Title

A Computational Approach to Exploring the Role of Chromatin Modifiers in Development and Disease

Author

Bondhus, Leroy Martin

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Computational Approach to Exploring the Role of Chromatin Modifiers in Development and Disease

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Human Genetics

by

Leroy Martin Bondhus

2024

ABSTRACT OF THE DISSERTATION

A Computational Approach to Exploring the Role of Chromatin Modifiers in Development and Disease

by

Leroy Martin Bondhus

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2024

Professor Valerie A. Arboleda, Chair

De novo mutations in chromatin modifier genes can lead to a variety of complex developmental syndromes that can have severe consequences for affected patients and their families. In this dissertation we will develop a computational framework for investigating the etiology of this diverse class of disorders, with the underlying motivation being that a deeper and more thorough understanding of the mechanisms underlying these disorders is essential to supporting the development of therapeutics that can improve the quality of life for those affected. In **Chapter 1** we will provide background information essential for understanding the work developed in this dissertation. We will begin this chapter with a rather broad overview of the basic biology that grounds our direction of investigation into chromatin modifier syndromes

and provide some definitions for key concepts. In **Chapter 2** we will then cover in some detail the methods in molecular biology that form the state of the art employed for studying chromatin modifier syndromes. In particular we will look at the various functional genomics assays that are used to measure the transcriptomic and epigenomic effects caused by mutations in chromatin modifier genes. Here we will also give a survey of existing computational methods for the analysis of data generated by these molecular biology assays. In this survey we will highlight several critical gaps that exist in current methods of analysis and note how these hinder investigations into the etiology of chromatin modifier syndromes. This will lead us into the subsequent chapters of the dissertation where we develop methods that address these gaps.

In **Chapter 3**, we will look at the gap that exists in our ability to use existing methods to identify the scale of changes over the genome and develop a method for the analysis of differential DNA methylation that addresses this problem. In **Chapter 4** we will look at the limitations of current methods for integrating analysis with the wealth of existing knowledge on the structure of and relationships between biological entities. This limitation we address in our development of a method to weight measures of gene expression specificity based on the similarity structure of the biological entities that compose the underlying sample set. The novel methods that we develop in **Chapter 3** and **Chapter 4** provide a framework for building a more systems level understanding of the molecular pathology of chromatin modifier syndromes that we believe will be essential in the pursuit of effective treatments and therapies for these diverse and complex disorders. To conclude in **Chapter 5**, we will summarize our main results and take a brief prospective look at the direction of the field of research into chromatin modifier syndromes making note of promising directions for future research to expand on the work developed here.

The dissertation of Leroy Martin Bondhus is approved.

Michael F. Carey

Jingyi Li

Bogdan Pasaniuc

Valerie A. Arboleda, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

LIST OF SYMBOLS, ACRONYMS, AND ABBREVIATIONS

Abbreviation	Full Description
ATAC	assay for transposase-accessible chromatin
AUCPR	area under the precision recall curve
bp	base pair
CAGE	cap analysis gene expression
cDNA	coding DNA
ChIP	chromatin immunoprecipitation
CI	confidence interval
CpG	cytosine guanine dinucleotide
DMR	differentially methylated region
DNA	deoxyribonucleic acid
DNAme	DNA methylation
eCDF	emperical cumulative density function
FDR	false discovery rate
FN	false negative
FP	false positive
Gb, Gbp	giga (billion) base pairs
GO	gene ontology
GTEx	gene-tissue expression project
HAT	histone acetyltransferase
HDAC	histone deacetylase
HDM	histone demethylase
HMT	histone methytransferase
IQR	interquartile range
kb, kbp	kilo (thousand) base pairs
lncRNA	long non-coding RNA
Mb, Mbp	mega (million) base pairs
MCC	Matthew's correlation coefficient
MNase	micrococcal nuclease
OR	odds ratio

Abbreviation (cont.)	Full Description (cont.)
PRC	polycomb repressive complex
PRD	polycomb repressive domain
PTM	post-translational modification
QC	quality control
RNA	ribonucleic acid
scRNA	single cell RNA seq
TAD	topologically associating domain
TN	true negative
TP	true positive
TPM	transcripts per million
VUS	variant of unknown significance
XX	XX sex chromosome genotype
XY	XY sex chromosome genotype

ACKNOWLEDGEMENTS

While the pursuit of a PhD is necessarily at times a solitary endeavor, it is also the product of a community of colleagues and mentors. I thank Dr. Valerie Arboleda for her mentorship and guidance through the many facets of the PhD, my family for their ever present support and encouragement, and my friends for making it all rather enjoyable. Thanks also to UCLA and the Human Genetics program for providing me the opportunity and environment in which to pursue this PhD.

Chapter 3 is a version of: Bondhus, L., Wei, A. and Arboleda, V.A., 2022. DMRscaler: a scale-aware method to identify regions of differential DNA methylation spanning basepair to multi-megabase features. BMC bioinformatics, 23(1), p.364. Reproduced with permission from Springer Nature

Chapter 4 is a version of: Bondhus, L., Varma, R., Hernandez, Y. and Arboleda, V.A., 2022. Balancing the transcriptome: leveraging sample similarity to improve measures of gene specificity. Briefings in Bioinformatics, 23(5), p.bbac158.

VITAE

EDUCATION

Institution	Degree	Start Date	Completion Date	Field of study
University of Minnesota - Twin Cities	BS	September 2014	May 2018	Genetics, Cell Biology, and Development (major) Biochemistry (major) Computer Science (minor)

PUBLICATIONS

2024.

1. (in review) **Bondhus L.,** Nava, A., Liu, L., Arboleda, V. 2024. Epigene Functional Diversity: Isoform Usage, Disordered Domain Content, and Variable Binding Partners
2. (in review) Lin, I., Awamleh, Z., Sinvhal, M, **Bondhus, L**., Wei, A., Russell, B., Arboleda, V. 2024. ASXL1 truncating mutations drive shared disruption of Wnt-signaling pathways across diseases but display distinct splicing of RUNX3.

2023

1. Singh, M., Spendlove, S.J., Wei, A., **Bondhus, L.M**., Nava, A.A., de L. Vitorino, F.N., Amano, S., Lee, J., Echeverria, G., Gomez, D. and Garcia, B.A., 2023. KAT6A mutations in Arboleda-Tham syndrome drive epigenetic regulation of posterior HOXC cluster. *Human genetics*, 142(12), pp.1705-1720.
2. (preprint) Nava, A.A., Jops, C.T., Vuong, C.K., Niles-Jensen, S., **Bondhus, L.**, Ong, C.J., de la Torre-Ubieta, L., Gandal, M.J. and Arboleda, V.A., 2023. KAT6A mutations drive transcriptional dysregulation of cell cycle and Autism risk genes in an Arboleda-Tham Syndrome cerebral organoid model. *bioRxiv*, pp.2023-06.

2022

3. **Bondhus, L.**, Wei, A. and Arboleda, V.A., 2022. DMRscaler: a scale-aware method to identify regions of differential DNA methylation spanning basepair to multi-megabase features. *BMC bioinformatics*, 23(1), pp.1-34.
4. **Bondhus, L.**, Varma, R., Hernandez, Y. and Arboleda, V.A., 2022. Balancing the transcriptome: leveraging sample similarity to improve measures of gene specificity. *Briefings in Bioinformatics*.
5. Spendlove, S.J., **Bondhus, L.**, Lluri, G., Sul, J.H. and Arboleda, V.A., 2022. Polygenic risk scores of endo-phenotypes identify the effect of genetic background in congenital heart disease. *Human Genetics and Genomics Advances*, *3*(3), p.100112.

ORAL PRESENTATIONS

1. **Bondhus L***,* Arboleda VA. Leveraging ontologies to query the functional genomic architecture of heart development and congenital heart disease*. NHLBI B2B Annual Meeting*. Arlington, Virginia. Oct 12, 2023.

2. **Bondhus L**, Arboleda VA. DMRscaler: A scale-aware method for differential methylation analysis. *Biological Chemistry Floor Meeting*. Los Angeles, California. February 22, 2021.
3. **Bondhus L**, Arboleda VA. Co-morbidities and Genetic Testing Rates Across Patients with Congenital Malformations. *Genomics + Health: Computational Medicine Meeting*. Los Angeles, California. February 26, 2020.
4. **Bondhus L**, Wei A, Arboleda VA. Developing methods geared towards rare disease cohorts: DMRmini. 2019. *Medical and Population Genomics*. Los Angeles, California. November 20, 2019.

ABSTRACTS

1. **Bondhus L**, Arboleda VA**.** Investigating the relationship between gene expression and phenotype in CHD associated syndromes via knowledge graph embeddings*. NHLBI B2B Annual Meeting*. Arlington, Virginia. Poster Presentation and Short Talk. Oct 12, 2023.
2. **Bondhus L**, Wei A, Arboleda VA. DMRscaler: A Scale-Aware Method to Detect Regions of Differential Methylation Spanning Basepairto Multi-MegabaseFeatures. *NHGRI Annual Training Meeting*. Duke University. Durham, North Carolina. Poster Presentation. April 4, 2022.
3. **Bondhus L**, Wei A, Arboleda VA. DMRscaler: A Scale-Aware Method to Detect Regions of Differential Methylation Spanning Basepairto Multi-MegabaseFeatures. *American Society of Human Genetics*. Virtual Meeting. Poster Presentation. October 18, 2021.
4. **Bondhus L**, Roshni V, Yenifer H, Arboleda VA. Balancing the Transcriptome: Integrating Sample Similarity Information To Improve Measures of Gene Specificity. *Department of Human Genetics Retreat*. Los Angeles, California. Poster Presentation. November 15, 2020.
5. **Bondhus L**, Wei A, Arboleda VA. Dysregulation of DNA Methylation at Homeobox Genes in KAT6A Syndrome. *Department of Human Genetics Retreat*. Los Angeles, California. Poster Presentation. November 7, 2019.

FELLOWSHIPS AND AWARDS

11/2022 - 11/2023	Pediatric Cardiac Genomics Consortium (PCGC) & Cardiovascular Development Data Resource Center (CDDRC) Fellows Program. NIH NHLBI External.
08/2020 – 08/2022	Genome Analysis Training Program. UCLA Intramural. (T32HG002536)
07/2019 – 07/2020	Biomedical Big Data Training Program. UCLA Intramural. (T32LM012424).

SCIENTIFIC COMMUNITY INVOLVEMENT

2022-2024	Served as invited reviewer for papers submitted to *Briefings in Bioinformatics*

CHAPTER 1

INTRODUCTION

Foundations to the study of chromatin modifier syndromes

Context in biology

The genome contains all of the information necessary to guide the development and continued existence of the organism, provided that the environmental conditions to life are met. For complex multicellular organisms such as humans, this implies that the genome must be capable of supporting the differentiation and maintenance of the numerous distinct cell types that compose the organism through each stage of life along with their varied functions and behaviors. Critical to supporting this diversity of cells, functions, and behaviors is the ability of information encoded in the genome to be differentially deployed depending on the context in which it exists.

We can distinguish between two types of context in which genomic information exists. One is the context provided by the genomic sequence itself. For instance, each gene exists somewhere within the sequence of the genome and is regulated with some degree of independence from all of the other genes encoded in the genome. Generalizing this, we can consider information encoded in the genome to be subdivided into distinct regions that can be differentially regulated. We will later explore some of the complexity that arises in trying to identify genomic regions that form the individual units of regulation and make note of how this complexity affects attempts to identify those genomic regions affected in disease. The second type of context in which genomic information exists is the spatial-temporal context in which the genome and the cell are found. Whereas the sequence of the genome is effectively constant over the course of the organism's life, the spatial-temporal context in which the cell and genome exist is dynamic. Temperature and nutrient availability varies, cells migrate and their neighbors change, pathogen exposure and injury occur. We will be particularly interested in the subdomain of spatio-temporal context that covers organismal development and anatomic structure which are in large part products of the genome itself.

These two basic types of context, the sequential and the spatial-temporal, are fundamentally linked. The spatial-temporal context of the cell and genome determines which signals will be present for the cell to respond to while the context encoded in the sequence of the genome broadly determines which regions of the genome will respond to those signals. In development, some information that comes from the spatial-temporal domain must be persistently influential over the genome. That is, cells must retain what is frequently described as a memory of past signals to which they have been exposed. For example, a cell that has differentiated into a neuron as a result of cell extrinsic developmental signals must maintain those gene programs necessary for neuronal function and not turn on those related to other cell identities. To achieve this memory, the secondary physical and chemical structure of chromatin, chromatin being the complex of genomic DNA and histone proteins, must be modified and these modifications maintained.

Chromatin modifiers

The broad class of genes that have a function in modifying chromatin structure, either by chemical modification of DNA or histones, physically repositioning histones, or by introducing higher order structure to chromatin such as loops, are called chromatin modifiers. The proteins encoded by these genes tend to function in multiprotein complexes which contain proteins with domains that have a variety of functional roles such as enzymatic activity to catalyze specific chromatin modifications, DNA or histone binding to target the complex to more or less specific regions of the genome, various adapters that enable binding of the different subunits of the complex, and domains to enable the integration of external signals for modulation of activity (**Figure 1-1**).

Through their joint ability to target specific regions of the genome [1,2], modify chromatin [3], and respond to and integrate signals containing spatial-temporal information [4], chromatin

modifiers serve as an interface between the information encoded in the genome and the information in the environment. We note here that we are using the term "environment" in a broad sense to indicate anything outside the cell itself as opposed to the narrow sense of anything outside the organism. In this sense, concepts such as "the heart environment" can be understood to mean and include the different anatomic domains of the heart, such as atria and ventricles, the pressures of fluids within the heart chambers and their rhythmic variations, the paracrine and juxtacrine signals produced by the cells of the heart, and so on. The normal functioning of chromatin modifiers is particularly critical for development, where signals of body position and the neighboring cell environment must be passed to the genome to ensure that the cell takes on and maintains its correct identity and contributes properly to the morphogenesis of the organism [3,5–8]. The disruptions that occur when these chromatin modifier genes are mutated, which we explore next, exemplify their importance in development.

Germline chromatin modifiers syndromes

Germline mutations in chromatin modifier genes can lead to a wide range of distinct developmental syndromes that are often characterized by effects in multiple body systems (**Figure 1-2**). Between these disorders there is considerable variability in the severity of their phenotypes. Some contribute to the early death of affected individuals, such as with mutations in *ASXL1*, a subunit of polycomb and trithorax chromatin modifying complexes, that lead to Bohring-Opitz syndrome in which patients usually do not survive past childhood [9,10]. Others are fairly benign, such as specific mutations in the chromatin remodeler *SMARCAD1* which can cause individuals to develop without fingerprints in the absence of any other more severe phenotypes [11,12]. Additionally, even within a particular disorder there is often a wide range in the penetrance and expressivity of phenotypes. For instance, many chromatin modifier syndromes are associated with congenital heart disease [13], but within most of these disorders only a subset

of affected individuals will have developed with a congenital heart defect, i.e. variable penetrance of the phenotype. Of those who develop with a heart defect there will often be variability in the severity of the defect that occurs, i.e. variable expressivity of the phenotype [14–17]. A unifying feature of chromatin modifier disorders is that they are expected to manifest primarily from dysregulation of the chromatin modifier's function in regulating chromatin structure. We note, however, that some of these genes have functions distinct from their role as chromatin modifiers such as non-histone acetylation by canonical histone acetyltransferases [18,19], and in cases it could be that disruption of these functions are the major contributor to disease phenotype.

While a great deal of progress has occurred in the past few decades in identifying the primary genetic defects that cause chromatin modifier disorders and diagnosing affected individuals, there are still many unknowns with respect to the etiology of these diseases. Consider, as a contrast, cystic fibrosis, a well studied disorder with a fairly complete etiology that has been worked out, as reviewed by [20]. Cystic fibrosis which was first identified as a distinct disease in 1938 with pathology of the pancreas resulting from mucus clogging the pancreatic ducts and pathology of the lungs from a similar effect of mucus [21]. In 1983, the cause of the mucus effects observed was identified as an issue with chloride ion transport [22,23], and in 1989 a mutation in the *CFTR* gene was identified as the most prevalent primary cause of cystic fibrosis [20,24]. Having worked out a great deal of detail regarding the mechanism of disease has enabled the focus of research to shift towards identifying therapies and treatments, for example in the development of drugs that can partially rescue the function of the defective CFTR protein [25]. This stands in contrast to most chromatin modifier syndromes for which there are many fundamental questions that remain unanswered, such as those questions around when and where the secondary defects that lead to the resulting set of phenotypes occur in development,

what regions of the genome are affected, and what intermediate processes are disrupted that then lead to or exacerbate disease.

Structural complexity of chromatin modifiers

One of the major challenges in answering even basic questions of the mechanisms of disease for chromatin modifier syndromes comes from various sources of complexity related to their normal function. For one, chromatin modifiers exist in multiprotein complexes that often have variable binding partners. That is, one chromatin modifier might exist in several distinct complexes and have its function modulated by changing the exact partners to which it is bound. For instance, KAT6A is known to function in complexes that includes one of the three distinct homologs of the BRPF protein (BRPF1, BRPF2, or BRPF3), however little has been reported about the divergent functions of these three distinct complexes [26,27]. Additionally, KAT6A's close homolog, KAT6B, can complex with the same set of factors as KAT6A, namely BRPF1/2/3, ING5, and MEAF6 [28]. Thus determining for instance the specific effects of loss of the KAT6A-BRPF1 species of complex compared to the KAT6A-BRPF2/3 species would be technically challenging because simply knocking out BRPF1 will remove not only the effects of the KAT6A-BRPF1 complex species but also those of the KAT6B-BRPF1 complex species. Compared to other proteins, chromatin modifiers tend to have a substantially inflated number of variable binding partners (**Figure 1-3**), which may add substantially to their individual functional diversity but also makes efforts to understand the breadth of their function challenging.

The complexity of the functional form of chromatin modifiers that comes with the variability in their binding partners is compounded when one considers the potential for splice variants and alternative transcripts in these genes. Chromatin modifier proteins tend to be fairly large and encoded by correspondingly large transcripts and genes. When controlling for overall transcript length we find that chromatin modifiers tend to have a disproportionate number of

exons making up their gene bodies (**Figure 1-4**). Alternative splicing is one means by which an individual gene can give rise to distinct transcripts and protein products with potentially distinct functions, so the increased number of exons in chromatin modifiers opens up transcript diversity as a potential means by which a relatively small number of chromatin modifiers genes could achieve additional functional diversity. The extent to which alternative splice forms of chromatin modifiers play a role in modulating their functions remains as an open direction for investigation, though recent studies have begun to identify isoforms of chromatin modifiers with apparently distinct functions [29,30]. Studies of alternative splicing in a number of other genes have shown alternative splicing as a prevalent mechanism for generating functional protein diversity, reviewed in [31]. Consequently, little is currently known about whether differential splicing for chromatin modifiers is a common mechanism for modulating the partners they will complex with or if, alternatively, alternative splice products tend to function with a common set of binding partners and modulate the function of the multiprotein complexes in other ways.

So far we have discussed the complexity of chromatin modifiers as protein products and in multiprotein complexes, leaving out discussion of post-translational modifications of the proteins in these complexes which may add another layer of functional complexity [32]. Taken together we can begin to appreciate the scale of the challenge that confront researchers in their attempts to identify all the functional varieties of chromatin modifiers and chromatin modifier complexes. We next turn our discussion from complexity in the form of chromatin modifiers to the complexity in the effects they have on chromatin.

Functional complexity of chromatin modifiers

As mentioned in the previous sections, chromatin modifiers are responsible for a variety of chemical and physical modifications to chromatin. Canonically, histone acetyltransferases (HATs) add acetyl groups to lysine residues of histones, histone deacetylases (HDACs) remove

these groups from lysine residues of histones, histone methyltransferases (HMTs) add methyl groups, histones demethylases (HDMs) remove them. Histones have a number of different amino acid residues that can be modified and each chromatin modifier that acts as a chemical modifier of histones can affect some subset of these residues. While many relationships between chromatin modifiers and their modifications have been worked out [33–35], a large number of histone post-translational modifications (PTMs) have been observed that remain unannotated in terms of both function and the specific modifiers that catalyze their addition or removal from histones [33].

Canonical HATs, HDAC, HMTS, and HDMs have also been reported to produce non-canonical modifications. For example, HAT1, a canonical lysine acetyl-transferase has been observed to also have lysine isobutyryl-transferase activity for histone H4 [36]. The precise modifications to histones that chromatin modifiers make are important for the regulation of the genomic information that these histones overlay. Transcription factors and other regulatory proteins may require specific histone marks in order to bind to a region of the genome and activate or repress transcription, or conversely specific histone marks may be required to block such binding and activity [37]. While we use the example here specifically of histone modifiers, these observations are similarly valid for other chromatin modifiers and their associated modifications, such as DNA methylation which can also affect the binding characteristics for a variety of effector proteins [38].

The specific modifications to chromatin that a chromatin modifier makes is one facet of its function, another is in the determination of the regions of the genome over which the chromatin modifier will carry out these effects. The ability of chromatin modifiers to target specific regions of the genome is critical for their function in turning on or off particular gene expression programs or in altering the capacity of a genomic region to respond to other regulatory signals. Such regional specificity is achieved by the integration of information from a

variety of sources. The sequence of DNA and recruitment by specific transcription factors [39,40], the histones' current set of chemical modifications [3], and higher order features of chromatin, such as location within the nucleus [41], together influence whether and to what degree a particular chromatin modifier complex may bind and carry out its activity over a particular genomic region.

The complementary functionality that chromatin modifiers possess of recognizing and binding to genomic regions that are in a specific chromatin state and then modifying the chromatin state of the same underlying regions, potentially in response to some additional signals, places them at a key position in choreographing the gene expression programs critical to cellular and organismal development. We explore some of the developmental functions chromatin modifiers have been associated with next.

Chromatin modifiers in development

Complementary to molecular biological studies into the function of chromatin modifiers are investigations into the role they play in development. While a complete understanding of the molecular biology of chromatin modifiers must account for the complexities described in the previous section to fully explain the molecular mechanisms of their activity, genetic experiments in model organisms have shed light on some of the diversity of developmental processes in which chromatin modifiers play a critical role.

Forward genetic screens in model organisms were foundational in identifying some of the key processes in which chromatin modifiers play a critical role. Some of the most dramatic include the early mutagenesis assays in drosophila that identified for instance the polycomb repressive complex (PRC) [42,43]. Reverse genetic approaches introducing mutations to known chromatin modifiers in mouse and zebrafish models have identified roles in diverse developmental processes such as hematopoiesis [44] and segmentation [45,46].

In addition to experimental genetic approaches, genetic studies aimed at diagnosing disorders provide another avenue for understanding the developmental impact of mutations in chromatin modifiers. Congenital heart defects [47], autism [48], and a variety of developmental syndromes [10,15,49,50].

Conclusion

In this chapter, we have attempted to provide the reader with a brief foundation to the biology of chromatin modifiers and frame some major open questions around the form and function of this diverse class of genes. Given the importance of chromatin modifiers as an interface between the information in the genome and that in the environment, especially as it relates to organismal development and various diseases of development, it is unsurprising that chromatin modifier biology has rapidly grown as an area of active inquiry. In the next chapter, we will survey some of the existing molecular assays and methods of analysis that are used to investigate chromatin modifier biology. There we will outline some of the major gaps that exist, particularly in modes of analysis for molecular biology data, that affect our ability to understand chromatin modifier biology. This will then take us to the remaining chapters of this dissertation where we will develop novel methods of analysis that aim to address some of these major gaps.

Figures and Tables

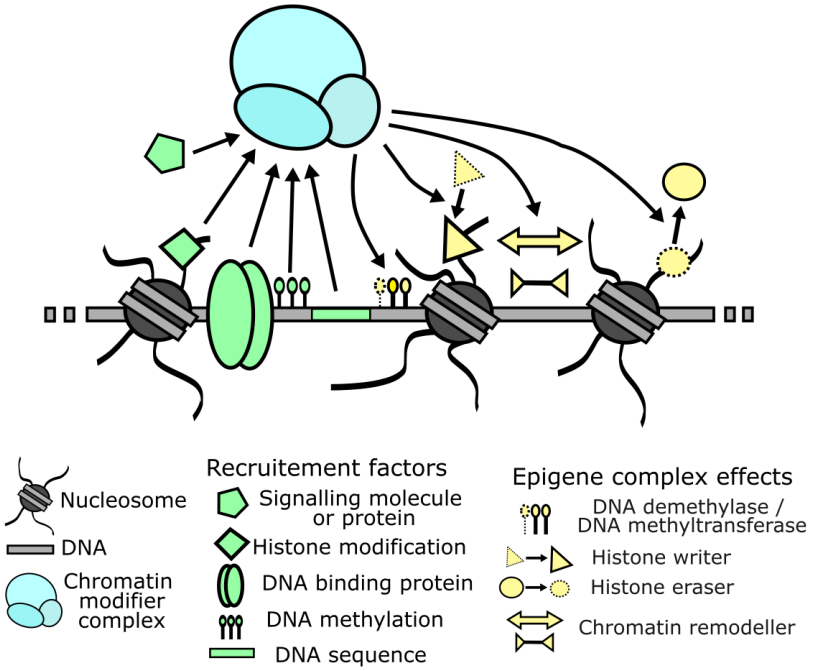


Figure 1-1: Toy figure representing some general functions of chromatin modifier complexes (blue) in reading signals (green) and writing/erasing chromatin features (yellow). Symbols of the figure are described in the legend. Note that chromatin modifier complexes do not individually catalyze all shown effects. The chromatin modifier complex shown above is an aggregate representation of potential functions for chromatin modifiers and not a model of an individual complex.

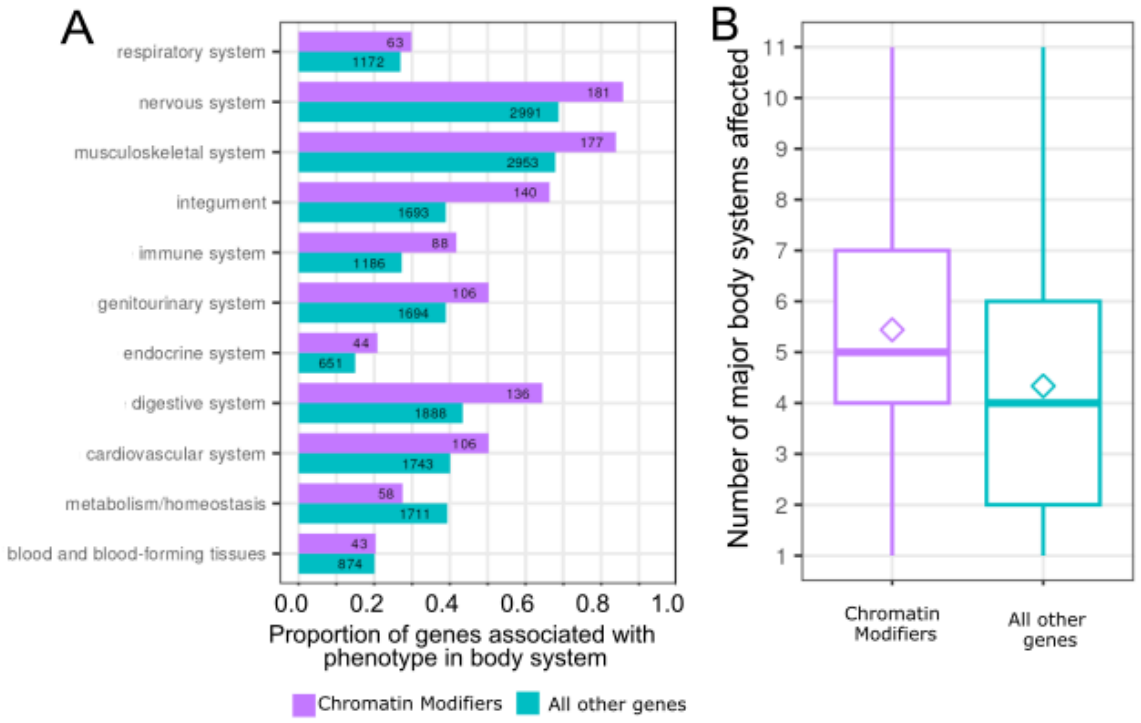


Figure 1-2. Chromatin modifier associated syndromes tend to affect a greater number of body systems than other mendelian diseases. Phenotype information was obtained from the online mendelian inheritance in man (OMIM) [51,52] and the human phenotype ontology (HPO)[53,54]. Chromatin modifier annotations were from [55] excluding histone and protamine genes as done in [56] **A)** proportion of genes associated with some mendelian disease that have a phenotype affecting the major body systems given as rows. Number in each box is number of gene's associated with phenotypes in each body system. **B)** boxplot of number of major body systems affected by genes' associated syndromes. Mean is plotted as diamond.

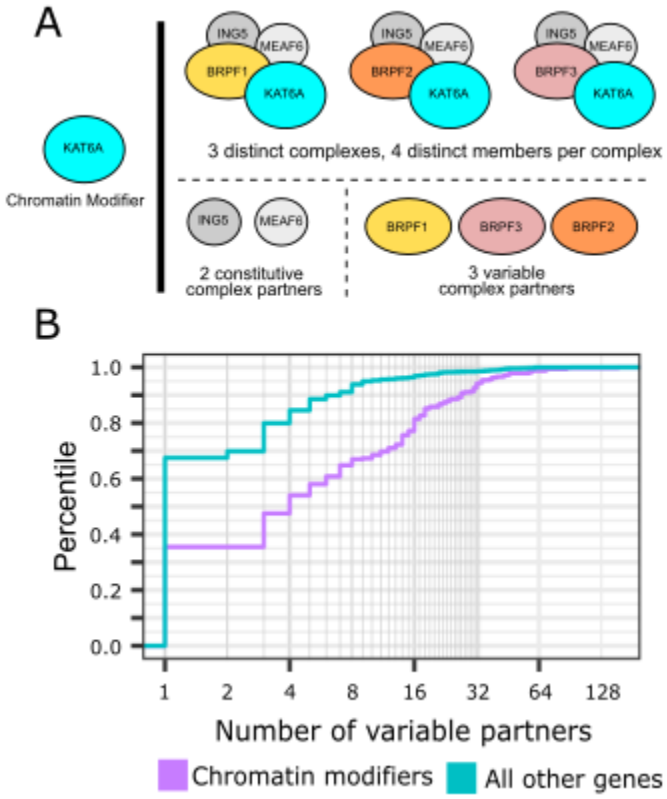


Figure 1-3. Chromatin modifiers have an inflated number of variable binding partners. Binding partner annotations were obtained from the EMBL-EBI Complex Portal [57,58]. Chromatin modifier annotations were from [55] excluding histone and protamine genes as done in [56]. A) a toy figure demonstrating how variable partners are counted. B) empirical cumulative density function (eCDF) comparing number of variable complex partners for chromatin modifiers to the same for all other genes that have an associated protein complex.

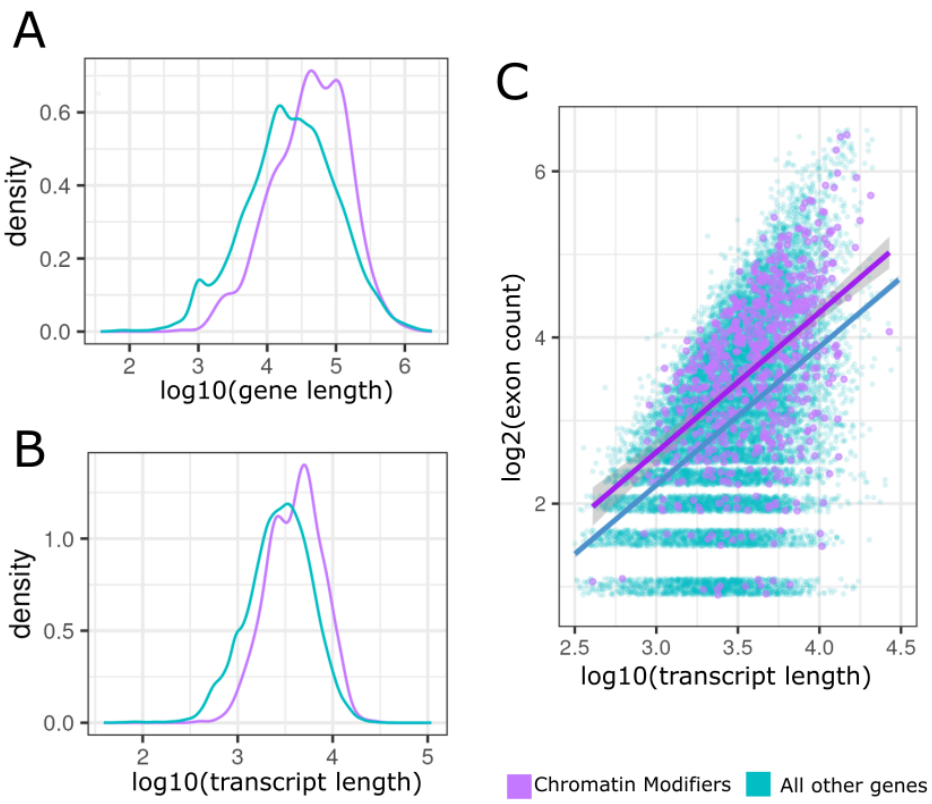


Figure 1-4. Chromatin modifiers are encoded by longer genes and transcripts with an inflated number of exons after accounting for transcript length. Gene structure annotations from Ensembl [59,60]. Chromatin modifier annotations were from [55] excluding histone and protamine genes as done in [56]. **A)** distribution of gene lengths for chromatin modifiers and all other genes **B)** distribution of transcript lengths for chromatin modifiers and all other genes **C)** number of exons vs transcript length where each point is the transcript length and exon count for the canonical gene transcript. Regression lines relating exon count to transcript length are shown.

References

1. Smith, E. & Shilatifard, A. The chromatin signaling pathway: diverse mechanisms of recruitment of histone-modifying enzymes and varied biological outcomes. *Mol. Cell* **40**, 689–701 (2010).

2. Davidovich, C. & Cech, T. R. The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA* **21**, 2007–2022 (2015).

3. Chen, T. & Dent, S. Y. R. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Genet.* **15**, 93–106 (2014).

4. Badeaux, A. I. & Shi, Y. Emerging roles for chromatin as a signal integration and storage platform. *Nat. Rev. Mol. Cell Biol.* **14**, 211–224 (2013).

5. Perino, M. & Veenstra, G. J. C. Chromatin Control of Developmental Dynamics and Plasticity. *Dev. Cell* **38**, 610–620 (2016).

6. Butler, J. S. & Dent, S. Y. R. The role of chromatin modifiers in normal and malignant hematopoiesis. *Blood* **121**, 3076–3084 (2013).

7. Miroshnikova, Y. A., Cohen, I., Ezhkova, E. & Wickström, S. A. Epigenetic gene regulation, chromatin structure, and force-induced chromatin remodelling in epidermal development and homeostasis. *Curr. Opin. Genet. Dev.* **55**, 46–51 (2019).

8. Tyssowski, K., Kishi, Y. & Gotoh, Y. Chromatin regulators of neural development. *Neuroscience* **264**, 4–16 (2014).

9. Hastings, R. *et al.* Bohring-Opitz (Oberklaid-Danks) syndrome: clinical study, review of the literature, and discussion of possible pathogenesis. *Eur. J. Hum. Genet.* **19**, 513–519 (2011).

10. Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat. Genet.* **43**, 729–731 (2011).

11. Nousbeck, J. *et al.* A mutation in a skin-specific isoform of SMARCAD1 causes autosomal-dominant adermatoglyphia. *Am. J. Hum. Genet.* **89**, 302–307 (2011).

12. Elhaji, Y. *et al.* Two SMARCAD1 Variants Causing Basan Syndrome in a Canadian and a Dutch Family. *JID Innov* **1**, 100022 (2021).

13. Fahrner, J. A. & Bjornsson, H. T. Mendelian disorders of the epigenetic machinery: tipping the balance of chromatin states. *Annu. Rev. Genomics Hum. Genet.* **15**, 269–293 (2014).

14. Kennedy, J. *et al.* KAT6A Syndrome: genotype–phenotype correlation in 76 patients with pathogenic KAT6A variants. *Genet. Med.* **21**, 850–860 (2019).

15. Banka, S. *et al.* How genetically heterogeneous is Kabuki syndrome?: MLL2 testing in 116 patients, review and analyses of mutation and phenotypic spectrum. *Eur. J. Hum. Genet.* **20**, 381–388 (2012).

16. Willemsen, M. H. *et al.* Update on Kleefstra Syndrome. *Mol. Syndromol.* **2**, 202–212 (2012).

17. Kingdom, R. & Wright, C. F. Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. *Front. Genet.* **13**, 920390 (2022).

18. Glozak, M. A., Sengupta, N., Zhang, X. & Seto, E. Acetylation and deacetylation of non-histone proteins. *Gene* **363**, 15–23 (2005).

19. Rokudai, S. *et al.* MOZ increases p53 acetylation and premature senescence through its complex formation with PML. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 3895–3900 (2013).

20. Davis, P. B. Cystic fibrosis since 1938. *Am. J. Respir. Crit. Care Med.* **173**, 475–482 (2006).

21. Andersen, D. H. CYSTIC FIBROSIS OF THE PANCREAS AND ITS RELATION TO CELIAC DISEASE: A CLINICAL AND PATHOLOGIC STUDY. *Am. J. Dis. Child.* **56**, 344–399 (1938).

22. Quinton, P. M. Chloride impermeability in cystic fibrosis. *Nature* **301**, 421–422 (1983).

23. Knowles, M. R. *et al.* Abnormal ion permeation through cystic fibrosis respiratory epithelium. *Science* **221**, 1067–1070 (1983).

24. Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).

25. Ong, T. & Ramsey, B. W. Cystic Fibrosis: A Review. *JAMA* **329**, 1859–1871 (2023).

26. Klein, B. J., Lalonde, M. E., Côté, J., Yang, X. J. & Kutateladze, T. G. Crosstalk between epigenetic readers regulates the MOZ/MORF HAT complexes. *Epigenetics* **9**, (2014).

27. Yang, X.-J. MOZ and MORF acetyltransferases: Molecular interaction, animal development and human disease. *Biochim. Biophys. Acta* **1853**, 1818–1826 (2015).

28. Ullah, M. *et al.* Molecular architecture of quartet MOZ/MORF histone acetyltransferase complexes. *Mol. Cell. Biol.* **28**, 6828–6843 (2008).

29. He, Q. *et al.* Isoform-specific involvement of Brpf1 in expansion of adult hematopoietic stem and progenitor cells. *J. Mol. Cell Biol.* **12**, 359–371 (2020).

30. Nazim, M. *et al.* Alternative splicing of a chromatin modifier alters the transcriptional regulatory programs of stem cell maintenance and neuronal differentiation. *Cell Stem Cell* **31**, 754–771.e6 (2024).

31. Kelemen, O. *et al.* Function of alternative splicing. *Gene* **514**, 1–30 (2013).

32. Dantas, A. *et al.* Biological Functions of the ING Proteins. *Cancers* **11**, (2019).

33. Zhao, Y. & Garcia, B. A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb. Perspect. Biol.* **7**, a025064 (2015).

34. Shah, S. G. *et al.* HISTome2: a database of histone proteins, modifiers for multiple organisms and epidrugs. *Epigenetics Chromatin* **13**, 31 (2020).

35. Zhang, Y. *et al.* Overview of Histone Modification. in *Histone Mutations and Cancer* (eds. Fang, D. & Han, J.) 1–16 (Springer Singapore, Singapore, 2021).

36. Zhu, Z. *et al.* Identification of lysine isobutyrylation as a new histone modification mark. *Nucleic Acids Res.* **49**, 177–189 (2021).

37. Xin, B. & Rohs, R. Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Res.* **28**, 321–333 (2018).

38. Héberlé, É. & Bardet, A. F. Sensitivity of transcription factors to DNA methylation. *Essays Biochem.* **63**, 727–741 (2019).

39. McManus, S. *et al.* The transcription factor Pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells. *EMBO J.* **30**, 2388–2404 (2011).

40. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).

41. Harr, J. C., Gonzalez-Sandoval, A. & Gasser, S. M. Histones and histone modifications in perinuclear chromatin anchoring: from yeast to man. *EMBO Rep.* **17**, 139–155 (2016).

42. Kassis, J. A., Kennison, J. A. & Tamkun, J. W. Polycomb and Trithorax Group Genes in Drosophila. *Genetics* **206**, 1699–1725 (2017).

43. Blackledge, N. P. & Klose, R. J. The molecular principles of gene regulation by Polycomb repressive complexes. *Nat. Rev. Mol. Cell Biol.* **22**, 815–833 (2021).

44. Ding, Y., Liu, Z. & Liu, F. Transcriptional and epigenetic control of hematopoietic stem cell fate decisions in vertebrates. *Dev. Biol.* **475**, 156–164 (2021).

45. Kok, F. O. *et al.* The role of the SPT6 chromatin remodeling factor in zebrafish embryogenesis. *Dev. Biol.* **307**, 214–226 (2007).

46. Miller, C. T., Maves, L. & Kimmel, C. B. moz regulates Hox expression and pharyngeal segmental identity in zebrafish. *Development* **131**, 2443–2461 (2004).

47. Watkins, W. S. *et al.* De novo and recessive forms of congenital heart disease have distinct

genetic and phenotypic landscapes. *Nat. Commun.* **10**, (2019).

48. Suliman, R., Ben-David, E. & Shifman, S. Chromatin regulators, phenotypic robustness, and autism risk. *Front. Genet.* **5**, 81 (2014).

49. Arboleda, V. A. *et al.* De novo nonsense mutations in KAT6A, a lysine acetyl-transferase gene, cause a syndrome including microcephaly and global developmental delay. *Am. J. Hum. Genet.* **96**, 498–506 (2015).

50. Hempel, M. *et al.* De Novo Mutations in CHAMP1 Cause Intellectual Disability with Severe Speech Impairment. *Am. J. Hum. Genet.* **97**, 493–500 (2015).

51. Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* **15**, 57–61 (2000).

52. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–98 (2015).

53. Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).

54. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).

55. Medvedeva, Y. A. *et al.* EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database* **2015**, bav067 (2015).

56. Nava, A. A. & Arboleda, V. A. The omics era: a nexus of untapped potential for Mendelian chromatinopathies. *Hum. Genet.* **143**, 475–495 (2024).

57. Meldal, B. H. M. & Forner-Martinez, O. The complex portal-an encyclopaedia of macromolecular complexes. *Nucleic acids* (2015).

58. Meldal, B. H. M., Perfetto, L. & Combe, C. Complex Portal 2022: new curation frontiers. *Nucleic acids* (2022).

59. Birney, E. *et al.* An overview of Ensembl. *Genome Res.* **14**, 925–928 (2004).

60. Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).

CHAPTER 2

Assays in molecular biology and modes of analysis for the study of chromatin modifier syndromes

Introduction

In this chapter we will elaborate on some of the various methods that are used to study chromatin modifiers, their functions, and the downstream effects that result when their normal activity is disrupted in disease. This chapter will be divided into two main parts. The first part will cover methods in molecular biology that are used to measure chromatin state and other molecular features important to chromatin modifier biology, the second will cover a range of computational methods used in the analysis of the data generated by these molecular biology methods. In the second part of this chapter we will also introduce some of the key gaps that exist in current methods of analysis and highlight those that we will address in the later chapters of this dissertation.

The analytic methods we develop in the later chapters of this dissertation primarily aim to improve upon methods for modeling features that map directly to genomic sequence and describing the anatomic and developmental context over which they occur. This chapter is primarily intended to give context for those later chapters, and so the scope of the material presented here will be limited to those assays and methods of analysis most relevant to these latter discussions. We focus our discussion on methods related to the study of chromatin modifier effects in cellular systems and model organisms, particularly those related to the measurement of features that map directly back to genomic sequence.

Survey of molecular biology assays used in the study of chromatin modifier function

As discussed in the first chapter of this dissertation, chromatin modifiers can act both as readers of chromatin, for instance requiring specific histone modifications to facilitate binding a region of chromatin, and as writers in the general sense, adding or removing chemical modifications to histones or DNA or remodeling chromatin or contributing to the formation of higher order chromatin structure such as loops. A variety of assays have been developed to

explore various features of chromatin state that collectively form the epigenetic landscape of the genome. In particular over the past few decades advances in DNA sequencing technology have made possible assays that can measure the distribution of these features in a genome-wide manner. In the following sections, we will direct much of our attention to these modern assays, beginning with a brief discussion of modern DNA sequencing technology that serves as their foundation.

DNA sequencing

The rapid advancement of DNA sequencing technology over the past few decades have revolutionized biology [1]. In the space of chromatin modifier disorders, whole exome sequencing, along with the collection of population level genomic data that serve as a reference for distinguishing between genetic variants that are likely benign and those that are likely pathogenic [2], have greatly expanded the collective diagnostic capacity of researchers and clinicians [3,4]. Indeed, many chromatin modifier disorders have only been described following the wide adoption of whole exome sequencing technology as a tool for seeking a genetic diagnosis in syndromic individuals without a family history or environmental risk factors for disease. Prior to whole exome sequencing, genetic mapping for many of these disorders was effectively impossible. This is because these disorders tend to be relatively severe with a dominant mode of inheritance and so the vast majority cases of disease are the result of *de novo* germline mutations. Prior to the era of whole exome sequencing, the primary means by which genes were associated with their respective disorders was through family based genetic mapping studies, which by their very nature cannot be used to map *de novo* mutations [5].

Beyond enabling the identification and diagnosis of chromatin modifier disorders, modern DNA sequencing technology also forms the basis for a wide range of molecular biology

assays that are used to measure chromatin state in a genome wide manner. We explore the variety of these derivative assays in the next sections.

Protein-DNA interactions

Chromatin modifiers, along with a variety of other proteins such as transcription factors, must localize to specific regions of the genome in order to carry out their functions. Given that a protein is expected to carry out some function by binding to specific regions of the genome, a natural question to ask is where in the genome is that protein binding. Two widely used methods for answering this question are Chromatin ImmunoPrecipitation followed by DNA sequencing, ChIP-seq [6,7], and the related more recently developed Cleavage Under Targets and Releasing Using Nuclease, CUT&RUN assay [8,9], both of which we briefly outline below.

ChIP-seq and CUT&RUN share a substantial part of their overall workflow, diverging primarily in the means by which the DNA bound to the protein of interest is separated from background DNA. Both methods rely on having some antibody to specifically target the protein of interest and share a first step of covalently crosslinking DNA and proteins with formaldehyde so that associations are preserved in the next steps. The DNA with its covalently associated proteins is then sheared by sonication, typically into fragments of a few hundred basepairs. In ChIP-seq, the antibody targeting the protein of interest is then added in order to cause the fragments of DNA bound to the protein of interest to aggregate and precipitate out of solution when centrifuged. In CUT&RUN, the antibody is added to permeabilized cells along with a DNA cleaving enzyme that binds to the antibody. When the DNA cleaving enzyme is activated it results in specific cleavage at the sites around where the protein of interest is crosslinked to the DNA. The antibody, cleaving enzyme, and the bound protein of interest with the associated DNA then diffuses out of the cell and is collected in the supernatant fraction following centrifugation. For both ChIP-seq and CUT&RUN, the crosslinks between the DNA and the associated protein

of interest are then reversed, the DNA is purified and can then be sequenced to identify the regions of DNA that were bound by the protein of interest [10].

Both ChIP-seq and CUT&RUN enable identification of the regions where a protein is bound genome-wide. The resulting data can then be analyzed by a variety of computational methods to characterize the bound DNA and the genomic regions it came from. This general strategy has been useful for identifying the specific DNA sequences, or motifs, that certain proteins such as transcription factors bind in the genome. However, while there has been some success applying these methods to chromatin modifiers [11,12], in many cases these methods have not been effective at identifying the regions where chromatin modifiers affect chromatin state. Part of this may be due to the interactions between chromatin modifiers and the regions of genomic DNA they affect being more transient or more indirect physically[13] as compared to other proteins such as transcription factors which tend to directly bind their respective target regions with high affinity [14–18]. The modifications that chromatin modifiers impose on chromatin, in line with their role as a persistent layer of information over the genome, are often more accessible to measurement.

Histone post-translational modifications

As mentioned in the first chapter, a subset of chromatin modifiers are defined by their enzymatic activity in adding or removing post-translational modifications (PTMs) to histones. Measuring the chemical modifications present on histones is therefore an obvious potential readout for measuring effects that occur when the function of these chromatin modifiers is disrupted, whether in an experimental context or in observing changes in samples from patients affected by chromatin modifier disorders.

While the chemical structure of DNA is mostly invariant, with a few exceptions such as DNA methylation, histones are a common substrate for PTMs. The post-translational

modification of histones is a critical mechanism by which histone function is modulated, for instance in specifying whether the histone will function to facilitate or restrict transcription of an underlying gene.

Histones are rich in positively charged amino acid residues such as lysine and arginine which enables their binding to the negatively charged DNA. Modifications to histones can act to decrease their affinity for DNA resulting typically in a more open chromatin state; acetylation of histone lysines achieves this by neutralizing the negative charge on lysine with an acetyl group. Alternatively modifications can preserve or enhance histone binding to DNA, for example by protecting the positive change on lysine residues through the addition of methyl groups which protects the lysine from acetylation. In addition to the changes PTMs can have on the charge state of histones, modifications to histones can also introduce or remove binding sites for other proteins such as transcription factors or chromatin modifiers which may have counterintuitive effects, such as addition of an acetyl group introducing a repressor binding site which can then facilitate a more condensed state in opposition to the typical role of acetylation of histones in opening chromatin [19].

There are two complementary types of approach for querying the histone code of the epigenome. The first is to use methods such as ChIP-seq or CUT&RUN, discussed in the previous section, to investigate where histones with a specific PTM are binding genome-wide. The advantage of these methods is that they enable genome-wide assessment of where a histone PTM occurs. Their main drawback is that they require antibodies specific to each PTMs of interest and as a result they cannot be used to identify novel PTMs [8,10].

The second type of approach for investigating the histone code are mass-spectrometry based methods. In contrast to the sequencing based methods of ChIP-seq and CUT&RUN, mass-spectrometry does not require specific antibodies and can be used to query effectively all histone PTMs present in a sample simultaneously, including those that might be novel. It is

however worth noting that rare PTMs may be difficult to detect without deep, and therefore expensive, sampling by mass-spectrometry. The drawback compared to ChIP-seq and CUT&RUN is that mass-spectrometry based methods do not inherently enable mapping histone PTMs to where they map genome-wide [20].

The sequencing based assays such as ChIP-seq and CUT&RUN along with the complementary mass-spectrometry based ones together can be used to answer a wide range of questions around which histone PTMs a chromatin modifier might add or remove, and where across the genome this activity might be occurring.

DNA methylation

Another feature of the epigenome that can be affected by chromatin modifiers is DNA methylation, typically in the context of cytosine guanine (CpG) dinucleotides [21], though non-CpG methylation of cytosine and other nucleotide modifications have also been reported [22–25]. While there are only three genes in humans with DNA methyltransferase activity (*DNMT1, DNMT3A, DNMT3B*)[26], and three primary drivers of active DNA demethylase activity (*TET1, TET2, TET3*) which do not directly catalyze DNA demethylation, but rather oxidize the methylated cytosine which then induced DNA base excision repair [27], DNA methylation is coupled to a variety of other chromatin modifier activities and chromatin features. For example, the polycomb repressive complex which includes histone ubiquitin transferase activity that establishes and maintains chromatin in a specific repressive state that has been found to prevent DNA methylation [28–30]. DNA methylation is also unique amongst chromatin features in that it can retain signatures of prior cell states, which in theory may be useful for investigating changes that occurred in difficult to access precursor cell lineages in individuals affected by chromatin modifier diseases [31]. In the space of chromatin modifier syndromes, DNA methylation has also been studied as a diagnostic tool. Many chromatin modifier disorders have been found to carry

specific DNA methylation signatures that can be used to diagnose otherwise ambiguous cases of disease, such as when attempting to interpret whether a particular missense variant in a gene is pathogenic and disease causing or benign [32].

There are two primary technologies used for assessing DNA methylation genome wide, those that are based on arrays of probes and those that are based on sequencing. For both arrays and sequencing, the specific conversion of unmethylated cytosine residues to uracil using bisulfite enables differentiation between methylated cytosine and unmethylated cytosine. The difference in these technologies is that array based methods use arrays of probes to measure the relative abundance of methylated and unmethylated cytosines at each particular site represented by a probe on the array, whereas the sequencing based methods sequence the bisulfite converted DNA directly [33].

DNA methylation arrays have been designed that can sample many thousands of CpG sites that can be selected to optimize some metric of informativeness. For instance, DNA methylation has been found to be a robust measure of aging and so arrays have been specifically designed to sample age correlated CpG sites as an economical means of measuring this epigenetic age feature across many biological samples [34]. Methylation arrays have the primary advantage of being substantially cheaper than sequencing methods of measuring methylation.

Direct DNA sequencing methods, typically referred to as whole genome bisulfite sequencing, WGBS, enable an agnostic and truly genome-wide survey of methylation at all CpG sites. The drawback is that to get estimates of CpG methylation proportion at individual sites with a similar accuracy as DNA arrays requires very deep sequencing and is therefore much more expensive per CpG queried than arrays [35].

