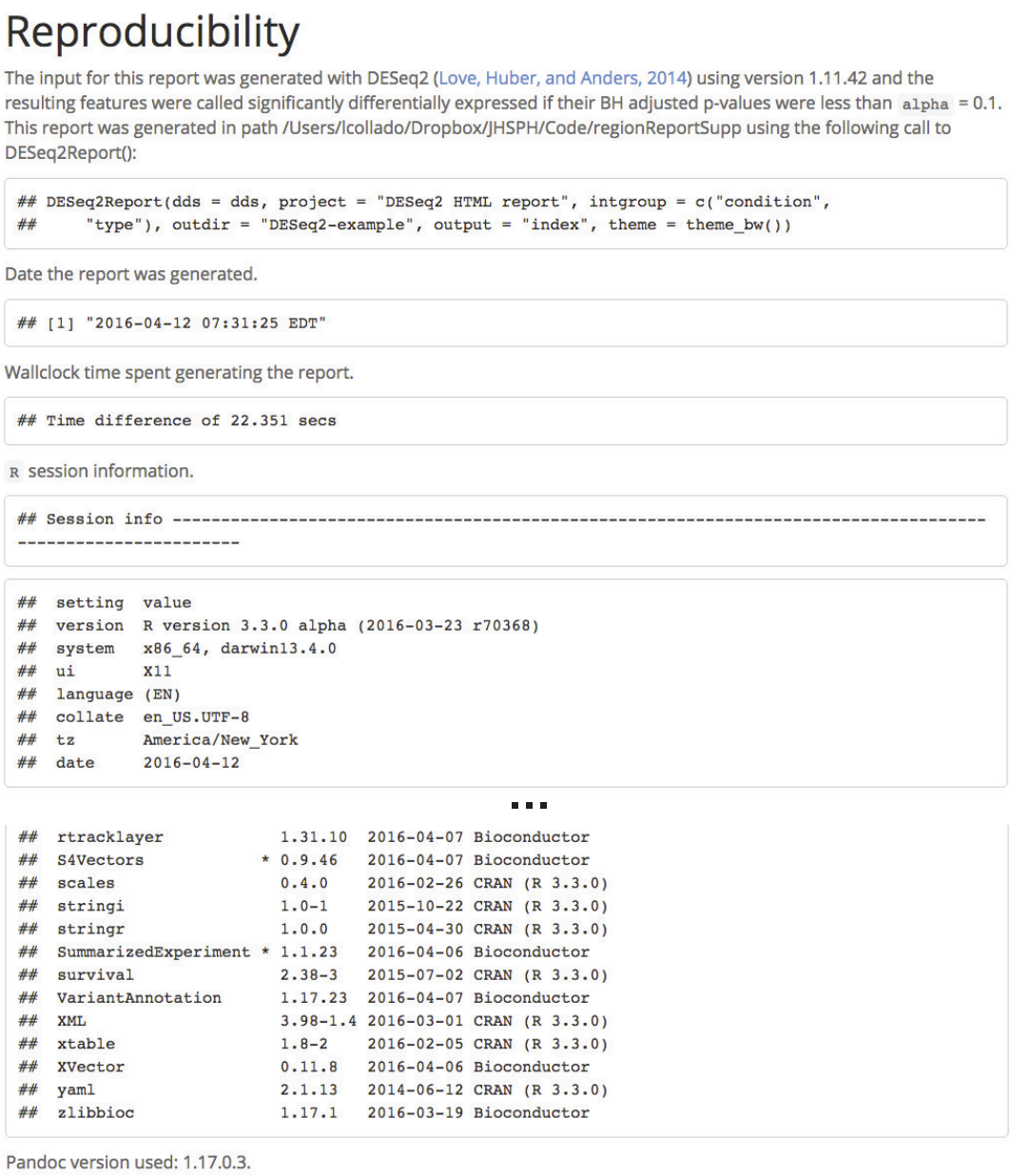
(B)

Show 10 entries Search: NR_046018

value	cluster	L	clusterL	name	annotation	description	region	distance	s
15.81	2	10	29.00	DDX11L1	NR_046018	upstream	upstream	9243	
3.20	1	3	29.00	DDX11L1	NR_046018	upstream	upstream	11348	
1.57	2	2	29.00	DDX11L1	NR_046018	upstream	upstream	9663	
1.20	2	1	29.00	DDX11L1	NR_046018	upstream	upstream	9033	
0.78	1	1	29.00	DDX11L1	NR_046018	upstream	upstream	11103	
0.77	2	1	29.00	DDX11L1	NR_046018	upstream	upstream	9628	
0.62	1	1	29.00	DDX11L1	NR_046018	upstream	upstream	11243	
0.61	1	1	29.00	DDX11L1	NR_046018	upstream	upstream	11873	
0.58	2	1	29.00	DDX11L1	NR_046018	upstream	upstream	8963	
0.52	2	1	29.00	DDX11L1	NR_046018	upstream	upstream	9068	

Showing 1 to 10 of 10 entries (filtered from 15 total entries) First Previous 1 Next Last

Figure 4.5: Interactive table with results for the top regions in the general use case example using `bumphunter` results. The interactive table can (**A**) show all the top regions or (**B**) a subset of the results by using the search box. The table can also be sorted by each of the different columns.



