Novel Computational Methods to Understand Epigenetic Variation and the Analysis of DNA Methylation Subtypes in Elderly Acute Myeloid Leukemia

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

John Taylor Lawson

December 2021

.

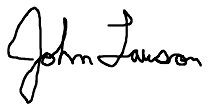
APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: 

This Dissertation has been read and approved by the examing committee:

Advisor: Nathan Sheffield

Advisor:

Committee Member: Francine Garrett-Bakelman

Committee Member: Jason Papin

Committee Member: Stefan Bekiranov

Committee Member: Kevin Janes

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

December 2021

Novel computational methods to understand epigenetic variation and the analysis of DNA methylation subtypes in elderly acute myeloid leukemia

By
John T. Lawson

A dissertation presented to the faculty of the School of Engineering and Applied Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Engineering
University of Virginia
Charlottesville, Virginia
December, 2021

Contents

List of Figures

Acknowledgments

I want to thank those who have supported me throughout graduate school. First, my family and especially my brother James have been indispensable. Second, my advisors Nathan Sheffield and Francine Garrett-Bakelman for their mentorship. Third, my friends Fadi and Andrew who help keep life fun. And for the many others who I had the chance to get to know in Charlottesville and at UVA, I am grateful. You all played a part in my journey so thank you!

Abstract

A key challenge in epigenetics is to determine the biological significance of high-dimensional genome-scale epigenetic data. In this dissertation, I present two methods that address this challenge and build on those methods to analyze DNA methylation subtypes in elderly acute myeloid leukemia (AML). The first method MIRA aggregates genome-scale DNA methylation data into a DNA methylation profile for independent region sets with shared biological annotation. Using this profile, MIRA infers and scores the collective regulatory activity for each region set. MIRA facilitates regulatory analysis in situations where classical regulatory assays would be difficult and allows public sources of open chromatin and protein binding regions to be leveraged for novel insight into the regulatory state of DNA methylation datasets. The second method COCOA is a computational framework that uses covariation of epigenetic signals across individuals and a database of region sets to annotate epigenetic heterogeneity. COCOA is the first such tool for DNA methylation data and can also analyze any epigenetic signal with genomic coordinates. I demonstrate COCOA's utility by analyzing DNA methylation, ATAC-seq, and multi-omic data in supervised and unsupervised analyses, showing that COCOA provides new understanding of intersample epigenetic variation. The MIRA and COCOA methods are available as Bioconductor R packages. Finally, I characterize DNA methylation subtypes in elderly AML. In AML, there is frequently epigenetic dysregulation, including dysregulation of DNA methylation. DNA methylation subtypes of AML can have different survival outcomes and mechanisms of dysregulation. In this study, I identify DNA methylation variation in regions related to stemness, differentiation, and hematopoietic cell type. Through a clustering analysis, I show that there are non-exclusive epigenetic subtypes depending on regions examined, with differing gene regulatory implications.

Chapter One: Introduction

The human body is composed of diverse tissues and cell types with essentially identical DNA. However, the tissue-specific function of the DNA is finely tuned to be expressed in the correct time and place in large part through epigenetics. Epigenetics describes the system of regulation that orchestrates chromatin structure and protein interactions to enable the context-specific transcription of genes as well as to maintain cell states across cell division. One epigenetic mechanism is DNA methylation, the covalent addition of a methyl group generally to a cytosine in DNA. DNA methylation can inhibit the DNA-binding of many transcription factors (TFs) and, therefore, is generally associated with decreased gene expression when present in gene promoters or enhancers (Tate and Bird, 1993). Other functions of DNA methylation include affecting transcriptional elongation and splicing when present in the gene body (Maor *et al.*, 2015; Lee *et al.*, 2015) but this dissertation will mostly focus on the effects of DNA methylation in relation to regulatory regions such as promoters and enhancers.

DNA methylation interacts with other regulatory features to control gene expression (Stadler *et al.*, 2011). DNA methylation can interfere with transcription factor (TF) binding (Domcke *et al.*, 2015) and is generally anti-correlated with transcriptional activity (Wiench *et al.*, 2011). The connection between methylation and TF binding goes both ways: TF binding affects and is affected by DNA methylation (Fujiki *et al.*, 2013; Zhu *et al.*, 2016), making it difficult to infer the causative factor; nevertheless, independent of directionality, the inverse correlation indicates that regulatory information can be derived from DNA methylation data. Multiple approaches have been used to relate DNA methylation to regulatory activity; for example, correlating differential methylation with expression of nearby genes (Yao *et al.*, 2015), or testing enrichment of TFs in differentially methylated regions (Wijetunga *et al.*, 2016; Yao *et al.*, 2015). These approaches are limited by arbitrary thresholds for differential methylation and do not make full use of genome-wide data. Also, factors other than DNA methylation levels, such as the shape of the DNA methylation profile around a site may be important to the site's activity (Kapourani and Sanguinetti, 2016).

Epigenetic data is inherently high-dimensional and often difficult to interpret. Because of the high dimensionality, it is common to group individual genomic loci into collections that share a functional annotation, such as binding of a particular transcription factor (Sheffield and Bock, 2015; Schep *et al.*, 2017; Lawson *et al.*, 2018). These genomic locus collections, or region sets, can be derived from experimental data such as ChIP-seq experiments that identify TF binding sites or locations of histone modifications across the genome. Other common sources are experiments such as ATAC-seq that find regions of accessible chromatin across the genome in certain cell types or biological conditions. Alternatively, region sets can be generated computationally, for example, by matching TF motifs to similar DNA sequences across the genome. Region sets are analogous to the more common gene sets, but relax the constraint that data must be gene-centric. While gene set approaches may be applied to epigenetic data by linking regions to nearby genes (McLean *et al.*, 2010), this linking process is ambiguous and loses information because a regulatory locus may affect the expression of multiple genes or more distant genes. Alternatively, a region-centric approach is often more appropriate for epigenetic data, and there are now many region-based databases and analytical approaches (Sheffield and Bock, 2015; Schep *et al.*, 2017; Sheffield *et al.*, 2013; Sheffield *et al.*, 2017; Dozmorov, 2017), such as using region set databases for enrichment analysis (Sheffield and Bock, 2015; Dozmorov, 2017; Layer *et al.*, 2018) or to aggregate epigenetic signals from individual samples across regions to assign scores of regulatory activity to individual samples or single cells (Schep *et al.*, 2017; Lawson *et al.*, 2018; Sheffield *et al.*, 2017; Boer and Regev, 2018).

Region-based methods have provided complementary ways to annotate and understand epigenomic data, but they suffer from three drawbacks: First, it is common to ignore covariation between the epigenetic signal and continuous patient phenotypes, relying instead on differential signals between discrete sample groups. This approach loses information about the differences among samples within a group. Second, the use of discrete cutoffs for identifying significant epigenetic differences between samples loses information about the strength of covariation between epigenetic features and sample phenotype. Third, existing approaches are generally specific to certain scenarios (e.g. unsupervised analysis) or data types (e.g. ATAC-seq), and therefore do not provide a generally applicable framework for covariation-based analysis. Chapter Two will address these issues.

Acute myeloid leukemia (AML) frequently involves epigenetic dysregulation, including dysregulation of DNA methylation (Figueroa, Lugthart, *et al.*, 2010). While AML has a very low mutational burden compared to

other common cancers, there are often mutations in epigenetic regulators (TCGAResearchNetwork, 2013; Papaemmanuil *et al.*, 2016). Epigenetic regulation is important for induction and maintenance of differentiation programs (Suzuki *et al.*, 2017) which are disrupted in AML (Mingay *et al.*, 2017). Multiple previous studies have identified AML DNA methylation subtypes (Figueroa, Lugthart, *et al.*, 2010; Glass *et al.*, 2017; Giacopelli *et al.*, 2021). DNA methylation subtypes of AML can have different survival outcomes and mechanisms of dysregulation (Figueroa, Lugthart, *et al.*, 2010). Previous studies have demonstrated that there are groups of patients with similar DNA methylation profiles and enrichment for certain mutations (Figueroa, Lugthart, *et al.*, 2010; Glass *et al.*, 2017).

The majority of previous work on DNA methylation subtypes in AML (Figueroa, Lugthart, *et al.*, 2010; Figueroa, Abdel-Wahab, *et al.*, 2010; Glass *et al.*, 2017; Rampal *et al.*, 2014) has focused on younger patients, defined here as under 60 years old, or mixed younger and elderly (>60 years old) populations (TCGAResearchNetwork, 2013; Lugthart *et al.*, 2010; Li *et al.*, 2016; Zhang *et al.*, 2018; Deneberg *et al.*, 2011; Bullinger *et al.*, 2009), despite AML having a median diagnosis age of 68 (National Cancer Institute, 2019). The epigenetic profiles of elderly AML patients are important not only because elderly people make up the majority of the patient population but because epigenetic changes occur with age (Maegawa *et al.*, 2014; Johnson *et al.*, 2017), which could affect AML pathways and contribute to the cancer in ways that have not yet been characterized.

The DNA methylation pathway is frequently dysregulated in AML, often through mutations in DNMT3A, TET2, or IDH1/2 (Figueroa, Abdel-Wahab, *et al.*, 2010; Papaemmanuil *et al.*, 2016). DNMT3A is a DNA methyltransferase. Compared to DNMT1, the enzyme mostly responsible for adding methylation to the newly synthesized DNA strand after DNA replication to maintain existing patterns (Ren *et al.*, 2018), DNMT3A is more often associated with de novo methylation of DNA. Patients with DNMT3A mutations often have significant hypomethylation of the genome (Glass *et al.*, 2017). In contrast, patients with TET2 and IDH1/2 mutations generally have increased DNA methylation levels (Figueroa, Abdel-Wahab, *et al.*, 2010; Glass *et al.*, 2017). TET2 is involved in the process of actively removing DNA methylation through a several step process (Rasmussen and Helin, 2016). The inhibition of the demethylation process through inactivation of TET2 results in genome hypermethylation. TET2 relies on the metabolite alpha-ketoglutarate to catalyze its reaction, connecting it to IDH1/2. The isocitrate dehydrogenase enzymes (IDH1 and IDH2) are metabolic enzymes that, with certain mutations, can produce an oncogenic metabolite 2-hydroxyglutarate (2HG) that is similar to alpha-ketoglutarate. 2HG inhibits TET2, leading to hypermethylation of the genome (Mingay *et al.*, 2017). 2HG can also inhibit other enzymes with similar mechanisms to TET2 but these effects are less well characterized. Some AML mutations affect other epigenetic pathways such as histone modification, for example, mutations in EZH2 (Rinke *et al.*, 2020), and chromatin organization, for example, mutations in RAD21 and STAG2 (Fisher *et al.*, 2017), but this dissertation will focus on DNA methylation in AML.

DNA methylation is important for regulating differentiation in hematopoiesis and can block or alter differentiation when dysregulated. For example, hematopoietic stem cells with TET2 knockout have skewed differentiation, forming fewer erythrocytes and common lymphoid progenitors and forming more HSCs and monocytes (Izzo *et al.*, 2020). In contrast, HSCs with DNMT3A knockout form fewer of a certain HSC subset and monocytes while forming more of certain erythrocytes, basophils, and megakaryocyte progenitors (Izzo *et al.*, 2020). These effects can be explained in part by the altered DNA methylation caused by these mutations. One study compared a mouse model with an oncometabolite-producing IDH1 mutation to the same model treated with vitamin C, which stimulates TET2 activity (Mingay *et al.*, 2017). This study found that vitamin C treatment resulted in lower DNA methylation at myeloid enhancers. These differentially methylated regions were also enriched for motifs of hematopoietic TFs. Binding of hematopoietic TFs can also be differentially affected by DNA methylation. Motifs of erythroid TFs generally have higher CpG frequency than monocytic TFs, a factor that may contribute to cell-type skewing seen when DNA methylation is increased or decreased by mutations in TET2 and DNMT3A (Izzo *et al.*, 2020). DNA methylation is coordinated during hematopoiesis and different defects in DNA methylation may have different functional consequences.

In addition, epigenetic state can change with age. Horvath et al. (Horvath, 2013) defined a multi-tissue "epigenetic clock" that predicted age from the DNA methylation of 353 CpGs. In hematopoietic stem cells specifically, age-associated epigenetic changes have been observed in DNA methylation, H3K27ac, H3K4me1, and H3K4me3 (Adelman *et al.*, 2019). Age-associated epigenetic changes can happen for multiple reasons. First, mutations accumulate with age, some of which affect the epigenome and may be selected for. An

example of this is age-related clonal hematopoiesis, where immune cells with a mutation in genes such as TET2 or DNMT3A are found more often in older individuals and can expand to make up a subpopulation of the hematopoietic system (Kandarakov and Belyavsky, 2020). Second, errors in replication of DNA methylation patterns in the process of cell division can lead to loss of DNA methylation (Petryk *et al.*, 2020). It is possible that age-related changes in DNA methylation contribute to age-related diseases such as AML.

Dissertation preview

During my PhD, I have done a combination of developing computational methods for epigenetic analysis and applying computational approaches to understand biology with a focus on DNA methylation in AML. This dissertation will follow a similar progression. This can be summarized broadly in two aims. **Aim 1:** Develop computational methods to gain insight into gene regulation from genome-scale epigenetic data. **Aim 2:** Characterize DNA methylation subtypes in elderly AML. The first section of Chapter Two will describe the computational method and Bioconductor package MIRA which analyzes the shape of the DNA methylation profile around regulatory regions to infer regulatory activity (Aim 1). The second section of Chapter Two will describe the computational method and Bioconductor package COCOA which takes advantage of the similar epigenetic states in regions with shared regulation and methods such as PCA that quantify covariation of features across samples to annotate epigenetic variation in a group of samples (Aim 1). Chapter Three will describe my analysis of DNA methylation subtypes in elderly AML (Aim 2), which builds on the methods described in Chapters Two. Finally, Chapter Four will provide a discussion of the projects and future directions. In all, these projects contribute to the advancement of the fields of epigenetic analysis and AML.

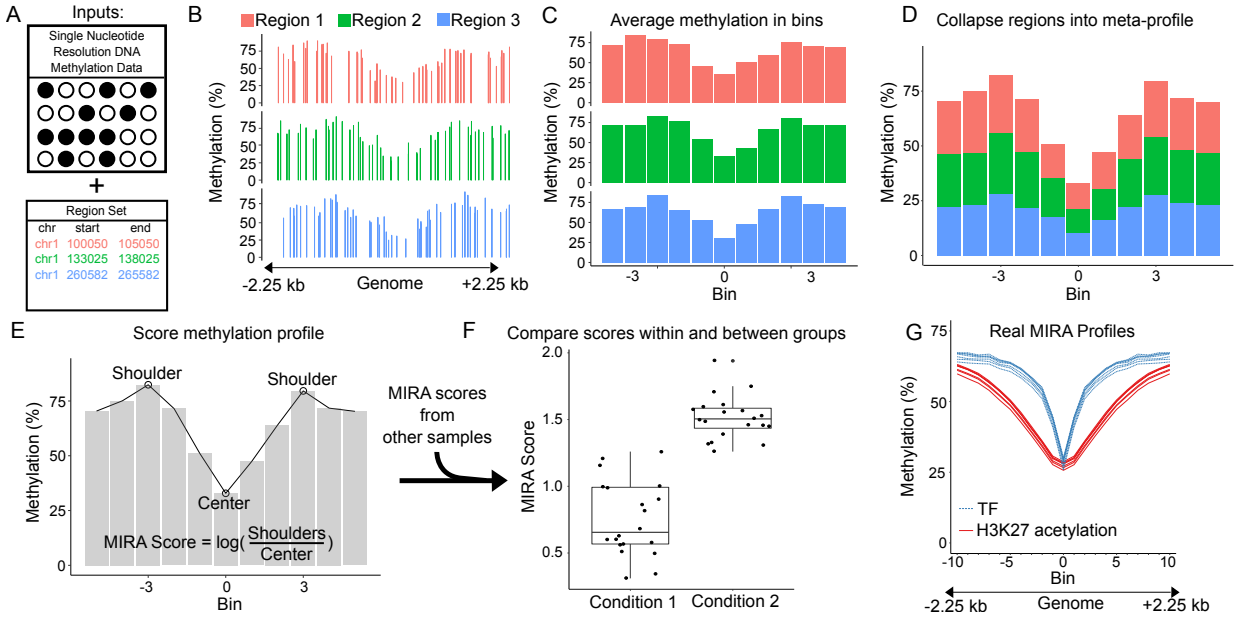
Chapter Two: Development of computational methods for epigenetics

Methylation-based inference of regulatory activity

The contents of this chapter section are a modification of the paper that is published here:

John T Lawson, Eleni M Tomazou, Christoph Bock, Nathan C Sheffield. MIRA: an R package for DNA methylation-based inference of regulatory activity, *Bioinformatics*. March 1, 2018. https://doi.org/10.1093/bioinformatics/bty083

We recently introduced and validated a novel method called MIRA (Methylation-based Inference of Regulatory Activity), which takes advantage of genome-scale DNA methylation data to assess regulatory activity (Sheffield *et al.*, 2017). We now present the MIRA R package which enhances this method and makes it broadly available. MIRA requires two inputs: 1) single-nucleotide-resolution, genome-scale DNA methylation data; and 2) a set of genomic regions that share a biological annotation (Fig. 1A). The DNA methylation data could come from sources such as whole genome or reduced representation bisulfite sequencing (WGBS or RRBS), or microarrays. Higher coverage of regulatory elements is ideal, but MIRA has been successfully tested with coverage as low as 450k array data. Genomic regions can be derived from sequencing assays such as ChIP-seq, DNase-seq, or ATAC-seq. Many region sets are publicly available through large-scale genomics projects and may be conveniently accessed through R packages like LOLA (Sheffield and Bock, 2015) and AnnotationHub (Morgan, 2019).



Dissertation figure 1: **Figure 1. MIRA workflow.** (A) Two inputs to MIRA: DNA methylation data for the sample of interest and a set of genomic regions that share a biological annotation (B) Three regions from the region set are shown for this example although a region set would normally be composed of thousands of regions. The DNA methylation level at individual CpGs is plotted for each 4.5 kb region, which is centered around a site of interest. (C) Each region is split into 11 bins of approximately equal size and an average methylation level is calculated based on the CpGs in each bin. (D) All regions are aggregated into a single DNA methylation profile by averaging methylation from the corresponding bins of each region. (E) The methylation profile is scored by taking the log of the ratio between the average methylation of the two shoulders and the methylation of the center. An algorithm determines the position of the shoulders. (F) As might be seen in an experiment that uses MIRA, the single score calculated from this sample is compared to scores from other samples of the same type – condition 1 – as well as to samples of a different type – condition 2. All scores were calculated using the same region set. The difference in scores between groups suggests differential activity of this region set. (G) Real MIRA profiles for a TF region set and for an H3K27 acetylation region set with DNA methylation data from 6 mesenchymal stem cell samples.

Using these two inputs, MIRA aggregates the DNA methylation levels of individual CpGs across all regions in a region set to create a summary profile. MIRA's 'aggregateMethyl' function does this through several steps: First, each region (Fig. 1B) is split into n (user-configurable) bins. Second, the DNA methylation level (0% to 100%) within a bin is averaged (Fig. 1C). Third, the regions are aggregated into a single summary profile by averaging the DNA methylation levels of each bin across all regions (all first bins averaged, all second bins averaged, etc.) (Fig. 1D). MIRA thus creates a 'meta-region profile' that provides general information about the activity of that region type across the genome Through aggregation, MIRA handles sparse DNA methylation data well. This makes MIRA well suited for low-coverage bisulfite sequencing (e.g. (Farlik *et al.*, 2015)). Once an aggregate profile is constructed (Fig. 1E), it is scored with MIRA's 'calcMIRAScore' function to quantify the regulatory activity of that type of region (Fig. 1F). MIRA assumes that genomic regions with lower DNA methylation levels have on average higher regulatory activity and gives a score based on the deepness of the characteristic "dip" in the middle of the 'meta-region profile'. MIRA automatically determines the location of the edges of the dip and calculates the score as the natural logarithm of the ratio between the DNA methylation level of the edges of the dip and the DNA methylation level of the center of the dip (Fig. 1F). The scores effectively reduce the DNA methylation profile to a number, which predicts the region set's aggregate regulatory activity. MIRA scores can be compared between samples to identify characteristic regulatory differences. MIRA supports a variety of applications depending on the context and what type of region set is used. For example, MIRA can be used to compare the chromatin states of different types of cells (Sheffield *et al.*, 2017). MIRA makes analysis of regulatory activity possible in cases where it would otherwise be infeasible. When sample amount or quality would not allow ATAC-seq or ChIP-seq but DNA methylation data can be obtained, regulatory analysis can be done with MIRA using existing ATAC-seq or ChIP-seq data (e.g., from a database). MIRA is also valuable for cases where it would be impractical in terms of time or cost to perform traditional regulatory assays, such as for large-scale cohort studies. The MIRA R package can be accessed via Bioconductor, and comes with multiple vignettes demonstrating how to apply it to biological data. MIRA provides a novel tool to enhance analysis of DNA methylation, enabling regulatory analysis in contexts where it was previously difficult and leveraging existing data from regulatory assays to gain new regulatory insights.

Coordinate covariation analysis

The contents of this chapter section are a modification of the paper that is published here:

John T. Lawson, Jason P. Smith, Stefan Bekiranov, Francine E. Garrett-Bakelman, and Nathan C. Sheffield. COCOA: coordinate covariation analysis of epigenetic heterogeneity, *Genome Biology*. Sept. 7, 2020. https://doi.org/10.1186/s13059-020-02139-4

Abstract

A key challenge in epigenetics is to determine the biological significance of epigenetic variation among individuals. We present Coordinate Covariation Analysis (COCOA), a computational framework that uses covariation of epigenetic signals across individuals and a database of region sets to annotate epigenetic heterogeneity. COCOA is the first such tool for DNA methylation data and can also analyze any epigenetic signal with genomic coordinates. We demonstrate COCOA's utility by analyzing DNA methylation, ATAC-seq, and multi-omic data in supervised and unsupervised analyses, showing that COCOA provides new understanding of intersample epigenetic variation. COCOA is available on Bioconductor (http://bioconductor.org/packages/COCOA).

Keywords: epigenetics; DNA methylation; chromatin accessibility; principal component analysis; dimensionality reduction; data integration; cancer; EZH2; multi-omics

Introduction

Here, we present Coordinate Covariation Analysis (COCOA), a method for annotating epigenetic variation across individuals using region sets. COCOA offers several advantages compared to existing methods: First, COCOA provides a flexible framework that supports both supervised and unsupervised analysis. Second, for supervised analysis, COCOA leverages covariation information by allowing continuous sample phenotypes as well as discrete groups. Third, COCOA incorporates epigenetic signal values instead of using binarized values (i.e. significant or not significant), further taking advantage of the covariation information. Finally, COCOA works with any epigenetic data that have a numerical value associated with genomic coordinates, such as DNA methylation data, chromatin accessibility data, or even multi-omics data. Importantly, no such tool that

leverages covariation of epigenetic signal across samples to annotate epigenetic variation previously existed for DNA methylation data. To demonstrate COCOA's utility, we applied it in three unsupervised analyses with DNA methylation, ATAC-seq, and multi-omics data, and a supervised analysis of DNA methylation and cancer stage. We found that across multiple data types and biological systems, COCOA is able to identify promising biological sources of epigenetic heterogeneity across sample populations.