Chromatin accessibility

Chromatin modifiers play an important role in shaping the physical structure of chromatin, for instance in opening regions of chromatin that are to be transcribed so that they are accessible to transcription factors and in closing regions of chromatin that must be silenced. Over the past few decades several technologies have emerged and subsequently replaced one another for assaying chromatin accessibility in a genome-wide manner [36]. MNase-seq was developed in 2007 [37] and coexisted for a while with DNAse-seq after its own development in 2008 [38], both have since largely been replaced by ATAC-seq, developed in 2013 [39,40], as a result of ATAC-seq's far lower input requirements. Chromatin accessibility information is useful for interpretation of a variety of functional genomic features. Regions where chromatin is accessible tend to be those where transcriptional machinery can bind and facilitate gene expression. Non-coding regions such as enhancers also need to have a degree of accessibility in order to function to promote transcription. Measuring accessibility of chromatin can also be used to assess high resolution features of the epigenetic landscape. For instance, identifying specific transcription factor binding sites that are occupied can be achieved by high resolution measurement of chromatin accessibility looking for gaps in peaks of accessible chromatin representing the "footprint" of the transcription factor [40].

Transcriptomics

The set of RNAs that are transcribed from the genome collectively form the transcriptome of the cell. Regulating the flow of information from the genome to the transcriptome is in many ways the primary functional readout of chromatin state. Measuring the transcriptome is therefore foundational to many studies of the ways that modifications to chromatin affect downstream biological processes.

Most methods of RNA sequencing include a step where RNA is reverse transcribed to DNA, and so much of the sequencing technology is shared between RNA and DNA sequencing. Different methods of sequencing the RNA content of a biological sample can focus on different questions. Whole RNA-seq with ribosome depletion and polyA RNA-seq are fairly standard methodologies for sampling all the RNAs or all polyadenylated RNAs in a sample [41]. Long-read RNA-sequencing is used when questions of isoform usage are of primary interest [42,43]. Cap analysis of gene expression sequencing, CAGE-seq, can be used to identify alternative transcription start sites with greater depth of sequencing than is typically cost effective for long-read RNA-sequencing [44]. GRO-Seq can identify genes actively being transcribed [45]. For comprehensive reviews of RNA-seq technologies see [46].

Single-cell data

Following the development of high-throughput DNA sequencing technology, one of the main advances in biology of the past decade has been in the development of microfluidic devices that enable high-throughput single-cell and single nuclei sequencing. Methods have been developed to couple many of the assays discussed above, e.g. ChIP-seq, ATAC-seq, RNA-seq, with single-cell technologies to enable observation of chromatin and transcriptomic state within individual cells. One of the main common features of these single-cell coupled assays is the use of cellular barcodes in the form of nucleotides that are ligated to the target molecules that are being assayed, for instance the cDNA generated from RNAs in scRNA-seq, or the genomic DNA fragments generated by the transposase in ATAC-seq. These barcodes enable pooling of fragments before sequencing so that the sequencing part of an experiment can use the same technologies as any other sequencing experiment [47–49].

Survey of current computational methods of analysis

Having surveyed a variety of the methods available for measuring features related to chromatin state (ATAC-seq, ChIP-seq, etc) and the transcriptional readout of chromatin state (RNA-seq), we next discuss some of the general modes of analysis that are available for interpreting the data generated by these assays.

Peak calling, regional enrichment testing, and the open challenge in identifying features varying dramatically in size

The result after sequencing and quality control for many of the assays discussed above is a file of all the reads or fragments that can then be mapped to the genome. For assays such as ATAC-seq and ChIP-seq that look at the distribution of some feature across the genome, the next step is to determine what features the data represent. This is the general role of peak calling methods, condensing a large file of reads mapped to the genome into a reduced set of features representing, for ATAC-seq [50], the regions of accessible chromatin, and for ChIP-seq [10,51] the regions where a protein of interest or a specific PTM was bound to the genome.

One of the technical aspects of many of these peak callers is their implementation of some sort of window size or bandwidth parameter that determines the size of the features or peaks they return [50–52]. Typically such peak calling algorithms are used by default to identify peaks on the order of tens of basepairs up to a few hundred basepairs. Many studies of chromatin features have observed that regulatory structures can effectively vary in size from individual basepairs [53–55] to essentially genome spanning effects [56,57], with intermediate features existing at nearly all scales. For instance polycomb domains may be tens of thousands or hundreds of thousands of basepairs in length [58], and topologically associated domains may be millions of basepairs in length [59–61]. Algorithms for peak calling can often be applied to identify these features by simply increasing the bandwidth parameter to more closely match the

expected size of the features of interest, however there is a lack of robust methods that can identify features without *a priori* knowledge of the feature size of interest. Additionally, when used in this manner, there are not well defined decision procedures for resolving for instance, whether a cluster of features called using a bandwidth of ten kilobases is better than a single feature returned when using a bandwidth of one hundred kilobases.

In **chapter 3** we will develop a novel method to address this issue for a similar problem of identifying regions of the genome that are enriched in differentially methylated CpGs for DNA methylation data.

Transcript quantification, defining the transcriptome, and the open challenge of describing features within the context of a broader system

One of the challenges with measuring functional genomic features is in setting a baseline for measurement. For case-control type studies where the contrast of interest is binary and well defined, the task of identifying differential features is somewhat simplified, for instance in identifying all genes upregulated in disease in a given tissue. However, for some questions, such as identifying all the genomic features that are specific to some context as opposed to all others, e.g. identifying genes specific to cardiac tissue or with broader specificity to all muscle tissues, the task is complicated by a number of factors. For one, measurement of transcript abundance is usually done relativistically, measuring transcript abundance against the background of all transcripts in the sample. This is easily seen in commonly used units of measurement for RNA abundance such as TPM or Transcripts (of the gene) Per Million (transcripts in the sample) [62]. This can complicate comparisons between tissues where for instance a few transcript species dominate the population of transcripts, such as in whole blood where most transcripts present are from just a few hemoglobin encoding genes, or pancreas where a similarly large part of the population of transcripts are just a few secreted digestive

enzymes. For such cases median normalization can make comparisons more meaningful [63]. One of the outstanding issues however, is that the composition of the sample set can drastically change estimates of gene expression characteristics, such as the specificity of gene expression [64].

In **chapter 4** we will develop a novel method for reweighting measures of gene specificity taking into account the similarity structure of the diverse tissue sample set, which can in principle be applied to a variety of statistics measured over datasets covering biological systems.

Conclusion

Modern sequencing technology and complementary molecular biology methods have opened up a large space in which we can begin to answer some of the most fundamental questions around chromatin modifier function. As we have highlighted in this chapter, it is now possible to design experiments to identify the types of modifications chromatin modifiers make to the epigenome, for instance through mass-spectroscopy based methods; map the regions of the genomes affected when chromatin modifier function is perturbed, by ChIP-seq, CUT&RUN, ATAC-seq, and other sequencing based technologies; and measure the transcriptomic output of those changes with the various flavors of RNA-seq that have been developed. Coupled with proper experimental design, these assays can generate a wealth of information. Converting that raw information into real biological knowledge however is far from trivial.

In the latter part of this chapter we briefly outlined some of the computational methods used in the analysis of the raw data generated by modern sequencing and molecular biology methods. In particular we focused on the methods of peak calling, to identify regions of the genome over which some feature exists, and methods for transcript quantification. We highlighted two major open challenges that exist in these general analysis tasks. First is the

inability of existing peak-caller methods to identify regions over which some feature is enriched when these regions may be hypervariable in size. Second is the lack of descriptive statistics that can take into account the relations that exist between biological entities which can cause issues of reproducibility and reporting of results. In the next chapters we develop and validate novel solutions to these problems, the implementation of which can improve our ability to investigate and understand the form and function of chromatin modifiers in health and disease.

References

1. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

2. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).

3. Chung, C. C. Y. *et al.* Rapid whole-exome sequencing facilitates precision medicine in paediatric rare disease patients and reduces healthcare costs. *Lancet Reg Health West Pac* **1**, 100001 (2020).

4. Nguyen, M. T. & Charlebois, K. The clinical utility of whole-exome sequencing in the context of rare diseases - the changing tides of medical practice. *Clin. Genet.* **88**, 313–319 (2015).

5. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).

6. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).

7. Collas, P. The current state of chromatin immunoprecipitation. *Mol. Biotechnol.* **45**, 87–100 (2010).

8. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, (2017).

9. Meers, M. P., Bryson, T. D., Henikoff, J. G. & Henikoff, S. Improved CUT&RUN chromatin profiling tools. *Elife* **8**, (2019).

10. Ma, S. & Zhang, Y. Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq. *Mol Biomed* **1**, 9 (2020).

11. Mazina, M. Y. & Vorobyeva, N. E. Chromatin Modifiers in Transcriptional Regulation: New Findings and Prospects. *Acta Naturae* **13**, 16–30 (2021).

12. Tyagi, M., Imam, N., Verma, K. & Patel, A. K. Chromatin remodelers: We are the drivers!! *Nucleus* **7**, 388–404 (2016).

13. Lloyd, S. M. & Bao, X. Pinpointing the Genomic Localizations of Chromatin-Associated Proteins: The Yesterday, Today, and Tomorrow of ChIP-seq. *Curr. Protoc. Cell Biol.* **84**, e89 (2019).

14. Jung, C. *et al.* True equilibrium measurement of transcription factor-DNA binding affinities using automated polarization microscopy. *Nat. Commun.* **9**, 1605 (2018).

15. Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D. & Patel, D. J. How chromatin-binding

modules interpret histone modifications: lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.* **14**, 1025–1040 (2007).

16. Yang, Y. & Li, G. Post-translational modifications of PRC2: signals directing its activity. *Epigenetics Chromatin* **13**, 47 (2020).

17. Patsialou, A., Wilsker, D. & Moran, E. DNA-binding properties of ARID family proteins. *Nucleic Acids Res.* **33**, 66–80 (2005).

18. Becht, D. C. *et al.* MORF and MOZ acetyltransferases target unmethylated CpG islands through the winged helix domain. *Nat. Commun.* **14**, 697 (2023).

19. Zhang, Y. *et al.* Overview of Histone Modification. in *Histone Mutations and Cancer* (eds. Fang, D. & Han, J.) 1–16 (Springer Singapore, Singapore, 2021).

20. Yuan, Z.-F., Arnaudo, A. M. & Garcia, B. A. Mass spectrometric analysis of histone proteoforms. *Annu. Rev. Anal. Chem.* **7**, 113–128 (2014).

21. Sinsheimer, R. L. The action of pancreatic deoxyribonuclease. II. Isomeric dinucleotides. *J. Biol. Chem.* **215**, 579–583 (1955).

22. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5237–5242 (2000).

23. Boulias, K. & Greer, E. L. Means, mechanisms and consequences of adenine methylation in DNA. *Nat. Rev. Genet.* **23**, 411–428 (2022).

24. Shen, C., Wang, K., Deng, X. & Chen, J. DNA N6-methyldeoxyadenosine in mammals and human disease. *Trends Genet.* **38**, 454–467 (2022).

25. Berney, M. & McGouran, J. F. Methods for detection of cytosine and thymine modifications in DNA. *Nature Reviews Chemistry* **2**, 332–348 (2018).

26. Lyko, F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.* **19**, 81–92 (2018).

27. Bhutani, N., Burns, D. M. & Blau, H. M. DNA demethylation dynamics. *Cell* **146**, 866–872 (2011).

28. Viré, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439**, 871–874 (2006).

29. Li, Y. *et al.* Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biol.* **19**, 18 (2018).

30. McLaughlin, K. *et al.* DNA Methylation Directs Polycomb-Dependent 3D Genome

Re-organization in Naive Pluripotency. *Cell Rep.* **29**, 1974–1985.e6 (2019).

31. Chater-Diehl, E. *et al.* Anatomy of DNA methylation signatures: Emerging insights and applications. *Am. J. Hum. Genet.* **108**, 1359–1366 (2021).

32. Awamleh, Z., Goodman, S., Choufani, S. & Weksberg, R. DNA methylation signatures for chromatinopathies: current challenges and future applications. *Hum. Genet.* **143**, 551–557 (2024).

33. Li, S. & Tollefsbol, T. O. DNA methylation methods: Global DNA methylation and methylomic analyses. *Methods* **187**, 28–43 (2021).

34. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).

35. Adusumalli, S., Mohd Omar, M. F., Soong, R. & Benoukraf, T. Methodological aspects of whole-genome bisulfite sequencing analysis. *Brief. Bioinform.* **16**, 369–379 (2015).

36. Mansisidor, A. R. & Risca, V. I. Chromatin accessibility: methods, mechanisms, and biological insights. *Nucleus* **13**, 236–276 (2022).

37. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**, 572–576 (2007).

38. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

39. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

40. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat. Protoc.* **17**, 1518–1552 (2022).

41. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.* **8**, 4781 (2018).

42. Uapinyoying, P. *et al.* A long-read RNA-seq approach to identify novel transcripts of very large genes. *Genome Res.* **30**, 885–897 (2020).

43. Cho, H. *et al.* High-resolution transcriptome analysis with long-read RNA sequencing. *PLoS One* **9**, e108095 (2014).

44. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211–222 (2006).

45. Gardini, A. Global Run-On Sequencing (GRO-Seq). *Methods Mol. Biol.* **1468**, 111–120 (2017).

46. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).

47. Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp. Mol. Med.* **52**, 1419–1427 (2020).

48. Jovic, D. *et al.* Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.* **12**, e694 (2022).

49. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).

50. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).

51. Jeon, H., Lee, H., Kang, B., Jang, I. & Roh, T.-Y. Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis. *Genomics Inform.* **18**, e42 (2020).

52. Mallik, S. *et al.* An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays. *Brief. Bioinform.* **00**, 1–12 (2018).

53. Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688 (2012).

54. Gaston, K. & Fried, M. CpG methylation has differential effects on the binding of YY1 and ETS proteins to the bi-directional promoter of the Surf-1 and Surf-2 genes. *Nucleic Acids Res.* **23**, 901–909 (1995).

55. Wiehle, L. *et al.* DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res.* **29**, 750–761 (2019).

56. Bao, J. & Bedford, M. T. Epigenetic regulation of the histone-to-protamine transition during spermiogenesis. *Reproduction* **151**, R55–70 (2016).

57. Brewer, L. R., Corzett, M. & Balhorn, R. Protamine-induced condensation and decondensation of the same DNA molecule. *Science* **286**, 120–123 (1999).

58. Schwartz, Y. B. *et al.* Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nat. Genet.* **38**, 700–705 (2006).

59. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Sci Adv* **5**, eaaw1668 (2019).

60. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome

Organization. *Mol. Cell* **62**, 668–680 (2016).

61. Bertero, A. *et al.* Chromatin compartment dynamics in a haploinsufficient model of cardiac laminopathy. *J. Cell Biol.* **218**, 2919–2944 (2019).

62. Zhao, Y. *et al.* TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *J. Transl. Med.* **19**, 269 (2021).

63. Zyprych-Walczak, J. *et al.* The Impact of Normalization Methods on RNA-Seq Data Analysis. *Biomed Res. Int.* **2015**, 621690 (2015).

64. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).

CHAPTER 3

DMRscaler: A Scale-Aware Method to Identify Regions of Differential DNA Methylation Spanning Basepair to Multi-Megabase Features

Abstract

Pathogenic mutations in genes that control chromatin function have been implicated in rare genetic syndromes. These chromatin modifiers exhibit extraordinary diversity in the scale of the epigenetic changes they affect, from single basepair modifications by DNMT1 to whole genome structural changes by PRM1/2. Patterns of DNA methylation are related to a diverse set of epigenetic features across this full range of epigenetic scale, making DNA methylation valuable for mapping regions of general epigenetic dysregulation. However, existing methods are unable to accurately identify regions of differential methylation across this full range of epigenetic scale directly from DNA methylation data.

To address this, we developed DMRscaler, a novel method that uses an iterative windowing procedure to capture regions of differential DNA methylation (DMRs) ranging in size from single basepairs to whole chromosomes. We benchmarked *DMRscaler* against several DMR callers in simulated and natural data comparing XX and XY peripheral blood samples. *DMRscaler* was the only method that accurately called DMRs ranging in size from 100 bp to 1 Mb (pearson's r = 0.94) and up to 152 Mb on the X-chromosome. We then analyzed methylation data from rare-disease cohorts that harbor chromatin modifier gene mutations in *NSD1*, *EZH2*, and *KAT6A* where *DMRscaler* identified novel DMRs spanning gene clusters involved in development.

Taken together, our results show DMRscaler is uniquely able to capture the size of DMR features across the full range of epigenetic scale and identify novel, co-regulated regions that drive epigenetic dysregulation in human disease.

Background

Genes that regulate chromatin structure and function are critical to coordination of complex developmental trajectories within an embryo. Mutations in these chromatin modifier genes are enriched in clinical cohorts with autism [1–4], congenital heart disease [5,6] and global developmental delay [3,5]. Pathogenic mutations in chromatin modifier genes can also result in specific syndromes that have both overlapping and distinct phenotypic features [7–10]. While clinical phenotypes often converge around a common set of chromatin modifier genes, the underlying molecular mechanisms driving these phenotypes are not well characterized.

Chromatin modifiers work in protein complexes to bind chromatin and shape the physical and chemical landscape of the genome, i.e. the epigenome. The regions within the genome where a particular chromatin modifier exerts its influence are critical to defining its role in development. The genomic region controlled by a chromatin modifier can be highly localized, as in methylation of individual cytosine nucleotides which modulates the binding affinity for certain transcription factors (TFs) [11–15], or it can extend across the chromatin landscape more globally, as occurs with the PRM1/2 mediated compaction of the genome during spermatogenesis [16,17] or *Xist* in condensing the X-chromosome in cells with multiple copies of the X-chromosome [18–20]. Between the local and the global are a diversity of epigenetic features that exist at intermediate scales from tens of kilobases to many megabases. These include features such as polycomb repressive domains (PRDs) [21–23] and topologically associated domains (TADs) [24] and co-regulated gene clusters. These intermediate-sized features coordinate higher order patterning events throughout the genome in development, such as PRD regulation of *Hox* segmentation patterning [25], or organization of olfactory receptor gene clusters into TADs [26] with interdependent epigenetic regulation of the member olfactory receptor genes [27,28]. A comprehensive understanding of chromatin modifiers requires understanding the scale of their effect on the epigenetic landscape.

While the direction of causality is still an open question for the interaction between many epigenetic features, changes in DNA methylation (DNAme) are often associated with changes in other epigenetic features across the range of epigenetic scale. DNAme is the covalent addition of a methyl group to a single cytosine nucleotide usually in the context of a CpG dinucleotide [29]. While DNAme directly alters the binding affinity for a set of DNA binding proteins [11–15], it is also associated with higher order epigenetic features. At promoters and enhancers DNAme tends to be inversely correlated with gene activity[30,31]. Over the tens to hundreds of kilobases of PRDs, DNA methylation is depleted by the antagonistic action of the polycomb repressive complex[32,33], and as a result changes in polycomb activity over PRDs are often associated with differential methylation[33]. Megabase scale domains of active and inactive chromatin can be reliably predicted from DNAme patterns[34], and in colon cancer, changes to DNAme have been reported to overlap with these megabase-sized inactive domains[35].

Phenotypic variability and genetic heterogeneity can make the diagnosis of rare syndromes challenging. Even more challenging is the interpretation of the clinical significance of rare genetic variants identified in whole genome sequencing studies in patients with rare disease. In the absence of clear functional data, these genetic variants are annotated as variants of unknown significance (VUSs). One method to distinguish between pathogenic and benign variants is to identify common patterns of differential DNAme from patients with known pathogenic mutations in the same gene, a methylation signature[9,10,36]. The presence of a DNAme signature suggests that common epigenetic marks are associated with pathogenic mutations in specific genes. However, directly linking observed DNAme change to the epigenetic mechanisms contributing to disease remains an open challenge.

Despite the known diversity in scale of differential DNA methylation features, no existing methods are designed to identify regions of differential methylation (DMRs) across the full range of scale from genome-wide methylation data. Instead, existing methods are designed to identify

DMRs on the scale of single genes or enhancers, which provides important but incomplete information towards understanding the full epigenetic architecture. This leaves a gap in using DNA methylation to understand the dynamics of co-regulated genes and regions in a broader epigenetic context.

Here we describe a method, *DMRscaler*, that accurately identifies regions of differential methylation that can span several basepairs up to those existing at much larger scales spanning many megabases of sequence across the global DNA methylation landscape. We demonstrate the dynamic range of our differential methylation caller by simulating DMRs varying in size from 100 bp to 1 Mb and testing its performance relative to existing methods. Additionally, we use real methylation data to test for sex differences in DNA methylation where *DMRscaler*, at its highest level calls the X-chromosome as a single differentially methylated feature while still calling small, gene-level DMRs on the autosomes. Finally, we show that pathogenic mutations in chromatin modifier genes are associated with differential methylation of large and highly conserved gene-clusters such as the *HOX* and *PCDH* gene clusters. By bridging the local and the global, *DMRscaler* can provide a broadened view of differential DNA methylation structure.

Implementation

The primary motivation for DMRscaler is to enable robust and accurate identification of regions of differential methylation that may exist at dramatically different scales. In DNA methylation data, DNA methylation is measured as the proportion of cytosines methylated at a given CpG site in the genome across all cells in a sample. This proportion is the β (beta) value of that site, with β=0 being completely unmethylated and β=1 being completely methylated. The distribution of β values for all CpGs across the genome follows a bimodal distribution (**Figure 3-1**). DMRscaler takes as input a set of CpG probes with their chromosome, genomic position, and pre-computed p-value for individual CpG level significance, and individual CpG level

p-value cutoff threshold for the desired Type I error control level. Estimation of the p-value cutoff for desired Type I error control should be done at the level of individual CpG level to avoid identifying DMRs that represent correlated blocks of CpGs that are not associated with the condition of interest. One method for estimation of this cutoff value, implemented with the *DMRscaler* package, is to repeat the individual CpG level significance testing with permutations of the case and control labels and compare the distributions of CpG significance values from these random permutations against the true case-control partition, however other methods could also be used. By working on an input of p-values, DMRscaler gives the user the flexibility to choose the statistical test that is most appropriate for their experimental design.

To identify regions that are characterized by differentially methylated CpGs (i.e differentially methylated regions or DMRs), DMRscaler uses a sliding window scheme (**Figure 3-2 A,B**). Windows are defined by a count of adjacent CpGs rather than by the span of the genomic region. The use of a count of adjacent CpGs for window definition makes *DMRscaler* agnostic to CpG density. This allows *DMRscaler* to scan regions with low CpG coverage, such as heterochromatin, that might be missed using a distance parameter between CpGs. Future iterations of the method may allow specification of fixed genomic interval widths for defining DMRs. However, a limitation here is that it is subject to the choice of methylation sites included on the currency DNA methylation chips.

Region-wide significance is taken to be the probability of obtaining within a window the set of CpG ranks or a set of more extreme ranks by random chance, given as a prior that the most significant CpG in the window has already been drawn. The null hypothesis then is that the ranks of the CpGs within a window are equally or less extreme than would be expected by a random draw from the complete set of CpG ranks given as a prior that the most significant CpG in the window has been drawn. The product of a sequence of hypergeometric tests is used to determine the region-wide significance of each window as described by the function

$$p_{region} = \prod_{i=1}^{m} hyper_{CDF}(k_i, n_i, N_i, K_i) \quad \text{(Eq. 1)}$$

where CpGs in the window, after excluding the most significant overlapping CpG from the window, are ordered from least to most significant going from $i = 1$ to $i = m$. Variables are defined as follows:

$$m = \text{total \# CpGs in the window}$$

$$k_i = \text{\# of CpGs in window with rank greater than or equal to } K_i$$

$$n_i = \begin{cases} m & \text{if } i = 1 \\ k_{i-1} - 1 & \text{otherwise} \end{cases}$$

$$N_i = \begin{cases} \text{total \# of CpGs in dataset} & \text{if } i = 1 \\ K_{i-1} - 1 & \text{otherwise} \end{cases}$$

$$K_i = \text{rank of } i\text{th CpG}$$

The hypergeometric cumulative distribution function, $hyper_{CDF}$, is set up to return the likelihood of obtaining k_i or more successes in n_i draws from a population of size N_i where there are K_i success cases total. At each step from $i = 1$ to $i = m$, the function determines the probability of having k_i or more CpGs of rank K_i or higher from n_i random draws in a population of N_i CpGs. The variables update at each step to account for how the likelihood has changed given the information contained in the previous step. The updates to the variables make the result of each hypergeometric test independent. The product of these independent tests then gives the region-wide significance value that effectively represents the probability of associating CpGs with the ranks observed, or ranks more extreme, by random chance in a window of the given size. This in effect then defines DMRs as regions that have a statistically significant association by adjacency of individually significant, by FDR or FWER control, CpGs. It should be noted that this procedure means that while region significance is linked to individual

CpG level significance, these two metrics of significance are distinct. For example, if a region's most significant individual level CpG significance value is p=0.01, but the region wide significance value is p=1e-12, then the region is almost certainly consisting of CpGs that are truly associated by adjacency (e.g. regulated in some respect as a unit relative to the set of all measured CpGs), however the determination of whether the differential methylation of this region is truly associated with the biological condition of interest should be based on individual CpG level significance.

Since there are nearly as many windows that can be tested for significance as there are CpGs included in the dataset, multiple testing must be accounted for to avoid excessive Type I errors. To do this, DMRscaler gives options to use Bonferroni correction procedure to control the family-wise error rate, or the Benjamini-Yekutieli procedure [37] to control the false discovery rate. With either of these the user supplies a region-wide significance threshold below which regions are considered significant. Both procedures are implemented so that the number of tests performed is equal to the number of measured probes below the user specified individual level CpG p-value cutoff. We observed conservative FDR control in simulations varying both the individual CpG level FDR threshold and the region-wide significance threshold (**Figure 3-3**)

In order to define DMRs that can vary dramatically in scale within the same analysis, we implemented the sliding window procedure iteratively increasing the size of the windows used to identify DMRs at each step of the iteration. The set of DMRs called at each step of the iteration is defined as a *layer* of the procedure, with layers named by the size of windows used for calling DMRs within that layer and indexed by the iteration step number. For example, if windows of 4 adjacent CpGs are used first to call DMRs, then *layer_1* (or $layer_1$) and the *4_adjacent_CpG_layer* are synonymous. An important step to accurately identify the scale of DMRs and avoid overinflation of DMR size is the inclusion of a step to integrate the results across these layers (**Figure 3-2 B,C**). This integration procedure works as the method iterates

from one layer to the next by testing whether a tentatively significant window called at the current layer remains statistically significant after the removal of CpGs from each overlapping DMR from the previous layer individually. For example, if a given window at the current layer has 100 CpGs and is considered tentatively significant, and there are 2 overlapping DMR from the previous layer with 20 CpGs and 30 CpGs, the DMR at the current layer is only retained if the 80 CpGs left after removal of the 20 CpG DMR from the first overlapping previous layer DMR are still considered significant as a region AND the 70 CpGs left after the removal of the 30 CpG DMR from the second overlapping previous layer DMR are still considered significant as a region. Otherwise, if either of these remaining 80 CpGs or 70 CpGs are not significant, then the current layer does not consider the 100 CpGs to be a DMR and instead the current layer is set to include the 20 CpG and 30 CpG DMRs from the previous layer, thereby propagating these DMRs from the previous layer.

It should be noted that no additional multiple testing correction is carried out to account for the number of layers used for identifying DMRs. As each layer is dependent on the same base layer of individual CpGs for estimation of significance, tests across layers are not independent and so our intuition is that it is reasonable to perform FDR or FWER correction only within layers. More rigorous statistical accounting for the dependency structure of this layer integration procedure is a challenge we have left as a future direction of study.

Here will elaborate on this procedure a little more formally. To begin, the smallest window size is used as a parameter to identify DMRs and build $layer_1$ of the output. As a note, the term 'layer' is used to describe the resulting set of DMRs constructed with a given window size parameter to suggest the relation of the results of each iteration of the algorithm. The construction of each successive layer either expands, adds, or retains DMRs from the *previous layer*, and so there is a hierarchical relation between DMRs across layers. DMRs in lower layers will always map to some DMR in upper layers, ensuring that $layer_i$ will always be a subset of

$layer_{i+1}$. To define the *next layer*, the following steps are repeated for windows of the next largest size: overlaying windows, identifying windows significantly enriched in differentially methylated CpGs, and merging significant windows (**Figure 3-2 A,B**). From the second layer onward, an additional step to integrate the results from the *previous layer* is performed. This is achieved by subtracting each *previous layer* DMR individually from any overlapping tentative DMR in the *next layer*, and retesting all generated reduced DMR sets in the *next layer* for significance (**Figure 3-2 C**). If a tentative DMR in the *next layer* remains significant following the subtraction of each overlapping *previous layer* DMR individually, then the tentative *next layer* DMR is retained in the new *integrated layer*. Otherwise, the overlapping *previous layer* DMRs are retained without change, replacing the tentative *next layer* DMR in the *integrated layer*. Additionally, any *previous layer* DMRs that did not overlap a DMR in the *next layer* are added to the *integrated layer*. In this way, at each iteration of the algorithm DMRs are either added, expanded, or consolidated from the *previous layer* to the *integrated layer* but never lost. As the algorithm proceeds, the *previous layer* is updated to the most recent *integrated layer* before the next step of integration is done with the *next layer*. Through iteratively calling DMRs using windows of increasing size and integrating the results, *DMRscaler* is able to identify DMRs that vary dramatically in terms of scale.

Methods

Cell culture. For Arboleda-Tham Syndrome data, fibroblast cell lines were derived from skin punch biopsies performed on the proband and one or both unaffected parents. This project was approved by the UCLA Institutional Review Board #11-001087. All individual level-data was de-identified prior to analysis. Fibroblast cell culture lines were created through the UCLA Pathology Research Portal and fibroblast cell lines were established and grown in DMEM (Gibco™), 10% FBS (Heat-inactivated Fetal Bovine Serum, Thermofisher), 1% Non-essential

Amino Acid (Gibco™) and 1% PenStrep at 37°C in 5% CO_2 incubators. Cell lines were tested for mycoplasma on a monthly basis.

DNA methylation studies. For Arboleda-Tham Syndrome methylation studies, DNA was extracted from patient-derived fibroblast cell lines. The specific mutation for each line is given in **Table 3-1**. DNA samples were bisulfite converted and run on the Illumina MethylationEPIC Array (850k EPIC array) as previously described[38] at the UCLA Neuroscience Genomics Core to generate idat files. QC on the resulting idat files was done using the MINFI package, and probes overlapping SNPs were removed[39]. After QC, 852,671 of 865,919 measured CpGs remiained, after removal of sex chromosome CpGs, 832,159 measured CpGs remained. Preprocessing and normalization of individual probes was done using background correction[40] and functional normalization[41].

Data sources. Publically available datasets of peripheral blood methylation data for control, Weaver Syndrome and Sotos Syndrome patients were downloaded from the Gene Expression Omnibus (GEO) resource[42,43] with accession number GSE74432[10].

Simulation. To demonstrate how *DMRscaler* distinguishes itself from other methods, we simulated differentially methylated regions (DMRs) ranging in size across several orders of magnitude (**Figure 3-4 A**).

DNA methylation measured on the Infinium HumanMethylation450 BeadChip (450K array) from whole blood for 53 controls from GEO (GSE74432) was used as a foundation for the simulation[10]. Real data was used as the foundation in order to capture the natural biological and technical variability present in DNA methylation array data. QC on the resulting idat files was

done using the MINFI package, and probes overlapping SNPs were removed[39]. After QC, 468,162 of 485,512 measured CpGs remained, and after removal of the sex chromosomes 456,514 measured CpGs remained. Preprocessing and normalization of individual probes was done using background correction[40] and functional normalization[41].

Regions for artificially introducing DMRs were selected at random across the genome but subject to the following constraints. DMRs specified as 0.1-1 kb in size were required to have at least 3 CpGs represented on the 450K array (CpGs), those 1-10 kb in size were required to have at least 6 CpGs, those 10-100 kb in size were required to have at least 9 CpGs, those 0.1-1 Mb in size were required to have at least 12 CpGs. Additionally, to avoid miscounting, DMRs were introduced such that they were spaced at least 10 CpGs apart from any other introduced DMR. Distribution of CpG counts in simulated DMRs against simulated DMR sizes are shown in **Figure 3-5.**

All 450k array samples used were from control whole blood DNA, so for each run of the simulation samples were pulled at random into one of two groups, Group1 and Group2. Each group consisted of 8 samples drawn without replacement from the pool of 53 samples.

Before artificially introducing the DMRs to the real data matrix, a proportion of CpGs within each DMR, excluding the first and last CpGs, specified by the *noise* parameter were randomly masked and kept at their original β values. This was done to model the variability of methylation state of neighboring CpGs in real data. Values for the *noise* parameter tested were 0, 0.25 and 0.5, corresponding to 0%, 25%, and 50% of CpGs overlapping a simulated DMR being masked. Then, the mean β value of CpGs within each DMR was measured for Group1 and Group2. The β values of the group with the greater mean β value would have all non-masked CpGs inflated by an amount specified by the *Δβ* parameter. Simulations were run with *Δβ* values of 0.1, 0.2, and 0.4 to model small, modest, and large effect DMRs respectively.

If this resulted in any samples having a β value greater than 1, the β values for that CpG were divided by the max β value for that CpG to bring values back to the range of 0-1.

Following the introduction of artificial DMRs into the dataset, *DMRscaler*, *bumphunter* [44], *comb-p* [45] and *DMRcate* [46] were run on the dataset and the results tabulated. *DMRscaler* was tested using a window_size_vector of c(4, 8, 16, 32, 64) adjacent CpGs, locs_pval_cutoff corresponding to the individual level CpG p-value at which FDR < 10% is achieved, region_signif_cutoff = 0.01 corresponding to the region level significance threshold for calling a region as a DMR after multiple testing correction, and region_signif_method = "benjamini-yekutieli" specifying the benjamini-yekutieli procedure as the method for Type I error control. *Bumphunter* was tested with MaxGap = 1e6 with loess smoothing enabled. *Comb-p* was tested with dist = 1e6, step = 5000, seed = 1e-3, region-filter-p = 0.1 (**Figure 3-6**). *DMRcate* was tested with lambda = 1e6, C=2000. Parameter sets for methods were chosen to facilitate identification of larger DMRs for output more comparable to *DMRscaler*.

To benchmark each method's performance several metrics were used including proportion of CpGs in DMRs that are differentially methylated, precision, recall, specificity, F1, Matthew's correlation coefficient (MCC), and the area under the precision recall curve (AUCPR). These metrics were recorded for analysis at the feature, basepair, and CpG probe level, where feature level assessment treated each simulated DMR as a single positive feature, the basepair level treated each basepair overlapping a simulated DMR as a positive feature, and the CpG probe level treated each CpG probe overlapping a simulated DMR as a positive feature. The basepair and CpG level assessments are based on direct counts of true and false positives and negatives. The feature level assessment was conducted following the framing of the problem for measuring precision and recall for time series proposed by Tatbul and colleagues, which is generally appropriate for other forms of range data when identification of individual features is of interest [47]. Each simulated DMR was considered as a true feature with true positive (TP) and

false negative (FN) attributes. Each called DMR was considered a called feature with TP and false positive (FP) attributes. Called DMRs were ordered by their p-value for all methods. The precision-recall curve was generated by measuring precision and recall with stepwise inclusion of the next highest scoring or most significant called DMR. At the n-th step precision and recall were measured as

$$Precision = \frac{TP_{Called\_DMRs}}{TP_{Called\_DMRs} + FP_{Called\_DMRs}} \quad \text{(Eq. 2)}$$

$$TP_{Called\_DMRs} = \frac{\sum_{i=1}^{n} TP_{Called\_DMR_i}}{n} \quad \text{(Eq. 2.1)}$$

$$FP_{Called\_DMRs} = \frac{\sum_{i=1}^{n} FP_{Called\_DMR_i}}{n} \quad \text{(Eq. 2.2)}$$

$$Recall = \frac{TP_{Simulated\_DMRs}}{TP_{Simulated\_DMRs} + FN_{Simulated\_DMRs}} \quad \text{(Eq. 3)}$$

$$TP_{Simulated\_DMRs} = \frac{\sum_{j=1}^{m} TP_{Simulated\_DMR_j}}{m} \quad \text{(Eq. 3.1)}$$

$$FN_{Simulated\_DMRs} = \frac{\sum_{j=1}^{m} FN_{Simulated\_DMR_j}}{m} \quad \text{(Eq. 3.2)}$$

where $TP_{Called\_DMR_i}$ is the proportion of the i-th called DMR overlapping simulated DMR regions, $FP_{Called\_DMR_i}$ is the proportion of the i-th called DMR not overlapping a simulated DMR region, $TP_{Simulated\_DMR_j}$ is the proportion of the j-th simulated DMR overlapping any of the 1 to n-th Called DMRs, $FN_{Simulated\_DMR_j}$ is the proportion of the j-th simulated DMR overlapping any of the 1 to n-th Called DMRs, n is the number of most significant called DMRs used at the n-th step, and m is the total number of simulated DMRs. The feature level measure of precision and recall gives equal weight to each simulated DMR so that large simulated DMRs do not dominate the signal.

In addition to precision and recall, several other metrics that were included in a recent benchmark of DMR callers by Mallik et al. [48] were used to assess method performance in the simulation. All of these metrics were measured on the set of called DMRs that had an adjusted region-wide significance p < 0.01 for each method. Specificity is a measure of the true negative rate, and is measured as the proportion of true negatives as a fraction of total negative features, and is calculated by the equation:

$$Specificity = \frac{1 - FN_{Simulated\_DMRs}}{N} \quad \text{(Eq. 4)}$$

False Discovery Rate as the inverse of precision gives expected proportion of false results and is given by:

$$FDR = 1 - Precision \quad \text{(Eq. 5)}$$

F1 is a measure of a test's accuracy and is given by:

$$F1 = 2 * \frac{(Precision * Recall)}{Precision + Recall} \quad \text{(Eq. 6)}$$

F1 ranges from 0 for worst accuracy to 1 for perfect classification. Finally, the Matthew's correlation coefficient, which is a measure of correlation between predicted and true class labels and is given by:

$$MCC = \frac{\sqrt{Recall * Specificity * Precision * \frac{TN_{Called_{DMRs}}}{N_{Calleld_{DMRs}}})}}{\sqrt{(1 - Recall) * (1 - Specificity) * (1 - Precision) * (1 - \frac{TN_{Called_{DMRs}}}{N_{Calleld_{DMRs}}})}} \quad \text{(Eq. 7)}$$

MCC values at +1 indicate perfect classification, 0 indicates equivalence with random classification, and -1 indicates perfect misclassification.

Rare Disease Data Analyses

For each real data analysis, *DMRscaler*, *bumphunter, comb-p*, and *DMRcate* were used to call DMRs. *DMRscaler* was tested using a window_size_vector of c(4, 8, 16, 32, 64) adjacent CpGs, locs_pval_cutoff corresponding to the individual level CpG p-value at which FDR < 10%

is achieved, region_signif_cutoff = 0.01 corresponding to the region level significance threshold for calling a region as a DMR after multiple testing correction, and region_signif_method = "benjamini-yekutieli" specifying the benjamini-yekutieli procedure as the method for Type I error control. *DMRcate* was tested with default parameters, as well as with lambda = 1e6, C=2000 to capture larger DMRs for output more comparable to *DMRscaler*. *Bumphunter* was tested with default parameters, as well as with MaxGap = 1e6 with loess smoothing enabled.

For the sex analysis, DNA methylation measured on the Infinium HumanMethylation450 BeadChip (450K array) from whole blood for 53 controls from GEO (GSE74432) with 29 female and 24 male samples was used [10]. QC on the raw idat files was done using the MINFI package, and probes overlapping SNPs were removed[39]. After QC, 468,162 of 485,512 measured CpGs remained. Preprocessing and normalization of individual probes was done using background correction[40] and functional normalization [41]. Individual level differential CpG significance between female and male samples was measured using the Wilcox test to serve as input for *DMRscaler* and *comb-p*.

For Arboleda-Tham Syndrome sample analysis, DNA methylation was measured on the Illumina MethylationEPIC Array (850k EPIC array) with 8 cases and 12 controls. QC on the resulting idat files was done using the MINFI package, and probes overlapping SNPs were removed [39]. After QC, 852,671 of 865,919 measured CpGs remained, after removal of sex chromosome CpGs 832,159 measured CpGs remained. Preprocessing and normalization of individual probes was done using background correction [40] and functional normalization [41]. Individual level differential CpG significance between female and male samples was measured using the Wilcox test to serve as input for *DMRscaler* and *comb-p*.

For Weaver analysis, DNA methylation measured on the Infinium HumanMethylation450 BeadChip (450K array) from whole blood for 8 patients with *EZH2* mutations and 53 controls from GEO (GSE74432) was used [10]. This data comes from a study that found an epigenetic

signature specific to Sotos syndrome from *NSD1* mutations using Weaver syndrome samples as a negative control for their classifier [10]. More recently, this data has been used to identify an epigenetic signature specific to Weaver syndrome [9]. QC on the raw idat files was done using the MINFI package, and probes overlapping SNPs were removed [39]. After QC, 468,162 of 485,512 measured CpGs remained, and after removal of the sex chromosomes 456,514 measured CpGs remiained. Preprocessing and normalization of individual probes was done using background correction[40] and functional normalization [41]. Individual level differential CpG significance between female and male samples was measured using the Wilcox test to serve as input for *DMRscaler* and *comb-p*.

For Sotos syndrome analysis, DNA methylation measured on the Infinium HumanMethylation450 BeadChip (450K array) from whole blood for 38 patients with *NSD1* mutations and 53 controls from GEO (GSE74432) was used[10]. QC on the raw idat files was done using the MINFI package, and probes overlapping SNPs were removed[39]. This comes from the same study as the Weaver syndrome data [10]. After QC, 468,162 of 485,512 measured CpGs remained, and after removal of the sex chromosomes 456,514 measured CpGs remiained. Preprocessing and normalization of individual probes was done using background correction[40] and functional normalization[41]. Individual level differential CpG significance between female and male samples was measured using the Wilcox test to serve as input for *DMRscaler* and *comb-p*.

Syndrome DMR Overlap Analysis

To test for overlapping regions of differential methylation between Arboleda-Tham, Sotos, and Weaver syndrome, the number of measured CpGs considered for DMR detection was downsampled to include only those CpGs measured on both the Infinium

HumanMethylation450 BeadChip (450K array) and the Illumina MethylationEPIC Array (850k EPIC array). This left 425,733 measured CpGs for calling DMRs.

Overlaps between DMRs were counted between syndromes as was the overlap of gene sets. Gene set overlaps were considered separately to identify genes that may be commonly differentially methylated but identified by non-overlapping regions of the gene, something the direct DMR overlap measure would miss.

To test whether CpGs identified as belonging to DMRs are enriched between syndromes, that is whether membership of a CpG to a DMR in one syndrome makes it more or less likely to also belong to a DMR in another syndrome, we computed the odds ratios (OR). The OR was calculated by forming a 2x2 contingency table with counts of CpGs belonging to DMRs in both syndromes, CpGs belonging to one and not the other, and CpGs belonging to DMRs in neither.

Results

DMRscaler Overview

Our goal in developing *DMRscaler* was to have a method capable of accurately identifying regions that demonstrate differential methylation across the full range of epigenetic scale, from small-promoter to whole-chromosome scale features. The major bottleneck to this goal is that regions of differential methylation show significant variability in methylation state between neighboring CpGs. For example, nearly 20% of neighboring CpGs between 0.5 - 1.0 kb away have a difference in the proportion of methylation greater than 50% (**Figures 3-7, 3-8**). When trying to identify DMRs that may span larger genomic regions, such as gene clusters, this variability makes the trivial method of taking contiguous blocks of significant CpGs as the DMRs ineffective. One approach to resolve this issue of high variability is to smooth differential methylation sites based on significance across adjacent CpGs or over some specified genomic

interval. However, the smoothing approach is sensitive to the choice of bandwidth parameter used for the smoothing window. Windows that are too small will fail to connect features over larger gaps, windows that are too large will result in excessively broad DMRs. Smoothing alone is therefore inappropriate when features are expected to vary dramatically in terms of scale. To capture potentially noisy features that may vary in size by several orders of magnitude, from the basepair to multi-megabase scale, we need a method that is both robust to noise and that can accurately determine the feature's size.

To address these limitations in determining the size of a DMR, *DMRscaler* uses an iterative sliding window over the genome (**Figure 3-2 A,B**), represented as a partially-ordered set of measured CpGs, and implements an integration step between each iteration of the sliding window (**Figure 3-2 C**). The windows at each step identify the set of regions that are enriched in CpGs with significantly different methylation values between cases and controls. By binning CpGs into windows and testing these windows for enrichment in significant CpGs (Eq. 1), the algorithm is robust to noise caused by variability in methylation of neighboring CpGs. To address the bias in feature size introduced by preselecting a window size parameter, *DMRscaler* calls significant windows iteratively with a variable increasing size parameter and integrates the result of each iteration with the results from the previous iterations. The integration step (**Figure 3-2 C**) is used between the previous (lower) layer, built from smaller windows, and the current (upper) layer to determine which features in the upper layer are already adequately represented by lower layer features and which upper layer features capture a statistically significant association missed by the lower layer features. If an upper layer feature captures a statistically significant association missed in the lower layer then that upper layer feature is retained and resolved with any overlapping lower layer features, otherwise the overlapping lower layer representation is carried through unmodified. For a more detailed description, see implementation.

DMRscaler provides a solution to the problem of identifying DMR features across the full range of epigenetic feature sizes, whether at the basepair level or across entire chromosomes. The integration of results across iterations of the windowing procedure *DMRscaler* implements is a novel mechanism for defining DMRs that could be generalized to other epigenetic features or one dimensional data where discontinuity in components defining a feature of interest is expected and where features of interest may exist at dramatically different scales.

Comparison of DMRscaler with existing methods

We next benchmarked *DMRscaler* to three commonly used methods in identification of differentially methylated regions: *bumphunter*[44], *comb-p*[45], and *DMRcate*[46] (**Table 3-2**). One significant difference between these methods is that our method, *DMRscaler,* and *comb-p* take pre-computed p-values as input while *bumphunter* and *DMRcate* use a t-test to determine individual level CpG significance. We observed that when running with a small sample size (n=8 per group) there is poor correlation between the significance of differential methylation determined by the non-parametric Wilcoxon and t-test (**Figure 3-9**). Since one of our goals was to develop a method that could detect DMRs in studies that compare rare disease datasets, the flexibility to choose the most appropriate statistical test for individual CpG significance based on experimental design and sample size constraints was desirable. While the t-test is appropriate when the sample size is sufficiently large (n > 30) or the sampling distribution is approximately normal, differential methylation analysis in small samples breaks these assumptions and therefore in our analysis of rare disease datasets the flexibility to use the Wilcoxon test was important.

The second difference between the differential methylation callers compared here (**Table 3-2**) lie in their modeling to identify differentially methylated regions. Briefly, *bumphunter* uses a linear regression model to identify CpG sites that are differentially methylated between case and

control conditions. Then to detect DMRs, *bumphunter* identifies stretches of adjacent CpGs that are above a specified significance threshold after smoothing. However, the methylation landscape of adjacent CpGs is complex, with CpGs with high, intermediate and low β values mixed together making definition of large and contiguous regions of differential methylation challenging (**Figure 3-7, 3-8**). *Comb-p* uses the Stouffer-Liptak method for p-value correction and then groups significant CpGs within a window or window interval defined by the *dist* and *step* parameters. *DMRcate* is similar to *bumphunter* in that it also implements linear modeling (**Table 3-2**). *DMRcate* uses a Guassian smoothing function on M transformed β values to identify DMRs in genome-wide data. This provides the user with control of a bandwidth parameter, lambda, and control parameter, C, that can be used to identify larger regions of differential methylation. However, the behavior of *DMRcate* at larger bandwidth is poorly defined and the size of DMRs returned tends to be sensitive to parameter choice. For an in-depth review of methods see[48].

The design of the *DMRscaler* method has several unique features that allow it to more accurately identify larger co-regulated regions. First, it deals with the intrinsic variability in methylation distribution across the genome by binning adjacent CpGs into windows before assigning significance. Second, *DMRscaler* integrates the results from layers of windows defined with a series of window sizes to consider regions that are dramatically different in scale as potential regions of differential methylation. To accommodate a variety of study designs and constraints, *DMRscaler* operates on pre-computed p-values for individual level CpG significance. While here we use the Wilcox test due to the small sample size of our rare disease cohorts, other methods of generating p-values can be used, for instance to model the effect of covariates. Together, these features allow for the robust detection of differentially methylated regions across a large dynamic range, spanning basepair to megabase resolution and allow for

detection of novel regions that are differentially methylated in rare disease cohorts and between other biological conditions such as chromosomal sex.

DMRscaler accurately captures the scale of epigenetic features from basepair (bp) to megabase (Mb) size in simulated methylation data

Except where stated otherwise, in the following sections DMRs are assumed to be those in the most inclusive top layer, Layer 5, which is built using all lower layers and is meant to be the most accurate representation of DMR features.

To benchmark our method against existing methods, we compared *DMRscaler* to *bumphunter, comb-p,* and *DMRcate* on several metrics that highlight behavior of calling DMRs across a wide range of simulated DMR sizes. These metrics include: the correlation between simulated DMRs and DMRs called by each method across a wide range of simulated DMR sizes, the mapping value or the degree to which each method was able to represent individual simulated DMRs as single unified features, and the run time of each method. Additionally we used standard metrics of evaluation such as precision, recall, specificity, F1, Matthew's correlation coefficient (MCC), and area under the precision-recall curve (AUCPR) to assess method performance.

We first simulated DMRs in methylation data from control blood samples (GSE74432)[10] ranging in size from 100 bp to 1 Mb (**Figure 3-4 A**, see method for details). In our simulation, we modeled the situation observed in real data where neighboring CpGs often have distinct methylation states with a *noise* parameter that represents the proportion of CpGs within simulated DMRs, excluding the first and last CpGs, that are left non-differentially methylated between the randomly selected samples placed in Group1 and Group2. In our simulations we tested *noise* parameter values of 0%, 25% and 50%. The $\Delta\beta$ parameter was used to control the magnitude of differential methylation with simulated DMRs, where $\Delta\beta$ is the difference in

methylation proportion at non-noise CpGs introduced between the samples in Group1 and Group2. Simulations were run with the *Δβ* parameter set to values of 0.1, 0.2, and 0.4 to model small, modest, and large effect sizes for differential methylation respectively. Results from simulations run with each combination of these parameters are included in **Figures 3-10,3-11,3-12**. The relative performance and behavior of methods was consistent across simulations run with each of these parameter combinations, so for space in the main text and figures we display results and report metrics from simulations run using *noise*=50% and *Δβ*=0.2.

DMRscaler was able to accurately call the size of the simulated DMRs (pearson's r = 0.94) relative to *bumphunter* (pearson's r = 0.04), *comb-p* (pearson's r = 0.69), and *DMRcate* (pearson's r = 0.85) (**Figure 3-4 B**). *DMRscaler* preserves a strong 1-to-1 relation between simulated and called DMRs, with 85% of simulated DMRs accurately called by *DMRscaler* with a 1-to-1 relation, compared with 19% for *bumphunter*, 44% for *comb-p*, and 69% for *DMRcate* (**Figure 3-4 C**).

To measure performance of our differential methylation caller, we calculate the AUCPR for each test. AUCPR combines a measure of precision of features called (ratio of true feature called to all features called) and recall (ratio of true features called to total number of true features) into a single value, with AUCPR = 0 representing no classification and AUCPR = 1 representing perfect classification of all features with no false positives. In our simulation, *DMRscaler* had an AUCPR of 0.79, *bumphunter* had an AUCPR value of 0.11, *comb-p* had an AUCPR of 0.34, and *DMRcate* had an AUCPR of 0.65 (**Figure 3-4 D**, see methods for details on AUCPR calculation). The low AUCPR of *bumphunter* is consistent with the low, slightly negative correlation observed between simulated and called DMR regions (**Figure 3-4 B**). This weak correlation is due to the fact that *bumphunter* has a strict requirement that significant differentially methylated CpGs are adjacent in order to belong to a common DMR and therefore breaks up simulated DMR features into many smaller features. The low AUCPR of *comb-p* is

the result of a low recall rate of features that are much smaller than the size set by the *dist* and *step* parameters. Setting lower values for the *dist* parameters increases the ability to detect smaller DMR features but at the expense of detecting larger DMR features (**Figure 3-6**), and at very large values for *dist* the run time becomes prohibitive especially as smaller step values are used (**Figure 3-6**). *DMRcate* had a reasonably high AUCPR, however there is a bias in the size of DMRs called based on the choice of the bandwidth parameter lambda, and the control parameter C. Specifically, there is an excess of false calls of DMRs around 1 Mb and 1 kb (**Figures 3-4 B** and **3-13**) which is related to the choice of bandwidth parameter λ (set to 1 Mb) and the scaling parameter C (ratio of λ/C set to 500) (**Figure 3-14**). Our data suggests that *DMRcate* is able to identify larger DMRs but also that called DMR size is sensitive to parameter choice for the lambda and C parameters. This is supported by the shape of the precision-recall curve for *DMRcate* that shows a modest drop in precision as recall increases, suggesting *DMRcate* incurs a steeper false positive penalty compared with *DMRscaler*.

While the Wilcoxon test was used to generate p-values for individual level CpG significance for the simulation and real data analysis, we note that the performance and behavior of *DMRscaler* in the simulation was comparable when the T-test was used (**Figure 3-16**). Additionally, while these results focus on the top layer of results from *DMRscaler*, behavior at each lower layer is shown in **Figure 3-16**.

Comparing each method on each combination of Δβ (0.1, 0.2, and 0.4) and noise (0%, 25%, and 50%) parameters on metrics of precision, recall, specificity, F1, MCC, and AUCPR, *DMRscaler* consistently outperformed competing methods on each metric, except specificity where bumphunter was consistently the best performing method though the difference between methods on specificity was generally small (**Table 3-3**). These results further demonstrate that *DMRscaler* performs well for accurately calling DMRs across a wide range of feature size.