Reproducibility

The input for this report was generated with DESeq2 (Love, Huber, and Anders, 2014) using version 1.11.42 and the resulting features were called significantly differentially expressed if their BH adjusted p-values were less than `alpha` = 0.1. This report was generated in path /Users/lcollado/Dropbox/JHSPH/Code/regionReportSupp using the following call to DESeq2Report():

```
## DESeq2Report(dds = dds, project = "DESeq2 HTML report", intgroup = c("condition",
##     "type"), outdir = "DESeq2-example", output = "index", theme = theme_bw())
```

Date the report was generated.

```
## [1] "2016-04-12 07:31:25 EDT"
```

Wallclock time spent generating the report.

```
## Time difference of 22.351 secs
```

`R` session information.

```
## Session info ------------------------------------------------------------------------------------
----------------------
```

```
##  setting  value
##  version  R version 3.3.0 alpha (2016-03-23 r70368)
##  system   x86_64, darwin13.4.0
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2016-04-12
```

▪▪▪

```
##  rtracklayer            1.31.10  2016-04-07 Bioconductor
##  S4Vectors            * 0.9.46   2016-04-07 Bioconductor
##  scales                 0.4.0    2016-02-26 CRAN (R 3.3.0)
##  stringi                1.0-1    2015-10-22 CRAN (R 3.3.0)
##  stringr                1.0.0    2015-04-30 CRAN (R 3.3.0)
##  SummarizedExperiment * 1.1.23   2016-04-06 Bioconductor
##  survival               2.38-3   2015-07-02 CRAN (R 3.3.0)
##  VariantAnnotation      1.17.23  2016-04-07 Bioconductor
##  XML                    3.98-1.4 2016-03-01 CRAN (R 3.3.0)
##  xtable                 1.8-2    2016-02-05 CRAN (R 3.3.0)
##  XVector                0.11.8   2016-04-06 Bioconductor
##  yaml                   2.1.13   2014-06-12 CRAN (R 3.3.0)
##  zlibbioc               1.17.1   2016-03-19 Bioconductor
```

Pandoc version used: 1.17.0.3.

Figure 4.6: Reproducibility section for a report using `DESeq2` results. The reproducibility information includes the actual function call used to generate the report, the path where the report was generated, the time it took to create the report, details about the R session information, and the pandoc version used for rendering the HTML report. For reports based on `DESeq2` results, the version used to perform the differential expression analysis and cutoff used are also displayed. Note that `DESeq2` version used for the analysis and for the report might differ.

4.3.2.5 Customization

`regionReport` allows users to customize the reports to their liking. This can be done in different ways depending on the amount of customization the user is looking for. Several plots are made with `ggplot2` and the user might want to change the default theme, for example to a black and white theme as shown in the function call in Figure 4.6. Another user might be interested in adding code that creates more plots than the ones included by default in the report. For example, the user might be interested in adding a MA and a PCA plot to the default report. This can be done via the `customCode` argument which results in new sections added to the menu as shown in Figure 4.1B compared to Figure 4.1A. Further customization can be achieved by modifying the templates included in `regionReport` and using the `template` argument.

4.3.3 `derfinder` report

When exploring `derfinder` results from the single base-level approach, for each of the best 100 (default) DERs a plot showing the coverage per sample is included in the report. These plots allow the user to visualize the differences identified by `derfinder` along known exons, introns and isoforms. The plots are created using `derfinderPlot` [16].

Due to the intrinsic variability in RNA-seq coverage data or mapping artifacts, in situations where there are two candidate DERs that are relatively close there might be reasons to consider them a single candidate DER and its important to visualize them. This tailored report groups candidate DERs into clusters based on a distance cutoff. After ranking them by their area, for

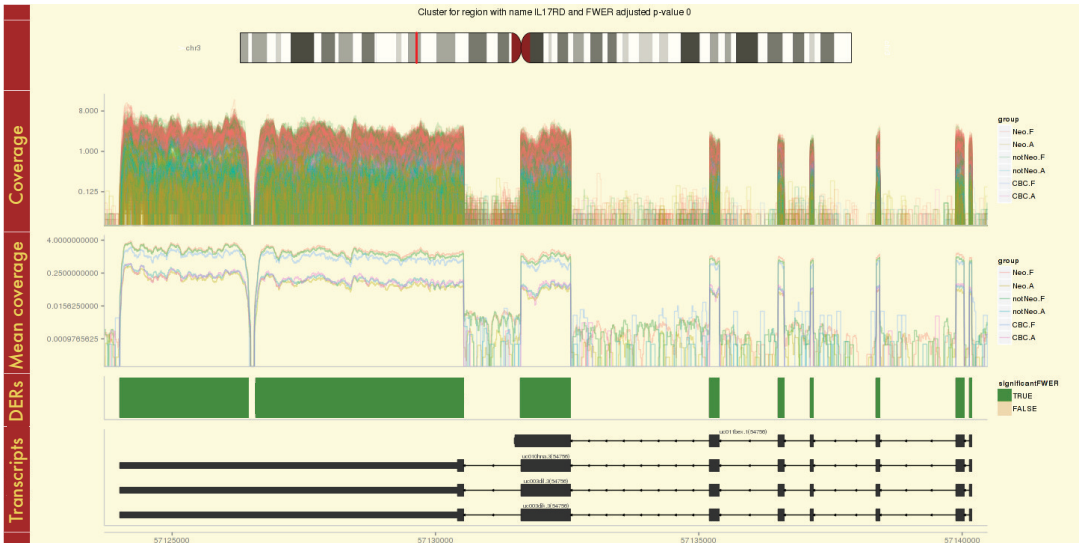
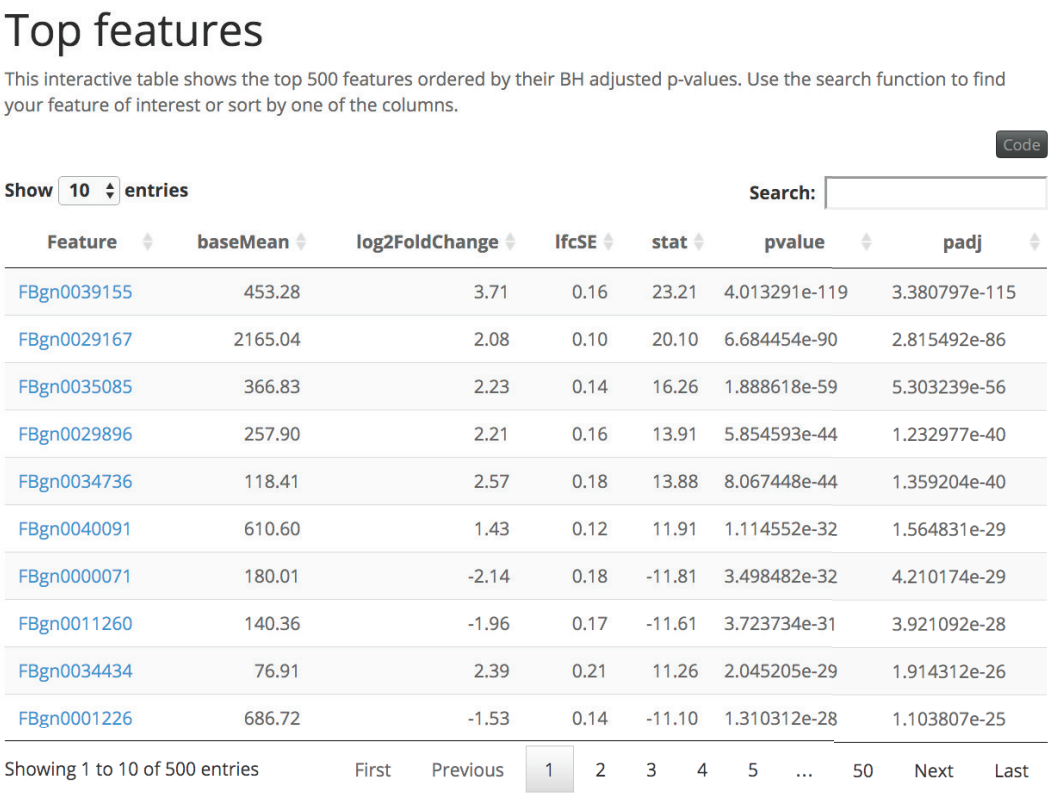


