Algorithms and Models for Genome Biology

A dissertation presented

by

James Yang Zou

to

The School of Engineering and Applied Sciences

in partial fulllment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Applied Mathematics

Harvard University

Cambridge, Massachusetts

October 2013

Thesis advisor: David C. Parkes Author: James Y. Zou

Abstract

New advances in genomic technology make it possible to address some of the most fundamental questions in biology for the first time. They also highlight a need for new approaches to analyze and model massive amounts of complex data. In this thesis, I present six research projects that illustrate the exciting interaction between high-throughput genomic experiments, new machine learning algorithms, and mathematical modeling. This interdisciplinary approach gives insights into questions ranging from how variations in the epigenome lead to diseases across human populations to how the slime mold finds the shortest path. The algorithms and models developed here are also of interest to the broader machine learning community, and have applications in other domains such as text modeling.

Acknowledgements

I would like to thank my Ph.D. mentors David Parkes, Michael Brenner and Brad Bernstein for their constant support and friendship. They were (and still are) essential for my development as a researcher, and their passion for science is truly inspiring. I have also learned a great deal from my undergraduate mentor, John Harer, and my advisor at the University of Cambridge, Frank Kelly. I hope I can be as good a mentor to future scientists as they have been to me.

I had the great fortune to collaborate with many friends, who made the last few years a wonderful and stimulating experience: Ryan Adams, Mazhar Adli, Martin Aryee, Jon Aster, Leonid Chindelevitch, Alon Goren, Roger Grosse, Sujit Gujar, CJ Ho, Daniel Hsu, Anders Johannson, Elliott Kieff, John Lai, Christoph Lippert, Jennifer Listgarten, Steve McCarroll, Eric Mendenhall, Tarjei Mikkelsen, Arvind Murugan, Swaprava Nath, Elan Pavlov, Aviv Regev, Pardis Sabeti, Matt Stremlau, David Sumpter, Andrew Tanentzap, Hongfang Wang, Bo Zhao, Jiang Zhu, Daniel Ziemek.

Special thanks also go to my friends and office/lab/group mates, who made 02138/02139 such an amazing and fun place to grow up: Ofra Amir, Elaine Angelino, Tiffany Chen, Audrey Fan, Michael Gelbart, Amneet Gulati, Anna Huang, Miriam Huntley, Varun Kanade, Richard Koche, Tom Liptay, Zhenming Liu, Shamin Nemati, Sana Nourani, Deriba Olana, Katharina Reinecke, Oren Rippel, Scott Linderman, Roni Stern, Finale Doshi-Velez, Harvey Xiao, Yuchen Yu, Zorana Zeravcic, Lihua Zou, the Maxwell-Dworkin board game group, all the coxes and rowers of the Dudley M2 boat, and the Dudley House.

Thanks to the National Science Foundation for the financial support of a graduate research fellowship.

Most importantly, I want to thank my parents and Yvonne for their love and support.

Contents

1

Introduction

The genomics era poses exciting opportunities and challenges. We need creative experiments to harness the power of genomic technology, new machine learning algorithms to extract robust patterns from complex data, and new mathematical models to understand the fundamental principles behind such patterns. The cross-breeding of ideas between these disciplines is extremely exciting. My research lies at this intersection of data modeling, mathematical analysis and biological discovery. In this thesis, I present six projects that illustrate this fruitful interaction.

A fascinating set of biological relate to the epigenomic regulation of cellular state and phenotype. To tackle these questions, I worked very closely with biologists to design experiments, and, in several projects, performed my own experiments with some successes and failures. By working in a wet lab everyday for many months, I gained a much better understanding of biology and data, as well as an appreciation for a different mode of problem solving. I think it's a very valuable experience for the analyst/theorist to design and perform some experiments. And some experiments are fun to do.

At the other end of the spectrum, I worked on developing new computational frameworks and proved theorems about rather abstract mathematical models. New tools and ideas are needed in computational biology. Currently a few basic tools like clustering, PCA and its variants are probably used in 95% of computational analysis. The limited toolkit might be sufficient for some problems, but it constrains the analyst's imagination in more complex settings. Without richer models, we fail to even ask the interesting questions. So I developed new algorithms and models based on rather artisanal techniques such as tensor decomposition and determinantal point process. They have elegant mathematical properties, which by itself is motivation enough for exploration, but I hope my work is a step towards making these powerful ideas a part of the compbio lingua. While motivated by biology, many of the algorithmic techniques that I developed here are broadly applicable and have led to interesting advances in domains such as text analysis and image classification.

1.1 Overview of the thesis

Chapter 2 is based on the paper (44). In it I present a novel and rigorous statistical framework to identify epigenomic changes in human populations that are potential drivers of disease. The DNA inherited by an individual encodes instructions for hundreds of distinct cell types. The cells of our brain and liver, for example, are very different, even though they all have identical DNA sequence. This is possible because the same 3 billion bases of DNA are packaged differently in brain cells and liver cells. The packaging of DNA in three-dimensions is regulated by epigenomics. Epigenomics refer to the chemical modifications to DNA and histones, around which DNA is wrapped. These modifications specify, in each cell, which subset of DNA sequences can be accessed and read by molecular machinery inside the cell. This in turn determines which subset of genetic instructions are performed by different cells.

Our predisposition to certain diseases can be linked to specific mutations to the DNA. It is now appreciated that many diseases also have epigenetic contributions. However it has been very challenging to systematically compare the epigenome across human population. This is because, unlike DNA, the epigenome is highly variable in different cells, and cell-type heterogeneity in samples lead to massive false association signals. To compound the issue, we often do not know what are the relevant cell types present in a sample, nor do we know what fraction of the sample is composed of each cell type. I develop a method, EWASher, which uses a linear mixed model to account for this type of confounders. The method pinpoints changes in the packaging of DNA that are potentially linked to disease. I apply it to discover regions of the genome that are abnormally shut-off or accessible in people with rheumatoid arthritis, breast cancer, or colon cancer. These regions give new insights into the progression of these diseases, and can potentially be used as diagnostic biomarkers.

Chapter 3 is based on the paper (69). Here I explore the changes in the epigenome during the transformation of T-cell leukemia. Using statistical analysis of epigenomic changes, I characterize an identity-theft strategy in leukemia whereby certain cancer-causing agents hijack the normal regulatory network within the cell. I show that a mutated form of a key protein, Notch1, hijack the DNA docking sites of normal proteins to change how T cells grow and divide. In addition, I find similar patterns in mouse T-cell leukemia, suggesting that this identity-theft strategy is conserved through evolution.

Chapter 4 is based on the paper (4). In a follow up to Chapter 3, I show that a similar identity-theft leads to uncontrolled cell proliferation in a different model systemB cell lymphoma caused by Epstein-Barr virus (EBV) infection. By analyzing epigenomic data, I show that in infected B cells, components of the EBV virus hijacked the same binding sites and regulatory elements that the mutated Notch1 co-opted in T cells. This resulted in immortal B cells that can lead to cancer. By analyzing DNA motifs, I identify potential proteins that partner with the viral proteins. Moreover, I present computational and experimental evidence that the viral protein targeted key oncogenes via 3-D DNA looping.

Chapter 5 is based on the paper (45). In the analysis of cancer above, as in many other settings, we have multiple data sets and want to understand the differences in statistical patterns between them. Motivated by this, I formalize a framework of *contrastive learning*, and introduce a new and general algorithm based on tensor decomposition to efficiently learn contrasts between data sets. This method learns a probabilistic, generative model that explains the difference between the foreground and background data. The chapter illustrates the applications of this method in contrastive epigenomics and in contrastive text modeling.

Chapter 6 is based on the paper (89). Here I introduce the notion of diversity as an objective in building statistical models. In many settings, we would like to have not just one good solution to a problem, but a diverse set of solutions that are mutually dissimilar. For example, think of the time you googled X. Google returns a set of high quality links relevant to X, but the links are very similar to each other (they might all use similar sources). A list of high quality results from diverse sources would be more informative. I use the determinantal point process as a prior to enforce diversity, and presented efficient learning methods. I show that diversified probabilistic models learned more interpretable and robust topics from corpus of texts. It also improves the performance of image classification.

Chapter 7 is based on the paper (75). It combines mathematical and algorithmic models to solve a fascinating mystery of the slime mold. Recent experiments observed that if we place a slime mold at one end of a maze and a food source (e.g. corn flake) at the other end, the slime grows over time and finds the shortest path through the maze. How does such a simple organism find the shortest path? I develop a system of differential equations to model the slime mold growth dynamics. By analyzing these equations, I prove that under general conditions, the slime mold is guaranteed to find the shortest path. Moreover, I show that a general family of optimization problems (linear programs) can be encoded as different slimes, and by growing them, it finds the optimal solutions. This inspires a new family of optimization solvers.

2

EWASher: Epigenome-Wide Association Studies Without the Need for Cell-Type Composition

2.1 Overview

Epigenome-wide association studies (EWAS) face many of the same challenges as genome-wide association studies (GWAS), but face an added challenge in that the epigenome can vary dramatically across cell types. When cell-type composition differs between cases and controls, this leads to spurious associations that bury true associations. While the current approach is to estimate the cell-type composition in each sample using laboriously assayed reference profiles, we propose to bypass this step altogether with EWASher. Our method automatically corrects for cell-type composition without explicit knowledge of it. On rheumatoid arthritis methylation data, EWASher performs as well as the state-of-art method, which explicitly uses cell-type composition. We further validate the approach using extensive simulations. Finally, we apply EWASher to the Cancer Genome Atlas breast and colon cancer methylation data, where no cell-type composition is available. Our work is a step toward placing EWAS on a solid statistical footing comparable to that of GWAS.

2.2 Introduction

With the era of next generation sequencing comes high-throughput measurement not only of the genome, but also of the epigenome, yielding complementary information that is critical to understanding, and then tackling, disease mechanisms. Epigenetics informs us about the structure and accessibility of DNA, which in turn yields information about regulation and transcriptionkey drivers of disease. Thus, epigenetics is a crucial mediating link between genetics and function. In many diseases, it is now appreciated that epigenetic changes complement genetic mutations in driving the disease (50, 76, 80, 83).

Currently, the measurement and analysis of epigenetic data through epigenome-wide association studies (EWAS) is a subject of much interest, as such analyses yield insights into the role of epigenetic regulation in disease (62). The goal of EWAS, analogous to GWAS, is to identify changes in the epigenome at particular loci that are correlated with some phenotype of interest, by scanning along the entire epigenome. While such analysis alone cannot establish causality, epigenetic association studies shed light on disease pathways and drivers, and also identify candidate biomarkers for diagnostics. The present work is a step towards placing epigenomic association studies on a more sound statistical footing, one comparable to that of GWAS. In particular, we seek to avoid spurious associations that arise due to cell type composition heterogeneity; to use as much of the data as possible (i.e., to not remove samples unnecessarily); and, to do so without the need for specialized auxiliary calibration data such as reference profiles from individual cell types.

EWAS faces many of the same challenges as traditional GWAS in finding the needles of signal in the haystack that is the genome. The shared challenges include confounding by batch effects, population structure and family relatedness, the large scale of multiple hypothesis testing, and the need to group together weak effects to find underpowered associations (5, 40). Importantly, EWAS faces an additional, major challenge in that the epigenome can be highly variable across different cell types, and the case and control samples in a study typically differ in their cell-type compositions. Such heterogeneity can give rise to spurious associations, which in turn hide the true associations. This undesirable effect was recently illustrated in a rheumatoid arthritis (RA) methylation study (67) (and analyzed in depth in Results). In this study, blood samples from cases contained a larger fraction of myeloid cells (granulocytes and monocytes) relative to lymphocytes (B, T and NK cell) than the controls. A naive EWAS that did not correct for this systematic cell-type heterogeneity was flooded by spurious associations arising from loci that were specifically methylated in either lymphocytes or myeloid cells (Fig 1), and the true disease-associated loci were not found. Including covariates for cell type composition corrected for this heterogeneity, thereby removing spurious associations ((67) and Fig 1). The critical bottleneck of this approach was in accurately measuring or estimating the cell-type composition of each sample; this would be the bottleneck for all studies of this nature. We propose to bypass this bottleneck entirely, while maintaining an accurate analysis, with the introduction of a new statistical approach for EWAS, called EWASher.

The method of adding cell type covariates, as used in (67), is the current state of art for tackling an EWAS in the presence of cell type heterogeneity. This approach requires knowledge of the reference methylome (measured from purified cells) of each of the main cell types in the sample. Given this information, it is then possible to estimate the cell type composition of every sample a statistical deconvolution algorithm. Although this approach worked well for the RA study, its dependence on purified reference profiles, which is the basis of all current procedures that correct for cell-type heterogeneity (33, 65), is problematic for several reasons:

1. For many diseases (e.g., cancer), complex mixture of many cell types are present for which it can be extremely difficult, laborious and expensive to obtain a reference population of cells. Furthermore, there may be cases where the dominant cell types are not known.

2. Reference cell type profiles generated in one lab or under one condition may not accurately capture the correct information for samples collected elsewhere or under different conditions.

3. The process of isolating cells from solid tissue is challenging and may perturb the cells (32).

A key insight in obtaining a general, reference free solution in EWAS is to recognize that the problem of confounding by cell-type heterogeneity is analogous to the problem of confounding due to population stratification and family relatedness in GWAS.

However, as we will demonstrate, confounding by cell type in EWAS is much more severe, because cell compositions can vary dramatically across samplesfar more so than the extent to which SNPs vary by population or family. Motivated by methods in GWAS, our solution makes use of the linear mixed model (LMM) (28, 36, 42), which we extend to perform a reference-free correction for EWAS in the presence of cell type heterogeneity.

Specifically, our new method, which we call EWASher, is a hybrid approach of (1) a feature-selected LMM (28, 29, 40) and (2) a principal components (PC) based approach (3, 20). As we will demonstrate on real and synthetic data, this method successfully corrects for confounding by cell-type composition, without any dependence on purified reference cell types, making it reference-free. Furthermore, the method is computationally efficient, scaling up to the large data sets produced with current technologies.

To validate our approach, we applied it to a bronze-standard RA dataset in which the methylomes of the reference cell types and estimates of the cell-type composition of each sample in the cohort are known. We found that our reference-free approach performed just as well as the state-of-the art reference-based approach. Next, we further validated our method using extensive simulations and mathematical analysis. Finally, we applied our method to breast and colon cancer methylation data from The Cancer Genome Atlas (TCGA) to identify candidate biomarkers and biologically significant genes in the presence of substantial confounding.

In summary, we make the following significant contributions:

1. We show that the standard methods used in GWAS to correct for confounders are not sufficient to remove spurious signals from EWAS.

2. We present a principled method that removes spurious associations that arise due to cell-type heterogeneity, while yielding true associations of interest. The approach does not require knowledge of cell-type compositions or the reference methylome of purified cell types.

3. We provide freely-available software for the academic community that implements our approach and is scalable to large data sets.

2.3 Results

2.3.1 Overview of the EWASher approach

EWAS confounding by cell-type composition arises because the cell-type composition is correlated with the phenotype, and also with many methylation loci. As a result, loci that are indicative of cell type will appear to be associated with the phenotype even though they are only correlated by way of cell type composition. To alleviate this problem, we use the data set itself to implicitly estimate and correct for this confounding. In particular, we use the genome-wide methylation data to construct a single similarity score between every two individuals. Jointly, these similarity scores are reflective of the relative cell type composition among individuals. Two individuals with similar sample cell type compositions will have a high similarity score, while those with different composition will have a low score. Together, these scores form a similarity matrix, which is then used within the linear mixed model (LMM) to remove associations due to cell-type compositions and to reveal true associations.

The core of EWASher is the LMM. However, when confounding by cell-type heterogeneity is large, the LMM is not able to fully correct for it . Therefore, we augment the LMM with principal component (PC)-based covariates computed from genome-wide methylation values. In particular, EWASher uses an iterative approach that identifies the best hybrid model by automatically selecting the best loci with which to measure similarity for the LMM, as well as the optimal number of PCs.

In line with our theoretical findings and simulations, we found that when confounding due to cell type composition was not too severe, as is the case with the RA data, the LMM alone was sufficient for correction. In solid cancer datasets that suffer from strong confounding, EWASher found that using the top two or three PCs was advantageous. Additionally, as we show, the use of PCs alone (without the LMM) is, in general, insufficient.

2.3.2 DNA methylation association with rheumatoid arthritis

First, we investigated in detail a bronze standard rheumatoid arthritis data set that has reference DNA methylation profiles available for each of the main cell types present in the samples. Next, we performed simulations with synthetic data based on the real data, serving as a gold standard. Finally, we applied our approach to two cancer

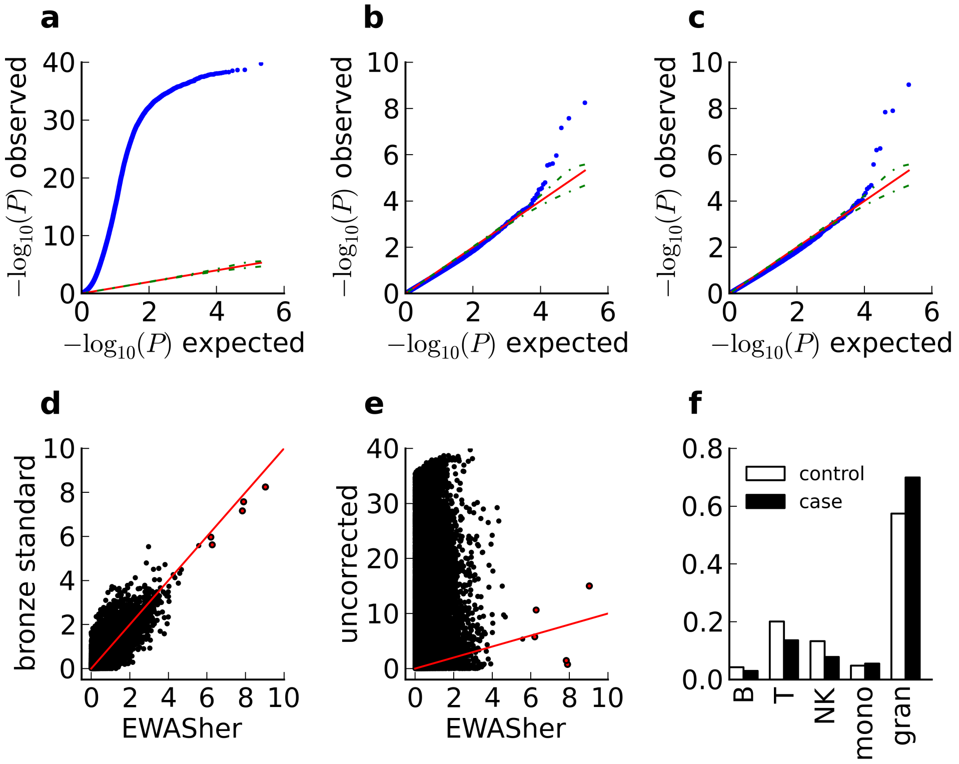


Figure 2.1: RA methylation association study analysis. In preprocessing, data were corrected for gender, batch and smoking status. (a) Quantile-quantile (qq) plot of the log10 P values for association without additional corrections. It shows severely inflated test statistics leading to many false positives. Green dashed lines show the 95% confidence intervals. Large deviations from the diagonal are indicative of inflation. (b) qq plot resulting from use of a bronze standard analysis in which estimated cell type composition was included as a covariate. (c) qq plot resulting from use of EWASher, which did not use knowledge of cell type composition. (d) Paired plot of the log10 P values from (b) and (c). The two approaches found the exact same five significant loci (with the same rankings), even though one required explicit knowledge of the cell type composition. (e) Paired plot of the values from (a) and (c), showing that the correction dramatically altered the rank order of the hypotheses, and that in the uncorrected method, the bronze standard significant loci were swamped by spurious associations. (f) Samples from RA patients had a higher proportion of myeloid cells (granulocytes and monocytes) and fewer lymphocytes (B, T, and NK cells) compared with control samples.

methylation data sets from TCGA that do not have reference profiles available, as is typical.

The models used in this section include: (1) an uncorrected analysis that adjusts only for standard covariates such as age, gender and batch, but does not correct for cell-type composition, (2) a reference-based analysis that corrects for known cell type composition by adding these as covariates to a linear regression, (3) a reference-free PC-based analysis that corrects for cell-type composition by adding PC covariates to the linear regression, (4) EWASher, our new reference-free hybrid LMM-PC approach.

Throughout our experiments we use quantile-quantile (qq) plots of the log10 quantiles to assess inflation of the test statistic in our experiments, as is common in the GWAS community. In these plots, the quantiles of the theoretical null distribution are plotted against the observed quantiles. Under the assumption that no methylation loci in the observed data are differentially expressed, the resulting plot should follow the diagonal and lie within the 95% confidence error bars. Because we expect some, but not too many, methylation sites to be differentially expressed, we expect to see only small deviations from this, and interpret greater deviations as inflation of the test statistic. We also use the genomic control factor, λ, a metric commonly used in GWAS to quantify how much the test statistics are inflated compared to the null distribution. A data set corrected for confounders has λ around 1, while λ significantly greater than 1 indicates confounding and potentially many false positives.

We obtained data from the recent study of DNA methylation association with rheumatoid arthritis (RA). The study collected blood samples from 354 cases and 312 controls, which were assayed on a 450k Illumina DNA methylation chip. After filtering out failed probes and probes that were constitutively methylated or unmethylated in all the samples, we retained 103,638 loci. Effects due to age, gender, smoking status, and batch were removed by regressing these factors on methylation and then using the residuals for all further analysis. In (67), the authors estimated the cell-type composition9 for each sample by using the carefully collected five known reference methylation profiles (CD14 monocytes, CD19 B cells, CD4 T cells, CD56 NK cells, and granulocytes). We use these estimated cell types for a bronze standard analysis to which we compare our approach.

First, we performed an uncorrected analysis to look for associations between each methylation marker and the RA phenotype. We observed severe global inflation of the test statistic, resulting in far too many low P values (Fig. 2.1a). In the RA data, the λ was 10.97, which is significantly more inflated than the values ($\lambda < 1.5$) observed in typical human GWAS with population structure. The blood samples of RA patients showed significant changes in the relative proportions of lymphocytes and myeloid cells compared with healthy control subjects (Fig. 2.1f). Therefore, if a marker was differentially methylated between blood cell types, then it also appeared to be correlated with the phenotype. These associations are biologically uninteresting because they simply tag known differences in cell type composition; the goal of an EWAS here is to find changes in methylation above and beyond these cell type composition associations.

Next, we performed a reference-based cell type corrected analysis by using the estimated cell-type composition as covariates in the regression model used for association analysis. We observed that this approach adequately corrected the genome-wide inflation of test statistics (Fig. 2.1b).

Finally, we applied our new reference-free approach, EWASher to these same data. Just as adding the reference-based covariates successfully corrected for cell type composition, so too did our reference-free approach (Fig. 1c). At the Bonferroni threshold of 5x10^{-7}, the reference-based approach, and the reference-free EWASher both find the same five significantly associated loci (Fig. 2.1d, Table 2.1). The top associated site is in the gene body of HLA-DQA2 for which gene expression levels are associated with RA disease severity. The other two significant loci are in the promoter of NLRC5 which has been shown to regulate inflammatory processes. In the uncorrected model, these biologically meaningful associations were swamped by spurious associations (Fig. 2.1e). For this data set, EWASher found that no PC covariates were needed with the LMM.

It is instructive to examine which loci were selected by the uncorrected association analysis and to understand why they lose significance in the corrected analysis. We selected the most significant loci from the uncorrected linear regression and plotted their methylation level in each of the five reference cell types. These sites were all highly methylated in B, T, and NK cells (lymphocytes) and barely methylated in monocytes and granulocytes (myeloid cells). Given that B, T, and NK cells make up a significantly smaller fraction of the cells in RA cases than in controls, it is not surprising that these loci were strongly associated with the phenotype. In contrast, EWASher did not pick up these lymphocyte-specific loci because it automatically controls for these systematic biases, without the need for explicit knowledge of lymphocytes and myeloid cells. EWASher can thus be used to identify loci that are associated with the phenotype of interest above and beyond the cell-type specific relationshipsknowledge of far greater interest.

A possible alternative approach to our hybrid EWASher approach would be to use only the top principal components (PCs) to adjust for confounding due to cell type distributions, similar to the manner in which EIGENSTRAT uses PCs to correct for population structure in GWAS, and Surrogate Variable Analysis uses PCs to correct for expression heterogeneity in gene expression. We applied such a model, PC covariates with linear regression, to the RA dataset, systematically varying the number of top PC covariates used. Even when a large number of PCs were used, the test statistics were still inflated. This result is consistent with the observation in GWAS that when samples exhibit complex relatedness, linear corrections using PCs cannot fully capture the confounders. In the RA data, complex structure amongst samples was caused by heterogeneous cell-type composition, which is a structure that EWASher accounted for by explicitly modeling the similarity between every pair of samples.