Results and Discussion

An overview of COCOA

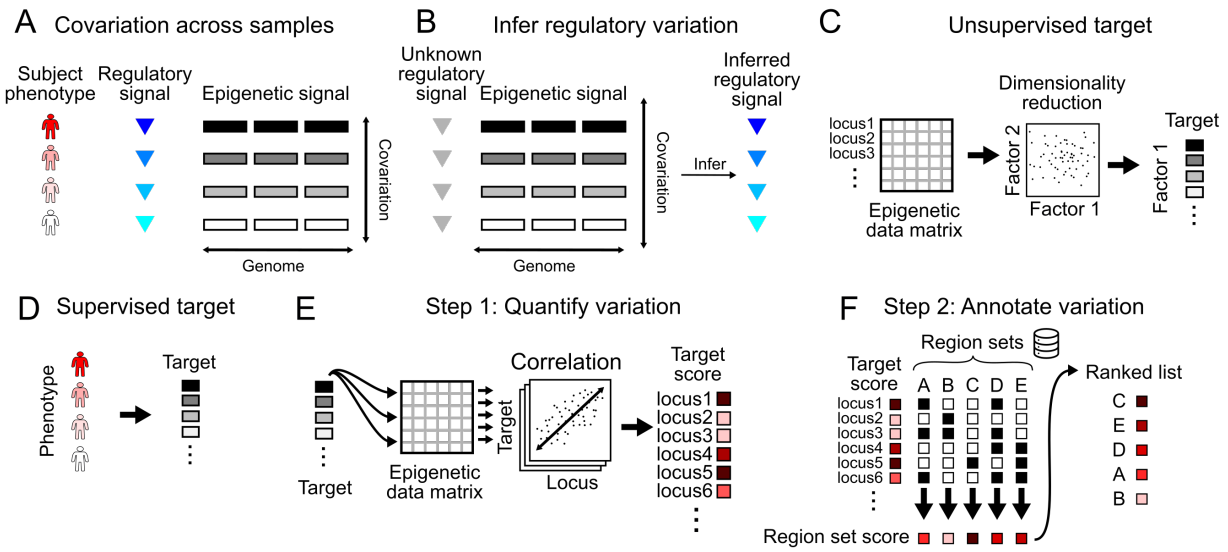
COCOA is an approach to understanding epigenetic variation among samples. COCOA derives its annotation power from a database of region sets that are grouped by function. This choice is rooted in the observation that a single effector, such as a transcription factor, often regulates many regions across the genome. Because the regions are coregulated, their epigenetic signal may covary across samples according to the activity of the effector (Fig. 1A), which can then be used to infer activity of the effector (Fig. 1B). This principle of covariation of coregulated loci or genes has been leveraged by other methods related to gene regulation (Schep *et al.*, 2017; Lawson *et al.*, 2018; Boer and Regev, 2018; Frost *et al.*, 2015; Subramanian *et al.*, 2005; Meng *et al.*, 2019; Odom *et al.*, 2019). To distinguish small differences among samples in the activity level of the effector, COCOA boosts statistical power by aggregating signal in region sets (Lawson *et al.*, 2018).

COCOA uses this aggregated region set approach to annotate the underlying source of epigenetic variation that relates to a "target variable," which can be either a supervised variable, like the phenotype of interest (Fig. 1C), or an unsupervised variable, like the primary latent factors in the data (Fig. 1D). COCOA annotates the inter-sample variation in the target variable by identifying region sets with variation patterns in epigenetic data that match the variation in the target variable. After a target variable is chosen, COCOA analysis consists of two main steps: first, for each locus, it computes the association of the inter-sample epigenetic variation with the target variable (Fig. 1E) and, second, it uses those associations to score a database of region sets (Fig. 1F). COCOA uses a permutation test to evaluate the statistical significance of each region set score. The result is a list of region sets ranked by how well the epigenetic signals in the region set correlate with the target variable. Highly scoring region sets have epigenetic signal that covaries across patients in the same way as the target variable, tying the functional annotation of the region set to the observed phenotypic variation.

COCOA annotates inter-sample variation in breast cancer DNA methylation data

We first evaluated COCOA in an unsupervised analysis to determine if COCOA could identify and annotate a driving source of variation. We applied COCOA to DNA methylation data from breast cancer patients in The Cancer Genome Atlas (TCGA). In breast cancer, estrogen receptor (ER) status is a major prognostic factor and is known to be associated with a specific DNA methylation profile (Ung *et al.*, 2014; Fleischer *et al.*, 2017). We first used Principle Component Analysis to identify the top four Principal Components (PCs), which we used as the target variables, and asked whether COCOA would be able to identify ER as an important source of inter-sample variation using only the DNA methylation data, without requiring the samples' ER status.

COCOA identified a strong ER-associated signature for Principal Component 1 (PC1). This signature included many ER-binding region sets as top hits, indicating that variation of the DNA methylation in these ER-binding regions is associated with PC1 (Fig. 2A, Additional file 1: Table S1). We also identified variation in region sets for FOXA1 and GATA3, which are known to be associated with ER status (Ung *et al.*, 2014; Fleischer *et al.*, 2017) (Additional file 1: Table S1). Furthermore, COCOA found the ER-associated histone modification H3R17me2 among the top scoring region sets (Frietze *et al.*, 2008) (Additional file 1: Table S1). When we test the association of each PC with ER status, PC1 scores have a highly significant association with ER status ($p < 10^{-46}$, Wilcoxon rank-sum test), whereas PC2 and PC3 are less associated (Fig. 2B). Therefore, COCOA clearly identified ER-related variation as relevant for the primary axis of inter-sample variation, despite not having access to ER status information. We found that PC4 was also associated with ER status to a lesser extent ($p < 10^{-20}$). For PC4, COCOA identified regions with repressive chromatin marks, including binding sites for polycomb components EZH2 and SUZ12 and repressive histone modifications H3K27me3 and H3K9me3 (Fig. 2A, Additional File 2: Fig. S1). Previous studies have linked polycomb expression to breast cancer: higher EZH2 expression is associated with ER- breast cancer (Guo *et al.*, 2016; Holm *et al.*, 2012), EZH2 interacts with the repressor of estrogen activity (REA) protein (Hwang *et al.*, 2007),



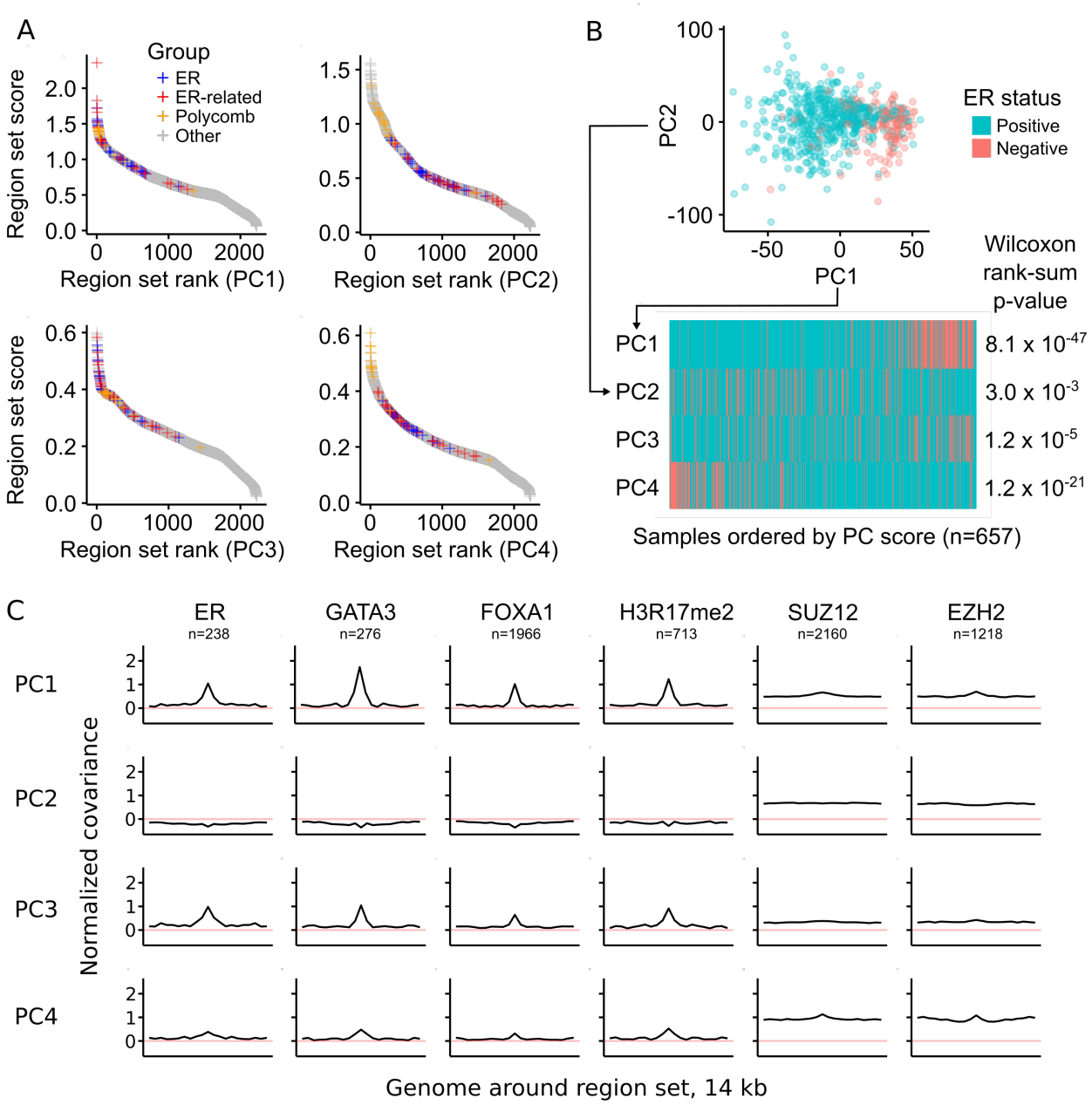
Dissertation figure 2: **Figure 1. Overview of COCOA**. A. A regulatory signal may covary with the epigenetic signal in the genomic regions it regulates. B. Covariation of the epigenetic signal in coregulated regions across individuals can be used to infer variation in the regulatory signal. C. COCOA can be used with an unsupervised target variable (latent factor), or D. with a supervised target variable (phenotype). E. The first step is to quantify the relationship between the target variable and the epigenetic data at each locus, resulting in a score for each locus. F. The second step is to annotate variation using a database of region sets. Each region set is scored to identify the region sets most associated with covariation between the epigenetic signal and the target variable. These top region sets can yield insight into the biological significance of the epigenetic variation.

and Suz12 binding sites have DNA methylation differences between ER+ and ER- breast cancer (Ung *et al.*, 2014). Therefore, PC4 represents an additional aspect of ER-related epigenetic variation. PC2 and PC3 had weaker associations with ER status ($p < 0.01$ and $p < 10^{-4}$ respectively); for PC3, the highest-ranking PC3 region sets include some ER-related region sets along with hematopoietic region sets (Additional file 1: Table S1). The hematopoietic region sets may represent inter-sample variation in the immune component of the tumors since breast cancer subtypes have been reported to be associated with differing immune cell profiles (Segovia-Mendoza and Morales-Montor, 2019). In summary, these results demonstrate that COCOA was able to identify relevant sources of inter-sample variation without requiring known sample groups and therefore reveal COCOA's usefulness for unsupervised analysis of DNA methylation data.

To visualize the inter-sample variation that drives the top region sets identified by COCOA, COCOA can also plot DNA methylation in a region set, ordered by PC value (Additional file 2: Fig. S2). Using this approach, we visualized how the DNA methylation in ER-related regions varies along PC1, demonstrating clear covariation across regions that drives the region set rankings (Additional file 2: Fig. S2). To further confirm the specificity of the region sets, COCOA can also plot variation in broader genomic regions around the regions of interest. We found that the DNA methylation close to the transcription factor binding regions shows stronger covariation with the PC score than DNA methylation in the surrounding genome (Fig. 2C). This visualization of specificity of the covariation to the binding regions provides additional evidence of association between the PC and region set. Other high-ranking transcription factors also showed this specificity (Fig. 2C, Additional file 2: Fig. S3). Some histone modifications, such as H3K9me3 and H3K27me3, where DNA methylation levels had high covariation with the PC showed broader regions of elevated covariation (Additional file 2: Fig. S3). Overall, these visualization functions reveal aspects of epigenetic variation in the top region sets that could not be captured by a single region set score.

COCOA annotates regulatory variation in ATAC-seq data

Next, we asked whether COCOA could be applied to ATAC-seq data. Unlike DNA methylation data, which annotates individual nucleotides, ATAC-seq data is summarized by accessibility values at "peak" regions (Corces *et al.*, 2018). COCOA handles either data type. To demonstrate the region-type analysis, we ran



Dissertation figure 3: **Figure 2. COCOA identifies sources of DNA methylation regulatory variation.** A. The COCOA score for each region set, ordered from highest to lowest. The ER-related group includes GATA3, FOXA1, and H3R17me2. The polycomb group includes EZH2 and SUZ12. B. The association of PC scores with ER status for PCs 1-4 based on a Wilcoxon rank-sum test. C. Meta-region profiles of several of the highest scoring region sets from PC1 (GATA3, ER, H3R17me2) and two polycomb group proteins (EZH2, SUZ12). Meta-region profiles show covariance between PC scores and the epigenetic signal in regions of the region set, centered on the regions of interest. A peak in the center indicates that DNA methylation in those regions covaries with the PC specifically around the sites of interest. The number of regions from each region set that were covered by the epigenetic data in the COCOA analysis (panel A) is indicated by "n".

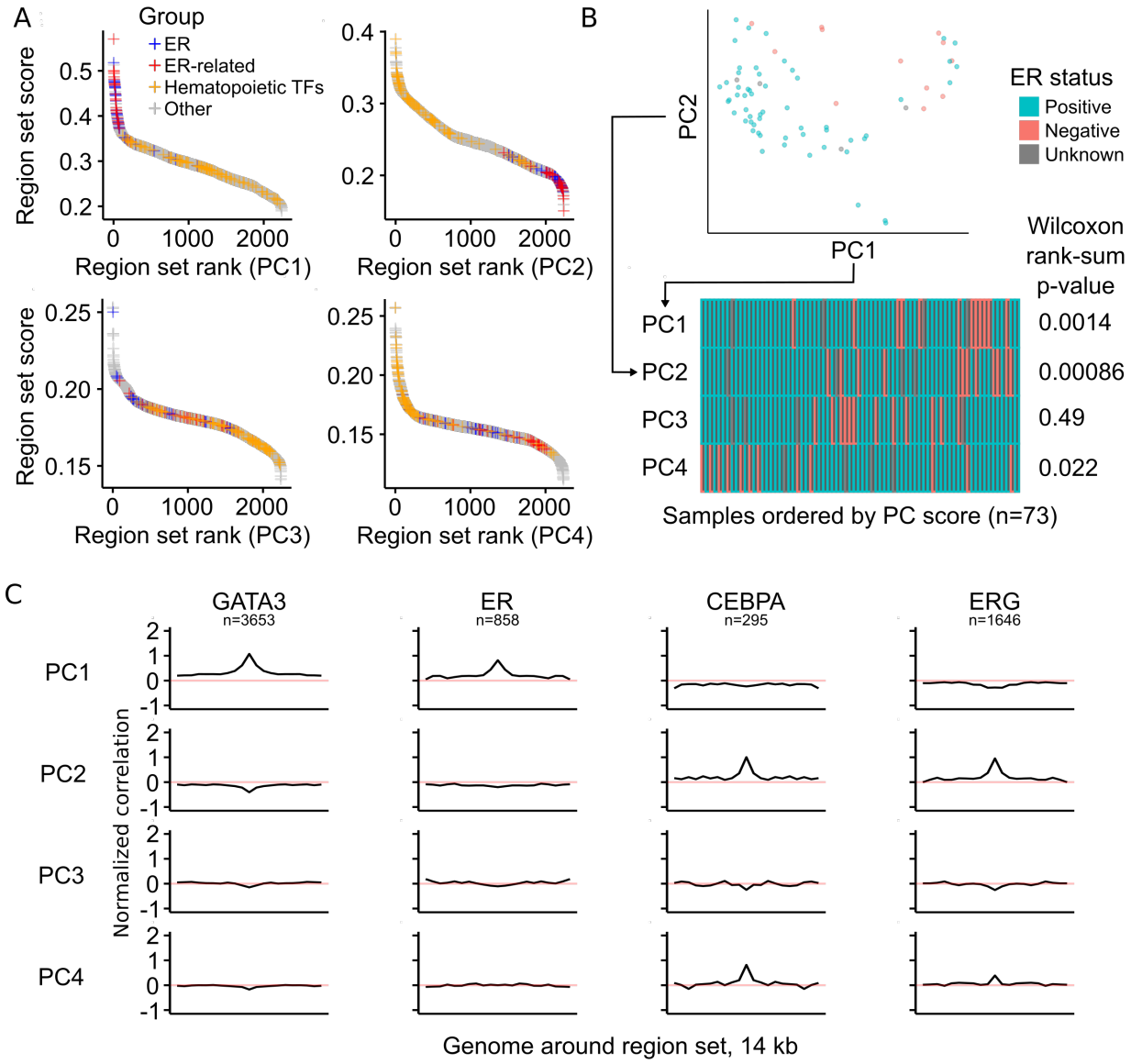
COCOA with ATAC-seq data from TCGA breast cancer patients (Corces *et al.*, 2018), expecting that ER-related region sets would be among our top results, similar to the DNA methylation data. As before, we used PCA on the ATAC-seq data and then applied COCOA to annotate the sources of variation for each PC. We identified many of the same region sets to be associated with epigenetic variation, despite far fewer samples (657 vs 73). We found ER-related region sets to be among the top ranked results for PC1 (Fig. 3A, Additional file 1: Table S2). PC2 was characterized by high-ranking hematopoietic transcription factors (Fig. 3A, Additional file 1: Table S2), once again potentially representing inter-sample variation in the immune component of the tumors (Segovia-Mendoza and Morales-Montor, 2019), as in PC3 of the DNA methylation data. A few other top PCs including PC4 also had high-ranking hematopoietic transcription factors (Fig. 3A, Additional file 2: Fig. S4). Consistent with our results, visual inspection of the chromatin accessibility signal in top ER-related and hematopoietic region sets also revealed correlation between the signal and PC scores for the PCs in which the region sets were highly ranked (Additional file 2: Fig. S5). Polycomb region sets did not rank as prominently for the ATAC-seq data as for the DNA methylation data but there were several polycomb region sets in the top 10% of region set scores for PC4 (Additional file 2: Fig. S6, Additional file 1: Table S2). These results are consistent with variation in ER status, which is significantly associated with PC1 and PC2 ($p < 0.01$, Wilcoxon rank-sum test, Fig. 3B) and to a lesser extent PC4 ($p < 0.05$). Visualization of the correlation between each PC and the ATAC-seq signal in the top region sets also shows specificity to the transcription factor-binding regions compared to the surrounding genome (Fig. 3C). Thus, COCOA can identify meaningful sources of variation in ATAC-seq data, providing a novel tool for regulatory analysis of ATAC-seq data.

COCOA identifies regulatory variation in multi-omics integration

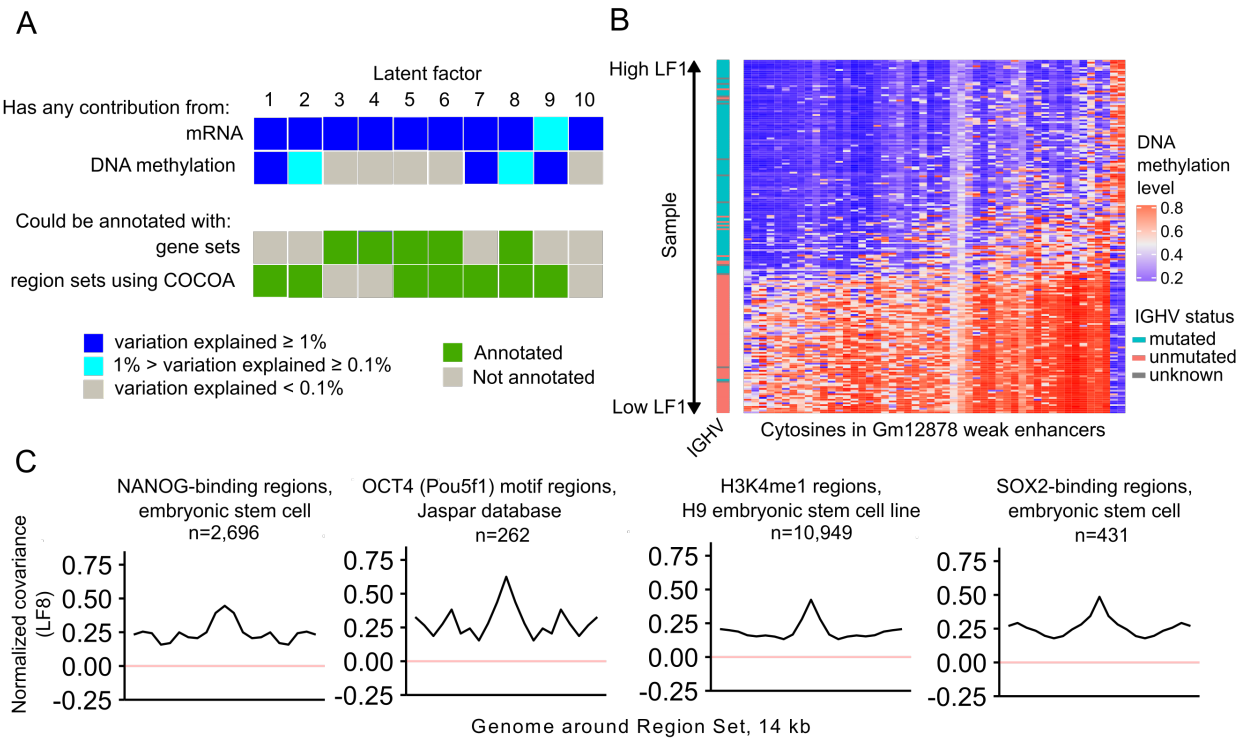
We also aimed to determine if COCOA could annotate inter-sample variation in multi-omics analyses that integrate epigenetic data with other data types. We therefore applied COCOA to a cohort of 200 chronic lymphocytic leukemia patients (Dietrich *et al.*, 2017) with gene expression, *ex vivo* drug response, somatic mutation, and DNA methylation data. We used preprocessed data from MOFA (Multi-Omics Factor Analysis), a multi-omics dimensionality reduction method that summarized the high-dimensional data into 10 new dimensions referred to as latent factors (LFs) (Argelaguet *et al.*, 2018). As part of the published analysis interpreting the 10 latent factors, the authors used a gene-centric method to annotate the latent factors with gene sets but only 5 could be associated with gene sets (Frost *et al.*, 2015; Argelaguet *et al.*, 2018). Because COCOA works with data associated with genomic coordinates, we were able to use the DNA methylation data from the MOFA analysis with COCOA to annotate the latent factors with region sets. Since only a subset of the DNA methylation data was used for the MOFA calculations, we calculated the correlation of each CpG in the 450k microarray with each latent factor and used this matrix as input for COCOA. Using COCOA, we are able to annotate 4 of the 5 latent factors that were not associated with gene sets, demonstrating that COCOA's region-centric approach complements the gene-centric approach applied by the MOFA authors (Fig. 4A). For latent factor 1 (LF1), we found variability in region sets for hematopoietic regulatory regions and transcription factors (Additional file 1: Table S3), consistent with the conclusions of the original paper that LF1 is related to the hematopoietic differentiation state of the leukemic cell of origin. The top region set for LF1 was enhancer regions in the GM12878 transformed B-lymphocyte cell line, which had stark differences in DNA methylation across samples that correlated with IGHV mutation status, a marker of mature B cells that have undergone somatic hypermutation (Fabbri and Dalla-Favera, 2016) (Fig. 4B). This result shows that COCOA was able to identify a plausible source underlying the latent factor, a result which was not identified using gene sets. As another example, we found region sets related to stem cell biology, including OCT4, NANOG, H3K4me1 from the H9 stem cell line, and SOX2, to be associated with LF8 (Fig. 4C, Additional file 1: Table S3). Since OCT4 and NANOG activity has been shown to be associated with beta catenin (Takao *et al.*, 2007; Faunes *et al.*, 2013; Ying *et al.*, 2015), a mediator of WNT signaling, our results support and further expand upon the original association between LF8 and WNT reported by the MOFA authors. These results demonstrate that COCOA can enable richer multi-omics analysis by annotating the epigenetic component of inter-sample variation.