Figure 4.7: Example region cluster plot for the `derfinder` use case example with the *BrainSpan* dataset. Coverage curves are shown for each sample colored by their group membership. Mean coverage curves by group, differentially expressed regions (DERs) and known transcripts are shown in the remaining tracks.

the top 20 (default) clusters it plots tracks with the coverage by sample, the mean coverage by group, the identified candidate DERs colored by whether they are statistically significant, and known alternative transcripts as shown in Figure 4.7. Figure 4.1C shows the main categories of the report generated from a richer region dataset than in the general case.

4.3.4 DESeq2 and edgeR reports

Feature-level differential expression analyses result in a set of features (genes, exons) with a p-value for each feature. To perform such analyses, some phenotype information about the samples is usually available. With this information, you can explore the raw data to identify potentially problematic samples using principal component analysis and sample distance plots. You can also explore



Top features

This interactive table shows the top 500 features ordered by their BH adjusted p-values. Use the search function to find your feature of interest or sort by one of the columns.

Code

Show 10 entries Search:

Feature	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
FBgn0039155	453.28	3.71	0.16	23.21	4.013291e-119	3.380797e-115
FBgn0029167	2165.04	2.08	0.10	20.10	6.684454e-90	2.815492e-86
FBgn0035085	366.83	2.23	0.14	16.26	1.888618e-59	5.303239e-56
FBgn0029896	257.90	2.21	0.16	13.91	5.854593e-44	1.232977e-40
FBgn0034736	118.41	2.57	0.18	13.88	8.067448e-44	1.359204e-40
FBgn0040091	610.60	1.43	0.12	11.91	1.114552e-32	1.564831e-29
FBgn0000071	180.01	-2.14	0.18	-11.81	3.498482e-32	4.210174e-29
FBgn0011260	140.36	-1.96	0.17	-11.61	3.723734e-31	3.921092e-28
FBgn0034434	76.91	2.39	0.21	11.26	2.045205e-29	1.914312e-26
FBgn0001226	686.72	-1.53	0.14	-11.10	1.310312e-28	1.103807e-25

Showing 1 to 10 of 500 entries First Previous 1 2 3 4 5 ... 50 Next Last

Figure 4.8: Interactive table for top features from the `DESeq2` use case example.

the results and check the features marked as differentially expressed with MA plots and a histogram of the p-values distribution. `regionReport` provides a template that allows you to create all these plots easily for `DESeq2` results (Figure 4.1D). It has similar components to the region-level reports such as an interactive table for the top features as shown in Figure 4.8, but also highlights specific exploratory plots for this type of results. `regionReport` can also be used for `edgeR` results (Figure 4.1E) resulting in very similar reports given the internal implementation. The only difference is that reports for `edgeR` results include sections for visualizing the biological coefficient of variation and the multidimensional scaling plot of distances between feature expression profiles. See the use cases for example reports from `DESeq2` and `edgeR` results.

4.3.5 Operation

4.3.5.1 Installation

`regionReport` and required dependencies can be easily installed from Bioconductor with the following commands:

```
source("http://bioconductor.org/biocLite.R")
biocLite("regionReport")
```

4.3.5.2 Input

To generate the report, the user first has to identify the regions of interest according to their analysis workflow. For example, by performing bumphunting to identify DMRs with `bumphunter`. The report is then created using `renderReport()` which is the main function in this package as shown in Figure

4.1A,B.