2.3.3 Simulation results

In this section, we used the real RA data as a basis for generating synthetic data sets that could then be used as a gold standard. Briefly: using the empirical distribution of cell types for the cases and controls in the RA dataset, in conjunction with the reference profiles themselves, we generated methylation profiles for a set of cases and controls that had no true associations with the case-controls status (that is, no associations above and beyond those due to cell type composition). On this null-only simulated data, EWASher was well-calibrated (yielded non-inflated P values) and did not produce any false positive markers (Fig. 2.2). Next, in order to examine power, we added in methylation site-specific signals to create true associations with some sites (roughly the same number as we believed to be present in the real data, as judged by using the reference-based analysis). In this setting, EWASher did not produce any false positives above the Bonferroni threshold among loci generated from the null distribution. Moreover, it was able to recover true signals just as well as a model that cheated by using the known (generative) reference cell type composition as covariates (Fig. 2.2f).

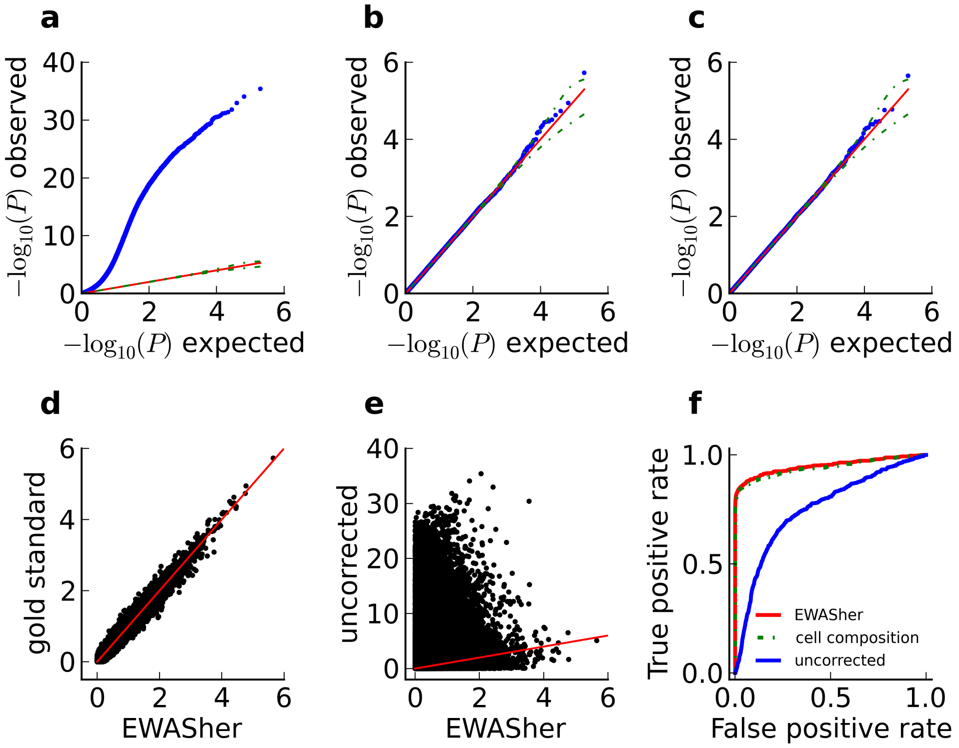


Figure 2.2: We generated synthetic data with cell type composition characteristics based on the actual RA data. EWASher yielded calibrated P values (no inflation) and did not produce any false positives. (a) qq plot of the log10 P values for association from the uncorrected model. Green dashed lines show the 95% confidence intervals. Large deviations from the diagonal are indicative of inflation. The test statistics were severely inflated similar to Fig. 2.1(a). (b) qq plot resulting from the inclusion of the cell type composition used to generate these data as covariates (with linear regression); the P values were calibrated. (c) qq plot resulting from the use of EWASher, which does not use cell type composition. (d) Paired plot of the values from (b) and (c). EWASher had good agreement with the gold standard. (e) Paired plot of the values from (a) and (c). A naive analysis has no agreement with the gold standard. (f) ROC curve demonstrating that EWASher recovered true associations as well as the gold standard method which used the cell-type composition.

Unsurprisingly, both our approach and the reference-based approach, significantly outperformed the uncorrected model. Note that, if the true association signal is present only in a relatively rare cell type, then both EWASher and the reference-based method lose power (as compared to a less rare cell type-based association). In all of these experiments, EWASher did not find the need for any PC covariates.

To test the model in settings with a larger number of cell types (above we used five, as in the original RA paper), we simulated additional datasets with up to 50 cell types. Across all these simulations, EWASher did not find any false positive associations (those passing the Bonferroni threshold). When the confounding effects are large, our mathematical analysis predicts that the LMM alone will be underpowered to correct for them completely. Indeed, in simulations we found that when there are large systematic differences in cell-type composition between cases and controls, P values were no longer adequately calibrated by the LMM alone (i.e., large inflation of the test statistic was observed). However, by adding top principal components to the LMM, EWASher did correct for all the spurious associations even when the confounding effects were large.

2.3.4 TCGA breast and colon cancer methylation analysis

DNA methylation and other epigenetic marks define cellular identity through their regulation of gene expression programs, and are known to change during normal tissue development and differentiation. Dysregulated DNA methylation is associated with disruptions in developmental processes and can lead to unchecked cell proliferation and oncogenesis. This is thought to occur through gains and losses of methylation that are associated with aberrant silencing of tumor suppressor genes and activation of oncogenes. Thus, investigation of epigenetic associations in cancers is of great importance toward understanding these diseases.

Solid tumors are particularly likely to contain a heterogeneous mixture of many cell types, most of which do not have reliable reference methylation profiles. Therefore, analyses of solid tumors methylation profiles are among those analyses most likely to benefit from application of EWASher. This difficulty with solid tumors has likely contributed to the dearth of clinically available DNA methylation biomarkers for the associated diseases, despite thousands of publications reporting methylation changes in hundreds of genes.

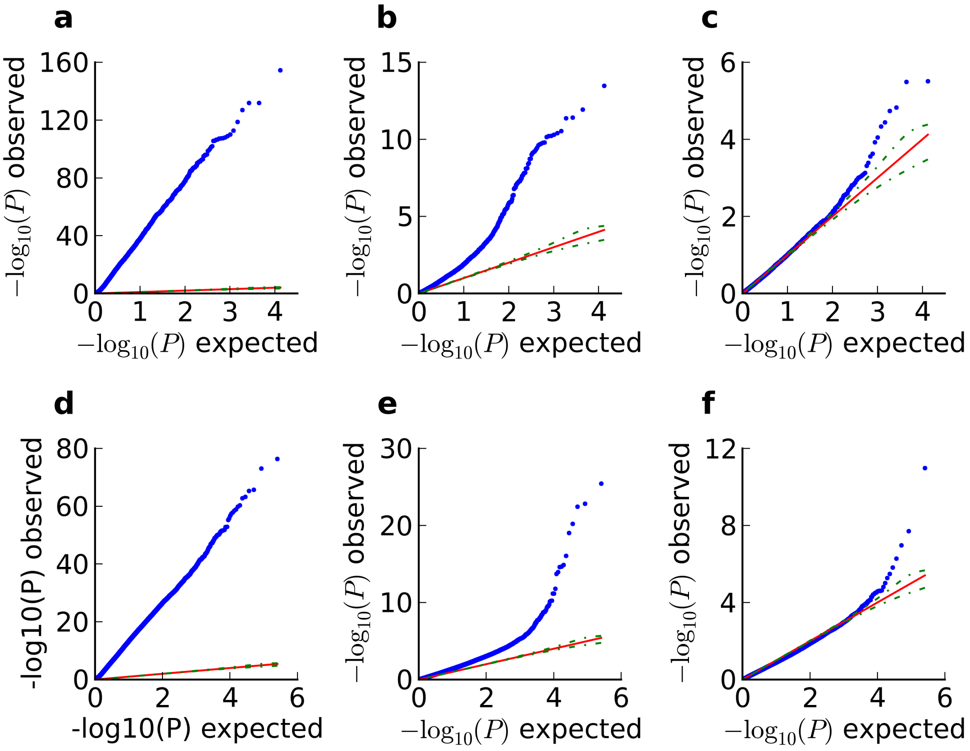


Figure 2.3: DNA methylation association with breast and colon cancer. Quantile-quantile (qq) plot of the log10 P values for association for various methods, with dashed green lines showing the 95% confidence intervals. Large deviations from the diagonal are indicative of inflation. The top row (a-c) are from breast cancer, while the bottom row (d-f) are from colon cancer. In pre-processing, breast cancer data were corrected for batch and breast cancer subtypes; colon cancer data were corrected for age, gender, and anatomic position of the sample. First column (a and d) shows the uncorrected model that did not account for cell type composition. The second column (b and e) shows the PC-only model (with the top 10 PCs as covariates). The test statistics were still inflated. The third column (c and f) shows EWASher, which used the top two (breast cancer) and three (colon cancer) PCs with the LMM.

We analyzed the TCGA breast cancer methylation data set, comprised of 816 cases and 124 controls. First, we performed an uncorrected EWAS analysis on this methylation data. We used the case-control status for the phenotype, and controlled only for batch and breast cancer subtypes (luminal A, luminal B, basal, and Her2). We observed severe inflation of the test statistic, yielding largely useless P values (Fig. 2.3a). Next, we investigated whether adding top PCs as covariates to a linear regression could fully control the inflation, finding that they could not (Fig. 2.3b). Finally, we applied EWASher, our hybrid method, which used two PCs with the LMM. This model suitably corrected the inflation (Fig. 2.3c), and yielded relevant associations as discussed below.

Two markers passed the Bonferroni threshold. The top marker, cg05127924, is in the gene body of FBXW10, an F-box protein. The second marker, cg21504624, is in the promoter of IL11RA, which has recently been suggested as a prognostic biomarker in human breast cancer. In the same TCGA samples, IL11RA had significantly lower expression in the cases, consistent with the increased methylation at its promoter. The next most significant markers, just below the Bonferroni threshold, were in the promoter or gene body of genes RUNX1, TAGLN, TNS1, OPRM1, and RUNX3. RUNX1 has been recently identified as a candidate breast cancer tumor suppressor, while multiple loss of function DNA mutations we observed in RUNX1 in the TCGA survey. OPRM1 is suggested to promote tumor growth and acts as a breast cancer tumor suppressor by targeting estrogen receptor alpha. RUNX3 acts as a tumor suppressor in breast cancer by targeting estrogen receptor alpha. In contrast, a similar literature search for the top ten associated genes in the uncorrected association analysis did not reveal any obvious connections to cancer. Note that global hypo-methylation can affect up to 50% of the genome of certain cancer cells, and such large-scale changes can mask the associations of biologically significant loci. EWASher was able to identify known biomarkers and drivers of breast cancer, above these global effects. We expect that as sample size increases in future EWAS, more loci will be associated above the Bonferroni threshold.

Finally, we used a gene ontology (GO) enrichment analysis to further validate our method. First we identified the nearest genes associated with the top 100 markers from both our method and the uncorrected method, and then analyzed the respective gene sets for enriched categories. The four most enriched GO categories (each with $p < 0.002$) for genes identified through EWASher were: immune response, defense response, induction of apoptosis, and cell proliferation. These findings are consistent with our understanding of the role of immune response in breast cancer, as well as with the general model of cancer-driven proliferating cells. In contrast, the four most enriched GO categories for genes from the uncorrected analysis were not particularly specific to cancer: disulfide bond, signal peptide, secreted, and signaling cascade.

One of the core ingredients of EWASher is the similarity matrix computed from selected loci (and used in the LMM component). This matrix, which contains the pairwise similarity between all individuals in a cohort, can be used to visualize the relationship amongst samples. In a heatmap of the clustered similarity matrix of the breast cancer data, the control samples clustered closely together, while there were three clusters among the cases. Upon closer examination, cluster 1 was enriched for samples diagnosed as Luminal B; cluster 2, for those diagnosed as Basal; and cluster 3, for those diagnosed with Luminal A and B. Visualization of the similarity matrix can thus reveal interesting biological subtypes among samples. It is interesting to note that we still observe cancer subtypes in the similarity matrix even though we had explicitly used the subtypes as covariates in correcting the data. This suggests breast cancer subtypes correlate with nonlinear changes in the methylome and cell-type composition, and cannot be fully accounted by linear covariates.

Next we analyzed the TCGA colon cancer methylation data set comprising 270 cases and 38 controls. After correcting for age, gender, and the anatomy of the sample, we observed severe inflation of the test statistic (Fig. 3.3d). As with breast cancer, adding PCs as covariates to a linear regression did not control all of the inflation (Fig. 3.3e). In contrast EWASher (which used the top three PCs with the LMM) was able to fully control inflation of the test statistic (Fig. 3.3f).

The EWASher methylation analysis yielded three markers that passed the Bonferroni significance threshold. These were in the gene bodies of MYBPC3, C9orf50, and SND1. SND1 is recognized as an oncogene in many cancers, and increase in its expression is correlated to colon cancer progression. Markers just below the Bonferroni threshold of $3.9x10^{-7}$ were in the promoter of SALL3 and the gene bodies of ZMIZ1, LRRC4, and NCOR2. SALL3 was recently discovered to be aberrantly methylated and down-regulated in human hepatocellular carcinoma. ZMIZ1 is an androgen receptor and is aberrantly expressed in a large fraction of human breast, ovarian and colon cancers. LRRC4 is a putative tumor suppressor gene, with potential to decrease growth rates. NCOR2 down regulates target genes by recruiting histone deacetylases, and aberrant expression of this gene is observed in several cancers. GO analysis of the top 100 markers did not reveal any particularly interesting categories for either EWASher or an uncorrected analysis, perhaps owing to the smaller sample size and consequently loss in power relative to the breast cancer data set, or to differences in disease architecture.

As in the breast cancer analysis, we again clustered and visualized the similarity matrix, which yielded two distinct clusters among the cases, and one cluster for the controls). Compared with cluster 1, cluster 2 was significantly enriched for tumor samples extracted from the right colon. This finding is consistent with previous reports that gene methylation varies between the two sides of the colon. We also found that cluster 2 was correlated with a higher incidence of lymphatic invasion ($p < 0.01$), which is supported by the observation that cancer in the right colon correlates with poorer prognosis.

2.4 Discussion

We have demonstrated the utility of our method on a bronze standard RA data set annotated with cell type composition. We showed that our reference-free EWAS approach works just as well as a reference-based approach, both in controlling for false positives and in finding associations of interest. We further validated our approach on gold standard synthetic data sets, demonstrating that EWASher successfully removes false positives, while maintaining just as much power as a reference-based approach. Finally, we applied our approach to TCGA breast cancer and colon cancer data sets, revealing relevant biological associations not obtainable by currently available analysis methods.

The core strength and uniqueness of our method is that it does not require knowledge of cell type compositions and reference cell methylation profiles, information that is both difficult and expensive to obtain, yet performs as well as methods that use such reference information. We believe this method can significantly expand the scope of EWAS, making it possible to conduct EWAS on samples for which it is extremely difficult to measure cell compositions. Additionally, the approach allows users to detect clusters amongst the samples and to visualize their relationships, possibly leading to further insights about the data.

One limitation of our approach is that we cannot analyze each cell type individually, as we do not explicitly decompose each sample into its constituent cell types. Investigating how to adapt our approach for such analyses is of interest. It should be possible to combine our method with a reference-based approach, for cases where not all (or a sufficient number) of the references are known. For example, if one had some reference profiles available, one could include them as covariates, and then use our approach to model the residual confounding. Thus, our model should allow for the use of available prior knowledge.

All experiments herein were univariate tests of one methylation site. However, it could be useful to jointly test multiple sites at once. All of our methods directly generalize to such a setting. Although we have applied a linear model to a binary phenotype, these types of models have previously been applied to case-control data with great success. Additionally, others have provided theoretical arguments for the use of linear models with case control data.

Finally, we note that although our experiments focused on methylation data, we believe that our method, or generalizations thereof, are likely to prove useful for other types of data as well, such as gene expression and DNA hypersensitivity. This is an interesting direction for future investigation.

Table 2.1: Methylation loci significantly associated with RA. - The same five loci passed the Bonferroni threshold in both EWASher and the bronze standard method which used cell type composition as a covariate. The two methods also gave the same estimates of effect sizes.

ID	**Chr**	**Gene**	**Effect (EWASher)**	**Effect (bronze standard)**
cg05428452	6	HLA-DQA2	-0.11	-0.11
cg07839457	16	NLRC5	-0.11	-0.11
cg16411857	16	NLRC5	-0.10	-0.10
cg25372449	6	HLA-DRB5	-0.09	-0.09
cg20821042	6	HLA-DQA2	-0.09	-0.09

2.5 Supplemental methods

RA data was obtained directly from the authors of (67), but is all publically available as indicated in their paper. The TCGA breast and colon cancer DNA methylation data was downloaded from the TCGA data portal: https://tcga-data.nci.nih.gov/tcga/. Our suite of tools includes functionalities to cluster and visualize the similarity matrix as a heatmap, in addition to performing the EWASher association analysis.

We deemed a site to be constitutively methylated if its average probe value across all samples (cases and controls) is above 0.8; and we call a site constitutively unmethylated if its average probe value across samples is below 0.2. Because we look for markers correlated with the phenotype, we remove such constitutive loci from our association analysis.

For GO enrichment analysis, we first had to assign markers to genes using the hg19 UCSC annotations. If a marker was in the promoter or gene-body, we assigned it to that gene. Intergenic markers were not assigned to genes. For the GO analysis, we collect genes associated with the top 100 markers from each method (EWASher and the uncorrected analysis). We then performed gene set enrichment analysis on these two sets of genes using DAVID, and report the most significantly enriched categories based on P values.

EWASher is computationally efficient. Each of the datasets analyzed (RA, breast and colon cancer) were analyzed by EWASher on a single laptop (Lenovo X1 Carbon with 8Gb of RAM) in 1-5 minutes. Furthermore, the linear mixed model backbone of EWASher has been successfully scaled to large GWAS datasets1416, and the EWASher running time is just a constant times the LMM running time. The constant is the number of PCs scanned (typically 1-5). Therefore as the size of EWAS approaches that of GWAS, EWASher will be a fast and memory efficient tool performing genome-wide analyses.

2.5.1 Simulations

To simulate null-only data in such a way as to most closely mimic the actual RA data, we obtained the previously inferred cell type composition for each sample in RA, as well as the reference DNA methylation maps for each of the five blood cell types. To simulate a sample, we took the weighted average of the five reference cell types, with weights given by the previously inferred cell type composition from the real data. We then we added independent and identically distributed Gaussian noise to each marker. The relative performance of EWASher and the method which used the reference-based cell type composition was robust to the amount of noise added (we tried noise standard deviation in the range 0.05 to 0.3). To add synthetic locus-specific signal, we selected a cell type (or a set of cell types) as being differentially expressed between case and control, and then created a case reference methylation profile from the null-only reference methylation profile by making the causal loci systematically higher (or lower) compared to the control reference profile. The samples were then simulated as before, only now using a weighted sum of either case, or control reference profiles, as appropriate.

For the simulations with additional cell types and individuals:

1. We generated synthetic reference cell-types by first breaking the five blood reference methylomes (from the RA dataset) into megabase blocks and then taking random combinations of these blocks to create a new cell type.

2. For simulation with N cell types, we set one N-dimensional Dirichlet distribution for cases and another for the controls.

3. For each case sample, we drew a set of mixture weights from the case Dirichlet, and then took the weighted average of the N reference cells. A similar procedure was conducted for the control samples.

4. We next added identically and independently distributed Gaussian noise to each marker, with a noise standard deviation of 0.1, as this matched the empirical noise distribution. We tried varying the standard deviation from 0.05 to 0.3 and EWASher continued to robustly control inflation of the test statistics.

The description of EWASher uses the concept of the genomic control factor, λ, which is defined as the ratio of the median observed to median theoretical test statistic. When there is no signal in the data, a calibrated result corresponds to $\lambda = 1$, and values of λ substantially greater than 1.0 are indicative of inflation. Methylation values were normalized to be between 0 and 1, and then, within the LMM, further normalized to have mean zero and unit variance.

2.5.2 The EWASher model

EWASher seeks to find the simplest combination of PCs and LMM that control for inflation of the test statistics. It works as follows:

1. Filter out markers that are constitutively high or low.

2. Run the uncorrected association analysis, and rank all the methylation loci by their significance. As in FaST-LMM-Select, we select the top K loci to construct the methylation similarity matrix, where K is automatically determined by maximizing cross-validation likelihood.

3. Using the similarity matrix determined from step 2 with the LMM, compute an association P value for each site. If the genomic control factor, λ, is still inflated (see note below), compute the PCs across all samples. Include the top PC as a covariate, and then rerun the linear regression model to rank all the loci by significance (now conditioned on the first PC). As in step 2, use the selected loci to construct the similarity matrix. This gives the EWASher model comprising the LMM and one PC, which is then used to compute association significance for each site. If λ is still inflated, use the top 2 PCs as fixed covariates, and iterate until the inflation is controlled.

On the colon cancer data set, EWASher with 1 or 2 PCs cannot correct for inflation of test statistics ($\lambda = 2.1$, and 1.1, respectively). With 3 PCs, EWASher sufficiently controls the inflation. For our datasets we set the inflation threshold to be $\lambda = 1$. Note that in GWAS studies with polygenic effects, it has been observed that λ can appear slightly inflated due to signal in the data; thus, setting the threshold of $\lambda = 1$ might be too conservative. In these cases λ values up to say 1.1 can be tolerated. Similarly in EWAS, if practitioners have prior belief that many loci might be truly associated with the phenotype beyond cell composition confounding, then they can experiment by setting the EWASher λ threshold to be above 1(i.e., the algorithm stops after λ falls below this threshold). In studies where controlling for all spurious associations is absolutely crucial, we suggest using the more conservative $\lambda = 1$ threshold.

As in the GWAS applications of the LMM, it is important to select methylation sites when constructing the similarity matrix. If we use all the methylation loci to construct the similarity matrix, then we are likely to introduce additional noise since many loci in the genome do not correlate with cell-type composition. When we used all loci to compute similarity for the RA data, the resulting LMM could not sufficiently correct for inflation of test statistics. Similarly, it is important to re-select markers after adding in an additional PC covariate. Combining PCs with LMM have been suggested in the GWAS context, although, to our knowledge, it has not fully been explored. The idea of performing feature selection for features in the LMM similarity matrix, iteratively, as we condition on increasingly more PCs covariates, is a novel contribution of this paper, and it significantly improves the performance of EWASher.

2.6 Mathematics of EWASher

2.6.1 Overview

EWASher uses a combination of a linear mixed model (LMM) and principal component covariates. LMMs tackle confounders by modeling the similarity between every pair of individuals. We first describe the LMM in Section 2. Then we discuss cell type composition in Section 3, which is a key confounder in EWAS. In this section we also show how our similarity score captures the similarity in cell composition between samples. In Section 4, we discuss the limitations of the LMM, and show mathematically that when the confounder is strong, the LMM alone does not adequately model the confounder. This failing motivates our development of the hybrid LMM + PC model. In Section 5, we discuss the connection between the LMM and PCA, while in Section 6, we describe EWASher.

2.6.2 Linear mixed models

In a linear mixed model (LMM), the phenotype $\mathbf{y}$ is expressed as a sum of fixed effects $\boldsymbol{\beta}, \beta_s$ and random effects $\mathbf{u}$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X_s}\beta_s + \frac{1}{\sqrt{M}}\mathbf{G}\mathbf{u} + \boldsymbol{\epsilon}. \tag{2.1}$$

- If there are N samples, then $\mathbf{y}$ is a N-by-1 vector. In this paper, $\mathbf{y}$ is a binary phenotype corresponding to case/control status (see main paper for discussion of application of a linear model to binary phenotypes).

- $\mathbf{X}$ is a matrix of known covariates, such as age, sex, batch, etc.

- $\mathbf{X_s}$ is a N-by-1 vector of the methylation values at site s across all samples. Site s is being tested for its univariate association with $\mathbf{y}$.

- $\mathbf{G}$ is a N-by-M matrix whose (i, j) entry is the methylation value of marker j in sample i. In practice, we assume that the markers are normalized so that the columns of $\mathbf{G}$ have zero mean and unit variance.

- $\mathbf{u}$ is a M-by-1 vector whose entries are identically and independently distributed (iid) samples from a Gaussian: $\mathbf{u} \sim N(0; \sigma_g^2 \mathbf{I_M})$, where $\mathbf{I_M}$ is the M-by-M identity matrix.

- $\boldsymbol{\epsilon} \sim N(0; \sigma_e^2 \mathbf{I_N})$ is iid random noise.

We would like to estimate $\beta_\mathbf{s}$ for every methylation site in the data, and compute the P value for its significance. The null model is that $\beta_s = 0$ and

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \frac{1}{\sqrt{M}}\mathbf{G}\mathbf{u} + \boldsymbol{\epsilon}. \tag{2.2}$$

The p value is computed by comparing the ratio of the alternative and the null likelihoods to the χ_1^2 distribution.