COCOA reveals associations between epigenetic state and variation in sample phenotype

The three examples thus far demonstrate how COCOA can be applied in an unsupervised analysis, which explores biological variation in the absence of known groups. To explore whether we could apply COCOA to a setting where groups or phenotypes are known, we extended COCOA to accommodate supervised analysis. For the supervised approach, we select a sample phenotype of interest (such as a molecular phenotype or a clinical



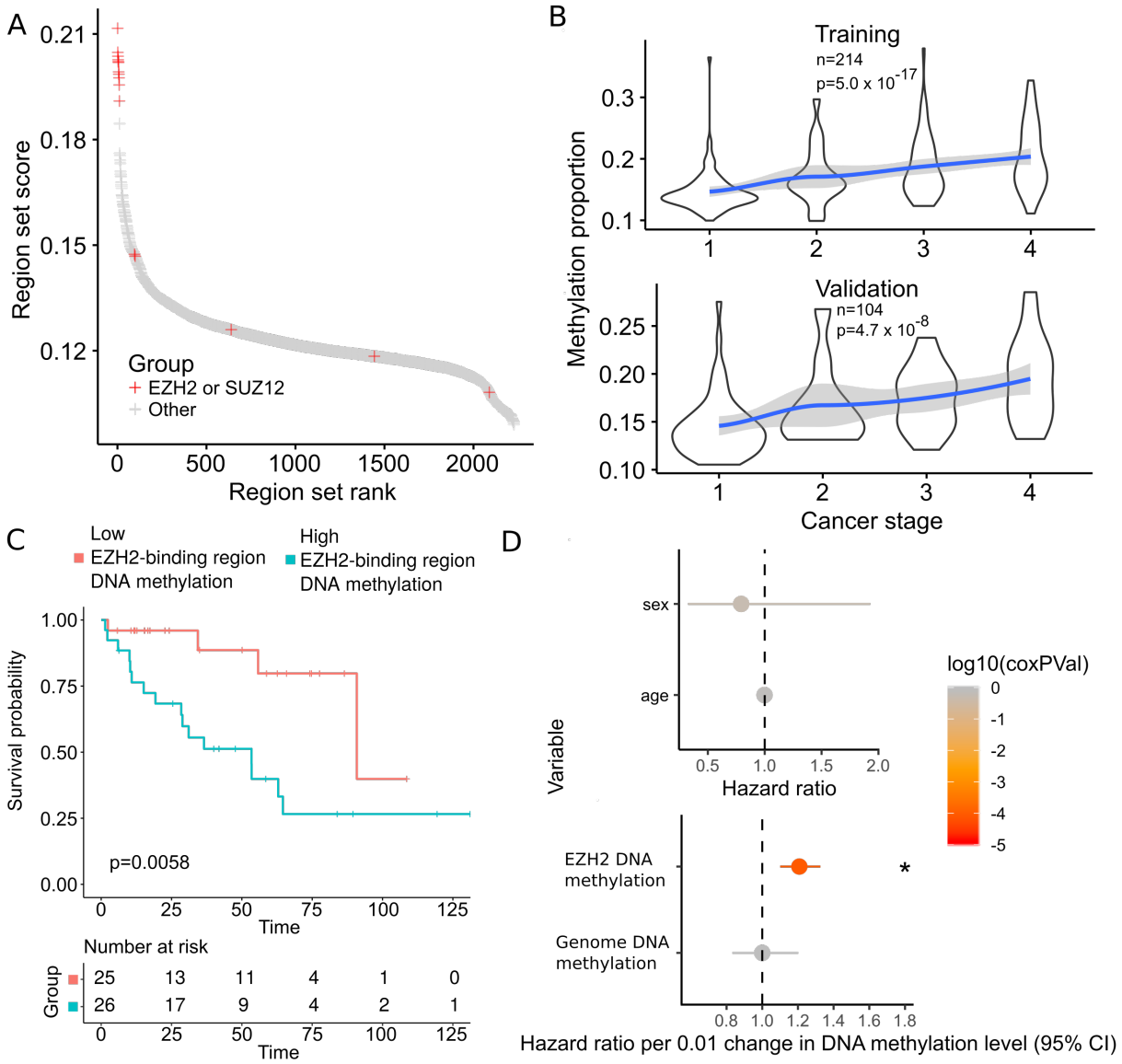
Dissertation figure 4: **Figure 3. COCOA can be used for region-based data such as ATAC-seq.** A. The COCOA score for each region set, ordered from highest to lowest. The ER-related group includes GATA3, FOXA1, and H3R17me2. For definition of the hematopoietic TF group, see "Region set database" in methods. B. The association of PC scores with ER status for PCs 1-4 based on a Wilcoxon rank-sum test. C. Meta-region profiles of the two highest scoring region sets from PC1 (GATA3, ER) and PC2 (CEBPA, ERG). Meta-region profiles show correlation between PC scores and the epigenetic signal in regions of the region set, centered on the regions of interest. A peak in the center indicates that chromatin accessibility in those regions correlates with the PC specifically around the sites of interest. The number of regions from each region set that were covered by the epigenetic data in the COCOA analysis (panel A) is indicated by "n".



Dissertation figure 5: **Figure 4. COCOA can be applied to multi-omics analyses that include epigenetic data.** A. COCOA can annotate latent factors that were not annotated by a gene set approach. In the top of panel A, dark blue indicates that the data type explained at least 1% of the variation of the latent factor while light blue indicates that the data type explained between 0.1% and 1% of the variation. Gray indicates less than 0.1% explained. In the bottom of panel A, green indicates that at least one statistically significant gene set or region set was found for the latent factor and gray indicates no significant gene or region sets were found. B. COCOA identifies an enhancer region set from a transformed B-lymphocyte cell line where DNA methylation is correlated with latent factor 1 and IGHV mutation status, a marker of mature B cells that have undergone somatic hypermutation. The 50 CpGs with the highest absolute correlation with LF1 from the region set are shown. C. Meta-region profiles show covariation between DNA methylation and LF8 score in certain regions bound by transcription factors functional in stem cell biology and by H3K4me1 in a stem cell line compared to the surrounding genome. The number of regions from each region set that were covered by epigenetic data in the COCOA analysis is indicated by "n".

outcome) and then measure the association of epigenetic variation with that parameter. To demonstrate a supervised COCOA analysis, we analyzed TCGA 450k methylation microarrays from kidney renal clear cell carcinoma (KIRC). This dataset includes a phenotypic annotation of cancer stage, which we used as our target variable. We hypothesized that COCOA could associate an epigenetic regulatory state with cancer stage and decreased survival. To test this hypothesis, we used COCOA to identify region sets where DNA methylation is correlated with cancer stage. We used a training-validation approach to assess significance of our results (see Methods). In the training samples, COCOA identified polycomb protein (EZH2 and Suz12)-binding region sets to have the highest correlation with cancer stage (Fig. 5A, Additional file 1: Table S4). Next, we tested whether the average DNA methylation level in the top EZH2 region set is associated with cancer stage. In both training and validation samples, average DNA methylation level in EZH2-binding regions had a significant positive correlation with cancer stage ($p < 10^{-16}$ and $p < 10^{-7}$, t approximation) showing that the COCOA result extends beyond the training set (Fig. 5B, Additional file 3: Table S5). Higher DNA methylation levels in EZH2-binding regions in advanced stages of cancer suggest that these regions could be repressed in advanced cancer stages, which would be consistent with higher activity of the repressive protein EZH2. This result is consistent with previous studies, which have found that higher EZH2 expression could promote metastasis in renal cell carcinoma (Zhang *et al.*, 2017) and other cancers [Varambally *et al.* (2002); Cheng2017; Kim and Roberts (2016)] and is associated with a more advanced cancer stage (Bachmann *et al.*, 2006; Melling *et al.*, 2015).

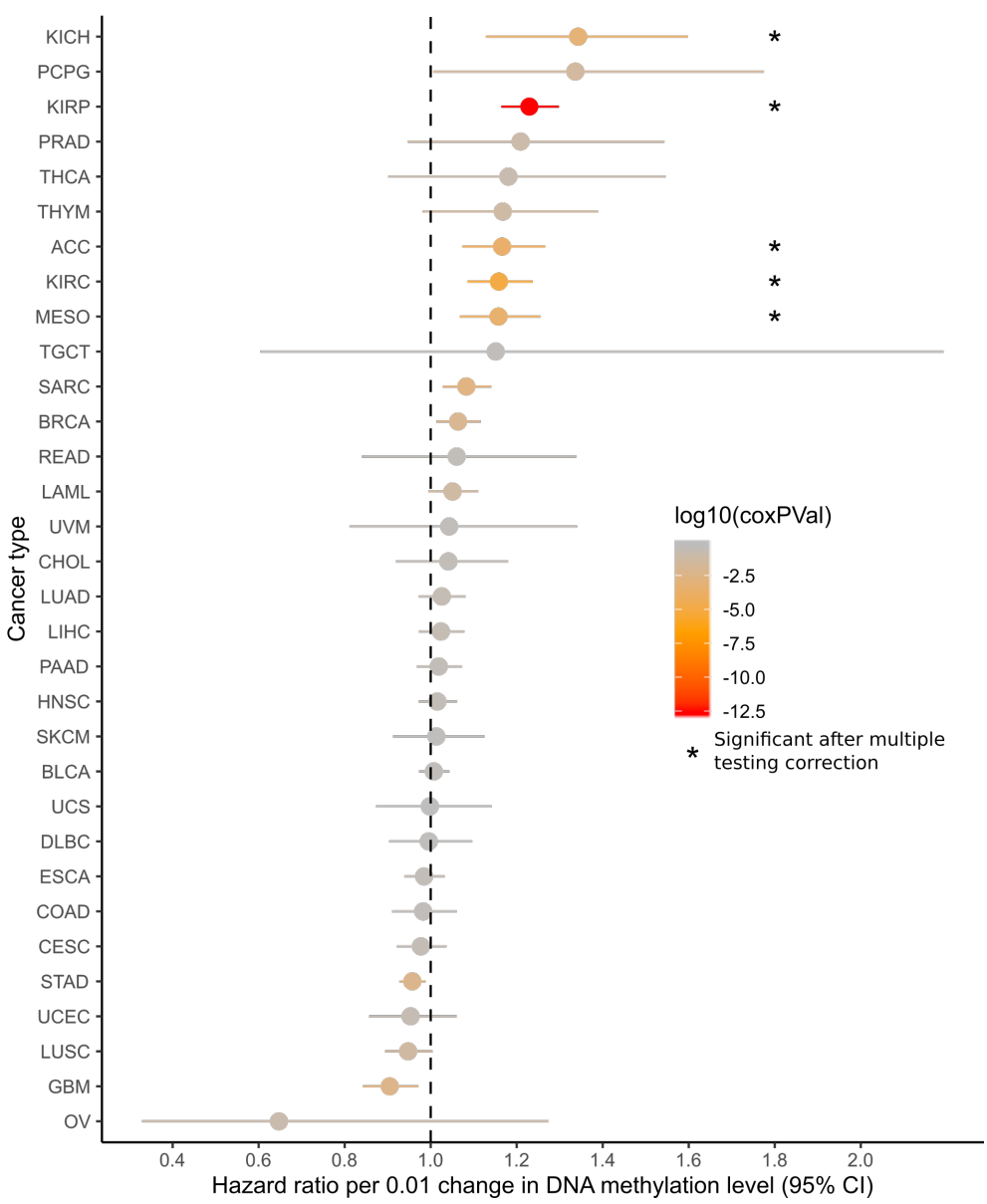
To further assess the relevance of our COCOA results, we tested the association between DNA methylation in our top EZH2 region set and patient survival. We compared the quartile of patients with highest average DNA methylation in the EZH2 region set to the quartile of patients with the lowest average DNA methylation, using a Kaplan-Meier estimate (Fig. 5C). Patients with higher EZH2 region set DNA methylation have significantly decreased survival compared to those with lower DNA methylation ($p < 0.01$, log-rank test, Fig. 5C). A Cox proportional hazards model correcting for age, gender and average genome methylation levels also revealed a significant association between average DNA methylation level in EZH2 binding regions and patient survival ($p < 10^{-4}$, Fig. 5D, Additional file 4: Table S6). Previous studies found EZH2 expression to be prognostic for survival in renal cell carcinoma and other cancers (Varambally *et al.*, 2002; Bachmann *et al.*, 2006; Liu *et al.*, 2013), but to our knowledge, this is the first demonstration that DNA methylation levels in EZH2 binding regions could be prognostic for survival in renal cell carcinoma. We further assessed for cancer stage and survival association for the top TF region sets from the COCOA analysis. We tested the two highest scoring TFs – JUND and TCF7L2. In the validation data, DNA methylation in JUND-binding regions had a significant negative correlation with cancer stage ($p=0.022$, t approximation) but we could not validate its association with survival because it did not satisfy the Cox proportional hazards assumption (Additional file 2: Fig. S7A, Additional file 4: Table S6). DNA methylation in TCF7L2-binding regions was not significantly correlated with cancer stage in the validation data (Additional file 2: Fig. S7B, Additional file 3: Table S5) but higher DNA methylation was significantly associated with better overall survival ($p= 0.038$, Cox proportional hazards model, Additional file 2: Fig. S7C, Additional file 4: Table S6). Through this supervised analysis, we demonstrate that COCOA can identify epigenetic variation related to a given sample phenotype of interest, providing a novel means for targeted analysis of epigenetic variation.



Dissertation figure 6: **Figure 5. COCOA identifies region sets related to a patient phenotype of interest, cancer stage.** A. Region sets for the polycomb proteins EZH2 and SUZ12 were the top region sets related to cancer stage. B. Average DNA methylation level in EZH2-binding regions (the top EZH2 region set) increases with cancer stage. P-values by t approximation with null hypothesis that correlation is zero. C. Kaplan-Meier curves of the validation samples, grouping samples by average DNA methylation in EZH2 binding regions (25% highest samples and the 25% lowest samples). P-value from log-rank test. E. Cox proportional hazards model of average DNA methylation in the top EZH2-binding region set, correcting for age, gender, and average genome methylation level.

DNA methylation in EZH2-binding regions is associated with cancer stage and survival in multiple cancers

Given that COCOA identified associations with EZH2 region sets in both our unsupervised analysis of breast cancer and our supervised analysis of kidney renal cell carcinoma, we wondered whether the link with EZH2 and DNA methylation would hold true for other cancer types. To test this, we performed a pan-TCGA analysis investigating the association between average DNA methylation in EZH2/SUZ12-binding regions and cancer stage as well as overall patient survival. We combined regions from the top group of 11 EZH2 and SUZ12 region sets from the KIRC analysis (Fig. 5A, Additional file 1: Table S4) to generate a single EZH2/SUZ12 region set, referred to hereafter simply as EZH2-binding regions. We then computed the average DNA methylation in this region set for each sample and tested its association with either cancer stage or overall survival. We found a significant correlation between DNA methylation in EZH2-binding regions and cancer stage in multiple cancer types (Additional file 2: Fig. S8, Additional file 5: Table S7, Additional file 6: Table S8). DNA methylation in EZH2-binding regions positively correlated with cancer stage in 5 of 21 tested cancers, but trended negative in 3 cancer types (Additional file 2: Fig. S8), of which colon adenocarcinoma (COAD) had a significant negative correlation ($p < 0.05$, t approximation, Holm-Bonferroni correction), consistent with a previous report (Chen *et al.*, 2017). To further investigate the significance of the EZH2-binding regions, we used a Cox proportional hazards model to test for association between survival and average DNA methylation in these regions and found a significant association in 5 cancer types (Fig. 6, Additional file 6: Table S8, Additional file 7: Table S9). Similar to the cancer stage analysis, higher DNA methylation level was more often associated with increased risk of death, but trended to lower risk in a few cancer types (Fig. 6). This result is consistent with previous reports that EZH2 can be either oncogenic or a tumor suppressor (Kim and Roberts, 2016; Wang *et al.*, 2017; Basheer *et al.*, 2019) and emphasizes the context-specific effects of EZH2. Our pan-cancer analysis also supports previous reports suggesting that polycomb activity may be commonly dysregulated in cancer (Kim and Roberts, 2016) and may influence survival in a variety of cancers, with some cancers having a positive and others a negative association (Kim and Roberts, 2016). Our results contrast with previous reports for several cancer types (Supplementary Discussion). This analysis identified a novel connection between EZH2 and survival in adrenocortical carcinoma (ACC), which has not been previously demonstrated. Furthermore, we have shown for the first time that variation in DNA methylation at EZH2-binding regions is associated with cancer stage and patient survival across a variety of cancers. Overall, this analysis demonstrates the ability of COCOA to annotate epigenetic variation and its potential to generate new mechanistic hypotheses about epigenetic heterogeneity and disease drivers.



Dissertation figure 7: **Figure 6. Pan-cancer survival analysis of DNA methylation in EZH2/SUZ12-binding regions.** The mean hazard ratio and 95% confidence interval for the average DNA methylation in EZH2/SUZ12-binding regions are shown for each cancer type. Color indicates the raw p-values and asterisks mark significance after Holm-Bonferroni correction.

Comparison of COCOA to other methods

COCOA distinguishes itself from other methods by being the only method of its type for DNA methylation data and by its flexibility in supporting a wide range of analyses for epigenetic data. We conceptualize COCOA as being in a class of methods that relies on covariation of epigenetic signal to annotate epigenetic variation. This separates COCOA from the methods that annotate epigenetic variation without taking into account covariation. To demonstrate the power of this approach, we compared COCOA to LOLA, a method that does not consider covariation. This analysis demonstrated that COCOA has superior ability to mitigate noise (Supplemental information, Additional file 2: Fig. S9, Additional file 8: Table S10, Additional file 1: Tables S11 and S12). Other methods that do take into account covariation have key differences from COCOA. First, while tools exist that aggregate signal in related groups such as gene sets or region sets and use PCA to identify covariation of signal across samples (Table 1), no existing tool does this for DNA methylation data. Second, COCOA creates a generalized framework for region set analysis which results in great flexibility in applications. This generalized framework allows COCOA to be used in analyses that other tools may not support: with multiple epigenetic data types, for supervised or unsupervised analyses, with a variety of mathematical metrics, and for single-omic or multi-omic analyses. For a brief description of each method from Table 1 and further comparison to COCOA, see "Comparison of COCOA to other region set or covariation-based methods" in the supplementary text. Of the epigenetic tools with similar goals to COCOA, chromVAR (Schep *et al.*, 2017) is the most widely used and most similar to COCOA in its input type. Therefore, we selected chromVAR for comparison to COCOA with the breast cancer ATAC-seq data. Each method revealed relevant but partially divergent aspects of inter-sample variation. COCOA had an improved ability to identify ER-related epigenetic variation and to separate biological signals with its use of PCA (Additional file 2: Fig. S10, Supplementary Information: "Comparison of COCOA to chromVAR", Additional file 1: Table S2, Additional file 9: Table S13). COCOA also extends beyond chromVAR in COCOA's analysis options and supported data types. COCOA thus provides a novel framework for flexible covariation-based analysis of DNA methylation and other epigenetic data.

	primary data type is DNA methylation data	primary data type is chromatin accessibility data	primary data type is gene or protein expression data	single cell focus	supports multi-omic analysis	region-centric	aggregates signal in genome-scale groups	uses PCA or matrix factorization	supervised (S), unsupervised (U), or both (B)	programming language
COCOA	✓	✓	×	×	✓	✓	✓	✓	B	R
chromVAR	×	✓	×	✓	×	✓	✓	×	U	R
BROCKMAN	×	✓	×	✓	×	✓*1	✓	✓	U	R/Python
PCGSE	×	×	✓	×	✓	×	✓	✓	U	R
MOGSA	×	×	✓*2	×	✓	×	✓	✓	U	R
pathwayPCA	×	×	✓*2	×	✓	×	✓	✓	B	R
coMethDMR	✓	×	×	×	×	✓*3	×	×	U	R

Dissertation figure 8: **Table 1. Features of COCOA and related methods.** *1BROCKMAN uses k-mer counts but the regions containing each k-mer can be conceptualized as a region set. *2MOGSA and pathwayPCA can involve multiple "omics" data types but including gene-centric data such as gene or protein expression is important for the methods. *3coMethDMR finds differentially methylated regions but often annotates them in reference to genes.

Conclusion

We created a flexible framework for identifying and understanding sources of regulatory variation in epigenetic data. COCOA could be applied to any epigenetic data that has a value associated with genomic coordinates, which includes both nucleotide-level data such as bisulfite sequencing and region-based data such as ATAC-seq data. Our results also demonstrate how COCOA can be integrated with multi-omics analyses that include epigenetic data. Our tool allows scientists to leverage publicly available regulatory data to annotate variation in their epigenetic data. In an unsupervised analysis, COCOA can annotate the major axes of inter-sample variation. In a supervised analysis, COCOA can annotate inter-sample variation related to a specific phenotype of interest. We have released COCOA as a Bioconductor package (Huber *et al.*, 2015), facilitating this new method of regulatory analysis. COCOA is a flexible and powerful method for interpreting regulatory variation between individuals.

Methods

COCOA algorithm

Overview

COCOA annotates variation in epigenetic data through two steps. In the first, we quantify the association between each feature in the epigenetic data and the target variable using a metric such as correlation (Fig. 1C). This gives a score to each epigenetic feature that represents how much it is associated with the target variable. Then in the second step, we use the epigenetic feature scores to score region sets from a large collection of region sets (Fig. 1D). Finally, we use a permutation test to assess statistical significance, and return a ranked list of region sets.

Step 1: Quantifying variation across samples

COCOA starts with a data matrix of epigenetic signal values in genomic regions, where each row is a genomic locus (e.g. a CpG or an ATAC-seq region), and each column is a sample. The values in the matrix correspond to signal intensity levels (e.g. DNA methylation level or chromatin accessibility) of a given sample at a given locus. The first step in a COCOA analysis is to transform the original data into a score for each locus measuring how much it contributes to the target inter-sample variation. We refer to the score for an epigenetic feature (locus) as a "feature contribution score" (FCS). This calculation can be either supervised or unsupervised (Fig. 1B):

Supervised. For supervised analyses, the goal is to identify sources of variation associated with a target sample phenotype of interest. Therefore, in addition to the epigenetic data matrix, we require a vector representing the target sample phenotype. We then quantify the association between the target sample phenotype and the epigenetic signal at each genomic locus using a method such as Pearson correlation. We end up with a vector of scores (which for correlation is the correlation coefficient) representing how strongly epigenetic variation at a genomic locus is associated with variation in the sample phenotype. Metrics other than Pearson correlation can be used to quantify variation, as long as they produce a score for each genomic locus. A detailed discussion of metric choice follows in the section, *Metric for quantifying variation.*

Unsupervised. For unsupervised analyses, we first apply a dimensionality reduction technique such as PCA or MOFA (Argelaguet *et al.*, 2018) to identify latent factors that represent significant sources of inter-sample variation (Frost *et al.*, 2015). Then, we treat these latent factors as target sample phenotypes and quantify the association between each latent factor and the epigenetic data as we would for the sample phenotype in the supervised analysis. In this case, the feature contribution score for each genomic locus represents how strongly epigenetic variation at that genomic locus is associated with variation in the latent factor.