While *DMRscaler* performs well compared to other methods at the task of identifying DMRs across a wide range of scales, the method also performs well in terms of computational time to the other methods that are time efficient with larger window size analogous parameters. On average *DMRscaler* required 30 seconds to 1 minute to complete a run, *bumphunter* required around 1 to 3 minutes, and *DMRcate* only required around 10 seconds to call DMRs. *Comb-p,* which uses a sliding window mechanism similar to *DMRscaler,* required an hour to complete each run with the given parameter set (**Figure 3-4E**).

The simulation results show that *DMRscaler* reconstructs the scale of DMR features more accurately than other methods across a wide range of DMR feature sizes as measured by called and simulated DMR size correlation, mapping value, and precision and recall. Additionally on other measures of performance including specificity, F1, MCC, and AUCPR, *DMRscaler* consistently performs well compared to other methods on simulated datasets with DMRs that vary widely in terms of scale.

Differential methylation between 46,XX and 46,XY individuals captures chromosome-wide and gene specific regulatory features in empiric data.

To test our hypothesis in real-world DNA methylation data, we sought to determine whether our method could capture both small regions of autosomal differential methylation as well as chromosome-wide features such as X-chromosome inactivation. Therefore, our test case is the natural occurrence of X-inactivation in females, where one copy of the X-chromosome is largely inactivated by the action of the lncRNA *Xist*[18,19]. This process of inactivation is correlated with a striking chromosome-wide difference in DNA methylation between males and females on the X-chromosome (Figure 3A, top) as compared to the autosomes, e.g. chromosome 2 where the size of differentially methylated regions span 103 bp to 873 bp (**Figure 3-17A, bottom**). Across all chromosomes, the size of DMRs called by

DMRscaler spans 98 bp to 152 Mb, representing a 1.5 million-fold difference in the scale of DMRs detected by *DMRscaler*.

With the visual intuition of the scale of differential methylation between sexes from **Figure 3-17A**, we next compared the result of differential methylation analysis using *DMRscaler*, *bumphunter*, *comb-p* and *DMRcate*. *DMRscaler* was the only method that consolidated the observed differential methylation into a single DMR that spanned 98% of the X-chromosome (**Table 3-4**, **Figure 3-17B**, **Figure 3-18**). Even with the maxWidth parameter set to 1 Mb, *Bumphunter* reported 1,162 unique DMRs on the X-chromosome with a median width of 531 bp (IQR: 1 bp - 1.21 kb) (**Figure 3-19**), likely due to a lack of mechanism for spanning non-differentially methylated CpGs. With a standard parameter set of dist = 1 kb, step = 100bp, *comb-p* reported 2,390 unique DMRs on the X-chromosome with a median width of 2 bp (IQR: 2 bp - 963 bp) (**Table 3-4**, **Figure 3-19**). With a wider parameter set of dist = 1 Mb, step = 100 kb, *comb-p* called 19 unique DMRs on the X-chromosome with a median width of 3.15 Mb (IQR: 512 kb - 8.54 Mb) (**Table 3-4**, **Figure 3-19**) . *DMRcate* with default settings reported 1,178 unique DMRs on the X-chromosome with a median width of 1.09 kb (IQR: 616 bp - 1.68 kb). When *DMRcate* was provided with a larger bandwidth parameter (lambda = 1 Mb, C = 2000) it improved in consolidating the DMRs, but still reported 15 unique DMRs (median width: 3.95 Mb, IQR: 1.00 Mb - 17.89 Mb). For complete distributions of called DMR sizes see **Figure 3-20**.

DMRscaler iteratively calls DMR-like regions using windows of increasing size while integrating the results of each iteration into the next layer of DMRs. While the top-most layer is the primary output of *DMRscaler*, this procedure produces a nested hierarchy of DMRs when considering the list of results across all layers that allows for a nuanced view of the differential methylation architecture. In **Figure 3-17B**, a subset of this hierarchy within the X-chromosome is shown. Here DMRs called at layer 1 are consolidated in the DMRs in layer 2, and then those DMRs in layer 2 are consolidated into DMRs in layer 3, consolidation of DMRs in layer 3 into

layer 4 results in the final consolidation of DMRs into a single feature spanning the entire X-chromosome .

While the entirety of the X-chromosome can be considered a differentially methylated feature, it has been well established that there is a small subset of genes on the X-chromosome that escape X-inactivation and DNA methylation[49]. The expectation when comparing methylation between females and males is that X-inactivation would result in differential methylation between sexes, with hypermethylation and to a lesser extent hypomethylation across the entire X-chromosome in females compared to males[50]. Therefore, we expect regions where the Δβ between the two groups is at or near zero would be enriched in regions that escape X-inactivation due to a relative lack of differential methylation at these sites. An example of one such region is the gap of two DMRs that persists until the integration between layer 3 and layer 4 which occurs at chrX: 71,459,274 - 71,521,494 which corresponds to the gene *RPS4X* (**Figure 3-17C**, **Figure 3-21**), which is known to escape X-inactivation [51]. To test whether this trend of gaps in DMRs mapping to held more generally across regions escaping X-inactivation, we performed an enrichment test for CpGs that overlapped genes known to escape X-inactivation and CpGs overlapping gaps in DMRs called at each layer of *DMRscaler*'s output. A consensus of genes known to escape or be silenced in X-inactivation reported in a 2015 study by Balaton et al. was used for the enrichment test [52]. In layers 1, 2, 3, and 4 which were defined by windows of 4, 8, 16, and 32 adjacent CpGs respectively, the odds ratio between CpGs overlapping gaps between DMRs and CpGs overlapping genes that escape X-inactivation were OR = 7.57 (95% CI 6.38 - 8.99; p-value = 1.04e-134 Fisher's exact test), OR = 7.24 (95% CI 6.07 - 8.65; p-value = 5.93e-100 Fisher's exact test), OR = 51.99 (95% CI 30.38 - 90.33; p-value = 4.34e-77 Fisher's exact test), and OR = 160.44 (95% CI 25.42 - 6,396.92; p-value = 5.93e-100 Fisher's exact test) respectively. At layer 5 no enrichment was detected, as the whole X-chromosome was consolidated into a single feature. *Bumphunter* similarly displayed

substantial enrichment, with odds ratios estimated between OR ≈ 10-20, however, as noted earlier, bumphunter could not consolidate DMRs on the X-chromosome to identify the whole of the X-chromosome as differentially methylated. *Comb-p* and *DMRcate* each observed much smaller associations of gaps between DMRs and genes escaping X-inactivation with ORs ≈ 1-3 (**Table 3-5**). These results demonstrate that while the top layer DMR, which spans the X-chromosome, correlates most intuitively with the phenomenon of X-inactivation, the exploration of the hierarchical structure of complex DMRs that is enabled by *DMRscaler* can reveal biologically meaningful features such as patterns of genic escape from X-inactivation.

The complex hierarchical relation of DMRs within the X-chromosome contrasts with the DMRs of the autosomal chromosomes. DMRs on the autosome show little to no branching, which implies that these DMRs are stable at each iteration of the algorithm (**Figure 3-18**). A genome view of one such DMR at chr9: 84,302,344-84,304,414 highlights this stability, where a feature identified as a DMR at the first layer of the algorithm is stable through each subsequent iteration (**Figure 3-17D**, **Figure 3-21**). The gene *TLE1* overlaps this DMR and has previously been identified as an autosomal gene that is differentially methylated between males and females[53,54].

The results of the differential methylation analysis between sexes highlights the utility of *DMRscaler* in identifying differential methylation features that exist at dramatically different scales in real data. This ability distinguishes *DMRscaler* from existing methods which either are unable to identify larger DMRs while preserving the stability of smaller DMRs, as in *DMRcate* and *comb-p*, or tend to fragment larger DMR into many smaller features, as in *bumphunter*. A brief analysis of the hierarchical structure that results from *DMRscaler*'s layer merging mechanism reveals how *DMRscaler* can capture biologically meaningful structure within a DMR, such as escape from X-inactivation. This ability to represent DMR structure more completely

highlights *DMRscaler*'s potential value as a tool for exploring the interactions between features of epigenetic regulation at different scales.

Rare chromatin modifier syndromes contain regions of differential methylation spanning gene clusters critical for development.

Next we analyzed DNA methylation datasets from several rare diseases of chromatin modifier genes to see whether *DMRscaler* revealed novel DMR features that might otherwise be missed by existing methods. Except where stated otherwise, in the following sections DMRs are assumed to be those in the most inclusive top layer, Layer 5, which is built using all lower layers and is meant to be the most accurate representation of DMR features.

First, we compared the DNA methylation profile from fibroblasts from Arboleda-Tham syndrome patients to control samples. This analysis consisted of 20 samples, with 8 patients and 12 controls (**Table 3-1**). All patients were previously reported by Kennedy, et al[7]. In our analysis, *DMRscaler* identified 390 unique DMRs with a median width of 144.59 kb (IQR: 21.1 kb - 481.2 kb), resulting in a total genomic coverage of 4.9% (151.35 Mb) (**Table 3-6**, **Figure 3-22**). Over the *HOXB* gene cluster three unique DMRs were identified. The first and second DMRs overlap regions of *HOXB2, HOXB3* and *HOXB4* and is hypomethylated in Arboleda-Tham Syndrome patients relative to controls. The second overlaps part of *HOXB5 and HOXB6* and is also hypomethylated in Arboleda-Tham Syndrome patients. The third spans *HOXB9* and is hypermethylated in Arboleda-Tham Syndrome patients relative to controls (**Figure 3-23A,B**, **Figure 3-24**). *Bumphunter* calls many more DMRs over this region that are highly fragmented, including regions missed by *DMRscaler*. This is likely due to the regions called by bumphunter having substantial variance that the Wilcox test used to pre-compute p-values for *comb-p* and *DMRscaler* being more conservative than the t-test used by bumphunter. *Comb-p* with the large distance parameter of 1 Mb calls the entire region as

differentially methylated. The relatively large coverage of the genome by DMRs is driven primarily by multi-megabase scale DMRs identified spanning relatively gene sparse regions, which other methods are unable to consolidate, with the exception of *comb-p* using a distance parameter of 1 Mb (e.g. **Figure 3-23C,D**, **Figure 3-24**).

Weaver syndrome (MIM# 277590), is a rare overgrowth disorder that is caused by de novo mutations in *EZH2*, a histone methyltransferase. Comparing Weaver syndrome patient samples to controls, *DMRscaler* identified 226 unique DMRs with a median width of 8.88 kb (IQR: 1.92 kb - 30.04 kb). These regions comprised a total of 0.40% (12.34 Mb) of the genome (**Table 3-7**, **Figure 3-25**).

Over the *HOXA* gene cluster, *DMRscaler* identified three distinct DMRs associated with Weaver Syndrome. The first spans *HOXA1-HOXA2* and is modestly hypermethylated in Weaver syndrome, the second covers *HOX5* and the last two exons of *HOX6* and is hypomethylated in Weaver syndrome cases relative to controls. The third DMR covers the first exon of *HOXA10,* as well as *HOXA11,* and *HOXA13*. This third DMR is generally weakly hypermethylated in Weaver Syndrome, with a small but significant region of hypomethylation just upstream of *HOXA11*. The other methods all report DMRs overlapping these clusters, however they are either fragmented or overly broad (**Figure 3-26 A,B**, **Figure 3-27**).

Finally, we also analyzed Sotos Syndrome (MIM# 117550), an overgrowth syndrome caused by truncating and missense mutation in the nuclear receptor binding SET domain protein 1 **(***NSD1)* gene[55]. Analysis with DMRscaler identified 1776 unique DMRs with a median width of 555.13 kb (IQR: 156 kb - 1.40 Mb), covering 71% of the genome (2.17 Gb), a similar degree of coverage was seen with *DMRcate* where 282 DMRs spanned 77% of the genome (**Table 3-8**, **Figure 3-28**). We identified three unique DMRs at the 32 Adj CpG layer that span gene clusters of protocadherins. These DMRs caused by mutations in *NSD1* cover the neighboring *Protocadherin* (*PCDH*) gene cluster *PCDHA*, *PCDHB*, and *PCDHGB* (**Figure**

3-26D,E, **Figure 3-27**), which encode large transmembrane proteins that are critical for a diverse range of processes ranging from cell-signaling to dendritic arborization[56]. One DMR spans the first exons of *PCDHA1-PCDHA12*, another spans from *PCDHB2* to *PCDHB19P*, and the third covers the first exons of *PCDHGA3-PCDHGA12* and *PCDHB1-PCDHGC5*. All of the DMRs covering these *PCDH* clusters are hypermethylated in Sotos Syndrome relative to controls, though it is notable that the β values of CpGs across these clusters are highly variable reflecting an example of the neighboring CpG heterogeneity described earlier (**Figure 3-26D**, **Figure 3-27**, **Figure 3-7** , **Figure 3-8**). Notably, only DMRcate with parameters λ=1 Mb, and C=2000 was also able to call a DMR over this region, however it lacks a mechanism to see the interior structure that shows that each of these three clusters is separated by regions of non-differentially methylated CpGs that is captured by *DMRscaler*'s hierarchical output.

These results in rare chromatin modifier syndromes highlight *DMRscaler's* utility in identifying patterns of differential methylation that exist over broader genomic features such as gene clusters.

Analysis of overlapping regions of differential methylation

Following analysis of each syndrome individually, we asked whether there was evidence of shared regions differentially methylated between Arboleda-Tham, Sotos, and Weaver syndrome. Between Arboleda-Tham and Sotos syndrome, we identified 652 regions with overlapping DMRs (77.3% of total DMRs for Arboleda-Tham, 4.7% of total DMRs for Sotos), and 458 genes overlapped by some DMR in both syndromes (11.4% of genes overlapping a DMR in Arboleda-Tham syndrome, 3.1% of genes overlapping a DMR in Sotos syndrome). Between Arboleda-Tham and Weaver syndrome, we identified 48 regions with overlapping DMRs (1.3% of total DMRs for Arboleda-Tham syndrome, 14.1% of total DMRs for Weaver syndrome), and 39 genes overlapped by some DMR in both syndromes (13.2% of genes

overlapping a DMR in Arbolelda-Tham syndrome, 5.6% of genes overlapping a DMR in Weaver syndrome). Between the two growth disorders, Sotos and Weaver syndrome, we identified 414 regions (0.7% of total DMRs for Sotos, 91.0% of total DMRs for Weaver) and 282 genes overlapped by some DMR in both syndromes (5.9% of genes overlapping a DMR in Sotos, 93.1% of genes overlapping a DMR in Weaver).

To test the significance of the overlap we tested the odds ratio (OR) of overlap between each pair of syndromes. To simplify the analysis and make the measure of the odds ratio closer in form to the *DMRscaler* method, we only used counts of measured CpGs (See methods for details). Essentially, the odds ratio tests whether there is enrichment of CpGs that are in DMRs in one syndrome in the set of CpGs found in DMRs in the other syndrome being compared. OR with a confidence interval (CI) overlapping 1 suggests no enrichment, closer to 0 or further from 1 indicates greater enrichment. The raw overlap counts used to calculate the OR are in (**Table 3-9**) and the odds ratios are reported in (**Table 3-10**). The highest odds ratio was between Sotos and Weaver with OR = 17.16 (95% CI: 12.27-23.99, p = 1.9e-32 Fisher's exact test) in Layer 1. The OR between Sotos and Weaver drops to OR = 1.55 (95% CI: 1.47-1.64, p = 4.6e-56 Fisher's exact test) in Layer 5, which is likely due to the extensive genomic coverage by DMR features in Sotos syndrome at Layer 5. The next highest odds ratio was between Arboleda-Tham and Weaver in Layer 1 with OR = 9.94 (95% CI: 5.83-16.94, p = 4.7e-10 Fisher's exact test) which gains in significance but drops in magnitude at Layer 5 with OR = 3.00 (95% CI: 2.71-3.31, p = 3.1e-77 Fisher's exact test). The OR between Arboleda-Tham and Sotos is relativately small at Layer 1 with OR = 1.72 (95% CI: 1.43-2.06, p = 5.6e-8 Fisher's exact test) and drops to non-significance at Layer 5 (**Table 3-10**). These results show how DMRs of each of the three syndromes analyzed here are enriched in CpGs in DMRs in each of the other syndromes tested here at the lower layers output by *DMRscaler*, where differential CpG density is required to be higher, and in particular the strong enrichment between the two

overgrowth disorders offers evidence for common epigenetic effects in these disorder and potentially common contributing factors.

One region of overlap between Sotos and Weaver syndrome was a region overlapping *INS,* and *INS-IGF2* proximal to *IGF2*. This region stood out as a region implicated in another growth disorder, Beckwith-Wiedemann syndrome (BWS) [57]. The DMR called for Sotos syndrome is just upstream of *IGF2* and overlaps *INS* and *INS-IGF2*, with sites of moderate effect hypomethylation ($\Delta\beta > 0.2$) (**Figure 3-29A**). The DMR called for Weaver syndrome overlaps the IGF2 gene and is composed of sites with a small effect size ($\Delta\beta \sim 0.05$), with hypermethylation over a region of the *IGF2* gene body and hypomethylation further upstream overlapping the *INS* and *INS-IGF2* genes. Upstream of *IGF2* overlapping *INS* and *INS-IGF2* the pattern of hypomethylation in Sotos and Weaver syndrome was consistent (**Figure 3-29B**).

Across all three syndromes there were 49 genes overlapping some DMR called by *DMRscaler*. Among these were *PCDHGA1, PCDHGA2, PCDHGA3, PCDHGA8, PCDHGA10, PCDHGB7, PCDHGA11*, *PCDHGA12,* and *PCDHGC3* of the *PCDHG* cluster genes, previously discussed in context of Sotos syndrome alone. These genes are worth noting as they are involved in neural development. The *PCDHG* cluster is broadly hypermethylated in Sotos, as noted earlier (Figure 5C,D). In Arboleda-Tham syndrome, there is a small DMR intergenic to most of the *PCDHG* genes in this cluster and positioned at the 5' end of *PCDHGC3*. In Weaver syndrome there is a DMR with minor hypomethylation at the stretching from the same 5' end of *PCDHGC3* Arboleda-Tham that spans through to the shared 3' end of *PCDHG* genes. Few of the other methods evaluated are able to identify this region of the *PCDHG* gene cluster in either Arboleda-Tham syndrome or Weaver syndrome (**Figure 3-30**).

The overlapping DMRs and the genes with overlapping DMRs between Arboleda-Tham, Sotos, and Weaver syndrome reveal a number of shared regions of differential methylation and shared genes with patterns of differential methylation. CpGs in DMRs in any one syndrome are

enriched in CpGs in DMRs of either of the other syndromes across all three pairs of syndromes, Arboleda-Tham:Sotos, Arboleda-Tham:Weaver, Sotos:Weaver, as measured by the odds ratio. However, these data are derived from different cell types (fibroblast vs blood) and exhibit cell-type specific changes in addition to those caused by the genetic mutation. Together these results suggest that while each syndrome has a distinct profile of differential methylation, there is also significant overlap in regions mirroring shared phenotypic features.

Discussion

The key development of our new method, *DMRScaler*, is a substantial improvement over existing methods in the ability to accurately identify the size of DMRs across the full range of epigenetic scale.

Differential methylation analysis between sexes performed with *DMRscaler* showed our algorithm could handle the full range of DMR features present in simulated and real-world samples. Looking at the DMRs between XX and XY individuals, *DMRscaler* was able to identify a small DMR 2.1 kb in length overlapping the autosomal gene *TLE1* that had previously been identified as differentially methylated between the sexes[53], while also consolidating the differential methylation of the X-chromosome into a single DMR 152.13 Mb in length, spanning 98% of the total length of the chromosome.

Additionally, *DMRscaler* provides the means for a hierarchical definition of a DMR that is built through the iterative procedure of merging layers built from increasing window sizes. A deeper analysis of the DMR spanning the X-chromosome showed that gaps in DMRs at lower layers that were consolidated in the upper layers were significantly enriched in genes known to escape X-inactivation, such as *RPS4X* [51,52], which is concordant with data showing that regions escaping X-inactivation should have a similar epigenetic landscape between sexes [52]. These enrichment results show how *DMRscaler*, in addition to providing an intuitive representation of

DMRs, also provides a mechanism for a hierarchical definition of DMRs that can be used to investigate the structure of the methylation landscape across larger epigenomic features. Together these behaviors of intuitive scaling and defining a hierarchical map of DMR features allow *DMRscaler* to be used to achieve greater flexibility and more meaningful interpretation of results in analyses of differential methylation than existing methods.

Finally, given our primary interest in leveraging this method for the smaller sample sizes in rare-disease studies, we tested *DMRscaler* on datasets from patients with rare chromatin modifier syndromes. Specimens that harbor known pathogenic mutations in chromatin modifier genes often display regional changes to epigenetic features, such as DNA methylation state[9,10]. Our study also explored three syndromes that are caused by pathogenic mutations in genes that directly control histone modifications.

Arboleda-Tham Sydrome (MIM# 616268), also known as KAT6A syndrome, is a genetic syndrome caused by mutations in the Lysine (K) acetyltransferase *KAT6A* characterized by global developmental delay, intellectual disability, speech delay or absence and phenotypes of variable expressivity such as congenital heart defects and gastrointestinal anomalies[7,58]. KAT6A acetylates histones K3K9, H3K14, and H3K23[59–61], but the genomic regions affected in Arboleda-Tham syndrome have not been comprehensively studied. Previously, deletion of *KAT6A* in model organisms has identified the *HOX* genes, including the *HOXB* cluster, as regulatory targets of KAT6A[60,62]. Three DMRs identified here were identified by *DMRscaler* spanning multiple genes of the *HOXB* cluster (**Figure 3-23**), which encompass 2 genes (*HOXB3, HOXB4)* found in a *KAT6A* knockout mouse model to have shifted domains of expression resulting in homeotic transformation of the axial skeleton [60]. The ability to highlight the extent of differential methylation beyond a single gene provides further context into the epigenetic change that occurs in Arboleda-Tham syndrome.

One limitation of our study is that cases are generally younger in age than the control groups. Previous studies have identified global hypomethylation as associated with aging[63,64]. For regions that are largely hypermethylated in Arboleda-Tham Syndrome patients relative to controls, we cannot exclude this as a potential confounding factor in our analysis.

Weaver syndrome and Sotos syndrome are rare overgrowth syndromes that can be difficult to distinguish without sequencing. They are caused by mutations in *EZH2*[65,66] and *NSD1* gene[55], respectively. Despite their common clinical phenotype of overgrowth, the regions of the genome that are identified as differentially methylated largely diverge between these two syndromes suggest distinct pathways to a common and complex phenotype. For Weaver Syndrome, *DMRscaler* identified differential methylation over the *HOXA* cluster genes in Weaver syndrome relative to controls. Genome-wide mapping of EZH2 binding domains shows EZH2 binds the *HOXA* cluster[67] and EZH2 overexpression in mantle cell lymphoma has been associated with hypermethylation over the *HOXA* cluster[68]. The key improvement is that rather than highlighting individual genes[9] as differentially methylated, *DMRscaler* is able to demonstrate the modular nature of the genetic regulation by highlighting the non-random spatial relation of these features as a pair of DMRs spanning several genes each.

Additionally, *DMRscaler* identified a novel finding in the neighboring *PCDHA, PCDHB, and PCDHG* clusters as broadly hypermethylated between Sotos syndrome patients relative to controls. The protocadherin family genes are critical in cell-cell adhesion and involved in the complex patterning of neural circuitry[56]. These same genes in the *PCDHGA/B* cluster were also identified as hypermethylated in Down syndrome human cortex relative to control cortex tissue[69]. From these results we can hypothesize that misregulation of the *PCDH* clusters in brain development may contribute to the neurodevelopmental phenotype of Sotos syndrome.

Notably, we observed that between the two overgrowth syndromes, Sotos and Weaver syndrome, the *IGFR2* region including the *INS* and *INS-IGFR2* genes was similarly differentially

methylated. Loss of normal imprinting regulation of *IGFR2* has been implicated in another overgrowth syndrome, Beckwith-Wiedemann Syndrome (BWS)[57]. Whether this common difference in DNA methylation proximal to the *IGFR2* locus represents an epigenetic contributor to the overgrowth phenotype or is a consequence of the overgrowth phenotype is worth further investigation.

The majority of real-world methylation data is in the form of reduced representation platforms that query CpGs in sites that are likely to play a role in gene regulation, such as known enhancers and transcriptional start sites. While the distance between the sites are variable on an array, our sex chromosome results demonstrate the ability of our method to call established DMRs that vary dramatically in size on this reduced representation platform. Whole genome bisulfite sequencing (WGBS) offers an alternative to array based technologies for querying DNA methylation that offers more complete coverage of the genome. While WGBS is technically and analytically challenging and remains prohibitively expensive for routine use, DMRscaler is platform agnostic and time efficient on array based data, and so should be readily portable to analysis of WGBS data. Due to the wider availability of array based DNA methylation datasets, particularly for rare disease cohorts, we decided to test DMRscaler on array data and have left validation on WGBS data as a future direction.

Conclusions

Here we have shown that *DMRscaler* is flexible yet robust in describing the scale of DMR features from the local scale of individual promoters and CpG sites, to the DMR features that represent chromosome level differences in methylation. All of the analyses described were run using a shared parameter set for *DMRscaler*, which highlights the utility to researchers who seek to explore these higher order epigenetic features while also describing the local changes with known biological implication, such as changes in methylation overlapping the promoter of a

gene. Importantly, *DMRscaler* serves as a proof of principle. The idea that important epigenetic features exist beyond the scale of a single gene is not new, however, existing methods for DNA methylation analysis do not capture this knowledge. Here *DMRscaler* proves that it is possible to computationally capture this intuition, and in doing so reveal novel biological insights.

Figures and Tables

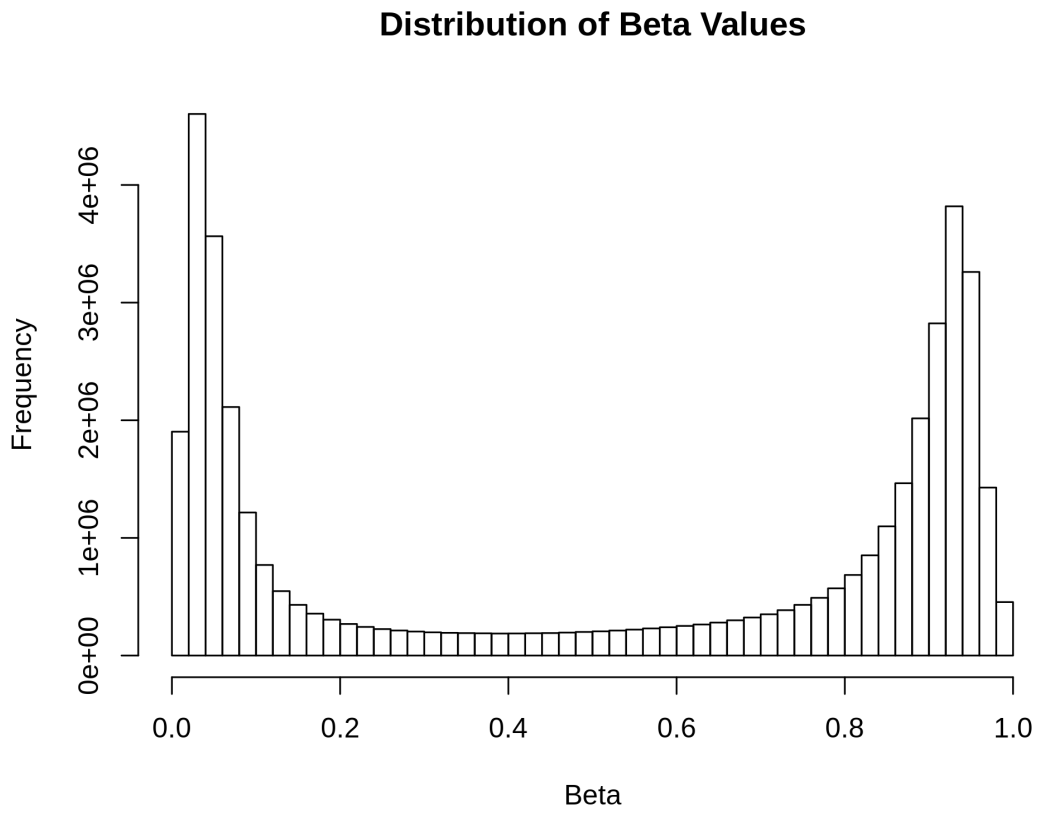


Figure 3-1. Beta Distribution of methylation data from Illumina Infinium Human Methylation450 Bead Chip 450 array. Beta value is the proportion of methylation at each CG site.

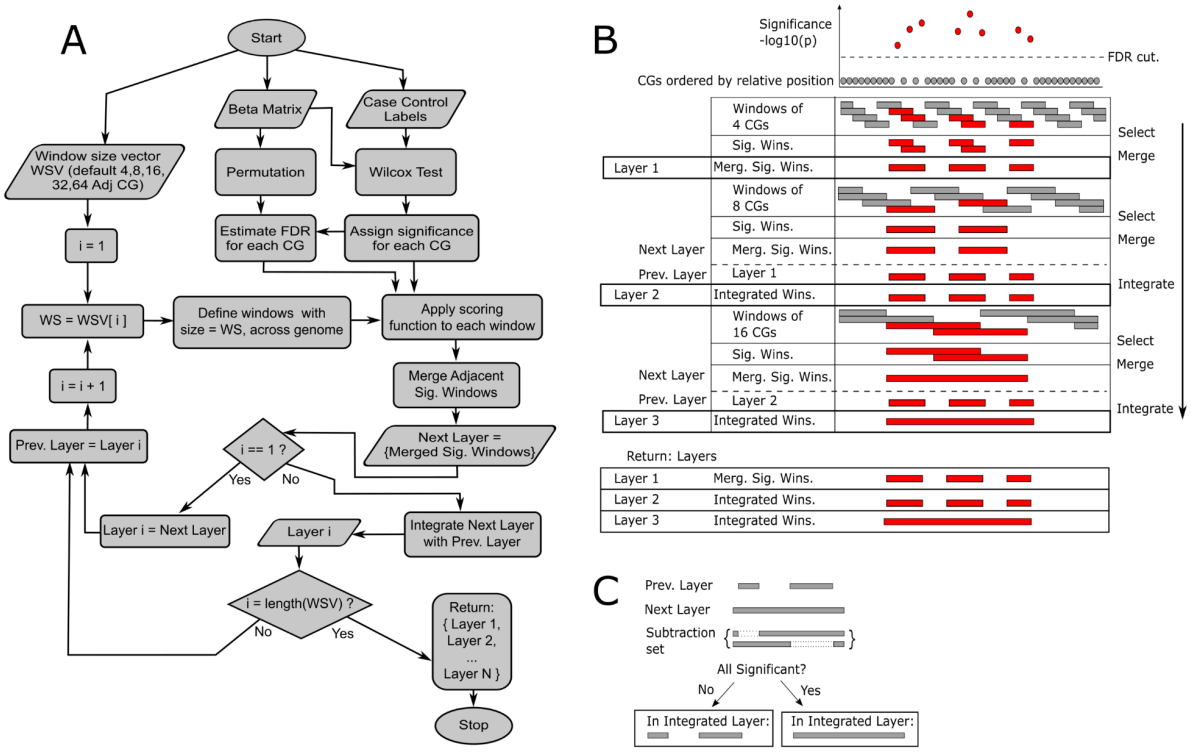


Figure 3-2: Outline of DMRscaler method. (A) Flowchart of decision tree for DMRscaler. Starts with Beta matrix, with individual CpG as rows and samples as columns and Beta values corresponding to methylation proportion, case-control labels, and a vector of window sizes in increasing order. Wilcoxon test and permutations are used to assign significance to CpGs as well as estimate false discovery rate FDR. The window scoring function is used with permutation to rank and assign significance to windows. Adjacent significant windows are merged forming the Next_Layer. For the first iteration, the returned Layer_1 is set to this Next_Layer, for subsequent iterations the returned Layer_i is set to the result of integration of Next_Layer with Prev_Layer. Integration of layers is described in 1C. Prev_Layer is updated to Layer_i before proceeding to iteration i+1. After the largest window size layer is generated, a list is returned of the results from each iteration of the algorithm. (B) Graphical description of algorithm. At the top

shows representation of CpGs ordered by position and associated with a significance value. Windows are laid over the ordered CpGs and selected if the window score is significant. Adjacent windows are then merged. If a Prev_layer has been assigned, then integration occurs. (C) Integration procedure. For each Next_Layer DMR, all overlapping Prev_Layer DMRs are identified. A subtraction set is generated by individually subtracting each overlapping Prev_Layer DMR from the Next_Layer DMR. Subtraction involves removing overlapping CpGs from the Next_Layer. If all elements of the subtraction set are significant when rescored with the window scoring function, then the Prev_Layer and Next_Layer regions are merged in the Integrated_Layer, otherwise the Prev_Layer DMRs are used in the Integrated_Layer. This procedure ensures that the broader Next_Layer DMRs are only included if no single Prev_Layer DMR was responsible for the significance of the region identified in the Next_Layer.

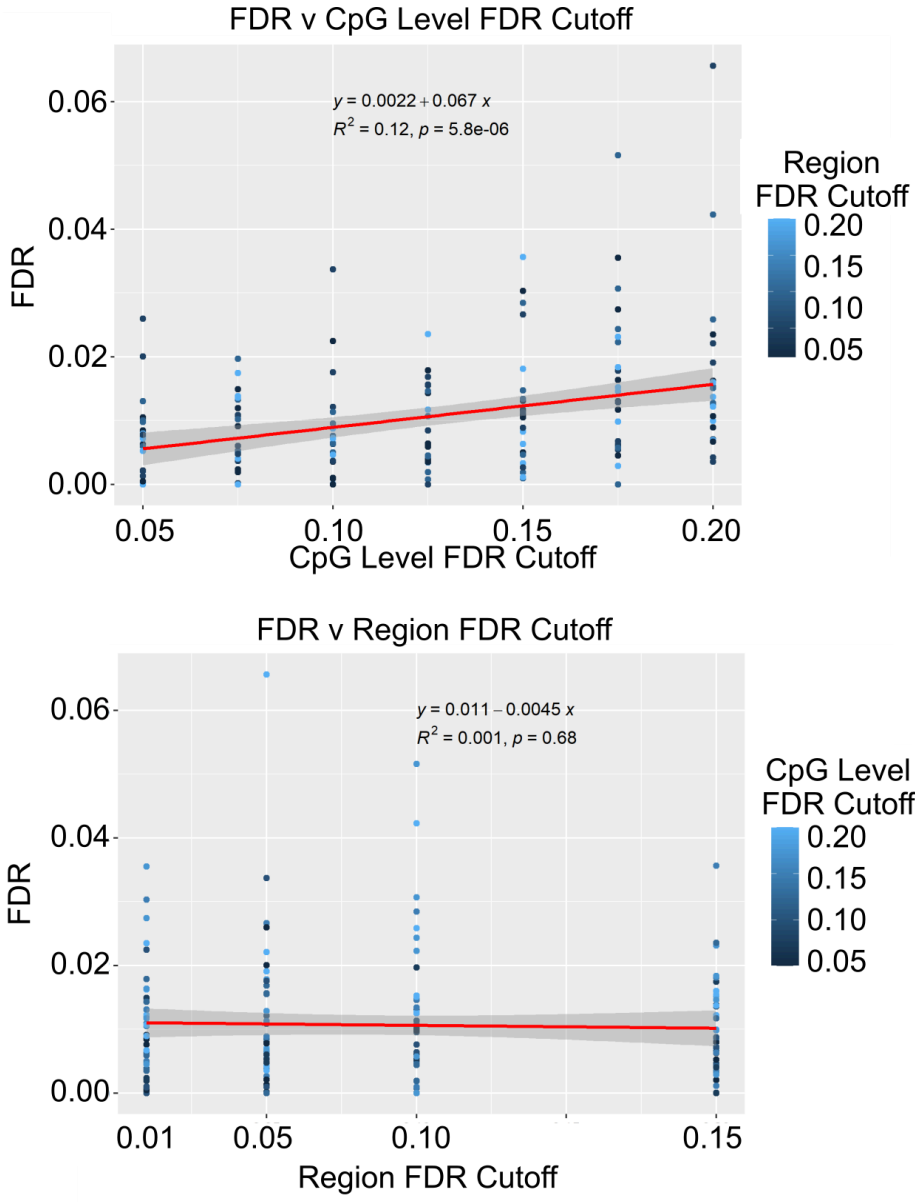


Figure 3-3: Test of FDR control from the CpG level FDR cutoff which is the p-value at which the set FDR is achieved based on permutation, and the region FDR cutoff, which is the FDR level set and determined by the Benjamini-Yekutieli procedure.

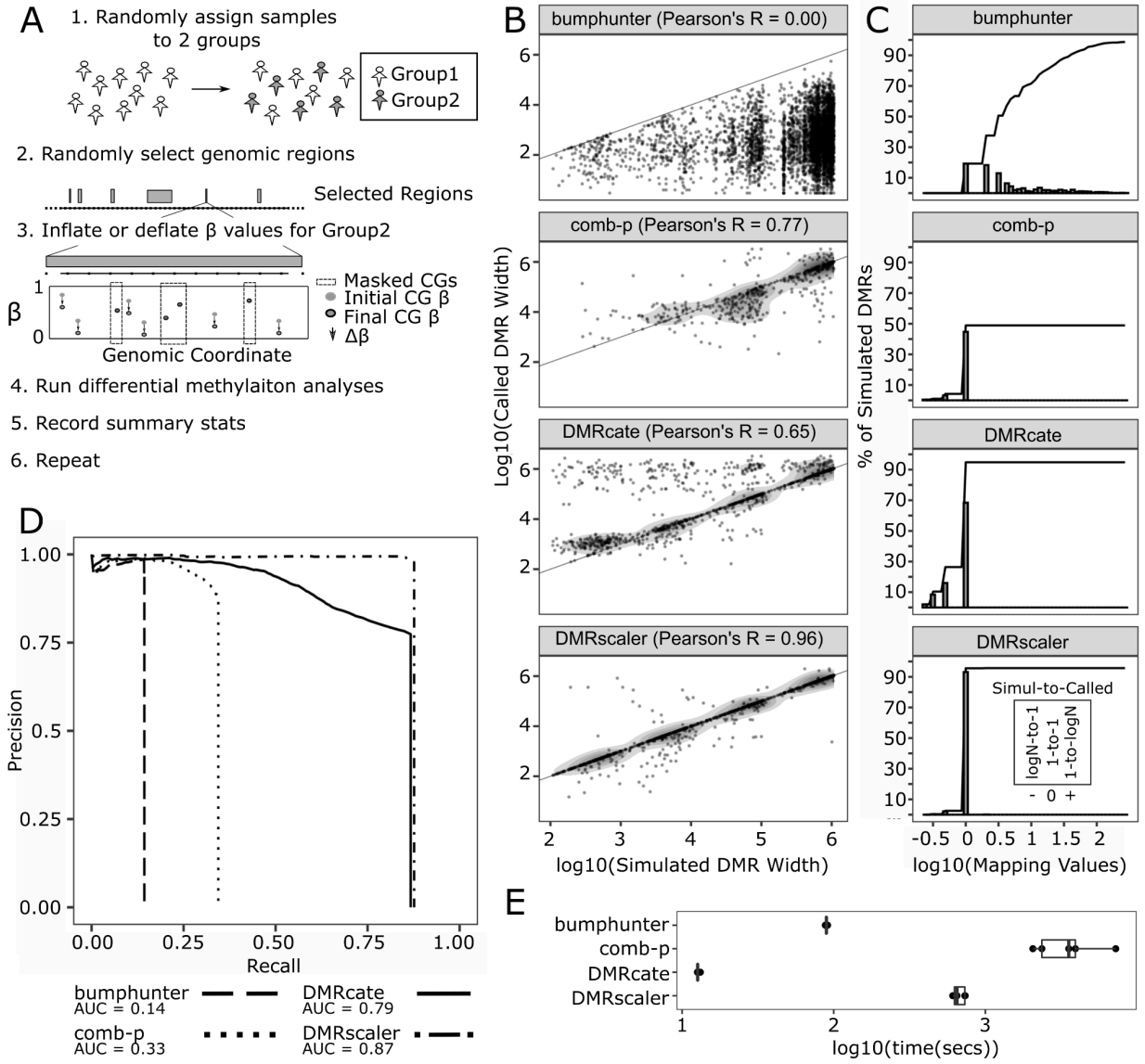


Figure 3-4: Simulation of DMRs ranging in size between 1kb to 1Mb for comparison of methods. (A) Graphical description of simulation design. First, samples are randomly assigned to one of two groups. Second, non-overlapping regions of the genome are randomly selected to be DMRs. Third, over selected DMRs one group has the β value of non-masked CpGs inflated or deflated by Δβ. Next all differential methylation methods are run and relevant summary statistics are recorded. This procedure is repeated a number of times to generate additional data points. (B) Simulated DMR Widths v Called DMR Widths plotted on log10 scale. Pairs are

formed between simulated and called DMRs if there is any overlap between the two. (C) Mapping Values plots. The mapping value is calculated for each simulated DMR and is either the inverse of the number of simulated DMRs sharing an overlapping called DMR or else it is the number of called DMRs overlapping the given simulated DMR, whichever is more extreme. Log values > 0 imply multiple DMRs called per simulated DMR. Value < 0 imply multiple simulated DMRs overlap single called DMR. Value = 0 implies one DMR called per DMR simulated. The plotted line indicates the cumulative proportion of simulated DMRs up to the given mapping value. (D) Feature level Precision-Recall Curves for each method, see methods for details on calculation. (E) Time for each method to run on the simulated dataset across 5 runs.

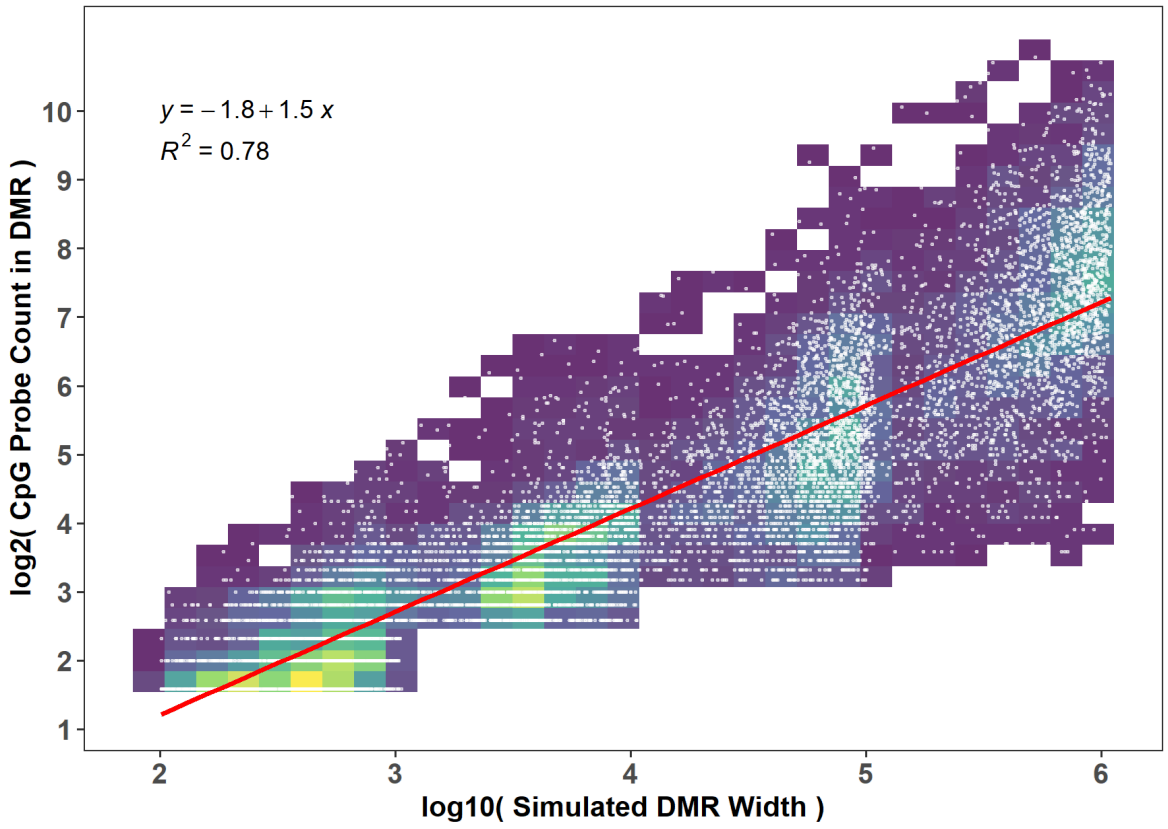


Figure 3-5: Simulated DMR width vs number of CpG probes. Plot shows distribution of CpG probes per simulated DMR against the simulated DMR width across all simulations run. Regression line and correlation included in upper, left corner of plot.

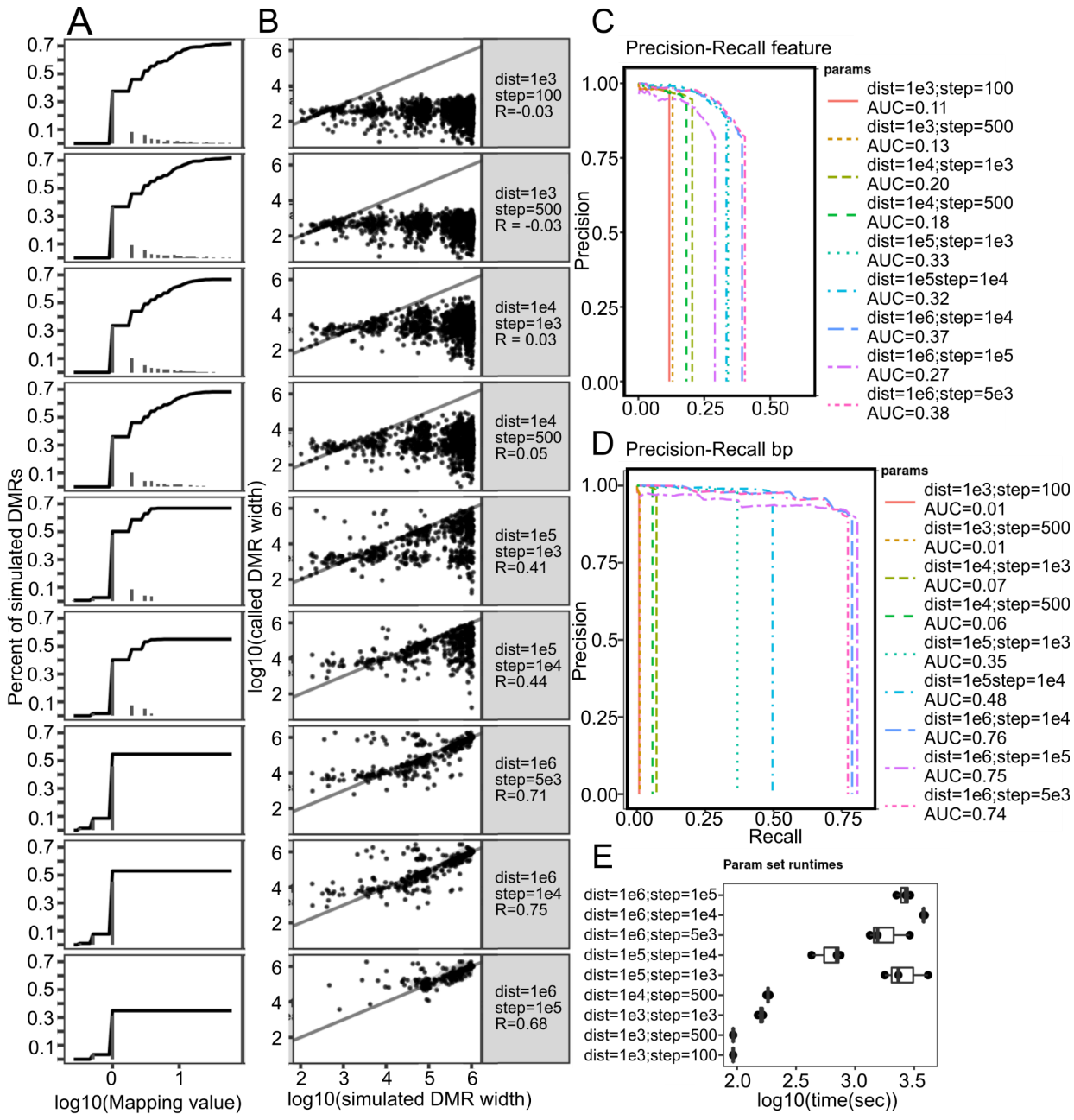


Figure 3-6: comb-p parameter testing: **(A)** Mapping Values plots. Log values > 0 imply multiple DMRs called per simulated DMR. Value < 0 imply multiple simulated DMRs overlap single called DMR. Value = 0 implies one DMR called per DMR simulated. The plotted line indicates the cumulative proportion of simulated DMRs up to the given mapping value. **(B)** Simulated DMR Widths v Called DMR Widths plotted on log10 scale.**(C)** Feature level

precision-recall curves, see methods for details on calculation. **(D)** basepair level precision recall curves **(E)** time for each parameter set run on the simulated dataset across 3 runs.

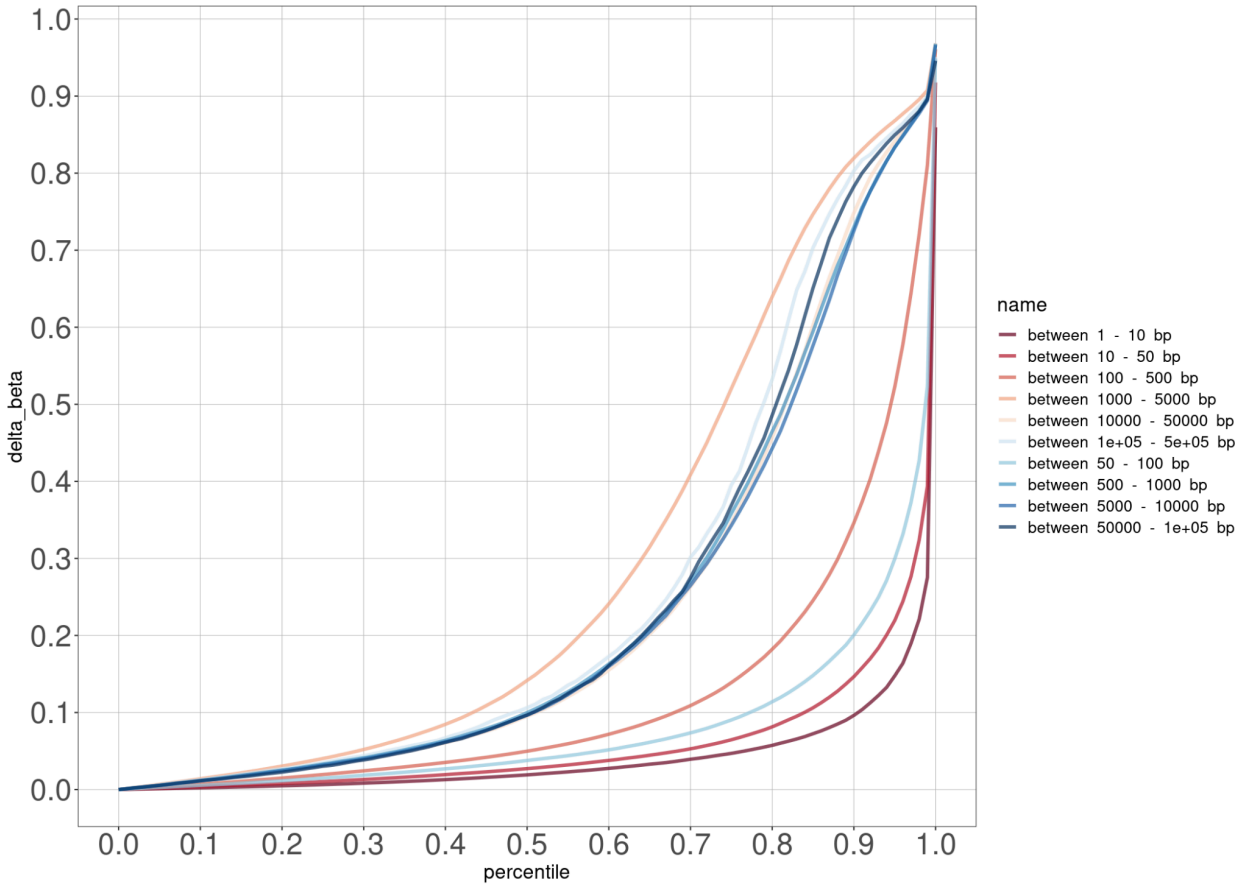


Figure 3-7: Cumulative Probability of difference in beta values between neighboring CG methylation. Y axis is the difference in Beta between neighboring CGs (CG_i and CG_i+1). E.g. 70% of adjacent measured CGs that are between 1000-5000 bp apart have a difference in beta value less than 0.40, said another way, 30% have a difference in beta greater than 0.4.