For the derfinder use case, the derfinderReport() function creates the recommended report that includes visualizations of the coverage information for the best regions and clusters of regions. Similarly DESeq2Report() and edgeReport() create reports for DESeq2 and edgeR results, respectively.

4.3.5.3 Output

A small example can be generated using:

```
example("renderReport", "regionReport", ask=FALSE)
```

The resulting HTML file will open in the users default browser when using R in an interactive session. Note that alternative output formats such as PDF files can also be generated, although they are not as dynamic and interactive as the HTML format.

4.4 Use Cases

The supplementary website contains reports using DiffBind, bumphunter, derfinder, DESeq2, and edgeR results. The derfinder use case is illustrated with datasets described previously [3] which a moderately sized dataset (25 samples) and a large dataset with 484 samples. We encourage you to explore the following example reports:

- general HTML report example using bumphunter results: leekgroup.github.io/regionReportSupp/bumphunter-example/index.html,

- customized general HTML report using `DiffBind` results with histograms instead of density plots: leekgroup.github.io/regionReportSupp/DiffBind-example/index.html,
- `DESeq2` HTML and PDF reports: leekgroup.github.io/regionReportSupp/DESeq2-example/index.html, leekgroup.github.io/regionReportSupp/DESeq2-example/DESeq2Report.pdf,
- `edgeR` HTML and PDF reports using the custom `ggplot2` theme `theme_linedraw()`: leekgroup.github.io/regionReportSupp/edgeR-example/index.html, leekgroup.github.io/regionReportSupp/edgeR-example/edgeReport.pdf,
- `edgeR`-robust HTML report: leekgroup.github.io/regionReportSupp/edgeR-robust-example/index.html,
- HTML report using `derfinder` results with the *BrainSpan* dataset (484 samples) and styled with `knitrBootstrap`: leekgroup.github.io/regionReportSupp/brainspan/basicExploration.html,
- HTML report using `derfinder` results with the *Hippo* dataset (25 samples) and styled with `knitrBootstrap`: leekgroup.github.io/regionReportSupp/hippo/basicExploration.html.

4.5 Summary

`regionReport` creates interactive reports from a set of regions and can be used in a wide range of genomic analyses. Reports generated with `regionReport` can easily be extended to include further quality checks and interpretation of the results specific to the dataset under study. These shareable documents are very powerful when exploring different parameter values of an analysis workflow or applying the same method to a wide variety of datasets. The reports allow users to visually check the quality of the results, explore the properties of the genomic regions under study, and inspect the best regions and interactively explore them.

Furthermore, `regionReport` promotes reproducibility of data exploration and analysis. Each report provides R code that can be used as the starting point for other analyses within a dataset. `regionReport` provides a flexible output for exploring and sharing results from high throughput genomics experiments.

4.6 Software availability

4.6.1 Sofware access

`regionReport` is freely available via Bioconductor at `bioconductor.org/packages/regionReport`.

The supplementary website `leekgroup.github.io/regionReportSupp/` hosts the code and output for generating all the use cases described. Versions of all software used are included in the reports.

4.6.2 Latest source code

The latest source code is available at via GitHub at `github.com/leekgroup/regionReport`. However, we highly recommend users to install `regionReport` directly from Bioconductor at `bioconductor.org/packages/regionReport`.

4.7 Author contributions

LCT conceived and developed the `regionReport` package, supervised by AEJ and JTL. All authors wrote and approved the final manuscript.

4.8 Competing interests

No competing interests were disclosed.

4.9 Grant information

JTL was partially supported by NIH Grant 1R01GM105705, LCT was supported by Consejo Nacional de Ciencia y Tecnología México 351535, AEJ was partially supported by 1R21MH109956.

4.10 Acknowledgements

We would like to acknowledge Michael I. Love for his feedback and input in creating the report specific to `DESeq2` results.

References

[1] R. Stark and G. Brown. *DiffBind: differential binding analysis of ChIP-Seq peak data*. 2011. URL: http://bioconductor.org/packages/DiffBind.

[2] A. E. Jaffe et al. "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies". In: *International journal of epidemiology* 41.1 (2012). PMID: 22422453 PMCID: PMC3304533, pp. 200–209. DOI: 10.1093/ije/dyr238.

[3] L. Collado-Torres et al. "Flexible expressed region analysis for RNA-seq with derfinder". In: *bioRxiv* (2016), p. 015370. DOI: 10.1101/015370.

[4] A. C. Frazee et al. "Differential expression analysis of RNA-seq data at single-base resolution". In: *Biostatistics* (2014). PMID: 24398039 PMCID: PMC4059460, kxt053. DOI: 10.1093/biostatistics/kxt053.

[5] M. Lawrence et al. "Software for computing and annotating genomic ranges". In: *PLoS computational biology* 9.8 (2013). PMID: 23950696 PMCID: PMC3738458, e1003118. DOI: 10.1371/journal.pcbi.1003118.

[6] K. R. Rosenbloom et al. "The UCSC Genome Browser database: 2015 update". In: *Nucleic acids research* 43.D1 (2015). PMID: 25428374 PMCID: PMC4383971, pp. D670–D681. DOI: 10.1093/nar/gku1177.

[7] W. Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature methods* 12.2 (2015). PMID: 25633503 PMCID: PMC4509590, pp. 115–121. DOI: 10.1038/nmeth.3252.

[8] M. I. Love, W. Huber, and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". eng. In: *Genome Biology* 15.12 (2014). PMID: 25516281 PMCID: PMC4302049, p. 550. ISSN: 1465-6914. DOI: 10.1186/s13059-014-0550-8.

[9] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". eng. In: *Bioinformatics (Oxford, England)* 26.1 (2010). PMID: 19910308 PMCID: PMC2796818, pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616.