Step 2: Annotate variation with the COCOA algorithm

After quantifying inter-sample variation, we are left with one or more vectors that assign FCS to each genomic locus in the original data matrix. COCOA next seeks to determine which region sets are associated with that variation. For this step, COCOA relies on a database of region sets. Here, we have used a subset of the LOLA database (Sheffield and Bock, 2015), which includes several thousand region sets that have been manually collected from several large-scale experiments and databases, including the ENCODE (Consortium, 2012; Davis *et al.*, 2017) and Roadmap Epigenomics projects (Bernstein *et al.*, 2010; Kundaje *et al.*, 2015). For the sample-specific data, COCOA can operate on two types of signal data: single-nucleotide data (e.g. DNA

methylation) or region-based data (e.g. ATAC-seq peaks). In either case, we will aggregate the scores for all individual genomic loci into a combined score for each region set (Fig. 1D). Due to different experiments testing the same TF or histone modification, some region sets share similar regions to each other and therefore their scores are not completely independent.

For single base-pair resolution data (e.g. DNA methylation data), the following algorithm is used for a single region set and a single FCS vector: First, we optionally take the absolute value of the FCS (Supplementary Methods). Then, we identify all features whose genomic coordinates overlap the given region set. Within each region from the region set, we average the FCS of any overlapping features to get a single average value for each region. We then average the region scores to get the final score for that combination of region set and FCS vector. This score represents how much that region set is associated with the latent factor or phenotype that corresponds to the FCS vector. We repeat this process for each pairwise combination of region set and latent factor/phenotype FCS vector.

For region-based data such as ATAC-seq data, the scoring is conceptually similar to single-nucleotide data, but with slight differences. We use the following algorithm: To score a region set for a given latent factor or phenotype, we first identify all overlaps between "data regions" (regions for the epigenetic signal data) and region set regions. For each overlap, we calculate what proportion of the region set region is overlapped by the data region. We then take a weighted average of the FCS of all the overlapping data regions, weighting each data region's FCS by the proportion that region overlaps a region set region and dividing by the sum of all overlap proportions. This weighted average is the region set score that represents how much the region set is associated with the latent factor or phenotype. We repeat this process for each combination of region set and latent factor/phenotype.

COCOA also offers alternative scoring methods including the option to use the median instead of the mean. We discuss this option in the Supplementary Discussion where we compare results for COCOA of breast cancer DNA methylation using median and mean scoring methods, finding overall similar results and high correlation between median and mean scores (Additional file 2: Fig. S11, Additional file 1: Tables S1, S14). Other scoring options can be found in the software documentation.

Metric for quantifying variation

Choosing an appropriate metric can help to effectively capture the relationship between epigenetic variation and variation in the target variable (Supplementary Methods). In this paper, we used covariance, Pearson correlation, Spearman correlation, PCA, and MOFA (Argelaguet *et al.*, 2018) to quantify variation, but other variation metrics and dimensionality reduction techniques can be used with COCOA for quantifying inter-sample variation, depending on the specific circumstances of a given analysis. The only requirement is that the metric must provide a score for each epigenetic locus that quantifies how much it is associated with variation in the target variable. The choice of metric can depend on the data type.

For DNA methylation data, since DNA methylation data is bounded from 0 to 1, we used covariance to give greater weight to CpGs with larger changes in DNA methylation across samples. Since the range of ATAC-seq counts could be very different between different peaks, we used Pearson correlation for the ATAC-seq data in order to give each peak a comparable score, regardless of the peak's range. This principle also applies to PCA. When performing PCA, we recommend scaling the data by dividing each variable by its variance for ATAC-seq data (equivalent to correlation) but not for DNA methylation data (equivalent to covariation). Then, when treating the principal components as the target variables, we use the corresponding metric – covariance or correlation – to get the feature scores. We recommend Spearman correlation when the relationships between the target variable and the epigenetic features are monotonic but not linear, as may occur when the target variable is ordinal (e.g. cancer stage).

Permutation test

To assess statistical significance of the COCOA results, we use a permutation test. For both supervised and unsupervised COCOA analyses, we have a target variable (i.e. the sample phenotype or latent factor) and want to understand the relationship between the target variable and the epigenetic data. For a single permutation, we randomly shuffle the samples' target variable values then recalculate the association between the epigenetic data and the target variable as done in Step 1 (Fig. 1C). This gives each epigenetic feature an FCS for the shuffled target variable. Then we run COCOA on the new feature contribution scores to score each region set in the database. This process is repeated for each permutation. The COCOA scores for a given region set

from the permutations form a region set-specific null distribution. Because the sample labels were shuffled instead of the epigenetic data, the null distributions can appropriately capture the correlation structure of the epigenetic data, accounting for the correlation between epigenetic features in a given region set. The region set-specific null distributions also protect against false positives that could arise from some region sets being more fully covered by the epigenetic assay than others because each score in a region set's null distribution is created from the same coverage profile. To reduce the computational burden, we calculated 300 permutations and applied a permutation approximation technique (Winkler *et al.*, 2016). We fit a gamma distribution to each null distribution using the method of moments in the fitdistrplus R package (Delignette-Muller and Dutang, 2015) and then calculated a p-value for each region set using its gamma distribution. To test the appropriateness of fit of the gamma approximation, we ran a simulation study with 100,000 permutations, and then subsampled and applied the approximation to see how close the approximation is to the true p-value. Our conclusion is that the gamma approximation is accurate for high p-values, but the gamma approximation may overestimate the significance of low p-values; therefore, we advise that it can be helpful for screening out region sets that are not significant (Fig S12; further discussion in Supplementary Information). To correct p-values for the number of region sets tested, we used Benjamini-Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg, 1995) with an FDR of 5%.

Meta-region profile plots

To visualize results, COCOA produces a plot we call the *meta-region profile* plot (e.g. Fig 2C). The goal of the meta-region profile is to compare the feature contribution scores in the regions of interest to the surrounding genome to assess how specific the captured signal is to a region set. We combine information from all regions of the region set into a single summary profile as has been done for DNA methylation data Sheffield *et al.* (2017). Each region in the region set is expanded on both sides to include the surrounding genome (e.g. expanded to 14 kb total, centered on the region of interest). This enlarged region is then split into bins of approximately equal size. Finally, the FCS for corresponding bins from each region of the region set are averaged to get a single "meta-region" FCS profile. By default, the type of average depends on whether the data is single-base pair resolution or region-based, with the same algorithm applied as for the COCOA score. A peak in the middle of the profile suggests that there is variation that is specific to this region set.

Region set database

To annotate variation in the epigenetic data, we used a subset of the LOLA database (Sheffield and Bock, 2015) (filtered with R script, see Supplementary Materials), totaling 2246 region sets from public sources. Sources included the ENCODE project (Consortium, 2012; Davis *et al.*, 2017), Roadmap Epigenomics (Bernstein *et al.*, 2010; Kundaje *et al.*, 2015), CODEX database (Sánchez-Castillo *et al.*, 2014), and the Cistrome database (Mei *et al.*, 2016). Additionally, we included some region sets derived from JASPAR motif (Sandelin, 2004) predictions. Examples of region sets include transcription factor binding sites from ChIP-seq experiments, histone modification regions from ChIP-seq experiments, and cell type or condition-specific accessible chromatin from ATAC-seq experiments. For a discussion of how to choose a region set database and other related considerations, see "Additional file 2". For each analysis, we only considered in the results region sets that had at least 100 regions with any coverage by the epigenetic data. Since the CLL MOFA data was in reference genome hg19 and the breast cancer data was in hg38, we used the corresponding hg19 or hg38 version of the region set database when analyzing each dataset. A brief description of the region sets can be found in the supplementary data (Additional file 1: Tables S1-S4) and the database is available at http://databio.org/regiondb (Sheffield and Bock, 2015). To designate region sets "hematopoietic TFs" for Fig. 3, we did a literature search, selecting three reviews: one focusing on myeloid TFs (Rosenbauer and Tenen, 2007), one focusing on lymphoid TFs (Somasundaram *et al.*, 2015) and one general hematopoietic TF (Orkin, 1995). The hematopoietic TFs identified from these reviews are the following: RUNX1, TAL1, PU.1, CEBPA, IRF8, GFI1, CEBPE (Rosenbauer and Tenen, 2007), TCF3, EBF1, PAX5, FOXO1, ID2, GATA3 (Somasundaram *et al.*, 2015), KLF1, GATA1, GATA2, IKZF1, CMYB, and NFE2 (Orkin, 1995). Since GATA3 was also identified as an ER-related TF, we did not consider GATA3 as a hematopoietic TF in plots to avoid confusion.

Breast cancer analyses

Datasets

For the unsupervised breast cancer analyses, we used DNA methylation and ATAC-seq datasets from The Cancer Genome Atlas (TCGA). We retrieved the DNA methylation and clinical data with the TCGAbi-

olinks R package (Colaprico *et al.*, 2015). We identified 657 patients with both 450k DNA methylation data and known ER and progesterone status. For the ATAC-seq data, we retrieved a peak count matrix for the consensus set of breast cancer ATAC-seq peaks identified by Corces et al. from the following location: https://atacseq.xenahubs.net/download/brca/brca_peak_Log2Counts_dedup. We used a sample ID lookup table to match the ATAC-seq IDs to the standard TCGA identifiers: https://gdc.cancer.gov/about-data/publications/ATACseq-AWG. We excluded one patient of the 74 patients with ATAC-seq data (TCGA-AO-A0J5) for whom we did not have sufficient metadata.

Data processing and quantifying variation

For the breast cancer DNA methylation data, we excluded the sex chromosomes. For the ATAC-seq data, we used the peak count matrix from Corces et al. (Corces *et al.*, 2018), without further processing. We performed PCA on the DNA methylation data and the ATAC-seq data separately with the 'prcomp' R function, with centering and without scaling. PCA is used to get covariance of features and to prioritize the largest sources of covariance. After PCA, we calculated the covariance or correlation coefficient for each epigenetic feature with each latent factor to get a value that represented how much each feature contributed to each latent factor. We used covariation for the DNA methylation data and correlation for the chromatin accessibility data. To test the association of ER status with PC score, we used the Wilcoxon rank-sum test with ER positive samples and ER negative samples as the two groups.

Comparison of COCOA and chromVAR

To compare COCOA and chromVAR (Schep *et al.*, 2017), we completed two tests with the breast cancer ATAC-seq data. First, we applied chromVAR with the same region set database used by COCOA in our ATAC-seq analysis. Second, we applied COCOA and chromVAR with the main motif database used by chromVAR in its publication, which is a curated version of the cisBP database (Weirauch *et al.*, 2014) and is available as the "human_pwms_v1" data object from the "chromVARmotifs" R package that can be downloaded from the"GreenleafLab/chromVARmotifs" Github repository. We applied chromVAR to the normalized data from Corces et al., adding a pseudocount to bring the minimum normalized signal up to zero. To use the motif database with COCOA, we identified peaks with motif hits using the "matchMotifs" function from the "motifmatchr" R package (Schep, 2018) with default parameters and took those regions as a region set. The "matchMotifs" function is the method chosen by chromVAR authors for identifying motif matches in the chromVAR Bioconductor vignette. For the chromVAR figure, we designated motifs as AP-1-related based on an AP-1 review (Figure 1 of review) (Eferl and Wagner, 2003).

Multi-omics chronic lymphocytic leukemia analysis

Datasets

For the unsupervised multi-omics analysis, we used preprocessed data that was included with the MOFA R package, specifically the latent factors from the multi-omics dimensionality reduction analysis of 200 chronic lymphocytic leukemia (CLL) patients as described by Argelaguet et al. (Dietrich *et al.*, 2017; Argelaguet *et al.*, 2018). We retrieved the 450k DNA methylation data for these patients using the ExperimentHub R package (Bioconductor Package Maintainer <Maintainer@Bioconductor.Org>, 2017) (CLLmethylation data package, ExperimentHub ID: EH1071) (Dietrich *et al.*, 2017).

Data processing and quantifying variation

For the multi-omics analysis, we used the dimensionality reduction results from the paper by Argelaguet et al. and then extended the results to CpGs that were not included in the dimensionality reduction. The original multi-omics analysis used only the most variable 1% of CpGs (4,248 CpGs) for calculation of the latent factors. Since COCOA benefits from higher coverage of CpGs across the genome, we calculated the correlation of each CpG from the DNA methylation microarrays (excluding sex chromosomes) with each latent factor. This yielded a matrix with CpG, latent factor correlations where each row is a CpG and each column is a latent factor, which can be used as input to COCOA.

Kidney renal clear cell carcinoma analysis

Dataset

For the supervised KIRC analysis, we used DNA methylation and clinical data from The Cancer Genome Atlas. We used 450k DNA methylation microarray data for 318 patients, retrieved with the curatedTCGAData R

package (Marcel Ramos, 2017). The clinical data included cancer stage and survival information that was used to label samples in the supervised analysis.

Data processing and quantifying variation

For the supervised analysis of KIRC methylation, we first split the data into two groups: training (2/3 of patients) and validation (1/3 of patients), keeping approximately equal proportions of each cancer stage in each group. With the COCOA samples, we first calculated the Spearman correlation between the DNA methylation levels and the sample phenotype of interest, cancer stage. This resulted in a correlation coefficient for each CpG. We then applied the COCOA algorithm on the absolute correlation coefficients.

Validation and survival analysis

After running COCOA on 2/3 of the samples, we did validation analyses on the remaining 1/3 of samples. First, we tested whether each patient's average DNA methylation level in the top EZH2 region set from COCOA was correlated with cancer stage, using the 'cor.test' R function (R Core Team, 2018) and Spearman correlation. To calculate correlation p-values for the null hypothesis that the correlation was zero, we used an asymptotic t approximation, the default method used by the 'cor.test' function. To calculate the average methylation, we first separately averaged DNA methylation within each EZH2 region, then averaged all the region averages. We also tested whether average DNA methylation in EZH2 regions was related to overall patient survival. We created Kaplan-Meier curves with two groups: the 25% of validation samples with highest DNA methylation in EZH2 regions and the 25% of samples with the lowest DNA methylation. We used a log-rank test from the 'survminer' R package's 'ggsurvplot' function (Kassambara *et al.*, 2019) to get a p-value for the Kaplan-Meier curves. We created a Cox proportional hazards model with all validation samples, relating average DNA methylation in EZH2 regions to patient survival and correcting for age, gender, and average genome methylation level. We also tested the two highest scoring TF region sets from the COCOA analysis – JUND and TCF7L2 -- for association with cancer stage and survival using the methods described above. We tested whether variables satisfied the proportional hazards assumption using the 'cox.zph' function in R (Grambsch and Therneau, 1994; Therneau, 2015; Terry M. Therneau and Patricia M. Grambsch, 2000) (Additional file 4: Table S6), considering variables with $p < 0.05$ as not satisfying the assumption. The JUND validation model did not meet the assumption for the variable of interest (average DNA methylation in EZH2/SUZ12-binding regions) and therefore was not considered.

Pan-cancer EZH2 analysis

In this analysis, we tested whether average DNA methylation level in EZH2-binding regions would be associated with cancer stage and patient survival in other cancer types than KIRC. We combined regions from the top group of 11 EZH2 and SUZ12 region sets from the KIRC analysis (Fig. 5A, Supplementary Data) to make a single "master" EZH2/SUZ12 region set (referred to as EZH2-binding regions). We took the union of all regions and merged regions that overlapped. We downloaded DNA methylation microarray data for 33 TCGA cancer types using the curatedTCGAData R package (Marcel Ramos, 2017). Then, for each sample, we calculated the average DNA methylation level in EZH2-binding regions. For each cancer type for which we had cancer stage information (21/33), we calculated the Spearman correlation between average EZH2-binding region DNA methylation and cancer stage, using the 'cor.test' R function (R Core Team, 2018). To calculate correlation p-values for the null hypothesis that the correlation was zero, we used an asymptotic t approximation, the default method used by the 'cor.test' function. Next, for each cancer type, we used a Cox proportional hazards model to test the association of average EZH2-binding region DNA methylation with survival, with the covariates patient age, sex, and average microarray-wide DNA methylation level as available. We tested whether variables satisfied the proportional hazards assumption using the 'cox.zph' function in R (Grambsch and Therneau, 1994; Therneau, 2015; Terry M. Therneau and Patricia M. Grambsch, 2000) (Additional file 7: Table S9). We considered variables with $p < 0.01$ as not satisfying the assumption, picking a more stringent cutoff because more models were tested. Models that did not meet the assumption for the variable of interest (average DNA methylation in EZH2/SUZ12-binding regions) were removed, in our case only one cancer type– low grade glioma (LGG). We corrected Spearman and Cox p-values for multiple testing using the Holm-Bonferroni method (Holm, 1979).

Availability of data and materials

The R scripts used for this analysis can be accessed at https://github.com/databio/COCOA_paper. The COCOA package can be accessed at http://bioconductor.org/packages/COCOA. An archived version of both

is also available (DOI: 10.5281/ZENODO.3973375). These are released under the open source BSD-3-Clause license.

We retrieved the BRCA DNA methylation and clinical data with the TCGAbiolinks R package (Colaprico *et al.*, 2015). We retrieved the KIRC DNA methylation and clinical data with the curatedTCGAData R package (Marcel Ramos, 2017).

For the BRCA ATAC-seq data, we used the peak count matrix from Corces et al. (Corces *et al.*, 2018), which we retrieved from https://atacseq.xenahubs.net/download/brca/brca_peak_Log2Counts_dedup. We used a sample ID lookup table to match the ATAC-seq IDs to the standard TCGA identifiers: https://gdc.cancer.gov/about-data/publications/ATACseq-AWG.

For the CLL multi-omics analysis, we retrieved the 450k DNA methylation data for these patients (Dietrich *et al.*, 2017) using the ExperimentHub R package (Bioconductor Package Maintainer <Maintainer@Bioconductor.Org>, 2017) (CLLmethylation data package, ExperimentHub ID: EH1071). We retrieved the MOFA latent factors and patients' mutation info from the MOFAdata R package (Ricard Argelaguet and Britta Velten and Damien Arnol and Florian Buettner and Wolfgang Huber and Oliver Stegle, 2020): https://bioconductor.org/packages/release/data/experiment/html/MOFAdata.html

To create hematopoietic chromatin accessibility region sets, we retrieved an ATAC-seq count matrix (GSE74912_ATACseq_All_Counts.txt.gz) from Gene Expression Omnibus with hematopoietic ATAC-seq data from Corces et al. (Corces *et al.*, 2016), which was further processed as described in the Methods section. The remaining region set database can be downloaded at http://databio.org/regiondb. The chromVAR motif database was retrieved from https://github.com/GreenleafLab/chromVARmotifs.

Acknowledgements

The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Funding

This work was supported by the National Institute of General Medical Sciences grant GM128636 (NCS). JTL was partially supported by an NIH training grant (NLM; 5T32LM012416) and the UVA Cancer Center. JPS was partially supported by an NIH training grant (5T32GM813633). FEG-B was supported by the UVA Cancer Center through the NCI Cancer Center Support Grant P30 CA44579 and a V-foundation scholar grant (V2017-013).

Author information

Contributions

JTL led the software development with contributions from JPS and NCS. JTL, FEG-B, and NCS contributed to the writing of the manuscript. FEG-B, SB, and NCS contributed technical expertise. All authors approved the final manuscript.

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Supplementary information

Additional file 1. Tables S1-S4, Tables S11-S12, Tables S14-S15

COCOA region set scores for the corresponding analysis. Table S1: BRCA DNA methylation data, Table S2: BRCA ATAC-seq data, Table S3: CLL multi-omics analysis, Table S4: KIRC cancer stage analysis, Table S11: Results for simulated DNA methylation data with low noise level, Table S12: Results for simulated DNA methylation data with high noise level, Table S14: BRCA DNA methylation data with median scoring method, Table S15: BRCA DNA methylation data when using loadings as FCSs.

Additional file 2.

Supplemental figures S1-S12, supplemental methods, and supplemental discussion.

Additional file 3. Table S5.

Spearman correlation between average DNA methylation in each region set and KIRC cancer stage for training and validation data.

Additional file 4. Table S6.

Results from Cox proportional hazards model for the three training models and three validation models that tested the relationship of average DNA methylation in each region set with overall patient survival.

Additional file 5. Table S7.

Spearman correlation between average DNA methylation in EZH2/SUZ12 region set and cancer stage for each cancer type.

Additional file 6. Table S8.

Association between average DNA methylation level in EZH2/SUZ12-binding regions and cancer stage or patient survival.

Additional file 7. Table S9.

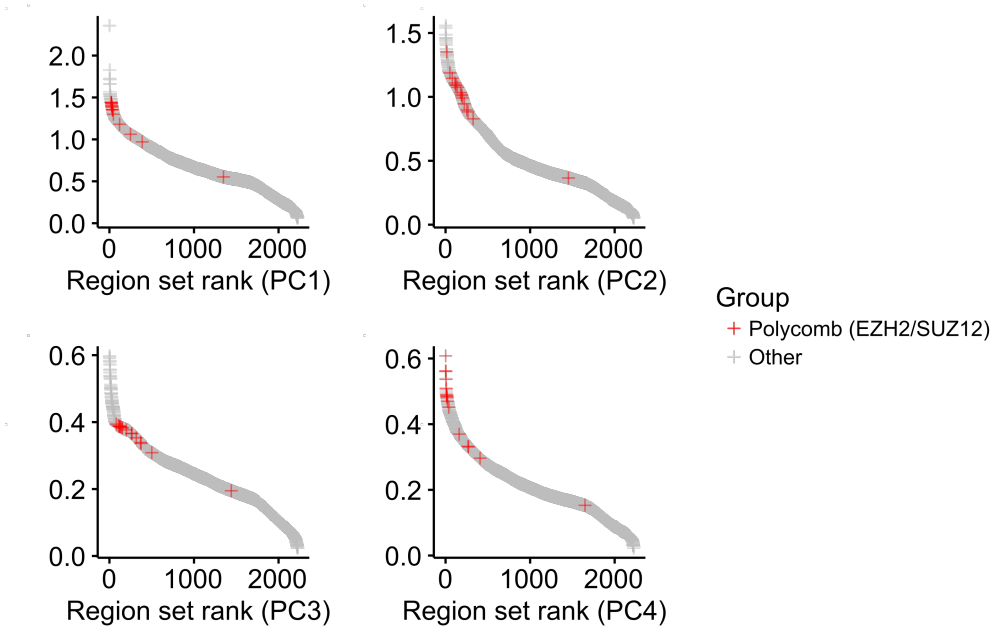
Results from Cox proportional hazards model for each cancer type that tested the relationship of average DNA methylation in the EZH2/SUZ12 region set with overall patient survival.

Additional file 8. Table S10.