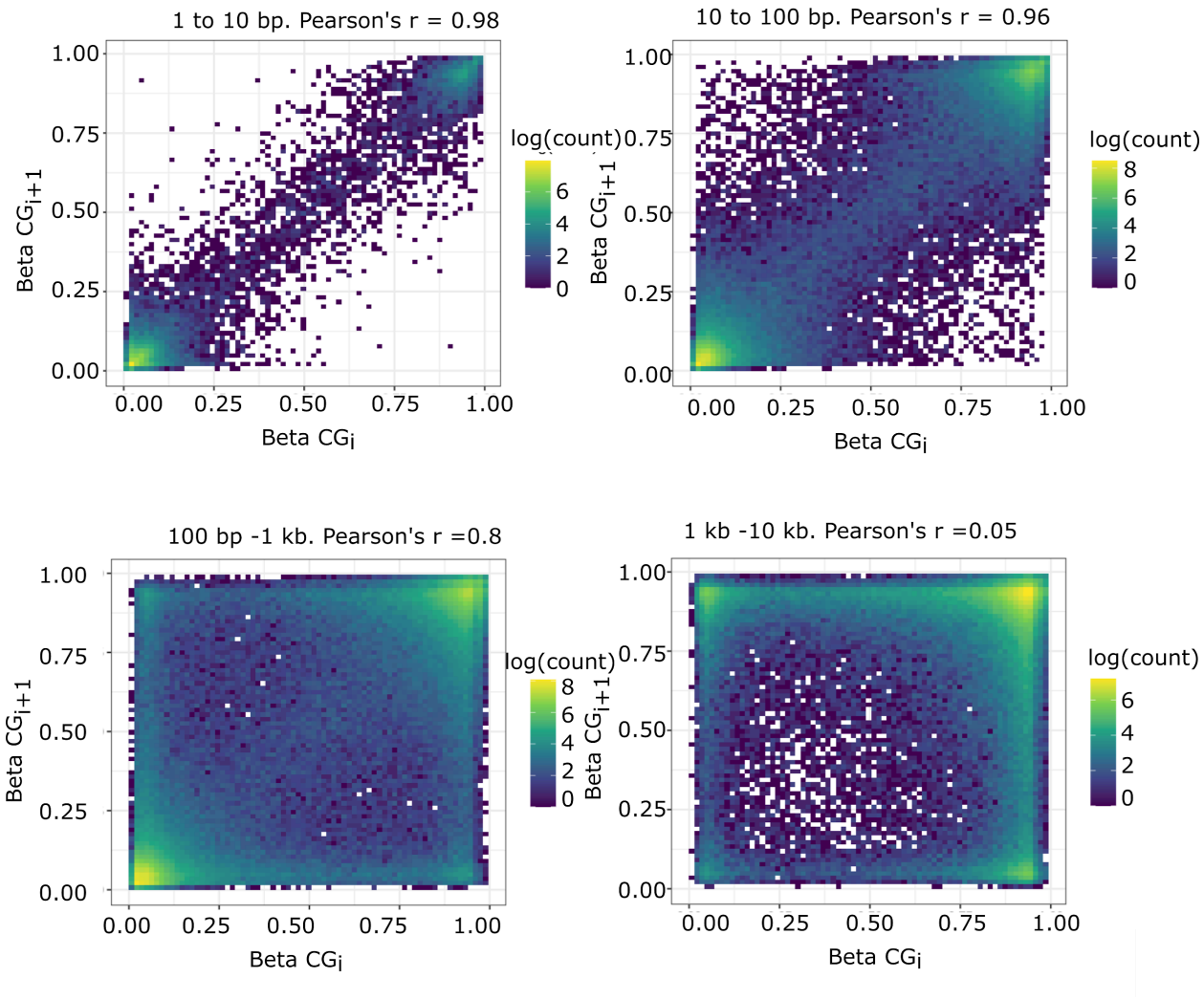


Figure 3-8: Neighboring CG methylation correlation. Plotted is beta value from 0-1, for CG_i on the x-axis and CG_i+1 on the y-axis. Neighboring CGs 1-10 bp apart and 10-100 bp apart have strong, but not perfect correlation (Pearson's r = 0.98, 0.96), those 100-1000 bp apart have modest correlation (Pearson's r = 0.8). At 1-10kb distance the correlation becomes weak (Pearson's r = 0.05).

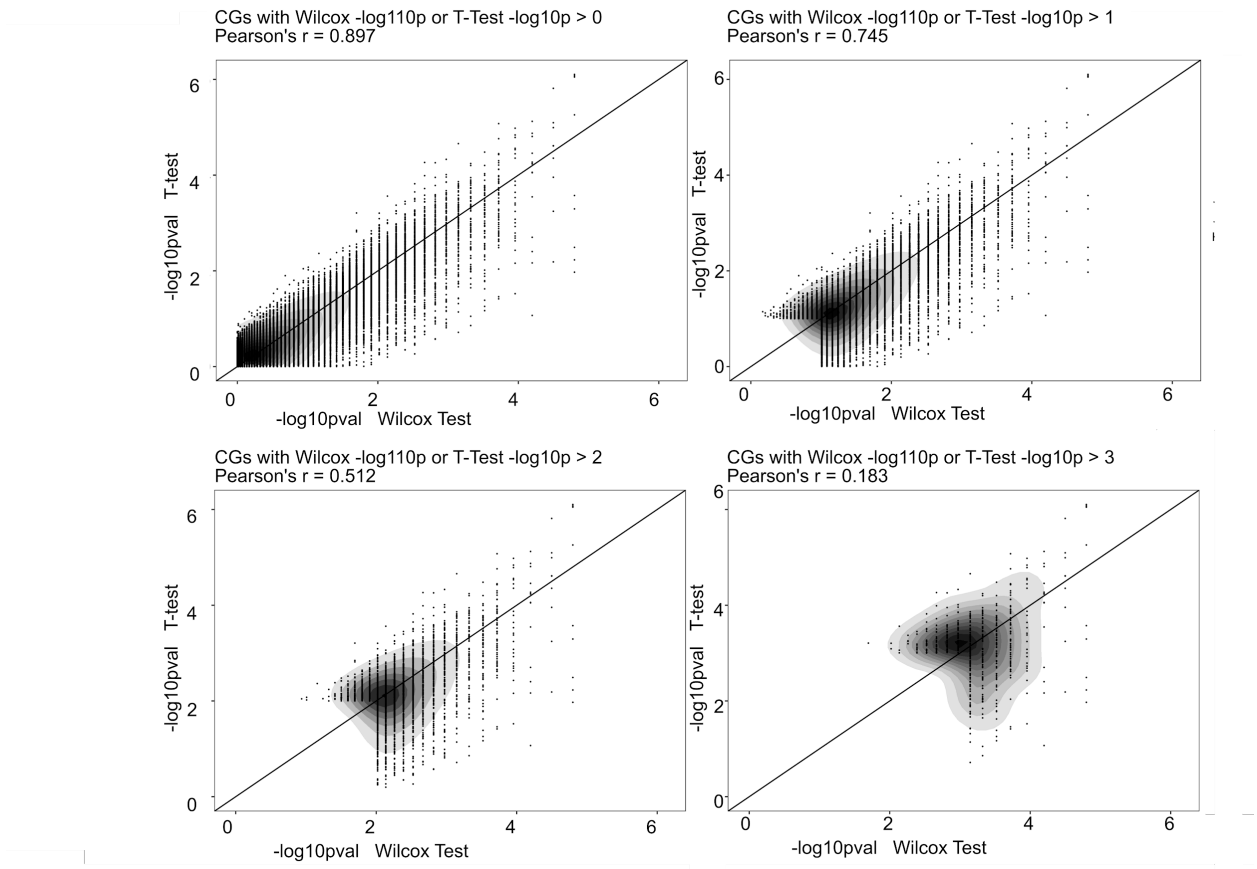


Figure 3-9: Wilcox - T-Test pearson's correlation (pearson's r) for each CG. From Arboleda-Tham Syndrome data, case-control labels determine group.

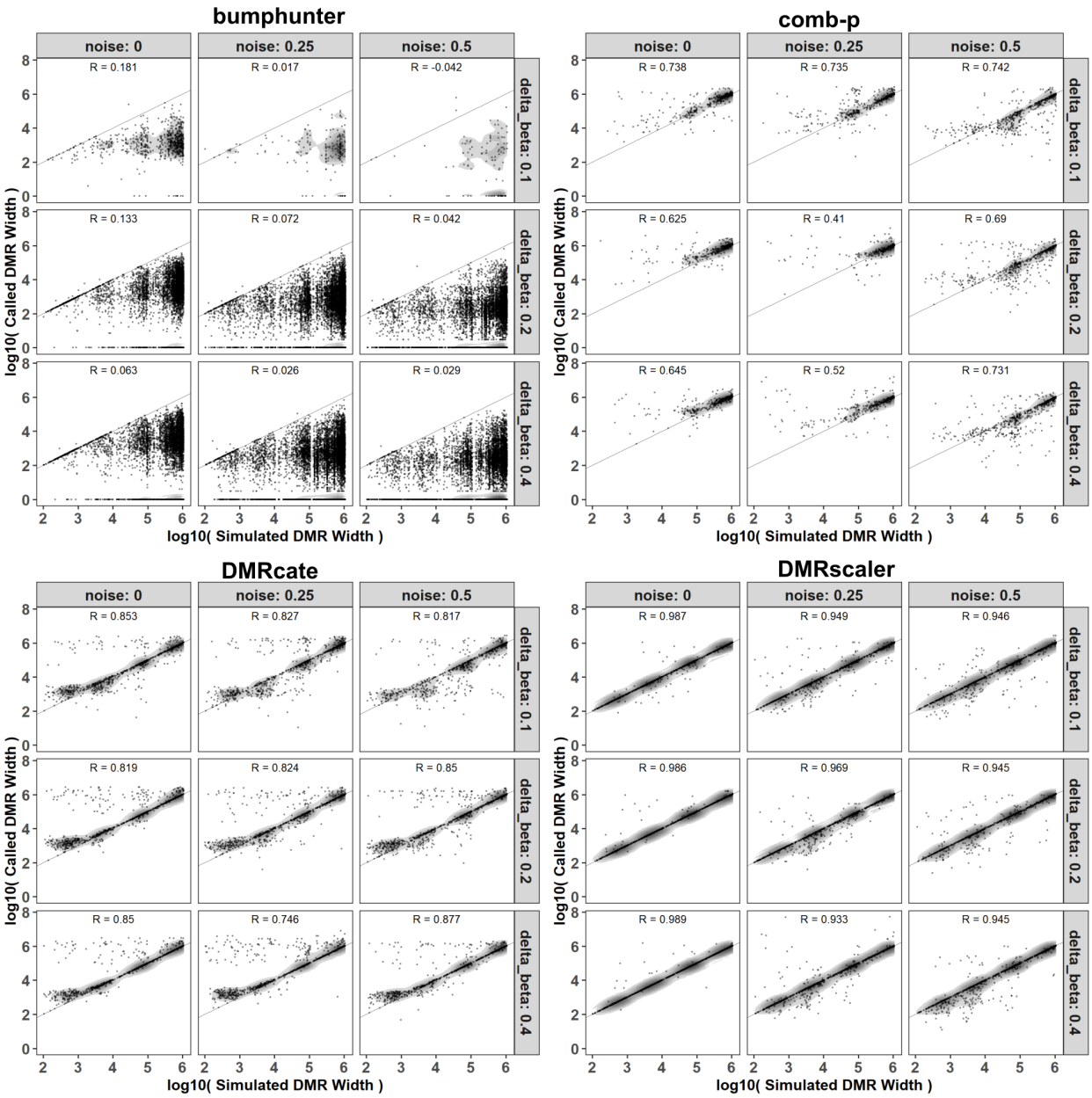


Figure 3-10: Simulated vs called DMR width for varied noise and delta_beta parameters. Simulated DMR Widths v Called DMR Widths plotted on log10 scale. Pairs are formed between simulated and called DMRs if there is any overlap between the two.

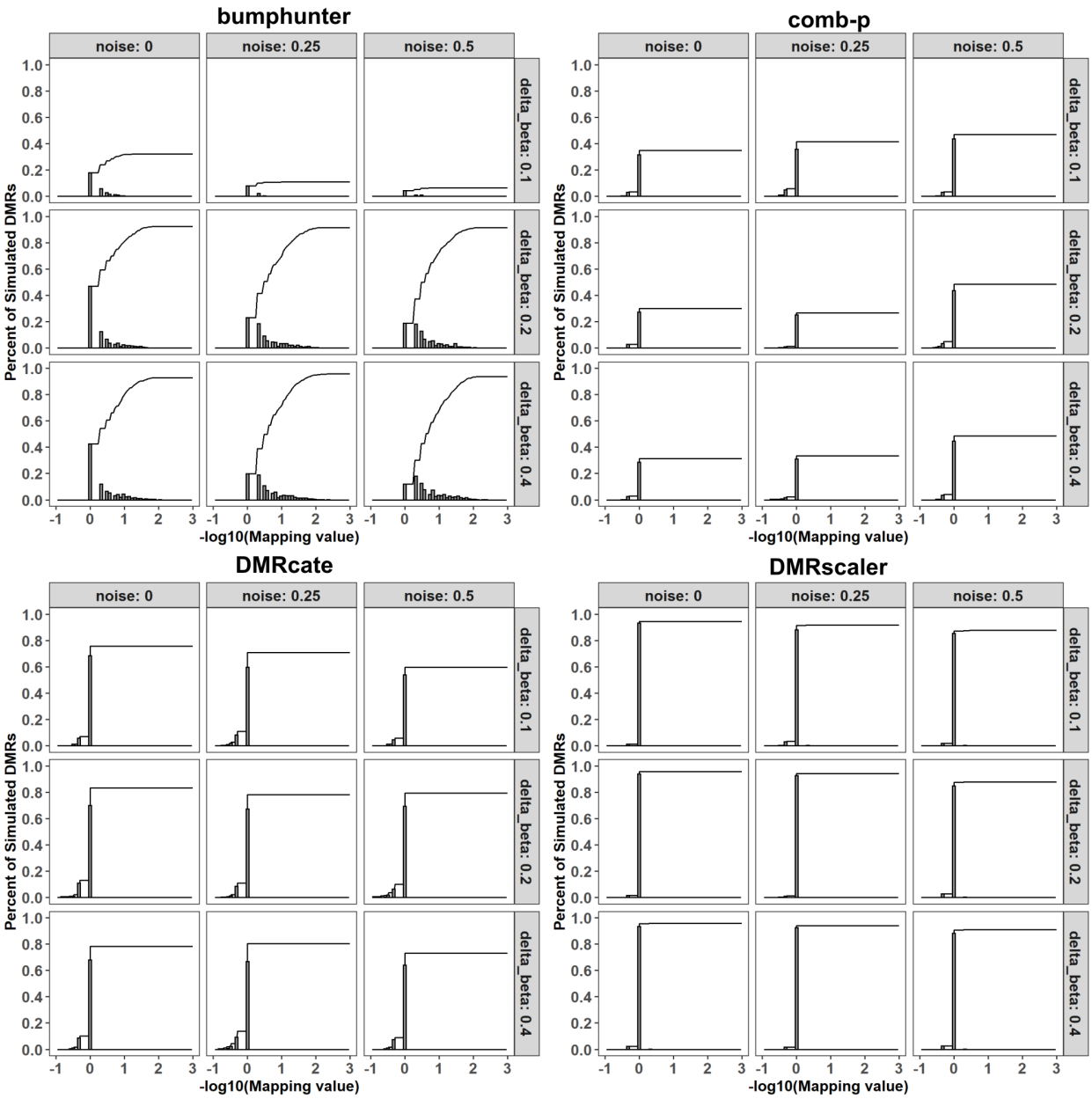


Figure 3-11: Mapping Values plots for varied noise and delta_beta parameters. The mapping value is calculated for each simulated DMR and is either the inverse of the number of simulated DMRs sharing an overlapping called DMR or else it is the number of called DMRs overlapping the given simulated DMR, whichever is more extreme. Log values > 0 imply multiple DMRs called per simulated DMR. Value < 0 imply multiple simulated DMRs overlap single called

DMR. Value = 0 implies one DMR called per DMR simulated. The plotted line indicates the cumulative proportion of simulated DMRs up to the given mapping value.

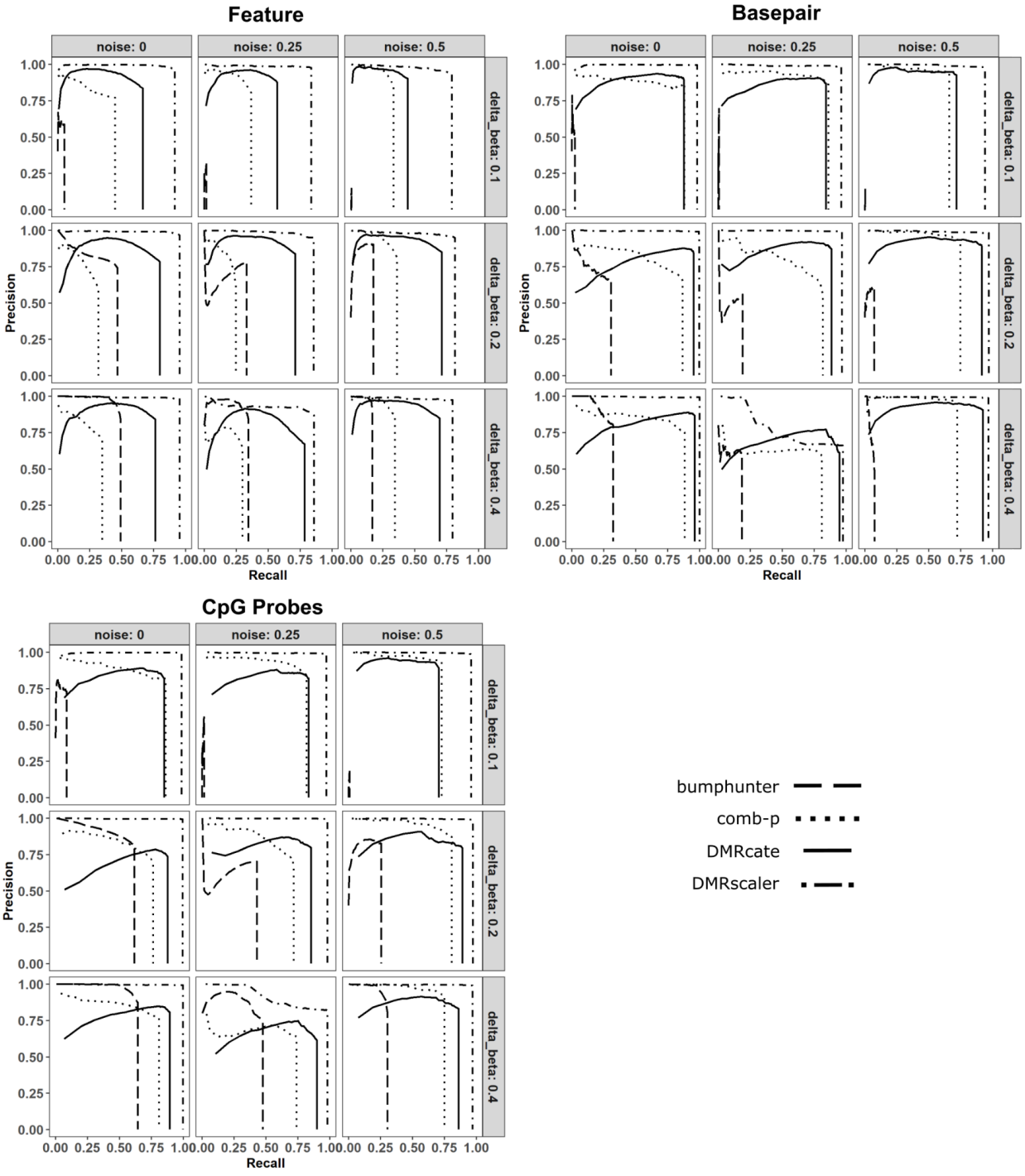


Figure 3-12: Precision and recall curves with varied simulation parameters for delta_beta and noise for each method, see methods for details on calculation.

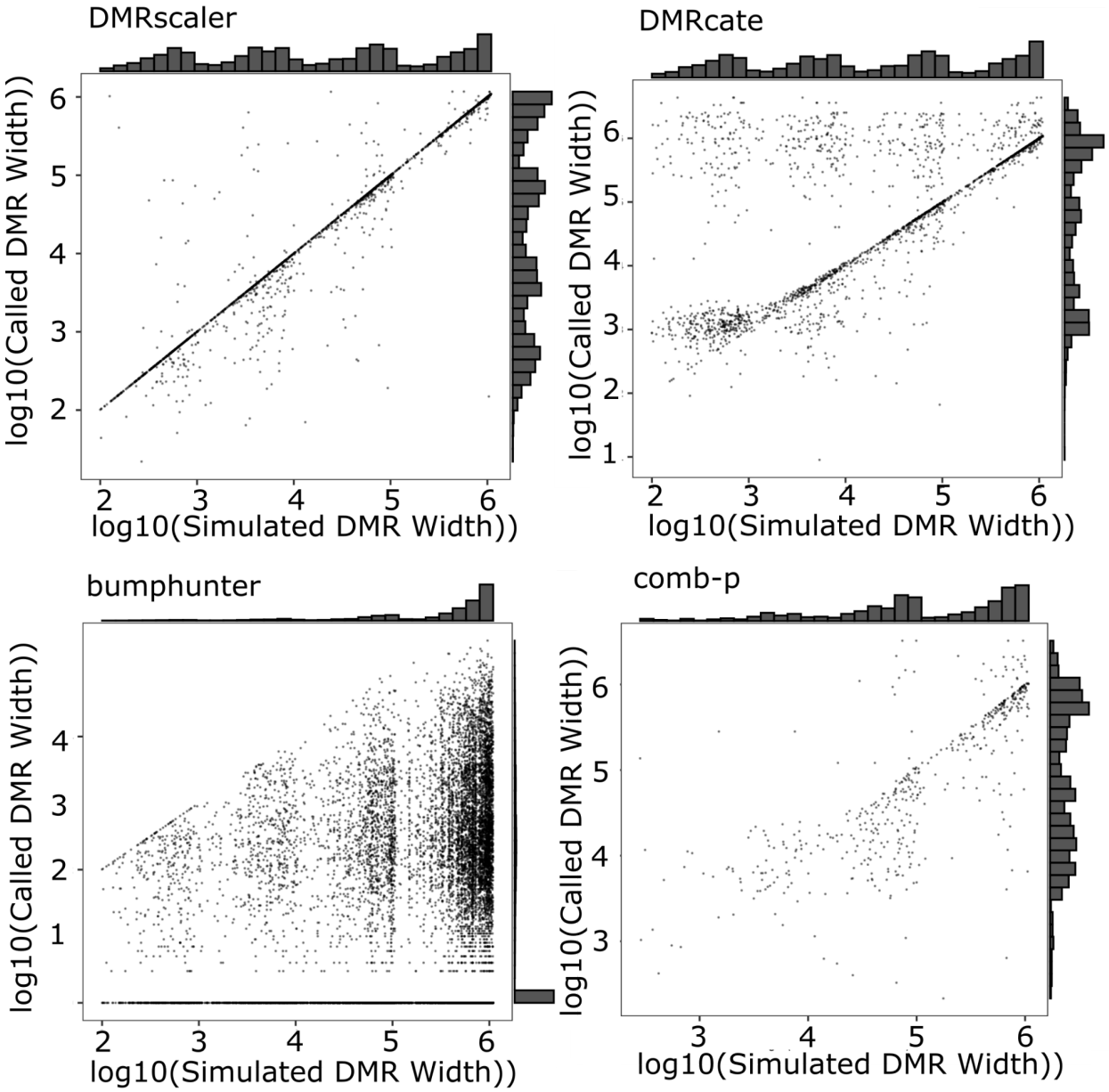


Figure 3-13. Simulated vs Called Widths with marginal density plots for noise = 50% and delta_beta = 0.2

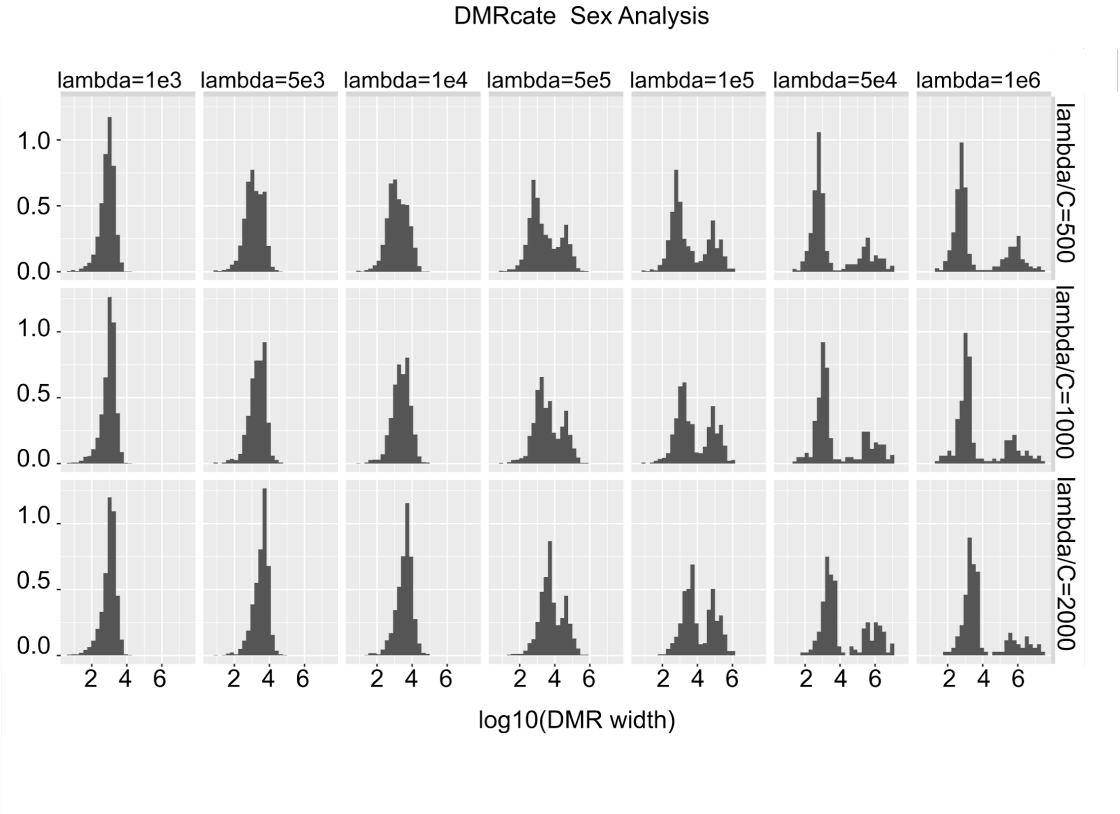


Figure 3-14: Testing variable lambda and C parameters on output from sex analysis for DMRcate. Only showing autosomal DMRs.DMRs called using a subset consisting of 8 XY and 8 XX controls.

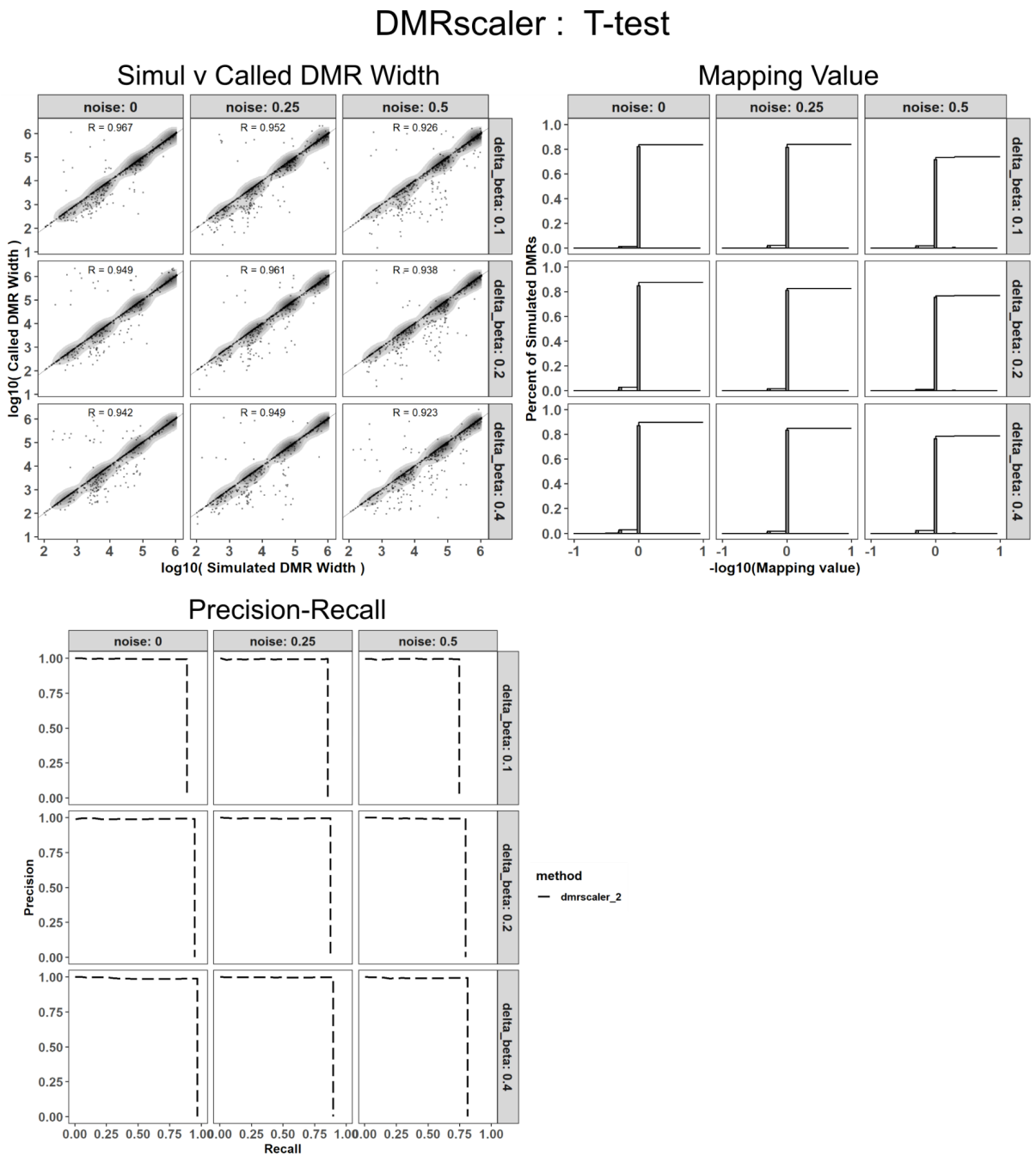


Figure 3-15: Simulation result for *DMRscaler* run on T-test derived p-values. Top left: Simulated vs called DMR width for varied noise and delta_beta parameters. Pairs are formed between simulated and called DMRs if there is any overlap between the two. Top right: The mapping value for each simulated DMR is either the inverse of the number of simulated DMRs

sharing an overlapping called DMR or the number of called DMRs overlapping the given simulated DMR, whichever is more extreme. Bottom: Precision and recall curves with varied simulation parameters for delta_beta and noise for each method, see methods for details on calculation.

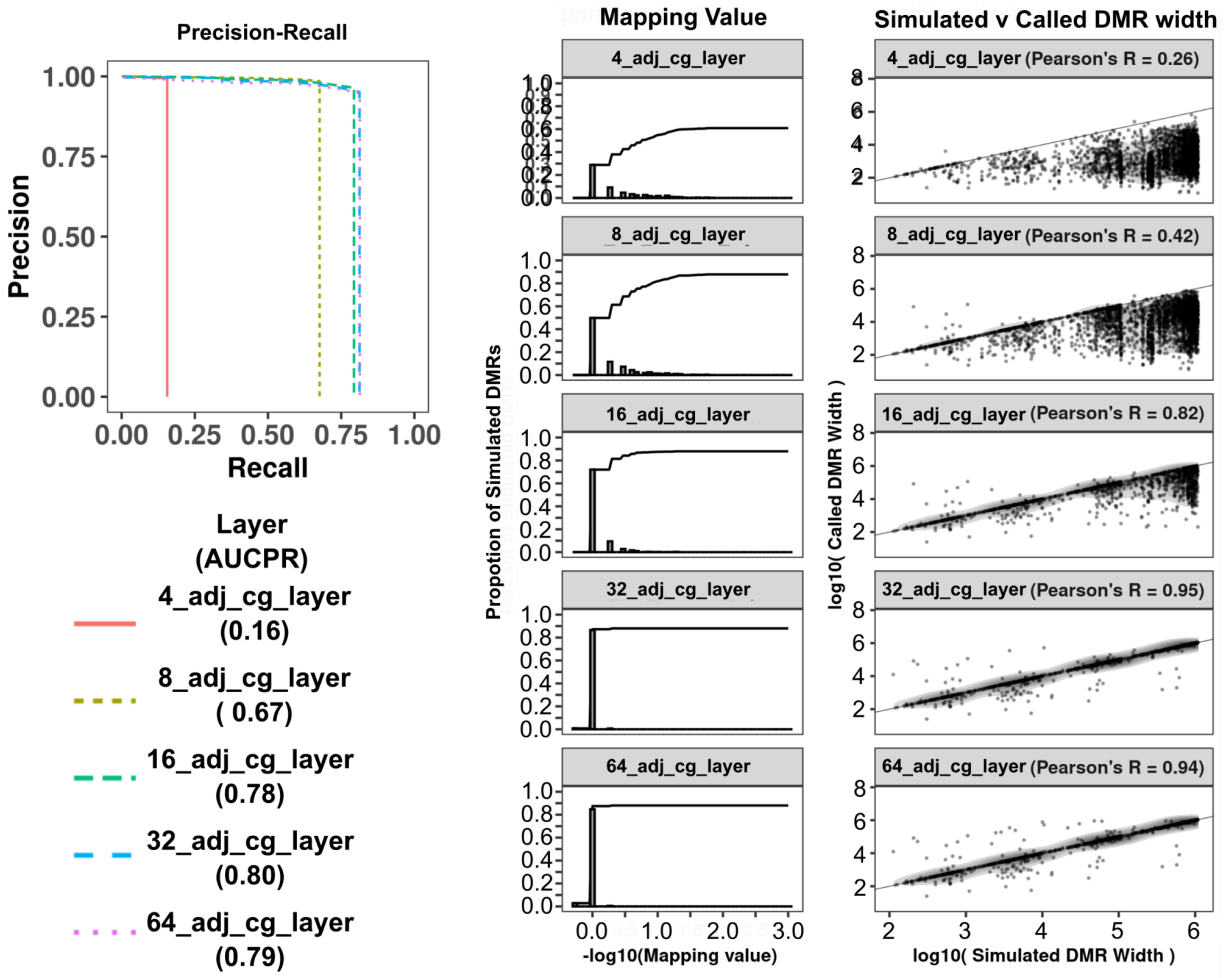


Figure 3-16: Performance measured at each layer of DMRscaler algorithm Top left: Precision and recall curves for each layer, see methods for details on calculation. Middle: The mapping value for each simulated DMR is either the inverse of the number of simulated DMRs sharing an overlapping called DMR or the number of called DMRs overlapping the given simulated DMR, whichever is more extreme. Right: Simulated vs called DMR width. Pairs are formed between simulated and called DMRs if there is any overlap between the two.

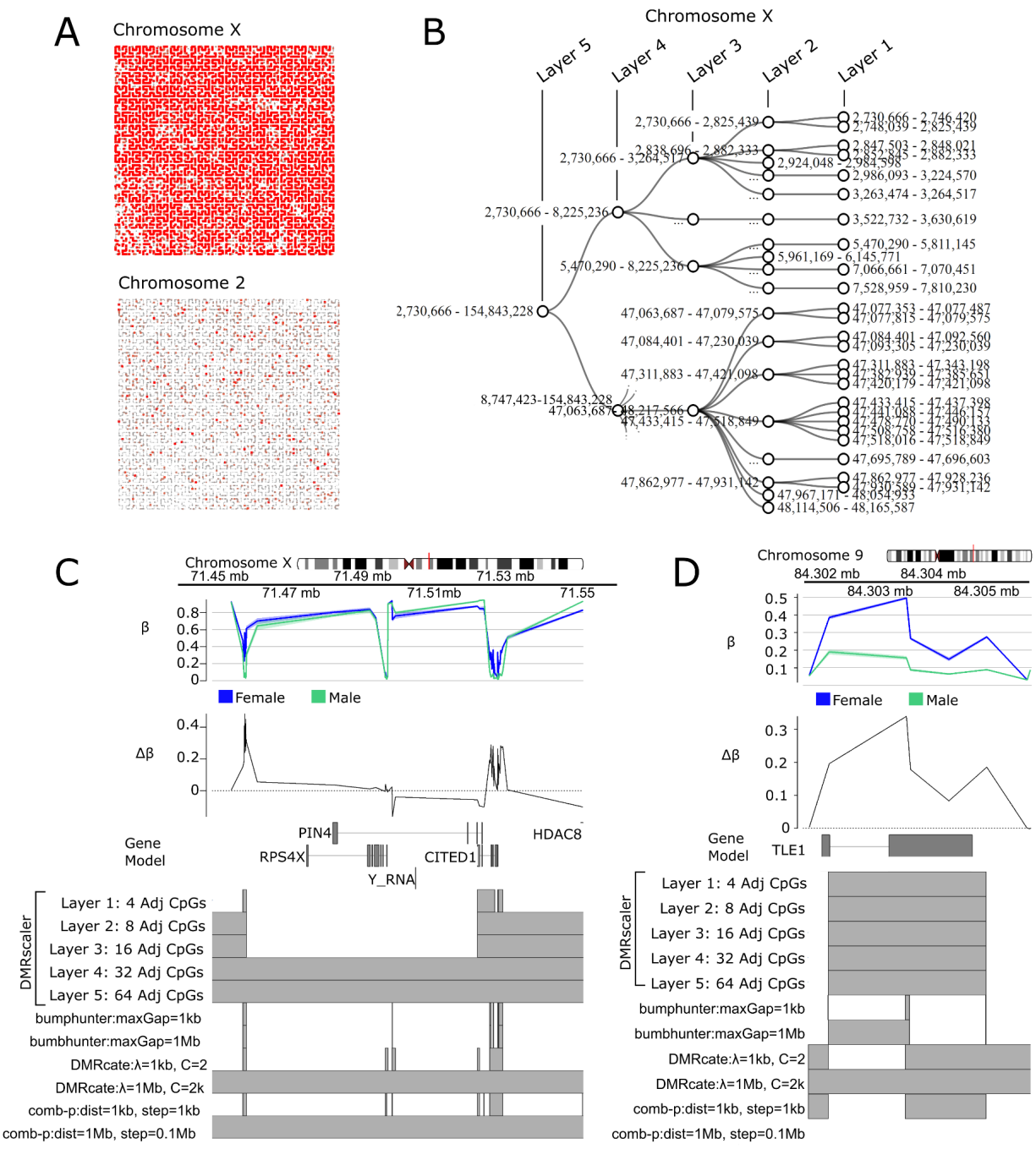


Figure 3-17: Differential Methylation Analysis Between XX and XY individuals. (A) Hilbert Curves of chrX and chr1. Hilbert curve is constructed by ordering CpGs by their position along the given chromosome. Red points are differentially methylated CpGs with FDR < 0.1. Point size scaled to max significance level (-log10 p-value). (B) Diagonal network plots showing

hierarchical relation of DMRs called by DMRscaler in Layers 4, 3, 2, and 1 (equivalent to 32 16, 8, and 4 Adj. CpG Layers respectively) for X-chromosome. (C) ChrX:71.4-71.6Mb. GVIZ track stack plot. Top track shows mean β value per group, next track shows Δβ, where Δβ = βfemale - βmale . Below the gene model track is the DMR track, highlighting the regions called as a DMR at each result layer from DMRscaler (Layers 1, 2, 3, 4, 5 are equivalent to 4, 8, 16, 32, 64 Adj. CpG layers) and from each competing method. (D) Chr9:84.302-84.306Mb. Tracks same as 3C.

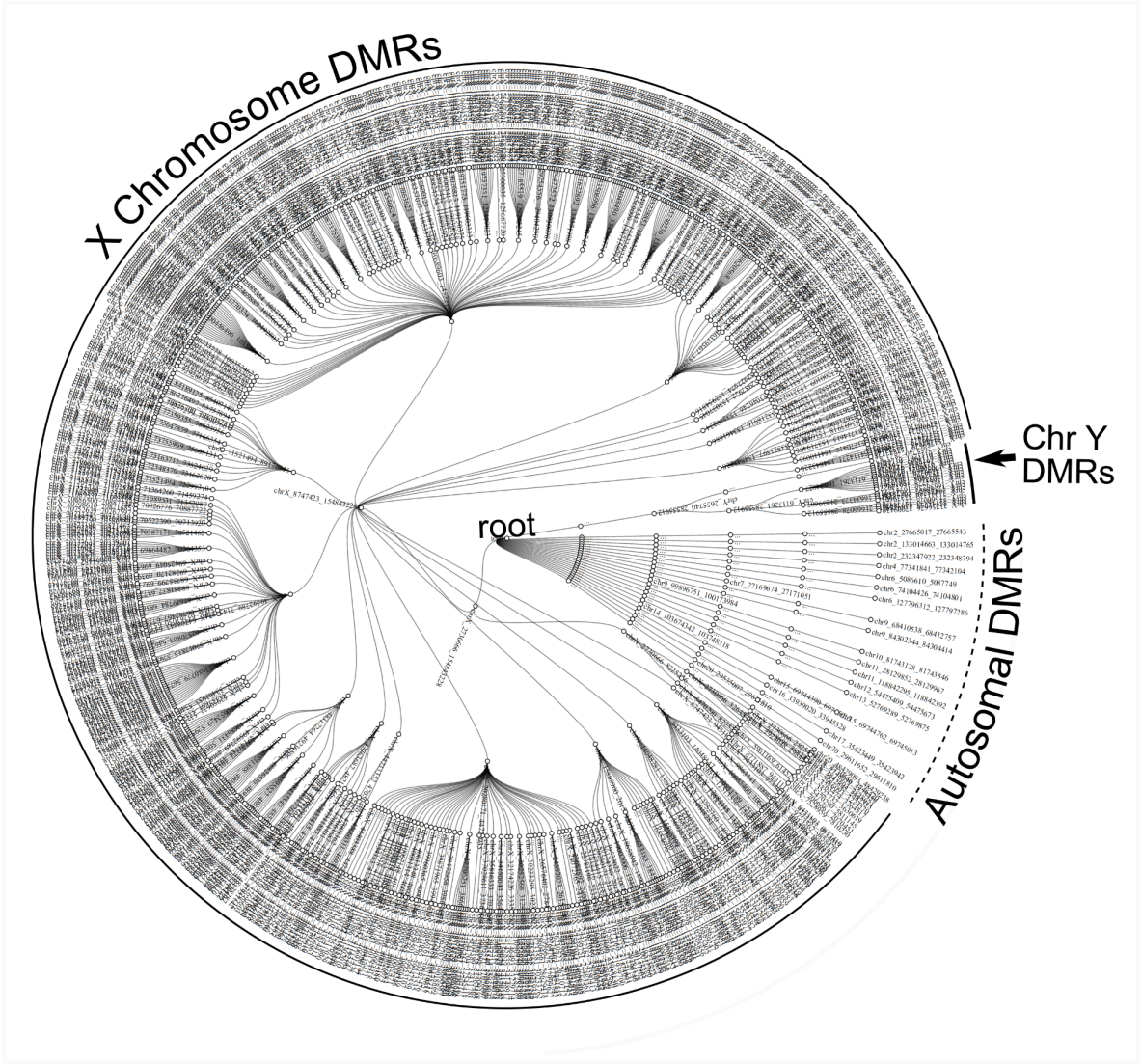


Figure 3-18: Radial Network Showing hierarchical structure of DMRs called across all layers of the DMRscaler algorithm from layer 1 (4 adjacent CG windows) at the edges nodes to layer 5 (64 CG windows) in the inner ring. Note, all are connected to the virtual root node which is only used for plotting purposes here. Each node is an individual DMR called. Coordinates for each DMR are printed at the earliest layer where that DMR appears. Unlabelled nodes are those that did not change from the previous layer.

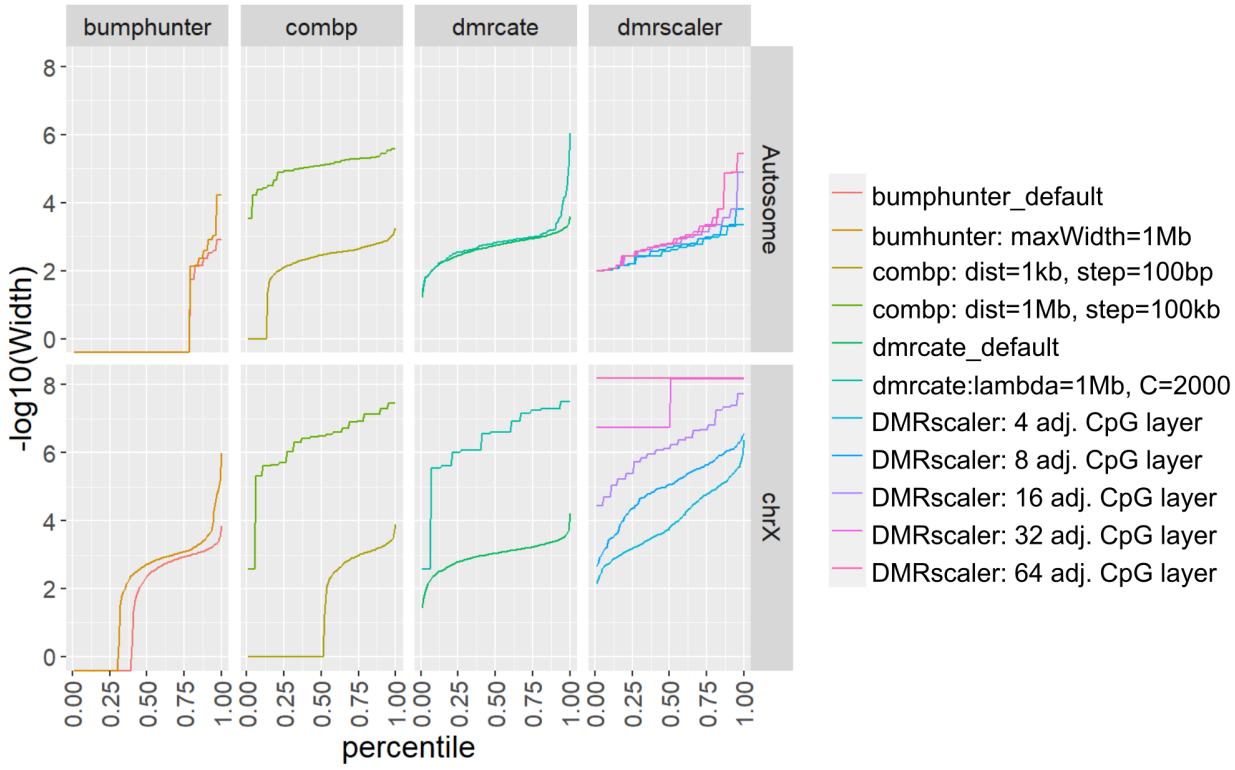


Figure 3-19: Sex analysis DMR width percentile plot. DMRs Called by each method for sex analysis ordered by dmr width.

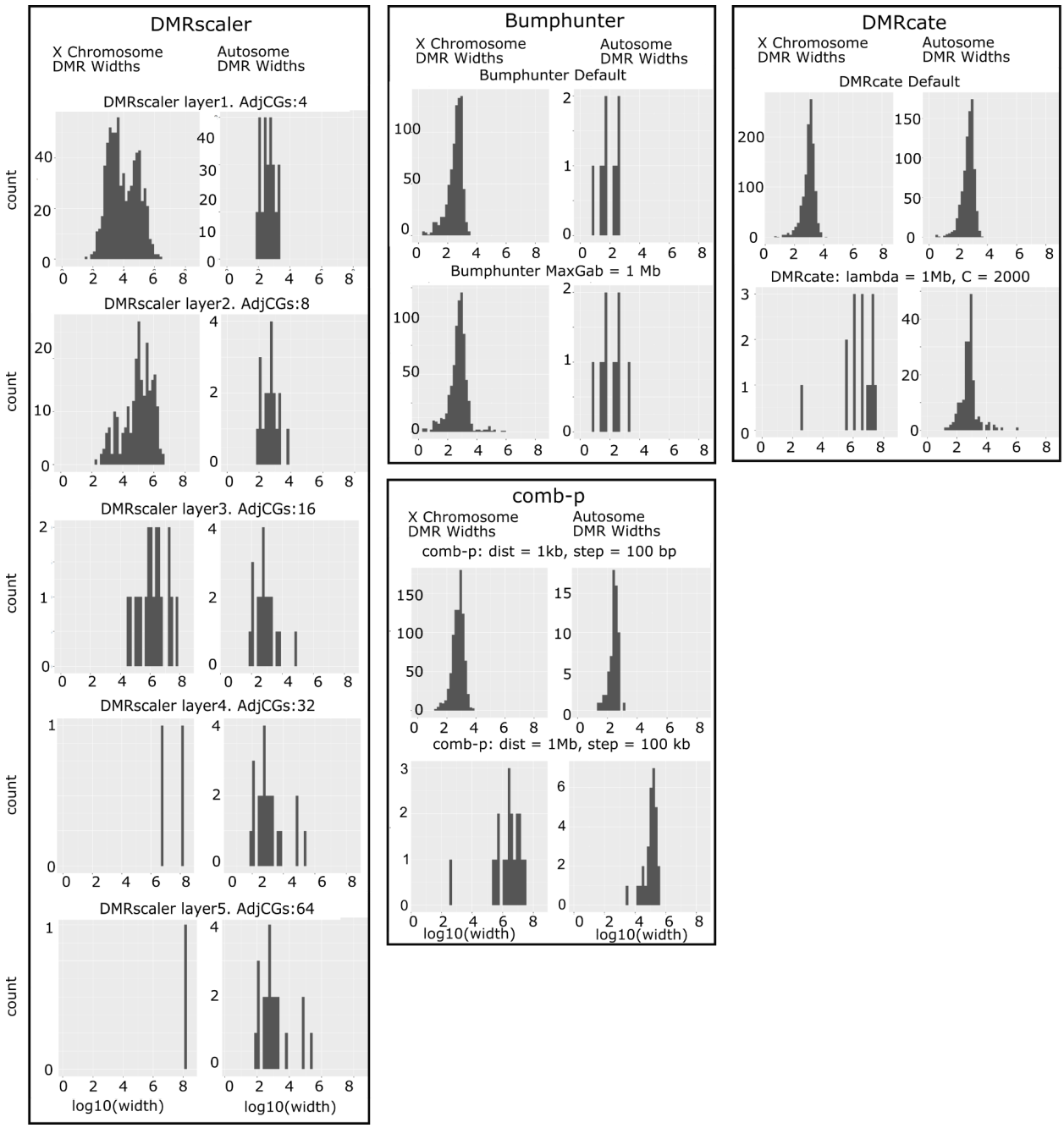


Figure 3-20: Distribution of DMR widths for each method called in XX vs XY sex chromosome analysis.

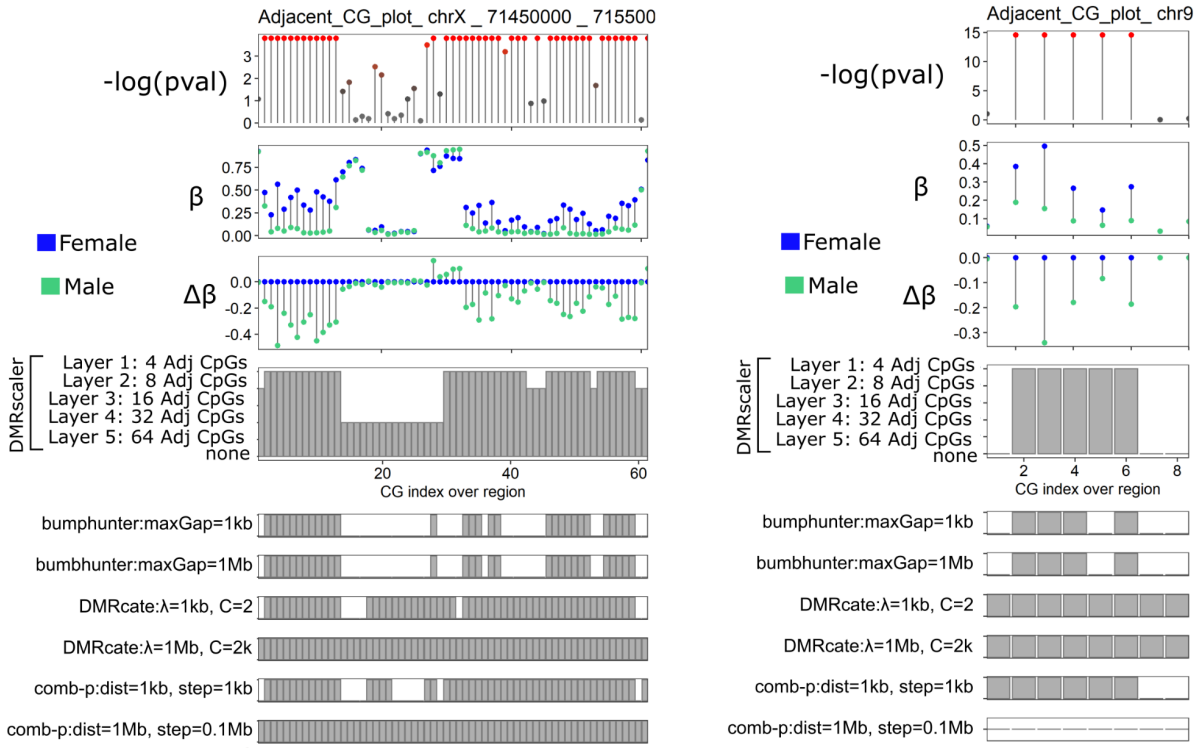


Figure 3-21: Sex analysis. Supplement to Figure 3-17C (left), 3-17D (right). Adjacency plot of CGs overlapping specified regions. Top panel is significance at individual CG level. Beta plot shows mean beta value for each group. Delta beta below shows mean beta value for each group relative to female values. Bottom plot shows in grey bars which layer or method a DMR was called in and from each competing method.