Two equivalent views of LMM. One way to think of the LMM is, as in Eqn. 2.1, a Bayesian regression that integrates over the random effect $\mathbf{u}$. Equivalently, we can also think of it as a multivariate Gaussian. The likelihood of $\mathbf{y}$ after integrating over $\mathbf{u}$ is

$$\int N(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{X_s}\beta_s + \frac{1}{\sqrt{M}}\mathbf{G}\mathbf{u}; \sigma_e^2\mathbf{I_N})N(\mathbf{u}|0; \sigma_g^2\mathbf{I_M})d\mathbf{u} = N(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{X_s}\beta_s; \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I_N})$$

where $\mathbf{K} = \frac{1}{M}\mathbf{G}\mathbf{G^T}$ is the similarity matrix. The entry (i, j) of $\mathbf{K}$ captures how similar the methylome of sample i is to that of sample j.

2.6.3 How the LMM captures cell type composition

Suppose there are T reference cell types for the EWAS problem on hand. Let $\mathbf{R}$ be the M-by-T matrix of methylation values for each reference type. That is, $R(s, t)$ is the methylation of site s in cell type t. Each sample i can, in principle, be characterized by a cell-type composition vector $\mathbf{W_i} = (w_{i1}, ..., w_{iT})^T$ (which may be known or unknown). The fraction of sample i that is cell type t is w_{it} and $\sum_t w_{it} = 1$. The methylation profile of sample i is then $\mathbf{G_i} = \mathbf{R}\mathbf{W_i}$. The similarity between samples i and j is $K(i, j) = \frac{1}{M}\mathbf{G_i}\mathbf{G_i^T} = \frac{1}{M}\mathbf{W_i^T}\mathbf{R^T}\mathbf{R}\mathbf{W_j}$. We can decompose $\mathbf{R^T}\mathbf{R}$ using a spectral decomposition, $\mathbf{R^T}\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where Λ is diagonal and contains the eigenvectors, while $\mathbf{U}$ contains the eigenvectors. What's important to observe is that if the cell types are equally dissimilar to one another, then all the diagonal entries (eigenvalues) are approximately equal. In this case, $K(i, j) \propto \mathbf{W_i^T W_j}$ (because $\mathbf{UU^T} = \mathbf{I}$), and therefore it can directly be seen that computation of the kernel $\mathbf{K}$ directly models the cell type composition. In practice, even with small deviations from this assumption, we expect the LMM to do well in modeling cell type heterogeneity (indeed, this bears out in our experiments in the main paper).

Note that on real and simulated data, we find that the LMM best removes spurious signals if, instead of using all the methylation sites to compute $\mathbf{K}$, we select for loci that are highly correlated with the phenotype as columns of $\mathbf{G}$. This selection tends to pick out loci that are tagging cell-type composition.

2.6.4 Limits of the LMM

We analyze a simple example below to illustrate that the LMM alone does not adequately capture strong confounders when the sample size is large. The main idea here will be to (1) assume a particular methylation site that is strongly predictive of the phenotype, due to confounding, (2) use the LMM to condition on just this site, while testing this site, and (3) show that such conditioning does not entirely remove the signal, and therefore that this site would appear as significant. The consequence of this conclusion is that conditioning on a variable by using it as a random effect (i.e. in the similarity matrix of the LMM), does not always always fully control for that variable. We now formally go through this argument.

Assume that we have a N-by-1 binary phenotype $\mathbf{y}$, and one marker that we want to test, $\mathbf{x} = \mathbf{y} + \boldsymbol{\delta}$, where, where $\boldsymbol{\delta} \sim N(0; \sigma^2\mathbf{I_N})$. It turns out to be more convenient to work in a different coordinate system, so we now rotate the data in such a way that $\mathbf{x} = [||\mathbf{x}||, 0, ..., 0]^T$. (The intuition here is that $\mathbf{x}$ is a vector that points to one particular direction in N-dimensional space, and therefore we can pick out this direction as our first coordinate, so that the now rotated $\mathbf{x}$ becomes nothing more than a vector with some magnitude in this one coordinate, and no magnitude in any other direction.) When N is large, $||\mathbf{x}||^2 \approx N_{case} + N\sigma^2$. Because $\boldsymbol{\delta}$ is sampled from a spherically symmetric Gaussian, its rotation, $\boldsymbol{\delta'}$, is also a sample from $N(0; \sigma^2\mathbf{I_N})$. Thus, we now have $\mathbf{y} = [||\mathbf{x}||, 0, ..., 0]^T - \boldsymbol{\delta'}$.

Suppose $\mathbf{x}$ is correlated with $\mathbf{y}$ due to very strong confounding. In this case, we would like the LMM to learn that it is not significant. An extreme scenario is to use the same $\mathbf{x}$ to construct the similarity matrix: $\mathbf{K} = \mathbf{x}\mathbf{x}^{\mathbf{T}}$. The null model likelihood (assuming no covariates for simplicity but without loss of generality) is $N(\mathbf{y}|0; \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I_N})$. The covariance matrix of the Gaussian is nicely diagonal $diag[\sigma_g^2||\mathbf{x}||^2 + \sigma_e^2, \sigma_e^2, ..., \sigma_e^2]$. When N is large, it will turn out that $\sigma_g^2||\mathbf{x}||^2 >> \sigma_e^2$. So to keep the exposition simple, we can approximate the covariance matrix by $diag[\sigma_g^2||\mathbf{x}||^2, \sigma_e^2, ..., \sigma_e^2]$. The log-likelihood under the null model can be expressed in the factored form

$$LL_{null} \approx -\frac{1}{2}\log 2\pi\sigma_g^2||\mathbf{x}||^2 - \frac{||\mathbf{x}||^2}{2\sigma_g^2||\mathbf{x}||^2} - \frac{N}{2}\log 2\pi\sigma_e^2 - \frac{||\boldsymbol{\delta}'||^2}{2\sigma_e^2}. \tag{2.3}$$

Solving for σ_e^2 and σ_g^2 that maximizes LL_{null} and plugging it in yields

$$LL_{null} \approx -\frac{1}{2}\log 2\pi||\mathbf{x}||^2 - \frac{N}{2}\log 2\pi||\boldsymbol{\delta}'||^2 - 1. \tag{2.4}$$

The alternative model likelihood is $N(\mathbf{y}|\mathbf{x}; \sigma_g^2\mathbf{K}+\sigma_e^2\mathbf{I_N})$. Using the fact that $\mathbf{x} = \mathbf{y} + \boldsymbol{\delta}$, this can be rewritten as $N(\boldsymbol{\delta}|0; \sigma_g^2\mathbf{K}+\sigma_e^2\mathbf{I_N})$. Because we optimize over σ_g^2 and σ_e^2, we can set $\sigma_g^2 = 0$ and obtain a lower bound on the alternative likelihood of $N(\delta|0; \sigma_e^2\mathbf{I_N})$. The log-likelihood is

$$LL_{alt} \geq -\frac{N}{2}\log 2\pi\sigma_e^2 - \frac{||\boldsymbol{\delta}||^2}{2\sigma_e^2}. \tag{2.5}$$

After optimizing σ_e^2, we have

$$LL_{alt} \geq -\frac{N}{2}\log 2\pi||\boldsymbol{\delta}||^2 - 0.5. \tag{2.6}$$

Since $||\boldsymbol{\delta}||^{\mathbf{2}} = ||\boldsymbol{\delta}'||^{\mathbf{2}}$ due to rotational invariance, the difference between the alternative and null log-likelihood is

$$LL_{alt} - LL_{null} \geq \frac{1}{2}\log 2\pi||\mathbf{x}||^2 \tag{2.7}$$

Therefore as the number of samples increases, the difference between the alternative and null model likelihoods increases approximately as $\log(N_{case} + N\sigma^2)$, at some point yielding a significant likelihood ratio test statistic (twice the difference in likelihoods). So even when we condition on $\mathbf{x}$ by way of including it as a random effect, we find that testing $\mathbf{x}$ as a fixed effect can be significant (the larger the sample size, the more significant). In other words, $\mathbf{x}$ remains a significant predictor of $\mathbf{y}$ even when conditioning on $\mathbf{x}$ itself by way of the LMM (i.e., use it to build the similarity matrix in the LMM). While very simple, this example sheds light on the limitations of the LMM in correcting for strong confounders when the sample size is large, and motivates our proposed hybrid PC + LMM approach. The setting of this example can arise when the cases all contain a significant fraction of one cell type, which is present only in low amount in the controls. Then if $\mathbf{x}$ is a site that is specifically methylated in this cell type, $\mathbf{x}$ will be strongly correlated with the phenotype, $\mathbf{y}$, as in our example above.

2.6.5 The connection between PCs and LMM.

In this section we show that using PCs as fixed effect covariates is equivalent to using the maximum likelihood approximation to the Bayesian LMM regression.

Recall from Eqn. 1 that the linear mixed model can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X_s}\beta_s + \frac{1}{\sqrt{M}}\mathbf{G}\mathbf{u} + \boldsymbol{\epsilon}$, where $\mathbf{u}$ are the random effects that correct for confounding structure in $\mathbf{G}$. After integration of $\mathbf{u}$, we're left with the marginal form, $N(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{X_s}\beta_s; \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I_N})$, where $\mathbf{K} = \frac{1}{M}\mathbf{G}\mathbf{G^T}$ is the similarity matrix. Also recall that when using PC covariates to correct for confounding, we would add the leading principal components from $\mathbf{G}\mathbf{G^T}$ to a linear regression.

Now consider the singular value decomposition of $\frac{1}{\sqrt{M}}\mathbf{G}$ given by $\frac{1}{\sqrt{M}}\mathbf{G} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{B}$, where $\mathbf{A}$ is the N-by-N left singular matrix, $\mathbf{B}$ is the M-by-M right singular matrix, and Λ is a diagonal matrix. The columns of $\mathbf{A}$ are the eigenvectors of $\mathbf{G}\mathbf{G^T}$ or, equivalently, the principle components of $\mathbf{G}$. Then we can rewrite $\frac{1}{\sqrt{M}}\mathbf{G}\mathbf{u}$ with $\mathbf{u} \sim \mathbf{N}(\mathbf{0}; \sigma_\mathbf{g}^\mathbf{2}\mathbf{I_M})$ as $\frac{1}{\sqrt{M}}\mathbf{G}\mathbf{u} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{v}$ with $\mathbf{v} \sim \mathbf{N}(\mathbf{0}; \sigma_\mathbf{g}^\mathbf{2}\mathbf{B}\mathbf{B^T})$. A low rank approximation is then $\frac{1}{\sqrt{M}}\mathbf{G}\mathbf{u} \approx \sum_{i=1}^{L}\mathbf{A_i}\lambda_i v_i$, where $\{\mathbf{A_i}\}_{i=1}^{L}$ are the top L principal components. If we were to take only take the top few PCs, we could estimate v_i explicitly without overfitting, and therefore we could treat v_i not as random variables to be integrated out, but instead estimate them by way of maximum likelihood. Therefore the low-rank maximum likelihood approximation to the linear mixed model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X_s}\beta_s + \sum_{i=1}^{L}\mathbf{A_i}\lambda_i v_i + \boldsymbol{\epsilon}, \tag{2.8}$$

which is precisely the PC-based correction that is used in GWAS and similar studies. Note that this PC approximation is only good if $\mathbf{G}$ is actually low-rank. When there are many latent cell types and when there are other hidden confounders, $\mathbf{G}$ is no longer low rank. In real data sets that contain such effects, such as rheumatoid arthritis, breast cancer and colon cancer, we find that the top PCs do not sufficiently model the confounders, resulting in inflated test statistics.

2.6.6 EWASher

EWASher augments the linear mixed model with top PCs as covariates so that it can correct for confounding even when it is strong. The model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X_s}\beta_s + \sum_{i=1}^{L} \mathbf{A_i}\lambda_i v_i + \frac{1}{\sqrt{M}}\mathbf{G}\mathbf{u} + \boldsymbol{\epsilon} \tag{2.9}$$

where $\boldsymbol{\beta}$, β_s and $\{v_i\}_{i=1}^{L}$ are fixed effects and $\mathbf{u}$ is the random effect. L is the number of top PCs used. We use an iterative procedure to find the minimum L that corrects for test statistic inflation as captured by the genomic control factor λ. This corresponds to finding the simplest model, i.e. one with the fewest number of parameters, that controls the inflation. The algorithm is initialized with $L = 0$ which is just the standard LMM of Eqn. 2.1. If this model has inflated test statistics ($\lambda > 1$), we then add the top one PC to the model. If this model is still insufficient, we set $L = 2$ and iterate. Note that because the LMM portion of the model does feature selection to build the similarity matrix, $\mathbf{K}$, one could choose to redo this feature selection upon addition of each new PC. In practice, we find that such redoing of the feature selection is essential.

To compute the PCs in EWASher, we represent each sample as an M-by-1 vector, where M is the total number of measured methylation loci, and perform the standard PCA. As an alternative approach, we have investigated performing feature selection for the PCA, whereby each sample is represented by a subset of loci. These loci can be selected through univariate regression against the phenotype, as with the LMM. Such an approach is similar to discriminative PCA. On real and simulated data, we did not observe significant differences in results from using the standard PCA or the feature selection PCA in EWASher, and therefore chose to go with PCA without feature selection.

3

Genome-wide Analysis Reveals Conserved and Divergent Features of Notch1/RBPJ Binding in Human and Murine T Lymphoblastic Leukemia Cells

3.1 Overview

Notch1 regulates gene expression by forming transcription activation complexes with the DNA-binding factor RBPJ and is oncogenic in murine and human T cell progenitors. We used ChIP-Seq to identify Notch1 and RBPJ binding sites in human and murine T-LL genomes. In both species, Notch1 binds preferentially to promoters, to RBPJ binding sites, and near imputed ZNF143, Ets and Runx sites. ChIP-Seq confirmed that ZNF143 binds to 40% of Notch1 sites. Notch1/ZNF143 sites are characterized by high Notch1 and ZNF143 signals, frequent co-binding of RBPJ (generally through sites embedded within ZNF143 motifs), strong promoter bias, and lower levels of activating chromatin marks than other classes of Notch1-binding sites. K-means clustering of Notch1 binding sites and associated motifs identified conserved Notch1-Runx, Notch1-Ets, Notch1-RBPJ, Notch1-ZNF143, and Notch1-ZNF143-Ets clusters with different genomic distributions and levels of chromatin marks. Although Notch1 binds mainly to gene promoters, 75% of direct target genes lack promoter binding and are presumably regulated by enhancers, which were identified near MYC, DTX1, IGF1R, IL7R and the GIMAP cluster. Human and murine T-LL genomes also have many sites that bind only RBPJ. Murine RBPJ only sites are highly enriched for imputed REST sites, whereas human RPBJ only sites lack REST motifs and are more highly enriched for imputed CREB sites. Thus, there is a conserved network of cis-regulatory factors that interacts with Notch1 to regulate gene expression in T-LL cells, as well as novel classes of divergent RBPJ only sites that also likely regulate transcription.

3.2 Introduction

Notch receptors participate in a highly conserved signaling pathway that regulates development and tissue homeostasis. Signaling is mediated through a series of proteolytic cleavages, the last carried out by γ-secretase, that permit the Notch intracellular domain (NICD) to translocate to the nucleus and form a transcriptional activation complex with the DNA-binding factor RBPJ (also known as CSL in vertebrates and Su(H) in flies; for recent review, see (79)). This complex in turn recruits a Mastermind (MAML) family member, other co-activators, and mediator complex components, leading to regulated target gene expression. NICD-containing transcription activation complexes are believed to turn over rapidly due to transcription-coupled degradation. In the absence of NICD, RBPJ can also bind several transcriptional co-repressors, supporting models in which RBPJ functions like a transcriptional switch. However, genome-wide studies of RBPJ and NICD binding have yet to be done in higher organisms and other more complex modes of RBPJ/Notch1 interaction with genomes can be envisioned.

Notch effects are variable, depending both on dosage and the chromatin state of the signal-receiving cell. Dysregulated Notch signaling has been implicated in human developmental disorders and cancers, most notably T lymphoblastic leukemia/lymphoma (T-LL), in which the majority of human and murine tumors acquire somatic gain-of-function mutations in Notch1 that increased nuclear levels of NICD1 (37). Dysregulated RBPJ-dependent gene expression is also seen in other cancers, particularly B cell tumors associated with Epstein-Barr virus (EBV), which encodes viral proteins that bind RBPJ and modify gene expression (34, 55, 56, 71). With few exceptions, however, activating Notch mutations have been confined to T-LL, suggesting that high-level Notch signaling is uniquely transforming within the epigenetic context of T cell progenitors.

Recent data indicate that transcription factors bind widely throughout genomes and that individual transcription factor binding sites show substantial evolutionary divergence (15, 84). To identify conserved relationships of likely fundamental importance, as well as potentially interesting points of evolutionary divergence, we performed ChIP-Seq for Notch1 and RBPJ in human and murine T-LL cell lines.

3.3 Results

Notch1 and RBPJ binding sites in human T-LL cell genomes. Initial ChIP-Seq studies were done with the Notch1-dependent human T-LL cell line CUTLL1 (60). Duplicate ChIP-Seq libraries prepared with antisera to RBPJ and Notch1 showed reproducible enrichment of DNA sequences ($10x10^6$ to $25x10^6$ reads per library) that aligned to the human genome. Analysis of Chip-Seq data pooled from the CUTLL1 Notch1 and RBPJ libraries identified 3846 high confidence Notch1 binding sites and 2112 high confidence RBPJ binding sites (false discovery rate (FDR)< 0.01) (Fig. 3.1A). Only 36% of Notch1 peaks overlapped with RBPJ binding sites, while 66% of RBPJ peaks overlapped with Notch1 binding peaks. Most Notch1 and RBPJ binding sites were found in promoters, defined as regions within 2 kb of annotated refseq gene transcriptional start sites (TSSs). In line with prior studies in Drosophila suggesting that genomic RBPJ binding is stabilized by NICD (17), RBPJ and Notch1 ChIP-Seq signals were higher at sites where both bound (Fig. 1B) ($p < 10^{-6}$ for both comparisons).

3.3.1 Transcription factor motifs enriched at sites of Notch1 and RBPJ binding in human T-LL cells.

To gain insight into factors that influence Notch1 and RBPJ binding to T-LL genomes, binding data for recombinant RBPJ on oligonucleotide arrays (protein binding microarrays, or PBMs) (26) were used to identify the highest affinity RBPJ binding site within 100bp of the center of each Notch1 binding site. Both RBPJ/Notch co-sites and Notch1 only sites were significantly enriched for high-affinity RBPJ binding sequences compared to random genomic sequences ($p < 10^{-10}$), suggesting that many Notch1 only sites also bind RBPJ. The promoter localization of RBPJ/Notch1 and Notch1

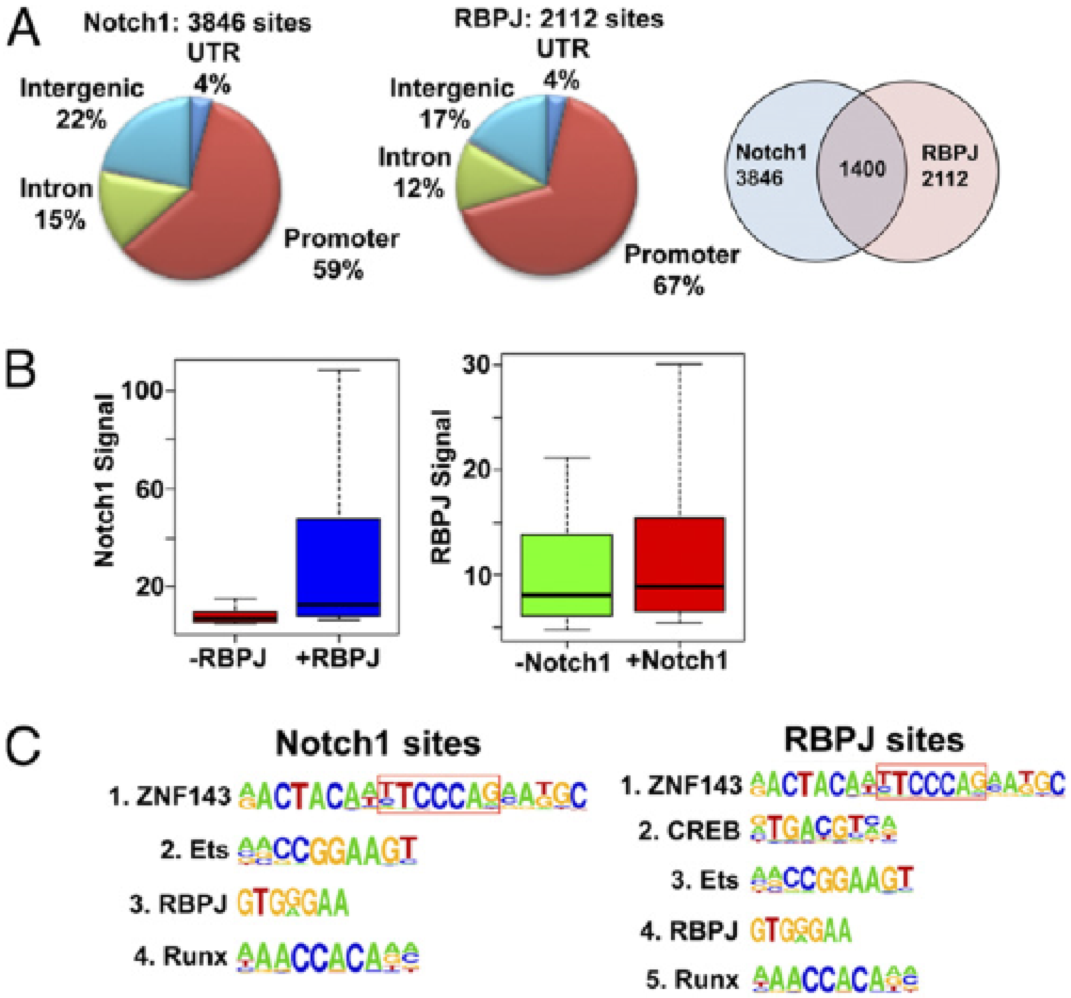


Figure 3.1: Notch1 and RBPJ binding sites in human TLL cells. (A) Distribution and overlap of RBPJ and Notch1 binding peaks in CUTLL1 cells. Promoter regions are defined as sequences within 2 kb of the TSS of annotated genes. (B) Notch1 and RBPJ ChIP-Seq signals (expressed as reads per kilobase per million aligned reads, RPKM) at Notch1-only sites, RBPJ-only sites, and RBPJ/ Notch1 cosites. (C) Enriched transcription factor motifs lying 250 bp of Notch1 and RBPJ binding sites. A RBPJ consensus motif embedded within the extended ZNF143 motif is boxed.

only sites was not significantly different (63% versus 59%, respectively). Failure to detect RBPJ at sites with RBPJ consensus sequences may stem from shielding of RBPJ epitopes within certain complexes. It is also possible that some Notch1 only sites result from RBPJ-independent binding of Notch1 to other chromatin-associated proteins or are an artifact of cross-linking of bona fide Notch1 sites to other regions of the genome via chromatin looping.

Additional insights came through a search for transcription factor motifs enriched within 250 base pairs of the center of each RBPJ and Notch1 binding site. For Notch1 sites the most enriched motif (compared to its overall genomic frequency) was that of the zinc finger protein ZNF143, followed by those for Ets and Runx factors ($p < 10^{-50}$ for each motif) (Fig. 3.1C); these imputed factors in the same rank order were associated with both Notch1/RBPJ co-sites and Notch1 only binding sites. ZNF143 was also the most enriched motif associated with RBPJ binding sites, followed by the motifs for CREB, Ets factors, and Runx factors ($p < 10^{-50}$) (Fig. 3.1C). Of note, the ZNF143 consensus motif contains an embedded high-affinity site for RBPJ, providing a likely explanation for association of RBPJ and Notch1 with these sites. Overall, the CREB consensus motif was found in association with 46% of the RBPJ only sites, as compared to 25% of Notch1/RBPJ co-sites ($p < 10^{-6}$). Compared to RBPJ sites without CREB motifs, the 588 RBPJ/CREB sites had lower affinity RBPJ binding sites (based on PBM score, $p < 10^{-10}$), yet also had higher RBPJ signals, suggesting that other factors (e.g., protein-protein interactions) contribute to RBPJ association with imputed CREB sites.

3.3.2 Confirmation of ZNF143 association with RBPJ and Notch1 binding sites.

Western blotting confirmed the presence of activated Notch1, RBPJ, ZNF143, the Ets factors GABPA and Ets1, Runx1, and CREB in 3 human and 2 murine Notch1-dependent T-LL cell lines. In local ChIP assays performed on 4 to 5 randomly selected Notch1 binding elements with imputed ZNF143, GABPA, Ets1, or Runx1 sites, 17 of 17 sites tested showed significant binding by each factor, suggesting a high fraction of imputed sites are bound by the corresponding transcription factor. This is expected for Ets factors and Runx factors, which bind widely to the genomes of Jurkat T-LL cells (51, 52), are required during early stages of T cell development (49), and interact functionally with Notch1 in early hematopoiesis (9).