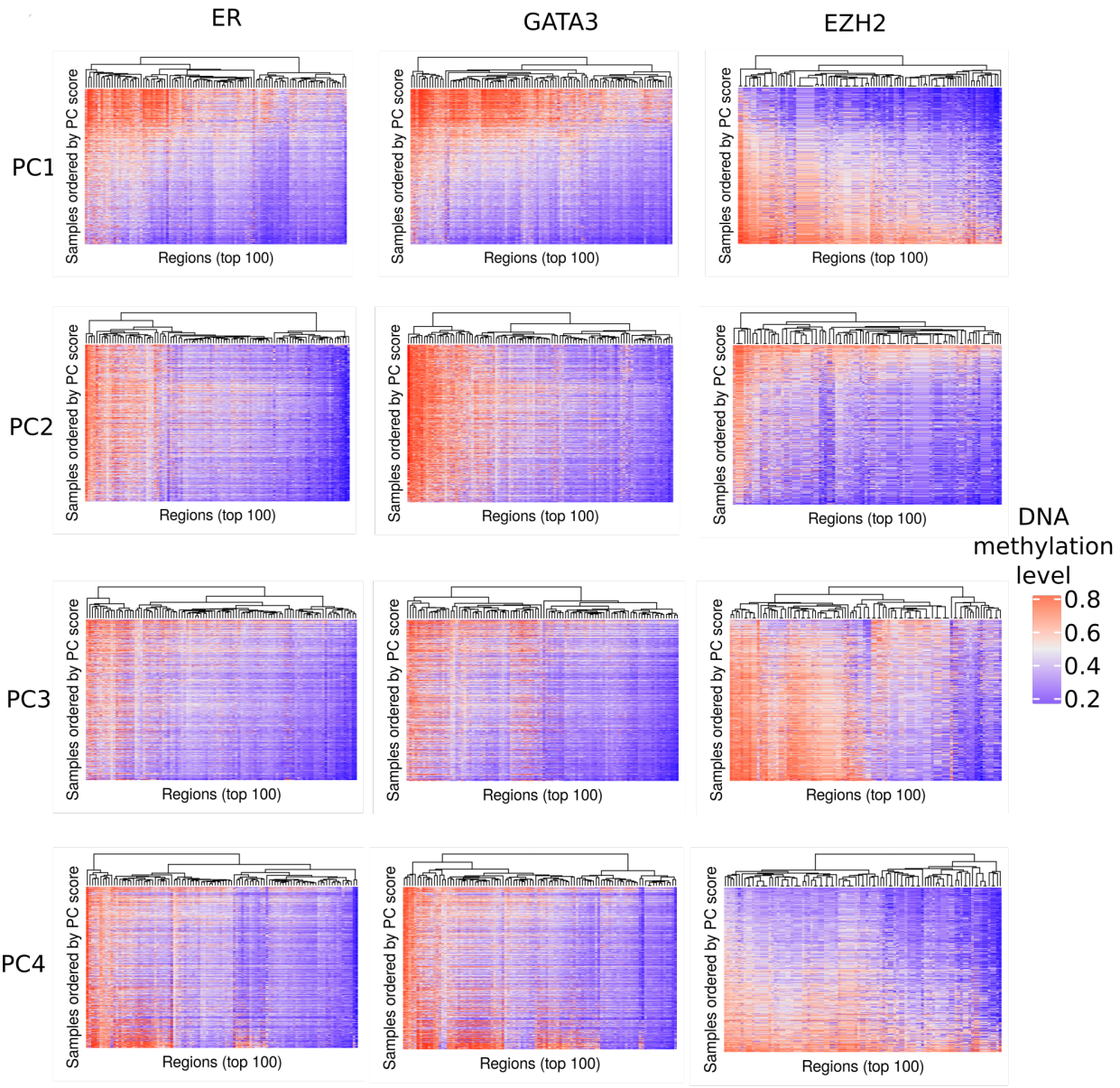
LOLA results from differentially methylated regions from simulated DNA methylation data with low noise.

Additional file 9. Table S13.

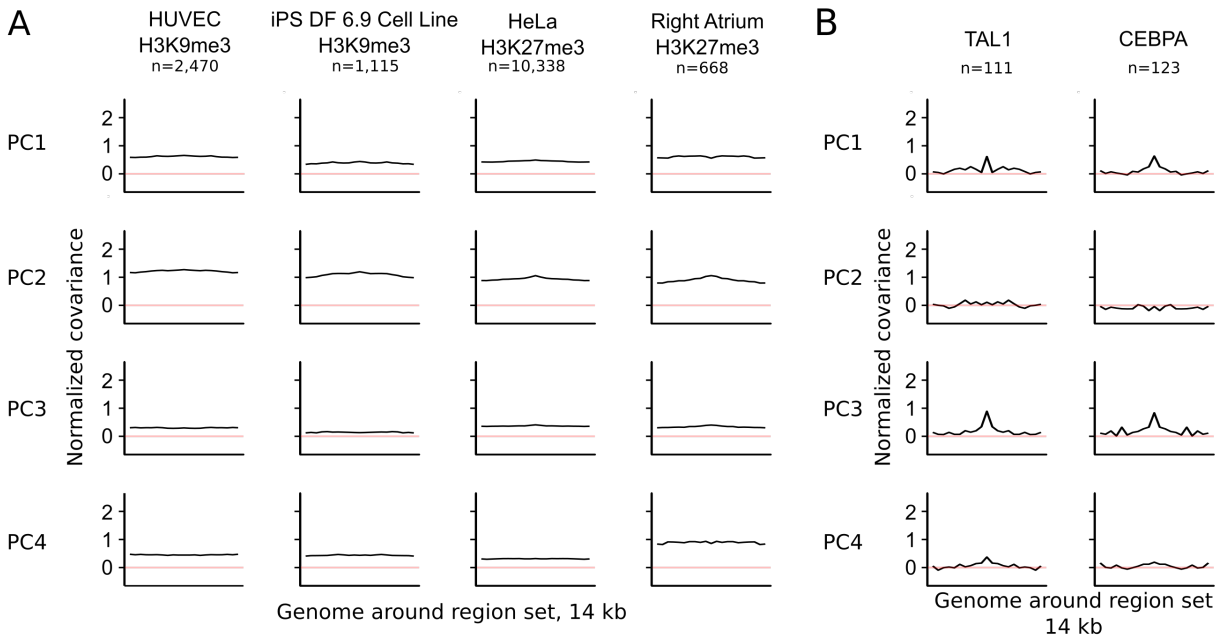
COCOA and chromVAR results for breast cancer ATAC-seq data using chromVAR motif database.



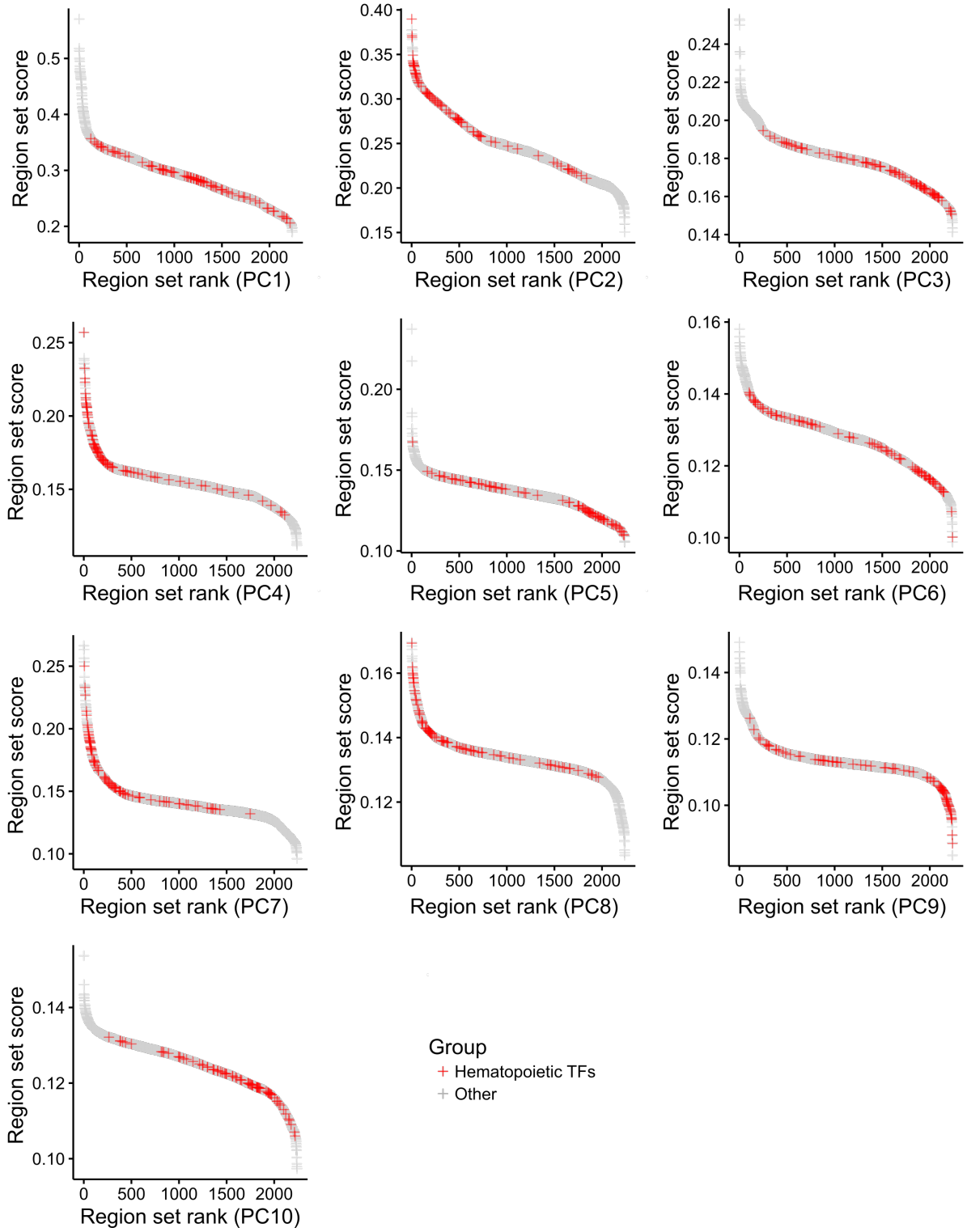
Dissertation figure 9: **Fig. S1. Region set scores for PCs 1-4 for the BRCA DNA methylation data.** This figure is included with only the polycomb group marked to allow clearer visualization of the polycomb region set group in comparison to Fig. 1 where several region set groups are marked.



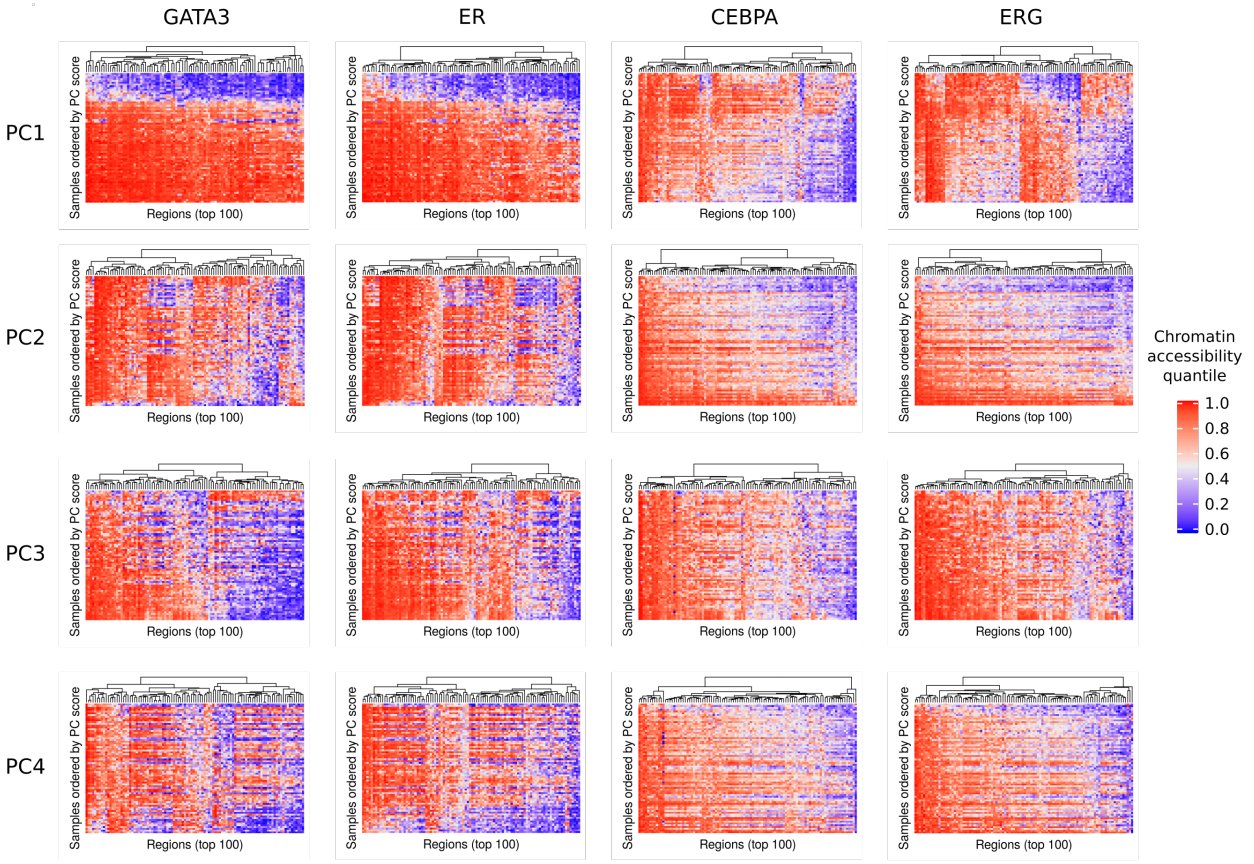
Dissertation figure 10: **Fig. S2. DNA methylation in some of the top scoring region sets for principal component 1.** Average DNA methylation levels are shown for the 100 regions from each region set that had the highest absolute FCS for each PC. Patients are ordered by PC scores.



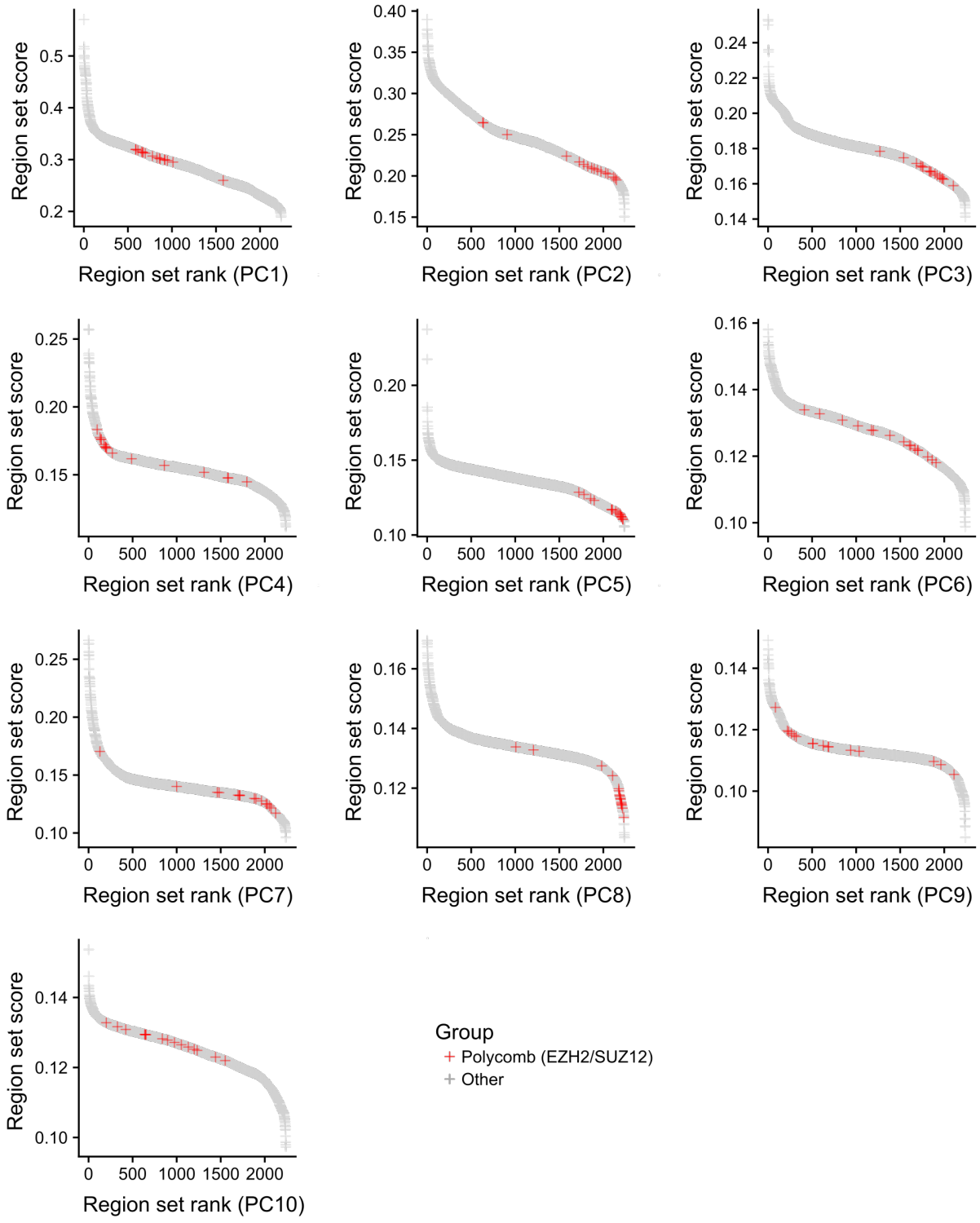
Dissertation figure 11: **Fig. S3. Meta-region profiles for region sets from the COCOA analysis of breast cancer DNA methylation data.** A. Profiles for the highest scoring H3K9me3 and H3K27me3 region sets from PC2. B. Profiles for the two highest scoring hematopoietic TFs in PC3. A peak in the center of the meta-region profile indicates that the DNA methylation level covaries with the PC more at the region of interest than in the surrounding genome. Profiles have been normalized to the mean and standard deviation of the covariance of all cytosines for each PC. The number of regions from each region set that were covered by the epigenetic data in the COCOA analysis (Fig. 2, panel A) is indicated by "n".



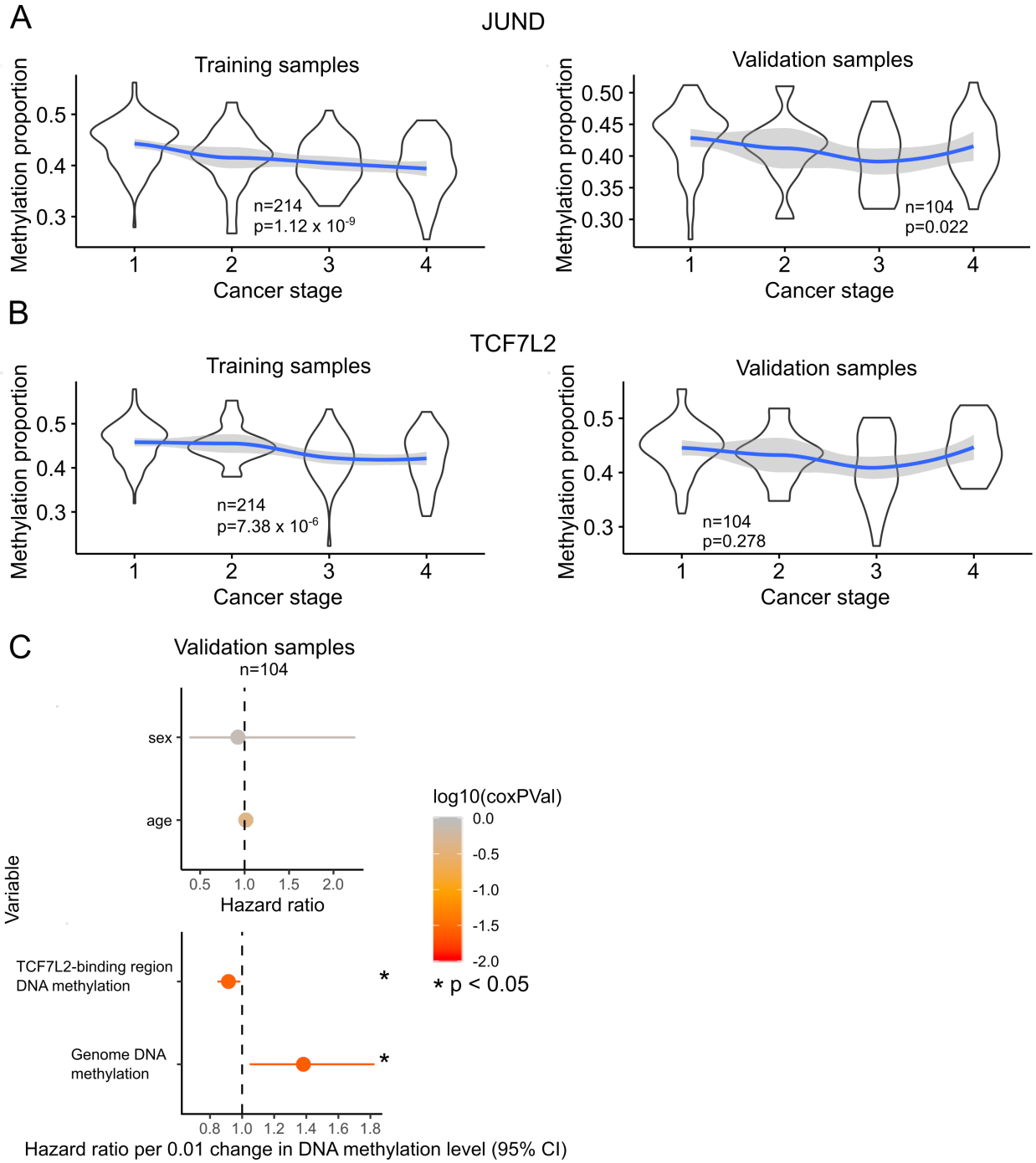
Dissertation figure 12: **Fig. S4. Hematopoietic transcription factor region sets have high scores for several of the top principal components.** Region set scores for each of the first 10 principal components of the BRCA ATAC-seq data.