[10] McCarthy et al. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". In: *Nucleic Acids Research* 40.10 (2012). PMID: 22287627 PMCID: PMC3378882, pp. –9. DOI: 10.1093/nar/gks042.

[11] X. Zhou, H. Lindsay, and M. D. Robinson. "Robustly detecting differential expression in RNA sequencing data using observation weights". eng. In: *Nucleic Acids Research* 42.11 (2014). PMID: 24753412 PMCID: PMC4066750, e91. ISSN: 1362-4962. DOI: 10.1093/nar/gku310.

[12] G. K. Sandve et al. "Ten simple rules for reproducible computational research". In: *PLoS computational biology* 9.10 (2013). PMID: 24204232 PMCID: PMC3812051, e1003285. DOI: 10.1371/journal.pcbi.1003285.

[13] Rstudio. *rmarkdown: Dynamic Documents for R*. 2016. URL: http://rmarkdown.rstudio.com.

[14] Y. Xie. *Dynamic Documents with R and knitr*. CRC Press, 2013.

[15] J. Hester. *knitrBootstrap: Knitr Bootstrap framework*. 2015. URL: https://github.com/jimhester/knitrBootstrap/.

[16] L. Collado-Torres, A. E. Jaffe, and J. T. Leek. *derfinderPlot: Plotting functions for derfinder*. 2015. URL: http://www.bioconductor.org/packages/derfinderPlot.

Chapter 5

Discussion and Conclusion

This thesis demonstrates that the DER Finder approach [1] can be applied to differential expression and differential binding analyses with the `derfinder` package [2] using RNA-seq and ChIP-seq data respectively. `derfinder` provides a complementary option to the feature counting and transcriptome assembly strategies. It has been already used to identify changes in the human brain transcriptome over the human lifespan [3]. `derfinder` also works with alignments from `Rail-RNA` [4] and expect that it will become widely used in the future. In particular, we are already working on creating an updated `ReCount` resource [5] which should be valuable to the research community.

`regionReport` [6] is based on the newest technologies in the R community and is part of a growing community of packages that are designed for the end user to explore results from complicated pipelines. `regionReport` is very flexible and customizable, which we believe will make it more attractive to different groups of users.

All the software we created is regularly tested, well documented and freely available as part of the `Bioconductor` project [7]. This speaks in favor

of the quality of the software and is our latest contribution to this thriving community.

References

[1] A. C. Frazee et al. “Differential expression analysis of RNA-seq data at single-base resolution”. In: *Biostatistics* 15 (2014). PMID: 24398039, pp. 413–426. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxt053.

[2] L. Collado-Torres et al. “Flexible expressed region analysis for RNA-seq with derfinder”. In: *bioRxiv* (2016), p. 015370. DOI: 10.1101/015370.

[3] A. E. Jaffe et al. “Developmental regulation of human cortex transcription and its clinical relevance at single base resolution”. In: *Nature Neuroscience* 18.1 (2015). PMID: 25501035 PMCID: PMC4281298, pp. 154–161. ISSN: 1546-1726. DOI: 10.1038/nn.3898.

[4] A. Nellore et al. “Rail-RNA: Scalable analysis of RNA-seq splicing and coverage”. In: *bioRxiv* (2015), p. 019067. DOI: 10.1101/019067.

[5] A. C. Frazee, B. Langmead, and J. T. Leek. “ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets”. In: *BMC bioinformatics* 12 (2011). PMID: 22087737 PMCID: PMC3229291, p. 449. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-449.

[6] L. Collado-Torres, A. E. Jaffe, and J. T. Leek. “regionReport: Interactive reports for region-level and feature-level genomic analyses [version2; referees: 2 approved, 1 approved with reservations]”. In: *F1000Research* 4 (2016), pp. 1–10. DOI: 10.12688/f1000research.6379.2.

[7] R. C. Gentleman et al. “Bioconductor: Open software development for computational biology and bioinformatics”. In: *Genome Biology* 5 (2004), R80.

Leonardo Collado-Torres

615 N. Wolfe Street, Room E3032 – 21205-2179 – United States

+1 (410) 955 0958 • lcollado@jhu.edu
http://lcolladotor.github.io/about.html • fellgernon
lcolladotor

Education

Johns Hopkins Bloomberg School of Public Health — **Baltimore, US**
PhD in Biostatistics — *2011-present*

National Autonomous University of Mexico (UNAM) — **Cuernavaca, MX**
Bachelor in Genomic Sciences (LCG), Grade 9.71/10 — *2005-2009*

ITESM Campus Cuernavaca — **Cuernavaca, MX**
High school, Grade 97.8/100 — *2002-2005*

PhD thesis

Title: *Annotation-Agnostic Differential Expression and Binding Analyses*.

Advisors: **Jeffrey T. Leek** and **Andrew E. Jaffe**.

Description: The goal is to develop statistical methods and software that enable researchers to differentiate the sources of variation observed in RNA-seq while minimizing the dependance on known annotation. This will allow researchers to correct for technological variation and study the biological variation driving their phenotype of interest. Then apply these methods to further our understanding of neuropsychiatric disorders using the Lieber Institute for Brain Development human brains collection ($>$ 1000 samples).

Honors and awards

2011: Awarded CONACyT Mexico scholarship for PhD studies outside Mexico.

2009: *Summa Cum Laude* for bachelor in Genomic Sciences studies at LCG-UNAM.

2005: Best high school average ($\sim$ 200 students): awarded ITESM system 90% scholarship for college studies, declined to join LCG-UNAM.

Experience

Industry

Winter Genomics **Cuernavaca, MX**

Scientific executive *2009-2011*

Responsible for recruiting and hiring new personnel, overseeing and supervising bioinformaticians, training new employees, writing research reports and presenting them to colleagues, and organizing all scientific projects.