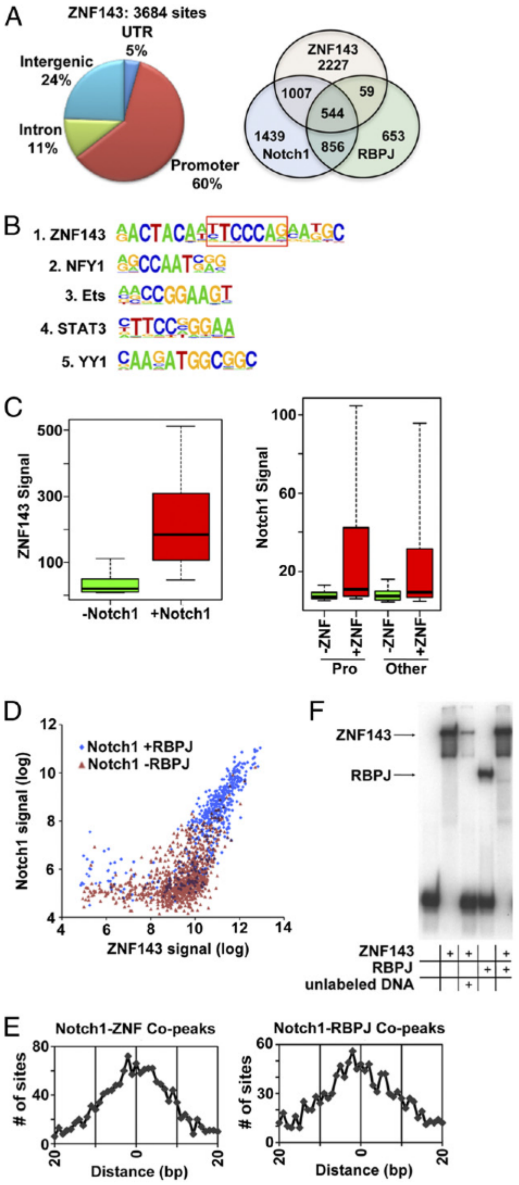
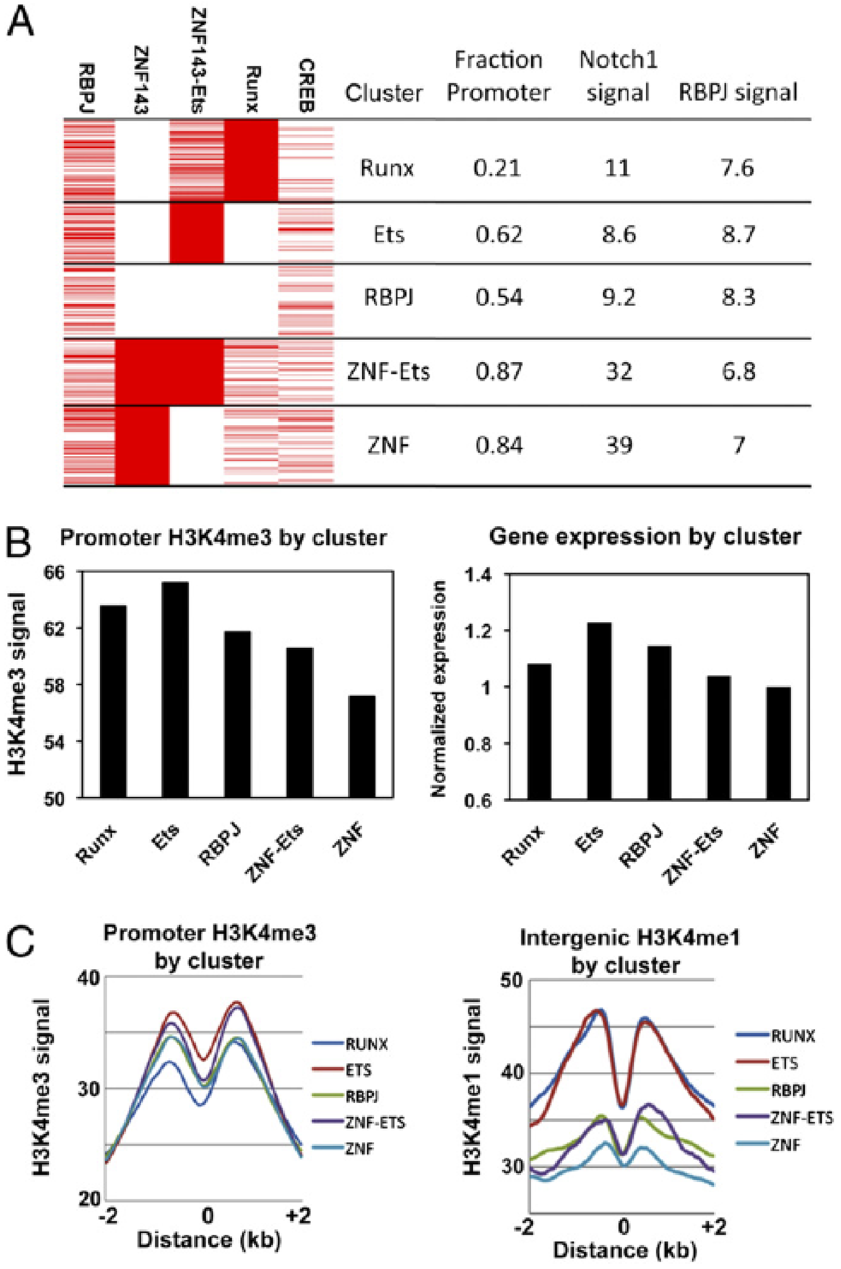


Figure 3.2: Frequent association of ZNF143 and Notch1 binding sites in human TLL cells. (A) Distribution of ZNF143 binding peaks and overlap with RBPJ and Notch1 sites in CUTLL1 cells. (B) Enriched transcription factor motifs lying 250 bp of ZNF143 binding sites. (C) ZNF143 and Notch1 ChIP-Seq signals (expressed as RPKM) at ZNF143 sites with and without Notch1, and Notch1 promoter (Pro), and nonpromoter sites with and without ZNF143. (D) Correlation of Notch1 and ZNF143 signals on individual cosites sites with or without RBPJ signals. (E) Overlap of Notch1 sites with ZNF143 and RBPJ binding sites in CUTLL1 cells, expressed as the distance in base pairs between the centers of Notch1 and ZNF143 copeaks (Left) and Notch1 and RBPJ copeaks (Right). (F) EMSA done with an oligonucleotide containing the SKP2 promoter ZNF143/Notch1 cosite. ZNF143 and RBPJ were added at concentrations of 280 and 407 nM, respectively. In one lane, an unlabeled probe was added in 200-fold excess over the labeled probe.

In contrast, little is known about ZNF143. To confirm the association of ZNF143 with Notch1 and RBPJ, ZNF143 ChIP-Seq was carried out in CUTLL1 cells (Table S1), resulting in the identification of 3684 high confidence ZNF143 binding sites (FDR¡0.01) (Fig. 3.2A). The genome-wide distribution of ZNF143 binding was similar to that of Notch1and RBPJ, with 60% of the binding sites located in promoters. Remarkably, 40% (n=1551) of the ZNF143 binding peaks lay within 100 bp of the center of Notch1 binding sites, and 15% (n=544) overlay RBPJ/Notch co-peaks (Fig. 3.2A). Motif analysis confirmed that the most highly enriched associated motif was that of ZNF143 (Fig. 3.2B). Among the ZNF143 sites overlapping with Notch1 peaks, 67% (n=1044) contained the ZNF143 consensus motif, and of these 57% (n=591) harbored an embedded high-affinity RBPJ binding site. ZNF143 signals were higher at sites where Notch1 also bound, as were Notch1 signals at sites with ZNF143 co-binding (Fig. 3.2C)(p¡10-100 for each). Consistent with these associations, Notch1 and ZNF143 signals correlated at co-sites (R2=0.66), with co-sites having high confidence RBPJ signals showing the highest Notch1/ZNF143 co-signals (Fig. 3.2D). In line with the possibility that Notch1 is associating with these sites via RBPJ binding sequences embedded within ZNF143 motifs, on average the positions of ZNF143 and Notch1 co-peaks coincided exactly (within 2bp), closely approximating the degree of overlap between RBPJ and Notch1 co-peaks (Fig. 3.2E). Prior in silico analysis and local ChIP studies suggest ZNF143 preferentially binds bi-directional promoters (39, 66). The ZNF143 sites identified in CUTLL1 cells confirmed this, as out of 1048 bi-directional promoters in the genome, 285 (27%) bound ZNF143, and of these, 173 also bound Notch1. Of note, NFY1, Ets, and YY1 motifs, each of which are enriched near ZNF143 sites (Fig. 3.2B), are all also enriched in bi-directional promoters.

3.3.3 Identification of distinct classes of Notch1 and RBPJ binding sites.

Cluster analysis was used to investigate whether associated transcription factors sort Notch1 binding sites into different classes. Each Notch1 binding site in CUTLL1 cells was represented as a five-dimensional vector corresponding to the presence or absence of RBPJ or ZNF143 binding, or imputed Ets, Runx, and CREB sites. K-means clustering of all 3846 Notch binding sites revealed 5 major clusters named Runx, Ets, ZNF-Ets, ZNF, and RBPJ for the predominant cofactor(s) (Fig. 3.3A). The Runx



Cluster	Fraction Promoter	Notch1 signal	RBPJ signal
Runx	0.21	11	7.6
Ets	0.62	8.6	8.7
RBPJ	0.54	9.2	8.3
ZNF-Ets	0.87	32	6.8
ZNF	0.84	39	7

Figure 3.3: Transcription factors associated with Notch1 binding sites define distinct classes of putative response elements in TLL cells. (A) K-means clus- tering of Notch1 binding sites in CUTLL1 cells. In the clusters shown at the left, each red line represents a genomic Notch1 binding site where cobinding (RBPJ, Znf143) or a motif (ETS, RUNX, CREB) for the indicated factor is seen. The numbers at the right are the fraction of binding sites within each cluster that are found within promoters of annotated genes and the associated Notch1 and RBPJ ChIP-Seq signal strengths (RPKM). (B) Mean histone H3K4me3 signals (Left) and relative gene expression (Right) for genes with Notch1 promoter binding by cluster designation. Normalized gene expres- sion was determined by expression profiling of CUTLL1 cells. (C) Distribution of H3K4me3 signals in Notch1-binding promoters (Left) and H3K4me1 sig- nals in Notch1-binding intergenic sites (Right) by cluster designation. 0 marks the center of the associated Notch1-binding peaks.

cluster showed a bias for non-promoter sites, whereas the ZNF143 and ZNF143-Ets clusters had a strong promoter bias and high Notch1 signals (Fig. 3.3A). To further characterize these classes, Chip-Seq for activating H3K4me1 (enhancer) and H3K4me3 (promoter) histone marks as well as repressive H3K27me3 marks were performed in CUTLL1 cells. Subtle but statistically significant differences in H3K4me3 signals were noted among the various clusters at promoters, with the highest marks (Fig. 3.3B, $p < 0.01$) and the greatest nucleosome displacement (Fig. 3.3C) being associated with the Ets and Runx clusters, and the lowest marks and least nucleosome displacement with the ZNF143 cluster. Consistent with these measurements, mean steady state gene expression (described below) was highest for genes with Ets sites in their promoters and lowest for genes with ZNF sites (Fig. 3.3B, $p < 0.01$). More striking differences were observed at non-promoter sites, with the highest H3K4me1 signals and greatest nucleosome displacement being associated with the Ets and Runx clusters and the lowest with the RBPJ, ZNF143-Ets and ZNF143 clusters (Fig. 3.3C, $p < 0.001$). Promoter and intergenic ZNF143 sites that did not bind Notch had lower active chromatin marks ($p < 0.0001$ for both comparisons) and higher repressive marks (H3K27me3) (Fig. S5C, $p < 10^{-100}$) compared to those that did, suggesting that ZNF143 only sites may be associated with repressive complexes.

The chromatin marks of RBPJ only (no Notch1) binding sites were also assessed. RBPJ sites without Notch1 had significantly lower intergenic H3K4me1 and promoter H3K4me3 signals ($p < 0.05$); furthermore, RBPJ only sites associated with CREB motifs had lower intergenic H3K4me1 and promoter H3K4me3 signals than RBPJ only sites without CREB motifs ($p < 0.05$). These findings are consistent with a repressive role for RBPJ in the absence of NICD.

3.3.4 Genomic Notch1 binding sites and gene regulation.

To identify genes that are dynamically regulated by canonical Notch1 signaling, we depleted CUTLL1 cells of NICD1 with the γ-secretase inhibitor (GSI) compound E, and then restored NICD1 through GSI washout, which reactivates Notch1 even in the presence of protein synthesis inhibitors, leading to re-loading of Notch1 on dynamic response elements and target gene activation (56). Additional controls included transduction of cells with dominant negative MAML1, a specific antagonist of canonical Notch1 signaling, prior to Notch1 reactivation, and a mock GSI washout to control for cycloheximide effects. High-confidence direct canonical Notch1 target genes were defined by a 2-fold or greater increase in expression within 4 hrs of GSI washout that was insensitive to cycloheximide and sensitive to dominant negative MAML1. Two hundred forty-five genes met these criteria, including many previously identified direct target genes such as HES1, HES4, HES5, DTX1, and MYC (45, 56). Of these 245 genes, 61 (25%) had Notch1 binding peaks in their promoters (Fig. 3.4), which is enriched ($p < 10^{-4}$, binomial test) over the total number of genes on the Affymetrix chip with Notch1 binding in their promoters (2325 of 15,340 genes, 15.1%). The remaining direct target genes are presumably regulated through enhancers, a possibility consistent with the presence of at least one Notch1 binding site within 100 kb of the promoters of 127 of 179 target genes (69%) lacking Notch1 promoter binding.

For each target gene, the Notch binding site closest to the transcriptional start site was designated the most likely regulatory element (summarized in Table S2). RBPJ binding was detected by ChIP-Seq at 52% of these candidate regulatory sites, an enrichment over the 36% overlap between RBPJ and Notch binding genome-wide ($p < 10^{-6}$, binomial test). The 179 target genes without promoter binding showed a larger expression change in response to Notch1 activation than the 66 target genes with promoter binding ($p < 0.05$), indicating that Notch1-responsive enhancers are relatively potent inducers of target gene expression.

RBPJ/Notch1 binding sites near some target genes merit comment. HES1, a gene involved in T-LL and normal T cell development, has two high-confidence (HC) and one low-confidence binding peaks. One HC peak is centered on a sequence-paired site 44 bp 5 of the TSS that loads Notch1/RBPJ dimers, while the second HC peak is found in the proximal promoter 1239 bp 5 of the TSS; this second site only loads RBPJ/Notch1 monomers (S. Blacklow, unpublished data). Mutations that disable Notch1 dimerization prevent activation of a reporter gene containing the sequence-paired site, yet the same mutated versions of Notch1 are fully competent to drive HES1 expression in T-LL cells (data not shown). The 5 RBPJ/Notch1 binding site near HES1 may compensate for defective Notch1 dimerization on the proximal sequence-paired site, providing an explanation for this discrepancy. HES1 is also unusual in having high levels of bivalent H3K4me3 and H3K27me3 chromatin marks, a combination that in embryonic stem cells defines genes that are poised to respond to differentiation cues;

the significance of this chromatin structure near HES1 in T-LL cells remains to be determined.

ChIP-Seq alignments identified candidate enhancers (defined by high levels of H3K4me1 marks) near other direct Notch1 target genes of interest, including: i) the clustered GIMAP genes, which have been implicated in T-LL cell survival; DTX1, a robust Notch1 target gene; and IL7R, IGF1R, and MYC, all of which contribute to Notch1-dependent T-LL cell growth. Local ChIP confirmed Notch1/RBPJ binding to the candidate MYC enhancer and also showed binding of the co-activators p300 and CREB binding protein (CBP) as well as RNA polymerase II. Although a promoter binding site for RBPJ/Notch1 has been reported in IL7R in CUTLL1 cells, we observed no binding to the IL7R promoter in CUTLL1 cells by ChIP-Seq, whereas local ChIP confirmed binding of Notch1 to a putative 3 IL7R enhancer in three Notch1-dependent cell lines, CUTLL1, HPB-ALL and KOP-TK1. Similarly, Notch1 binding to a candidate IGF1R intronic enhancer was also observed in these lines. These examples highlight the utility of ChIP-Seq in identifying regulatory elements near target genes of interest.

Conservation of factors associated with Notch1 binding sites and divergence of factors associated with RBPJ only binding elements. To determine conserved aspects of RBPJ/Notch1 interactions with T-LL genomes, we did Chip-Seq analyses on two Notch1-dependent murine T-ALL cell lines; T6E, which expresses a membrane-tethered form of human Notch1, and G4A2, which has a 5 deletion in an endogenous notch1 allele leading to constitutive Notch1 signaling. We identified 1587 high confidence Notch1 sites and 2776 high confidence RBPJ sites in the T6E cell genome (FDR¡0.01). The distribution of Notch1 binding sites was similar to that in CUTLL1 cells, with a bias for promoters. Furthermore, the transcription factor motifs associated with Notch1 binding sites in T6E cells were the same as in CUTLL1 cells and in the same rank order: ZNF143, Ets, and Runx (not shown). The fraction of Notch sites overlapping with each co-factor was almost identical in mouse and human (Fig. 3.5B), and K means clustering identified the same five classes of Notch1 binding clusters (Fig. 3.5C). For unclear reasons, fewer high confidence Notch1 and RBPJ binding sites (548 and 668, respectively) were identified in G4A2 cells, but the same associations were found (not shown).

In contrast to the conservation of motifs and co-factor combinations, Notch1 binding to orthologous elements was quite divergent; for example, of the 245 direct Notch1 target genes identified in human CUTLL1 cells, only 50 genes (20.4%) showed Notch1 binding to orthologous elements within or flanking the same genes in murine T6E cells. Well-characterized target genes with conserved Notch1 binding sites include HES1 and HES5, while divergent Notch1 binding sites were observed in other conserved target genes, such as DTX1. Conserved RBPJ binding to orthologous elements was highly associated with conserved Notch1 binding (Fishers test, $p < 10^{-16}$), while conserved imputed ZNF143, Ets and Runx sites showed smaller, but still significant, associations (Table 3.1A).

Unexpectedly, the distribution of RBPJ binding sites was also highly divergent between human and murine T-LL cells. Only 36% of RBPJ sites were in promoters and only 35% of RBPJ sites overlapped with Notch1 sites in murine T6E cells (Fig. 3.5A); similar distributions were observed in murine G4A2 cells. The motif most highly enriched near RBPJ binding sites in T6E cells (Fig. 3.5D) and G4A2 cells (not shown) was that of REST, a DNA-binding transcriptional repressor originally identified as a regulator of neurogenesis, followed by the motifs for CREB and Ets. Overall, 895 of 2776 (32%) RBPJ binding sites in T6E cells were associated with imputed REST sites. Most RBPJ/REST sites were outside of promoters and none of these sites bound Notch1; thus, these sites are responsible for the divergence in RBPJ binding distributions between human and murine T-LL genomes. Analyses using PBM data showed that RBPJ/REST sites generally lack high-affinity RBPJ binding motifs, suggesting that RBPJ binding to these sites involves protein-protein interactions. By contrast, only 6 of 2112 (0.3%) of the RBPJ sites in human CUTLL1 cells were associated with a REST motif. Furthermore, ChIP-Seq studies have mapped 2319 REST and 10529 RBPJ binding sites (FDR < 0.01) in EBV-transformed lymphoblastoid cell lines, only 12 of which (0.1%) overlapped. Thus, RBPJ/REST co-sites appear to be restricted to murine cells. A second interspecies difference in RBPJ only sites was the presence of a higher fraction of RBPJ/CREB sites in human T-LL cells (Fig. 3.5E, Fishers test $p < 10^{-16}$). Conservation of CREB motifs was strongly associated with conserved RBPJ binding to orthologous elements (Table 1B, Fishers test $p < 10^{-16}$). Thus, unlike RBPJ/REST sites, there appears to be selective pressure favoring conservation of RBPJ only sites associated with CREB motifs.

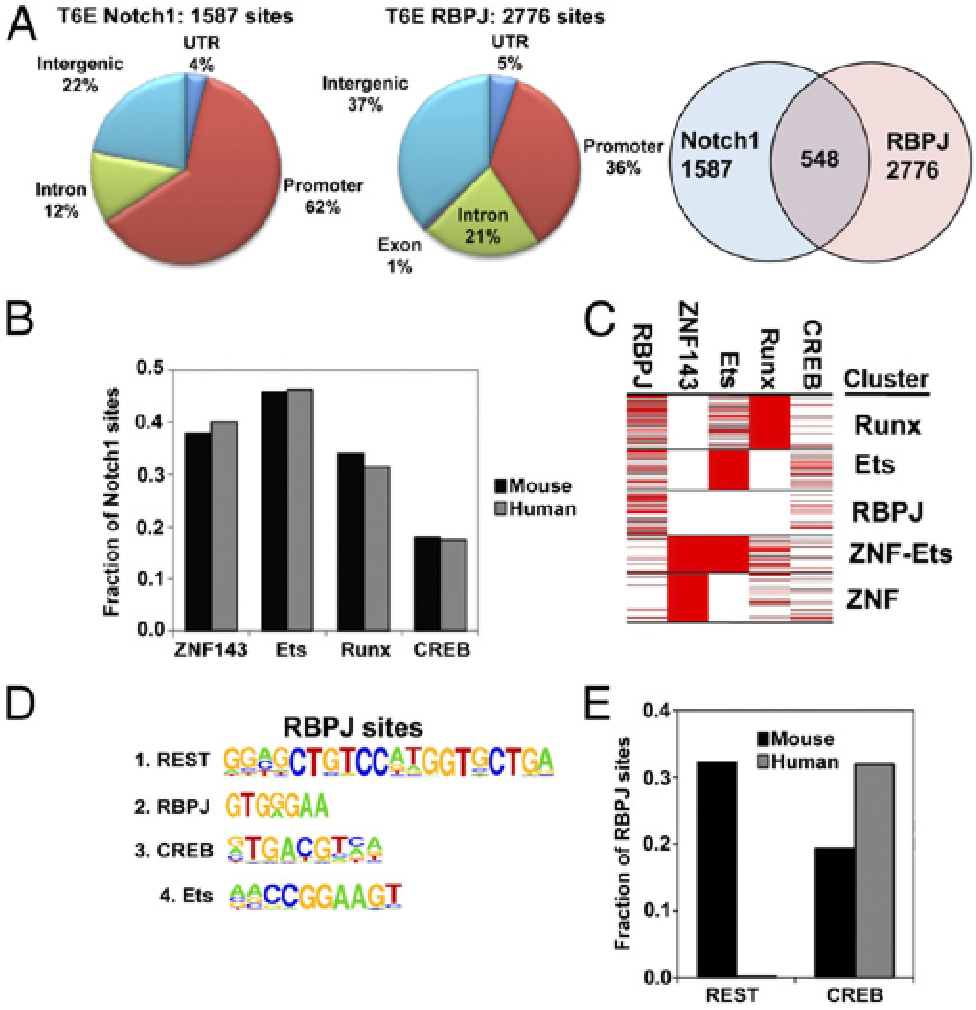


Figure 3.4: Conserved and divergent features of Notch1 and RBPJ binding sites in murine and human TLL cells. (A) Distribution and overlap of RBPJ and Notch1 binding sites in the genome of the murine TLL cell line T6E. Promoter regions are defined as sequences within 2 kb of the TSS of annotated genes. (B) Conservation of the proportions of Notch1 binding sites associated with ZNF143, ETS, RUNX, and CREB motifs in the genomes of human CUTLL1 cells and murine T6E cells. The fraction of Notch1 binding sites that contain the indicate motif 250 bp of Notch1 peaks is shown. (C) Conservation of RUNX, ETS, RBPJ, ZNF143-Ets, and ZNF143 clusters in murine T6E cells. (D) Enriched transcription factor motifs lying 250 bp of RBPJ peaks in murine T6E cells. (E) Divergence of REST and conservation of CREB motifs near RBPJ peaks in murine T6E cells and human CUTLL1 cells. The fraction of RBPJ binding sites that contain the indicated motif 250 bp of RBPJ peaks is shown.