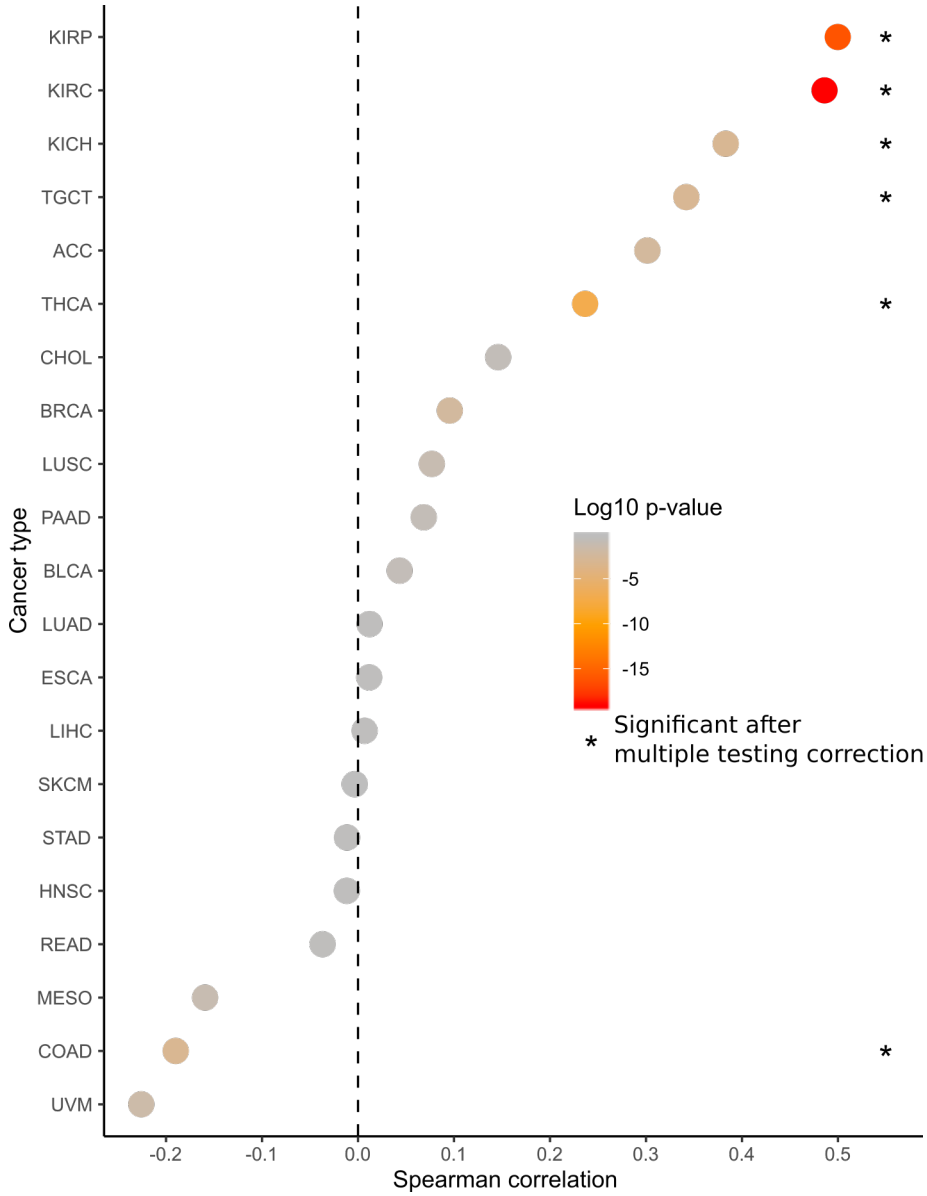
Dissertation figure 13: **Fig. S5. Chromatin accessibility signal in some of the top scoring region sets from COCOA analysis of breast cancer ATAC-seq data.** GATA3 and ER were the top scoring region sets for PC1 while CEBPA and ERG were the top scoring region sets for PC2. Average chromatin accessibility quantiles are shown for the 100 regions from each region set that had the highest absolute FCS for each PC. Patients are ordered by PC scores.



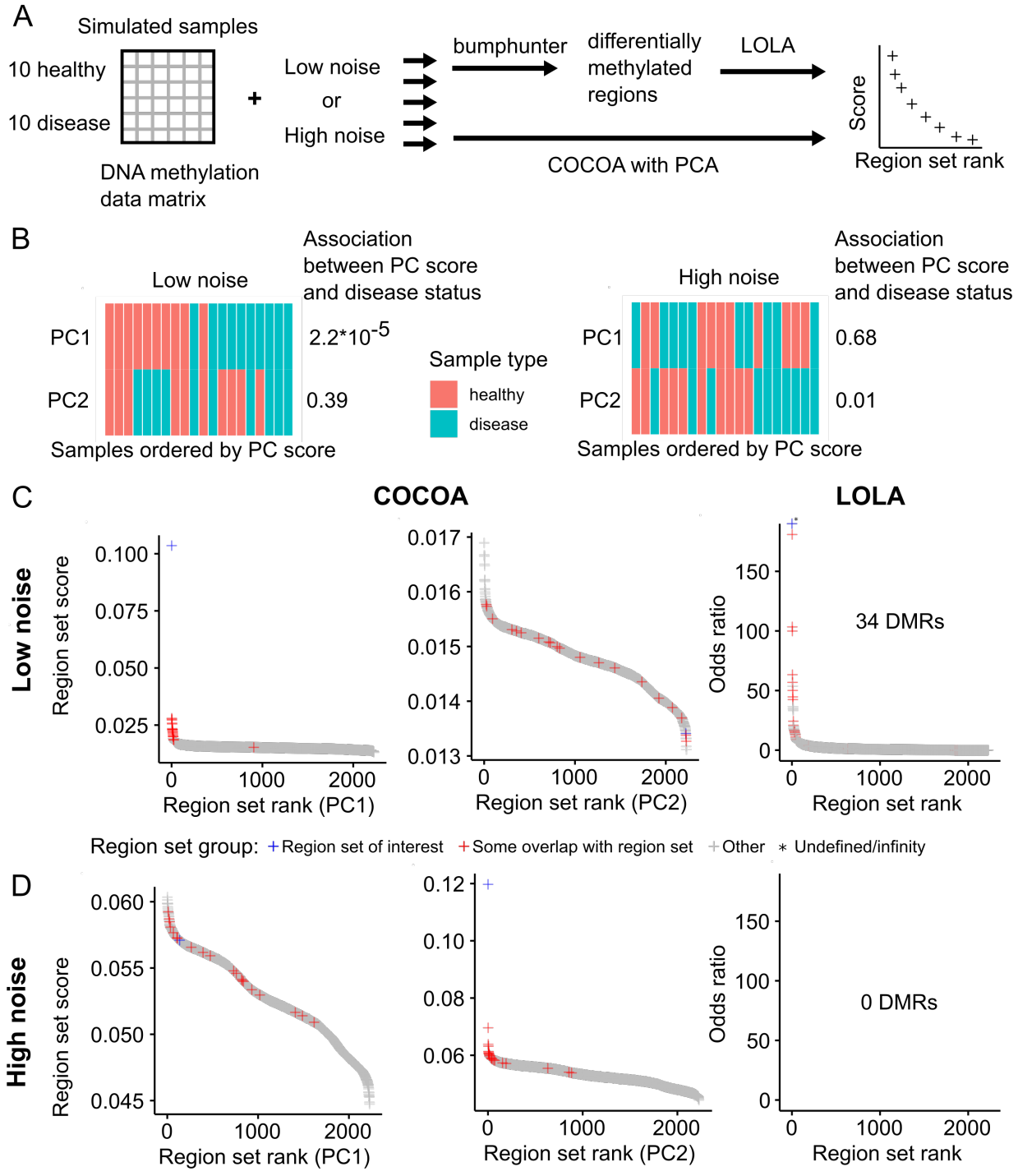
Dissertation figure 14: **Fig. S6. Region set scores for each of the first 10 principal components of the BRCA ATAC-seq data, with polycomb region sets (EZH2/SUZ12-binding regions) indicated.**



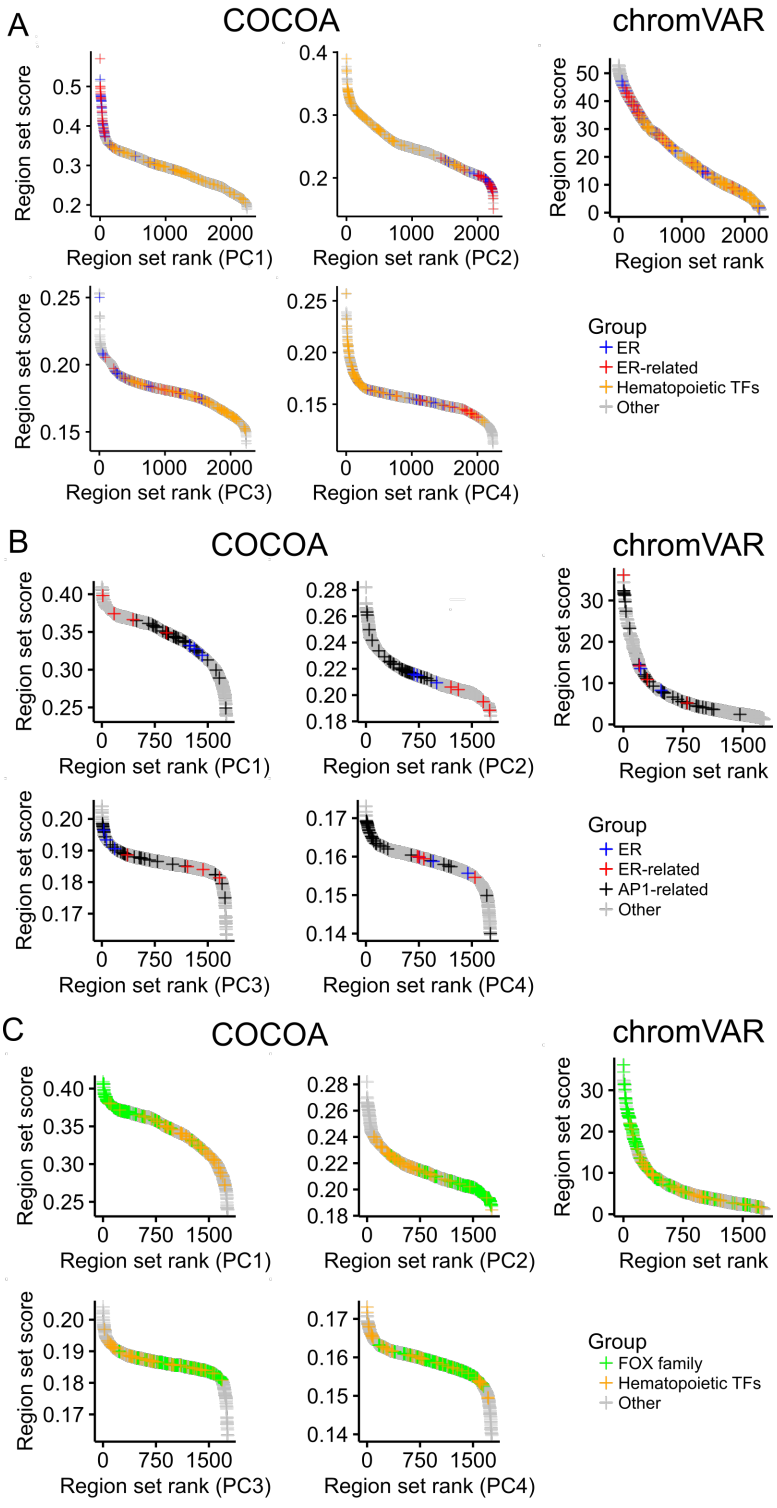
Dissertation figure 15: **Fig. S7. Association of average DNA methylation level in JUND and TCF7L2-binding regions with KIRC cancer stage and overall survival.** A. The Spearman correlation of cancer stage with the average DNA methylation in JUND-binding regions. The JUND region set used is the highest scoring transcription factor region set from the KIRC COCOA analysis. The JUND Cox proportional hazards model did not meet the proportional hazards assumption and is therefore not included in the figure. B. The Spearman correlation of cancer stage with the average DNA methylation in TCF7L2-binding regions. The TCF7L2 region set used is the second highest scoring transcription factor region set from the KIRC COCOA analysis. C. Hazard ratios for Cox proportional hazards model of the association between overall patient survival and average DNA methylation in TCF7L2-binding regions.



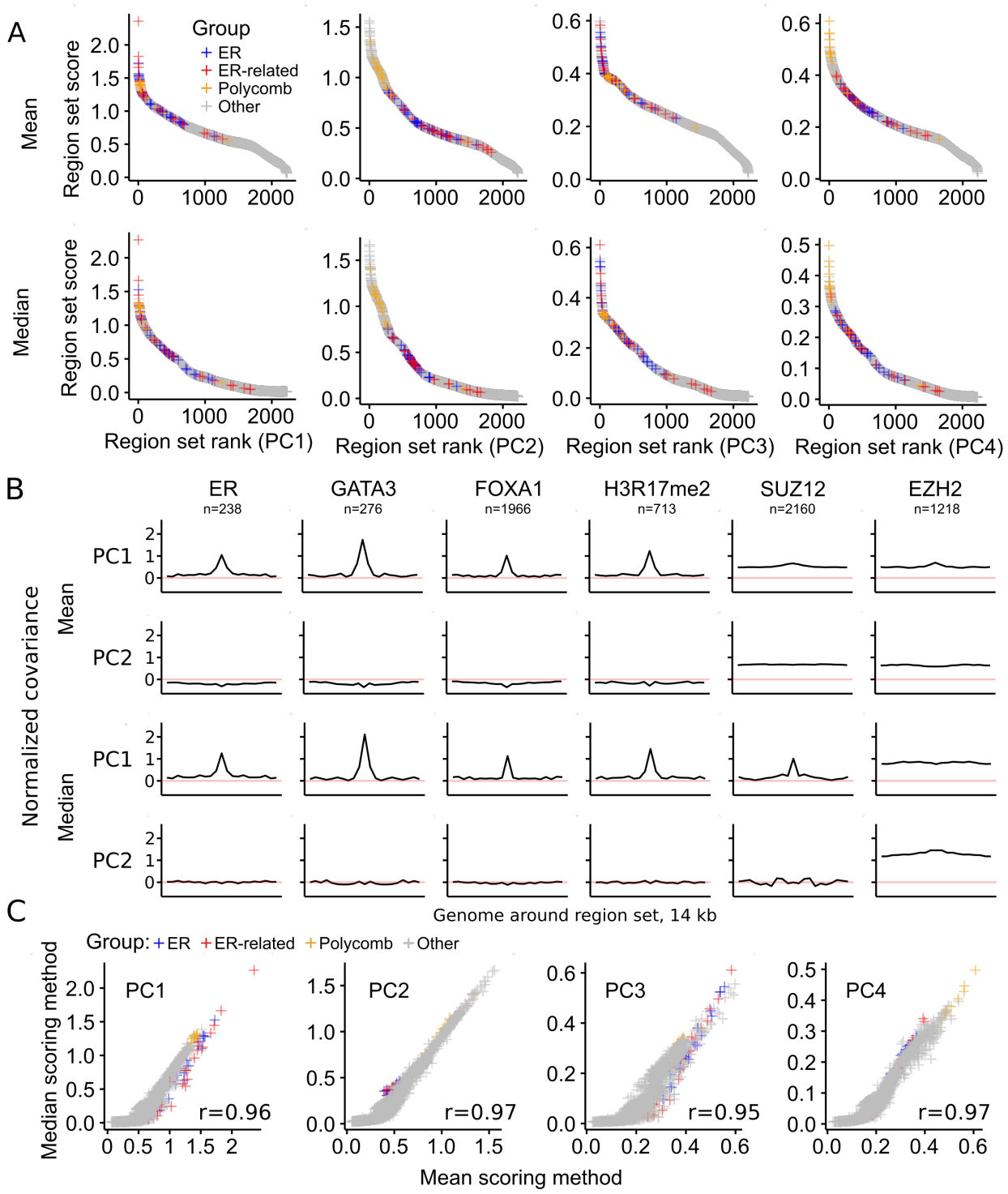
Dissertation figure 16: **Fig. S8. Correlation between average EZH2/SUZ12-binding region DNA methylation and cancer stage.** Color is based on the raw Spearman p-values and asterisks mark significant correlations after Holm-Bonferroni correction to account for testing 21 cancer types.



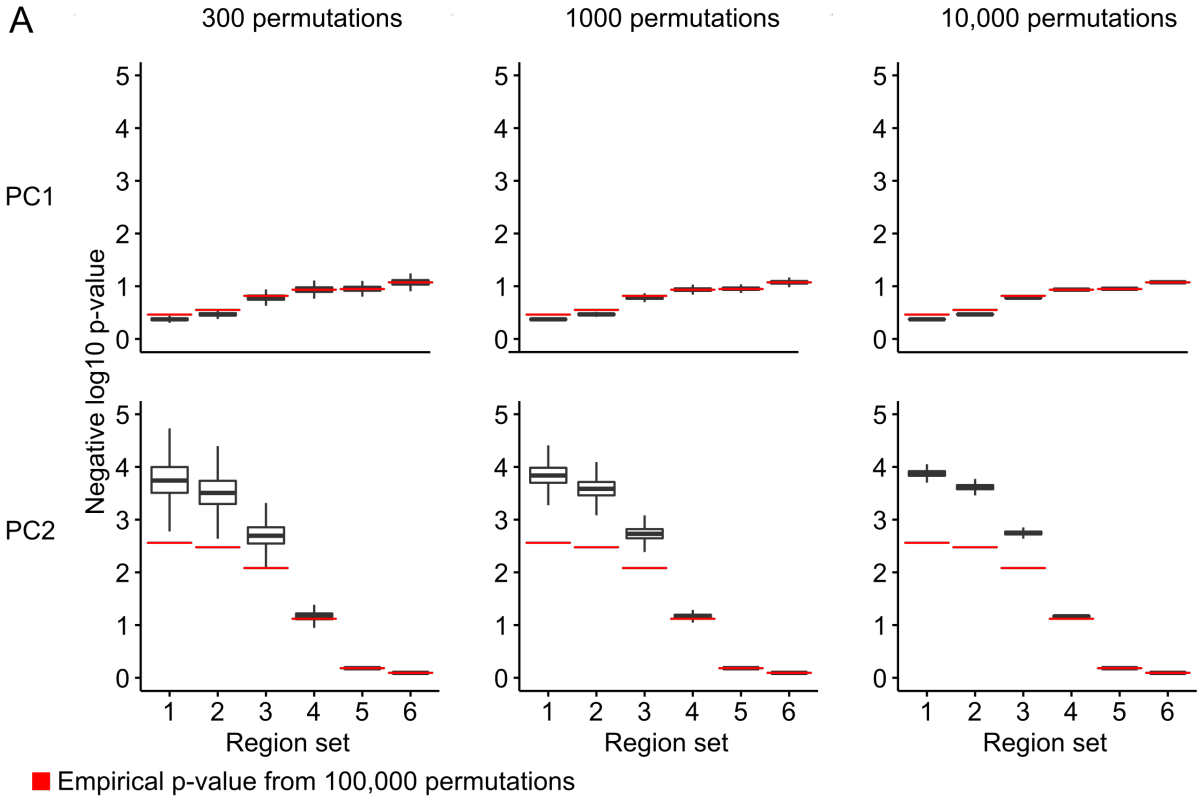
Dissertation figure 17: **Figure S9. Comparison of COCOA and LOLA.** A. The workflow for comparison of the methods. B. Association of PC scores with disease status. For low noise, PC1 is associated with disease status but for high noise, PC2 is associated with disease status (Wilcoxon rank-sum test). C. Results with a low level of noise added to samples. The COCOA score or LOLA odds ratio for each region set, ordered from highest to lowest. C. Results with a high level of noise added to samples. The COCOA score or LOLA odds ratio for each region set, ordered from highest to lowest. There are no scores for LOLA because bumphunter did not identify any significant DMRs (FDR <= 0.05).



Dissertation figure 18: **Fig. S10. Comparison of COCOA and chromVAR on breast cancer ATAC-seq data.** A. COCOA and chromVAR scores for the region set database (see "Region set database" in methods). The chromVAR score for a region set is the standard deviation of all samples' chromatin accessibility z-scores for that region set. The ER-related region set group includes FOXA1, GATA3, and H3R17me2. For definition of the hematopoietic TF group, see "Region set database" in methods. B. COCOA and chromVAR scores for a curated version of the cisBP motif database. The ER-related region set group includes FOXA1 and GATA3. For the definition of the AP1-related group, see "Comparison of COCOA and chromVAR" in methods. C. The same COCOA and chromVAR scores for a curated version of the cisBP motif database but indicating FOX family motifs and hematopoietic TF motifs.



Dissertation figure 19: **Figure S11. Comparison of median and mean scoring methods.** A. The COCOA score for each region set, ordered from highest to lowest. The ER-related group includes GATA3, FOXA1, and H3R17me2. The polycomb group includes EZH2 and SUZ12. B. Meta-region profiles of several of the highest scoring region sets from the breast cancer analysis. Meta-region profiles show covariance between PC scores and the epigenetic signal in regions of the region set, centered on the regions of interest. The number of regions from each region set that were covered by the epigenetic data in the COCOA analysis (panel A) is indicated by "n". The line at zero marks the mean or median respectively of the FCS for each PC. C. The relationship between region set scores for each scoring method. The Spearman correlation is shown.



Dissertation figure 20: **Figure S12. Comparison of empirical p-values to gamma distribution p-value approximation.** COCOA was run on PC1 and PC2 of PCA of simulated data with six region sets. The empirical p-values from 100,000 permutations are shown. P-values were also calculated with a gamma distribution approximation after sampling either 300, 1000, or 10,000 permutations from the 100,000 that were calculated. 500,000 such samples were taken to get a distribution of gamma p-values for each region set (outliers not shown).

Chapter Three: DNA methylation subtypes in acute myeloid leukemia

The contents of this chapter are in preparation for an upcoming submission.

Introduction

In this study, we characterize DNA methylation subtypes in aged AML patients (aAML). We use data from the Eastern Cooperative Oncology Group 3999 clinical trial #NCT00046930 (Cripe *et al.*, 2010). This study included 449 patients age 60 or older who had AML or high-risk myelodysplastic syndrome. It tested the drug zosuquidar which modulates P-glycoprotein. The study concluded that zosuquidar did not improve outcome. In this paper, we will use DNA methylation and mutation data from a subset of these patients. We use consensus clustering to identify robust clusters suggesting that epitypes identified depends on the regions considered. Through this, we identify epigenetic signatures that can yield valuable prognostic and mechanistic information.

Results

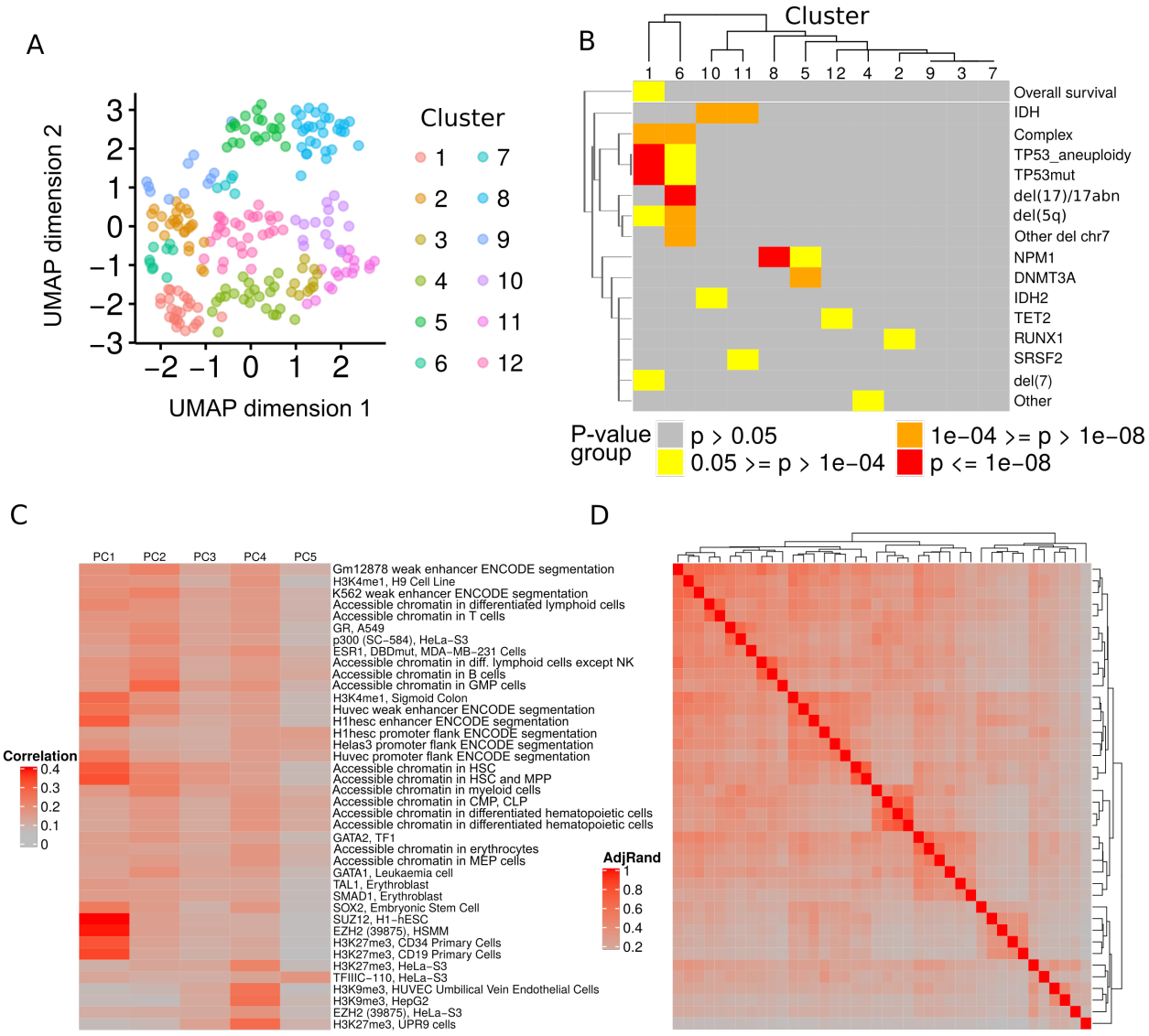
Annotation of DNA methylation variation identified variation in regulatory regions

We performed consensus clustering on the 25% most variable CpGs shared by all samples to segregate patient specimens into general sample "epitypes" (subtypes) (Fig. 1A). We found similar epitypes to those previously described (Figueroa, Lugthart, *et al.*, 2010; Glass *et al.*, 2017; Giacopelli *et al.*, 2021), with most clusters associated with specific somatic mutations or cytogenetics events (Fig. 1B).