- First scientific staff member at Winter Genomics;
- Projects completed:
  - de novo genome assembly simulations,
  - assembly and annotation of the *phiVC8* bacteriophage genome,
  - integrated analysis of more than 20 RNA-seq samples for determination of transcription initiation in *Escherichia coli* reported in Gama-Castro et al., PMID 21051347,
  - de novo assembly of four *Escherichia coli* strains and lead to Aguilar et al., PMID 22884033;
- Designed training material for new employees.

Research

Enrique Morett lab **IBT-UNAM, Cuernavaca, MX**

Bioinformatician *2009–2011*

Identified transcriptions start sites and transcription units in *Escherichia coli* and *Geobacter sulfurreducens* with RNA-seq data. Developed the *BacterialTranscription* R package.

Guillermo Dávila lab **CCG-UNAM, Cuernavaca, MX**

Undergraduate research assistant *2007–2009*

Determined bacteriophage ecological groups by developing a method based on codon distribution of all phage sequenced genomes. Joint work with Sur Herrera Paredes.

Roberto Kolter lab **Harvard, Boston, US**

Undergraduate research assistant *2007*

Supervisor: Elizabeth Shank. Carried out screenings to identify bacteria that activate the production of exopolysaccharide through the activation of the gene tasA in *Bacillus subtilis*.

Publications

Peer-reviewed

1. **Collado-Torres L**, Jaffe AE and Leek JT. regionReport: Interactive reports for region-level and feature-level genomic analyses [version2; referees: 2 approved, 1 approved with reservations]. *F1000Research* 2016, 4:105. doi: 10.12688/f1000research.6379.2.
2. Jaffe AE, Shin J, **Collado-Torres L**, Leek JT, et al. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat. Neurosci.* 2015. doi: 10.1038/nn.3898.
3. Shank EA, Klepac-Ceraj V, **Collado-Torres L**, Powers GE, Losick R, Kolter R. Interspecies interactions that result in Bacillus subtilis forming biofilms are mediated mainly by members of its own genus. *Proc. Natl. Acad. Sci.* U.S.A. 2011 Nov;108(48):E1236–1243. doi: 10.1073/pnas.1103630108.

4. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, García-Sotelo JS, López-Fuentes A, Porrón-Sotelo L, Alquicira-Hernández S, Medina-Rivera A, Martínez-Flores I, Alquicira-Hernández K, Martínez-Adame R, Bonavides-Martínez C, Miranda-Ríos J, Huerta AM, Mendoza-Vargas A, **Collado-Torres L**, Taboada B, Vega-Alvarado L, Olvera M, Olvera L, Grande R, Morett E, Collado-Vides J. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.* 2011 Jan;39(Database issue):D98–105. doi: 10.1093/nar/gkq1110.

Pre-prints (unpublished)

1. **Collado-Torres L**, Nellore A, Frazee AC, Wilks C, Love MI, Irizarry RA, Leek JT, Jaffe AE. Flexible expressed region analysis for RNA-seq with derfinder. *bioRxiv* 015370 (2016). doi: 10.1101/015370.
2. Nellore A, Jaffe AE, Fortin JP, Alquicira-Hernández J, **Collado-Torres L**, Wang S, Phillips RA, Karbhari N, Hansen KD, Langmead B, Leek JT. Human splicing diversity across the Sequence Read Archive. *bioRxiv* 038224 (2016). doi: 10.1101/038224.
3. Nellore A, **Collado-Torres L**, Jaffe AE, Morton J, Pritt J, Alquicira-Hernández J, Leek JT, Langmead B. Rail-RNA: Scalable analysis of RNA-seq splicing and coverage. *bioRxiv* 019067 (2015). doi: 10.1101/019067.

Pre-prints (published)

1. **Collado-Torres L**, Jaffe AE, Leek JT. regionReport: Interactive reports for region-based analyses. *bioRxiv* 016659 (2015). doi: 10.1101/016659.

Books

1. Frazee AC, **Collado-Torres L**, Jaffe AE, Langmead B, Leek JT. Measurement, Summary, and Methodological Variation in RNA-sequencing in Statistical Analysis of Next Generation Sequencing Data, *Springer*, 2014, 115-128.

Public profiles

Google Scholar: h57-MykAAAAJ

ORCID: 0000-0003-2140-308X

Impactstory: 0000-0003-2140-308X

GitHub: lcolladotor

Twitter: fellgernon

SlideShare: lcolladotor

LinkedIn: lcollado

Epernicus: lc40

Professional service

Develop and maintain open-source biostatistical software.

Peer review

Bioinformatics: Since 2015

Biostatistics: Since 2013

Professional memberships

2015-2016: American Statistical Association

2014-2016: ENAR student member

2014: American Public Health Association

Presentations

Talks at conferences

2016: **Collado-Torres L**, et al. Annotation-agnostic differential expression analysis, *ENAR*, Austin – US. (slides)

2015: **Collado-Torres L**, Frazee AC, Love MI, Irizarry RA, Jaffe AE, Leek JT. Annotation-agnostic differential expression analysis, *Genomics and Bioinformatics Symposium*, Center for Computational Genomics, Hopkins, Baltimore – US.

2015: Jaffe AE, Shin J, **Collado-Torres L**, Leek JT, et al. Dissecting human brain development at high resolution using RNA-seq, *ENAR*, Miami – US. (slides)

2014: Jaffe AE, Shin J, **Collado-Torres L**, Leek JT, et al. Developmental regulation of human cortex transcription at base-pair resolution, *is3b*: 1st International Summer Symposium on Systems Biology, INMEGEN, Mexico City – MX.