3.4 Discussion

Our detailed analysis of RBPJ/Notch1 interactions with the genomes of T-LL cells has identified conserved features with implications for transcriptional regulation by RBPJ and Notch, as well as unexpected points of evolutionary divergence. While Notch1 binds predominantly to promoters, we find that promoter binding is a poor predictor of Notch1 responsiveness; all told, 97% of genes with Notch1 promoter binding do not change expression appreciably when Notch1 signaling is chemically inhibited. This is in line with a study of 263 transcription factors in yeast, which found only 3% overlap between promoter occupancy and measurable response to transcription factor perturbations. Similar findings have been noted in Drosophila. In non-responsive genes with promoters bound by Notch1, loss of Notch1 may be compensated for by other transcription factors, or Notch1 might only affect gene expression when other signaling pathways are also altered. The only significant positive predictor of Notch1 responsiveness in T-LL cells was RBPJ co-binding, but only a small subset of genes with RBPJ/Notch1 promoter binding respond to Notch1 inhibition, and other determinants of responsiveness must exist. The ChIP-Seq studies described here also revealed a large number of Notch1 binding sites outside of promoters that have chromatin marks consistent with enhancers, which presumably are involved in regulation of the 75% of Notch1 target genes that lack Notch1 promoter binding. Additional studies employing methods such as chromosome conformation capture will be necessary to definitively link these elements to regulation of Notch1 target genes.

Another variable likely to influence the responsiveness of genes to Notch1 are co-factors that bind to nearby sites. We identified a significant enrichment for overlapping ZNF143 binding sites and Ets and Runx factor motifs near sites of Notch1 binding. These co-factors occur in non-random combinations with Notch1 binding sites in both human and murine T-LL cells, pointing to a conserved cis-regulatory transcription factor network that forms the basis for a new understanding of Notch1 function in this particular cellular context. Ets factors are known to be involved in T cell specification and development. Similarly, Runx factors interact with Notch1 to promote hematopoietic stem cell development, act upstream of Notch1 in early T cell development, and when deficient predispose mice to T-LL. An important question in understanding Notch1 roles in T cell progenitors is the identity of factors that condition the epigenome, allowing NICD1 access to sites that mediate the expression of genes that are required for T cell development or, when dysregulated, contribute to T-LL. Ets and Runx family members, by virtue of their frequent association with Notch1 binding sites in T-LL genomes, are candidates for such roles.

Unlike Ets and Runx factors, ZNF143 has no known role in T cell development, T-LL, or Notch signaling. ZNF143 is a ubiquitously expressed zinc finger protein of unknown function originally identified as a regulator of selenomethionine tRNA gene expression. Our ChIP-Seq analyses confirm bioinformatic and local ChIP studies suggesting that ZNF143 localizes to promoters, particularly those driving transcription bi-directionally. Gene pairs regulated by bidirectional promoters are highly conserved in vertebrates; whether this is a consequence of convergent functions necessitating coordinate regulation or chance insertion of a coding sequence adjacent to a promoter with bidirectional activity during early vertebrate evolution (promoter capture) is uncertain. A motif resembling that of ZNF143 is also enriched at sites of SREBP-1 promoter binding in hepatocytes, suggesting that ZNF143 is a general transcription factor with functions in many contexts. A central question emerging from our work is whether sites where RBPJ/Notch1/ZNF143 co-localize load RBPJ/Notch1/ZNF143 ternary complexes, or instead are occupied sequentially by distinct RBPJ/Notch1 and ZNF143 complexes that undergo exchange on co-sites. We have recently observed that binding of purified RBPJ and ZNF143 to DNA oligonucleotides containing ZNF143/RBPJ co-binding sequences, such as that from the promoter of the Skp2 gene, is mutually exclusive (K. Arnett and S. Blacklow, unpublished data), a finding that supports the exchange model, but further work is needed to test these two models in cells and to determine the functional importance of this unusual and pervasive class of RBPJ/Notch1 genomic binding site in T-ALL cells.

Finally, our studies have identified major classes of genomic RBPJ sites that fail to bind Notch1 or do so very inefficiently. Both murine and human T-LL genomes contain numerous RBPJ only binding sites that are enriched for CREB motifs. These sites have relatively high levels of repressive chromatin marks, tend to flank genes with low expression levels, and often lack high-affinity RBPJ binding motifs. A second RBPJ only site is characterized by motifs for the DNA-binding factor REST, best known for its role in regulating neurogenesis (5) but which also has other functions, such as regulating herpes simplex virus latency. RBPJ/REST sites are common in murine T-LL cells, but essentially undetectable in human T-LL cells or EBV transformed human B cells, presumably due to divergence in factors that mediate recruitment of RBPJ to REST sites. One important Notch-independent RBPJ-dependent function in mammals involves the association of RBPJ with the transcription factor Ptf1a. The identification of CREB/RBPJ only and REST/RBPJ only sites in T-LL cells suggests that other Notch-independent RBPJ functions exist and that the RBPJ transcriptional switch model of gene regulation is overly simplistic, at least in higher vertebrates.

4

Epstein-Barr Virus Exploits Intrinsic B-Lymphocyte Transcription Programs to Achieve Immortal Cell Growth

4.1 Overview

Epstein-Barr Virus Nuclear Antigen 2 (EBNA2) regulation through the cell RBPJ transcription factor (TF) is essential for resting B-lymphocyte (RBL) conversion into Lymphoblast Cell Lines (LCLs). ChIP-seq of EBNA2 and RBPJ sites in LCL DNA found EBNA2 at 5151 and RBPJ at 10,529 sites. EBNA2 localized 72% with RBPJ, at intergenic and intronic sites and only 14% at promoter sites. EBNA2/RBPJ sites were enriched for Early B-cell Factor (EBF, 60%), RUNX(41%), ETS(35%), NFkB(31%), and PU.1(17%) motifs. Using ENCyclopedia Of DNA Elements (ENCODE) LCL data, EBF, RELA, and PU.1 were at 54%, 31%, and 17% of EBNA2 sites. K-Means clustering of EBNA2 site associated RBPJ, EBF, RELA, PU.1, and ETS or RUNX motifs identified RELA-ETS, EBF-RUNX, EBF, ETS, RBPJ, and repressive RUNX clusters, which then ranked highest to lowest in H3K4me1 signal distribution and nucleosome depletion around the EBNA2 site, marks of active chromatin. Surprisingly, although quantitatively less, the same sites in RBLs exhibited high level H3K4me1 signals, with similar nucleosome depletions. The EBV genome also has an EBF site in the LMP1 promoter, which was critical for EBNA2 activation. LCL HiC data mapped intergenic EBNA2 sites to EBNA2 up-regulated genes. Fluorescence In Situ Hybridization (FISH), Chromatin Conformation Capture, and 3C q-PCR, linked EBNA2/RBPJ enhancers 428 kb 5' of MYC to MYC. These data indicate that EBNA2 evolved to target RBL H3K4me1 modified, nucleosome depleted, non-promoter sites to drive B-lymphocyte proliferation in primary human infection. These RBL sites likely support antigen-induced proliferation.

4.2 Introduction

Although Epstein Barr Virus (EBV) infection is highly prevalent in all human populations, EBV is also an important cause of endemic Burkitts Lymphoma (48), Hodgkins Lymphomas (24), and Lymphoproliferative Diseases in immune suppressed (49) and HIV infected people (21). In early primary human infection, EBV infects peripheral resting B lymphocytes (RBLs) and expresses 6 nuclear proteins (EBNAs) and two integral membrane proteins (LMPs). EBNAs and LMPs, principally EBNA2 and LMP1, cause RBL proliferation, which can be malignant in the absence of effective T-cell responses. Proliferating RBLs enter tonsils and other lymphoid organs. Since EBNAs and LMPs are encoded by over 4,000 amino acids and have many T-cell epitopes, normal T-cell immune responses eliminate these cells. Infected cells, with less complex or absent EBV protein expression, persist and are reservoirs for reactivated infection (21). EBV conversion of RBLs into continuously proliferating Lymphoblasts (LCLs) (64) is a relevant and experimentally useful model for EBV proliferative effects in B-lymphocytes.

In converting RBLs to LCLs, EBNA2 and EBNALP are expressed first. EBNA2 up-regulates EBV and cell gene expression, including EBV LMP1 and cell MYC, CD23, and CD21 (25, 27, 35, 53). EBNA2 associates with DNA through RBPJ, a sequence specific cell transcription factor that also mediates NOTCH DNA binding (68). The B-lymphocyte and macrophage lineage ETS protein, PU.1, is also important in EBNA2 activation of the EBV LMP1 promoter. The EBNA2 C-terminal acidic domain recruits basal and activating transcription factors, including p300/CBP and PCAF Histone Acetyl Transferases. EBNA2 transactivation of MYC causes continuous cell proliferation. EBNA2 transactivation of EBV LMP1 induces cell NFkB and MAP kinase activations, which up-regulate pro-survival BCL2 family gene expression and enable LCL survival. Overall, EBV conversion of RBLs to LCLs mimics antigen induced RBL clonal expansion in germinal centers, where antigen binding to surface Ig induces MYC mediated proliferation (19) and T-cell expressed CD40 ligand activates B-lymphocyte CD40 receptors to up-regulate NFkB, MAP kinases, and anti-apoptotic BCL2 family protein expression.

A challenge in understanding the mechanisms through which EBNA2 and RBPJ regulate gene expression is the finding that an RBPJ motif (G)TGGGAA(A) near a promoter is a poor predictor of EBNA2 dependence. EBNA2 and RBPJ Chromatin Immune Precipitation, Deep-Sequencing (ChIP-seq) and bioinformatic analyses were therefore undertaken to identify and potentially characterize EBNA2 and RBPJ binding sites, genome wide, in LCLs.

4.3 Results

4.3.1 EBNA2 and RBPJ Binding Sites in LCLs.

Independent, duplicate large-scale, EBNA2 and RBPJ ChIPs were deep-sequenced and 107 EBNA2 and 107 RBPJ DNA reads were mapped to the human genome (hg.18) with 36-54% efficiency. After normalization for input DNA, QUEST identified 5151 EBNA2 and 10529 RBPJ sites based on bi-directional sequence reads and FDR< 0.01. EBNA2 and RBPJ, sites were similarly distributed relative to annotated genes. Only 14% of EBNA2 and 13% of RBPJ sites were at promoters, defined as 2kb of a transcription start site (TSS). Instead, EBNA2 and RBPJ sites were 86% and 87%, respectively, at more distal sites; 43% and 42% were intergenic sites, 8% and 9%, UnTranslated Regions (UTR), 34% and 35% introns, and 1% exons. Overall, 72% (3710) of EBNA2 sites were within 100b of a significant RBPJ site (hereafter, EBNA2/RBPJ site). The 1441 EBNA2 sites lacking significant RBPJ signal are hereafter EBNA2 only sites and the 6819 RBPJ sites lacking EBNA2 are RBPJ sites. The finding of some EBNA2 sites without significant RBPJ signal may be due to RBPJ being occluded from antibody recognition, less stable RBPJ DNA association, or EBNA2 tethering to DNA by another transcription factor.

RBPJ at EBNA2/RBPJ sites is more highly DNA associated. On average, at non-promoter EBNA2/RBPJ sites, RBPJ had 34 DNA Reads Per Kilobase per Million aligned reads (hereafter, signals) versus 10 signals per RBPJ site and the difference was highly significant ($p < 10^{-10}$). Similarly, albeit to a lesser extent, at promoters, EBNA2/RBPJ sites had 24 signals versus 11 signals for RBPJ sites ($p < 10^{-10}$). Overall, EBNA2/RBPJ had more RBPJ signals than RBPJ sites genome wide, consistent with the previous finding that EBNA2 significantly increased RBPJ association.

EBNA2 sites are enriched for RBPJ, ETS, EBF, RUNX, PU.1, and NFkB motifs. Analysis of a 500bp window around EBNA2 sites for Transfac database motif enrichment, identified significant ($p < 10^{-100}$) enrichment not only for RBPJ (78%), but also for ETS (39%), EBF (39%), RUNX (43%), PU.1 (22%), and NFkB (22%) motifs, over a control set of sequences with similar GC content. These data are consistent with EBF, RELA, and Pu1, having a significant affects on EBNA2 or RBPJ binding and transcription regulation. (Fig. 4.1A).

To evaluate the extent to which EBNA2 associated TF site predictions correlate with TF occupancy, we used the RBPJ and ENCODE ChIP-seq GM12878 LCL EBF, NFkB subunit RELA, and PU.1 data to evaluate RBPJ, EBF, PU.1, and RELA occupancy at EBNA2 sites. EBF mapped to 27,552, RelA to 24,796, and PU.1 to 24,343 LCL sites. EBNA2 sites were 54% within 100b of an EBF site, 31% of a RELA site, and 17% of a PU.1 site (Fig.4.1B). Overall, the EBNA2 associated TF motif analysis correlated surprisingly well with TF presence by ChIP-Seq. Of EBNA2 sites, RBPJ was predicted at 78% and detected at 72%. EBF was predicted at 39% and detected at 54%. PU.1 was predicted 22% and detected 17%. NFkB was predicted at 22% and was detected at 31%. Thus, EBNA2 and RBPJ had accessed sites that were highly accessible to each of the associated TFs.

EBNA2/RBPJ, EBNA2 only, and RBPJ were associated with different TF complexes. EBF was strongly enriched at EBNA2/RBPJ sites (60%) compared to EBNA2 only sites (36%) or RBPJ sites (37%) ($p < 10^{-10}$). In contrast, ETS motifs were enriched at EBNA2 only sites (45%) compared to EBNA2/RBPJ or RBPJ sites (35%) ($p < 10^{-10}$). P300 sites strongly correlated with EBNA2/RBPJ and EBNA2 only sites, but were significantly less at RBPJ sites ($p < 10^{-10}$) (Table S2). The 262 EBNA2 only sites with smallest RBPJ signals were even more enriched for ETS motifs (57%), consistent with the possibility that ETS or an ETS associated factor may tether EBNA2 to these sites.

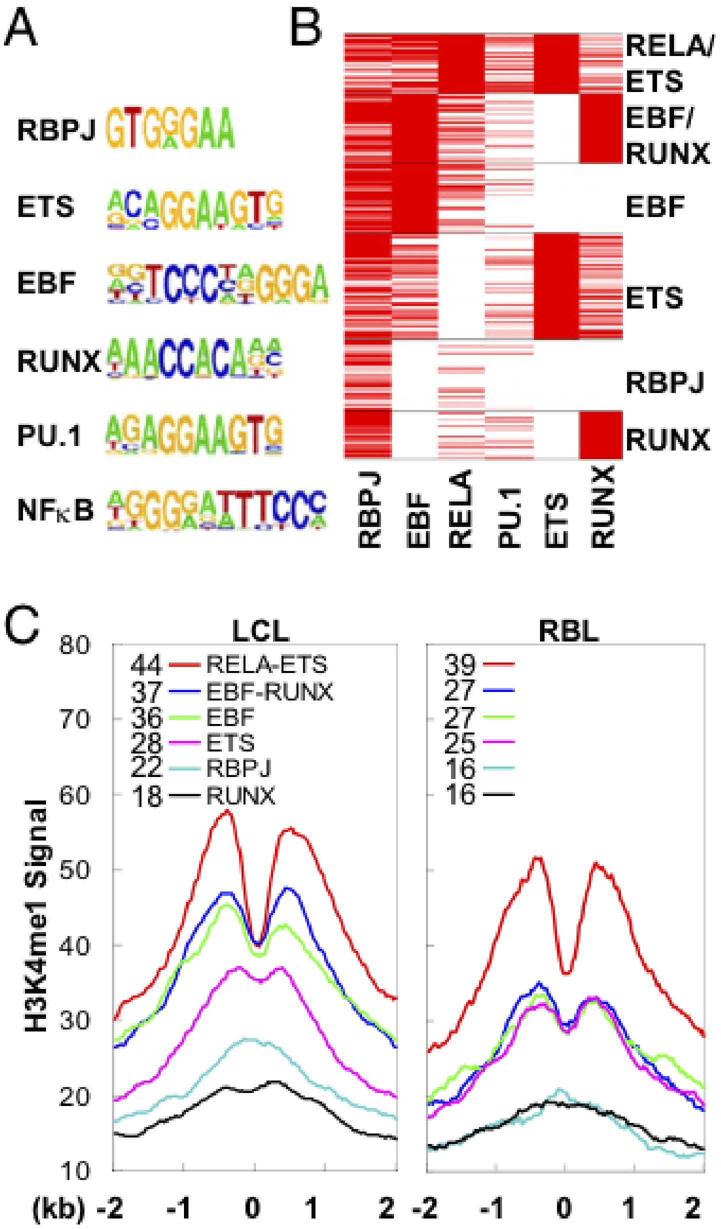


Figure 4.1: EBNA2 binding site enriched TF motifs, K-means TF clusters, and associated H3K4me1 signals in LCLs and at the same genome sites in RBLs. (A) EBNA2 binding site (250 bp) enriched TF motifs are ranked by enrich- ment over random sequences controlling for GC content from top to bottom. All had corrected $p < 10^{-50}$. (B) EBNA2 binding sites (5,151) were K-means clustered based on the presence of RBPJ, EBF, PU.1, and NFkB subunit RELA binding by LCL ChIPseq and Homer-imputed ETS and RUNX motifs. EBNA2 site clusters are named by the principal TF components. (C) LCL EBNA2 site clusters have distinct H3K4me1 signal distributions (Left). H3K4me1 distributions at the same sites in RBLs (Right). The numbers in the upper left corners of the panels are the total signals (reads per kilobase per million mapped reads) under each curve.

4.3.2 EBNA2 co-factors correlate with distinct chromatin footprints.

K-means clustering of the 5151 EBNA2 sites for associated factors identified six clusters with distinct co-factor combinations of RBPJ, EBF, RELA, PU.1 and imputed ETS and RUNX (Fig. 4.1B). EBNA2 site clusters were named by their principal co-factors. RBPJ was a major component of all clusters. One cluster was comprised of EBNA2 sites in which RBPJ was the only dominant component and is referred to as RBPJ. RELA-ETS, EBF-RUNX, EBF, ETS, RBPJ, and RUNX clusters correlated with most (top) to least (bottom) ENCODE LCL H3K4me1 signals within +/- 2 Kb of EBNA2 sites (Fig. 4.1C Left Panel). Differences in H3K4me1 signals between RELA-ETS and EBF-RUNX, EBF and ETS, ETS and RBPJ, and RBPJ and RUNX were each highly statistically significant ($p < 10^{-6}$ Mann-Whitney-U test). Importantly, H3K4me1 enriched clusters, characteristic of RELA-ETS, EBF-RUNX, EBF, and ETS, had fewer H3K4me1 signals at the EBNA2 site, indicative of EBNA2 localization at nucleosome depleted open chromatin sites (Fig. 4.1C, left panel).

Sites of significant p300 signals, indicative of active transcription, also correlated with H3K4me1 signals, as noted. EBNA2 clusters with RelA-ETS, EBF-RUNX, EBF, ETS, RBPJ, and RUNX had 41%, 37%, 30%, 15%, 6%, and 5% of sites with significant p300 signals. These correlative data further support the hypothesis that RELA-ETS, EBF-RUNX, EBF, ETS, RBPJ, and RUNX clusters are ranked from highest to lowest in transcription activation.

The higher H3K4me1 and p300 association of EBNA2 with EBF-RUNX versus EBF alone (Fig. 4.1C, Left Panel, upper blue and green curves) and the very low H3K4me1 effect of RUNX alone (Fig. 4.1C, Left Panel, lower black curve) correlate an activating-RUNX effect with EBF and a repressive-RUNX effect unassociated with EBF, at other sites. Opposing RUNX isoforms have been described in LCLs, where RUNX3 repression of RUNX1 is important for continued growth.

As a complement to clustering, which detected an EBF associated activating-RUNX effect and a repressive-RUNX effect, an additive linear model was used to correlate H3K4me1 signals levels at EBNA2 sites with the presence of EBF, RELA, PU.1, or imputed ETS or RUNX as single factors. EBF increased H3K4me1 signals 1.68 fold, RELA 1.22 fold, PU.1 1.21 fold, and imputed ETS 1.08 fold. Each effect was highly significant ($p < 0.0001$). RUNX as a single factor had no significant overall effect, consistent with the positive RUNX effect with EBF being similar in magnitude to the negative RUNX effect without EBF.

Ontogeny of H3K4me1 modifications. To understand how EBV infection alters RBL transcription, H3K4me1 modifications at EBNA2 sites in LCLs were compared to H3K4me1 modifications at the same sites in RBLs. Roadmap Epigenomics Mapping Center ChIP-seq H3K4me1 signals in CD19(+) RBLs (Fig. 4.1C, Right panel) were compared with H3K4me1 signals at EBNA2 sites in LCLs (Fig. 4.1C, Left Panel). Although H3K4me1 signals were less elevated in RBLs, they mirrored the symmetrical elevation with central depletion seen in LCLs and followed the LCL cofactor cluster site hierarchy. For example, the EBNA2, RELA, and imputed ETS cluster in LCLs had 44 H3K4me1 signals, whereas the same sites in RBLs had 39 H3K4me1 signals (Fig. 4.1C Left versus Right Panels). Furthermore, the EBNA2 EBF-RUNX, EBF, and ETS cluster sites in LCLs had 37, 36, and 28 H3K4me1 signals, whereas the same sites in RBLs had 27, 27, and 25 H3K4me1 signals. Also, the EBNA2 RBPJ and RUNX cluster sites in LCLs had 22 and 18 H3K4me1 signals, whereas the same sites in RBLs had 16 and 16 H3K4me1 signals. Importantly, nucleosome depletion was also evident at the same sites in RBLs, indicating that EBNA2 targets sites of pre-existing nucleosome depeletion. These data strongly support a model in which EBNA2 target H3K4me1 modified and nucleosome depleted RBL RELA-ETS, EBF-RUNX, EBF, and ETS sites for transcription up-regulation.

Differences between EBNA2 Promoter and putative-enhancer effects. In contrast to LCL putative-enhancer sites where RELA-ETS, EBF-RUNX, and EBF were associated with higher H3K4me1 signals ($p < 0.01$) versus ETS, RBPJ, and RUNX, and as a single factor EBF had the largest up-regulatory effect (Fig. 4.1C, left panel), at promoter sites, EBNA2 and RELA-ETS, ETS, or RBPJ were associated with higher H3K4me3 signals and RUNX, EBF-RUNX, and RBPJ were associated with higher H3K27me3 signals. These data are consistent with EBF having the largest up-regulatory effects at putative enhancer sites, and with ETS having the largest up-regulatory effects at promoter sites. A repressive RUNX effect was also evident at promoter sites.

EBF is critical for EBNA2 activation of the EBV LMP1 promoter. EBF is essential in B-lymphocyte development and regulates gene expression from pro-B cell through mature B-cell stages and therefore a likely pioneering factor in EBNA2 up-regulated B-lymphocyte gene expression, given the 54% co-localization of EBNA2 at EBF sites

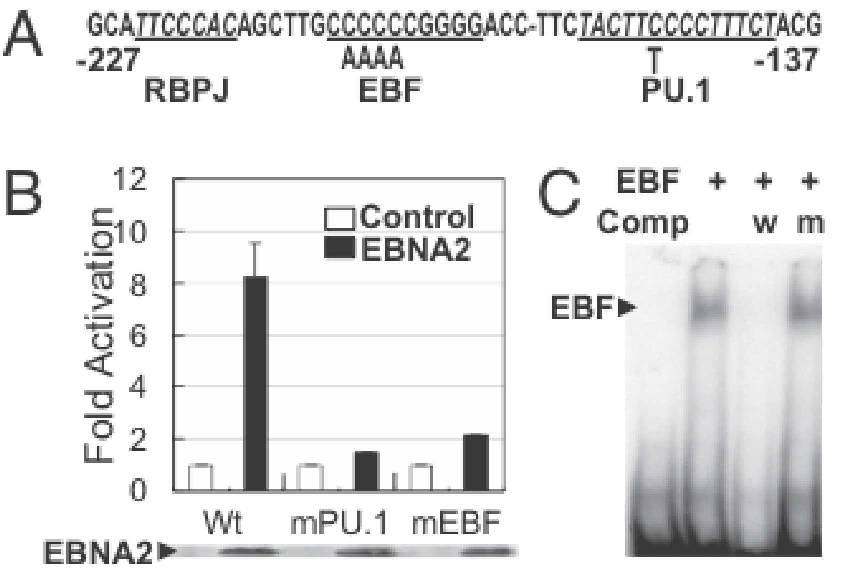


Figure 4.2: EBF is important for EBNA2 activation of the EBV LMP1 promoter. (A) Schematic of the LMP1 promoter EBNA2 response element (?227/?137), with RBPJ, EBF, and PU.1 binding sites underlined in the continuous se- quence. Null EBF or PU.1 point mutations are indicated below the continu- ous sequence. (B) EBNA2 activated the LMP1 promoter-luciferase reporter eightfold, the EBF mutant site twofold, and the PU.1 mutant site onefold in BJAB cells. Luciferase activities were normalized to cotransfected EBNA2-independent ?-galactosidase activity. EBNA2 was expressed at similar levels in all experiments as shown in the EBNA2 immune blot at the bottom of B. (C) In vitro translated EBF bound to a 32P labeled oligonucleotide that in- cluded the EBF binding site. Binding was completed by a 500-fold excess cold wild-type oligonucleotide (w), but not by 500-fold cold mutant nucleotide (m). Protein-DNA complexes were separated by PAGE and visualized by phosphoimager.

in LCLs, the embedding of RBPJ TGGGA(A) core motifs in EBF motifs, and the similarity of H3K4me1 at EBNA2 sites in LCLs with the same sites in RBLs (Fig. 4.1C).