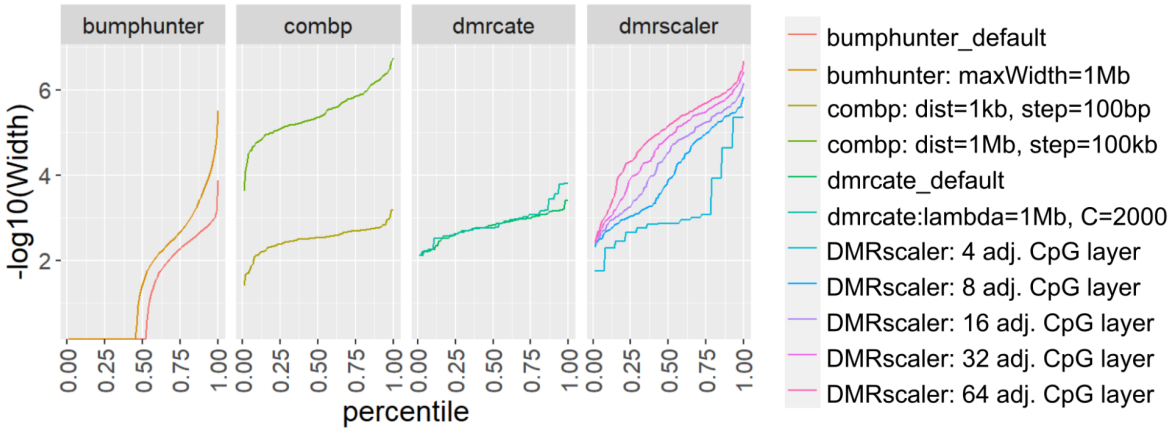


Figure 3-22: Arboleda-Tham Syndrome analysis DMR width percentile plot. DMRs Called by each method for Arboleda-Tham Syndrome analysis ordered by dmr width.

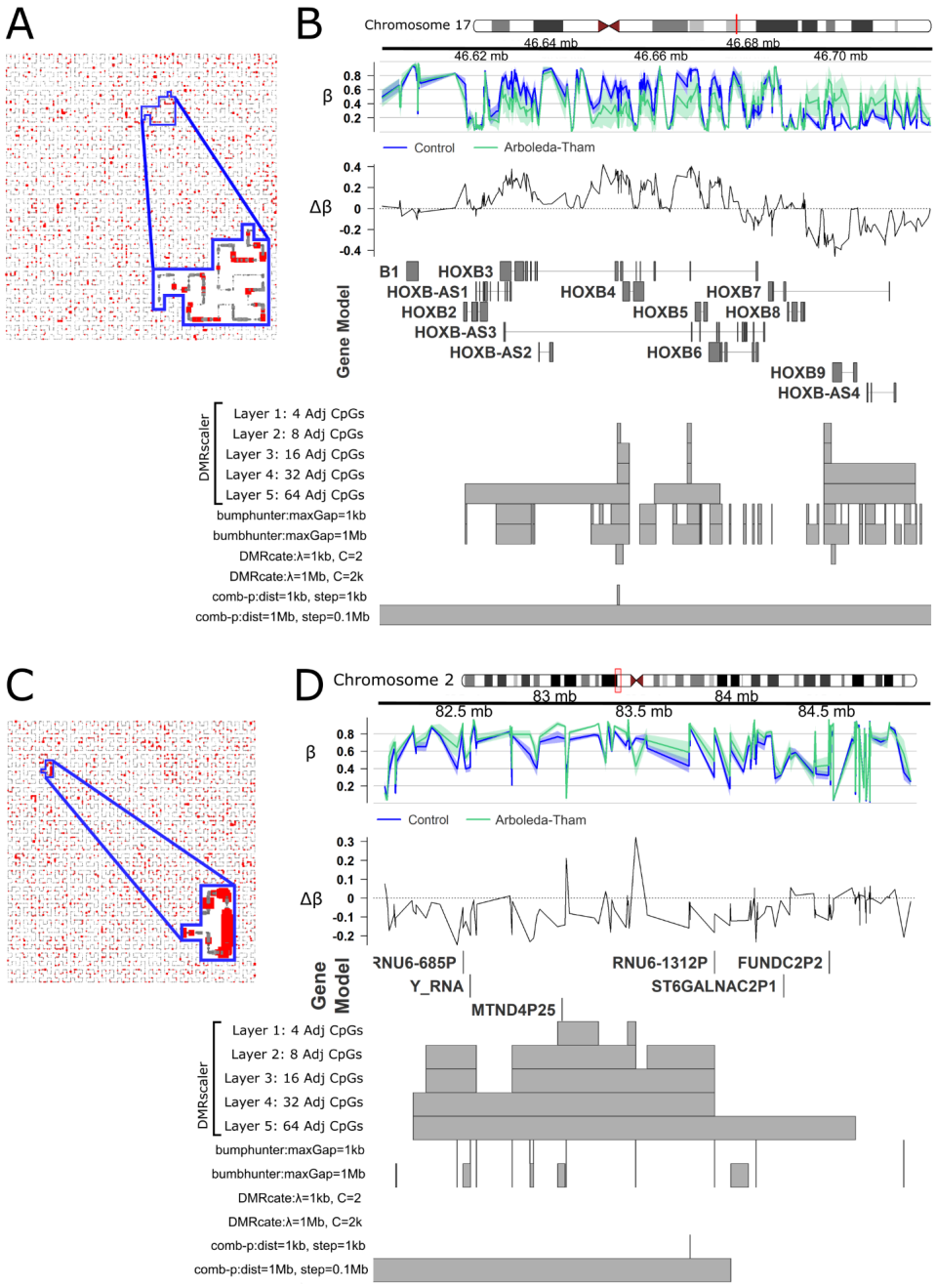


Figure 3-23: Differential Methylation Analysis in Arboleda-Tham Syndrome. (A) Hilbert curve of CpGs from chr17, outlined is the region corresponding to Chr17:46.59-46.73Mb, the HOXB cluster. CpGs with FDR < 0.1 are highlighted red. Point size is scaled to maximum significance value. (B) Chr17:46.59-46.73Mb. HOXB cluster. GVIZ track stack plot. Top track

shows mean β value per group, next track shows Δβ, where Δβ = βControl - βArboleda-Tham. Below the gene model track is the DMR track, highlighting the regions called as a DMR at each result layer from DMRscaler and from each competing method. (C) Chr2:81.5-84.5 Mb. Design same as 4A. (D) Chr2:81.5-84.5 Mb. Tracks same as 4B.

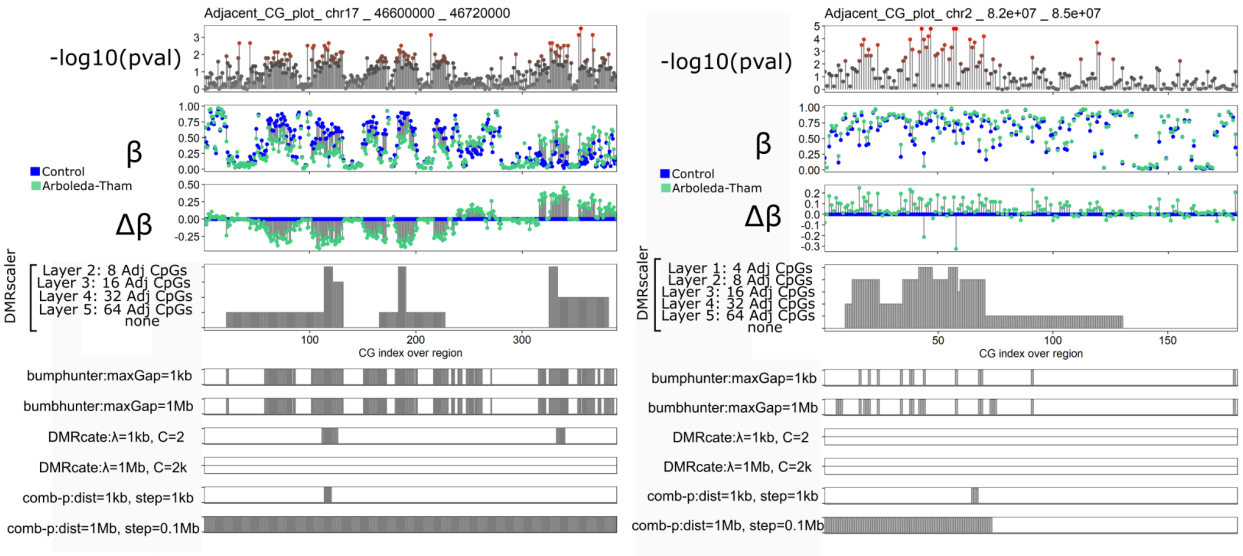


Figure 3-24: Arboleda-Tham analysis. Supplement to Figure 3-23B (left), 3-23D (right). Adjacency plot of CGs overlapping specified regions. Top panel is significance at individual CpG level. Beta plot shows mean beta value for each group. Delta beta below shows mean beta value for each group relative to female values. Bottom plot shows in grey bars which layer or method a DMR was called in and from each competing method.

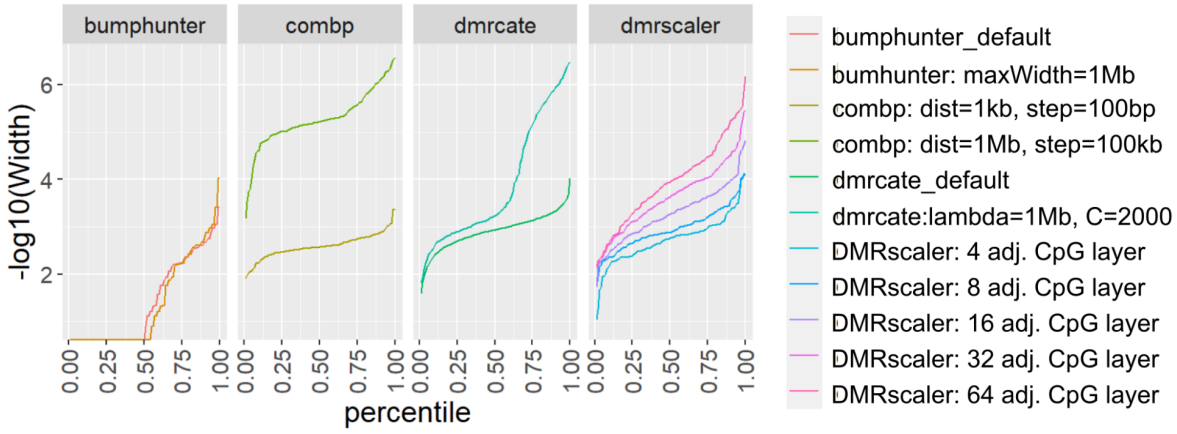


Figure 3-25: Weaver syndrome analysis DMR width percentile plot. DMRs Called by each method for Weaver syndrome analysis ordered by dmr width.

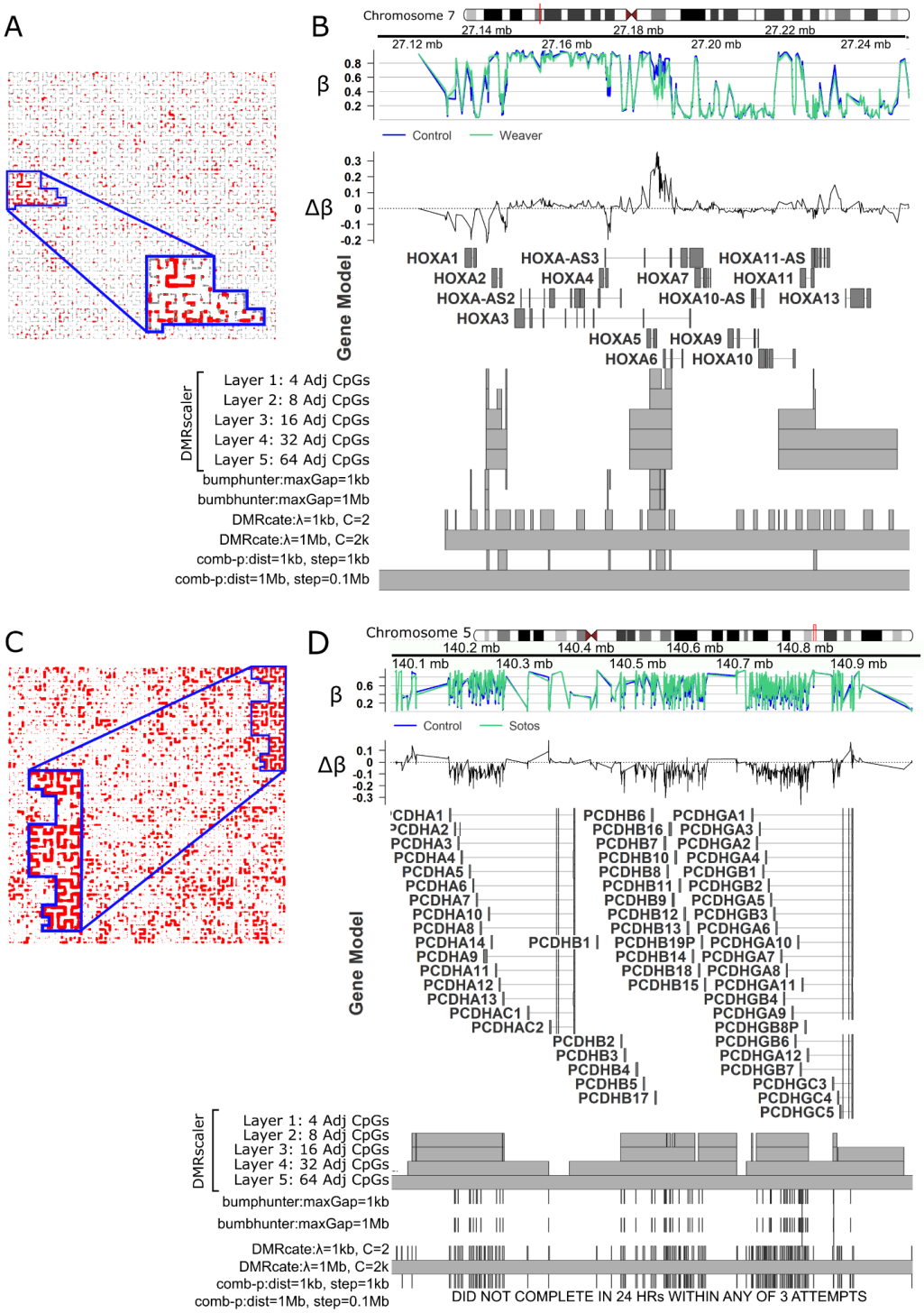


Figure 3-26: Differential Methylation Analysis in Weaver (A,B,C) and Sotos Syndrome (D,E,F). (A) Hilbert curve of CpGs from chr7, outlined is the region

corresponding to Chr7:27.1-27.3Mb, the *HOXA* cluster. CpGs with FDR < 0.1 are highlighted red. Point size is scaled to maximum significance value. (B) Chr7:27.1-27.3Mb. *HOXA* cluster. GVIZ track stack plot. Top track shows mean β value per group, next track shows Δβ, where $\Delta\beta = \beta_{Control} - \beta_{Weaver}$. Below the gene model track is the DMR track, highlighting the regions called as a DMR at each result layer from *DMRscaler* and from each competing method. (C) Chr5:140.1-140.8Mb over the *PCDH* clusters. Design same as 4A. (D) .Chr5:140.1-140.8Mb over the *PCDH* clusters. Tracks same as 4B. except that $\Delta\beta = \beta_{Control} - \beta_{Sotos}$

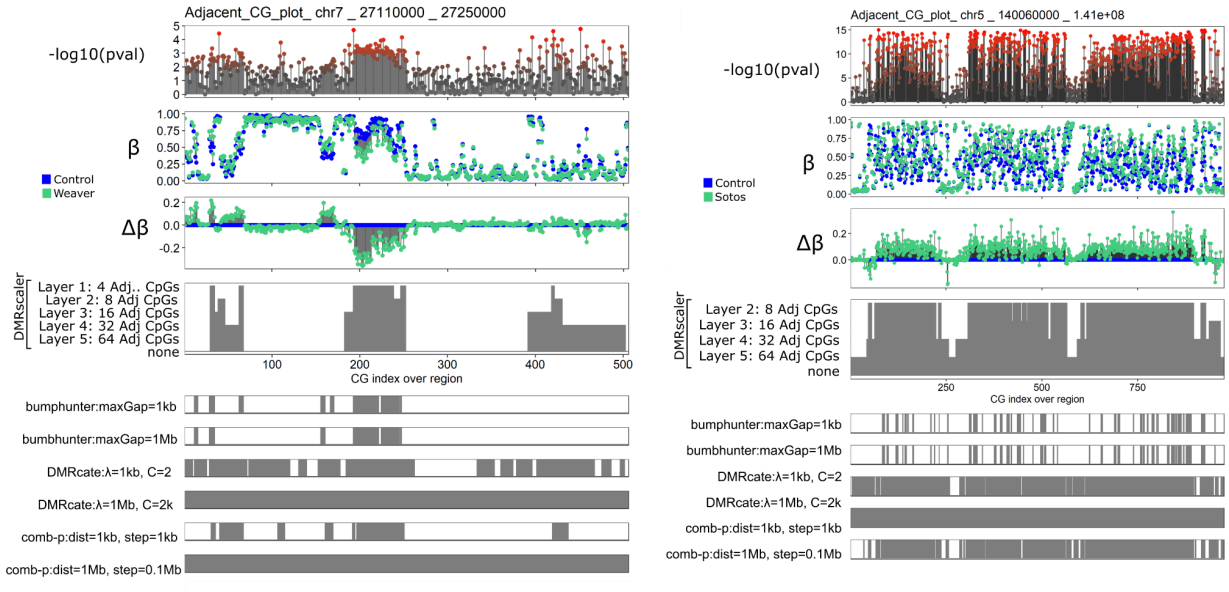


Figure 3-27: Weaver (left) and Sotos (right) analyses. Supplement to Figure 3-26B (left), 3-26D (right). Adjacency plot of CGs overlapping specified regions. Top panel is significance at individual CG level. Beta plot shows mean beta value for each group. Delta beta below shows mean beta value for each group relative to female values. Bottom plot shows in grey bars which layer or method a DMR was called in and from each competing method..

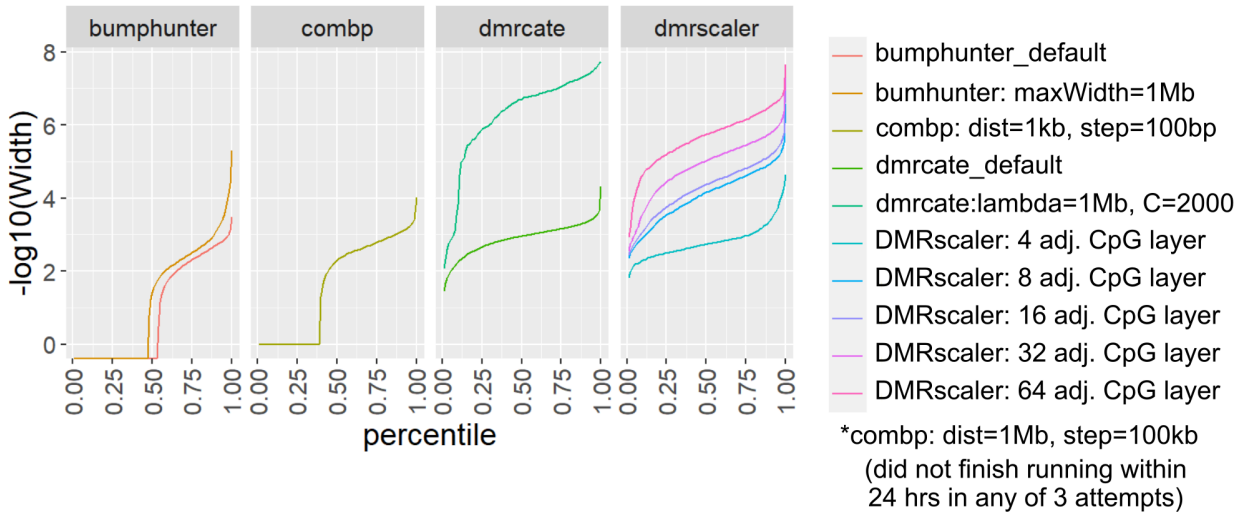


Figure 3-28: Sotos syndrome analysis DMR width percentile plot. DMRs Called by each method for Sotos syndrome analysis ordered by dmr width.

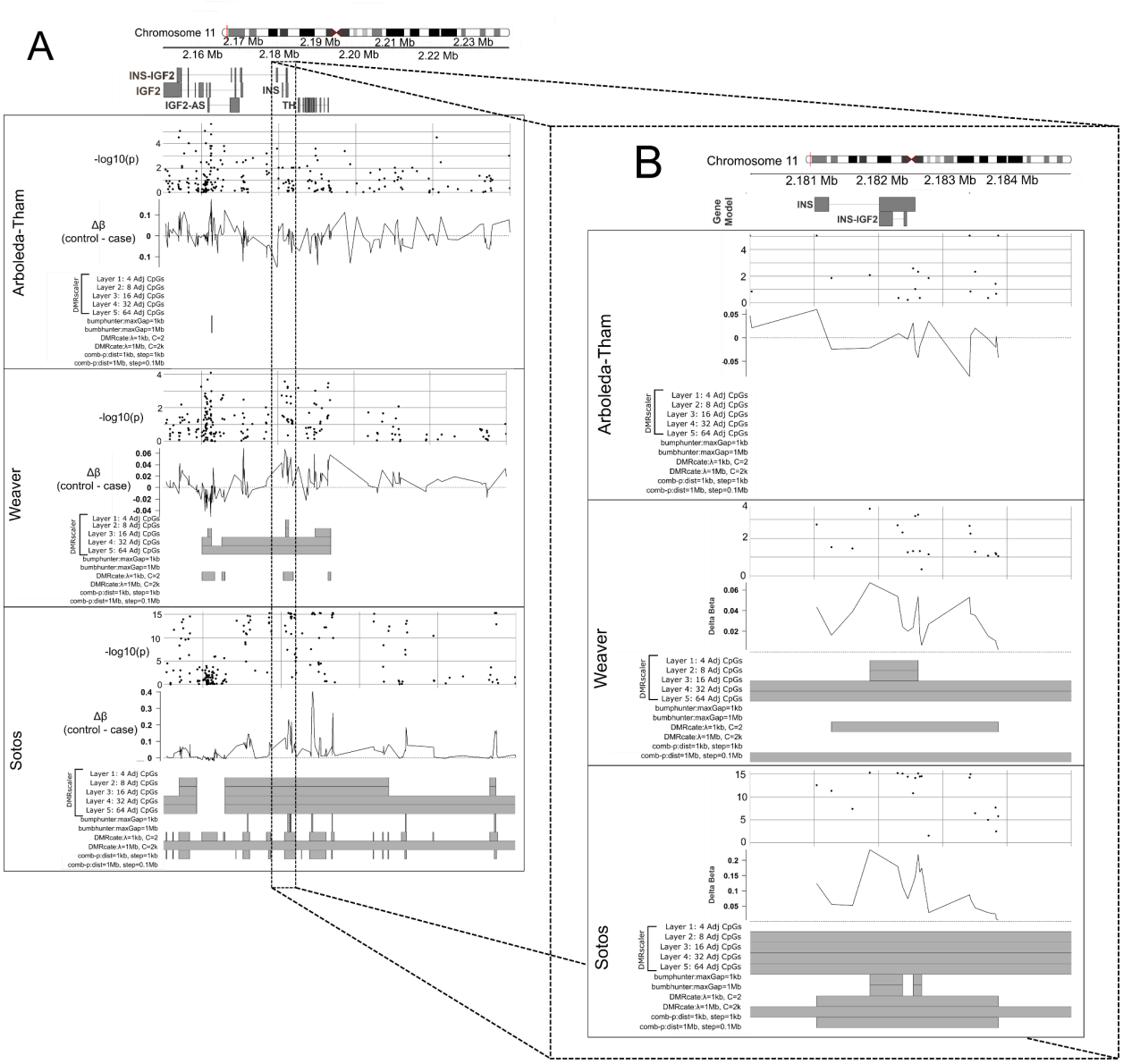


Figure 3-29: *INS, IGF2, INS-IGF2* region. Overlap between the Sotos and Weaver Syndromes, two overgrowth syndromes, identified a region proximal to and overlapping *INS, IGF2*, and *INS-IGF2*. Top track shows -log10(p) significance value for the dataset specified at left, next track shows Δβ, where $\Delta\beta = \beta_{Case} - \beta_{Control}$. Below the gene model track is the DMR track, highlighting the regions called as a DMR at each result layer from DMRscaler and from each competing method.

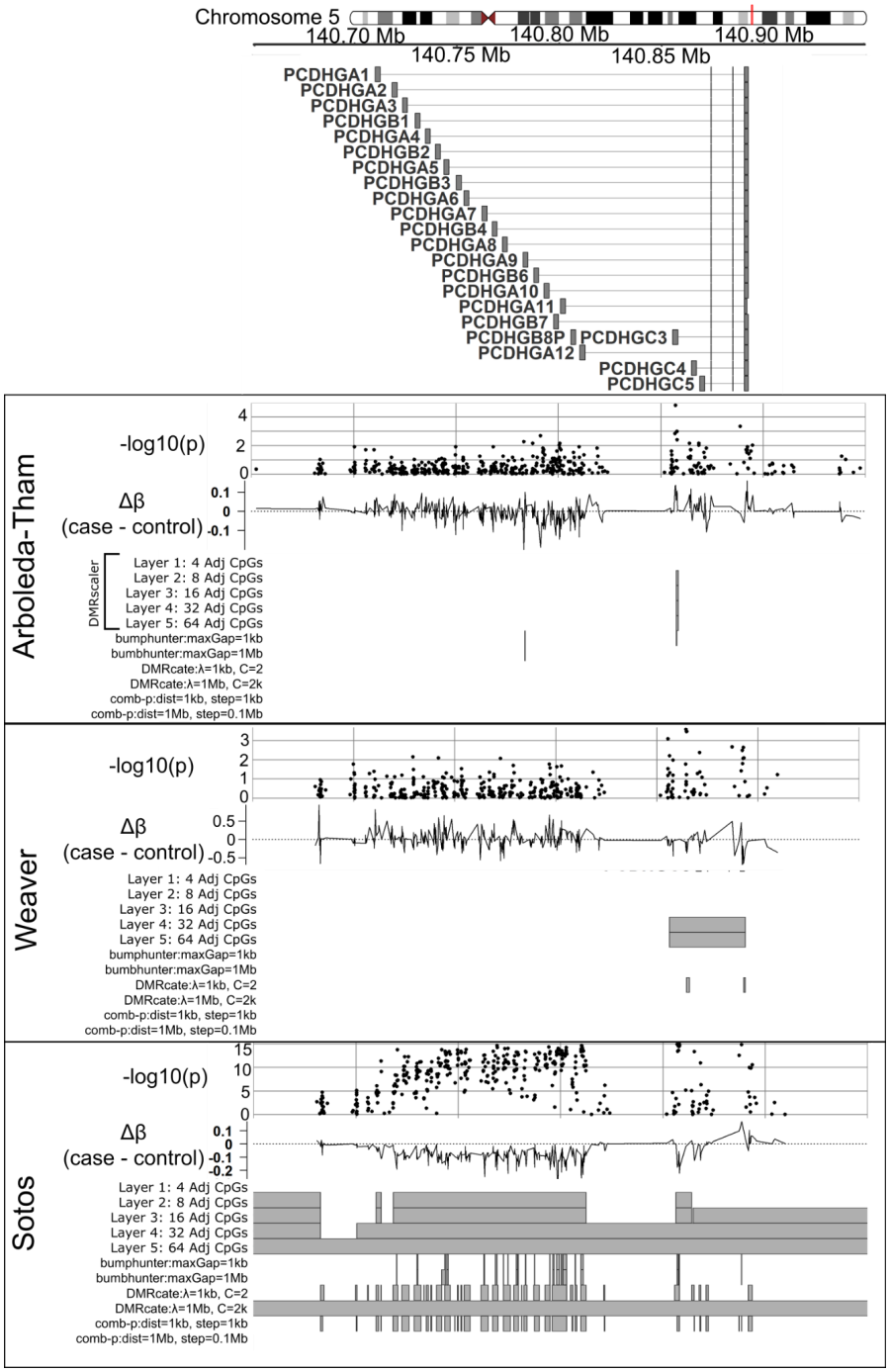


Figure 3-30 : *PCDHG* gene cluster GVIZ plot. PCDHG genes were identified as genes overlapped by some DMR in each of the syndrome datasets analyzed. Within each box, top track is -log10(p) from Wilcox test of beta values between cases and controls.Middle track

shows -log10(p) significance value for the dataset specified at left. Next track shows Δβ, where $\Delta\beta = \beta_{Case} - \beta_{Control}$. Below the gene model track is the DMR track, highlighting the regions called as a DMR at each result layer from DMRscaler and from each competing method.

Arboleda-Tham samples	genomic position	coding change (NM_006766.3)	protein change (NP_006757.2)
Patient 1	chr8:g.41792353G>A	c.3385C>T	p.R1129*
Patient 2	chr8:g.41795056G>A	c.3070C>T	p.R1024*
Patient 3	chr8:g.41792353G>A	c.3385C>T	p.R1129*
Patient 4	chr8:g.41791630C>A	c.4108G>T	p.E1370*
Patient 5	chr8:g.41834753G>T	c.1136C>G	p.S379*
Patient 6	chr8:g.41794839_41794840insG	c.3286_3287insC	p.C1096Sfs*6
Patient 7	chr8:g.41791376dupC	c.4362dupG	p.T1455Dfs*9
Patient 8	chr8:g.41791085A>C	c.4653T>G	p.S1551R

Table 3-1: Arboleda-Tham Syndrome Patient Mutations

	DMRscaler	***bumphunter***	***comb-p***	***DMRcate***
Individual CpG significance	NA (takes p-value as input)	T-Test	NA (takes p-value as input)	T-Test, using M transformed Beta values
DMRs definition	Iterative enrichment testing for significant CpGs within window.	Consecutive CpGs above significance threshold	Groups significant CGs if within window or window interval	Gaussian smoothed regions above significance threshold
DMR significance	Hypergeometric Test	Stouffer's Method	Stouffer-Liptak	Permutation Test
Parameters controlling size (default)	window_size (4,8,16,32,64 adj CpGs)	maxGap (500 bp)	dist, step (no default)	lambda (1 kb), C (2)

Table 3-2 : Comparison of Differential Methylation Methods

method	Δβ	noise	Proportion CGs Diff Methylated	Precision = 1-FDR (±1 SD)	Recall (±1 SD)	Specificity (±1 SD)	F1 (±1 SD)	MCC (±1 SD)	AUCPR (±1 SD)
bumphunter	0.1	0	0.56(±0.1)	0.56(±0.1)	0.06(±0.03)	**1(±6e-04)**	0.11(±0.04)	0.14(±0.04)	0.033(±0.02)
comb-p	0.1	0	0.77(±0.1)	0.71(±0.1)	0.31(±0.1)	0.98(±0.01)	0.41(±0.1)	0.45(±0.09)	0.255(±0.1)
DMRcate	0.1	0	0.78(±0.2)	0.77(±0.2)	0.67(±0.1)	0.95(±0.05)	0.69(±0.07)	0.68(±0.06)	0.58(±0.07)
DMRscaler	0.1	0	**0.99(±0.02)**	**0.98(±0.02)**	**0.89(±0.03)**	**1(±0.006)**	**0.94(±0.02)**	**0.93(±0.02)**	**0.885(±0.03)**
bumphunter	0.2	0	0.76(±0.2)	0.76(±0.2)	0.5(±0.03)	**1(±8e-04)**	0.59(±0.05)	0.36(±0.04)	0.409(±0.08)
comb-p	0.2	0	0.64(±0.3)	0.61(±0.3)	0.2(±0.1)	0.97(±0.004)	0.3(±0.2)	0.33(±0.2)	0.161(±0.1)
DMRcate	0.2	0	0.77(±0.02)	0.77(±0.03)	0.82(±0.03)	0.92(±0.01)	0.8(±0.009)	0.76(±0.01)	0.695(±0.03)
DMRscaler	0.2	0	**0.99(±0.008**)	**0.99(±0.009)**	**0.94(±0.02)**	0.99(±0.006)	**0.96(±0.01)**	**0.96(±0.01)**	**0.932(±0.02)**
bumphunter	0.4	0	0.89(±0.1)	0.89(±0.1)	0.51(±0.02)	**1(±4e-04)**	0.65(±0.04)	0.35(±0.02)	0.507(±0.03)
comb-p	0.4	0	0.78(±0.05)	0.71(±0.07)	0.29(±0.1)	0.98(±0.008)	0.4(±0.08)	0.44(±0.06)	0.239(±0.08)
DMRcate	0.4	0	0.81(±0.03)	0.82(±0.04)	0.81(±0.08)	0.94(±0.03)	0.81(±0.03)	0.78(±0.02)	0.704(±0.06)
DMRscaler	0.4	0	**0.99(±0.008)**	**0.99(±0.008)**	**0.97(±0.02)**	0.99(±0.007)	**0.98(±0.009)**	**0.97(±0.01)**	**0.956(±0.02)**
bumphunter	0.1	0.25	0.35(±0.2)	0.35(±0.2)	0.012(±0.01)	**1(±6e-04)**	0.023(±0.02)	0.053(±0.03)	0.003(±0.003)
comb-p	0.1	0.25	0.68(±0.02)	0.83(±0.02)	0.36(±0.1)	0.98(±0.006)	0.5(±0.1)	0.53(±0.08)	0.337(±0.1)
DMRcate	0.1	0.25	0.71(±0.03)	0.88(±0.03)	0.54(±0.06)	0.96(±0.03)	0.67(±0.04)	0.67(±0.02)	0.508(±0.04)
DMRscaler	0.1	0.25	**0.83(±0.007)**	**0.99(±0.008)**	**0.86(±0.05)**	0.99(±0.006)	**0.92(±0.03)**	**0.92(±0.03)**	**0.848(±0.06)**
bumphunter	0.2	0.25	**0.87(±0.1)**	0.87(±0.1)	0.32(±0.03)	**1(±9e-04)**	0.47(±0.03)	0.24(±0.03)	0.268(±0.05)

comb-p	0.2	0.25	0.49(±0.3)	0.59(±0.3)	0.26(±0.1)	0.97(±0.01)	0.35(±0.2)	0.37(±0.2)	0.226(±0.1)
DMRcate	0.2	0.25	0.66(±0.03)	0.82(±0.04)	0.76(±0.07)	0.94(±0.03)	0.79(±0.03)	0.76(±0.02)	0.672(±0.05)
DMRscaler	0.2	0.25	0.82(±0.02)	**0.97(±0.02)**	**0.85(±0.03)**	**1(±0.004)**	**0.91(±0.02)**	**0.91(±0.02)**	**0.841(±0.02)**
bumphunter	0.4	0.25	**0.96(±0.05**)	0.96(±0.05)	0.32(±0.01)	**1(±7e-04)**	0.48(±0.01)	0.19(±0.02)	0.32(±0.01)
comb-p	0.4	0.25	0.65(±0.04)	0.79(±0.06)	0.25(±0.09)	0.98(±0.01)	0.37(±0.09)	0.43(±0.05)	0.215(±0.08)
DMRcate	0.4	0.25	0.66(±0.03)	0.82(±0.03)	0.8(±0.09)	0.93(±0.04)	0.8(±0.04)	0.78(±0.02)	0.682(±0.06)
DMRscaler	0.4	0.25	0.83(±0.008)	**0.99(±0.006)**	**0.85(±0.03)**	0.99(±0.006)	**0.92(±0.02)**	**0.92(±0.02)**	**0.844(±0.03)**
bumphunter	0.1	0.5	0.15(±0.1)	0.16(±0.1)	0.004(±0.001)	**1(±8e-04)**	0.007(±0.002)	0.017(±0.007)	0.001(±3e-04)
comb-p	0.1	0.5	0.5(±0.07)	0.83(±0.1)	0.32(±0.04)	0.97(±0.01)	0.46(±0.05)	0.5(±0.05)	0.288(±0.07)
DMRcate	0.1	0.5	0.55(±0.02)	0.91(±0.02)	0.41(±0.05)	0.99(±0.01)	0.56(±0.05)	0.6(±0.03)	0.39(±0.05)
DMRscaler	0.1	0.5	**0.63(±0.02)**	**1(±0.004)**	**0.76(±0.01)**	0.99(±0.004)	**0.86(±0.008)**	**0.86(±0.008)**	**0.754(±0.01)**
bumphunter	0.2	0.5	**0.93(±0.05)**	0.93(±0.05)	0.18(±0.01)	**1(±4e-04)**	0.31(±0.02)	0.17(±0.01)	0.169(±0.01)
comb-p	0.2	0.5	0.52(±0.02)	0.86(±0.02)	0.35(±0.04)	0.98(±0.007)	0.49(±0.03)	0.54(±0.02)	0.33(±0.03)
DMRcate	0.2	0.5	0.52(±0.02)	0.9(±0.04)	0.59(±0.1)	0.98(±0.03)	0.7(±0.06)	0.71(±0.04)	0.565(±0.09)
DMRscaler	0.2	0.5	0.64(±0.02)	**0.99(±0.008)**	**0.81(±0.02)**	0.99(±0.006)	**0.89(±0.01)**	**0.89(±0.01)**	**0.797(±0.02)**
bumphunter	0.4	0.5	**0.9(±0.08)**	0.9(±0.08)	0.17(±0.02)	**1(±7e-04)**	0.29(±0.02)	0.16(±0.02)	0.173(±0.01)
comb-p	0.4	0.5	0.51(±0.04)	0.85(±0.03)	0.37(±0.04)	0.99(±0.01)	0.51(±0.04)	0.55(±0.03)	0.351(±0.04)
DMRcate	0.4	0.5	0.5(±0.02)	0.85(±0.03)	0.74(±0.09)	0.96(±0.02)	0.79(±0.04)	0.77(±0.03)	0.679(±0.06)
DMRscaler	0.4	0.5	0.63(±0.01)	**0.99(±0.009)**	**0.79(±0.009)**	0.99(±0.007)	**0.88(±0.008)**	**0.88(±0.009)**	**0.78(±0.01)**

Table 3-3: Feature level evaluation of methods in simulation on proportion of CpGs differentially methylated in DMRs, precision, recall, specificity, F1, RCC, and AUCPR metrics on several choices of noise and Δβ parameters. Noise is the proportion of CpGs within a simulated DMR that are not differentially methylated, Δβ is the difference in the proportion of methylation introduced at non-noise CpGs within the simulated DMR region. Shown are the mean values across five replicates for each measure and the standard deviation. Bolded are the best performing method for the given combination of noise and Δβ parameters.

Sex Analysis DMR Summary Table								
Method	**Chromosome X**				**Autosomes**			
	# DMRs	**Mean DMR width**	**Median DMR width**	**% total width**	**# DMRs**	**Mean DMR width**	**Median DMR width**	**% total width**
DMRscaler								
Layer 1: 4 adj CpGs	694	67.91 kb	6.29 kb	30%	18	616 bp	398 bp	0.00038 %
Layer 2: 8 adj CpGs	246	381.32 kb	114.54 kb	61%	19	935 bp	494 bp	0.00062 %
Layer 3: 16 adj CpGs	20	7.30 Mb	1.54 Mb	94%	21	4.73 kb	587 bp	0.0035 %
Layer 4: 32 adj CpGs	2	75.80 Mb	75.80 Mb	98%	23	19.59 kb	624 bp	0.016%
Layer 5: 64 adj CpGs	1	152.11 Mb	152.11 Mb	98%	22	20.32 kb	606 bp	0.016%
bumphunter								
Default: maxGap = 1kb kb	1258	527 bp	238 bp	0.43%	32	67 bp	1 bp	0.00007 5%
maxGap = 1 Mb	1162	6.76 kb	531 bp	5.1%	32	611 bp	1 bp	0.00068 %
comb-p								
dist = 1 kb, step= 100 bp	2390	567 bp	2 bp	0.87%	580	330 bp	292 bp	0.0067 %
dist = 1 Mb, step = 100 kb	19	6.13 Mb	3.15 Mb	75%	29	140.59 kb	127.23 kb	0.14%
DMRcate								
Default: lambda = 1kb, C=2	1178	1.30 kb	1.09 kb	0.99%	826	658 bp	558 bp	0.019%
Lambda = 1 Mb, C = 2000	15	8.53 Mb	3.95 Mb	83.0%	197	7.63 kb	676 bp	0.052%

Table 3-4 : Sex Analysis DMR Summary Table

Method	Odds Ratio (95% CI)	p-value (Fisher's Exact)
DMRscaler		
Layer 1: 4 adj CpGs	7.57 (6.38-8.99)	1.04e-134
Layer 2: 8 adj CpGs	7.24 (6.07-8.65)	5.93e-100
Layer 3: 16 adj CpGs	51.99 (30.38-90.33)	4.34e-77
Layer 4: 32 adj CpGs	160.44 (25.42-6,396.92)	6.59e-18
Layer 5: 64 adj CpGs	0 (0-319.59)	1
bumphunter		
Default: maxGap = 1kb	14 (11.18-17.61)	7.00E-185
maxGap = 1 Mb	15 (11.95-18.94)	5.36E-193
comb-p		
dist = 1 kb, step= 100 bp	2.07 (1.63-2.62)	5.72E-09
dist = 1 Mb, step = 100 kb	0 (0-Inf)	1
DMRcate		
Default: lambda = 1kb, C=2	1.27 (1.03-1.57)	0.023608
Lambda = 1 Mb, C = 2000	0 (0-43.86)	1

Table 3-5: Enrichment test for association between genes silenced by X-inactivation and DMRs, and genes that escape from X-inactivation and gaps between DMRs. Only CpGs on the X-chromosome overlapping genes are used in the enrichment test.

Arboleda-Tham Analysis DMR Summary Table				
method	# DMRs	mean DMR width	median DMR width	% total width
dmrscaler				
4_loc_window_layer	14	20.41 kb	748 bp	0.0092%
8_loc_window_layer	151	73.92 kb	7.88 kb	0.36%
16_loc_window_layer	224	138.65 kb	34.66 kb	1.0%
32_loc_window_layer	293	261.62 kb	81.35 kb	2.5%
64_loc_window_layer	390	388.07 kb	144.59 kb	4.9%
bumphunter				
bumphunter_1	6443	161 bp	1 bp	0.034%
bumphunter_2	6674	2.93 kb	29 bp	0.63%
dmrcate				
dmrcate_1	45	771 bp	685 bp	0.0011%
dmrcate_2	30	1.22 kb	600 bp	0.0012%
combp				
combp_1	90	381 bp	344 bp	0.0011%
combp_3	263	599.10 kb	226.85 kb	5.1%

Table 3-6: Summary of Arboleda-Tham analysis results

Weaver Analysis DMR Summary Table				
method	# DMRs	mean DMR width	median DMR width	% total width
dmrscaler				
4_loc_window_layer	83	1.10 kb	580 bp	0.0030%
8_loc_window_layer	123	1.55 kb	755 bp	0.0062%
16_loc_window_layer	182	4.67 kb	1.85 kb	0.028%
32_loc_window_layer	218	15.75 kb	4.28 kb	0.11%
64_loc_window_layer	226	54.64 kb	8.88 kb	0.40%
bumphunter				
bumphunter_1	58	177 bp	5 bp	0.00033%
bumphunter_2	55	359 bp	1 bp	0.00064%
dmrcate				
dmrcate_1	2560	1.10 kb	858 bp	0.092%
dmrcate_2	466	235.42 kb	1.71 kb	3.6%
combp				
combp_1	152	457 bp	374 bp	0.0023%
combp_3	396	416.17 kb	165.42 kb	5.4%

Table 3-7: Summary of Weaver analysis results

Sotos Analysis DMR Summary Table				
method	# DMRs	mean DMR width	median DMR width	% total width
dmrscaler				
4_loc_window_layer	507	1.52 kb	548 bp	0.025%
8_loc_window_layer	4295	38.12 kb	14.24 kb	5.3%
16_loc_window_layer	4754	69.20 kb	23.46 kb	11%
32_loc_window_layer	3845	258.23 kb	105.29 kb	32%
64_loc_window_layer	1776	1.22 Mb	555.13 kb	71%
bumphunter				
bumphunter_1	10336	155 bp	1 bp	0.052%
bumphunter_2	9819	1.11 kb	27 bp	0.36%
dmrcate				
dmrcate_1	26817	1.09 kb	903 bp	0.96%
dmrcate_2	282	8.37 Mb	5.26 Mb	77%
combp				
combp_1	34189	416 bp	202 bp	0.46%

Table 3-8: Summary of Sotos analysis results

Syndrome Pair	Layer1 CGs in DMR	Layer2 CGs in DMR	Layer3 CGs in DMR	Layer4 CGs in DMR	Layer5 CGs in DMR
Arboleda-Tham	63	789	1472	3478	9231
Sotos	2671	52004	76386	177375	343144
Weaver	429	874	2101	4542	7797
Arboleda-Tham: Sotos	0	139	310	1351	6827
Arboleda-Tham: Weaver	0	14	18	143	424
Sotos:Weaver	38	333	601	2179	6308
Arboleda-Tham: Sotos : Weaver	0	5	25	93	469

Table 3-9 : Raw Count of Measured CGs in DMRs called by *DMRscaler* : Layer1,2,3,4,5 are equivalent to 4,8,16,32,64 Adjacent CG Layers Respectively. CGs in DMR in each syndrome at each layer. Where multiple syndromes are listed, count represents CGs overlapped by some DMR in measured in each method using *DMRscaler*. Only the 425,733 Measured CGs present on both the Illumina 450k array, used for Sotos and Weaver, and the Illumina EPIC 850k array were used for overlap analysis.

Syndrome Pair	Layer1 OR (OR 95% CI)	Layer2 OR (OR 95% CI)	Layer3 OR (OR 95% CI)	Layer4 OR (OR 95% CI)	Layer5 OR (OR 95% CI)
Arboleda-Tham: Sotos	no overlap	1.72 (1.43-2.06) p=5.6e-8	1.37 (1.21-1.55) p=1.8e-6	1.04 (0.97-1.11) p=0.26	1.03 (0.98-1.08) p=0.20
Arboleda-Tham:Weaver	no overlap	9.94 (5.83-16.94) p=4.7e-10	2.8 (1.76-4.47) p=1.4e-4	4.58 (3.86-5.43) p=1.5e-46	3.00 (2.71-3.31) p=3.1e-77
Sotos : Weaver	17.16 (12.27-24.0) p=1.9e-32	4.95 (4.32-5.67) p=1.3e-95	2.06 (1.88-2.27) p=3.1e-45	1.52 (1.43-1.61) p=1.0e-43	1.55 (1.47-1.64) p=4.6e-56

Table 3-10 : Odds Ratio (OR) for CGs found in DMR at each Layer of *DMRscaler* between all pairs of syndromes : Layer1,2,3,4,5 are equivalent to 4,8,16,32,64 Adjacent CG Layers respectively. Odds ratios (OR) are computed by labeling each measured CG as either in a DMR or not in a DMR for each syndrome to create a 2x2 contingency table to perform the odds ratio test on. Counts in Table 3-9. An confidence interval (CI) of the OR overlapping 1 implies no significant enrichment of CGs from one syndrome in the other.

References

1. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* 94, 677–694 (2014).

2. Sun, W. *et al.* Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* 167, 1385–1397 e11 (2016).

3. Lopez, A. J. & Wood, M. A. Role of nucleosome remodeling in neurodevelopmental and intellectual disability disorders. *Front. Behav. Neurosci.* 9, 100 (2015).

4. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221 (2014).

5. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 498, 220–223 (2013).

6. Watkins, W. S. *et al.* De novo and recessive forms of congenital heart disease have distinct genetic and phenotypic landscapes. *Nat. Commun.* 10, (2019).

7. Kennedy, J. *et al.* KAT6A Syndrome: genotype–phenotype correlation in 76 patients with pathogenic KAT6A variants. *Genet. Med.* 21, 850–860 (2019).

8. Zhang, L. X. *et al.* Further delineation of the clinical spectrum of KAT6B disorders and allelic series of pathogenic variants. *Genet. Med.* 22, 1338–1347 (2020).

9. Choufani, S. *et al.* DNA Methylation Signature for EZH2 Functionally Classifies Sequence Variants in Three PRC2 Complex Genes. *Am. J. Hum. Genet.* 106, 596–610 (2020).

10. Choufani, S. *et al.* NSD1 mutations generate a genome-wide DNA methylation signature. *Nat. Commun.* 6, 10207 (2015).

11. Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* 22, 1680–1688 (2012).

12. Wiehle, L. *et al.* DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res.* 29, 750–761 (2019).

13. Gaston, K. & Fried, M. CpG methylation has differential effects on the binding of YY1 and ETS proteins to the bi-directional promoter of the Surf-1 and Surf-2 genes. *Nucleic Acids Res.* 23, 901–909 (1995).

14. Comb, M. & Goodman, H. M. CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. *Nucleic Acids Res.* 18, 3975–3982 (1990).

15. Prendergast, G. C., Lawe, D. & Ziff, E. B. Association of Myn, the murine homolog of max,

with c-Myc stimulates methylation-sensitive DNA binding and ras cotransformation. *Cell* 65, 395–407 (1991).

16. Bao, J. & Bedford, M. T. Epigenetic regulation of the histone-to-protamine transition during spermiogenesis. *Reproduction* 151, R55–70 (2016).

17. Brewer, L. R., Corzett, M. & Balhorn, R. Protamine-induced condensation and decondensation of the same DNA molecule. *Science* 286, 120–123 (1999).

18. Brown, C. J. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 71, 527–542 (1992).

19. Clemson, C. M., McNeil, J. A., Willard, H. F. & Lawrence, J. B. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J. Cell Biol.* 132, 259–275 (1996).

20. Disteche, C. M. & Berletch, J. B. X-chromosome inactivation and escape. *J. Genet.* 94, 591–599 (2015).

21. Pauler, F. M. *et al.* H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res.* 19, 221–233 (2009).

22. Schwartz, Y. B. *et al.* Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nat. Genet.* 38, 700–705 (2006).

23. Brown, J. L., Sun, M.-A. & Kassis, J. A. Global changes of H3K27me3 domains and Polycomb group protein distribution in the absence of recruiters Spps or Pho. *Proc. Natl. Acad. Sci. U. S. A.* 115, E1839–E1848 (2018).

24. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Sci Adv* 5, eaaw1668 (2019).

25. Soshnikova, N. & Duboule, D. Epigenetic temporal control of mouse Hox genes in vivo. *Science* 324, 1320–1323 (2009).

26. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell* 62, 668–680 (2016).

27. Magklara, A. *et al.* An epigenetic signature for monoallelic olfactory receptor expression. *Cell* 145, 555–570 (2011).

28. Lyons, D. B. *et al.* An epigenetic trap stabilizes singular olfactory receptor expression. *Cell* 154, 325–336 (2013).

29. Sinsheimer, R. L. The action of pancreatic deoxyribonuclease. II. Isomeric dinucleotides. *J.*

Biol. Chem. 215, 579–583 (1955).

30. Mcghee, J. D. & Ginder, G. D. Specific DNA methylation sites in the vicinity of the chicken β-globin genes. *Nature* 280, 419–420 (1979).

31. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480, 490–495 (2011).

32. Viré, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* 439, 871–874 (2006).

33. Li, Y. *et al.* Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biol.* 19, 18 (2018).

34. Fortin, J.-P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.* 16, 180 (2015).

35. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* 44, 40–46 (2011).

36. Butcher, D. T. *et al.* CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions. *Am. J. Hum. Genet.* 100, 773–788 (2017).

37. Benjamini, Y. & Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann. Stat.* 29, 1165–1188 (2001).

38. Mansell, G. *et al.* Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics* 20, 366 (2019).

39. Aryee, M. J. *et al.* Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30, 1363–1369 (2014).

40. Triche, T. J., Jr, Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* 41, e90 (2013).

41. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* 15, 503 (2014).

42. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210 (2002).

43. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–D995 (2012).

44. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* 41, 200–209 (2012).

45. Pedersen, B. S., Schwartz, D. A., Yang, I. V. & Kechris, K. J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* 28, 2986–2988 (2012).

46. Peters, T. J. *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* 8, 6 (2015).

47. Tatbul, N., Lee, T. J., Zdonik, S., Alam, M. & Gottschlich, J. Precision and Recall for Time Series. *arXiv [cs.LG]* (2018).

48. Mallik, S. *et al.* An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays. *Brief. Bioinform.* 00, 1–12 (2018).

49. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434, 400–404 (2005).

50. Sharp, A. J. *et al.* DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.* 21, 1592–1600 (2011).

51. Fisher, E. M. *et al.* Homologous ribosomal protein genes on the human X and Y chromosomes: escape from X inactivation and possible implications for Turner syndrome. *Cell* 63, 1205–1218 (1990).

52. Balaton, B. P., Cotton, A. M. & Brown, C. J. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol. Sex Differ.* 6, 35 (2015).

53. Liu, J., Morgan, M., Hutchison, K. & Calhoun, V. D. A study of the influence of sex on genome wide methylation. *PLoS One* 5, e10028 (2010).

54. Yousefi, P. *et al.* Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. *BMC Genomics* 16, 911 (2015).

55. Douglas, J. *et al.* NSD1 mutations are the major cause of Sotos syndrome and occur in some cases of Weaver syndrome but are rare in other overgrowth phenotypes. *Am. J. Hum. Genet.* 72, 132–143 (2003).

56. Peek, S. L., Mah, K. M. & Weiner, J. A. Regulation of neural circuit formation by protocadherins. *Cell. Mol. Life Sci.* 74, 4133–4157 (2017).