2014: **Collado-Torres L**, Frazee AC, Love MI, Irizarry RA, Jaffe AE, Leek JT. Fast differential expression analysis annotation-agnostic across groups with biological replicates, LCG 10 year anniversary, LCG-UNAM, Cuernavaca – MX.

2013: **Collado-Torres L**, Frazee AC, Irizarry RA, Jaffe AE, Leek JT. Differential expression analysis of RNA-seq data at base-pair resolution in multiple biological replicates, *useR2013*, Albacete – Spain.

2010: **Collado-Torres L**, Reyes-Quiroz A, Cuéllar-Partida G, Moreno-Mayar V, Vargas-Chávez C, Collado-Vides J. BacterialTranscription: a R package to identify Transcription Start Sites and Transcription Units, *Bioconductor Developer Meeting*, EMBL, Heidelberg – Germany.

Posters

2015: **Collado-Torres L**, Frazee AC, Love MI, Irizarry RA, Jaffe AE, Leek JT. Annotation-agnostic RNA-seq differential expression analysis software, *ASHG2015* and *IDIES2015*, Baltimore – US.

2014: **Collado-Torres L**, Frazee AC, Love MI, Irizarry RA, Jaffe AE, Leek JT. Fast annotation-agnostic differential expression analysis, *ENAR* and *Delta Omega Poster Competition (JHBSPH)*, Baltimore – US.

2013: **Collado-Torres L**, Jaffe AE, Leek JT. Fast annotation-agnostic differential expression analysis, *Genomics and Bioinformatics Symposium*, Center for Computational Genomics, Hopkins, Baltimore – US.

2010: **Collado-Torres L**, Reyes-Quiroz A, Cuéllar-Partida A, Moreno-Mayar V, Taboada B, Vega-Alvarado L, Jiménez-Jacinto V, Mendoza-Vargas A, Grande R, Olvera L, Olvera M, Vargas-Chávez C, Júarez K, Collado-Vides J, Morett E. Global Analysis of Transcription Start Sites and Transcription Units in Bacterial Genomes, *From Functional Genomics to Systems Biology*, EMBL, Heidelberg – Germany.

2010: **Collado-Torres L**, Reyes-Quiroz A, Cuéllar-Partida A, Moreno-Mayar V, Taboada B, Vega-Alvarado L, Jiménez-Jacinto V, Mendoza-Vargas A, Grande R, Olvera L, Olvera M, Vargas-Chávez C, Júarez K, Collado-Vides J, Morett E. Global Analysis of Transcription Start Sites and Transcription Units in Bacterial Genomes, *BioC2010*, FHCRC, Seattle – US.

Other talks

2015: dbFinder, *Joint Genomic Meeting*, JHBSPH, Baltimore – US.

2015: Easy parallel computing with BiocParallel and HTML reports with knitrBootstrap, *Biostatistics Computing Club*, JHBSPH, Baltimore – US.

2015: Does mapping simulated RNA-seq reads provide information?, *Joint Genomic Meeting*, JHBSPH, Baltimore – US.

2014: Git for research, *Biostatistics Computing Club*, JHBSPH, Baltimore – US.

2013: Introduction to ggbio, *Genomics for Students*, JHBSPH, Baltimore – US.

2013: Introduction to knitr, *Biostatistics Computing Club*, JHBSPH, Baltimore – US.

2013: Introduction to High-Throughput Sequencing and RNA-seq, *Genomics for Students*, JHBSPH, Baltimore – US.

2012: DEXSeq paper discussion, *Genomics for Students*, JHBSPH, Baltimore – US.

2012: Introduction to R and Biostatistics, LCG-UNAM via Skype.

2012: Introducing Git while making your academic webpage, *Biostatistics Computing Club*, JHBSPH, Baltimore – US.

2011: Introducing Biostatistics to first year LCG students, LCG-UNAM via Skype.

2010: Introduction to using Bioconductor for High Throughput Sequencing Analysis, *National Bioinformatics Week*, CCG-UNAM, Cuernavaca – MX.

2009: Bacteriophages: analyzing their diversity, *LCG third generation symposium*, CCG-UNAM, Cuernavaca – MX.

Courses and Meetings Attendance

2016: *ENAR*, Austin – US.

2015: *ENAR*, Miami – US.

2014: *is3b*, INMEGEN, Mexico City – MX.

2014: *BioC2014*, Harvard, Boston – US.

2014: *IDIES2014*, Johns Hopkins University, Baltimore – US.

2014: *ENAR*, Baltimore – US.

2014: *Delta Omega Poster Competition*, Johns Hopkins University, Baltimore – US.

2014: LCG 10 year anniversary, LCG-UNAM, Cuernavaca – MX.

2013: *Genomics and Bioinformatics Symposium*, Johns Hopkins University, Baltimore – US.

2013: *useR2013*, Albacete – Spain.

2011: *BioC2011*, FHCRC, Seattle – US.

2010: *From Functional Genomics to Systems Biology*, EMBL, Heidelberg – Germany.

2010: *BioC2010*, FHCRC, Seattle – US.

2009: *BioC2009*, FHCRC, Seattle – US.

2009: Course on Oral Communication taught by the master Rafael Popoca, CCG-UNAM, Cuernavaca – MX.

2008: *BioC2008*, FHCRC, Seattle – US.

2008: *A Short R/Bioconductor Course* by James Bullard from UC Berkeley, LCG-UNAM, Cuernavaca – MX.

2007: *Boston Bacterial Meeting*, Boston – US.

2007: *Retreat of the Department of Microbiology and Molecular Genetics - Harvard*, Boston – US.

2006: *Winter School in Genomics*, CCG-UNAM, Cuernavaca – MX.