The putative EBF role in EBNA2 mediated transcription was evaluated in the context of the EBV LMP1 promoter, which is EBNA2, RBPJ, and PU.1 dependent. Since the LMP1 promoter RBPJ, PU.1, and EBF sites are non-overlapping (Fig. 4.2A), independent null EBF, RBPJ, or PU.1 mutations could be readily evaluated. In BJAB Lymphoma cells, EBNA2 activated the (-335 to +40) LMP1 promoter 8 fold over control expression vector. A single point mutation in the PU.1 site reduced EBNA2 activation to background (Fig. 4.2B). Mutation of the EBF motif CCCCCCGGGG to CAAAACGGGG abolished EBF binding and reduced EBNA2 activation from 8 to 2 fold (Fig. 4.2B,C), similar to the RBPJ binding site mutation effect. These data indicate that EBF is important for LMP1 promoter activation.

The hypothesis that EBF is a ”pioneer” B-lymphocyte transcription factor that opens chromatin for EBNA2 and RBPJ binding, was further evaluated by scanning EBNA2 ChIP-seq RBPJ binding sites through a 500bp window to identify the 8 bp DNA sequence that has the strongest, in vitro, affinity for RBPJ, based on a RBPJ Protein Binding Matrix. At 436 EBNA2 bound sites, the strongest binding sites for RBPJ were in an EBF motif.

EBNA2 up-regulated genes are enriched for EBNA2 binding ¿2kbp from the TSS. EBNA2 sites were 43% intergenic, 35% intron, and only 14% at promoter sites. Intergenic EBNA2 sites correlated with LCL RNA expression levels in 4 gene sets (Fig. 4.3A). A conditional EBNA2 regulated 81 gene set had 2.4 EBNA2 sites per gene within 2-152 kb of the TSS. The 2.4 sites were distributed, with 0.7 sites at 2-32kb, 0.5 at 32-62kb, 0.5 at 62-92kb, 0.3 at 92-122kb, and 0.4 at 122-152kb from the TSS (Fig. 4.3A). A second gene set was comprised of RNAs that are 5 fold up-regulated during RBL conversion to LCLs. This set had 1.2 EBNA2 putative enhancer sites per gene 0.3 at 2-32kb, 0.2 at 32-62kb, 0.2 at 62-92kb, 0.3 at 92-122, and 0.2 at 122-152kb. Genes 2-5 fold up-regulated during RBL conversion to LCLs had 1.1 EBNA2 putative-enhancer sites, evenly split within 2-152 kb from the TSS. Even genes up-regulated less than 2 fold had 0.8 EBNA2 putative-enhancer sites, evenly split 2-152 kb from the TSS. These data indicate that EBNA2 binding to distal enhancer sites significantly correlates with up-regulated LCL gene expression ($p < 0.01$).

Multiple EBNA2 sites have long range interactions with promoters of the 81 EBNA2 regulated gene set. Chromatin Conformation Capture followed by deep-sequencing (HiC) from GM06990 LCLs revealed paired-end sequences representative of both long range genome interaction components (36). Many of these LCL HiC sequences mapped with one end within +/-2kb of a non-promoter EBNA2 site and the other end within -1 to +5 kb of the TSS of 1 of 56 EBNA2 regulated genes. Excluding EBNA2 sites < 5kb from the TSS, 3.3 different long range interactions were detected per EBNA2 regulated gene. The EBNA2 regulated gene set was significantly enriched (corrected $p < 10^{-4}$) over a control gene set, with similar expression level and distance from EBNA2 sites to the TSS, but not conditionally EBNA2 affected (Fig. 4.3B). Overall, 61% of the

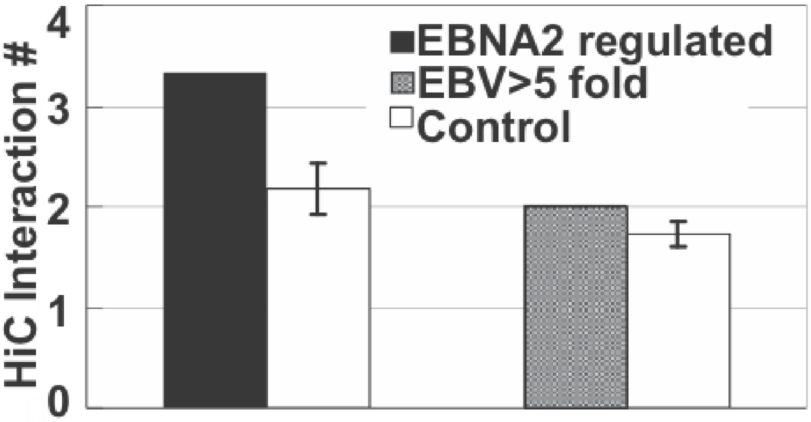


Figure 4.3: Nonpromoter EBNA2 binding sites are enriched for long-range interactions with EBNA2 regulated or LCL > 5 upregulated. HiC detected 3.2 long-range interactions between intergenic EBNA2 binding sites and EBNA2 conditionally regulated genes versus control genes (corrected $p < 0.0001$) and two long-range interactions between EBV > 5 up-regulated genes (corrected $p < 0.05$). Control genes had similar RNA levels and distance to EBNA2 binding sites as the 81 conditionally regulated or ¿fivefold up-regulated genes.

interactions were within the same chromosome and 39% were inter-chromosomal. For intra-chromosome interactions, the median distance from EBNA2 enhancer to an affected gene was 330kb, with 90% greater than 150kb. EBNA2 enhancer sites overlapped 66% with an EBF site and 30% had significant p300 signals, both significantly enriched over the 54% of EBNA2 sites that have an EBF site (corrected $p < 0.01$) and the 22% of EBNA2 sites with p300 signals (corrected $p < .05$). A greater than 5 fold up-regulated LCL RNA data set was also enriched for interactions, with 2 different HiC interactions per gene (corrected $p < 0.05$) (Fig. 4.3B). Interactions of EBNA2 enhancer sites with the EBNA2 regulated or > 5 fold up-regulated gene sets peaked within 2kb of the EBNA2 binding site. EBNA2 binding site interactions were progressively less frequent within neighboring 4kb windows. These data indicate that EBNA2 enhancers are significantly enriched for distal target interactions with EBNA2 regulated or 5-fold up-regulated genes. As an additional control, the same procedure was applied to compute HiC interactions between RBPJ sites without EBNA2 and EBNA2 regulated genes. The number of detected interactions was significantly fewer than interactions with EBNA2 sites (corrected $p < 0.01$).

4.3.3 EBNA2 up-regulates MYC gene expression through long range enhancer and promoter looping.

EBNA2 and RBPJ LCL ChIP-seq failed to identify significant EBNA2 or RBPJ signals at or near the MYC locus. However, strong PolII and H3K9ac signals were evident at the MYC promoter, consistent with known EBNA2 up-regulated MYC expression. Strong CTCF signals were also evident at +.2 kb, -1.9kb, and -10kb. However, the nearest significant EBNA2 and RBPJ signals were at -168kb to -186kb of the MYC TSS and were associated with high H3K4me1. Moreover, the strongest EBNA2 and RBPJ signals were from 11 significant EBNA2/RBPJ sites at -428 to -556 kb of the MYC TSS (Fig. 4.4A). These sites had high H3k4me1, H3K9ac signals and coincided with PolII and p300 signals, indicative of highly active enhancers (Fig. 4.4A). Some -428 to -556kb EBNA2/RBPJ sites also coincided with EBF and RELA and were near CTCF signals. In LCLs, EBNA2 induced LMP1 substantially up-regulates canonical and non-canonical NFkB. RELA is a major component of the -428 and -556 MYC regulatory complexes and is second to EBF in overall H3K4me1 signal effects and likely coordinately up-regulates MYC transcription with EBNA2. Surprisingly, RBL H3K4me1 signals at the principal -428 and -556 MYC regulatory sites were very similar to LCL H3K4me1, consistent with the supposition that RBLs are poised for MYC transcription activation and are then activated by EBNA2 and RBPJ with EBF and possibly RELA.

FISH assays with green fluorescent Bacmid probe centered around the 428-556 kb enhancer site and an orange fluorescent Bacmid probe centered around MYC were used to investigate interaction between the EBNA2 enhancer and MYC in LCLs versus RBLs. Overall, 94/100 LCLs and 92/100 RBLs had both alleles discernible. Of these, 43 LCLs and 41 RBLs had 1 co-localized and 1 separated allele. In contrast, 49 LCLs and only 1 RBL had 2 co-localized alleles, whereas 2 LCLs and 50 RBLs had 2 separated alleles (P¡.001) (Fig. 4.4B and 4.4C). In summary, LCLs significantly differed from RBLs in bi-allelic EBNA2 enhancer site association with MYC.

Conditional EBNA2 expressing LCLs were used to investigate the effect of EBNA2 inactivation on EBNA2 enhancer site association with MYC. After 4 days in medium permissive (+) versus non-permissive (-) for EBNA2 expression, 96 EBNA2 (+) versus 69 EBNA2 (-) LCLs were evaluated. EBNA2 (+) LCLs significantly differed from

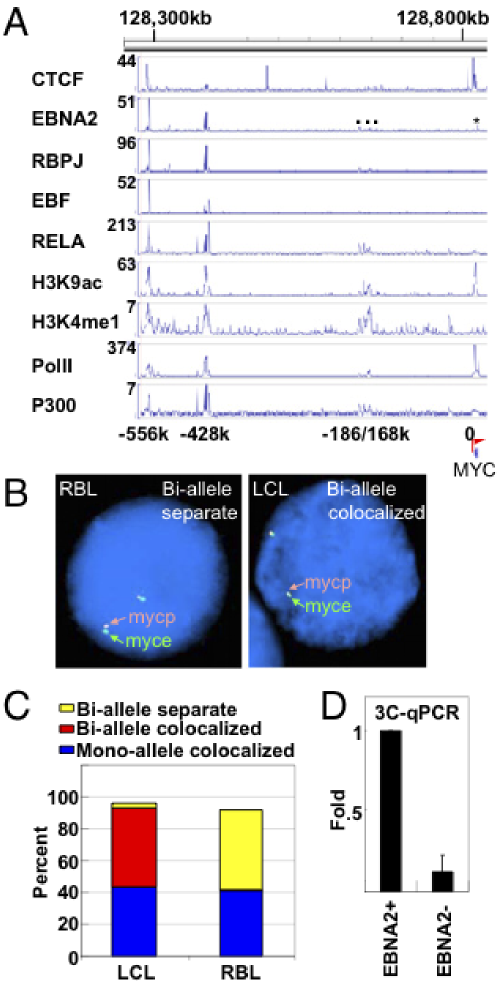


Figure 4.4: EBNA2 activates MYC via long-range enhancer and promoter interactions. (A) CTCF, EBNA2, RBPJ, EBF, RELA, H3K9ac, H3K4me1, PolII, and p300 signals and MYC promoter (arrow) are shown. Normalized signals are at the left end of each track. The asterisk indicates unidirectional EBNA2 reads. The three small dots at 168 kb to 186 kb indicate three significant EBNA2/RBPJ peaks. Major EBNA2/RBPJ peaks are 428 kb to 556 kb. For enlarged view, see Fig. S5B. (B) FISH for IB4 LCL and RBLs. Fixed LCLs or RBLs were hybridized with fluorescent labeled BACmid probes for DNA 428 to 556 kb upstream of MYC, (green) and 150 kb cen- tered around MYC (orange). Cell DNA is stained blue with DAPI. Mycp indicates myc promoter and myce indicates myc enhancer. (C) One hundred LCLs or RBLs were scored for colocalization of EBNA2 up-stream sites and MYC. Blue indicates the percentage of cells with one allele colocalized and one allele separate. (D) Chromatin conformation capture qPCR assay using conditional EBNA2 LCLs grown under permissive or nonpermissive conditions for 4 d. Cells were cross-linked with formaldehyde, DNA digested with Csp6I, diluted, ligated, and quantitated by Taqman qPCR. Fold differences were determined using the Ct method with EBNA2- expressing cells set at 1.

EBNA(-) LCLs in having 62% versus 36% single allele co-localization and in 34% versus 63% bi-allele separation. These data support the hypothesis that the 428-556 kb upstream EBNA2 enhancer site association with MYC is EBNA2 dependent.

Chromosome conformation capture or 3C was used to further investigate the proximity of the -428 kb upstream EBNA2 enhancer site to MYC. After generating the 3C library, PCR was used to detect the -428kb EBNA2 enhancer site ligated to the MYC promoter or first intron. Sequencing of the PCR product revealed the -428kb EBNA2 enhancer linked through a Csp61 site to the MYC promoter and first intron.

LCLs conditional for EBNA2 expression were used to further investigate conditional EBNA2 effects on the -428 kb EBNA2 enhancer proximity to the MYC first intron. 3C libraries were made from EBNA2(+) and EBNA2(-) conditional LCLs. Taqman qPCR and Taqman probe for the putative EBNA2 Enhancer and MYC intron Csp6I DNA fragment ligation junction were used to assess product abundance. EBNA2 (+) LCLs were 7.8 fold enriched for the -428kb EBNA2 enhancer and MYC intron Csp6I ligation product over EBNA2 (-) LCLs (P¡.01) (Fig. 4.4D). These further support an EBNA2 dependent role for the -428 kb enhancer in MYC regulation.

4.4 Discussion

Previous studies of the molecular genetics and pathogenesis of EBV induced B-cell growth support a model that EBNA2 and LMP1 constitutively activate RBL growth and survival pathways that are normally regulated by RBL B-cell receptor antigen activation and T-cell help. RBLs have rearranged surface Immunoglobulin (sIg) and are likely programmed to proliferate in response to cognate antigen and T-cell help in lymph node germinal centers. Antigen signaling through sIg induces MYC-driven B-lymphocyte proliferation. B-cell MHC-class II antigen presentation to germinal center CD4+ T-cells induces CD40-ligand expression. CD40 ligand induces B-cell NFkB activation and anti-apoptotic BCL2 family protein expression, which prevents MYC-proliferation induced cell death. EBNA2/RBPJ activation of MYC and of LMP1 provide similar high level MYC and NFkB activation, enabling LCL proliferation and survival.

The investigations described here elucidate the basic mechanisms through which EBNA2, genome-wide, exploits and enhances the RBL transcription program to cause continuous B-cell proliferation. EBNA2 targeted 5000 RBL open chromatin sites, mostly through RBPJ and EBF. EBNA2, RBPJ, and EBF sites were predominantly distal intergenic or intron enhancers. Only 14% were at promoter sites. EBNA2 associated with six different RBPJ, EBF, ETS, RUNX, NFkB RELA, or PU.1 complexes. Specific complex composition determined much of the EBNA2 effect on nucleosome depletion, H3K4me1 signal, and other transcription associated effects. The key EBF, PU.1, ETS, NFkB RELA, and activating RUNX1 roles in determining EBNA2 and RBPJ B-cell transcription effects is consistent with their pioneering roles in developing and mature B-lymphocyte gene expression. Component analyses detected activating and repressive RUNX effects. RUNX1 is the likely activator since EBF, PU.1, ETS and RUNX1 are frequently associated with active B-cell enhancers.

The most surprising finding is the extent to which EBV and particularly EBNA2 have evolved to exploit the RBL transcription regulatory framework. Indeed, uninfected RBLs had very similar nucleosome depletion and H3K4me1 chromatin modification at sites subsequently targeted by EBNA2. The RBL sites were poised to activate gene expression in response to antigen stimulation and T-cell help. RBL nucleosome depletion was enabled by and efficiently enabled cell TF access. EBNA2 and RBPJ exploited efficient access. EBNA2 further increased nucleosome depletion and H3K4me1 signals, and activated gene expression. EBNA2 actuating effects were clearly evident at the MYC locus, where EBNA2 moderately increased chromatin associated pro-transcription effects and markedly up-regulated transcription.

Not only has EBNA2 evolved to activate transcription through RBPJ and EBF, but the EBV genome also evolved to require EBF for EBNA2 regulation of the LMP1 promoter. EBF developmentally activates PAX5 and is frequently linked to PAX5 in transcription up-regulatory effects. PAX5 was previously shown to be important for initial transcription from the Wp EBNA promoter in B-lymphocytes.

We were surprised to note that RBPJ and EBF share core DNA binding sequences and that EBNA2 associated EBF sites are enriched for high affinity RBPJ sites, consistent with site evolution in response to both EBF and RBPJ, at different points in development, cell cycle, hormonal, or tissue specific effects. While the structure of EBF dimers on DNA appears to preclude RBPJ association with the embedded EBF motif, RBPJ occupancy with an EBF monomer might be possible.

Since RBPJ is also downstream of Notch in regulating MYC and cell survival gene expression in T-cell Lymphocytic Leukemia (TLL) and may have similar roles in thymic T-cell and B-cell marginal zone development (44), comparison of EBNA2 and Notch effects will be important for improved understanding and better control of these pathways. In contrast to EBNA2, which are most frequently distant from affected genes, Notch tends to be within affected gene promoter sites. Similar to EBNA2 targeting of EBF and embedded RBPJ sites in B-cells, NOTCH frequently targeted ZNF143 and embedded RBPJ sites in TLL cells. However NOTCH interactions with ZNF143 sites were less associated with activated transcription than EBNA2 interactions with EBF. While EBNA2 localized with EBF and RELA to sites 100kb, 428kb, and 556kb up-stream of MYC in LCLs, the strongest Notch site was even further up-stream of MYC. Most interesting in terms of fundamental mechanism for EBNA2/RBPJ or Notch/RBPJ association with MYC, the 428kb and 556kb, MYC promoter sites have nearby CTCF binding sites that could enable upstream site association with the MYC promoter. Long range enhancer MYC regulation is also characteristic of human colon, prostate and breast cancer.

Interestingly, 31% of LCL EBNA2 sites had significant RELA binding, including MYC enhancer sites. The role of LMP1 up-regulated RELA in co-regulating MYC expression with EBNA2 requires further study. LMP1 activation of canonical and non-canonical NFkB and related MAP kinase pathways is critical for LCL survival. The data presented here indicate that EBNA2 and NFkB RELA have many other targets in common. EBNA2 and LMP1 gene co-regulation creates opportunities to investigate the extent to which these transcriptional effects are EBNA2 and/or LMP1 dependent. Inhibitors of either pathway may have dominant or synthetic lethal effects on LCL growth or survival.

5

Contrastive Learning Using Spectral Methods

5.1 Overview

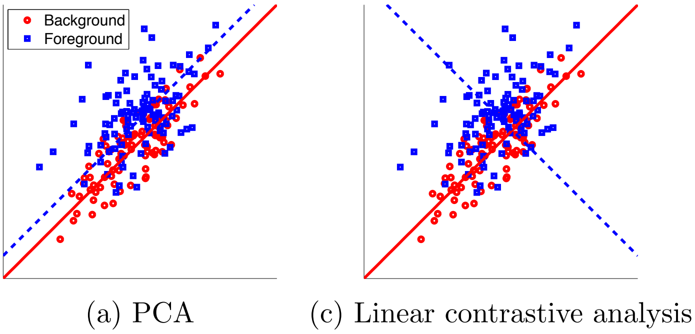
In many natural settings, the analysis goal is not to characterize a single data set in isolation, but rather to understand the difference between one set of observations and another. For example, given a background corpus of news articles together with writings of a particular author, one may want a topic model that explains word patterns and themes specific to the author. Another example comes from genomics, in which biological signals may be collected from different regions of a genome, and one wants a model that captures the differential statistics observed in these regions. This paper formalizes this notion of contrastive learning for mixture models, and develops spectral algorithms for inferring mixture components specific to a foreground data set when contrasted with a background data set. The method builds on recent moment-based estimators and tensor decompositions for latent variable models, and has the intuitive feature of using background data statistics to appropriately modify moments estimated from foreground data. A key advantage of the method is that the background data need only be coarsely modeled, which is important when the background is too complex, noisy, or not of interest. The method is demonstrated on applications in contrastive topic modeling and genomic sequence analysis.

5.2 Introduction

Generative latent variable models offer an intuitive way to explain data in terms of hidden structure, and are a cornerstone of exploratory data analysis. Popular examples of generative latent variable models include Latent Dirichlet Allocation (LDA) (9) and Hidden Markov Models (HMMs) (7), although the modularity of the generative approach has led to a wide range of variations. One of the challenges of using latent variable models for exploratory data analysis, however, is developing models and learning techniques that accurately reflect the intuitions of the modeler. In particular, when analyzing multiple specialized data sets, it is often the case that the most salient statistical structure—that most easily found by fitting latent variable models—is shared across all the data and does not reflect interesting specific local structure. For example, if we apply a topic model to a set of English-language scientific papers on computer science, we might hope to identify different co-occurring words within subfields such as theory, systems, graphics, *etc.* Instead, such a model will simply learn about English syntactic structure and invent topics that reflect uninteresting statistical correlations between stop words (88). Intuitively, what we would like from such an exploratory analysis is to answer the question: *What makes these data different from other sets of data in the same broad category?*

To answer this question, we develop a new set of techniques that we refer to as *contrastive learning* methods. These methods differentiate between *foreground* and *background* data and seek to learn a latent variable model that captures statistical relationships that appear in the foreground but do not appear in the background. Revisiting the previous scientific topics example, contrastive learning could treat computer science papers as a foreground corpus and (say) English-language news articles as a background corpus. As both corpora share the same broad syntactic structure, a contrastive foreground topic model would be more likely to discover semantic relationships between words that are specific to computer science. This intuition has broad applicability in other models and domains as well. For example, in genomics one might use a contrastive hidden Markov model to learn amplify the signal of a particular class of sequences, relative to the broader genome.

Note that the objective of contrastive learning is not to discriminate between foreground and background data, but to learn an interpretable generative model that cap-



(a) PCA (c) Linear contrastive analysis

Figure 5.1: These figures show foreground and background data from Gaussian distributions. The foreground data has greater variance in its minor direction, but the same variance in its major direction. The means are slightly different. Different projection lines are shown for different methods, to illustrate the difference between (a) the purely unsupervised variance-preserving linear projection of principal component analysis, (b) the contrastive foreground projection that captures variance that is not present in the background.

tures the differential statistics between the two data sets. To clarify this difference, consider the difference between principal component analysis and linear discriminant analysis. Principal component analysis finds the linear projection that maximally preserves variance without regard to foreground versus background. Linear discriminant analysis also finds a linear projection, but instead tries to maximize a measure of the margin between foreground and background. A contrastive approach, however, would try to find a linear projection that maximally preserves the foreground variance that is not explained by the background. Figure 5.1 illustrates the differences between these ideas.

Our contributions. We formalize the concept of contrastive learning for mixture models and present new spectral contrast algorithms. We prove that by appropriately "subtracting" background moments from the foreground moments, our algorithms recover the model for the foreground-specific data. To achieve this, we extend recent developments in learning latent variable models with moment matching and tensor decompositions. We demonstrate the effectiveness, robustness, and scalability of our method in contrastive topic modeling and contrastive genomics.

5.3 Contrastive learning in mixture models

Many data can be naturally described by a mixture model. The general mixture model has the form

$$p(\{x_n\}_{n=1}^N; \{(\mu_j, w_j)\}_{j=1}^J) = \prod_{n=1}^{N} \left[\sum_{j=1}^{J} w_j f(x_n|\mu_j) \right] \tag{5.1}$$

where $\{\mu_j\}$ are the parameters of the mixture components, $\{w_j\}$ are the mixture weights, and $f(\cdot|\mu_j)$ is the density of the j-th mixture component. Each μ_j is a vector in some parameter space, and a common estimation task is to infer the component parameters $\{(\mu_j, w_j)\}$ given the observed data $\{x_n\}$.

In many applications, we have two sets of observations $\{x_n^{\mathsf{f}}\}$ and $\{x_n^{\mathsf{b}}\}$, which we call the foreground data and the background data, respectively. The foreground and background are generated by two possibly overlapping sets of mixture components. More concretely, let $\{\mu_j\}_{j\in A}$, $\{\mu_j\}_{j\in B}$, and $\{\mu_j\}_{j\in C}$ be three disjoint sets of parameters, with A, B, and C being three disjoint index sets. The foreground $\{x_n^{\mathsf{f}}\}$ is generated from the mixture model $\{(\mu_j, w_j^{\mathsf{f}})\}_{j\in A\cup B}$, and the background $\{x_n^{\mathsf{b}}\}$ is generated from $\{(\mu_j, w_j^{\mathsf{b}})\}_{j\in B\cup C}$.