57. Murrell, A. *et al.* An association between variants in the IGF2 gene and Beckwith-Wiedemann syndrome: interaction between genotype and epigenotype. *Hum. Mol. Genet.* 13, 247–255 (2004).

58. Arboleda, V. A. *et al.* De novo nonsense mutations in KAT6A, a lysine acetyl-transferase

gene, cause a syndrome including microcephaly and global developmental delay. *Am. J. Hum. Genet.* 96, 498–506 (2015).

59. Mishima, Y. *et al.* The Hbo1-Brd1/Brpf2 complex is responsible for global acetylation of H3K14 and required for fetal liver erythropoiesis. *Blood* 118, 2443–2453 (2011).

60. Voss, A. K., Collin, C., Dixon, M. P. & Thomas, T. Moz and Retinoic Acid Coordinately Regulate H3K9 Acetylation, Hox Gene Expression, and Segment Identity. *Dev. Cell* 17, 674–686 (2009).

61. Huang, F., Abmayr, S. M. & Workman, J. L. Regulation of KAT6 Acetyltransferases and Their Roles in Cell Cycle Progression, Stem Cell Maintenance, and Human Disease. *Mol. Cell. Biol.* 36, 1900–1907 (2016).

62. Miller, C. T., Maves, L. & Kimmel, C. B. moz regulates Hox expression and pharyngeal segmental identity in zebrafish. *Development* 131, 2443–2461 (2004).

63. Xiao, F.-H., Wang, H.-T. & Kong, Q.-P. Dynamic DNA Methylation During Aging: A 'Prophet' of Age-Related Outcomes. *Front. Genet.* 10, 107 (2019).

64. Wilson, V. L. & Jones, P. A. DNA methylation decreases in aging but not in immortal cells. *Science* 220, 1055–1057 (1983).

65. Weaver, D. D., Graham, C. B., Thomas, I. T. & Smith, D. W. A new overgrowth syndrome with accelerated skeletal maturation, unusual facies, and camptodactyly. *J. Pediatr.* 84, 547–552 (1974).

66. Gibson, W. T. *et al.* Mutations in EZH2 cause Weaver syndrome. *Am. J. Hum. Genet.* 90, 110–118 (2012).

67. Bracken, A. P., Dietrich, N., Pasini, D., Hansen, K. H. & Helin, K. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev.* 20, 1123–1136 (2006).

68. Kanduri, M. *et al.* A key role for EZH2 in epigenetic silencing of HOX genes in mantle cell lymphoma. *Epigenetics* 8, 1280–1288 (2013).

69. El Hajj, N. *et al.* Epigenetic dysregulation in the developing Down syndrome cortex. *Epigenetics* 11, 563–578 (2016).

CHAPTER 4

Balancing the Transcriptome: Leveraging Sample Similarity To Improve Measures of Gene Specificity

Abstract

The spatial and temporal domain of a gene's expression can range from ubiquitous to highly specific. Quantifying the degree to which this expression is unique to a specific tissue or developmental timepoint can provide insight into the etiology of genetic diseases. However, quantifying specificity remains challenging as measures of specificity are sensitive to similarity between samples in the sample set. For example, in the Gene-Tissue Expression project (GTEx), brain subregions are overrepresented at 13 of 54 (24%) unique tissues sampled. In this dataset, existing specificity measures have a decreased ability to identify genes specific to the brain relative to other organs. To solve this problem, we leverage sample similarity information to weight samples such that overrepresented tissues do not have an outsized effect on specificity estimates. We test this reweighting procedure on 4 measures of specificity, Z-score, Tau, Tsi, and Gini in the GTEx data and in single cell datasets for zebrafish and mouse. For all of these measures, incorporating sample similarity information to weight samples results in greater stability of sets of genes called as specific and decreases the overall variance in the change of specificity estimates as sample sets become more unbalanced. Further, the genes with the largest improvement in their specificity estimate's stability are those with functions related to the overrepresented sample types. Our results demonstrate that incorporating similarity information improves specificity estimates' stability to the choice of the sample set used to define the transcriptome, providing more robust and reproducible measures of specificity for downstream analyses.

Introduction

The transcriptome is the set of potential or realized states of gene expression in a cell, tissue, or organism. In human adults there are estimated to be over 400 distinct cell-types that each develop along a unique developmental trajectory [1]. Add to this the diversity of progenitor cells and intermediate transition cell-states that occur earlier in development and one begins to appreciate the complexity of information relayed through the transcriptome. To guide development through this diversity of cell-types and states requires the ubiquitous expression of genes with global functions for cell proliferation and survival as well as the precise expression of genes that control specialized developmental programs. The full extent of a gene's functions are not known *a priori*, so investigating spatial and developmental patterns of gene expression, i.e the context of expression, can provide insight into the gene's function. This context of gene expression can partially explain the phenotype that results when a given gene is mutated [2–5] or be used to investigate whether the gene is involved in the specialized functions of a given cell, tissue, or developmental event and what those specialized functions might be [6–9]. A useful summary of the degree to which a gene's expression leans towards ubiquity or specialization is the aim of gene expression specificity measurements.

While methods for quantifying gene expression are well established [10], measuring the specificity of gene expression requires addressing additional challenges (for a review of current methods of measuring specificity of gene expression see [11]). We address here one emergent challenge associated with the choice of a transcriptomic data sample set on which to measure specificity. Often, tissues and organs can be subdivided in numerous ways, such as dividing the brain into distinct functional domains or along different developmental axes, which often include gradients of expression changes rather than discrete transition points. In the brain, the transcription profiles of these subregions tend to be highly correlated with one another reflecting the common functions and developmental origins of these subregions [12]. This leads to a

problem, however, as measures of specificity are sensitive to the sampling depth into any particular organ system or timepoint [11]. A consequence of this sensitivity is that the ability of measures of gene specificity to detect genes specific to regions or timepoints is diminished if they are highly similar to other regions or timepoints that are overrepresented in the sample set. In the case of the brain, this means that sampling multiple brain subregions decreases a specificity measure's ability to detect brain-specific genes.

One potential means of alleviating the problems associated with adding sampling depth to a particular organ system when building a representative sample set is by using sample similarity information to weight samples to adjust each sample's contribution to the measure of specificity. To establish the intuition for this, consider a sample set that includes biological replicates. Biological replicates of the same sample type tend to have very high similarity and so the weight of any given replicate should be inversely proportional to the number of replicates coming from the same sample type. By extension, the weight of individual samples from different regions of a common organ should be inversely related to the number of regions sampled from that organ. Collectively these examples point to the intuition that the weight assigned to a sample should be inversely related to its similarity to the other samples in the sample set suggesting that sample similarity is a natural metric on which to assign sample weight for measures such as specificity.

Presently, no existing methods for measuring gene specificity take into account the similarity between samples in the sample set used to define the transcriptome. This leads to instability in existing measures of specificity when called on datasets that vary in the depth of sampling of particular biological contexts. Here we propose a generalizable procedure for integrating sample similarity information with measures of gene specificity and demonstrate how this natural integration of sample similarity information stabilizes specificity measures.

Results

Description of problem and proposed solution

The similarity between cells and tissues can distort measures of specificity for gene expression. A balanced sample set where all sampled tissues share nearly the same level of similarity with one another facilitates specificity measures that match intuition (**Figure 4-1A**). However, balancing the sample set by considering only tissues that are at approximately the same level of similarity occurs at the expense of deeper sampling of individual tissue subregions. Adding depth to one sample type (e.g. brain), without an equivalent addition of sample types for other organs can substantially change the measured degree of specificity of gene expression. This problem is demonstrated in the toy example in **Figure 4-1A**, where the brain marker *OLIG1* has specificity comparable to other marker genes, such as *PRSS2* in the pancreas and *MYH6* in the heart, when the sample set is relatively balanced, but when the sample set is expanded to include samples from additional brain subregions *OLIG1* ceases to appear specific to any brain tissues.

For the present study, we compare several measures of gene specificity that are amenable to incorporation of sample weights. These measures are Z-score [13], Tau [14], Tsi [15], and Gini [16,17] which were previously compared in a benchmarking study of measures of gene specificity [11] (see Methods for details). We chose to look at these several measures of specificity to test whether incorporation of sample similarity information to measures of gene specificity could improve a variety of different measures. Throughout the manuscript, we refer to raw specificity measures that do not incorporate sample similarity information as *flat* measures and measures that do incorporate sample similarity information via weights as *weighted* measures.

To test the effect of incorporating sample similarity derived weights on measures of specificity, we used three RNA-seq datasets including a human tissue sample set from the Gene-Tissue Expression (GTEx) project [18], and single cell datasets from zebrafish [19] and mouse

[20]. From GTEx, the matrix of the median gene expression values across all individuals for each of 54 unique tissue types was used. Brain-region samples are overrepresented in the GTEX dataset, making up 24% (13/54) of the different available tissue samples. This enabled us to explore how specificity values varied when measured on balanced sample sets where only one tissue from each organ system was included compared to unbalanced sample sets where there is an overrepresentation of brain regions. The zebrafish single cell dataset comes from [19] and includes 220 unique cell clusters from four developmental time points, subsets of which are used for our analyses. The mouse single cell dataset comes from [20] and includes cells from 98 major cell clusters representing over 50 mouse tissues and cultures, subsets of these clusters were used for our analyses. As a note, we use the term *sample* throughout to refer to either unique tissues or cell clusters as opposed to biological or technical replicates.

To test the effects of measuring specificity on an unbalanced sample set we looked first at the correlation of specificity values measured on sample sets that were either balanced or unbalanced with respect to the set of tissues or cell clusters included. In the balanced GTEx subsets, we include only one brain subregion in the sample set compared to the unbalanced GTEx subset that includes all brain-region samples in the sample set. The correlation between balanced and unbalanced sample sets was repeated for each choice of brain subregion for the balanced sample set. All non-brain samples were included in both the balanced and unbalanced sample sets. Genes with Z-score greater than 2 were considered specific with higher values indicating greater specificity. Tau, Tsi, and Gini measures are all on the scale between 0 and 1, where 0 indicates non-specific expression and 1 indicates the maximum specificity in the tissue or cell type with the highest overall level of expression for that gene. Overall, we observe a strong correlation ($R > 0.9$) between specificity values calculated using the balanced sample set and those calculated using the unbalanced sample set for all measures of gene specificity with the highest average correlation observed for Gini $R = 0.991$ (95% CI: 0.991-0.991, one sample

t-test) and lowest for Tsi R = 0.939 (95% CI: 0.938-0.940, one sample t-test). For Z-score the average correlation was R = 0.962 (95% CI: 0.961-0.963, one sample t-test), and for Tau, R = 0.989 (95% CI: 0.989- 0.989, one sample t-test). A representative example of the specificity scores measured on the balanced and unbalanced sample sets is shown for each measure in **Figure 4-1B**. Relatively strong correlations were similarly found in the single cell datasets.

However, the genes with the largest difference in specificity scores measured between the balanced and unbalanced sample sets are not a random subset of genes. For example, in the GTEx dataset, the top 1% of genes with the greatest positive differences in each specificity score between the balanced and unbalanced sample sets (i.e. where specificity in the balanced sample set is greater than specificity in the unbalanced sample set) are highlighted in **Figure 4-2** in red and represent genes that are the most variable as the sample set becomes unbalanced. Gene ontology enrichment analysis performed on these genes showed substantial enrichment in genes that function in brain related processes (**Figure 4-2**). There was less consistency in terms associated with the genes where specificity in the balanced sample set was set less than specificity in the unbalanced sample set across measures (**Figure 4-3**). The enrichment of brain related terms in the top 1% of genes with the greatest positive difference between the balanced and unbalanced sample sets highlights how overrepresentation of particular sample types can introduce systematic biases into measures of specificity reducing the power of these measures to identify genes specific to the overrepresented sample type.

To address this problem, we decided to leverage sample similarity information to reweight samples in the sample set defining the transcriptome such that similar samples tend to share their weight while more distinct samples tend to retain more of their full weight. The general workflow proposed, which we call the Specificity-Similarity Integration (SSI) procedure, is given in **Figure 4-1C** and discussed in detail in the Methods. In the GTEx dataset, we measured tissue-tissue similarity using each tissue's respective gene expression profile and

found brain samples clustered together with a high degree of intragroup similarity, though cerebellar samples had a lesser degree of similarity than other brain regions (**Figure 4-4**). Following the SSI procedure proposed in **Figure 4-1C**, this sample similarity information was used to generate a sample similarity (or dissimilarity) tree on which Equation 1 (adapted from [21]) was applied to assign a weight to each sample. After incorporating sample similarity information into weights, brain subregions were found to have lower weights compared to more distinctive tissues, such as testis and pituitary (**Figure 4-5**).

We proceeded to incorporate these weights to each of the specificity measures and compared the correlation of specificity values measured on the balanced and unbalanced sample sets to the correlations obtained before weights were applied. For this, each choice of a single brain subregion was used to generate a distinct balanced sample set (n=13) that included a single brain subregion and all non-brain samples, using these 13 replicates we tested whether using the weighted specificity measure resulted in a stronger correlation between balanced and unbalanced sample sets. For all of the specificity measures tested, the correlation between the balanced and unbalanced sample sets increased when the weighting approach was applied compared to when weights were not applied. $P(R\ Z\text{-}score_{weighted} \leq R\ Z\text{-}score_{flat}) = 2.2e\text{-}16$; $P(R\ Tau_{weighted} \leq R\ Tau_{flat}) = 2.2e\text{-}16$; $P(R\ Tsi_{weighted} \leq R\ Tsi_{flat}) = 1.1e\text{-}4$; $P(R\ Z\text{-}score_{weighted} \leq R\ Z\text{-}score_{flat}) = 2.2e\text{-}16$; using welch's t-test for each test (**Figure 4-6**). Additionally, the improved correlation for the weighted measure held as the sample size varied while holding the proportion of the sample set composed of brain samples constant, except for Tsi which has previously been shown to be sensitive to sample size [11] (**Figure 4-7**). This trend of improved correlation between the balanced and unbalanced sample set was further replicated in the zebrafish and mouse single cell datasets. For the zebrafish dataset the test results are as follows: $P(R\ Z\text{-}score_{weighted} \leq R\ Z\text{-}score_{flat}) = 5.2e\text{-}9$; $P(R\ Tau_{weighted} \leq R\ Tau_{flat}) = 4.4e\text{-}8$; $P(R\ Tsi_{weighted} \leq R\ Tsi_{flat}) = 5.0e\text{-}3$; $P(R\ Z\text{-}score_{weighted} \leq R\ Z\text{-}score_{flat}) = 1.8e\text{-}8$, using welch's t-test for each test. For the

mouse dataset the test results are as follows: $P(R\ Z\text{-}score_{weighted} \leq R\ Z\text{-}score_{flat}) = 5.2e\text{-}8$; $P(R\ Tau_{weighted} \leq R\ Tau_{flat}) = 6.9e\text{-}8$; $P(R\ Tsi_{weighted} \leq R\ Tsi_{flat}) = 1.8e\text{-}5$; $P(R\ Z\text{-}score_{weighted} \leq R\ Z\text{-}score_{flat}) = 2.0e\text{-}8$, using welch's t-test for each test (**Figure 4-8, Figure 4-9**).

When gene ontology enrichment analysis was performed on the weighted measures, the enrichment of brain related terms in the top 1% of genes with the largest positive difference in specificity measured between the balanced and unbalanced sample sets decreased for all measures, except for Tsi which did not change substantially (**Figure 4-10**). As was the case for the flat measures, for the weighted specificity measures there was less consistency in terms associated with the genes where specificity in the balanced sample was set less than specificity in the unbalanced sample set across measures. However, for Tau there was enrichment of brain related terms, possibly representing brain subregion specific genes being called as more specific as the sample set size increases with the inclusion of more brain subregions (**Figure 4-11**).

These results suggest that incorporating sample similarity information via weights allows one to include additional samples enriching the transcriptomic diversity within the sample set without necessarily sacrificing the ability to identify particular tissue- and cell-type specific genes.

Validation of similarity-weighted specificity scores

To further test whether integration of similarity information improves the stability of gene specificity measures across variable sample sets, we used the GTEx dataset to quantify the degree of change in specificity scores as the proportion of the sample set composed of brain subregions increased for both the weighted and flat measures. The procedure to calculate the change in specificity is outlined in **Figure 4-12A** and **Figure 4-12B** and a more detailed description of the procedure is included in the Methods section.

Following the procedure outlined in **Figure 4-12A** and **Figure 4-12B**, we observed that the weighted measures exhibited a marked reduction in the variance of the change in specificity measures as additional brain samples were added to the sample set (**Figure 4-12C**). When the similarity-weighting procedure was applied, specificity measurements were more stable as sampling depth into brain regions increased than when the procedure was not applied. For the non-brain samples, assigning weights based on similarity to each tissue resulted in 73.1% (95% CI: 71.9 - 74.3%, paired t-test) lower variance in the change in specificity scores across all genes between the baseline with 1 brain sample included and the full set of 13 brain samples included than when weights were not used. For the brain samples, assigning weights based on similarity to each tissue resulted in 31.6% (95% CI: 25.4 - 37.3%, paired t-test) lower variance in the change in specificity scores across all genes between the baseline with 1 brain sample included and the full set of 13 brain samples included than when weights were not used (**Figure 4-12C**). A similar reduction in variance of specificity values between the baseline of 1 brain sample and inclusion of the full sample set was observed for Tau, Tsi, and Gini as well (**Figure 4-12C**). In contrast, when a similar procedure was used, substituting the brain partitioning with a random partitioning, we observed a much less dramatic difference in the change in variance between the weighted and flat measures. For the random partition, the non-brain sample set was replaced by a random sample set of the same size, called P1, and the brain sample set was replaced with a random sample generated following the same incrementing procedure described in Figures **4-12A**, called P2. We observed a <20% difference in the variance in specificity values in P1 and ~0% difference in the variance in specificity values in P2 between weighted and flat measures in the random partition compared to the 73.1% difference in P1 and the 31.6% difference in P2 between the weighted and flat measures in the brain partitioned sample set when the number of samples in P2 increased from 1 to 13 (**Figure 4-13**, **Figure 4-12C**).

As cut-off values are often used to binarize genes as either specific or non-specific, we wanted to test whether incorporating sample similarity information would also improve the stability of gene sets called as specific as the sample set becomes more unbalanced. To do this, we compared the sets of genes that would be called as specific using different cutoff values as the number of brain subregions included in the sample set increased (**Figure 4-12D**). The Jaccard index, which is the ratio of the intersection and the union of two sets, was used to measure similarity of the gene sets. The Jaccard index ranges from 0, with no elements common to both sets, to 1, with all elements being shared between both sets. We observed that the set of genes specific to brain samples changed substantially over typical Z-score cutoff values between 2 and 3 standard deviations. For example, at a Z-score cutoff of 2 standard deviations the Jaccard index dropped to 0.24 (95% CI: 0.14 - 0.34, t-test) for the flat measure, compared to a Jaccard index of 0.59 (95% CI: 0.54 - 0.63, t-test) at the same cutoff for the weighted measure as the number of brain samples included increased from 1 to 13 (**Figure 4-12D**). The change in the Jaccard index for the set of non-brain sample specific genes was also substantial. At a Z-score cutoff of 2 standard deviations the Jaccard index dropped to 0.65 (95% CI: 0.64 - 0.67, t-test) for the flat measure, compared to a Jaccard index of 0.78 (95% CI: 0.77 - 0.79, t-test) at the same cutoff for the weighted measures as the number of brain samples included increased from 1 to 13 (**Figure 4-12D**). Similar but less dramatic trends were observed for Tau, Tsi, and Gini measures (**Figure 4-12D**). In contrast, when the partition was random such that expanding the sample set included adding samples without high similarity to those already in the set, there was no significant difference in the change in Jaccard statistics between the weighted and flat measures (**Figure 4-14**).

Effect of integrating weights on patterns of specificity

We next wanted to explore the factors which influenced how a gene's specificity score changed in response to integration of weighted similarity information. We first looked at the GTEx dataset. The most striking change in specificity scores that occurred as the sampling depth of brain subregions increased were in genes with brain related functions (**Figure 4-2**). If incorporating sample similarity information reduced the bias introduced by increasing sampling depth in the brain, then we would expect that most of the differences between the weighted and flat specificity scores would occur in genes primarily expressed in the brain and with brain related functions. Indeed when we looked at the top 10 genes with the largest positive difference between the weighted and flat Z-score in each tissue (i.e. where the weighted specificity score was greater than the flat score), the genes with the largest change in specificity value were genes specific to brain samples (**Figure 4-15A**). Even in non-brain tissues, the largest changes in gene specificity were in genes where expression was shared with brain samples (**Figure 4-15A**, **Figure 4-16**). While we focus on protein coding genes for most of our analyses, we also found these patterns to be consistent for lncRNA (**Figure 4-17, Figure 4-18**), which have been observed to have strong patterns of tissue and cell type specificity [22].

For the top 10 genes with the largest negative difference in specificity value between the weighted and flat Z-score in each tissue (i.e. where the weighted specificity score was less than the flat score), the largest effects were in non-brain samples (**Figure 4-15B**). These changes in non-brain samples tended to be in genes with high values of specificity from the flat measure being measured as slightly less specific by the weighted measure (**Figure 4-16**). This effect is likely due to the decrease in the effective sample size caused by downweighting individual brain samples. In the brain samples, while the effect size of the negative difference between the weighted and flat Z-score was modest (**Figure 4-15B**), this difference was associated with genes specifically depleted in brain samples becoming more specifically depleted, suggesting

an increase in the power of the weighted Z-score to detect genes specifically depleted in brain tissues (**Figure 4-16**).

We next looked at the behavior of the flat and weighted specificity scores for genes known to have tissue-specific expression patterns. Brain specific genes *OLIG1* and *OLIG2* are markers for oligodendrocytes, cells which are restricted to the spinal cord and brain [23] though they are less abundant in the cerebellum than other brain regions [24]. With the flat Z-score, *OLIG1* had specificity values in brain samples between 1.02-1.06 SDs in the cerebellar subregions and 1.56-2.03 SDs in the other brain regions and *OLIG2* had specificity values between 0.91-0.97 SDs in the cerebellar subregions and 1.56-2.11 SDs in the other brain regions. When measured using the weighted Z-score *OLIG1* had specificity values between 1.54-1.60 in cerebellar subregions and 2.22-2.75 SDs in other brain regions and *OLIG2* had specificity values between 1.49-1.57 SDs in cerebellar subregions, and 2.33-3.03 SDs in other brain regions for the weighted Z-score. Specificity estimates for the more specific basal ganglia marker, *DRD1* [25] also increased when weights were applied, up to 4.77-5.00 SDs in brain basal ganglia subregions from the 3.22-3.30 SDs by the flat Z-score (**Figure 4-15C**). The specificity scores between the flat and weighted measures were similar for genes specific to uniquely represented tissue types. For example, *PRSS2*, a pancreas specific protease, had a flat Z-score of 6.38 SDs and a weighted Z-score of 5.40 SDs in the pancreas, and *MYH6*, a heart specific myosin heavy chain, had a flat Z-score of 5.16 SDs and a weighted Z-score of 4.77 in the atrial appendage of the heart (**Figure 4-15C**). Similar trends for these marker genes were observed for Tau, Tsi, and Gini coefficients. Further quantification of the differences in genes called as specific between the flat and weighted measure showed a general increase in the number of genes called as specific to brain tissues when weights were applied, and modest differences in the set of genes called as specific between the flat and weighted measures (**Figure 4-19**).

We next performed gene ontology enrichment analysis on the genes that changed from nonspecific to specific in either direction using a Z-score cutoff for classification as specific of 2 SDs. The top 10 terms were those related to synaptic and neurotransmitter function (e.g. synapse organization, neurotransmitter secretion, signal release from synapse) (**Figure 4-15D**). This is consistent with the expectation that weighting based on sample similarity would increase power to detect genes that are specific to tissues that are more deeply sampled and overrepresented in the sample set.

We next repeated these analyses for the zebrafish and mouse single cell datasets. In zebrafish, there was an overrepresentation of cell clusters from brain related cell types as well as a secondary overrepresentation of cell clusters from skeletal muscle cell types. When looking at the top 5 genes from each cell cluster with the greatest positive difference in specificity value measure between the weighted and flat specificity scores, the genes specific to brain and skeletal muscle clusters had the largest absolute change in specificity values measures (**Figure 4-20A**). Those genes with the largest change in other cell clusters tended to be those with expression shared with brain or skeletal muscle (**Figure 4-20A**). For the bottom 5 genes with the largest negative difference in specificity between the weighted and flat specificity scores, the genes with the largest absolute change in specificity value measures were those with high specificity to non -brain or -skeletal muscle cell clusters being measured as slightly less specific (**Figure 4-20B**), likely due to a decrease in the effective sample size between the flat and weighted measures. In the brain and skeletal muscle cell clusters, the largest negative change in specificity values occurred in genes depleted in brain and skeletal muscle cell clusters being measured as more specifically depleted (**Figure 4-20B**, **Figure 4-21**) suggesting an increase in the power to detect specifically depleted genes in these overrepresented cell types when using the weighted specificity measure. In the mouse, the same patterns were observed where genes specific to the overrepresented myeloid lineage cell clusters and the kidney cell clusters were

those that had the greatest positive difference between the weighted and flat measures (**Figure 4-22A**), and those with the more modest negative difference overlapped more with genes specifically depleted in the myeloid and kidney cell types being measured as more specifically depleted (**Figure 4-22B**, **Figure 4-23**).

As observed in the GTEx dataset, when we looked at markers genes in the zebrafish and mouse cell types that were either uniquely represented or overrepresented, we found similar levels of specificity between the weighted and flat specificity measures for those cell types that were uniquely represented and an increase in specificity for markers for overrepresented cell types (**Figure 4-20C**, **Figure 4-22C**). In the zebrafish dataset, gene ontology enrichment analysis of the terms associated with genes that were called as non-specific with the flat measure and specific with the weighted measure found enrichment in terms related to the overrepresented brain and skeletal muscle cell types (e.g. muscle cell development, muscle contraction, brain development, head development) (**Figure 4-20D**). In the mouse datasets a similar trend was observed with the top terms being those related to the overrepresented myeloid cell types, however no terms related to the other overrepresented cell type, kidney cells, were observed in the top 15 enriched terms (**Figure 4-22D**).

Overall, these results demonstrate that the use of our sample similarity weighting procedure improves the stability of gene specificity measures across a variety of sample sets that are balanced or unbalanced with particular tissue or cell types overrepresented. This enables the identification of genes specific to more deeply sampled biological contexts and reduces bias that is otherwise introduced by variation in sampling depth. Implementing this weighting procedure can give researchers more flexibility in building a sample set, allowing greater sampling depth into a cell type, tissue, or organ of interest without sacrificing the ability to detect genes specific to that same cell type, tissue, or organ.

Discussion

Previous work developing and implementing measures of specificity have had a variety of aims including imputation of expression levels for cell and tissue precursors [14], investigating mechanisms of dosage compensation [15], and characterizing conservation of gene expression patterns across evolutionary time [26,27]. While existing measures have been used successfully, we identified a limitation in that these measures lack a mechanism to account for the similarities that exist between cells or tissues. The absence of a mechanism to account for sample similarity makes existing specificity measures sensitive to the choice of sample set used and can introduce bias into analyses, an issue that has been previously noted [11,28]. A feature of this sensitivity to the sample set composition is a loss of measure robustness as the sampling depth of particular developmental lineages increases, particularly for the features that are specific to the more deeply sampled lineage. Greater depth of sampling is necessary for a more complete view of transcriptome diversity and therefore the antagonistic relationship between sampling depth and the stability of specificity measures is problematic.

To address this, we utilized sample similarity information to weight each sample's contribution to measures of gene specificity. In this work we have shown that accounting for similarity between biological samples in the manner proposed makes measures of specificity more robust to sample set variation and improves the ability of these measures to detect features specific to different cell and tissue types, even when the cell or tissue type is overrepresented within the larger sample set.

One component of the procedure proposed here for integrating sample similarity information into measures of gene specificity is the use of a similarity (or dissimilarity) tree structure to partition weight across the sample set, analogous to the method for assigning sequence weight used by the multiple sequence alignment algorithm *ClustalW* [21]. This mechanism is a natural choice when samples can be defined along a natural hierarchy, such as

when the developmental relation between a set of cells is known, however for tissues, which are often composites of cells from distinct lineages, this model is imprecise. While we have demonstrated that using this model to weight samples improves existing measures of gene specificity for tissues, more general graph-based methods that can account for heterogeneous tissue composition may be able to improve upon the method proposed here by refining the weighting of samples for heterogeneous samples.

Applying this workflow on single-cell data avoids the issue of dealing with heterogeneous composites and also provides a higher resolution view of patterns of specificity for gene expression. However, single-cell analysis requires dealing with problems of low read depth and accurate transcript estimation amongst others [29]. Further, as the method proposed here involves calculating a similarity matrix between samples which requires $O(n^2)$ time, performing the calculation on a large dataset of tens of thousands or more cells becomes, though feasible, somewhat resource intensive without additional optimizations. Clustering cells is a common part of most workflows for single cell analysis and provides a convenient work around for these issues [30]. Here we have shown that the Specificity-Similarity Integration procedure can be used on clusters of single cells to achieve improvements to specificity estimates within single cell analyses.

As additional RNA-seq datasets come online, particularly those spanning various stages of development, our method for calculating specificity that is robust to expansion of the sample set will be invaluable. The Developmental Genotype-Tissue Expression (dGTEx) project has recently been announced and will expand on the GTEx project to include samples from neonatal, pediatric, and adolescent individuals. dGTEx will add depth to a large range of developmental stages for many cells, tissues, organs, and will provide a unique opportunity to broadly investigate transcriptomic changes through development [31]. The method for calculating

gene specificity proposed here is a natural model for hierarchical developmental relationships that will be captured in this dataset and that currently exist in datasets for model organisms [19,20,32–34]. We expect that our method can be used to facilitate improved investigations into the dynamics of gene expression across development in a transcriptome-wide context.

Here we have demonstrated that integrating sample similarity information into measures of gene expression specificity in cells and tissues improves the robustness of these measures to variation in the underlying sample set. By improving the stability of specificity measures to deeper sampling of particular biological contexts of interest, the proposed procedure can facilitate the analysis of patterns of gene expression that captures both the broad, by including a diverse set of cell or tissue types, as well as the focused perspective, by allowing greater depth of sampling of highly similar cell or tissue types. This procedure for integrating sample similarity can easily be extended to measure the specificity of other functional measures of the genome and epigenome such as histone modification or DNA methylation features.

Methods

Data availability

The GTEx data used for the analyses described in this manuscript were obtained from the Genotype-Tissue Expression (GTEx) Project which was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS [18].

website: https://www.gtexportal.org/home/datasets

access date: March 1, 2022

file: GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz

The zebrafish single cell dataset comes from [19].

website: https://cells.ucsc.edu/zebrafish-dev

access date: March 1, 2022

cell annotation file: meta.tsv

cell expression file: exprMatrix.tsv.gz

The mouse single cell dataset comes from [20].

website : https://figshare.com/articles/dataset/HCL_DGE_Data/7235471

access date: March 1, 2022

cell annotation file: annotation_rmbatch_data_revised417.zip

cell expression file: dge_rmbatch_data.tar.gz

Data preprocessing

For the data from GTEx, transcripts per million (TPM) values were recalculated after removing mitochondrial gene reads, to prevent signal driven by relative mitochondrial abundance in tissues, and after removing non-protein coding genes. Expression values in TPM were then log transformed as log10(TPM + 1). The addition of 1 to TPM value before taking the log was done to avoid the issue of taking log of 0, and also because very low TPM estimates are unstable across replicates at standard sequencing depths in the tens of millions of reads. Following this transformation, log10(TPM +1) values for gene expression were scaled with median normalization across all samples [35].

For the single cell data from the zebrafish and mouse datasets, cell cluster annotations were obtained from their respective source studies. These cluster annotations can be found in their respective "cell annotation file" linked in the Data Availability section above. RNA read counts were obtained from their respective "cell expression file" and reads were aggregated across all cells within each cluster. Reads from mitochondrial and non-coding RNAs were filtered out. Clusters with less than 100k total reads after this filtering were then removed from further analysis. Genes with read counts < 10 for each cluster were set to 0 to reduce noise

caused by low read counts. Read counts for each gene were then transformed to TPM values by multiplying read counts by 1e6 and dividing by the sum of read counts for each cluster. These TPM values were then log10(TPM + 1) transformed. These transformed log10(TPM + 1) values were scaled with median normalization across all samples [35].

General algorithm for incorporating sample similarity information into measures of gene specificity

The Specificity-Similarity Integration (SSI) Algorithm in **Figure 4-1C** outlines the general workflow for integrating sample similarity information with an arbitrary measure of specificity. Beginning with a matrix of log transformed gene expression values for a set of samples (genes as rows, samples as columns) sample similarity is measured (SSI step a.). The use of the gene expression matrix for measuring sample similarity is suggested as the gene expression matrix is already required for measuring gene specificity, however other feature sets could be used to assign sample similarity. The important component is to have a mechanism for generating a meaningful sample similarity matrix. Several measures of similarity (cosine, canberra, euclidean, manhattan) were tested and each of the similarity measures tested produced similar intuitive sample similarity structure. For example, each measure found brain samples to have a high degree of similarity with one another. The major difference in measures of similarity was the average similarity across all pairs of samples **Figure 4-24**. For downstream analyses, cosine similarity was used as it has previously been shown to be robust in high dimensional datasets in benchmarking studies [36,37]. The next step is to apply a hierarchical clustering algorithm on the sample similarity matrix (SSI step b.). Single, average, and complete clustering were tested and each produced similar intuitive clusters of samples (e.g. brain samples clustered together; tibial, aortic, and coronary arteries clustering together; etc) (**Figure 4-25**). Average linkage clustering was used as it has previously been shown to be robust when the size of cluster groups vary

substantially [38]. Other methods could be substituted so long as a suitable tree structure is generated for sample representation, where suitability can be determined, for example, on metrics such known developmental relations between tissues or cells. The dissimilarity tree is then used to determine the sample weights (SSI step c.) with the recursive function given in Eq. 1 and described in the section below. The final step is to use a specificity function that allows sample weights with the initial log transformed expression value matrix (SSI step d.). The specificity functions used in this paper are discussed below.

Assignment of sample weights

Sample weights are assigned using the recursive function:

$$w_i = \frac{d_{i,p(i)}}{n_i} + w_{p(i)} \tag{1}$$

where w_i is the weight of node i in the dissimilarity tree (where dissimilarity = 1 - similarity). $p(i)$ is the parent of node i. $d_{i,p(i)}$ is the distance between node i and its parent node $p(i)$. n_i is the number of descendant leaf nodes for node i, where a leaf node is considered a descendant of itself. Weight of the root node is set to zero. Weighting method is based on that introduced for the guide tree implemented in the *ClustalW* sequence alignment algorithm [21]. **Figure 4-26** provides an example calculation.

Specificity measures tested

Four different specificity scores were used to measure how changes in the depth of sampling of certain regions affected the variance in specificity scores assigned to genes. For each equation, n is the number of tissues and x_i is the expression of a gene of interest in tissue i.

The first measure is Z-score [13], which determines specificity by calculating how many standard deviations away gene expression in a given tissue is from the mean expression value across all tissues for that gene. It is calculated as:

$$Z_i = \frac{x_i - \bar{x}}{s} \tag{2}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.2}$$

where Z_i is the Z-score in tissue i, x_i is the gene expression value in tissue i, $\bar{x}$ is mean expression of the gene of interest across all tissues and s is the standard deviation in expression of the gene of interest across tissues. The more positive the Z-score, the more specific a certain gene is to a certain tissue.

The weighted version of this equation is given by:

$$Z_{wi} = \frac{x_i - \bar{x}_w}{s_w} \tag{3}$$

where, from [39],

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{3.1}$$

$$s_w = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2}{\sum_{i=1}^{n} w_i}} \tag{3.2}$$

and w_i is the weight of a given tissue, and other variables are the same as in Eqs. 2, 2.1 - 2.2.

The second measure is tau (τ)[14], which is a tissue specificity measure ranging from 0 to 1, with genes with tau near 0 being more ubiquitously expressed and scores near 1 being more

specifically expressed. At the extremes, a score of 0 corresponds to a gene with equal expression across all tissues, while a score of 1 represents a gene only expressed in one tissue. In a benchmark of measures for gene specificity, tau was found to be consistently the most robust measure of gene specificity on several metrics [11]. Tau is calculated as:

$$\tau = \frac{\sum_{i=1}^{n}(1 - \hat{x_i})}{n-1} \tag{4}$$

where

$$\hat{x_i} = \frac{x_i}{\max\limits_{j \in \{1,\ldots,n\}} x_j} \tag{4.1}$$

with the weighted version of the equation being

$$\tau_{weighted} = \frac{\sum_{i=1}^{n}(w_i - w_i\hat{x_i})}{(\sum_{i=1}^{n} w_i) - w_k} \tag{5}$$

with

$$k \text{ such that } x_k = \max_{j \in \{1,\ldots,n\}} x_j \tag{5.1}$$

where $\hat{x_i}$ is the same as that for Eq. 4 given in Eq. 4.1. The domain of the weighted tau is the same as the unweighted tau.

The third measure is tissue specificity index (Tsi)[15], which measures specificity on a scale of $1/n$ to 1. For any given gene, $1/n$ represents equal gene expression across tissues, while 1 represents expression only in one tissue. Tsi is calculated as:

$$tsi = \frac{\max\limits_{j \in \{1,\ldots,n\}} x_j}{\sum_{i=1}^{n} x_i} \tag{6}$$

with the weighted version of the equation being

$$tsi_{weighted} = \frac{w_j \max\limits_{j \in \{1,\ldots,n\}} x_j}{\sum_{i=1}^{n} w_i x_i} \tag{7}$$

The weighted Tsi has a similar domain as the unweighted version, except that the lower bound is $w_k / \sum_{i=1}^{n} w_i$ with $k \text{ such that } x_k = \max_{j \in \{1,\ldots,n\}} x_j$, instead of $1/n$.

The fourth specificity measure was the Gini coefficient [16,17], a measure of inequality commonly used in economics. Existing on a 0 to $(n-1)/n$ scale, for any gene of interest, a score of 0 represents uniform distribution of gene expression across tissues, while a score of $(n-1)/n$ would indicate that a gene is only expressed in one tissue. The Gini coefficient is calculated as:

$$Gini = \frac{n+1}{n} - \frac{2\sum_{i=1}^{n}(n+1-i)x_i}{n\sum_{i=1}^{n} x_i} \tag{8}$$

where x_i are ordered from least to greatest.

The weighted version from [40] is given by

$$Gini_{weighted} = 2\sum_{i=1}^{n} w_i(x_i - \bar{x})(\hat{F}_i - \bar{F})/\bar{x} \tag{9}$$

where

$$\hat{F}_i(x) = \sum_{j=0}^{i-1} w_j + w_i/2 \tag{9.1}$$

with $w_0 = 0$ and again with x_i ordered from least to greatest . $\bar{F}$ is the mean of $\hat{F}_i$. The domain of the weighted Gini index is similar to the unweighted version except that the upper bound is $((\sum_{i=1}^{n} w_i) - 1)/\sum_{i=1}^{n} w_i$ instead of $(n-1)/n$.

Specificity measure robustness testing

The GTEx dataset was used for the specificity measure robustness testing. To test the robustness of measures of specificity, the change in specificity estimates as the dataset came to contain an increasing proportion of brain samples was followed. For this the GTEx dataset was used which consists of 54 tissue types in total, of which 13 (25%) are from different brain subregions. The GTEx dataset was partitioned into 41 non-brain tissues, P1, and 13 brain tissues, ¬P1. The following procedure was then repeated 8 times using a unique brain subregion sample for the baseline and a unique order of addition for the remaining brain subregion samples:

To begin, one brain sample was selected at random and placed in P2, this is $P2_{baseline}$. The union of P1 and $P2_{baseline}$, P1 U $P2_{baseline}$, was then taken as the sample set. Specificity was then measured using the P1 U $P2_{baseline}$ sample set with each of the flat and weighted specificity measures. The results generated using a single randomly selected brain sample serve as the baseline to compare estimates of specificity as additional brain samples were added to the sample set.

Next brain samples were added successively to $P2_{baseline}$ and specificity recalculated on P1 U $P2_{n=i}$, where *i* is the number of brain samples in P2 in the current iteration. The variance in the change in specificity between specificity measured on P1 U $P2_{n=i}$ and P1 U $P2_{baseline}$ across all genes was recorded and used in generating Figure 2C. The sets of genes called as specific at various cutoff values from the specificity values measured on P1 U $P2_{n=i}$ and on P1 U $P2_{baseline}$ were compared using the Jaccard index. The Jaccard index was recorded and used in generating Figure 2D. This was repeated until the sample set included all 13 brain tissues.

Gene Ontology analysis

The clusterProfiler package [41] in R was used to perform enrichment analyses and generate gene ontology [42] plots. Sets of genes were defined as specified in relevant sections of text or figure captions and enrichment was tested against the set of all genes in the GTEx, mouse, or zebrafish expression matrix after filtering non-protein coding and mitochondrial genes. Benjamini-Hochberg procedure [43] was used to adjust p-values for significance. The Biological Process set of GO terms was used throughout.

Code Availability

All code used for analyses in this manuscript are available at: https://github.com/leroybondhus/gene_specificity

Figures and Tables

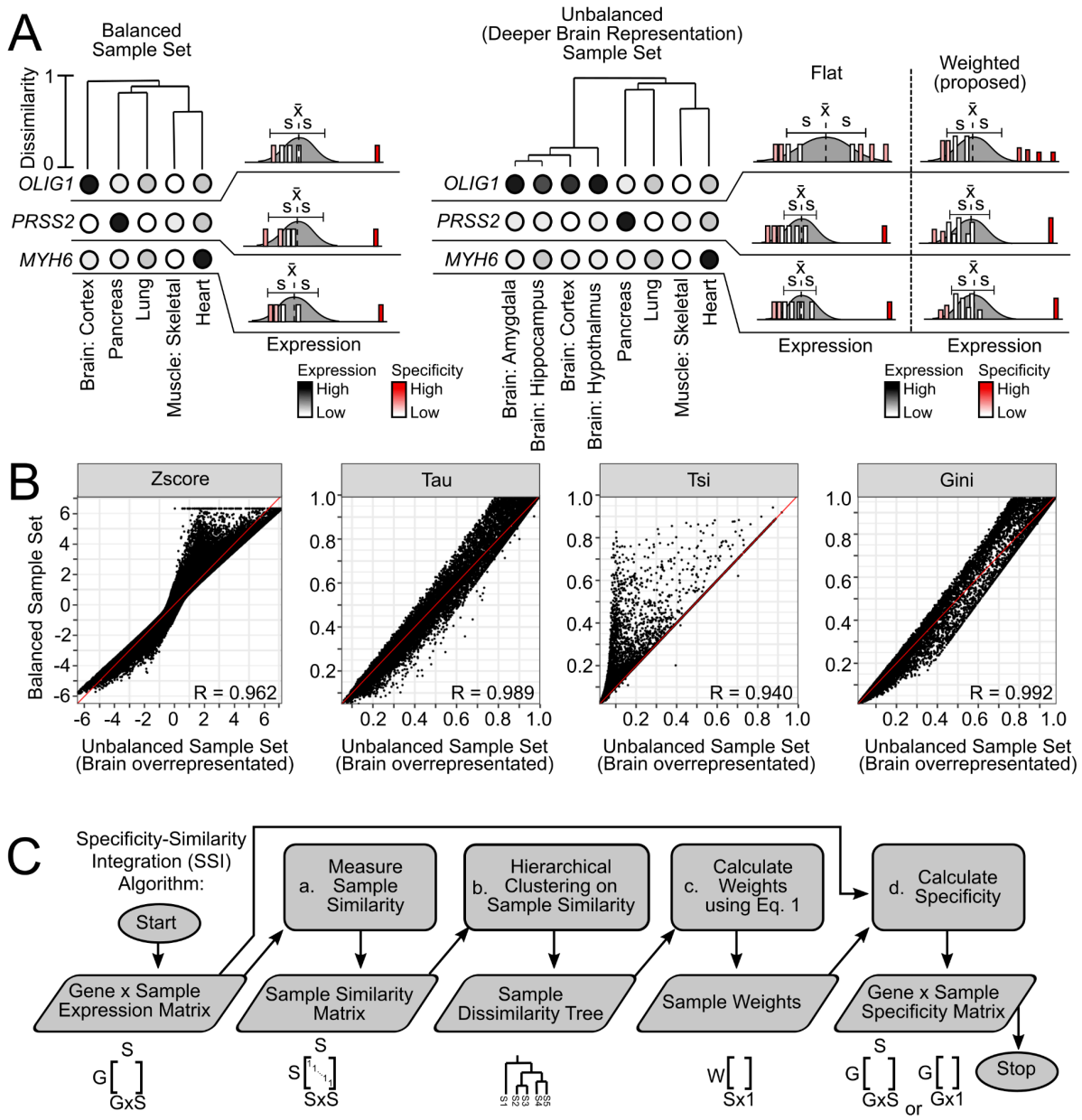


Figure 4-1. Problem with unbalanced sample sets for measuring gene specificity and the proposed solution A) Toy diagram of problem addressed. Global (dis)similarity of tissues is represented as a dendrogram for the balanced sample set (left) and the unbalanced sample set (right) that has an excess of brain subregions included. The color of each dot represents the

relative expression of the gene in the given tissue sample. Fitted normal curve is shown to the right with sample mean ($\bar{x}$) and sample standard deviation (s) for log expression values. Bars plotted with the fitted normal curves each represent an individual tissue sample's expression and the bar's relative height represents that sample's relative weight. Specificity, as measured by the Z-score, is the number of standard deviations of the bar from the sample mean for the given gene-sample pair and is represented by the color of the bar. **B)** change in specificity measures with deeper brain sampling. On x- and y-axes are specificity values measured on the unbalanced and balanced sample sets respectively for each gene (or gene-tissue pair for Z-score). The unbalanced sample set includes all non-brain samples and all brain subregion samples while the balanced sample set includes all non-brain and one randomly selected brain subregion sample from the GTEx dataset **C)** Specificity-Similarity Integration (SSI) Algorithm or workflow proposed for integrating sample similarity information into specificity measures.

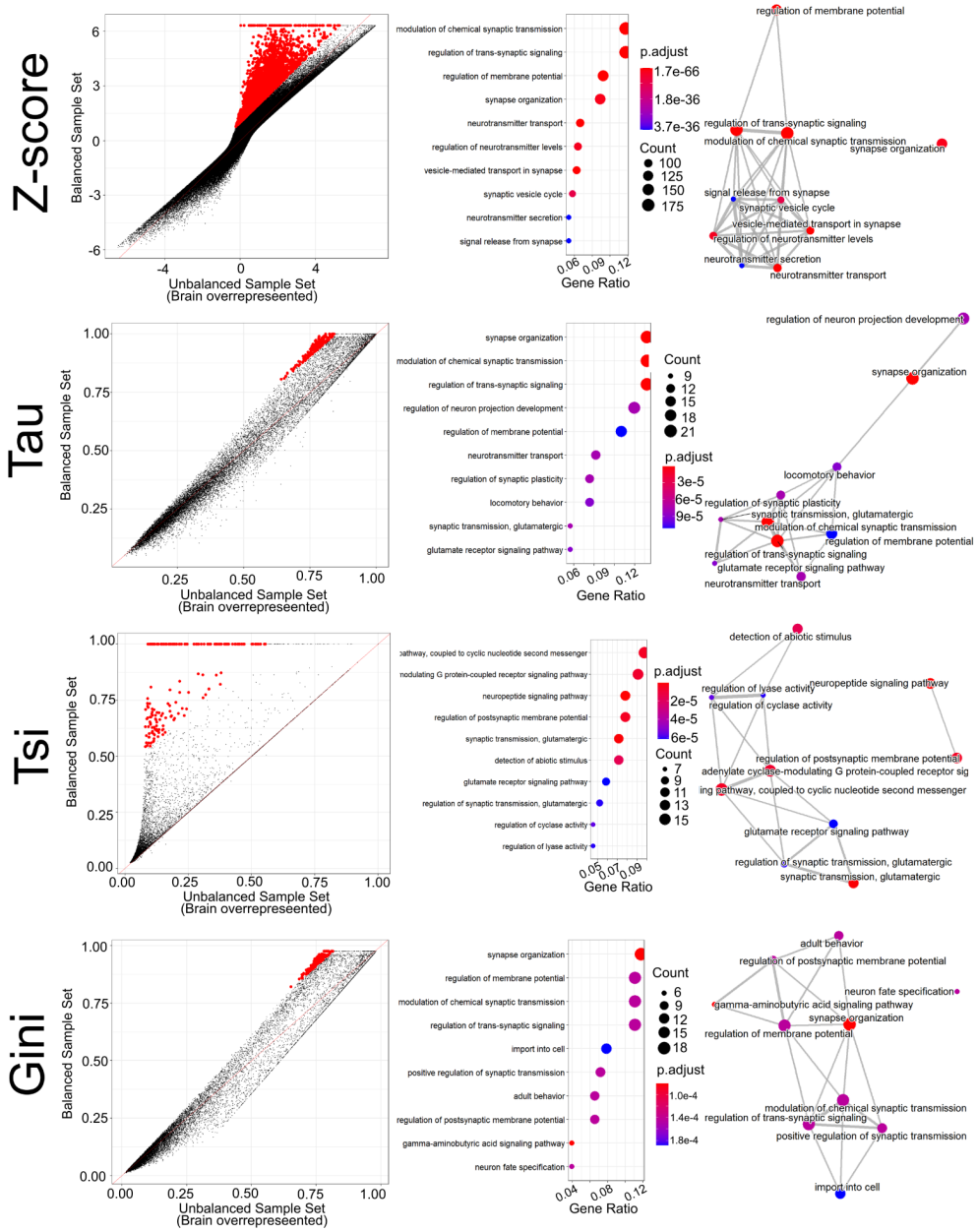


Figure 4-2. From the GTEx dataset, genes with greater specificity measured in balanced than unbalanced sample set for the flat specificity measures. Left column) On the x- and y-axes are specificity values measured on the unbalanced and balanced sample sets respectively for each gene (or gene-tissue pair for Z-score). The unbalanced sample set

includes all non-brain samples and all brain subregion samples while the balanced sample set includes all non-brain and one brain subregion sample from the GTEx dataset. Highlighted in red are the top 1% of genes with the greatest difference in specificity between balanced and unbalanced sample sets with a difference of at least 1 SD for Z-score and 0.1 for other measures. **Middle column)** Gene Ratio, i.e. the proportion of the gene set with the given GO term, for the top 10 GO terms enriched in the highlighted set from the left column **Right column)** Network plots between the top 10 GO terms from middle column, where the edge width indicates the number of shared genes between a connected pair of terms.

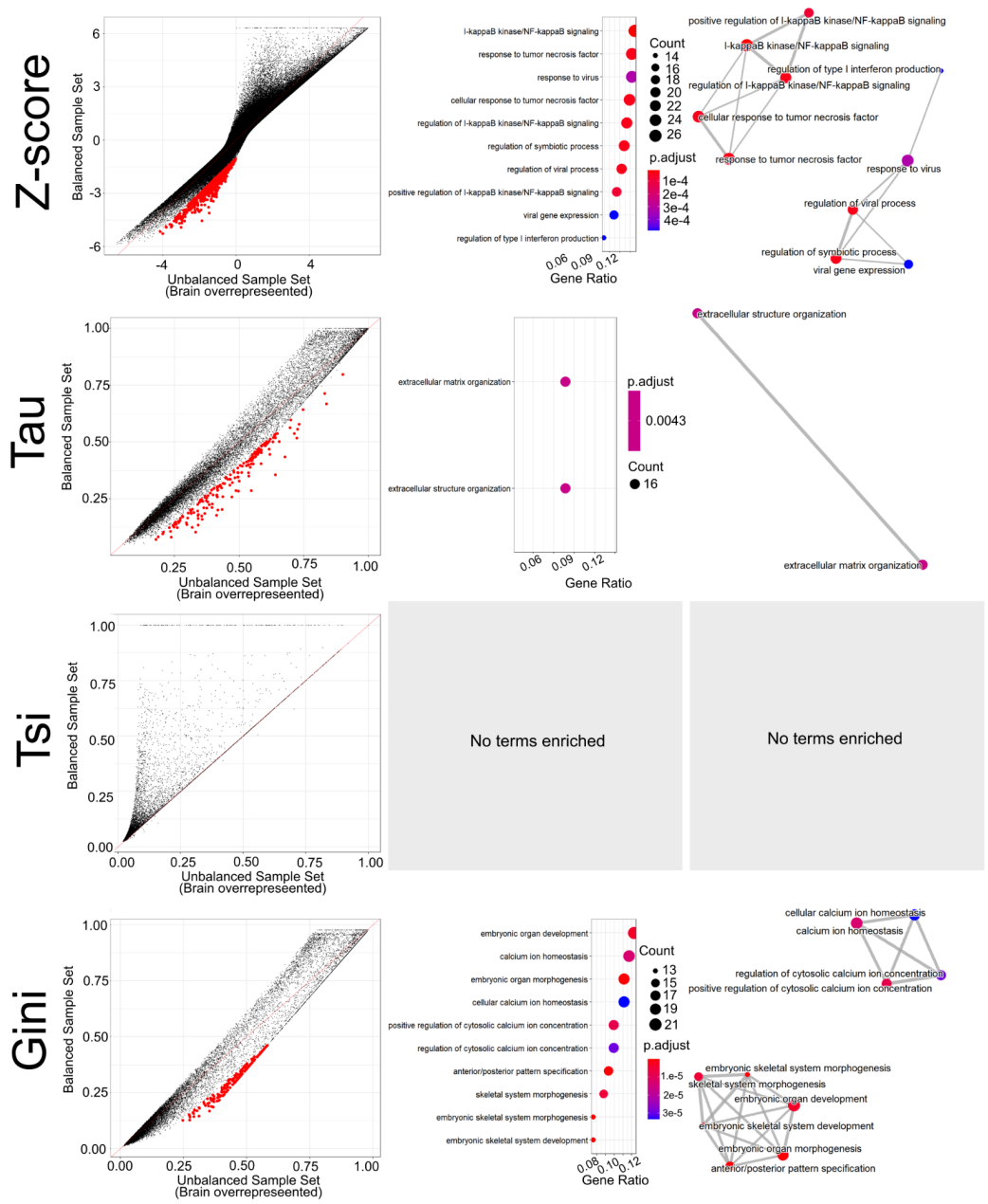


Figure 4-3: From the GTEx dataset, genes with greater specificity measured in unbalanced than balanced sample set for the flat specificity measures. Left column) On the x- and y-axes are specificity values measured on the unbalanced and balanced sample sets respectively for each gene (or gene-tissue pair for Z-score). The unbalanced sample set includes all non-brain samples and all brain subregion samples while the balanced sample set

includes all non-brain and one brain subregion sample from the GTEx dataset. Highlighted in red are the top 1% of genes with the greatest difference in specificity between balanced and unbalanced sample sets with a difference of at least 1 SD for Z-score and 0.1 for other measures. **Middle column)** Gene Ratio, i.e. the proportion of the gene set with the given GO term, for the top 10 GO terms enriched in the highlighted set from the left column **Right column)** Network plots between the top 10 GO terms from middle column, where the edge width indicates the number of shared genes between a connected pair of terms.