2005: *HUGO 2005*, Kyoto – Japan.

Software

Bioconductor – main author

2014: derfinder: Annotation-agnostic differential expression analysis of RNA-seq data at base-pair resolution via the DER Finder approach – 5745 downloads.

2014: derfinderPlot: plotting functions for derfinder results – 4477 downloads.

2014: regionReport: Generate HTML or PDF reports for a set of genomic regions or DESeq2/edgeR results – 4254 downloads.

2014: derfinderHelper: helper functions for derfinder package – 5168 downloads.

2014: derfinderData: data for derfinder examples – 1112 downloads.

Bioconductor – contributor role

2015: bumphunter

2014: ballgown

Other R packages

2016: recount: Explore and download data from the recount project.

2014: enrichedRanges: identify enrichment between two sets of genomic ranges.

2014: dots: simplify function calls.

2013: fitbitR: visualize your FitBit data.

2011: BacterialTranscription: identify TSSs and TUs from RNA-seq data.

shiny applications

2014–2016: MPH capstone TA office hours sign up

2014: Simple mortgage calculator

Miscellaneous

2016: Updated the JHU thesis template available at GitHub and Overleaf

Computer skills

all-purpose: R *Ranked 152/5661 in the US and 418/53325 worldwide by GitHub Awards as of January 28, 2016. Does not take into account contributions at leekgroup organization.*

statistics: Stata

scripting: bash, Perl

markup: LaTeX, markdown

OS: Linux

cluster queue: Sun Grid Engine

Teaching Experience

Instructor

- PDCB-UNAM, Cuernavaca, MX
  2011: Invited instructor for the course *Introduction to R and Biostatistics* (website) ~ 10 enrollment.

  2010: *Analysis of High-Throughput Sequencing data with Bioconductor* for Biomedical Sciences PhD Program students (website) ~ 10 enrollment.
- CCG-UNAM, Cuernavaca, MX
  2010: *Introduction to Using Bioconductor for High-Throughput Sequencing Analysis* practice lab at the *National Bioinformatics Week* ~ 40 enrollment.
- IBT-UNAM, Cuernavaca, MX
  2010: *Introduction to R and plotting with R* course for Morett's lab ~ 10 enrollment.

  2010: Organized and gave a lecture for the course on *Statistical Methods and Analysis of Genomic Data* (website) ~ 20 enrollment.

  2009: Organized the course *Introduction to Bioinformatics* for Morett's lab and served as instructor for the *Introduction to R and plotting with R* module (website) ~ 10 enrollment.
- LCG-UNAM, Cuernavaca, MX
  2009: *Seminar III: R/Bioconductor*. In-depth Bioconductor course (website) ~ 30 enrollment.

Guest lecturer

- JHBSPH, Baltimore, US
  2015: *Introduction to R for Public Health Researchers: Reproducible research module* ~ 20 enrollment.
- LCG-UNAM, Cuernavaca, MX
  2012: *Introduction to R and Biostatistics* lecture for *Seminar 1: Introduction to Bioinformatics* course ~ 30 enrollment.

  2011: *Introduction to R and Biostatistics* lecture for *Seminar 1: Introduction to Bioinformatics* course ~ 30 enrollment.

Lead teaching assistant

- JHBSPH, Baltimore, US
  2015–2016: *Statistical Methods in Public Health II* ~ 550 enrollment.

  2014–2015: *Statistical Methods in Public Health I* and *II* ~ 550 enrollment.

Teaching assistant

- JHBSPH, Baltimore, US
  2014–2016: *MPH capstone project*: 30 min one-on-one consulting sessions (biostatistics, Stata coding) ~ 500 enrollment. Develop and maintain the MPHcapstoneTA shiny application.

  2015–2016: *Statistical Methods in Public Health I* ~ 550 enrollment.

  2015: *Introduction to R for Public Health Researchers* ~ 20 enrollment.

  2013–2014: *Statistical Methods in Public Health I* and *II* ~ 550 enrollment.

  2012–2013: *Statistical Methods in Public Health I, II, III,* and *IV* ~ 550 enrollment.
- LCG-UNAM, Cuernavaca, MX
  2009: *Principles of Statistics*. Basic R (website) ~ 30 enrollment.

  2008: *Bioinformatics and Statistics I*. R and Bioconductor overview (website) ~ 40 enrollment.

Mentoring

2015: Mentored Alquicira-Hernández J, LCG-UNAM student visiting Jeff Leek's group.

2009–2011: Advised and trained 13 LCG-UNAM students and alumni while working at *Winter Genomics*: Riveros-McKay F, Vargas-Chávez C, Dulanto-Acevedo V, Romero-Martínez S, Samaniego-Castruita J, Zepeda-Mendoza L, Vargas-Velázquez A, Noé-González M, Soto Jiménez LM, López Moyado I, Medina-Abarca H., Izquierdo-Rangel E, and Berrocal-Quezada NA.

2009: Trained 3 LCG-UNAM students to take over the R/Bioconductor course: Reyes-Quiroz A, Moreno-Mayar V, and Reyes-López J.

Other

2016: Student representative for the Centennial celebration of the Department of Biostatistics.

2012–2016: Organized *Cultural Mixer* events for the Department of Biostatistics with Amanda Mejia for raising cultural awareness.

2012–2014: Organized the *Genomics for Students* group (website)

2009–2011: Organized a Genomics Journal Club at IBT-UNAM.

2008–2009: Elected class representative for the LCG Academic Committee.

2008–2009: Class representative for Administration Unit for Technology Information committee.

2008: Helped start the National Node of Bioinformatics online forum.

Languages

Native: Spanish

Bilingual: English

Basic: French