The goal of contrastive learning is to infer the parameters $\{(\mu_j, w_j^{\mathsf{f}})\}_{j\in A}$, which we call the *foreground-specific model*. The direct approach would be to infer $\{(\mu_j, w_j^{\mathsf{f}})\}_{j\in A\cup B}$ just from $\{x_n^{\mathsf{f}}\}$, and in parallel infer $\{(\mu_j, w_j^{\mathsf{b}})\}_{j\in B\cup C}$ just from $\{x_n^{\mathsf{f}}\}$, and then pick out the components specific to the foreground. However, this involves explicitly learning a model for the background data, which is undesirable if the background is too complex, if $\{x_n^{\mathsf{b}}\}$ is too noisy, or if we do not want to devote computational power to learn the background. In many applications, we are only interested in learning a generative model for the difference between the foreground and background, because that contrast is the interesting signal.

In this paper, we introduce an efficient and general approach to learn the foreground-specific model without having to learn an accurate model of the background. Our approach is based on a method-of-moments that uses higher-order tensor decompositions for estimation (2); we generalize the tensor decomposition technique to deal with our task of contrastive learning. Many other recent spectral learning algorithms for latent variable models are also based on the method-of-moments (*e.g.*, (1, 6, 11, 13, 14, 72, 73, 85)), but their parameter estimation can not account for the asymmetry between foreground and background.

We demonstrate spectral contrastive learning through two concrete applications: contrastive topic modeling and contrastive genomics. In contrastive topic modeling we are given a foreground corpus of documents and a background corpus. We want to learn a fully generative topic model that explains the foreground-specific documents (the contrast). We show that even when the background is extremely sparse—too noisy to learn a good background topic model—our spectral contrast algorithm still recovers foreground-specific topics. In contrastive genomics, sequence data is modeled by HMMs. The foreground data is generated by a mixture of two HMMs; one is foreground-specific, and the other captures some background process. The background data is generated by this second HMM. Contrastive learning amplifies the foreground-specific signal, which have meaningful biological interpretations.

5.4 Contrastive topic modeling

To illustrate contrastive analysis and introduce tensor methods, we consider a simple topic model where each document is generated by exactly one topic. In LDA (9), this corresponds to setting the Dirichlet prior hyper-parameter $\alpha \to 0$. The techniques here can be extended to the general $\alpha > 0$ case using the moment transformations given in (1). The generative topic model for a document is as follows.

- A word x is represented by an indicator vector $e_x \in \mathbb{R}^D$ which is 1 in its x-th entry and 0 elsewhere. D is the size of the vocabulary. A document is a bag-of-words and is represented by a vector $\mathsf{c} \in \mathbb{R}^D$ with non-negative integer word counts.
- A topic is first chosen according to the distribution on $[K] := \{1, 2, \ldots, K\}$ specified by the probability vector $w \in \mathbb{R}^K$.
- Given that the chosen topic is t, the words in the document are drawn independently from the distribution specified by the probability vector $\mu_t \in \mathbb{R}^D$.

Following previous work (*e.g.*, (1)) we assume that $\mu_1, \mu_2, \ldots, \mu_K$ are linearly independent probability vectors in $\mathbb{R}^D$. Let the foreground corpus of documents be generated by the mixture of $|A| + |B|$ topics $\{(\mu_t, w_t^{\mathsf{f}})\}_{t \in A} \cup \{(\mu_t, w_t^{\mathsf{f}})\}_{t \in B}$, and the background topics generated by the mixture of $|B| + |C|$ topics $\{(\mu_t, w_t^{\mathsf{b}})\}_{t \in B} \cup \{(\mu_t, w_t^{\mathsf{b}})\}_{t \in C}$

(here, we assume (A, B, C) is a non-trivial partition of $[K]$, and that $w_t^{\mathsf{f}}, w_t^{\mathsf{b}} > 0$ for all t). Our goal is to learn $\{(\mu_t, w_t^{\mathsf{f}})\}_{t\in A}$.

5.4.1 Moment decompositions

We use the symbol $\otimes$ to denote the tensor product of vectors, so $a \otimes b$ is the matrix whose (i, j)-th entry is $a_i b_j$, and $a \otimes b \otimes c$ is the third-order tensor whose (i, j, k)-th entry is $a_i b_j c_k$. Given a third-order tensor $T \in \mathbb{R}^{d_1\times d_2\times d_3}$ and vectors $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, and $c \in \mathbb{R}^{d_3}$, we let $T(I, b, c) \in \mathbb{R}^{d_1}$ denote the vector whose i-th entry is $\sum_{j,k} T_{i,j,k} b_j c_k$, and $T(a, b, c)$ denote the scalar $\sum_{i,j,k} T_{i,j,k} a_i b_j c_k$.

We review the moments of the word observations in this model (see, *e.g.*, (1)). Let $x_1, x_2, x_3 \in [D]$ be the first, second, and third words in a random document generated by the foreground model (the discussion here also applies to the background model). The second-order (cross) moment matrix $M_2^{\mathsf{f}} := \mathbb{E}[e_{x_1} \otimes e_{x_2}]$ is the matrix whose (i, j)-th entry is the probability that $x_1 = i$ and $x_2 = j$. Similarly, the third-order (cross) moment tensor $M_3^{\mathsf{f}} := \mathbb{E}[e_{x_1} \otimes e_{x_2} \otimes e_{x_3}]$ is the third-order tensor whose (i, j, k)-th entry is the probability that $x_1 = i, x_2 = j, x_3 = k$. Observe that for any $t \in A \cup B$, the i-th entry of $\mathbb{E}[e_{x_1}|\mathsf{topic} = t]$ is precisely the probability that $x_1 = i$ given $\mathsf{topic} = t$, which is i-th entry of μ_t. Therefore, $\mathbb{E}[e_{x_1}|\mathsf{topic} = t] = \mu_t$. Since the words are independent given the topic, the (i, j)-th entry of $\mathbb{E}[e_{x_1} \otimes e_{x_2}|\mathsf{topic} = t]$ is the product of the i-th and j-th entry of μ_t, *i.e.*, $\mathbb{E}[e_{x_1} \otimes e_{x_2}|\mathsf{topic} = t] = \mu_t \otimes \mu_t$. Similarly, $\mathbb{E}[e_{x_1} \otimes e_{x_2} \otimes e_{x_3}|\mathsf{topic} = t] = \mu_t \otimes \mu_t \otimes \mu_t$. Averaging over the choices of $t \in A \cup B$ with the weights w_t^{f} implies that the second- and third-order moments are

$$M_2^{\mathsf{f}} = \mathbb{E}[e_{x_1}\otimes e_{x_2}] = \sum_{t\in A\cup B} w_t^{\mathsf{f}}\, \mu_t\otimes\mu_t \quad \text{and} \quad M_3^{\mathsf{f}} = \mathbb{E}[e_{x_1}\otimes e_{x_2}\otimes e_{x_3}] = \sum_{t\in A\cup B} w_t^{\mathsf{f}}\, \mu_t\otimes\mu_t\otimes\mu_t.$$

(We discuss how to efficiently use documents of length > 3 in Section 5.6.2.) We can similarly decompose the background moments M_2^b and M_3^b in terms of tensors products of $\{\mu_t\}_{t\in B\cup C}$. These equations imply the following proposition.

Proposition 1. *Let M_2^{f}, M_3^{f} and M_2^{b}, M_3^{b} be the second- and third-order moments from the foreground and background data, respectively. Define*

$$M_2 := M_2^{\mathsf{f}} - \gamma M_2^{\mathsf{b}} \quad \textit{and} \quad M_3 := M_3^{\mathsf{f}} - \gamma M_3^{\mathsf{b}}.$$

If $\gamma \geq \max_{j\in B} w_j^{\mathsf{f}}/w_j^{\mathsf{b}}$, then

Algorithm 1 Contrastive Topic Model estimator

```
input Foreground and background documents $\{\mathsf{c}^{\mathsf{f}}_n\}$, $\{\mathsf{c}^{\mathsf{b}}_n\}$; parameter $\gamma > 0$; number of topics $K$.
output Foreground-specific topics $\mathsf{Topics}_{\mathsf{f}}$.
Let $\hat{M}^{\mathsf{f}}_2$ and $\hat{M}^{\mathsf{f}}_3$ ($\hat{M}^{\mathsf{b}}_2$ and $\hat{M}^{\mathsf{b}}_3$) be the foreground (background) second- and third-order moment estimates based on $\{\mathsf{c}^{\mathsf{f}}_n\}$ ($\{\mathsf{c}^{\mathsf{b}}_n\}$), and let $\hat{M}_2 := \hat{M}^{\mathsf{f}}_2 - \gamma \hat{M}^{\mathsf{b}}_2$ and $\hat{M}_3 := \hat{M}^{\mathsf{f}}_3 - \gamma \hat{M}^{\mathsf{b}}_3$.
Run Algorithm 2 with input $\hat{M}_2, \hat{M}_3$, $K$, and $N$ to obtain $\{(\hat{a}_t, \hat{\lambda}_t) : t \in [K]\}$.
$\mathsf{Topics}_{\mathsf{f}} := \{(\hat{a}_t / \|\hat{a}_t\|_1, 1/\hat{\lambda}_t^2) : t \in [K], \hat{\lambda}_t > 0\}$.
```

$$M_2 = \sum_{t=1}^{K} \omega_t \ \mu_t \otimes \mu_t \quad \textit{and} \quad M_3 = \sum_{t=1}^{K} \omega_t \ \mu_t \otimes \mu_t \otimes \mu_t \tag{5.2}$$

where $\omega_t = w^{\mathsf{f}}_t > 0$ for $t \in A$ (foreground-specific topic), and $\omega_t \leq 0$ for $t \in B \cup C$.

Using tensor decompositions. Proposition 1 implies that the modified moments M_2 and M_3 have low-rank decompositions in which the components t with positive multipliers ω_t correspond to the foreground-specific topics $\{(\mu_t, w^{\mathsf{f}}_t)\}_{t \in A}$. A main technical innovation of this paper is a generalized tensor power method, described in Section 5.6, which takes as input (estimates of) second- and third-order tensors of the form in (5.2), and approximately recovers the individual components. We argue that under some natural conditions, the generalized power method is robust to large perturbations in M^{b}_2 and M^{b}_3, which suggests that foreground-specific topics can be learned even when it is not possible to accurately model the background. We use the generalized tensor power method to estimate the foreground-specific topics in our Contrastive Topic Model estimator (Algorithm 1). Proposition 1 gives the lower bound on γ; we empirically find that $\gamma \approx \max_{j \in B} w^{\mathsf{f}}_j / w^{\mathsf{b}}_j$ gives good results. When γ is too large, the convergence of the tensor power worsens. Where possible in practice, we recommend using prior belief about foreground and background compositions to estimate $\max_{j \in B} w^{\mathsf{f}}_j / w^{\mathsf{b}}_j$, and then vary γ as part of the exploratory analysis.

5.4.2 Experiments with contrastive topic modeling

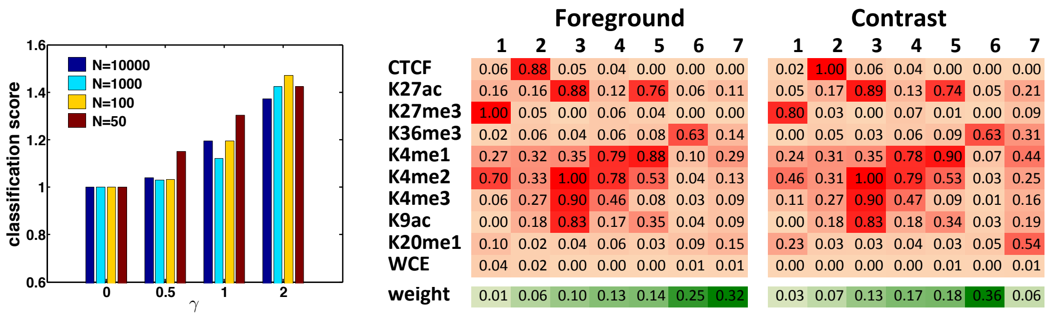
We test our contrastive topic models on the RCV1 dataset, which consists of ≈ 800000 news articles. Each document comes with multiple category labels (*e.g.*, economics,

Table 5.1: Top words from representative topics: foreground alone (left); foreground/background contrast (right). Each column corresponds to one topic.

USA foreground				USA foreground, Economics background			
lbs	bond	million	stock	play	research	result	basketball
usda	municipal	week	price	round	science	hockey	game
hog	index	sale	close	golf	cancer	nation	nation
gilt	year	export	trade	open	cell	cap	la
barrow	trade	total	index	hole	study	ny	association
China foreground				China foreground, Economics background			
share	billion	shanghai	yuan	china	panda	earthquake	china
market	reserve	yuan	year	east	china	china	office
percent	bank	firm	bank	typhoon	year	office	court
million	balance	china	foreign	storm	xinhua	richt	smuggle
trade	trade	exchange	invest	flood	zoo	scale	ship

entertainment) and region labels (*e.g.*, USA, Europe, China). The corpus spans a large set of complex and overlapping categories, making this a good dataset to validate our contrastive learning algorithm.

In one set of experiments, we take documents associated with one region as the foreground corpus, and documents associated with a general theme, such as economics, as the background. The goal of the contrast is to find the region-specific topics which are not relevant to the background theme. The top half of Table **??** shows the example where we take USA-related documents as the foreground and Economics as the background theme. We first set the contrast parameter $\gamma = 0$ in Algorithm 1; this learns the topics from the foreground dataset alone. Due to the composition of the corpus, the foreground topics for USA is dominated by topics relevant to stock markets and trade; representative topics and keywords are shown on the left of Table **??**. Then we increase γ to observe the effects of contrast. In the right half of Table **??**, we show the heavily weighted topics and keywords for when $\gamma = 2$. The topics involving market and trade are also present in the background corpus, so their weights are reduced through contrast. Topics which are very USA-specific and distinct from economics rise to the top: basketball, baseball, scientific research, *etc.* A similar experiment with China-related articles as foreground, and the same economics themed background is shown in



(a)

	Foreground							Contrast						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
CTCF	0.06	0.88	0.05	0.04	0.00	0.00	0.00	0.02	1.00	0.06	0.04	0.00	0.00	0.00
K27ac	0.16	0.16	0.88	0.12	0.76	0.06	0.11	0.05	0.17	0.89	0.13	0.74	0.05	0.21
K27me3	1.00	0.05	0.00	0.06	0.04	0.00	0.00	0.80	0.03	0.00	0.07	0.01	0.00	0.09
K36me3	0.02	0.06	0.04	0.06	0.08	0.63	0.14	0.00	0.05	0.03	0.06	0.09	0.63	0.31
K4me1	0.27	0.32	0.35	0.79	0.88	0.10	0.29	0.24	0.31	0.35	0.78	0.90	0.07	0.44
K4me2	0.70	0.33	1.00	0.78	0.53	0.04	0.13	0.46	0.31	1.00	0.79	0.53	0.03	0.25
K4me3	0.06	0.27	0.90	0.46	0.08	0.03	0.09	0.11	0.27	0.90	0.47	0.09	0.01	0.16
K9ac	0.00	0.18	0.83	0.17	0.35	0.04	0.09	0.00	0.18	0.83	0.18	0.34	0.03	0.19
K20me1	0.10	0.02	0.04	0.06	0.03	0.09	0.15	0.23	0.03	0.03	0.04	0.03	0.05	0.54
WCE	0.04	0.02	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01
weight	0.01	0.06	0.10	0.13	0.14	0.25	0.32	0.03	0.07	0.13	0.17	0.18	0.36	0.06

(b)

Figure 5.2: (a) Relative AUC as function of γ. (b) Emission probabilities of HMM states.

the bottom of Table **??**.

These examples illustrate that Algorithm 1 learns topics which are unique to the foreground. To quantify this effect, we devised a specificity test. Using the RCV1 labels, we partition the foreground USA documents into two disjoint groups: documents with any economics-related labels (group 0) and the rest (group 1). Because Algorithm 1 learns the full probabilistic model, we use the inferred topic parameters to compute the marginal likelihood for each foreground document given the model. We then use the likelihood value to classify each foreground document as belonging to group 0 or 1. The performance of the classifier is summarized by the AUC score.

We first set $\gamma = 0$ and compute the AUC score, which corresponds to how well a topic model learned from only the foreground can distinguish between the two groups. We use this score as the baseline and normalize so it is equal to 1. The hope is that by using the background data, the contrastive model can better identify the documents that are generated by foreground-specific topics. Indeed, as γ increases, the AUC score improves significantly over the benchmark (dark blue bars in Figure 5.2(a)). For $\gamma > 2$ we find that the foreground specific topics do not change qualitatively.

A major advantage of our approach is that we do not need to learn a very accurate background model to learn the contrast. To validate this, we down sample the background corpus to 1000, 100, and 50 documents. This simulates settings where the background is very sparsely sampled, so it is not possible to learn a background model very accurately. Qualitatively, we observe that even with only 50 randomly sampled background documents, Algorithm 1 still recovers topics specific to USA and not related to Economics. At $\gamma = 2$, it learns sports and NASA/space as the most prominent foreground-specific topics. This is supported by the specificity test, where contrastive topic models with sparse background better identify foreground-specific documents relative to the $\gamma = 0$ (foreground data-only) model.

5.5 Contrastive Hidden Markov Models

Hidden Markov Models (HMMs) are commonly used to model sequence and time series data. For example, a biologist may collect several sequences from an experiment; some of the sequences are generated by a biological process of interest (modeled by an HMM), while others are generated by a different "background" process—*e.g.*, noise or a process that is not of primary interest.

Consider a simple generative process where foreground data are generated by a mixture of two HMMs: $(1-\gamma)\,\mathrm{HMM}^A + \gamma\,\mathrm{HMM}^B$, and background data are generated by HMM^B. The goal is to learn the parameters of HMM^A, which models the biological process of interest. As we did for topic models, we can estimate a contrastive HMM by taking appropriate combinations of observable moments. Let $x_1^{\mathsf{f}}, x_2^{\mathsf{f}}, x_3^{\mathsf{f}}, \ldots$ be a random observation sequence in $\mathbb{R}^D$ generated by the foreground model $(1-\gamma)\,\mathrm{HMM}^A + \gamma\,\mathrm{HMM}^B$, and $x_1^{\mathsf{b}}, x_2^{\mathsf{b}}, x_3^{\mathsf{b}}, \ldots$ be the sequence generated by the background model HMM^B. Following (2), it suffices to estimate the following cross moment matrices and tensors: $M_{1,2}^{\mathsf{f}} := \mathbb{E}[x_1^{\mathsf{f}} \otimes x_2^{\mathsf{f}}]$, $M_{1,3}^{\mathsf{f}} := \mathbb{E}[x_1^{\mathsf{f}} \otimes x_3^{\mathsf{f}}]$, $M_{2,3}^{\mathsf{f}} := \mathbb{E}[x_2^{\mathsf{f}} \otimes x_3^{\mathsf{f}}]$, $M_{1,2,3}^{\mathsf{f}} := \mathbb{E}[x_1^{\mathsf{f}} \otimes x_2^{\mathsf{f}} \otimes x_3^{\mathsf{f}}]$, as well as the corresponding moments for the background model (here, we only use the first three observations in the sequence, but it is also justifiable to average over all consecutive observation triplets (78)). Then, an analogue of Proposition 1 (with suitable non-degeneracy assumptions as discussed in (2)) implies that we can recover the foreground-specific model HMM^A by using an asymmetric generalization of Algorithm 2 (see Appendix 5.8.4) with $M_{u,v} := M_{u,v}^{\mathsf{f}} - \gamma M_{u,v}^{\mathsf{b}}$ (for $\{u, v\} \subset \{1, 2, 3\}$) and $M_{1,2,3} := M_{1,2,3}^{\mathsf{f}} - \gamma M_{1,2,3}^{\mathsf{b}}$.

Application to contrastive genomics. For many biological problems, it is important to understand how signals in certain data are enriched relative to some related background data. For instance, we may want to contrast foreground data composed of gene expressions (or mutation rates, protein levels, *etc*) from one population against background data taken from (say) a control experiment, a different cell type, or a different time point. The contrastive analysis methods developed here can be a powerful exploratory tool for biology.

As a concrete illustration, we use spectral contrast to refine the characterization of chromatin states. The human genome consists of ≈ 3 billion DNA bases, and has recently been shown that these bases can be naturally segmented into a handful of chromatin states (38, 43). Each state describes a set of genomic properties: several states describe different active and regulatory features, while other states describe repressive features. The chromatin state varies across the genome, remaining constant for relatively short regions (say, several thousand bases). Learning the nature of the chromatin states is of great interest in genomics. The state-of-the-art approach for modeling chromatin states uses an HMM (38). The observable data are, at every 200 bases, a binary feature vector in $\{0, 1\}^{10}$. Each feature indicates the presence/absence of a specific chemical feature at that site (assumed independent given the chromatin state). This correspond to ≈ 15 million observations across the genome, which are used to learn the parameters of an HMM. Each chromatin state correspond to a latent state, characterized by a vector of 10 emission probabilities.

We take as foreground data the observations from exons, introns and promoters, which account for about 30% of the genome; as background data, we take observations from intergenic regions. Because exons and introns are transcribed, we expect the foreground to be a mixture of functional chromatin states and spurious states due to noise, and expect more of the background observations to be due to non-functional process. The contrastive HMM should capture biologically meaningful signals in the foreground data. In Figure 5.2(b), we show the emission matrix for the foreground HMM and for the contrastive HMM. We learn $K = 7$ latent states, corresponding to 7 chromatin states. Each row is a chemical feature of the genome. The foreground states recover the known biological chromatin states from literature (38). For example, state 6, with high emission for K36me3, is transcribed genes; state 5 is active enhancers; state 4 is poised enhancers. In the contrastive HMM, most of the states are the same as before. Interestingly, state 7, which is associated with feature K20me1, drops from the largest component of the foreground to a very small component of the contrast. This finding suggests that state 7 and K20me1 are less specific to gene bodies than previously thought (30), and raises more questions regarding its function, which is relatively unknown.

5.6 Generalized tensor power method

We now describe our general approach for tensor decomposition used in Algorithm 1. Let $a_1, a_2, \ldots, a_K \in \mathbb{R}^D$ be linearly independent vectors, and set $A := [a_1|a_2|\cdots|a_K]$. Let $M_2 := \sum_{i=1}^K \sigma_i a_i \otimes a_i$ and $M_3 := \sum_{i=1}^K \lambda_i a_i \otimes a_i \otimes a_i$, where $\sigma_i = \mathrm{sign}(\lambda_i) \in \{\pm 1\}$. The goal is to recover $\{(a_t, \lambda_t) : t \in [K]\}$ from (estimates of) M_2 and M_3.

The following proposition shows that one of the vectors a_i (and its associated λ_i) can be obtained from M_2 and M_3 using a simple power method similar to that from (2, 81) (note that which of the K components is obtained depends on the initialization of the procedure). Note that the error ε is exponentially small in 2^t after t iterations, so the number of iterations required to converge is very small. Below, we use $(\cdot)^\dagger$ to denote the Moore-Penrose pseudoinverse.

Proposition 2 (Informal statement)**.** *Consider the sequence $u^{(0)}, u^{(1)}, \ldots$ in $\mathbb{R}^D$ determined by $u^{(i+1)} := M_3(I, M_2^\dagger u^{(i)}, M_2^\dagger u^{(i)})$. Then for any $\varepsilon \in (0,1)$ and almost all $u^{(0)} \in \mathrm{range}(A)$, there exists $t^* \in [K]$, $c_1, c_2 > 0$ such that*

$$\|\tilde{u}^{(i)} - a_{t^*}\|^2 \leq \varepsilon$$

and

$$|\tilde{\lambda} - |\lambda_{t^*}|| \leq |\lambda_{t^*}|\varepsilon + \max_{t \neq t^*} |\lambda_t| \varepsilon^{3/2}$$

for $\varepsilon := c_1 \exp(-c_2 2^i)$, where $\tilde{u}^{(i)} := \sigma_{t^*} u^{(i)} / \|A^\dagger u^{(i)}\|$, $\tilde{\lambda} := M_3(M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)})$.

See below for the formal statement and proof which give explicit dependencies. We use the iterations from Proposition 2 in our main decomposition algorithm (Algorithm 2), which is a variant of the main algorithm from (2). The main difference is that we do not require M_2 to be positive semi-definite, which is essential for our application, but requires subtle modifications. For simplicity, we assume we run Algorithm 2 with exact moments M_2 and M_3 — a detailed perturbation analysis would be similar to that in (2) but is beyond the scope of this paper. Proposition 2 shows that a single component can be accurately recovered, and we use deflation to recover subsequent components (normalization and deflation is further discussed in Appendix 5.8.2). As

Algorithm 2 Generalized Tensor Power Method

```
input $\hat{M}_2 \in \mathbb{R}^{D \times D}$; $\hat{M}_3 \in \mathbb{R}^{D \times D \times D}$; target rank $K$; number of iterations $N$.
output Estimates $\{(\hat{a}_t, \hat{\lambda}_t) : t \in [K]\}$.
1: Let $\hat{M}_2^{\dagger}$ := Moore-Penrose pseudoinverse of rank $K$ approximation to $\hat{M}_2$; initialize $T := \hat{M}_3$.
2: for $t = 1$ to $K$ do
3:     Randomly draw $u^{(0)} \in \mathbb{R}^D$ from any distribution with full support in the range of $\hat{M}_2$.
4:     Repeat power iteration update $N$ times: $u^{(i+1)} := T(I, \hat{M}_2^{\dagger} u^{(i)}, \hat{M}_2^{\dagger} u^{(i)})$.
5:     $\hat{a}_t := u^{(N)} / |\langle u^{(N)}, \hat{M}_2^{\dagger} u^{(N)} \rangle|^{1/2}$; $\hat{\lambda}_t := T(\hat{M}_2^{\dagger}\hat{a}_t, \hat{M}_2^{\dagger}\hat{a}_t, \hat{M}_2^{\dagger}\hat{a}_t)$; $T := T - |\hat{\lambda}_t| \hat{a}_t \otimes \hat{a}_t \otimes \hat{a}_t$.
6: end for
```

noted in (2), errors introduced in this deflation step have only a lower-order effect, and therefore it can be used reliably to recover all K components. For increased robustness, we actually repeat steps 3–5 in Algorithm 2 several times, and use the results of the trial in which $|\hat{\lambda}_t|$ takes the median value.