We used the Coordinate Covariation Analysis (COCOA) method to annotate DNA methylation variation between patients using region sets. We found that the top components of variation in the data were associated with hematopoietic and polycomb-related region sets (Fig. 1C). This is consistent with previous literature that suggests certain subtypes have different states of hematopoietic differentiation (Giacopelli *et al.*, 2021).

Consensus clustering reveals non-exclusive sample epitypes

We hypothesized that DNA methylation subtypes might depend on the regions considered and that there might be multiple non-exclusive groupings of patient specimens that each reflect different aspects of epigenetic regulation. Therefore, we focused on region sets that showed the most inter-sample variation from the COCOA analysis to further define DNA methylation subtypes. For each top region set, we performed consensus clustering using subsampling and multiple dimensionality reduction and clustering methods to group samples into robust clusters. We show that the grouping of samples does depend on the regions considered, with very different possible groupings, as quantified by the adjusted Rand index (Fig. 1D). While many studies define a single clustering of samples as the DNA methylation subtype, we believe this definition is incomplete. It may be more helpful to allow samples to have multiple non-exclusive DNA methylation epitypes, depending on the regions examined, that may each give independent information about gene regulatory activity in those samples.

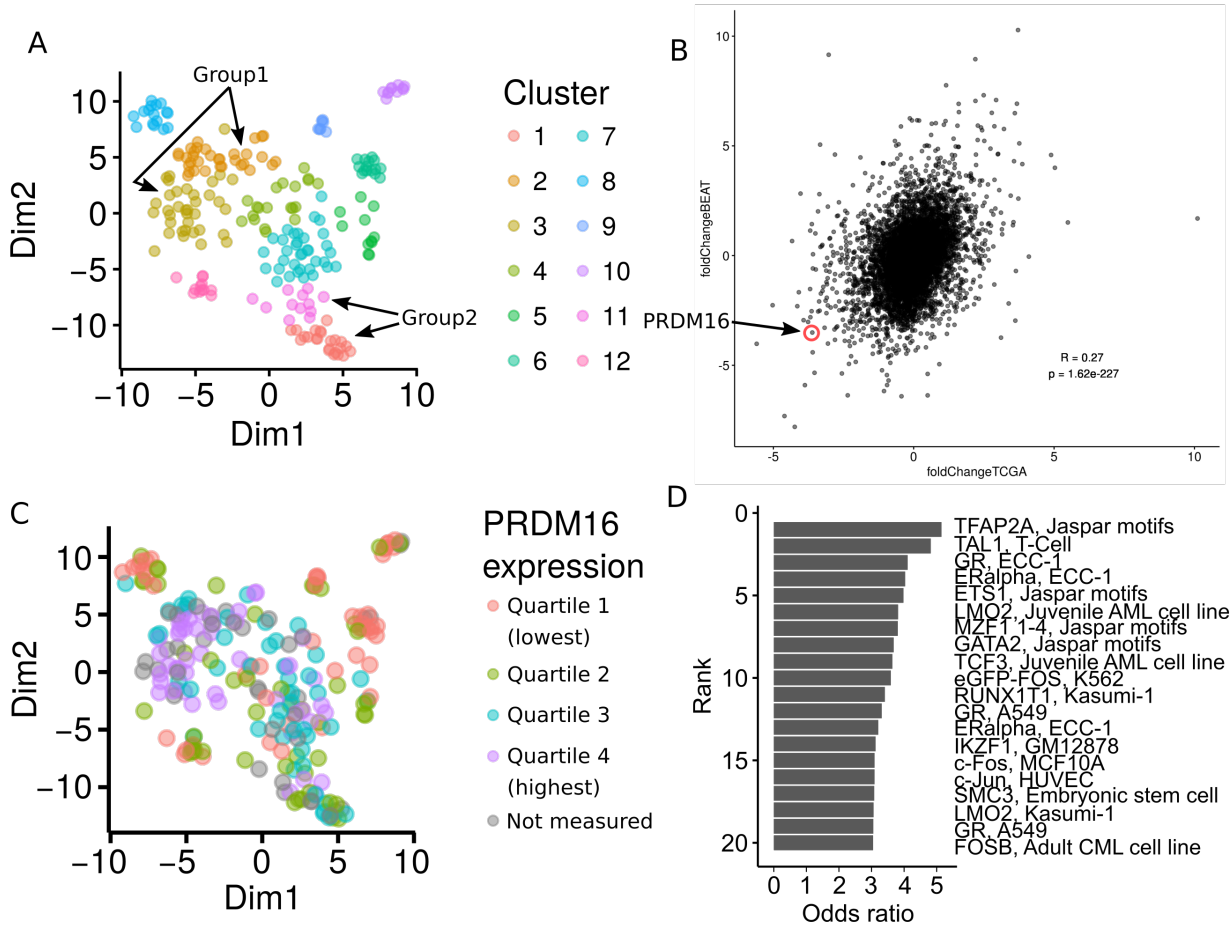


Dissertation figure 21: **Figure 1. DNA methylation subtypes and variation in aged AML.** A. DNA methylation subtypes visualized with UMAP dimensionality reduction. B. Mutation composition of each cluster. C. COCOA correlation score of DNA methylation in the top COCOA region sets for the first five PCs. D. Adjusted Rand index quantifying the similarity of consensus clustering results using the CpGs in each of the top COCOA region sets. Region sets are in the same order as C.

Cell type DNA methylation influences DNA methylation profiles

Multiple highest ranking region sets from the COCOA analysis relate to hematopoietic cell type. Previous studies have characterized AML samples by their similarity to hematopoietic stem cells (HSCs) or monocytes (Giacopelli *et al.*, 2021). We observed DNA methylation variation in HSC regions and myeloid regions, consistent with previous reports (Giacopelli *et al.*, 2021). Since we observed DNA methylation variation in lymphoid accessible chromatin regions in our COCOA analysis, we further examined whether some AML samples have a DNA methylation profile that is more similar to lymphoid cells than others. To clarify this, we selected distinct clusters that were enriched for NPM1 mutations based on the DNA methylation in the lymphoid regions and compared them. There were two main groups: the DNMT3Amut/NPM1mut cluster and the NPM1mut, IDH or TET mutated cluster. We calculated the differentially methylated regions between the NPM1 mutated samples in these clusters. Furthermore, we identified the corresponding clusters in two other cohorts, the BEAT AML and TCGA cohorts (Fig. 2A), and did differential methylation and gene expression analysis on those clusters for each cohort. The gene expression analysis fold changes had a high correlation between the BEAT and TCGA cohorts (Fig. 2B).

These comparisons revealed differential DNA methylation and expression related to a network of TFs involved in differentiation and proliferation. The PRDM16 gene encodes a key regulatory protein that is involved in HSC function and can form a fusion protein that can cause AML (Aguilo *et al.*, 2011; Gudmundsson *et al.*, 2020). PRDM16 had increased expression in the DNMT3Amut/NPM1mut group in both TCGA and BEAT cohorts (Fig. 2B, C). TET-mediated DNA demethylation of this gene's promoter has been suggested to control its expression in response to alpha-ketoglutarate level (Yang *et al.*, 2016). We found differential methylation of PRDM16 between the two groups in e3999, BEAT, and TCGA cohorts. DMRs were also enriched for the binding regions of PRDM16-related TFs including SPI1 which has been shown to be under regulatory control of PRDM16 (Deneault *et al.*, 2013; Hu *et al.*, 2019) and FOS (Deneault *et al.*, 2013; Gudmundsson *et al.*, 2020) (Fig. 2D).



Dissertation figure 22: **Figure 2. Non-exclusive epitypes give insight into epigenetic dysregulation.** A. t-SNE dimensionality reduction of CpGs in regions accessible in lymphoid cells but not myeloid cells for the BEAT AML cohort. Cluster membership was identified by consensus clustering. The two groups chosen for comparison are indicated. B. Differential gene expression fold change values are plotted for the comparison of Group1 and Group2 in the BEAT and TCGA cohorts. For the TCGA cohort, the corresponding clusters to Group1 and Group2 were chosen based on consensus clustering of the TCGA DNA methylation data. The correlation was calculated with Spearman correlation. C. The t-SNE plot from panel A, colored by PRDM16 gene expression quartile. D. LOLA results for the BEAT cohort. We used the LOLA software to annotate differentially methylated CpGs from the Group1/Group2 comparison.

Methods

Study data and quality control

All patients enrolled in ECOG-ACRIN clinical trial NCT00046930 (study cohort) (Cripe *et al.*, 2010) signed informed consent according to the declaration of Helsinki for collection and use of sample materials in research protocols. Study protocols were approved by the Institutional Review Boards at all participating institutions. Information on treatments received were reported in Cripe et al. (Cripe *et al.*, 2010) and will be included in metadata submitted to the NCTN. Information about treatments received after the clinical trial (including stem cell transplantation) was not available. Mononuclear cells were isolated centrally on the day after collection and viably frozen in a 90:10% mixed solution of heat inactivated fetal bovine serum: dimethyl sulfoxide. De-identified specimens (n=239) were then provided to the study investigators. Samples were thawed and depleted of lymphocytes (CD3+ and CD19+ cells) using magnetic beads (Miltenyi Biotec, order No. 130-050-101 and 130-050-301 respectively) using the depleteS program on an autoMACS Pro Separator (Miltenyi Biotec, Bergisch Gladbach, Germany). DNA was isolated from the myeloid cell fraction using Qiagen's Gentra Puregene Kit (Cat. No. 158667) per manufacturer's recommendation. Quality control was performed using standard agarose gel visualization and DNA was quantified using ThermoScientific's Qubit 2.0 Fluorometer and QubitTMdsDNA HS Assay Kit (ThermoScientific, Cat. No. Q32854). Specimen processing was performed and coordinated by Francine Garrett-Bakelman and Ria Kleppe. ERRBS was performed as previously described (Garrett-Bakelman *et al.*, 2015) on DNA from 239 patient specimens and 10 age-matched controls (CD34+ bone marrow cells) by Caroline Sheridan and Francine Garrett-Bakelman. Sequencing was performed on an Illumina Hiseq 2500 with 50 base pair single reads.

In ERRBS, the first three bases of each read often have low complexity (Garrett-Bakelman *et al.*, 2015). Therefore, a 'dark sequencing' ERRBS approach (Boyle *et al.*, 2012) was used, in which the first three bases of the read are not used for cluster localization and instead bases four through seven are used. Then, the read is sequenced again to get the first three bases (Boyle *et al.*, 2012). We processed the 'dark sequencing' reads with a custom script to put together each full 50 base pair read. We aligned reads to the hg38 genome with the Bismark software (Krueger and Andrews, 2011). We generated DNA methylation counts with the "epilog" software (https://github.com/databio/epilog).

We screened DNA methylation samples based on their median coverage, requiring a median of at least 33 reads/CpG. We limited analysis to CpGs that had coverage of at least 10 reads in every sample, resulting in a 1,960,684 CpGs and 221 AML samples.

Mutation data for study patients was generated from exome capture as described in Rapaport et al. (Franck Rapaport, In preparation.).

TCGA AML data

We used the *curatedTCGAData* Bioconductor package to download RNA-seq counts and DNA methylation data for AML (Marcel Ramos, 2017).

BEAT AML data

We retrieved the DNA methylation beta value matrix for the BEAT AML cohort from the Gene Expression Omnibus (GSE159907) (Giacopelli *et al.*, 2021). We retrieved patient metadata and mutation information from Table S5 of the 2018 BEAT AML study (Tyner *et al.*, 2018). We retrieved RNA-seq gene counts (counts per million) from Table S9 of the 2018 study (Tyner *et al.*, 2018).

COCOA analysis

We performed unsupervised COCOA on the matrix of DNA methylation for e3999 patients. First, we performed PCA with centering but not scaling. We then used the PC scores from the first 10 PCs as input for COCOA, quantifying the relationship between the PC scores and DNA methylation with covariance and scoring region sets with the "regionMean" method. We used the region set database from the COCOA paper (Lawson *et al.*, 2020) that includes region sets from the ENCODE project (Consortium, 2012; Davis *et al.*, 2017), Roadmap Epigenomics (Bernstein *et al.*, 2010; Kundaje *et al.*, 2015), CODEX database (Sánchez-Castillo *et al.*, 2014), the Cistrome database (Mei *et al.*, 2016), JASPAR motif (Sandelin, 2004) predictions, and manually curated hematopoietic chromatin accessibility region sets (Lawson *et al.*, 2020). In addition, we used a manually curated group of region sets related to AML or the hematopoietic system. To get p-values

for COCOA, we calculated 10,000 random permutations to get a null distribution and used a gamma p-value approximation, as described in the COCOA paper (Lawson *et al.*, 2020).

Consensus clustering

We clustered DNA methylation within region sets of interest to identify patient clusters. First, we selected our region sets of interest. We took the top ten region sets from the COCOA analysis for each of the first ten PCs. Within each PC, we took only the highest ranking instance of a given TF or histone modification in order to prevent selection of multiple instances of similar region sets. Second, we did consensus clustering separately for each region set, with twelve clusters. We clustered using each unique combination of the following parameters. We used three dimensionality reduction techniques: PCA from which we used the first ten PCs, t-SNE with default parameters, and UMAP with default parameters. We used four different clustering algorithms: k-means clustering, hierarchical clustering with complete linkage, hierarchical clustering with single linkage, and hierarchical clustering with weighted average linkage. We used two distance metrics: euclidean and manhattan distance. For each combination of dimensionality reduction, clustering algorithm, and distance metric, we clustered 100 times. With all the clustering results for a region set, we created a consensus matrix. We did a final clustering of the consensus matrix using hierarchical clustering with Ward linkage and euclidean distance.

Differential methylation analysis

We performed differential methylation analysis on the e3999 RRBS data with the methylKit (Akalin *et al.*, 2012) and edmr packages (Li *et al.*, 2013). We performed differential methylation analysis on the TCGA and BEAT AML data with the bumphunter R package (Jaffe *et al.*, 2012). We annotated DMRs with the LOLA R package (Sheffield and Bock, 2015).

Differential gene expression analysis

We used the DeSeq2 package to calculate differentially expressed genes (Love *et al.*, 2014).

Data availability

De-identified patient-level clinical and molecular data from NCT00046930 will be deposited into the National Institutes of Health's NCTN/NCORP Data Archive (https://nctn-data-archive.nci.nih.gov) 6 months after publication of the corresponding paper. De-identified patient-level next generation sequencing data files will be deposited into the National Cancer Institute's Cancer Data Service Portal (https://datacommons.cancer.gov/repository/cancer-data-service) 6 months after publication of the corresponding paper. All analyses and conclusions in this dissertation are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, the NCORP or the NCI.

Chapter Four: Discussion and Future Directions

Summary

This research contributes to the fields of epigenetic analysis and AML. To facilitate epigenetic analysis, I developed two methods and R packages. First, I developed the MIRA R package that uses the shape of the DNA methylation profile around regions of interest to infer regulatory activity (Lawson *et al.*, 2018). Second, I developed the COCOA method and R package that annotate inter-sample epigenetic variation with region sets (Lawson *et al.*, 2020). I demonstrated COCOA in breast cancer, chronic lymphocytic leukemia, and renal cancer, in unsupervised and supervised analyses (Lawson *et al.*, 2020). I applied COCOA to both DNA methylation and ATAC-seq data, demonstrating the potential of COCOA to be used with a variety of epigenetic data types that have a signal associated with genomic regions and even to be integrated into a multi-omics analysis that includes such data. Observations from the COCOA analyses prompted a pan-cancer analysis of DNA methylation in polycomb complex-related regions that found that DNA methylation in these regions is significantly associated with patient survival and cancer stage in multiple cancer types. The MIRA and COCOA methods are available as R packages.

In AML, I analyzed DNA methylation subtypes in elderly patients. I used COCOA to identify DNA methylation variation between patients in genomic regions related to hematopoietic differentiation, the polycomb group proteins, and hematopoietic transcription factors. Through consensus clustering, I demonstrated that epitype depends on the regions examined. I compared lymphoid-accessible-chromatin epitypes to find a regulatory network potentially involving PRDM16 that is affected by DNA methylation.

Limitations

The MIRA and COCOA methods provide useful ways to understand epigenetic regulation but could still be improved. With MIRA, it was difficult to create an algorithm that would measure the shape of the DNA methylation profile despite the heterogeneity of different region sets. While I created an algorithm that would find the edges of the MIRA 'dips' automatically, I believe this is an area that could be improved for MIRA. A limitation of COCOA is that it requires a matrix of data points shared by all samples. This is not a problem when DNA methylation data comes from microarrays that have standard coverage but, for next generation sequencing-based DNA methylation results, some CpGs are only covered in a subset of samples and thus are dropped when generating a matrix of CpGs in all samples. However, I believe it would be possible to modify the COCOA algorithm to allow missing data. In addition, COCOA is limited by the region sets in what regulatory annotations are available to annotate the data but the number of publicly available region sets is increasing. In contrast, having too many region sets can be inconvenient if many region sets are similar to each other, for example, many ChIP-seq region sets for the same TF and cell line. This challenge is not unique to COCOA but is a general problem that the field needs to address. Another limitation of COCOA is its use of permutations to calculate p-values, which is computationally intensive. However, since publishing the COCOA paper, I improved the COCOA algorithm to increase speed by 1-2 orders of magnitude. This was accomplished by using matrix calculations internally. Once the data is in the format of a matrix of shared variables, I convert the region set database to a matrix of 'n' region sets by 'm' epigenetic variables (e.g. CpGs) based on whether an epigenetic variable overlaps a given region. This region set overlap matrix only needs to be calculated once and then can be used for each permutation when calculating COCOA scores. This method greatly increased the speed but a p-value approximation method is still required as discussed in the COCOA paper (Lawson *et al.*, 2020).

In the AML study, it was difficult to analyze the association of age and DNA methylation because the frequency of certain mutations that affect DNA methylation changes with age. If this is not corrected for, the age-associated DNA methylation changes may actually reflect the epigenetic signature of the mutations that become more or less common with age. If the frequency of relevant age-associated mutations is taken as a covariate to adjust the analysis, this could detract from the analysis because those mutation frequencies are correlated with the variable of interest, age. In hindsight, I see a way to potentially overcome this challenge: by grouping by epitype when testing the association of age with DNA methylation. In this way, the test within each epitype would be locally normalized.

Impact

A major challenge with high dimensional biological data is to interpret differences between individuals and understand their biological meaning. Methods like MIRA and COCOA annotate differences in systems-level gene regulation. MIRA can convert sparse data into a more robust signal to give sample-level scores that reflect gene regulatory information, allowing regulatory analysis of low-coverage samples. To understand regulatory variation within a group of samples, COCOA takes advantage of the shared regulation of regulatory regions across the genome and the covariation of these regions across samples to create a more robust signal. Then, this robust signal can be annotated with a rich collection of public and private region sets that represent gene regulatory information. These methods enable researchers to draw new insights from high dimensional epigenetic data.

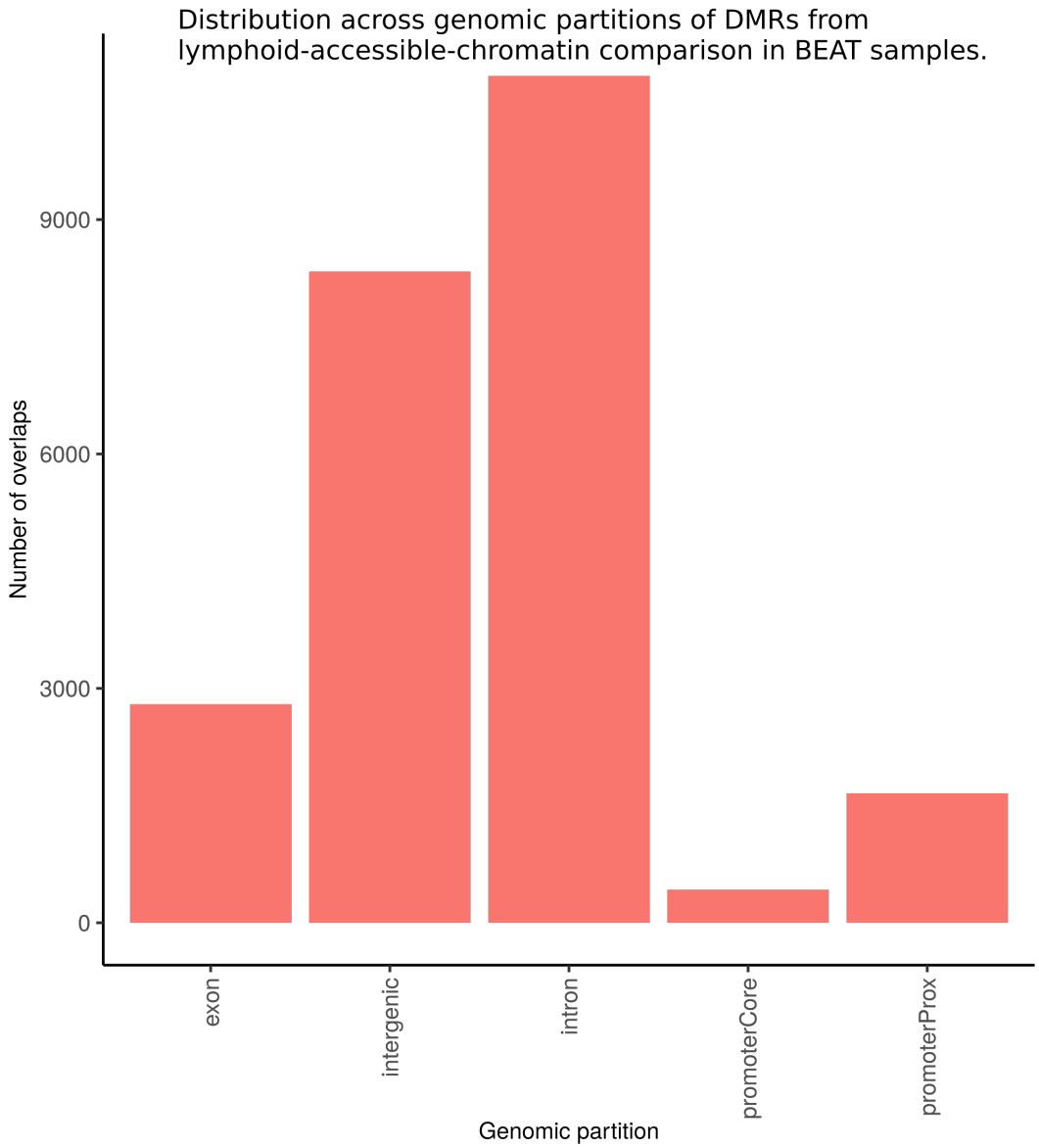
I made the methods MIRA and COCOA available as R packages, released through the Bioconductor organization. Bioconductor supports development and maintenance of R packages for biological analysis as well as training for scientists. Bioconductor screens packages before acceptance in order to ensure a quality standard, with requirements including documentation for functions and at least one vignette to detail use of the package. Bioconductor tracks and makes available statistics about the downloads of its packages. From this, I can calculate the average number of downloads from unique IP addresses per month for my packages. For MIRA, the average unique monthly downloads per year are as follows for 2018 through 2021 respectively: 61, 85.17, 43.5, and 46.78 (for 2021, January through September). For COCOA, the average unique monthly downloads per year are as follows for 2019, 2020, and 2021: 50.42, 36.42, and 40.22 (for 2021, January through September). These statistics demonstrate that the packages I developed are being disseminated to potentially hundreds or even thousands of people whose research can be impacted by them.