Figure 4-4: Similarity structure between samples for GTEx dataset with clear clustering of brain samples with one another. Sample similarity heatmap and dendrogram. Similarity matrix generated by cosine similarity, dendrogram clustering at top generated by average linkage clustering

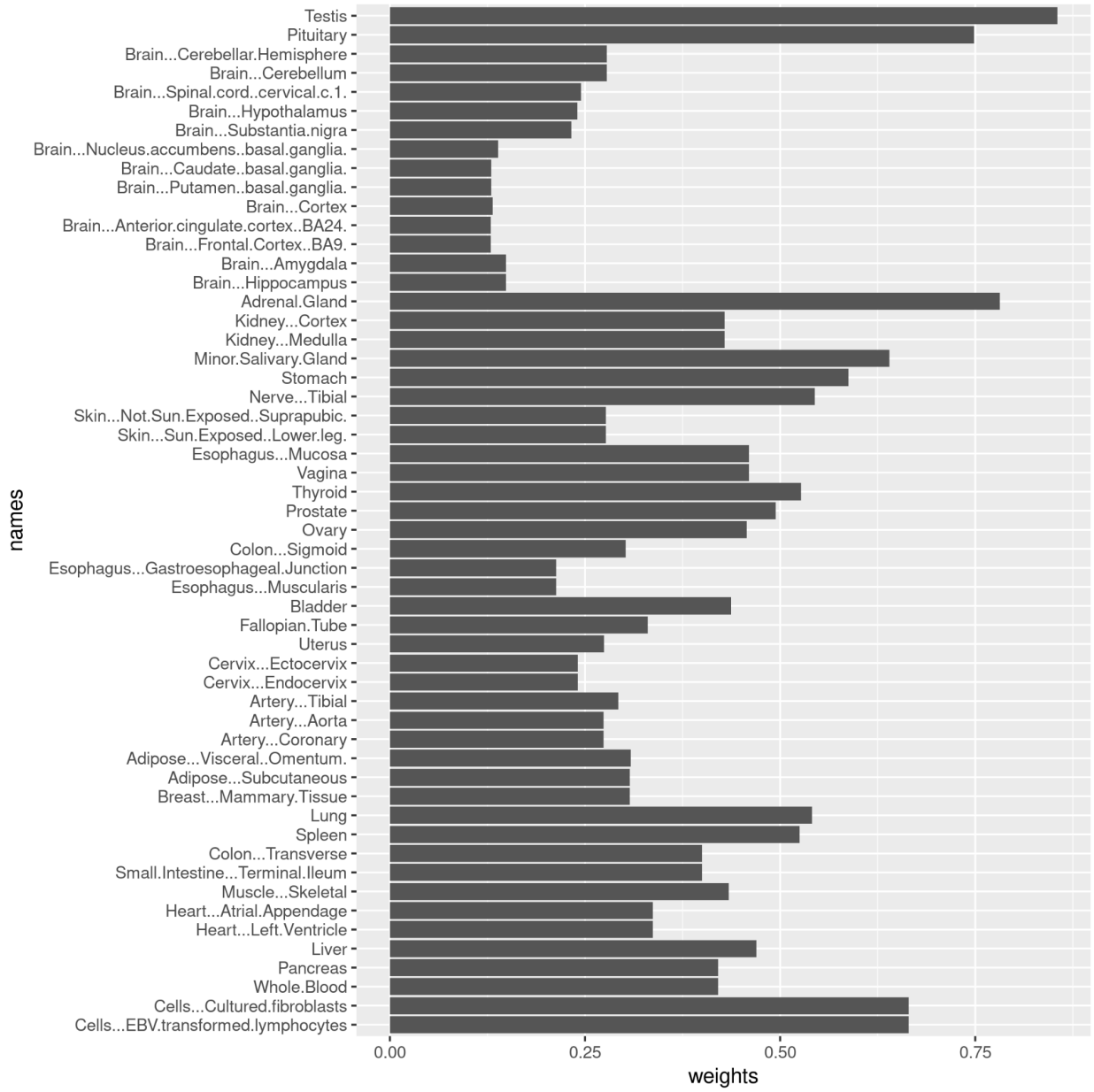


Figure 4-5: **From the GTEx dataset, SSI algorithm decreases weight assigned to individual brain samples relative to other sample types.** Sample weights calculated following the SSI algorithm (Figure 1C) using cosine similarity to generate similarity matrix and average linkage clustering to generate (dis)similarity tree.

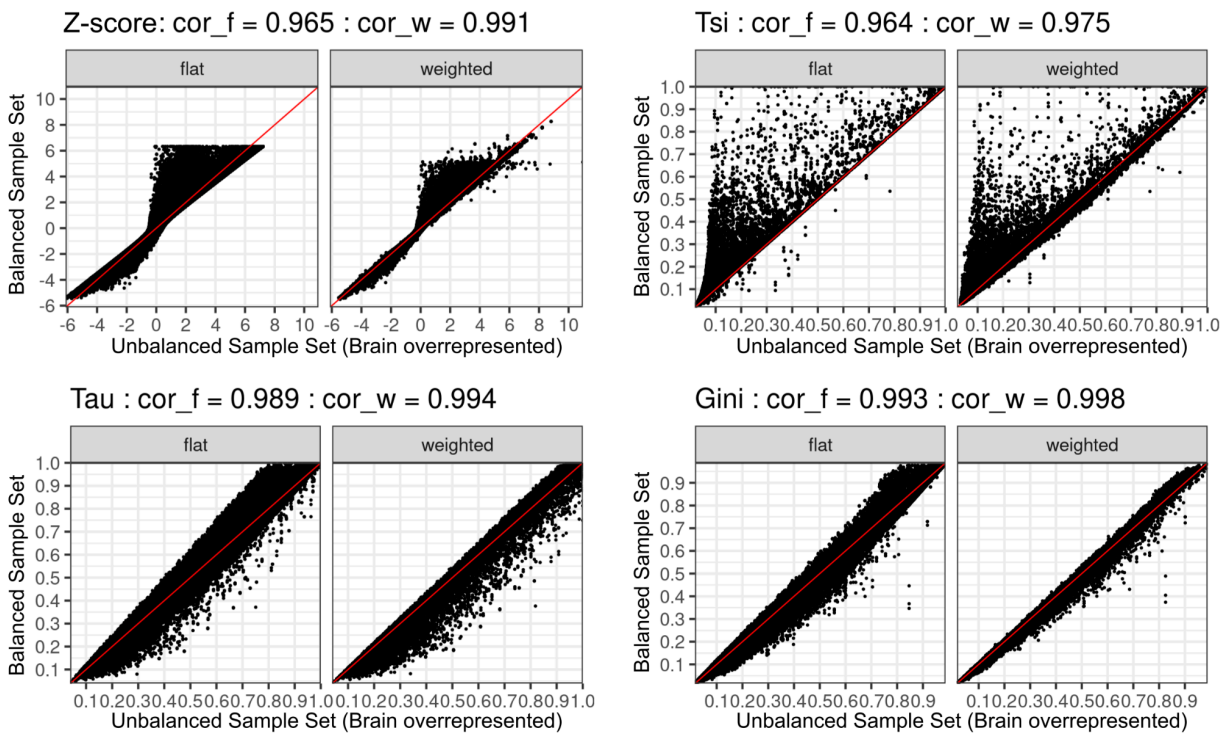


Figure 4-6: **From the GTEx dataset, incorporating sample similarity information increases correlation between specificity values measured on the balanced and unbalanced sample sets.** On the x- and y-axes are specificity values measured on the unbalanced and balanced sample sets respectively for each gene (or gene-tissue pair for Z-score). The unbalanced sample set includes all non-brain samples and all brain subregion samples while the balanced sample set includes all non-brain and one brain subregion sample from the GTEx dataset. For each specificity measure, Z-score, Tsi, Tau, and Gini , specificity was calculated without weight (i.e. the flat measure) and with weights (i.e. the weighted measure). cor_f is the pearson correlation between the balanced and unbalanced sample sets for the flat measure, cor_w is the pearson correlation between the balanced and unbalanced sample sets for the weighted measure.

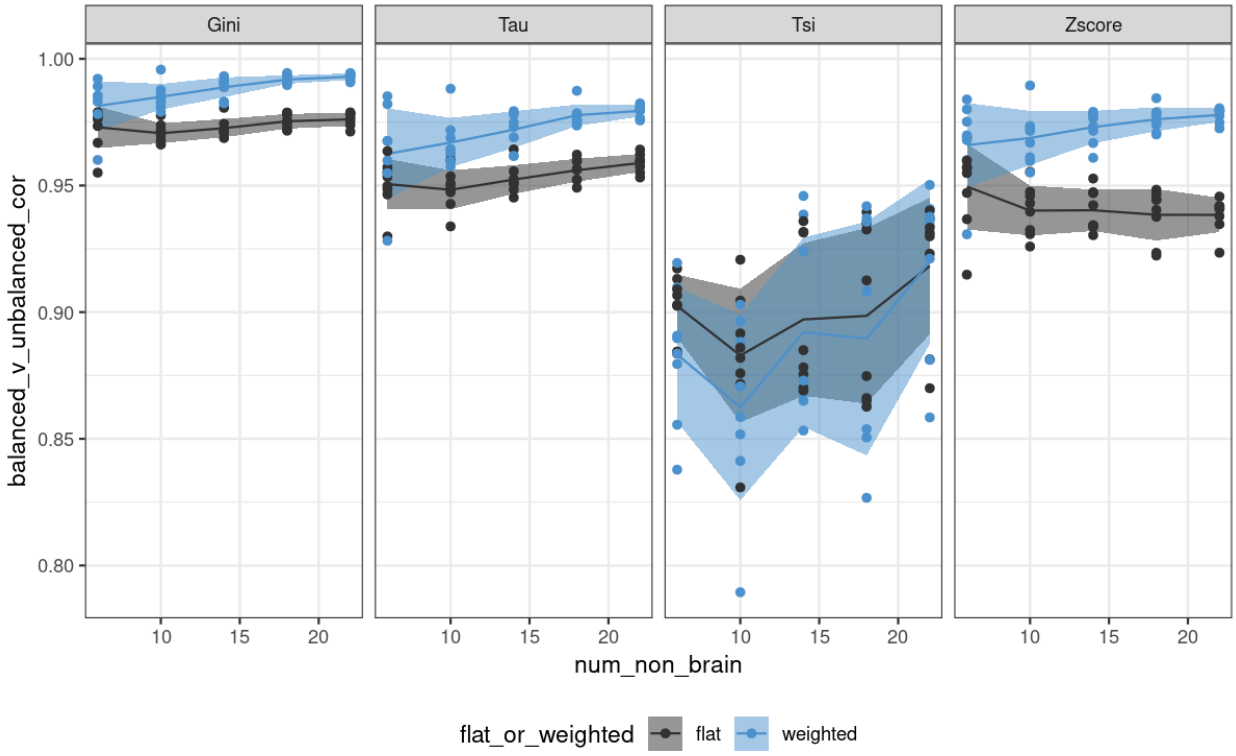
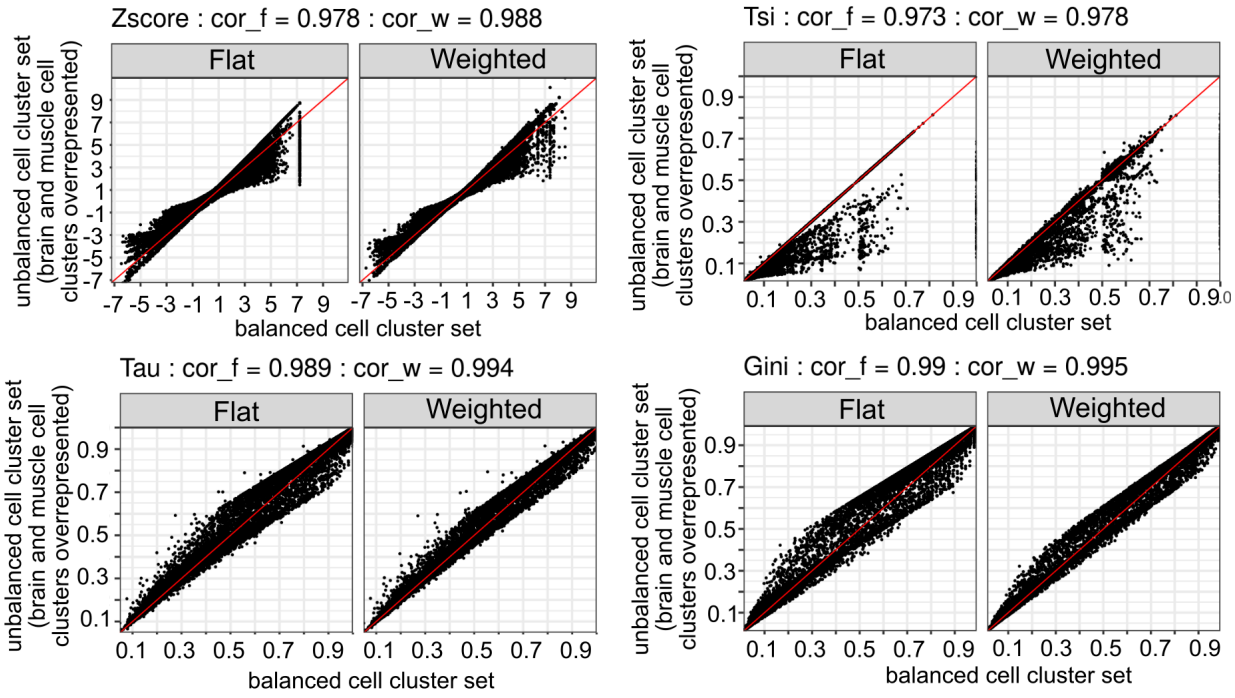


Figure 4-7: Correlation between balanced and unbalanced sample sets as sample set size changes. On the y-axis is the correlation between balanced and unbalanced sample sets from GTEx, where the balanced is a subset of the unbalanced for each replicate. Proportion of sample set composed of brain tissues is held at 50% so change observed is related to sample set size and not the proportion of the sample set that comes from a particular overrepresented context .



Figures 4-8: In the zebrafish single cell dataset incorporating sample similarity information increases correlation between specificity values measured on the balanced and unbalanced sample sets. On the x- and y-axes are specificity values measured on the unbalanced and balanced sample sets respectively for each gene (or gene-cell cluster pair for Z-score). The unbalanced sample set includes 20 clusters from brain cell types, 8 clusters of skeletal muscle cell types and 50 other distinct cell type clusters. The balanced sample set includes 1 brain cell type cluster, 1 skeletal muscle cell type cluster and the same 50 other distinct cell type clusters included in the unbalanced sample set. For each specificity measure, Z-score, Tsi, Tau, and Gini , specificity was calculated without weight (i.e. the flat measure) and with weights (i.e. the weighted measure). cor_f is the pearson correlation between the balanced and unbalanced sample sets for the flat measure, cor_w is the pearson correlation between the balanced and unbalanced sample sets for the weighted measure.

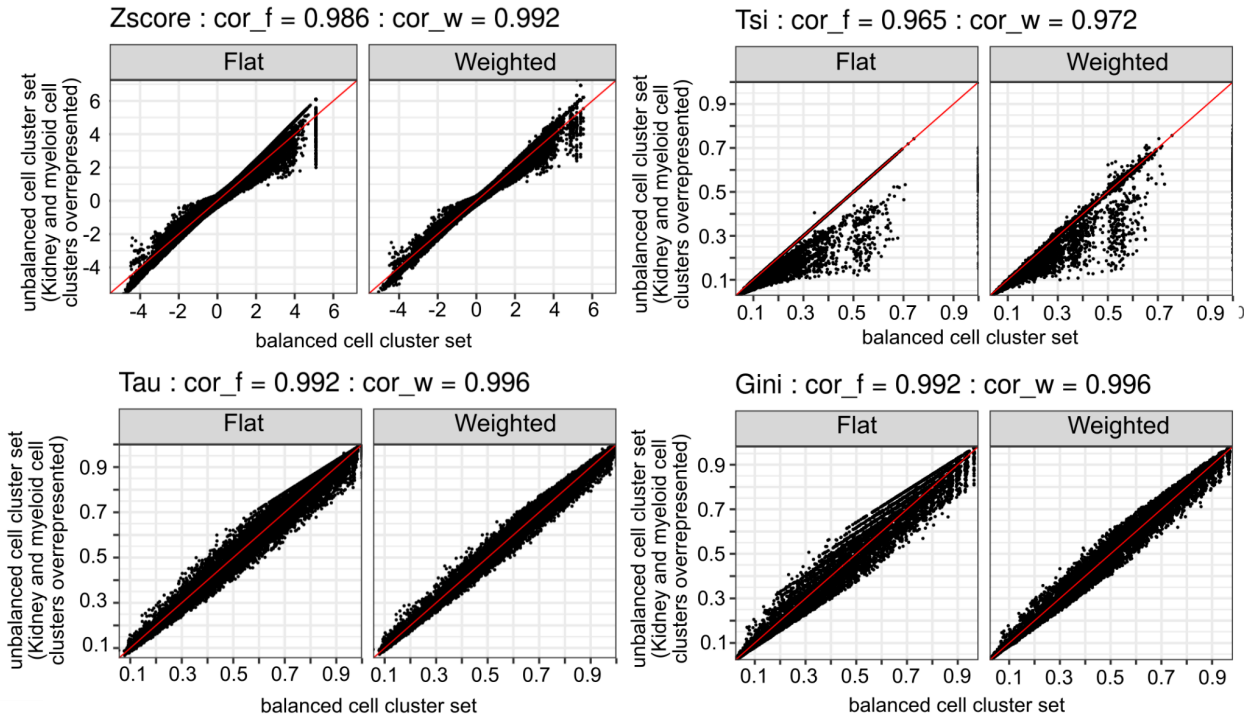


Figure 4-9: In the mouse single cell dataset incorporating sample similarity information increases correlation between specificity values measured on the balanced and unbalanced sample sets. On the x- and y-axes are specificity values measured on the unbalanced and balanced sample sets respectively for each gene (or gene-cell cluster pair for Z-score). The unbalanced sample set includes 11 clusters from kidney cell types, 7 clusters of myeloid blood lineage cell types and 21 other distinct cell type clusters. The balanced sample set includes 1 kidney cell type cluster, 1 myeloid blood lineage cell type cluster and the same 21 other distinct cell type clusters included in the unbalanced sample set. For each specificity measure, Z-score, Tsi, Tau, and Gini , specificity was calculated without weight (i.e. the flat measure) and with weights (i.e. the weighted measure). cor_f is the pearson correlation between the balanced and unbalanced sample sets for the flat measure, cor_w is the pearson correlation between the balanced and unbalanced sample sets for the weighted measure.

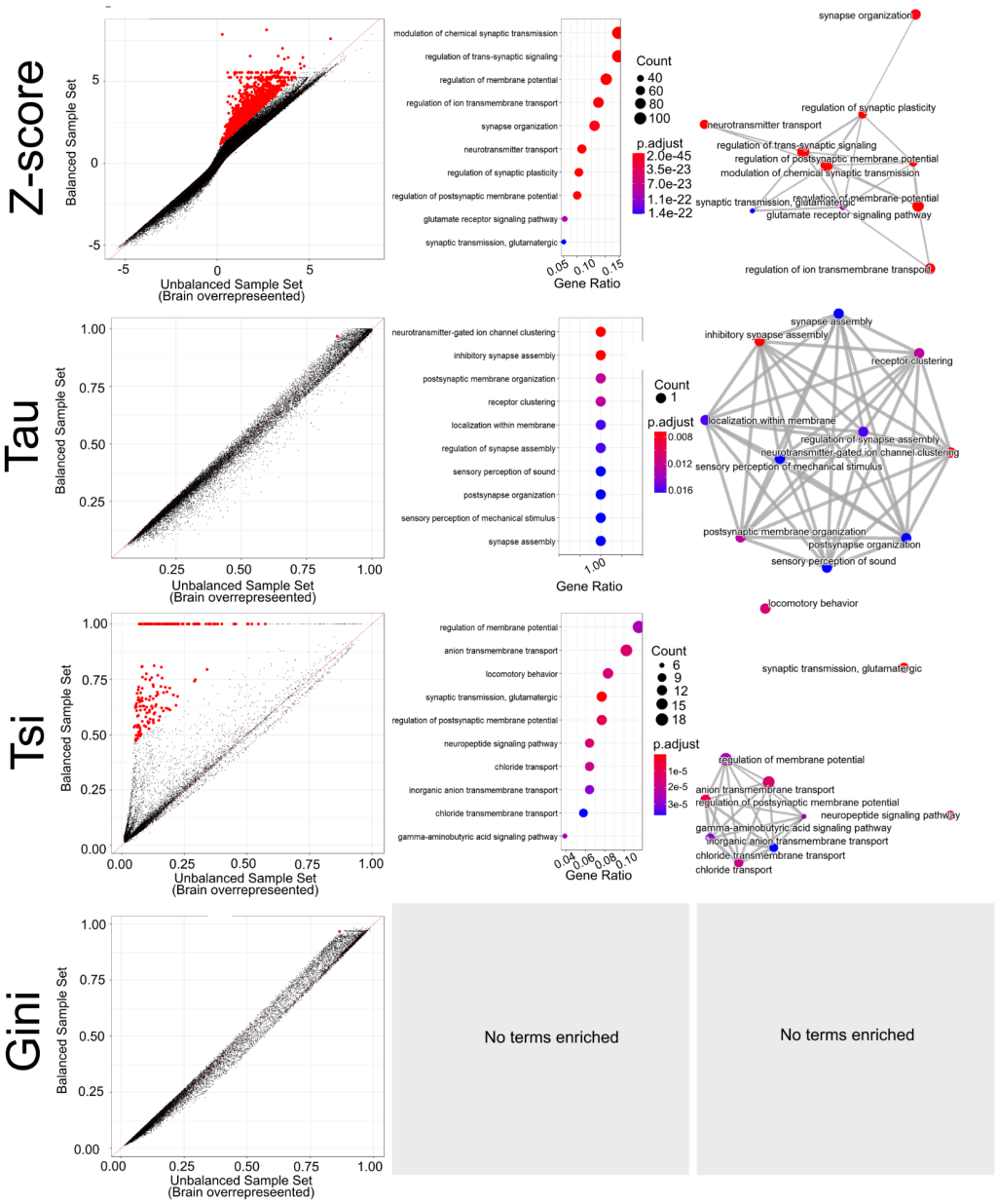


Figure 4-10: From the GTEx dataset, genes with greater specificity measured in balanced than unbalanced sample set for the weighted specificity measures. Left column) On the x- and y-axes are specificity values measured on the unbalanced and balanced sample sets respectively for each gene (or gene-tissue pair for Z-score). The unbalanced sample set includes all non-brain samples and all brain subregion samples while the balanced sample set

includes all non-brain and one brain subregion sample from the GTEx dataset. Highlighted in red are the top 1% of genes with the greatest difference in specificity between balanced and unbalanced sample sets with a difference of at least 1 SD for Z-score and 0.1 for other measures. **Middle column)** Gene Ratio, i.e. the proportion of the gene set with the given GO term, for the top 10 GO terms enriched in the highlighted set from the left column **Right column)** Network plots between the top 10 GO terms from middle column, where the edge width indicates the number of shared genes between a connected pair of terms.

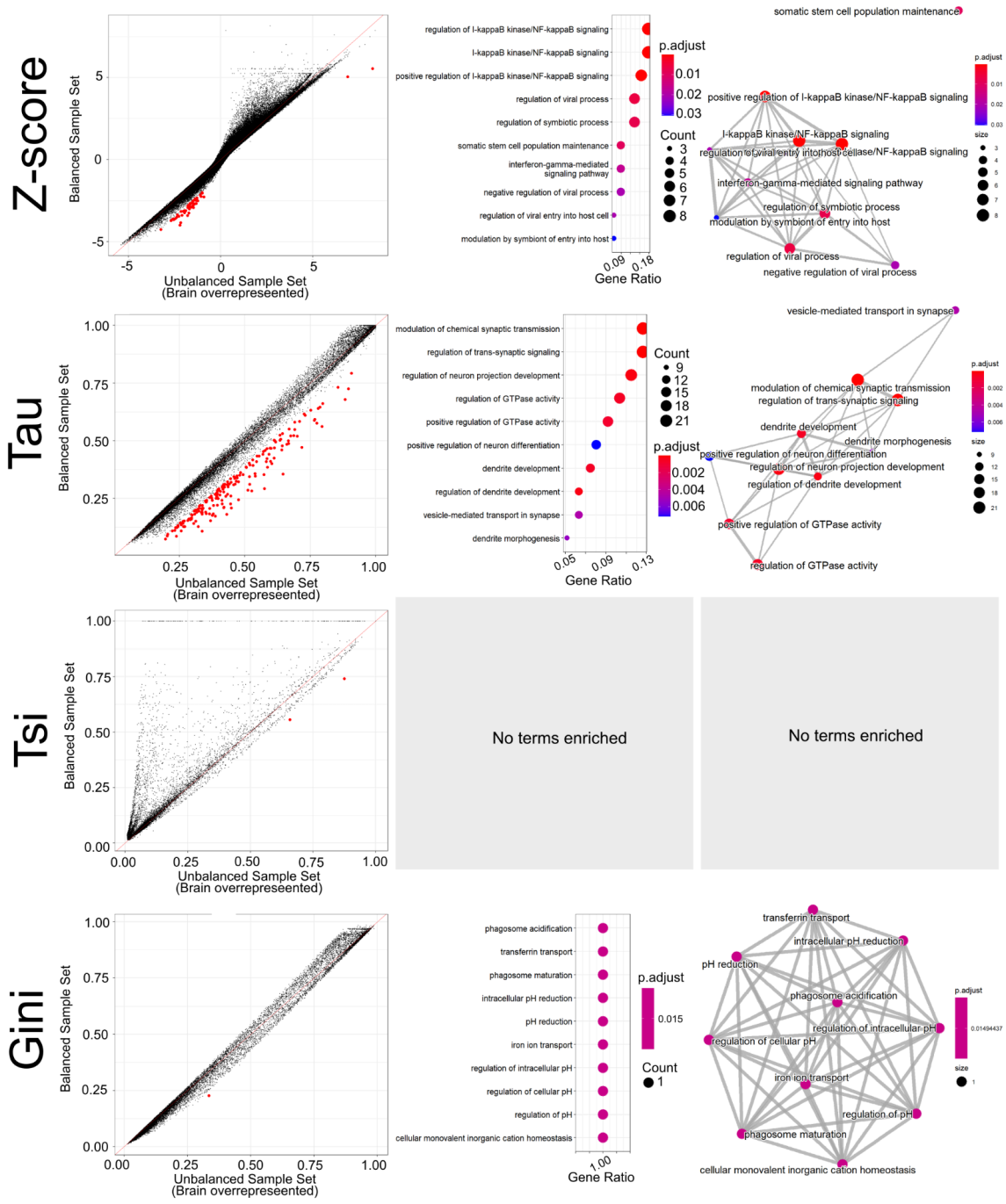


Figure 4-11: From the GTEx dataset, genes with greater specificity measured in unbalanced than balanced sample set for the weighted specificity measures. Left

column) On the x- and y-axes are specificity values measured on the unbalanced and balanced sample sets respectively for each gene (or gene-tissue pair for Z-score). The unbalanced sample set includes all non-brain samples and all brain subregion samples while the balanced sample set includes all non-brain and one brain subregion sample from the GTEx dataset. Highlighted in red are the top 1% of genes with the greatest difference in specificity between balanced and unbalanced sample sets with a difference of at least 1 SD for Z-score and 0.1 for other measures. **Middle column)** Gene Ratio, i.e. the proportion of the gene set with the given GO term, for the top 10 GO terms enriched in the highlighted set from the left column **Right column)** Network plots between the top 10 GO terms from middle column, where the edge width indicates the number of shared genes between a connected pair of terms.

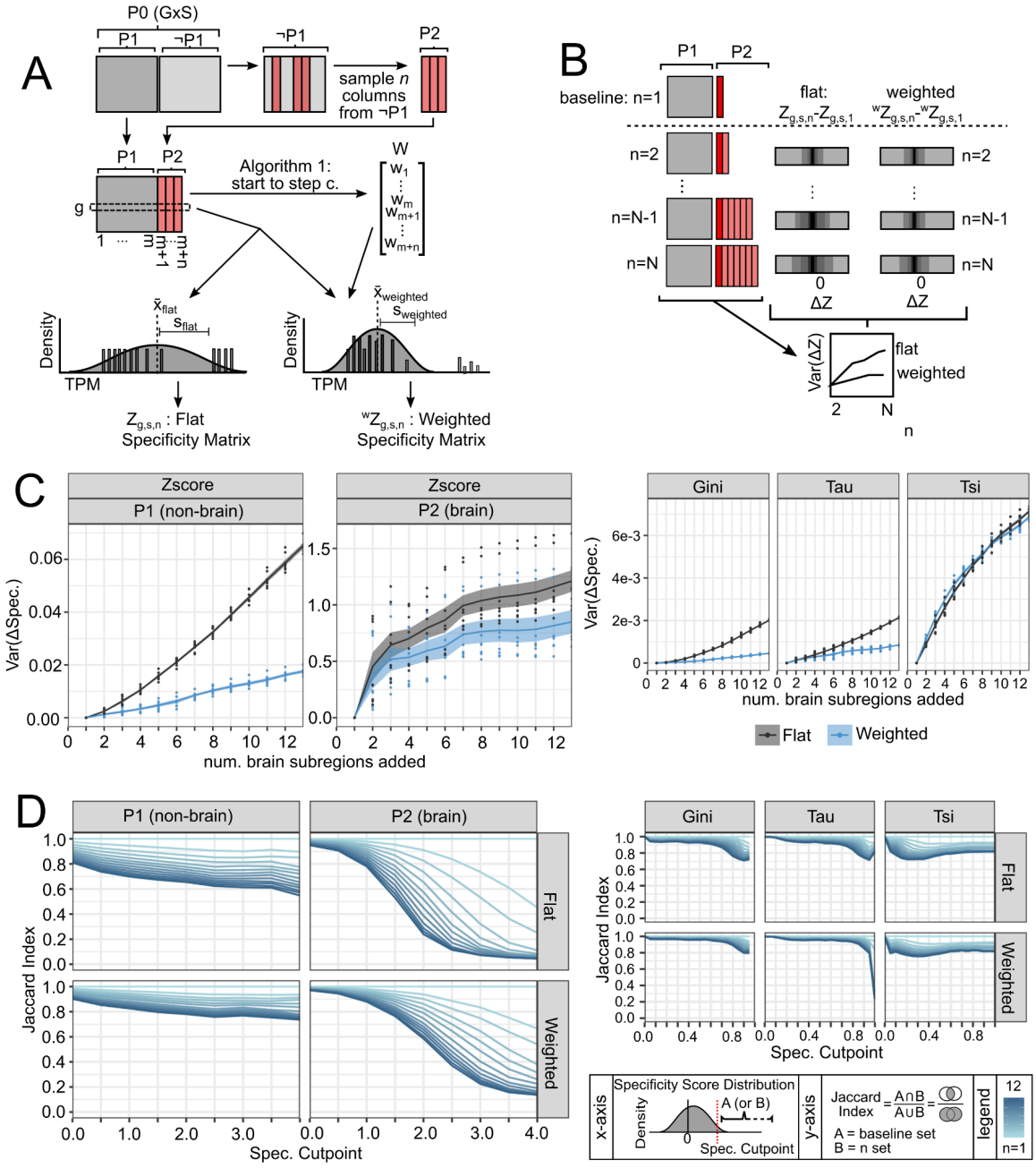


Figure 4-12. Quantification of robustness of specificity measures as sampling depth into brain subregions increases A) Workflow for generating each specificity matrix with the validation sampling procedure. P0 is the full dataset. G is the number of genes and S is the number of samples. P1 is the set of non-brain origin tissue samples. ¬P1 is the set of all brain origin tissue samples. P2 is a random selection of n brain samples. **B)** Z-score is illustrated but

a similar procedure was used for each specificity measure. Each specificity matrix ($Z_{g,s,n}$) where n > 1 is compared to a baseline where n=1, where 1 brain sample was included in P2. Plotted on the x-axis is the density of the change in Z-scores (ΔZ) for all genes in all samples in either P1 or the brain sample initially selected as baseline from P2, with darker color indicating increased density. This data is finally summarized in the change in variance of the ΔZ values as the sample set increases to include more brain samples. Note: for P2, only the change in specificity scores associated with the brain sample selected for the baseline (darker red in figure) is recorded for c. and d. for each of 8 permutations of the procedure where each permutation involves selection of a different brain tissue sample for the baseline and a different ordering of the addition of the remaining brain samples to the sample set **C)** On the y-axis is the variance of change in specificity measure compared to a baseline dataset using 1 brain sample when n additional brain samples are added. The number of additional brain samples in P2 is given on the x-axis. For Z-score, specificity values associated with samples in P1 and P2 are plotted separately since Z-score can be associated with each tissue individually; other specificity measures aggregate across all samples so resolution of specificity between samples in P1 and P2 is not possible for these measures. Points are values from each of the 8 permutations of the procedure, lines are the mean values for each value of n, and the shaded area is the 95% confidence interval **D)** On the x-axis is the cutoff above which a gene is called as specific. On the y-axis is the jaccard index comparing overlap of the set of genes called as specific at the cutoff given on the x-axis relative to the baseline set where P2 includes only 1 randomly selected brain region sample. For Z-score, the jaccard index is the average over all samples in the sample set (P1 or P2), for all other measures which aggregate across all samples the jaccard index is obtained directly. Line color corresponds to the value of n. Note: at n=1 the sample set used is the same as the sample set used in the baseline resulting in the line at jaccard index=1 for n=1 in all cases.

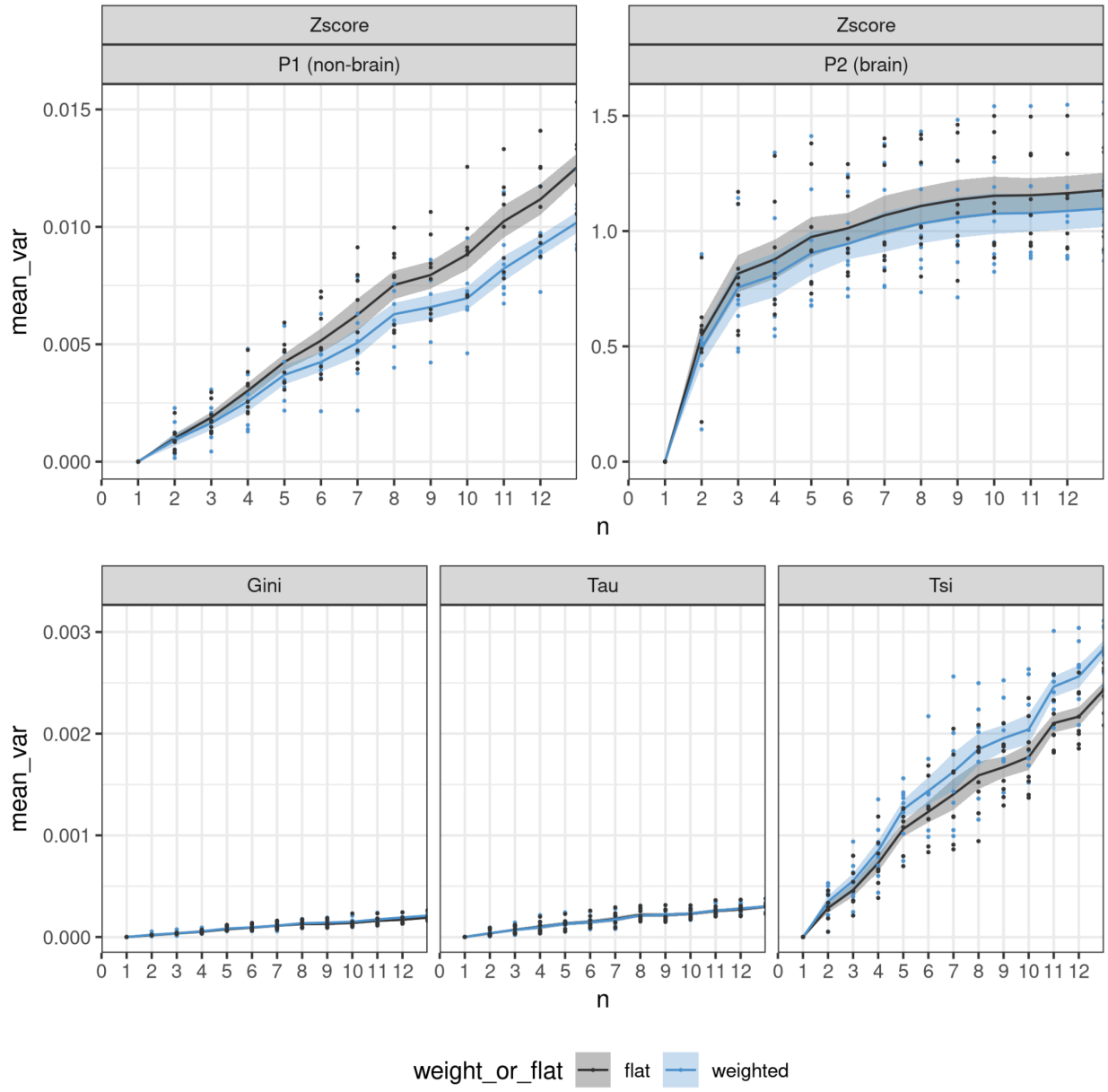


Figure 4-13: **Minor difference between weighted and flat measures in the variance of change in specificity measured as random samples are added to the sample set**. Same procedure as used in Figure 2, except P1 and P2 are random partitions of the same size as in Figure 2 (i.e. samples were not first partitioned into brain and non-brain subsets). For Z-score P1 and P2 are plotted separately since the specificity measure of Z-score can be associated

with each tissue individually; other specificity measures aggregate across all samples so resolution between samples in P1 and P2 is not possible for these measures. Plotted is the variance of the change in specificity measured on the baseline dataset P1 U $P2_{baseline}$, where $P2_{baseline}$ includes only the first random sample, and $P2_n$ with n random samples added successively such that $P2_{n+1}$ includes all the samples in $P2_n$. P1 is held constant in each permutation. Permutation was repeated 8 times to generate individual data points. Lines are drawn between the mean value of all permutations for each value of n. The shaded area is 95% confidence interval around mean estimate.

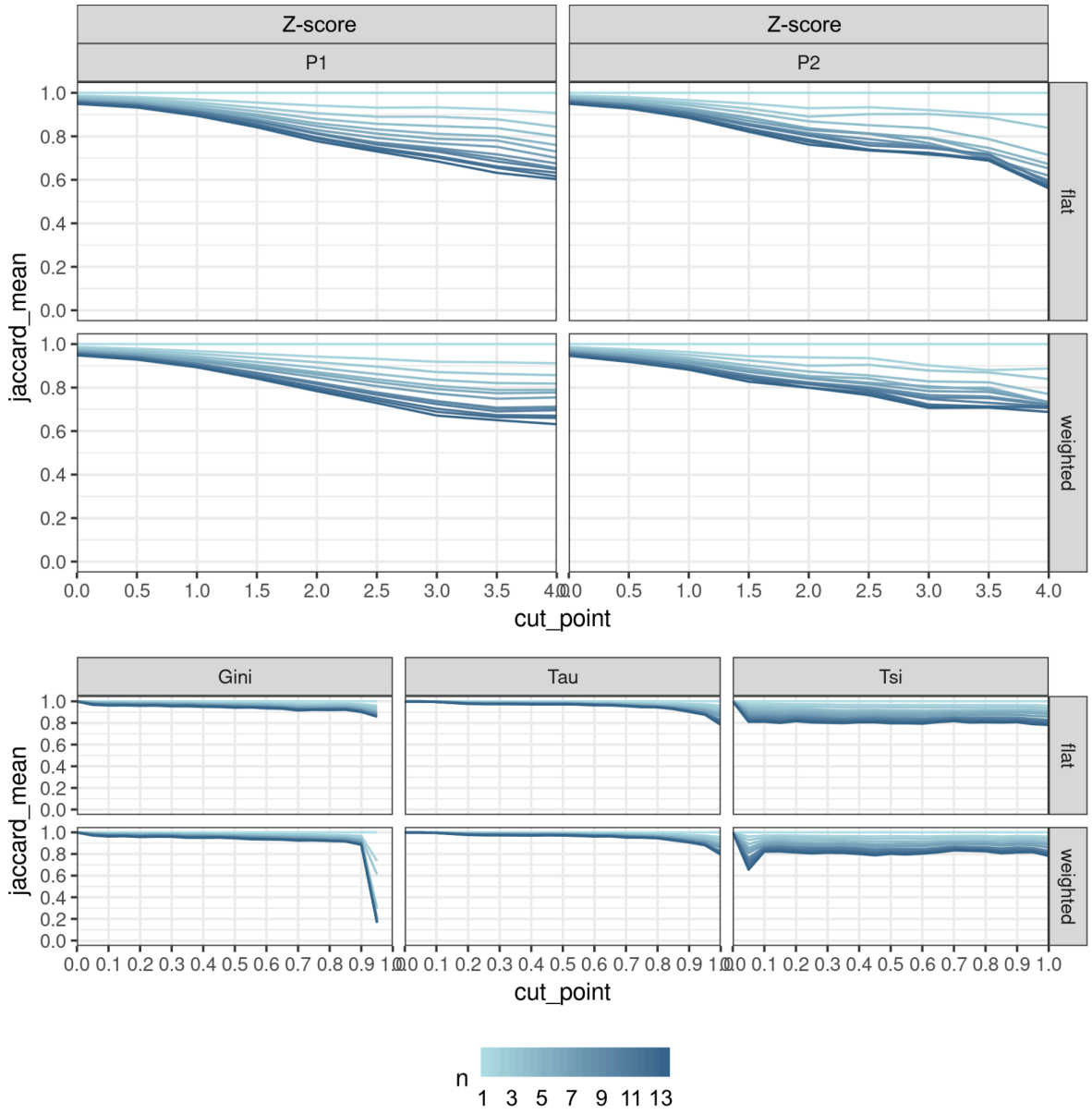


Figure 4-14: **Minor difference between weighted and flat measures in the sets of genes called as specific compared to baseline at variable specificity cutoff values as random samples are added to the sample set as measured by Jaccard index**. Same procedure as used in Figure 2, except P1 and P2 are random partitions of the same size as in Figure 2 (i.e. samples were not first partitioned into brain and non-brain subsets). For Z-score P1 and P2 are

plotted separately since the specificity measure of Z-score can be associated with each tissue individually; other specificity measures aggregate across all samples so resolution between samples in P1 and P2 is not possible for these measures. The baseline is the set of genes called as specific using the sample set P1 U $P2_{baseline}$ to calculate specificity where P2 includes only the first random sample. Plotted is the Jaccard index comparing P1 U $P2_{baseline}$ to P1 U $P2_n$ when $P2_n$ is composed of n random samples added successively such that $P2_{n+1}$ includes all the samples in $P2_n$. For Z-score, where each tissue has its own set of genes considered specific, the Jaccard index is the average over all samples in the sample set (P1 or P2). For all other measures the Jaccard index is obtained directly. The x-axis gives the cutoff above which a gene is considered specific. Line color corresponds to the value of n. Each line is the mean jaccard index from 8 permutations as the cutpoint varies for the given value of n, where each permutation is a distinct sampling of tissues for P1 and P2.

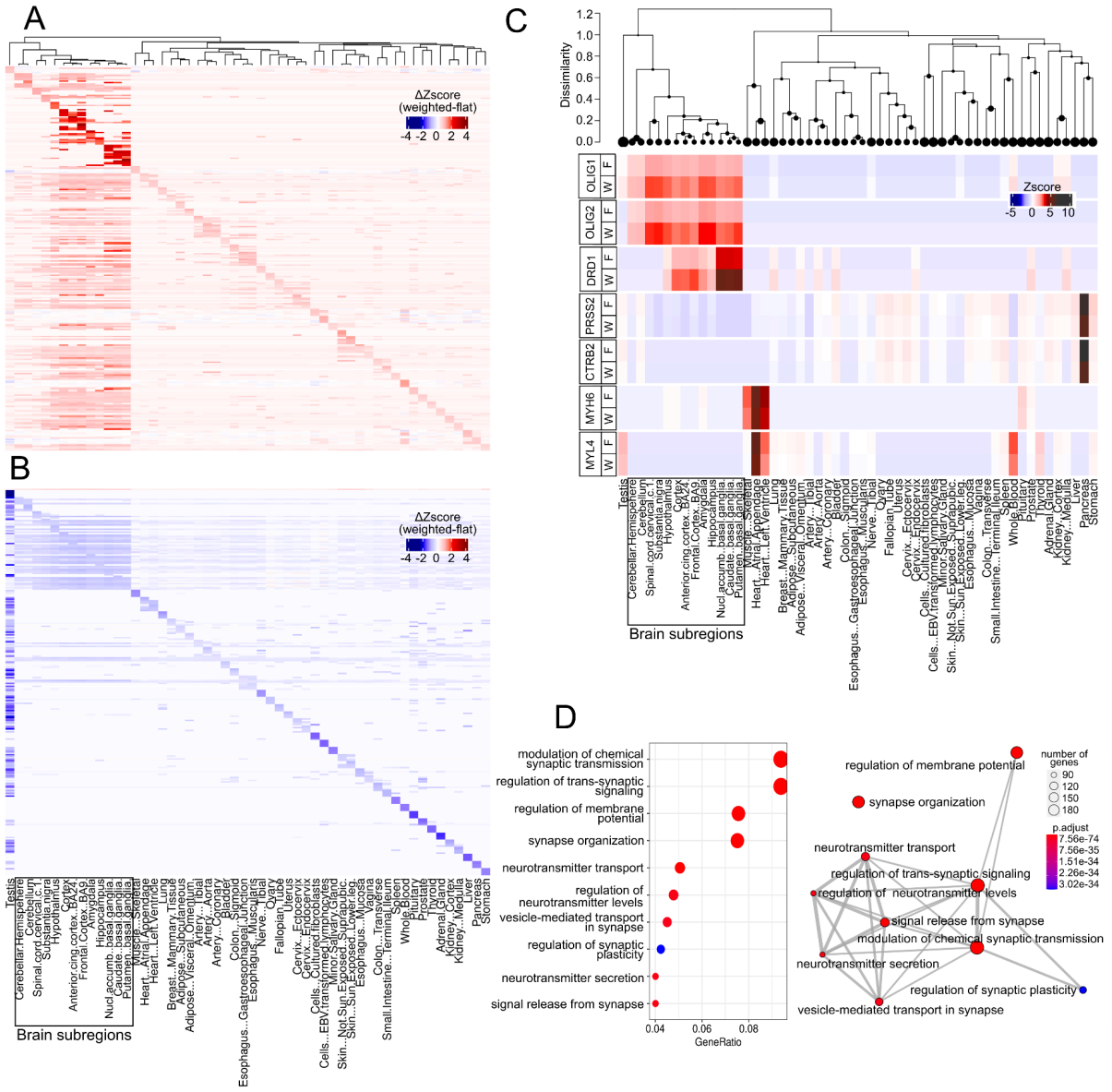


Figure 4-15. Biological context of differences between flat and weighted measures of specificity in GTEx dataset. A) Top 10 genes with greatest difference between weighted and flat Z-score for each tissue. Rows correspond to individual genes, columns to tissues. Note: diagonal produced by having top 10 genes from leftmost tissue on x-axis as first 10 rows, top 10 from next leftmost tissue as next 10 rows, etc. **B)** Bottom 10 genes with greatest difference between weighted and flat Z-score for each tissue **C)** genes known to be specific to brain

regions, *OLIG1* and *OLIG2* , pancreas, *PRSS1* and *CTRB2*, and heart *MYH6* and *MYL4*. For each gene, top row is flat (F) Z-score, bottom row is weighted (W) Z-score. Dendrogram at top shows the dissimilarity tree used to generate sample weights which are shown as the area of the leaf nodes of the dendrogram. **D)** Gene ontology results highlighting top 10 terms in the set of genes that have specificity values < 2 standard deviations by the flat Z-score, and >2 standard deviations by the weighted Z-score. On the left is the Gene Ratio, i.e. the proportion of the gene set with the given GO term. On the right is the network plot, where the edge width indicates the number of shared genes between a connected pair of terms.

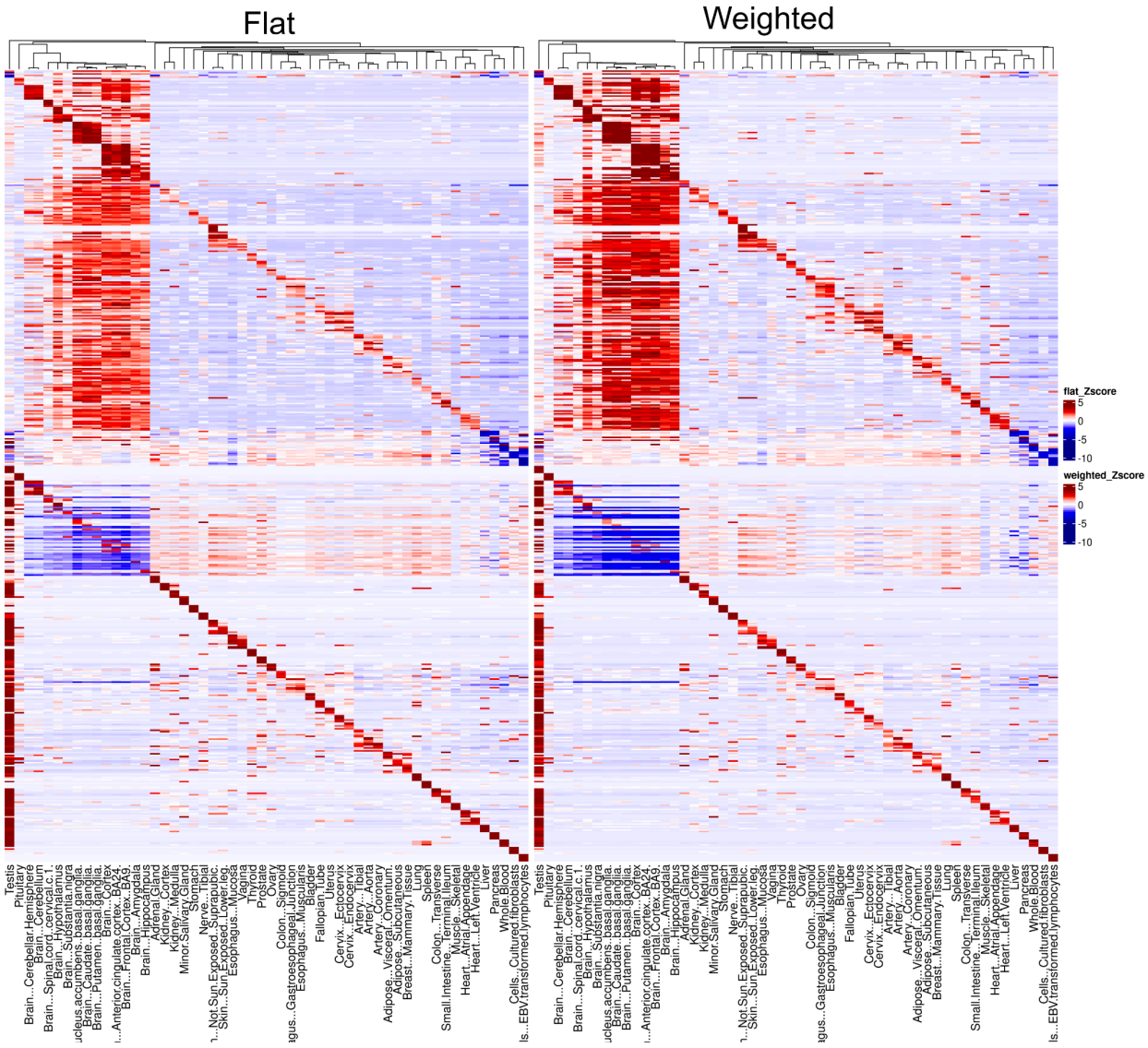


Figure 4-16: **GTEx dataset flat and weighted Z-score for top 10 genes with greatest positive (top) and negative (bottom) difference between weighted and flat Z-score for each tissue**. Genes on rows, tissue samples on columns. Gene and sample order is the same as in Figure 2A/B. Note: diagonal produced by having top 10 genes from leftmost tissue on x-axis as first 10 rows, top 10 from next leftmost tissue as next 10 rows, etc.

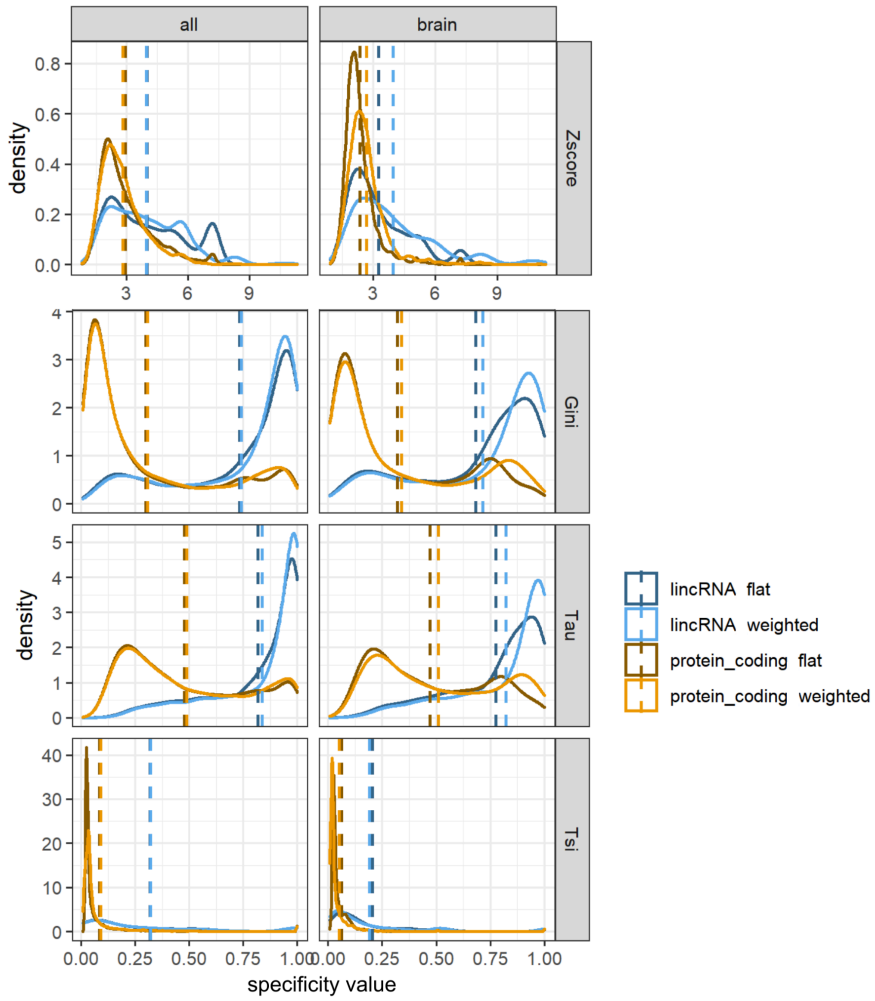


Figure 4-17: In the GTEx dataset, distributions of specificity values for lncRNA and protein coding genes. Left column shows distribution of specificity values for all genes, excluding those with their highest expression in the testis. Exclusion of testis was done as this tissue expresses a disproportionate number of non-protein coding genes relative to all other tissues and so inclusion of these genes specific to testis would mask the global trend of specificity for all other tissues. Right column shows distribution of specificity values for genes where the tissue with the maximum expression value was one of the brain subregions. For Zscore, specificity value is the maximum specificity value across all tissues for each gene. Dotted lines show the center of mass, i.e. the mean, for each distribution.

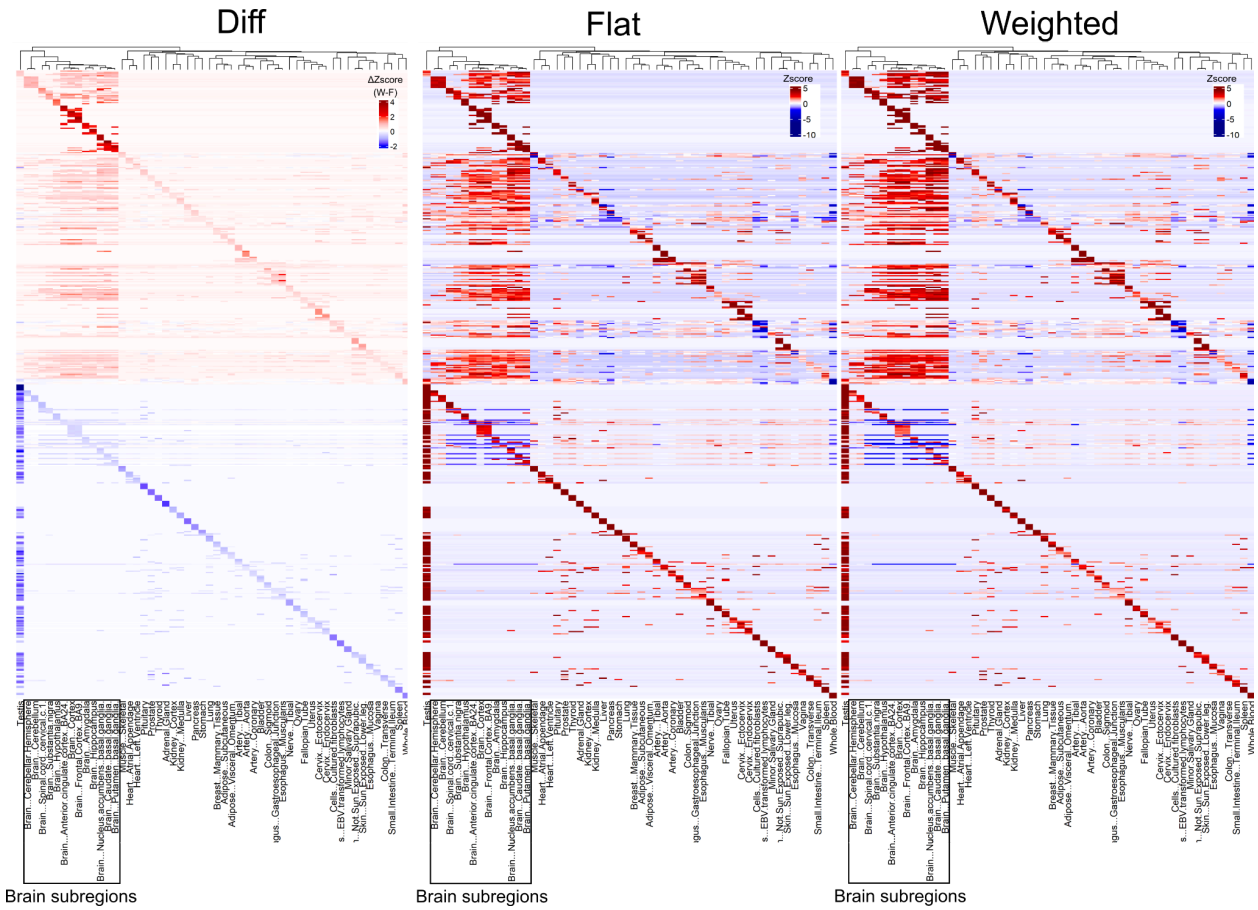


Figure 4-18: In the GTEx dataset, the difference between weighted and flat, flat, and weighted Z-score for top 5 lncRNA genes with greatest positive (top) and negative (bottom) difference between weighted and flat Z-score for each tissue. Genes on rows, tissue samples on columns. Gene and sample order is the same for left, middle and right plots . Note: diagonal produced by having top 5 genes from leftmost tissue on x-axis as first 5 rows, top 5 from next leftmost tissue as next 5 rows, etc.

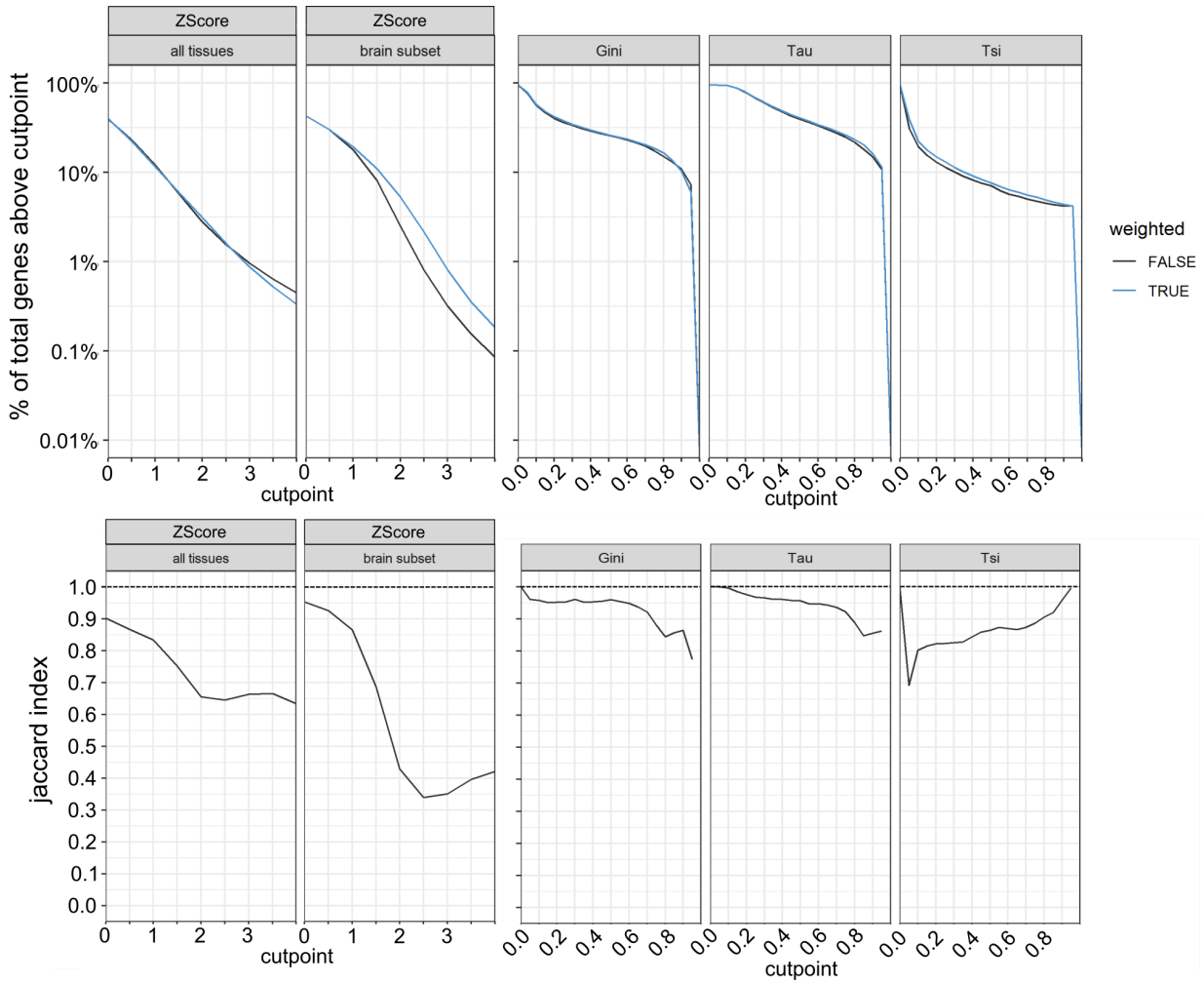


Figure 4-19: Quantitative summary of the difference in genes called as specific for the weighted and flat measures of specificity in the GTEx dataset. The proportion of genes called as specific as the cutpoint for specificity varies for each specificity measure (top) and the jaccard index between the sets of genes called as specific at each cutpoint for the weighted and flat measures (bottom).

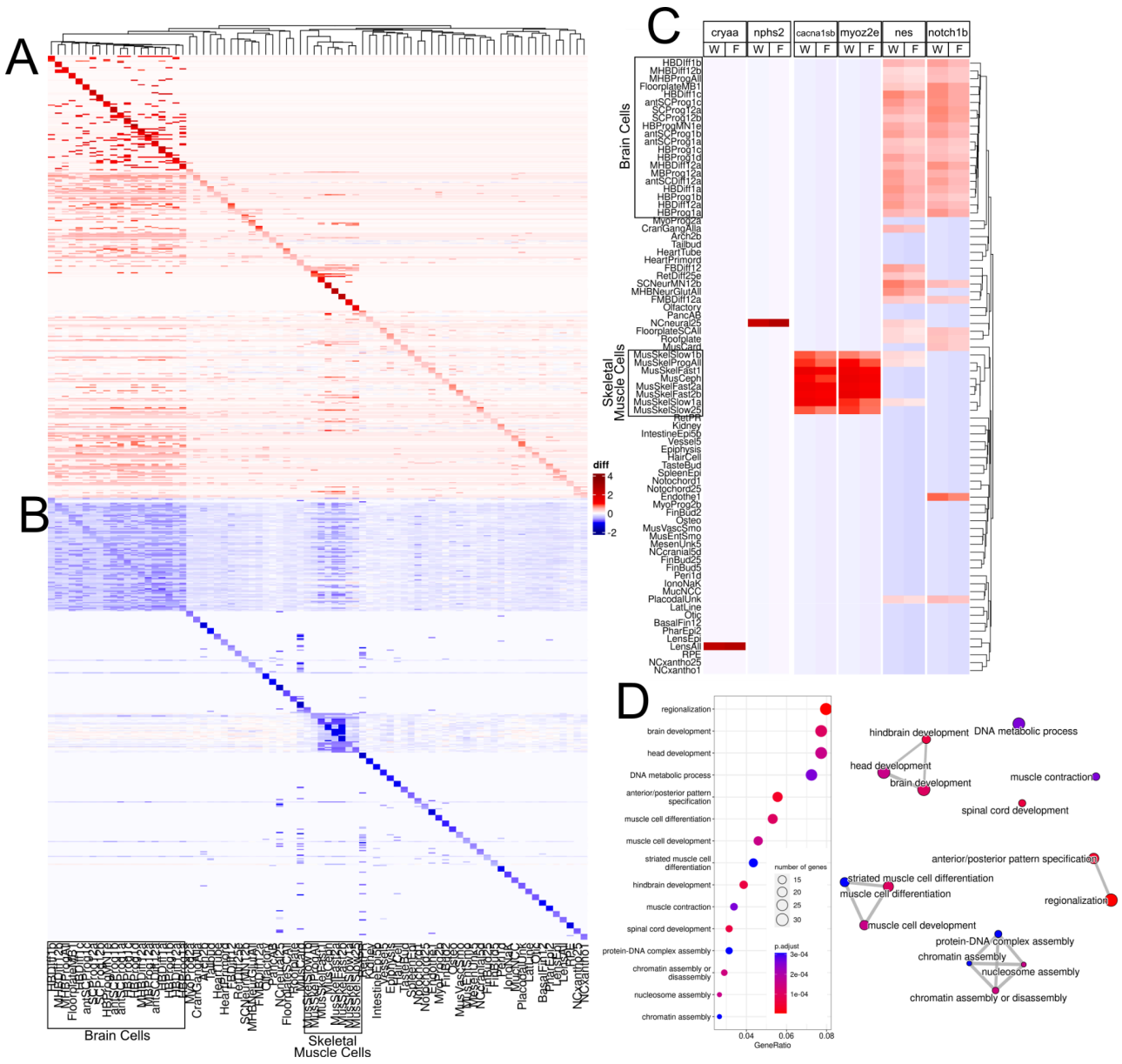


Figure 4-20: Biological context of differences between flat and weighted measures of specificity in zebrafish single cell dataset. A) Top 10 genes with greatest difference between weighted and flat Z-score for each tissue. Rows correspond to individual genes, columns to tissues. Note: diagonal produced by having top 10 genes from leftmost cell type cluster on x-axis as first 10 rows, top 10 from next leftmost cell cluster type as next 10 rows, etc. **B)** Bottom 10 genes with greatest difference between weighted and flat Z-score for each cell type cluster

C) genes known to be specific lens cells, *cryaa*, to kidney cells, *nph2,* to skeletal muscle, *cacna1sb* and *myoz2e*, and to brain *nes* and *notch1b*. For each gene, top row is flat (F) Z-score, bottom row is weighted (W) Z-score. Dendrogram at right shows the dissimilarity tree used to generate sample weights **D)** Gene ontology results highlighting top 15 terms in the set of genes that have specificity values < 2 standard deviations by the flat Z-score, and >2 standard deviations by the weighted Z-score. On the left is the Gene Ratio, i.e. the proportion of the gene set with the given GO term. On the right is the network plot, where the edge width indicates the number of shared genes between a connected pair of terms.

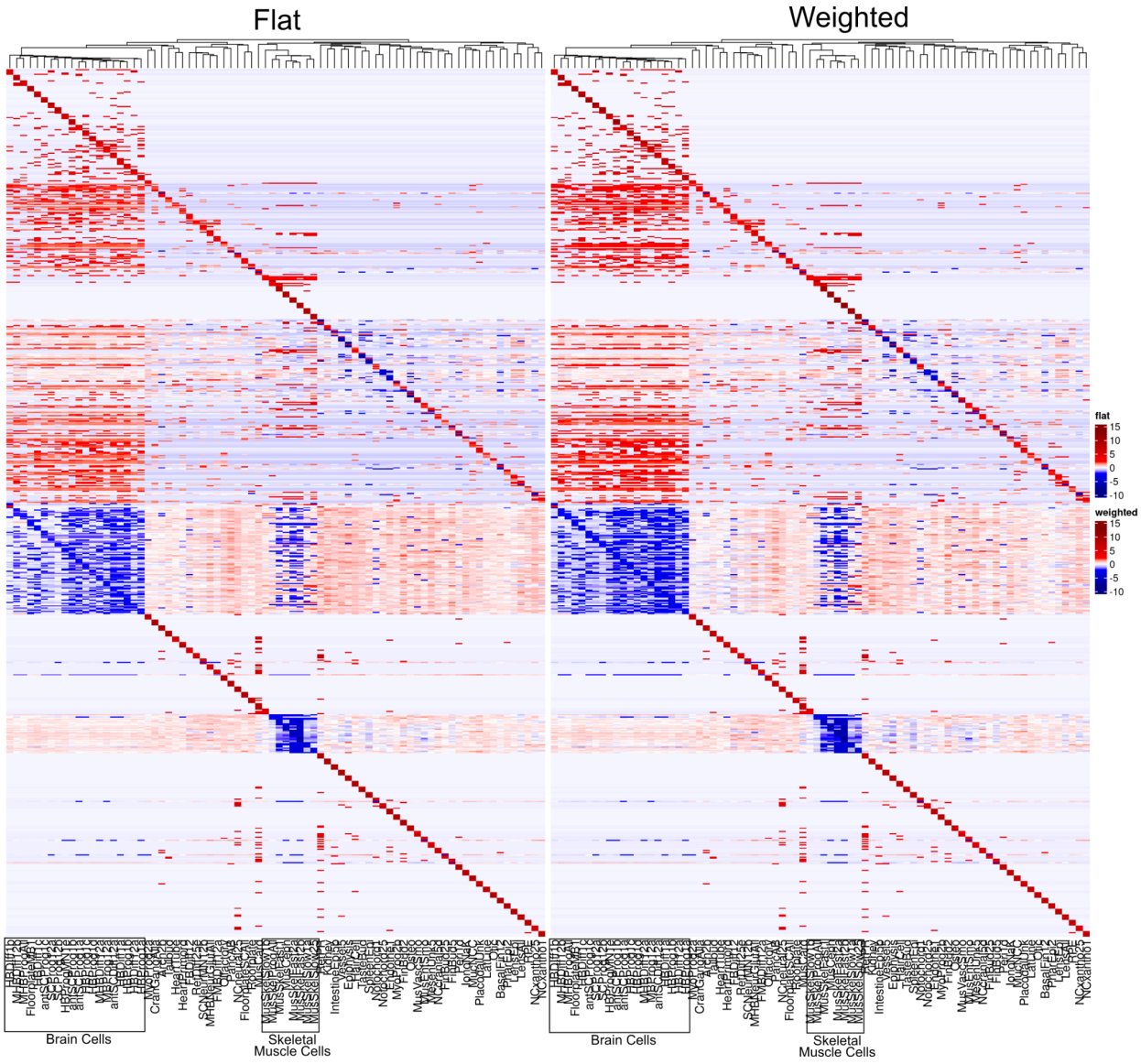


Figure 4-21: Zebrafish single cell dataset flat and weighted Z-score for top 10 genes with greatest positive (top) and negative (bottom) difference between weighted and flat Z-score for each cell type cluster. Genes on rows, cell type clusters on columns. Gene and sample order is the same as in Supplemental Figure S17A/B. Note: diagonal produced by having top 10 genes from leftmost cell type cluster on x-axis as first 10 rows, top 10 from next leftmost cell type cluster as next 10 rows, etc.

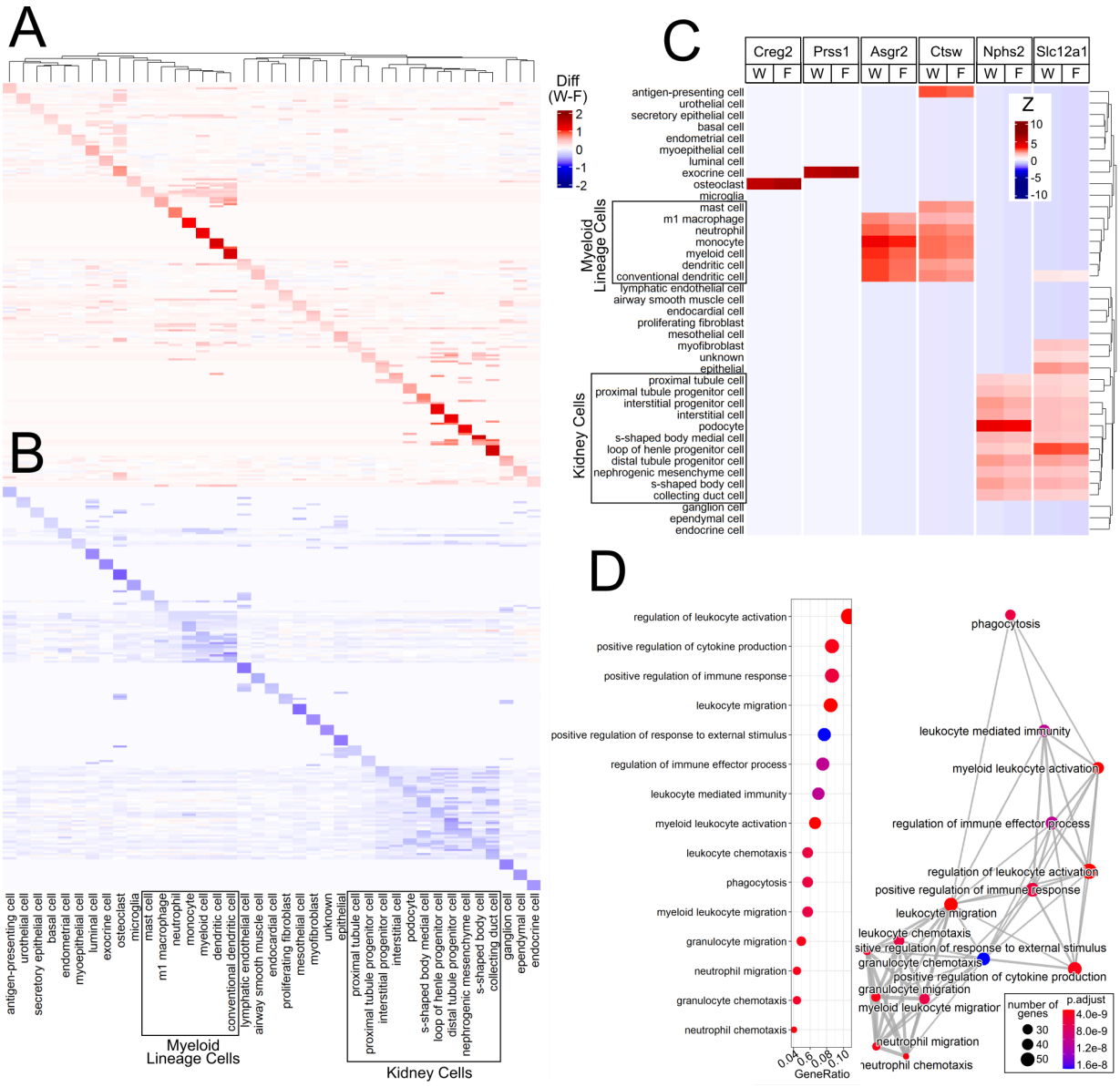


Figure 4-22: Biological context of differences between flat and weighted measures of specificity in mouse single cell dataset. A) Top 10 genes with greatest difference between weighted and flat Z-score for each tissue. Rows correspond to individual genes, columns to tissues. Note: diagonal produced by having top 10 genes from leftmost cell type cluster on x-axis as first 10 rows, top 10 from next leftmost cell cluster type as next 10 rows, etc. **B)** Bottom

10 genes with greatest difference between weighted and flat Z-score for each cell type cluster **C)** genes known to be specific osteoclast cells, *Creg2*, pancreas, *Prss1,* to kidney, *Slc12a1* and *Nphs2*, and myeloid blood lineages. For each gene, top row is flat (F) Z-score, bottom row is weighted (W) Z-score. Dendrogram at right shows the dissimilarity tree used to generate sample weights **D)** Gene ontology results highlighting top 15 terms in the set of genes that have specificity values < 2 standard deviations by the flat Z-score, and >2 standard deviations by the weighted Z-score. On the left is the Gene Ratio, i.e. the proportion of the gene set with the given GO term. On the right is the network plot, where the edge width indicates the number of shared genes between a connected pair of terms.

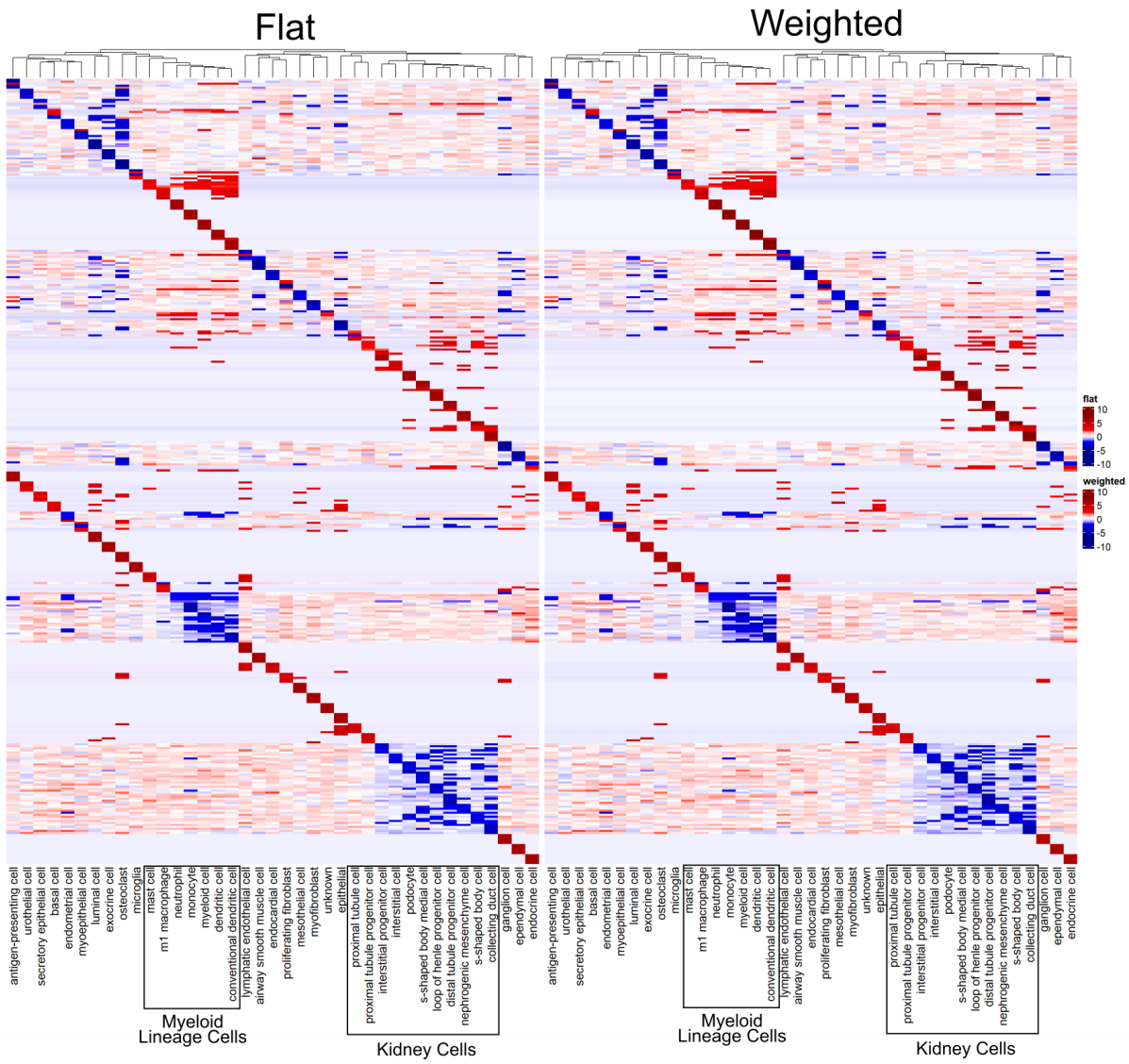


FIgure 4-23: Mouse single cell dataset flat and weighted Z-score for top 10 genes with greatest positive (top) and negative (bottom) difference between weighted and flat Z-score for each cell type cluster. Genes on rows, cell type clusters on columns. Gene and sample order is the same as in Supplemental Figure S19A/B. Note: diagonal produced by having top 10 genes from leftmost cell type cluster on x-axis as first 10 rows, top 10 from next leftmost cell type cluster as next 10 rows, etc.

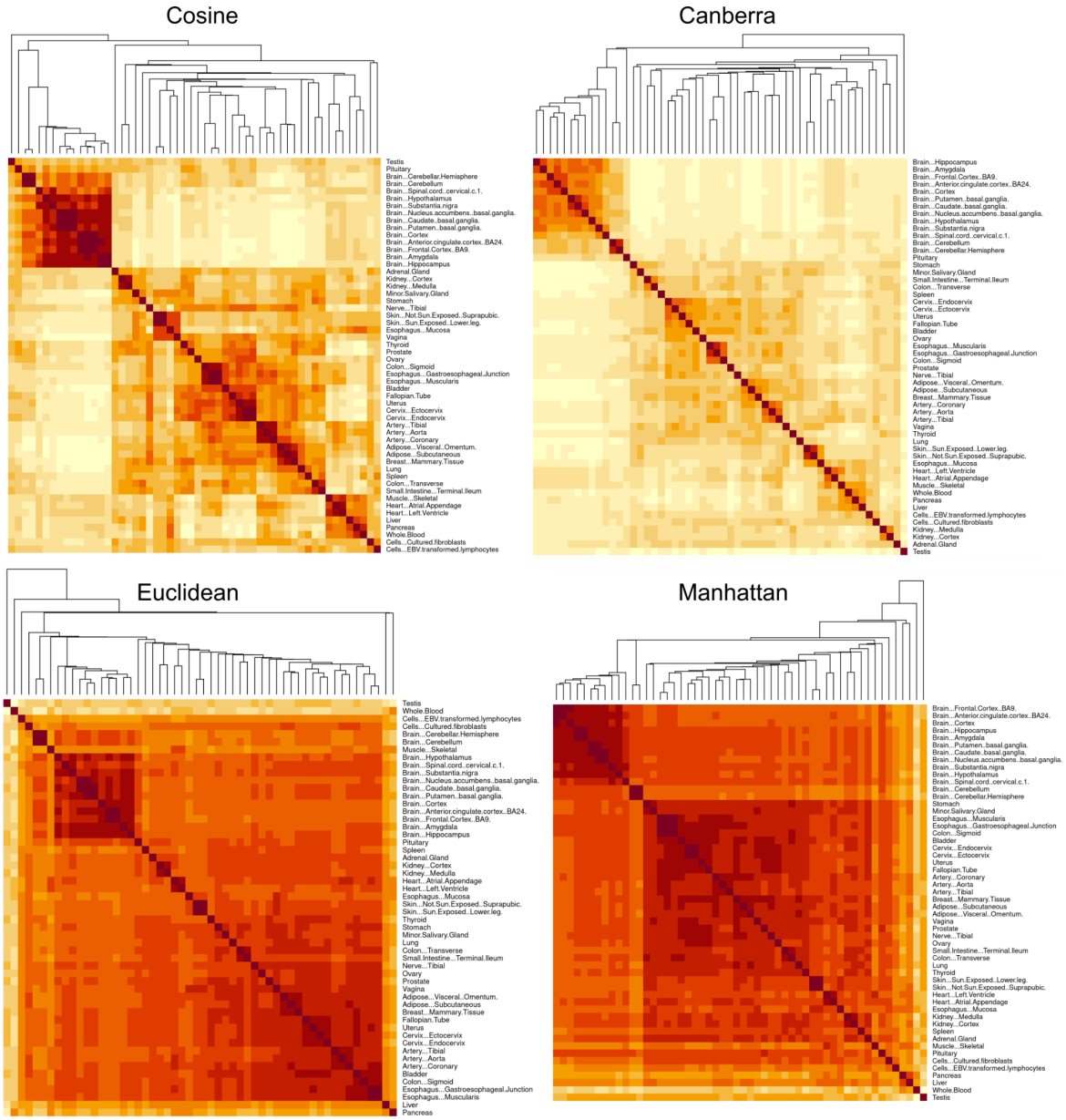


Figure 4-24: **Comparison of similarity measures for tissue-tissue similarity.** Comparison of cosine similarity, canberra distance, euclidean distance, and manhattan distance as similarity measures. All measures were normalized to the domain of [0-1] so that min=0 and max=1 for each measure, distances were converted to similarity analogs by subtracting the normalized value from 1.

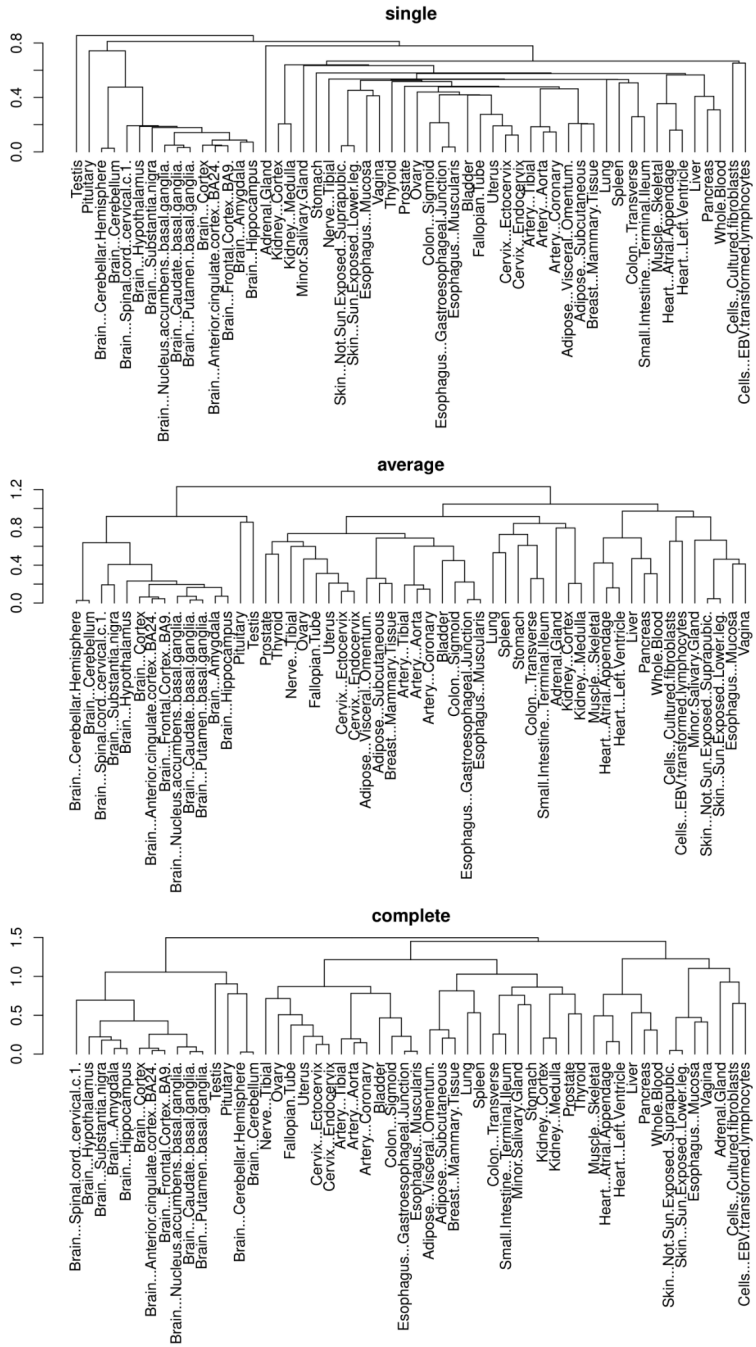


Figure 4-25: **Comparison of clustering methods.** Comparison of single-linkage, average-linkage, and complete-linkage clustering methods computed on the cosine similarity matrix.

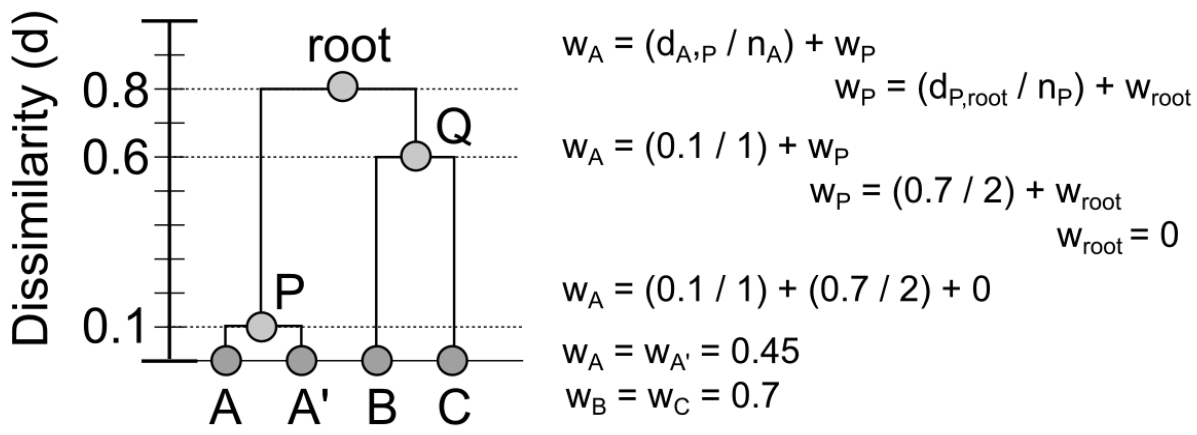


Figure 4-26: Toy example demonstrating behavior of Eq. 1, the recursive function used for determination of weights from a dissimilarity tree. Left shows the dissimilarity tree with similarity between tissues *A, A', B*, and *C*. Parent nodes *P* and *Q* are generated by clustering. Right top down: step-through of the calculations for determination of weight for tissue *A*.

References

1. Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* **81**, 425–455 (2006).

2. Lima Cunha, D., Arno, G., Corton, M. & Moosajee, M. The Spectrum of PAX6 Mutations and Genotype-Phenotype Correlations in the Eye. *Genes* **10**, (2019).

3. Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20870–20875 (2008).

4. Hekselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat. Rev. Genet.* **21**, 137–150 (2020).

5. Barshir, R., Shwartz, O., Smoly, I. Y. & Yeger-Lotem, E. Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS Comput. Biol.* **10**, e1003632 (2014).

6. Saitou, M. *et al.* Functional Specialization of Human Salivary Glands and Origins of Proteins Intrinsic to Human Saliva. *Cell Rep.* **33**, 108402 (2020).

7. Genuth, N. R. & Barna, M. Heterogeneity and specialized functions of translation machinery: from genes to organisms. *Nat. Rev. Genet.* **19**, 431–452 (2018).

8. Herrmann, H. Mechanisms of cell specialization. *Invest. Ophthalmol.* **8**, 17–25 (1969).

9. Arboleda, V. A. *et al.* Regulation of sex determination in mice by a non-coding genomic region. *Genetics* **197**, 885–897 (2014).

10. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).

11. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).

12. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

13. Vandenbon, A. & Nakai, K. Modeling tissue-specific structural patterns in human and mouse promoters. *Nucleic Acids Res.* **38**, 17–25 (2010).

14. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

15. Julien, P. *et al.* Mechanisms and evolutionary patterns of mammalian and avian dosage

compensation. *PLoS Biol.* **10**, e1001328 (2012).

16. Gini, C. *Variabilità E Mutabilità*. (ui.adsabs.harvard.edu, 1912).

17. Ceriani, L. & Verme, P. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J. Econ. Inequality* **10**, 421–443 (2012).

18. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

19. Farnsworth, D. R., Saunders, L. M. & Miller, A. C. A single-cell transcriptome atlas for zebrafish development. *Dev. Biol.* **459**, 100–108 (2020).

20. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091–1107.e17 (2018).

21. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).

22. Gloss, B. S. & Dinger, M. E. The specificity of long noncoding RNA expression. *Biochim. Biophys. Acta* **1859**, 16–22 (2016).

23. Miller, R. H. Regulation of oligodendrocyte development in the vertebrate CNS. *Prog. Neurobiol.* **67**, 451–467 (2002).

24. Valério-Gomes, B., Guimarães, D. M., Szczupak, D. & Lent, R. The Absolute Number of Oligodendrocytes in the Adult Mouse Brain. *Front. Neuroanat.* **12**, 90 (2018).

25. Cadet, J. L., Jayanthi, S., McCoy, M. T., Beauvais, G. & Cai, N. S. Dopamine D1 receptors, regulation of gene expression in the brain, and neurodegeneration. *CNS Neurol. Disord. Drug Targets* **9**, 526–538 (2010).

26. Assis, R. & Bachtrog, D. Neofunctionalization of young duplicate genes in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 17409–17414 (2013).

27. Piasecka, B., Robinson-Rechavi, M. & Bergmann, S. Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. *Bioinformatics* **28**, 1865–1872 (2012).

28. Martínez, O. & Reyes-Valdés, M. H. Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9709–9714 (2008).

29. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).

30. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.* **7**, 1141 (2018).

31. National Advisory Council for Human Genome Research (NACHGR). Concept Clearance for FOAs Developmental Genotype-Tissue Expression (dGTEx). https://www.genome.gov/sites/default/files/media/files/2020-02/Concept_Document_Developmental_GTEx.pdf (2020).

32. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

33. Leader, D. P., Krause, S. A., Pandit, A., Davies, S. A. & Dow, J. A. T. FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. *Nucleic Acids Res.* **46**, D809–D815 (2018).

34. Smith, C. M. *et al.* The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids Res.* **47**, D774–D779 (2019).

35. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).

36. Shirkhorshidi, A. S., Aghabozorgi, S. & Wah, T. Y. A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *PLoS One* **10**, e0144059 (2015).

37. Deshpande, R., Vandersluis, B. & Myers, C. L. Comparison of profile similarity measures for genetic interaction networks. *PLoS One* **8**, e68664 (2013).

38. Ferreira, L. & Hitchcock, D. B. A comparison of hierarchical methods for clustering functional data. *Commun. Stat. Simul. Comput.* **38**, 1925–1949 (2009).

39. Price, G. R. Extension of covariance selection mathematics. *Ann. Hum. Genet.* **35**, 485–490 (1972).

40. Lerman, R. I. & Yitzhaki, S. Improving the accuracy of estimates of Gini coefficients. *J. Econom.* **42**, 43–47 (1989).

41. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

42. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

43. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

CHAPTER 5

Conclusions

Recap of motivations for works discussed

As we introduced in **Chapter 1**, chromatin modifiers are critical to a variety of normal developmental processes [1–5] and mutations in the genes encoding these proteins cause a wide range of rare disorders [6–10]. The etiologies of these disorders are just now beginning to be worked out, but many factors exist that complicate our ability to define the functional species of modifiers that function in normal development and that are affected in disease. Particularly, challenges exist in identifying the precise composition of multiprotein chromatin modifier complexes with respect to their member proteins and isoforms [11,12] , and the post-translational modifications that may further modulate their function [13]. On top of the complexity in form of the chromatin modifiers are many unresolved questions around their functional output: what is the full set of modifications they catalyze, what modifiers are responsible for particular orphan histone marks [14,15], what are the precise genomic targets of modifier complexes, and how do any or all of these things vary between biological contexts. We discussed in **Chapter 2** key assays and modes of analysis that are being used to investigate these questions and highlighted fundamental gaps in existing methods. Specifically, we noted the inability of existing methods to robustly characterize the scale of epigenetic changes that occur in development and disease, and the inability of existing methods to precisely and robustly describe the context over which functional genomic changes occur. We addressed these gaps in the novel methods of analysis we developed in **Chapter 3** and **Chapter 4**.

Summary of key results

In **Chapter 3** we developed DMRscaler[16], a novel method for identifying regions of differential methylation across a wide range of genomic scale, from small basepair level features to those spanning whole chromosomes. While focused on DNA methylation data, generalization of this method provides a novel means of analyzing chromatin features to understand how they

are organized across genomic scales. Emergent properties in biology occur in all domains of study, and the ability to study how features interact across scale will be critical to generating a comprehensive understanding of how the genome is organized in development and the ways that alterations to this organization contribute to various pathologies. Beyond the usage of our method in applications analyzing function genomic elements that map to the genome, one can imaging further generalizations that encompass features in time or, with some imagination and some extra dimensions, features in spatial or other more abstract coordinate systems.

In **Chapter 4** we developed a novel method for measuring the specificity of gene expression when provided a diverse sample of tissues and cells as a reference set[17]. The rapid advances in sequencing and single cell technologies have made acquisition of vast datasets covering developmentally and anatomically diverse contexts possible [18,19]. While the quantity and richness of data available to researchers has exploded, methods for contextualizing it and gleaning biological insights have not kept pace. In particular, existing methods for describing functional genomics features such as the transcriptome and chromatin states have not taken into account our knowledge of the developmental relations between distinct tissues and cells. In our work to develop a method for measuring gene specificity that accounts for the similarity structure of the base sample set, we have created a portable framework that can readily generalize to describing other functional genomic features in a more system-aware manner.

We hope our work here will contribute to the advancement of systems levels of analysis and, through this, aid in the development of a deeper understanding of chromatin modifier diseases, and, eventually, the discovery of therapies and treatments.

Future direction for research and conclusion

Many diseases of human development that have so far eluded effective treatment or therapy have done so, in part, through the complexity of their pathology. This is particularly

resonant for chromatin modifier syndromes. It is possible that understanding these diseases to the depth necessary to devise effective treatments is beyond the reach of unaided human cognition. However, human cognition need not go unaided. Just as the index of a book or that of a library enables the individual to transcend the memory capacity of the human brain, computational tools allow us to ask questions that are otherwise beyond our own minds' cognitive limits.

In this work we have expanded on existing methods for the analysis and characterization of functional genomic features which are key facets of chromatin modifier biology. Going beyond our work to identify and characterize the scale of DNA methylation changes, future work to investigate the relations of features across different scales will be an essential part of understanding mechanistically how epigenetic features are established and maintained. For instance, identifying cases where a local feature gives rise to higher order organization, such as CTCF producing broad topologically associating domain (TAD)[20] structures, or in contrast to when a broad pattern then results in local changes such as broad repressive domains facilitating the methylation of individual CpG sites[21]. Mechanistic studies aimed at discovering the relations between functional genomic features are already well underway. Generalizing and automating methods for the discovery of such relations will enable their more rapid discovery and characterization in analogous systems, such as in studying gene regulation in novel non-model organisms.

Part of the motivation for generalizability, and therefore scalability, of analysis is also to go deeper into understanding the development of epigenetic organization in the course of organismal development. This was a core part of our motivation in working to improve measures of gene expression specificity to account for sample similarity information. The method we developed for balancing statistics of gene expression specificity in diverse sample sets can be extended to other descriptive statistics, such as the complexity of a gene's expression over

development. Additionally, the method we developed could be modified and optimized for other modes of data to enable a more systems level perspective of the whole of the functional genomic architecture of the cell and organism. Achieving systems level perspectives will require the development and adoption of methods of visualizing the functional genome and querying its structure. Summary statistics, which can look at a set of data and return a small, digestible, and intuitive handle on some aspect of the data's structure are an important component of this effort to see the functional genome, both as it exists in health as well as in disease.

In addition to the development of tools and methods to identify features of the functional genome and contextualize them, such as through summary statistics, the adoption of robust languages for describing and querying biological systems will be essential for the implementation of systems levels of analysis. In particular, we expect there is a substantial underutilization of formal biological ontologies, such as the gene ontology [22] and the anatomic and developmental UBERON ontology [23], in this space of data description, query, and analysis. The standard use case for the rich structure of ontologies is to perform set enrichment type analyses [24,25]. However, we expect ontology structures will prove useful as a foundation for ever deeper levels of analysis, richer reporting of discoveries, and more efficient integration of learned human knowledge. The development of novel computational tools and technologies will be required to see the full potential of ontology systems in biology come to fruition, but we are excited by their potential.

Biology is fundamentally the study of systems that: have evolved over a vast period of time, that develop and function on timescales spanning orders of magnitude, and that respond to and shape their environment at both microscopic and global scales. Handling the complexity of living systems to intelligently and efficiently ask questions, including those important for the advancement of medicine, will necessarily require augmenting our own analytic capacity with computation tools and methods. We hope that future research will continue to build on some of

the themes that we have developed here, particularly in advancing modes of analysis that can provide a more holistic and systems level perspective of biological phenomena.

It is an exciting time to be involved in biological research. Technologies have advanced rapidly and data is more plentiful and richer than at any other point in human history. Never has a more full appreciation for the structure, complexity, and beauty inherent to life been possible to strive for. There is much to be grateful for.

References

1. Kassis, J. A., Kennison, J. A. & Tamkun, J. W. Polycomb and Trithorax Group Genes in Drosophila. *Genetics* 206, 1699–1725 (2017).

2. Ding, Y., Liu, Z. & Liu, F. Transcriptional and epigenetic control of hematopoietic stem cell fate decisions in vertebrates. *Dev. Biol.* 475, 156–164 (2021).

3. Miller, C. T., Maves, L. & Kimmel, C. B. moz regulates Hox expression and pharyngeal segmental identity in zebrafish. *Development* 131, 2443–2461 (2004).

4. Watkins, W. S. *et al.* De novo and recessive forms of congenital heart disease have distinct genetic and phenotypic landscapes. *Nat. Commun.* 10, (2019).

5. Suliman, R., Ben-David, E. & Shifman, S. Chromatin regulators, phenotypic robustness, and autism risk. *Front. Genet.* 5, 81 (2014).

6. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* 42, 790–793 (2010).

7. Kennedy, J. *et al.* KAT6A Syndrome: genotype–phenotype correlation in 76 patients with pathogenic KAT6A variants. *Genet. Med.* 21, 850–860 (2019).

8. Lederer, D. *et al.* Deletion of KDM6A, a histone demethylase interacting with MLL2, in three patients with Kabuki syndrome. *Am. J. Hum. Genet.* 90, 119–124 (2012).

9. Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat. Genet.* 43, 729–731 (2011).

10. Nava, A. A. & Arboleda, V. A. The omics era: a nexus of untapped potential for Mendelian chromatinopathies. *Hum. Genet.* 143, 475–495 (2024).

11. He, Q. *et al.* Isoform-specific involvement of Brpf1 in expansion of adult hematopoietic stem and progenitor cells. *J. Mol. Cell Biol.* 12, 359–371 (2020).

12. Nazim, M. *et al.* Alternative splicing of a chromatin modifier alters the transcriptional regulatory programs of stem cell maintenance and neuronal differentiation. *Cell Stem Cell* 31, 754–771.e6 (2024).

13. Niessen, H. E. C., Demmers, J. A. & Voncken, J. W. Talking to chromatin: post-translational modulation of polycomb group function. *Epigenetics Chromatin* 2, 10 (2009).

14. Zhao, Y. & Garcia, B. A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb. Perspect. Biol.* 7, a025064 (2015).

15. Shah, S. G. *et al.* HISTome2: a database of histone proteins, modifiers for multiple organisms and epidrugs. *Epigenetics Chromatin* 13, 31 (2020).

16. Bondhus, L., Wei, A. & Arboleda, V. A. DMRscaler: a scale-aware method to identify regions of differential DNA methylation spanning basepair to multi-megabase features. *BMC Bioinformatics* 23, 364 (2022).

17. Bondhus, L., Varma, R., Hernandez, Y. & Arboleda, V. A. Balancing the transcriptome: leveraging sample similarity to improve measures of gene specificity. *Brief. Bioinform.* (2022) doi:10.1093/bib/bbac158.

18. Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp. Mol. Med.* 52, 1419–1427 (2020).

19. Jovic, D. *et al.* Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.* 12, e694 (2022).

20. Wiehle, L. *et al.* DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res.* 29, 750–761 (2019).

21. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* 44, 40–46 (2011).

22. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29 (2000).

23. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* 13, R5 (2012).

24. Zhou, T., Yao, J. & Liu, Z. Gene ontology, enrichment analysis, and pathway analysis. in *Bioinformatics in Aquaculture* 150–168 (John Wiley & Sons, Ltd, Chichester, UK, 2017).

25. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550 (2005).