Through my AML analysis, I contribute an important idea to the field and increase understanding of AML biology. Generally, samples are characterized as having only one DNA methylation subtype/epitype (Figueroa, Abdel-Wahab, *et al.*, 2010; Glass *et al.*, 2017; Giacopelli *et al.*, 2021). This condenses the information from millions of CpGs and, for bulk samples, a multitude of cells, into a single characterization. Given the complexity of the data and the many cellular pathways that can impact DNA methylation differently or at different CpGs, it seems likely that valuable gene regulatory information may be lost by assigning patients to a single epitype. In contrast, I suggest in my AML study that patients could have multiple epitypes that each reveal different gene regulatory information. I believe that, if this convention is practiced, it would allow patients to be better separated into groups for targeted therapies because it could give more detailed information about the cellular pathways or axes that are active in each patient. Therefore, this concept could be valuable to the AML field. In addition, I further characterized AML biology through my epitype analysis. In the comparison of lymphoid-accessible-chromatin epitypes, I uncovered a potential mechanism by which DNA methylation can impact expression of the PRDM16 gene and the related regulatory network. By understanding the regulatory axes that are dysregulated in certain epitypes, we come closer to targeted therapies and to understanding the mechanisms that result in convergent epitypes in the absence of epitype-defining mutations.

Future work

Since a major reason for interest in DNA methylation is how it impacts gene expression, it would be valuable to do a follow-up analysis relating the e3999 DNA methylation to gene expression, which was beyond the scope of this project. A basic form of this would be to characterize the gene expression profiles of the various epitypes. However, a promising future direction is to analyze the association between DNA methylation epitypes and gene splicing. Mutations in splicing-related proteins are fairly common in AML. In addition, DNA methylation can affect splicing and many of the differentially methylated regions between the lymphoid-accessible-chromatin epitypes examined in the previous chapter were in introns (Fig. 1). Protein isoforms can have different biological effects in AML, as in the case of PRDM16 in which the short isoform of the protein is predicted to have a specific effect in contributing to AML (Corrigan *et al.*, 2018).

Additional analysis of the consensus clustering data could suggest mechanisms behind cases of convergent and divergent epigenetic states. For convergent epigenetic states (e.g. when samples do not share the cluster-defining mutations, yet still cluster with those samples), it would be interesting to determine potential reasons for this. For instance, over-expression or under-expression of an important gene could cause this. In this case, that could potentially be determined by separating the samples with cluster-defining mutations from samples



Dissertation figure 23: **Figure 1.** Distribution across genomic partitions of DMRs from lymphoid-accessible-chromatin comparison in BEAT samples.

that do not have them and doing a differential gene expression analysis if sample size was large enough. Similarly, it would be interesting to understand divergent epigenetic states (different clustering) for samples that share the same mutations. One way to address this would be to give more attention to the specific mutations that occurred in the commonly mutated genes. In this study, I generally grouped mutations to a given gene into a single category but the different mutations in the same gene may have different effects and result in different epitypes. For instance, some samples with mutations in the RUNX1 gene often grouped together in DNA methylation clustering but others did not. Among many possible reasons for this could be that the samples that do not cluster with the main RUNX1 group have different types of RUNX1 mutations. It may be possible to group mutations for a given gene into subcategories based on the protein domain that was affected or the predicted result. Then, the samples with mutations in the same gene but which have divergent epigenetic states (different clustering) could be compared to see if the type of mutation differs between clusters.

My project identified DNA methylation variation in some CTCF regions in the differential methylation analysis between lymphoid-accessible-chromatin epitypes. CTCF helps to orchestrate chromatin looping, including between promoters and enhancers, to ensure proper gene expression (Oh *et al.*, 2021). Disruption of CTCF binding sites can result in enhancers not being directed toward the primary promoter but instead increasing the expression of nearby genes, a phenomenon termed "enhancer release and retargeting" (Oh *et al.*, 2021). Additionally, CTCF binding is DNA methylation sensitive (Bell and Felsenfeld, 2000) and a study found that increased DNA methylation at a CTCF binding site in glioma due to an IDH mutation was associated with an enhancer crossing domain boundaries and upregulating an oncogene (Flavahan *et al.*, 2015). In regard to my project, since I found differential methylation at CTCF binding sites, a next step would be to test whether that DNA methylation is correlated with gene expression of the nearby genes and also whether any correlated genes are associated with AML biology.

To translate the findings of my project to a clinical application, future research could focus on identifying a subset of CpGs that can distinguish epitypes or separate patients according to outcome. The rationale that region set epitypes may be relevant for prognosis or targeted therapy is defined as follows: We start with the assumption that differences between samples in epitype for a given region set may reflect differences in gene regulatory activity for a regulatory pathway or axis that affects multiple, perhaps many, genes. Some of these genes could affect survival or be important for a targeted therapy. Thus, it is the ability to potentially get information about the activity of many genes from a single epitype classification that makes region set-based epitypes an attractive potential biomarker. Epitypes from different region sets may give information about different genes. In some cases the usefulness of region set-based epitypes may only become clear in a specific treatment context. For a theoretical example, an epitype might not show ability to predict survival when tested in patients with standard chemotherapy treatment but, when tested in patients given a different treatment, there may be differential activity of a treatment-resistance gene related to the gene regulatory pathway the epitype represents that gives the epitype prognostic value for survival. In other words, the prognostic value of each region set epitype may be treatment-specific. As a way of representing systems-level gene regulatory differences, region set epitypes offer potential for better context-specific prognosis or targeted therapies.

Previous studies about DNA methylation signatures in AML provide insight into the creation of prognostic DNA methylation signatures for clinical use. A prognostic DNA methylation signature was defined by Figueroa et al. in 2010 (Figueroa, Lugthart, *et al.*, 2010). This signature predicted patient survival from DNA methylation levels at eighteen CpGs (Figueroa, Lugthart, *et al.*, 2010). A cost-effective, microsphere-based assay to measure this signature was developed by Wertheim et al. (Wertheim *et al.*, 2014; Wertheim *et al.*, 2015) to facilitate the use of this signature in clinical settings. In regard to my research, it should be possible to define a DNA methylation signature that could distinguish epitypes. Since a sample may have various epitypes depending on the CpGs considered, a separate signature could be developed for each region set's epitypes if they become relevant for prognosis or targeted therapy.

References

Adelman,E.R. *et al.* (2019) Aging human hematopoietic stem cells manifest profound epigenetic reprogramming of enhancers that may predispose to leukemia. **9**, 1080–1101.

Aguilo,F. *et al.* (2011) Prdm16 is a physiologic regulator of hematopoietic stem cells. *Blood*, **117**, 5057–5066.

Akalin,A. *et al.* (2012) methylKit: A comprehensive r package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, **13**, R87.

Argelaguet,R. *et al.* (2018) Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, **14**, e8124.

Bachmann,I.M. *et al.* (2006) EZH2 expression is associated with high proliferation rate and aggressive tumor subgroups in cutaneous melanoma and cancers of the endometrium, prostate, and breast. *Journal of Clinical Oncology*, **24**, 268–273.

Basheer,F. *et al.* (2019) Contrasting requirements during disease evolution identify EZH2 as a therapeutic target in AML. *The Journal of Experimental Medicine*, **216**, 966–981.

Bell,A.C. and Felsenfeld,G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482–485.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

Bernstein,B.E. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, **28**, 1045–1048.

Bioconductor Package Maintainer <Maintainer@Bioconductor.Org> (2017) ExperimentHub.

Boer,C.G. de and Regev,A. (2018) BROCKMAN: Deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinformatics*, **19**.

Boyle,P. *et al.* (2012) Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biology*, **13**, R92.

Bullinger,L. *et al.* (2009) Quantitative DNA methylation predicts survival in adult acute myeloid leukemia. *Blood*, **115**, 636–642.

Chen,Z. *et al.* (2017) Expression of EZH2 is associated with poor outcome in colorectal cancer. *Oncology Letters*.

Colaprico,A. *et al.* (2015) TCGAbiolinks: An r/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, **44**, e71–e71.

Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Corces,M.R. *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, **48**, 1193–1203.

Corces,M.R. *et al.* (2018) The chromatin accessibility landscape of primary human cancers. *Science*, **362**, eaav1898.

Corrigan,D.J. *et al.* (2018) PRDM16 isoforms differentially regulate normal and leukemic hematopoiesis and inflammatory gene signature. *Journal of Clinical Investigation*, **128**, 3250–3264.

Cripe,L.D. *et al.* (2010) Zosuquidar, a novel modulator of p-glycoprotein, does not improve the outcome of older patients with newly diagnosed acute myeloid leukemia: A randomized, placebo-controlled trial of the eastern cooperative oncology group 3999. *Blood*, **116**, 4077–4085.

Davis,C.A. *et al.* (2017) The encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, **46**, D794–D801.

Delignette-Muller,M.L. and Dutang,C. (2015) Fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, **64**.

Deneault,E. *et al.* (2013) Identification of non-cell-autonomous networks from engineered feeder cells that enhance murine hematopoietic stem cell activity. *Experimental Hematology*, **41**, 470–478.e4.

Deneberg,S. *et al.* (2011) Prognostic DNA methylation patterns in cytogenetically normal acute myeloid leukemia are predefined by stem cell chromatin marks. *Blood*, **118**, 5573–5582.

Dietrich,S. *et al.* (2017) Drug-perturbation-based stratification of blood cancer. *Journal of Clinical Investigation*, **128**, 427–445.

Domcke,S. *et al.* (2015) Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, **528**, 575–579.

Dozmorov,M.G. (2017) Epigenomic annotation-based interpretation of genomic data: From enrichment analysis to machine learning. *Bioinformatics*, **33**, 3323–3330.

Eferl,R. and Wagner,E.F. (2003) AP-1: A double-edged sword in tumorigenesis. *Nature Reviews Cancer*, **3**,

859–868.
Fabbri,G. and Dalla-Favera,R. (2016) The molecular pathogenesis of chronic lymphocytic leukaemia. *Nature Reviews Cancer*, **16**, 145–162.
Farlik,M. *et al.* (2015) Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Reports*, **10**, 1386–1397.
Faunes,F. *et al.* (2013) A membrane-associated beta-catenin/Oct4 complex correlates with ground-state pluripotency in mouse embryonic stem cells. *Development*, **140**, 1171–1183.
Figueroa,M.E., Lugthart,S., *et al.* (2010) DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell*, **17**, 13–27.
Figueroa,M.E., Abdel-Wahab,O., *et al.* (2010) Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell*, **18**, 553–567.
Fisher,J.B. *et al.* (2017) Cohesin mutations in myeloid malignancies. **3**, 282–293.
Flavahan,W.A. *et al.* (2015) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**, 110–114.
Fleischer,T. *et al.* (2017) DNA methylation at enhancers identifies distinct breast cancer lineages. *Nature Communications*, **8**.
Franck Rapaport,Y.N.,Kenneth Seier (In preparation.) A female-linked AML patient group older than sixty years of age has better risk in ECOG-ACRIN's clinical trial E3999.
Frietze,S. *et al.* (2008) CARM1 regulates estrogen-stimulated breast cancer growth through up-regulation of E2F1. *Cancer Research*, **68**, 301–306.
Frost,H.R. *et al.* (2015) Principal component gene set enrichment (PCGSE). *BioData Mining*, **8**.
Fujiki,K. *et al.* (2013) PPARgamma-induced PARylation promotes local DNA demethylation by production of 5-hydroxymethylcytosine. *Nature Communications*, **4**.
Garrett-Bakelman,F.E. *et al.* (2015) Enhanced reduced representation bisulfite sequencing for assessment of DNA methylation at base pair resolution. *Journal of Visualized Experiments*.
Giacopelli,B. *et al.* (2021) DNA methylation epitypes highlight underlying developmental and disease pathways in acute myeloid leukemia. *Genome Research*, **31**, 747–761.
Glass,J.L. *et al.* (2017) Epigenetic identity in AML depends on disruption of nonpromoter regulatory elements and is affected by antagonistic effects of mutations in epigenetic modifiers. *Cancer Discovery*, **7**, 868–883.
Grambsch,P.M. and Therneau,T.M. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–526.
Gudmundsson,K.O. *et al.* (2020) Prdm16 is a critical regulator of adult long-term hematopoietic stem cell quiescence. *Proceedings of the National Academy of Sciences*, **117**, 31945–31953.
Guo,S. *et al.* (2016) EZH2 overexpression in different immunophenotypes of breast carcinoma and association with clinicopathologic features. *Diagnostic Pathology*, **11**.
Holm,K. *et al.* (2012) Global H3K27 trimethylation and EZH2 abundance in breast tumor subtypes. *Molecular Oncology*, **6**, 494–506.
Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.
Horvath,S. (2013) DNA methylation age of human tissues and cell types. **14**, R115.
Hu,T. *et al.* (2019) PRDM16s transforms megakaryocyte-erythroid progenitors into myeloid leukemia-initiating cells. *Blood*, **134**, 614–625.
Huber,W. *et al.* (2015) Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, **12**, 115–121.
Hwang,C. *et al.* (2007) EZH2 regulates the transcription of estrogen-responsive genes through association with REA, an estrogen receptor corepressor. *Breast Cancer Research and Treatment*, **107**, 235–242.
Izzo,F. *et al.* (2020) DNA methylation disruption reshapes the hematopoietic differentiation landscape. **52**, 378–387.
Jaffe,A.E. *et al.* (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, **41**, 200–209.
Johnson,K.C. *et al.* (2017) Normal breast tissue DNA methylation differences at regulatory elements are associated with the cancer risk factor age. *Breast Cancer Research*, **19**.
Kandarakov,O. and Belyavsky,A. (2020) Clonal hematopoiesis, cardiovascular diseases and hematopoietic stem cells. **21**, 7902.
Kapourani,C.-A. and Sanguinetti,G. (2016) Higher order methylation features for clustering and prediction in

epigenomic studies. *Bioinformatics*, **32**, i405–i412.
Kassambara,A. *et al.* (2019) Survminer: Drawing survival curves using 'ggplot2'.
Kim,K.H. and Roberts,C.W.M. (2016) Targeting EZH2 in cancer. *Nature Medicine*, **22**, 128–134.
Krueger,F. and Andrews,S.R. (2011) Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, **27**, 1571–1572.
Kundaje,A. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
Lawson,J.T. *et al.* (2020) COCOA: Coordinate covariation analysis of epigenetic heterogeneity. *Genome Biology*, **21**.
Lawson,J.T. *et al.* (2018) MIRA: An r package for DNA methylation-based inference of regulatory activity. *Bioinformatics*, **34**, 2649–2650.
Layer,R.M. *et al.* (2018) GIGGLE: A search engine for large-scale integrated genome analysis. *Nature Methods*, **15**, 123–126.
Lee,S.-M. *et al.* (2015) The regulatory mechanisms of intragenic DNA methylation. *Epigenomics*, **7**, 527–531.
Li,S. *et al.* (2013) An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics*, **14**.
Li,S. *et al.* (2016) Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nature Medicine*, **22**, 792–799.
Liu,L. *et al.* (2013) Prognostic value of EZH2 expression and activity in renal cell carcinoma: A prospective study. *PLoS ONE*, **8**, e81484.
Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**.
Lugthart,S. *et al.* (2010) Aberrant DNA hypermethylation signature in acute myeloid leukemia directed by EVI1. *Blood*, **117**, 234–241.
Maegawa,S. *et al.* (2014) Age-related epigenetic drift in the pathogenesis of MDS and AML. *Genome Research*, **24**, 580–591.
Maor,G.L. *et al.* (2015) The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, **31**, 274–280.
Marcel Ramos,L.W. (2017) curatedTCGAData: Curated data from the cancer genome atlas (TCGA) as MultiAssayExperiment objects.
McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, **28**, 495–501.
Mei,S. *et al.* (2016) Cistrome data browser: A data portal for ChIP-seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research*, **45**, D658–D662.
Melling,N. *et al.* (2015) Overexpression of enhancer of zeste homolog 2 (EZH2) characterizes an aggressive subset of prostate cancers and predicts patient prognosis independently from pre- and postoperatively assessed clinicopathological parameters. *Carcinogenesis*, **36**, 1333–1340.
Meng,C. *et al.* (2019) MOGSA: Integrative single sample gene-set analysis of multiple omics data. *Molecular & Cellular Proteomics*, **18**, S153–S168.
Mingay,M. *et al.* (2017) Vitamin c-induced epigenomic remodelling in IDH1 mutant acute myeloid leukaemia. *Leukemia*, **32**, 11–20.
Morgan,M. (2019) AnnotationHub: Client to access AnnotationHub resources.
Odom,G.J. *et al.* (2019) pathwayPCA: An r package for integrative pathway analysis with modern PCA methodology and gene selection.
Oh,S. *et al.* (2021) Enhancer release and retargeting activates disease-susceptibility genes. *Nature*, **595**, 735–740.
Orkin,S.H. (1995) Transcription factors and hematopoietic development. *Journal of Biological Chemistry*, **270**, 4955–4958.
Papaemmanuil,E. *et al.* (2016) Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine*, **374**, 2209–2221.
Petryk,N. *et al.* (2020) Staying true to yourself: Mechanisms of DNA methylation maintenance in mammals. **49**, 3020–3032.
R Core Team (2018) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
Rampal,R. *et al.* (2014) DNA hydroxymethylation profiling reveals that WT1 mutations result in loss of TET2 function in acute myeloid leukemia. *Cell Reports*, **9**, 1841–1855.
Rasmussen,K.D. and Helin,K. (2016) Role of TET enzymes in DNA methylation, development, and cancer.

Genes & Development, **30**, 733–750.
Ren,W. *et al.* (2018) Structural basis of DNMT1 and DNMT3A-mediated DNA methylation. *Genes*, **9**, 620.
Ricard Argelaguet and Britta Velten and Damien Arnol and Florian Buettner and Wolfgang Huber and Oliver Stegle (2020) MOFAdata: Data package for Multi-Omics Factor Analysis (MOFA).
Rinke,J. *et al.* (2020) EZH2 in myeloid malignancies. **9**, 1639.
Rosenbauer,F. and Tenen,D.G. (2007) Transcription factors in myeloid development: Balancing differentiation with transformation. *Nature Reviews Immunology*, **7**, 105–117.
Sandelin,A. (2004) JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, **32**, 91D–94.
Sánchez-Castillo,M. *et al.* (2014) CODEX: A next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Research*, **43**, D1117–D1123.
Schep,A. (2018) Motifmatchr: Fast motif matching in r.
Schep,A.N. *et al.* (2017) chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods*, **14**, 975–978.
Segovia-Mendoza,M. and Morales-Montor,J. (2019) Immune tumor microenvironment in breast cancer and the participation of estrogen and its receptors in cancer physiopathology. *Frontiers in Immunology*, **10**.
Sheffield,N.C. *et al.* (2017) DNA methylation heterogeneity defines a disease spectrum in ewing sarcoma. *Nature Medicine*, **23**, 386–395.
Sheffield,N.C. *et al.* (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research*, **23**, 777–788.
Sheffield,N.C. and Bock,C. (2015) LOLA: Enrichment analysis for genomic region sets and regulatory elements in r and bioconductor. *Bioinformatics*, **32**, 587–589.
Somasundaram,R. *et al.* (2015) Transcription factor networks in b-cell differentiation link development to acute lymphoid leukemia. *Blood*, **126**, 144–152.
Stadler,M.B. *et al.* (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545–15550.
Suzuki,T. *et al.* (2017) RUNX1 regulates site specificity of DNA demethylation by recruitment of DNA demethylation machineries in hematopoietic cells. *Blood Advances*, **1**, 1699–1711.
Takao,Y. *et al.* (2007) Beta-catenin up-regulates nanog expression through interaction with oct-3/4 in embryonic stem cells. *Biochemical and Biophysical Research Communications*, **353**, 699–705.
Tate,P.H. and Bird,A.P. (1993) Effects of DNA methylation on DNA-binding proteins and gene expression. *Current Opinion in Genetics & Development*, **3**, 226–231.
TCGAResearchNetwork (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, **368**, 2059–2074.
Terry M. Therneau and Patricia M. Grambsch (2000) Modeling survival data: Extending the Cox model Springer, New York.
Therneau,T.M. (2015) A package for survival analysis in s.
Tyner,J.W. *et al.* (2018) Functional genomic landscape of acute myeloid leukaemia. *Nature*, **562**, 526–531.
Ung,M. *et al.* (2014) Effect of estrogen receptor alpha binding on functional DNA methylation in breast cancer. *Epigenetics*, **9**, 523–532.
Varambally,S. *et al.* (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, **419**, 624–629.
Wang,Y. *et al.* (2017) Ezh2 acts as a tumor suppressor in kras-driven lung adenocarcinoma. *International Journal of Biological Sciences*, **13**, 652–659.
Weirauch,M.T. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
Wertheim,G.B.W. *et al.* (2014) Microsphere-based multiplex analysis of DNA methylation in acute myeloid leukemia. **16**, 207–215.
Wertheim,G.B.W. *et al.* (2015) Validation of DNA methylation to predict outcome in acute myeloid leukemia by use of xMELP. **61**, 249–258.
Wiench,M. *et al.* (2011) DNA methylation status predicts cell type-specific enhancer activity. *The EMBO Journal*, **30**, 3028–3039.
Wijetunga,N.A. *et al.* (2016) A pre-neoplastic epigenetic field defect in HCV-infected liver at transcription factor binding sites and polycomb targets. *Oncogene*, **36**, 2030–2044.

Winkler,A.M. *et al.* (2016) Faster permutation inference in brain imaging. *NeuroImage*, **141**, 502–516.
Yang,Q. *et al.* (2016) AMPK/alpha-ketoglutarate axis dynamically mediates DNA demethylation in the Prdm16 promoter and brown adipogenesis. *Cell Metabolism*, **24**, 542–554.
Yao,L. *et al.* (2015) Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biology*, **16**.
Ying,L. *et al.* (2015) OCT4 coordinates with WNT signaling to pre-pattern chromatin at the SOX17 locus during human ES cell differentiation into definitive endoderm. *Stem Cell Reports*, **5**, 490–498.
Zhang,D. *et al.* (2017) EZH2 enhances the invasive capability of renal cell carcinoma cells via activation of STAT3. *Molecular Medicine Reports*.
Zhang,H. *et al.* (2018) Identification of DNA methylation prognostic signature of acute myelocytic leukemia. *PLOS ONE*, **13**, e0199689.
Zhu,H. *et al.* (2016) Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics*, **17**, 551–565.