5.6.1 Robustness to sparse background sampling

Algorithm 1 can recover the foreground-specific $\{\mu_t\}_{t \in A}$ even with relatively small numbers of background data. We can illustrate this robustness under the that the support of the foreground-specific topics $S_0 := \cup_{t \in A} \operatorname{supp}(\mu_t)$ is disjoint from that of the other topics $S_1 := \cup_{t \in B \cup C} \operatorname{supp}(\mu_t)$ (similar to Brown clusters (10)). Suppose that M_2^{f} is estimated accurately using a large sample of foreground documents. Then because S_0 and S_1 are disjoint, Algorithm 1 (using sufficiently large γ) will accurately recover the topics $\{(\mu_t, w_t^{\mathsf{f}}) : t \in A\}$ in $\mathsf{Topics}_{\mathsf{f}}$. The remaining concern is that sampling errors will cause Algorithm 1 to mistakenly return additional topics in $\mathsf{Topics}_{\mathsf{f}}$, namely the topics $t \in B \cup C$. It thus suffices to guarantee that the *signs* of the $\hat{\lambda}_t$ returned by Algorithm 2 are correct. The sample size requirement for this is *independent of the desired accuracy level for the foreground-specific topics*—it depends only on γ and fixed properties of the background model.[1] As reported in Section 5.4.2, this robustness is borne out in our experiments.

[1] For instance, if the background model consists only of one topic μ, then the analyses from (1, 2) can be adapted to bound the sample size requirement by $O(1/\|\mu\|^6)$.

5.6.2 Scalability

Our algorithms are scalable to large datasets when implemented to exploit sparsity and low-rank structure (each experiment we report runs on a standard laptop in a few minutes). Two important details are (i) how the moments M_2 and M_3 are represented, and (ii) how to compute execute the power iteration update in Algorithm 2. These issues are only briefly mentioned in (2) and without proof, so in this section, we address these issues in detail.

Efficient moment estimates for topic models. We first discuss how to represent empirical estimates of the second- and third-order moments M_2^{f} and M_3^{f} for the foreground documents (the same will hold for the background documents). Let document $n \in [N]$ have length ℓ_n, and let $\mathsf{c}_n \in \mathbb{N}^D$ be its word count vector (its i-th entry $\mathsf{c}_n(i)$ is the number of times word i appears in document n).

Proposition 3 (Estimator for M_2^{f})**.** *Assume $\ell_n \geq 2$. For any distinct $i, j \in [D]$, $\mathbb{E}[(\mathsf{c}_n(i)^2 - \mathsf{c}_n(i))/(\ell_n(\ell_n - 1))] = [M_2^{\mathsf{f}}]_{i,i}$ and $\mathbb{E}[\mathsf{c}_n(i)\mathsf{c}_n(j)/(\ell_n(\ell_n - 1))] = [M_2^{\mathsf{f}}]_{i,j}$.*

By Proposition 3, an unbiased estimator of M_2^{f} is $\hat{M}_2^{\mathsf{f}} := N^{-1} \sum_{n=1}^{N} (\ell_n(\ell_n - 1))^{-1} (\mathsf{c}_n \otimes \mathsf{c}_n - \operatorname{diag}(\mathsf{c}_n))$. Since $\hat{M}_2^{\mathsf{f}}$ is sum of sparse matrices, it can be represented efficiently, and we may use sparsity-aware methods for computing its low-rank spectral decompositions. It is similarly easy to obtain such a decomposition for $\hat{M}_2^{\mathsf{f}} - \gamma \hat{M}_2^{\mathsf{b}}$, from which one can compute its pseudoinverse and represent it in factored form as $PQ^\top$ for some $P, Q \in \mathbb{R}^{D \times K}$.

Proposition 4 (Estimator for M_3^{f})**.** *Assume $\ell_n \geq 3$. For any distinct $i, j, k \in [D]$, $\mathbb{E}[(\mathsf{c}_n(i)^3 - 3\mathsf{c}_n(i)^2 + 2\mathsf{c}_n(i))/(\ell_n(\ell_n - 1)(\ell_n - 2))] = [M_3^{\mathsf{f}}]_{i,i,i}$, $\mathbb{E}[(\mathsf{c}_n(i)^2\mathsf{c}_n(j) - \mathsf{c}_n(i)\mathsf{c}_n(j))/(\ell_n(\ell_n - 1)(\ell_n - 2))] = [M_3^{\mathsf{f}}]_{i,i,j}$, $\mathbb{E}[(\mathsf{c}_n(i)\mathsf{c}_n(j)\mathsf{c}_n(k))/(\ell_n(\ell_n - 1)(\ell_n - 2))] = [M_3^{\mathsf{f}}]_{i,j,k}$.*

By Proposition 4, an unbiased estimator of $M_3^{\mathsf{f}}(I, v, v)$ for any vector $v \in \mathbb{R}^D$ is $\hat{M}_3^{\mathsf{f}}(I, v, v) := N^{-1} \sum_{n=1}^{N} (\ell_n(\ell_n - 1)(\ell_n - 2))^{-1} (\langle \mathsf{c}_n, v \rangle^2 \mathsf{c}_n - 2\langle \mathsf{c}_n, v \rangle (\mathsf{c}_n \circ v) - \langle \mathsf{c}_n, v \circ v \rangle \mathsf{c}_n + 2\mathsf{c}_n \circ v \circ v)$ (where $\circ$ denotes component-wise product of vectors). Each term in the sum takes only $O(\operatorname{nnz}(\mathsf{c}_n))$ operations to compute, and each only has $\operatorname{nnz}(\mathsf{c}_n)$ non-zero entries. So the time to compute $\hat{M}_3^{\mathsf{f}}(I, v, v)$ is proportional to the number of non-zero entries of the term-document matrix, using just a single pass over the document corpus.

Power iteration computation. Each power iteration update in Algorithm 2 just requires the evaluating $\hat{M}_3^{\mathsf{f}}(I,v,v)-\gamma\hat{M}_3^{\mathsf{b}}(I,v,v)$ (one-pass linear time, as shown above) for $v := \hat{M}_2^{\dagger}u^{(i)}$, and computing the deflation $\sum_{\tau<t}\hat{\lambda}_\tau\langle\hat{a}_\tau,v\rangle^2\hat{a}_\tau$ ($O(DK)$ time). Since $\hat{M}_2^{\dagger}$ is kept in rank-K factored form, v can also be computed in $O(DK)$ time.

5.7 Discussion

In this paper, we formalize a model of contrastive learning and introduce efficient spectral methods to learn the model parameters specific to the foreground. Experiments with contrastive topic modeling show that Algorithm 1 can learn foreground-specific topics even when the background data is noisy. Our application in contrastive genomics illustrates the utility of this method in exploratory analysis of biological data. The contrast identifies an intriguing change associated with K20me1, which can be followed up with biological experiments. While we have focused in this work on a natural contrast model for mixture models, we also discuss an alternative approach below.

5.8 Supplemental analysis

5.8.1 Proof of Proposition 1

Proof of Proposition 1. This follows from the observation that

$$\begin{aligned} M_2 &= \sum_{t\in A} w_t^{\mathsf{f}}\ \mu_t\otimes\mu_t \ + \ \sum_{t\in B}(w_t^{\mathsf{f}}-\gamma w_t^{\mathsf{b}})\ \mu_t\otimes\mu_t \ + \ \sum_{t\in C}(-\gamma w_t^{\mathsf{b}})\ \mu_t\otimes\mu_t \\ M_3 &= \sum_{t\in A} w_t^{\mathsf{f}}\ \mu_t^{\otimes 3} \ + \ \sum_{t\in B}(w_t^{\mathsf{f}}-\gamma w_t^{\mathsf{b}})\ \mu_t^{\otimes 3} \ - \ \sum_{t\in C}\gamma w_t^{\mathsf{b}}\ \mu_t^{\otimes 3} \end{aligned}$$

and

$$w_t^{\mathsf{f}}-\gamma w_t^{\mathsf{b}} \le 0\ \forall t\in B \qquad \Longleftrightarrow \qquad \gamma \ge \max_{t\in B} w_t^{\mathsf{f}}/w_t^{\mathsf{b}}. \qquad \square$$

5.8.2 Generalized tensor power method

Normalization and deflation. By Proposition 2, the first for-loop iteration of Algorithm 2 recovers $u^{(N)}$ very close to $\sigma_{i^*}a_{i^*}$ for some $i^*\in[K]$, up to positive scaling $s := \|A^{\dagger}u^{(N)}\|$. Because

$$1 = (\sigma_{i^*}\langle a_{i^*}, M_2^{\dagger}a_{i^*}\rangle)^{1/2} = |\langle a_{i^*}, M_2^{\dagger}a_{i^*}\rangle|^{1/2},$$

this scaling s is close to $|\langle u^{(N)}, M_2^\dagger u^{(N)}\rangle|^{1/2}$, which is the normalization used in Algorithm 2. Thus, the estimates $\hat{a}_1$ and $\hat{\lambda}_1$ are close to $\sigma_{i^*} a_{i^*}$ and λ_{i^*}, respectively. For the next for-loop iteration, we want to execute the power iteration with a tensor close to $T - \lambda_{i^*} a_{i^*} \otimes a_{i^*} \otimes a_{i^*}$ in order to recover a component different from a_{i^*}. Therefore we use

$$M_3 - |\hat{\lambda}_1| \hat{a}_1 \otimes \hat{a}_1 \otimes \hat{a}_1 \approx M_3 - \sigma_{i^*} |\lambda_{i^*}| a_{i^*} \otimes a_{i^*} \otimes a_{i^*} = M_3 - \lambda_{i^*} a_{i^*} \otimes a_{i^*} \otimes a_{i^*}$$

(the crucial detail is the absolute value on $\hat{\lambda}_1$).

Convergence analysis.

Proposition 5. *Let $u^{(0)} \in \operatorname{range}(A)$, and consider the sequence determined by*

$$u^{(i+1)} := M_3(I, M_2^\dagger u^{(i)}, M_2^\dagger u^{(i)}).$$

Define

$$t^* := \arg\max_{t \in [K]} |\lambda_t \langle e_t, A^\dagger u^{(0)}\rangle|, \quad \rho := \max_{t \neq t^*} \left| \frac{\lambda_t \langle e_t, A^\dagger u^{(0)}\rangle}{\lambda_{t^*} \langle e_{t^*}, A^\dagger u^{(0)}\rangle} \right|, \quad \varepsilon := \rho^{2^{i+1}} \lambda_{t^*}^2 \sum_{t \neq t^*} \lambda_t^{-2},$$

$$\tilde{u}^{(i)} := \sigma_{t^*} u^{(i)} / \|A^\dagger u^{(i)}\|.$$

Then

$$\|A^\dagger(\tilde{u}^{(i)} - a_{t^*})\|^2 \leq 2\varepsilon,$$
$$\left| M_3(M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)}) - |\lambda_{t^*}| \right| \leq |\lambda_{t^*}| \cdot \varepsilon + \max_{t \neq t^*} |\lambda_t| \cdot \varepsilon^{1.5}.$$

Proof. Define $f_t := \langle e_t, A^\dagger u^{(0)}\rangle$, and without loss of generality, assume $|\lambda_1 f_1| \geq |\lambda_2 f_2| \geq \cdots \geq |\lambda_K f_K|$. Then, using the definition $u^{(1)} = M_3(I, M_2^\dagger u^{(0)}, M_2^\dagger u^{(0)})$ and the facts that A has full column rank and Σ is invertible, we have

$$\begin{aligned} u^{(1)} &= \sum_{t=1}^K \lambda_t \langle a_t, M_2^\dagger u^{(0)}\rangle^2 a_t \\ &= \sum_{t=1}^K \lambda_t \langle a_t, (A^\top)^\dagger \Sigma^{-1} A^\dagger u^{(0)}\rangle^2 a_t \\ &= \sum_{t=1}^K \lambda_t \sigma_t^{-2} \langle e_t, A^\dagger u^{(0)}\rangle^2 a_t \\ &= \sum_{t=1}^K \lambda_t f_t^2 a_t, \end{aligned}$$

which implies $\langle e_t, A^\dagger u^{(0)}\rangle = \lambda_t f_t^2$. By induction, $\langle e_t, A^\dagger u^{(i)}\rangle = \lambda_t^{2^i-1} f_t^{2^i}$. Therefore

$$1-\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle^2 = 1-\frac{\langle e_1, A^\dagger u^{(i)}\rangle^2}{\sum_{t=1}^K \langle e_t, A^\dagger u^{(i)}\rangle^2} = 1-\frac{|\lambda_1|^{2^{i+1}-2} f_1^{2^{i+1}}}{\sum_{t=1}^K |\lambda_t|^{2^{i+1}-2} f_t^{2^{i+1}}} \le \rho^{2^{i+1}} \lambda_1^2 \sum_{t=2}^K \lambda_t^{-2} = \varepsilon.$$

Moreover, $\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle = |\lambda_1|^{2^i-1} f_1^{2^i} / \sqrt{\sum_{t=1}^K |\lambda_t|^{2^{i+1}-2} f_t^{2^{i+1}}} \in [0,1]$, so $\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle \ge \langle e_1, A^\dagger \tilde{u}^{(i)}\rangle^2$. Therefore, using the fact that $\|A^\dagger \tilde{u}^{(i)}\| = \|A^\dagger a_1\| = 1$, we can bound $\|A^\dagger(\tilde{u}^{(i)} - a_1)\|^2$ as

$$\|A^\dagger(\tilde{u}^{(i)} - a_1)\|^2 = 2(1-\langle a_1, (AA^\top)^\dagger \tilde{u}^{(i)}\rangle) = 2(1-\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle) \le 2(1-\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle^2) \le 2\varepsilon.$$

It remains to show that $M_3(M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)})$ is close to $|\lambda_1|$. We have that

$$\begin{aligned} M_3(M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)}) &= \sum_{t=1}^K \lambda_t \langle a_t, M_2^\dagger \tilde{u}^{(i)}\rangle^3 \\ &= \sum_{t=1}^K |\lambda_t| \langle e_t, A^\dagger \tilde{u}^{(i)}\rangle^3 \\ &= |\lambda_1| \langle e_1, A^\dagger \tilde{u}^{(i)}\rangle + \sigma_1 \sum_{t=2}^K |\lambda_t| \left(\frac{\langle e_t, A^\dagger u^{(i)}\rangle}{\sqrt{\sum_{j=1}^K \langle e_j, A^\dagger u^{(i)}\rangle^2}} \right)^3. \end{aligned}$$

Since $(1+\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle)(1-\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle) = 1-\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle^2 \le \varepsilon$ and $\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle \in [0,1]$, it follows that $|1-\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle| = (1-\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle) \le \varepsilon/(1+\langle e_1, A^\dagger \tilde{u}^{(i)}\rangle) \le \varepsilon$. Furthermore, by Hölder's inequality, the triangle inequality, and the fact that $(\sum_t |v_t|^3)^{1/3} \le (\sum_t v_t^2)^{1/2}$,

$$\begin{aligned} \left| \sum_{t=2}^K |\lambda_t| \left(\frac{\langle e_t, A^\dagger u^{(i)}\rangle}{\sqrt{\sum_{j=1}^K \langle e_j, A^\dagger u^{(i)}\rangle^2}} \right)^3 \right| &\le \left(\max_{t>1} |\lambda_t| \right) \frac{\sum_{t=2}^K |\langle e_t, A^\dagger u^{(i)}\rangle|^3}{\left(\sum_{j=1}^K \langle e_j, A^\dagger u^{(i)}\rangle^2\right)^{3/2}} \\ &\le \max_{t>1} |\lambda_t| \left(\frac{\sum_{t=2}^K \langle e_t, A^\dagger u^{(i)}\rangle^2}{\sum_{j=1}^K \langle e_j, A^\dagger u^{(i)}\rangle^2} \right)^{3/2} \\ &\le \max_{t>1} |\lambda_t| \varepsilon^{3/2}. \end{aligned}$$

Thus, again by the triangle inequality,

$$\left| M_3(M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)}, M_2^\dagger \tilde{u}^{(i)}) - |\lambda_1| \right| \le |\lambda_1| \varepsilon + \max_{t>1} |\lambda_t| \varepsilon^{3/2}. \qquad \square$$

5.8.3 Moment estimators

In the proofs of Propositions 3 and 4, we let $x_{n,1}, x_{n,2}, \ldots, x_{n,\ell_n} \in [D]$ be the words in document n, so $\mathsf{c}_n := \sum_{i=1}^{\ell_n} e_{x_{n,i}}$.

Proof of Proposition 3. For any $i \in [D]$,

$$\begin{aligned}
\mathbb{E}\Big[\mathsf{c}_n(i)^2 - \mathsf{c}_n(i)\Big] &= \mathbb{E}\left[\left(\sum_{p=1}^{\ell_n} x_{n,p}(i)\right)^2 - \sum_{p=1}^{\ell_n} x_{n,p}(i)\right] \\
&= \mathbb{E}\left[\sum_{p=1}^{\ell_n} x_{n,p}(i)^2 + 2\sum_{p<q} x_{n,p}(i)x_{n,q}(i) - \sum_{p=1}^{\ell_n} x_{n,p}(i)\right] \\
&= 2\sum_{p<q} \mathbb{E}\Big[x_{n,p}(i)x_{n,q}(i)\Big] \quad (\text{since } x_{n,p}(i)^2 = x_{n,p}(i)) \\
&= \ell_n(\ell_n - 1)[M_2^{\mathsf{f}}]_{i,i}.
\end{aligned}$$

For $i \neq j$,

$$\begin{aligned}
\mathbb{E}\Big[\mathsf{c}_n(i)\mathsf{c}_n(j)\Big] &= \mathbb{E}\left[\sum_{p=1}^{\ell_n} x_{n,p}(i) \sum_{q=1}^{\ell_n} x_{n,q}(j)\right] \\
&= \mathbb{E}\left[\sum_{p=1}^{\ell_n} x_{n,p}(i)x_{n,p}(j) + \sum_{p\neq q} x_{n,p}(i)x_{n,q}(j)\right] \\
&= \sum_{p\neq q} \mathbb{E}\Big[x_{n,p}(i)x_{n,q}(j)\Big] \quad (\text{since } x_{n,p}(i)x_{n,p}(j) = 0 \text{ for } i \neq j) \\
&= \ell_n(\ell_n - 1)[M_2^{\mathsf{f}}]_{i,j}.
\end{aligned}$$

□

5.8.4 Asymmetric generalized tensor power method

Let $\{a_1, a_2, \ldots, a_K\} \subset \mathbb{R}^{D_a}$, $\{b_1, b_2, \ldots, b_K\} \subset \mathbb{R}^{D_b}$, and $\{c_1, c_2, \ldots, c_K\} \subset \mathbb{R}^{D_c}$ be sets of linearly independent vectors. Let $M_{u,v} := \sum_{t=1}^{K} \sigma_t u_t v_t^\top$ for $(u, v) \in \{a, b, c\} \times \{a, b, c\}$, and $M_{a,b,c} := \sum_{t=1}^{K} \lambda_t a_t \otimes b_t \otimes c_t$, where $\sigma_t = \operatorname{sign}(\lambda_t) \in \{\pm 1\}$. Given (estimates of) $M_{a,b}, M_{a,c}, M_{b,c}, M_{a,b,c}$, Algorithm 3 approximates $\{(a_t, b_t, c_t, \lambda_t) : t \in [K]\}$. The proof of convergence (assuming exact estimates of $M_{a,b}, M_{a,c}, M_{b,c}, M_{a,b,c}$) is very similar to Proposition 2 and is thus omitted.

5.8.5 An alternative model of contrast

We consider a different generative model in which the topic of a document is a simple (fixed) mixture of a foreground topic and a background topic (say, $0.9\mu_t^{\mathsf{f}} + 0.1\mu_{t'}^{\mathsf{b}}$ with probability $w_{t,t'}$, for $t \in [K^{\mathsf{f}}]$ and $t \in [K^{\mathsf{b}}]$). One can treat this using the previous model with $K = K^{\mathsf{f}}K^{\mathsf{b}}$ topics, but there are really only $K^{\mathsf{f}} + K^{\mathsf{b}}$ topics. Using auxiliary background data which is modeled by a topic model over just the background topics

Algorithm 3 Asymmetric Generalized Tensor Power Method

```
input $\hat{M}_{a,b} \in \mathbb{R}^{D_a \times D_b}$; $\hat{M}_{a,c} \in \mathbb{R}^{D_a \times D_c}$; $\hat{M}_{b,c} \in \mathbb{R}^{D_b \times D_c}$; $\hat{M}_{a,b,c} \in \mathbb{R}^{D_a \times D_b \times D_c}$; target rank $K$; number of iterations $N$.
output Estimates $\{(\hat{a}_t, \hat{b}_t, \hat{c}_t, \hat{\lambda}_t) : t \in [K]\}$.
Let $\hat{M}^\dagger_{a,b}$ := Moore-Penrose pseudoinverse of rank $K$ approximation to $\hat{M}_{a,b}$; similarly define $\hat{M}^\dagger_{a,c}$ and $\hat{M}^\dagger_{b,c}$; let $\hat{M}^\dagger_{a,a} := \hat{M}^\dagger_{b,a}\hat{M}_{b,c}\hat{M}^\dagger_{a,c}$; initialize $T := \hat{M}_{a,b,c}$.
for $t = 1$ to $K$ do
    Randomly draw $u^{(0)} \in \mathbb{R}^D$ from any distribution with full support in the range of $\hat{M}_{a,b}$.
    Repeat power iteration update $N$ times: $u^{(i+1)} := T(I, \hat{M}^\dagger_{a,b}u^{(i)}, \hat{M}^\dagger_{a,c}u^{(i)})$.
    $\hat{a}_t := u^{(N)}/|\langle u^{(N)}, \hat{M}^\dagger_{a,a}u^{(N)}\rangle|^{1/2}$; $\hat{b}_t := \hat{M}_{b,c}\hat{M}^\dagger_{a,c}\hat{a}_t$; $\hat{c}_t := \hat{M}_{c,b}\hat{M}^\dagger_{a,b}\hat{a}_t$; $\hat{\lambda}_t := T(\hat{M}^\dagger_{a,a}\hat{a}_t, \hat{M}^\dagger_{a,b}\hat{a}_t, \hat{M}^\dagger_{a,c}\hat{a}_t)$; $T := T - |\hat{\lambda}_t|\hat{a}_t \otimes \hat{b}_t \otimes \hat{c}_t$.
end for
```

$\{\mu^{\mathsf{b}}_{t'} : t' \in [K^{\mathsf{b}}]\}$, it is possible to determine an orthogonal projector $\Pi \in \mathbb{R}^{D \times D}$ for the range of the second-order moments, which approximately captures the span of the $\{\mu^{\mathsf{b}}_{t'}\}$. Then, the projection $I - \Pi$ can be applied to the second- and third-order moments of the foreground documents (which is generated by mixed topics) to annihilate the background topic contributions: $(I - \Pi)(0.9\mu^{\mathsf{f}}_t + 0.1\mu^{\mathsf{b}}_{t'}) = 0.9(I - \Pi)\mu^{\mathsf{f}}_t$. If, in addition, the support of the foreground topics and background topics are disjoint (as in Brown clusters), then $(I - \Pi)\mu^{\mathsf{f}}_t = \mu^{\mathsf{f}}_t$. Therefore, one can directly estimate the K^{f} foreground topics using the foreground data. Moreover, we do not need to fully estimate the model for the background documents, as we only need the second-order (but not third-order) moments to determine Π.

We used this model to conduct experiments similar to those reported in Section 5.4.2 on the RCV1 dataset, and observed qualitatively similar results, but it was less numerically stable compared to Algorithm 1. Developing better estimators for this model is a promising direction of future research.

6

Priors for Diversity in Generative Latent Variable Models

6.1 Overview

Probabilistic latent variable models are one of the cornerstones of machine learning. They offer a convenient and coherent way to specify prior distributions over unobserved structure in data, so that these unknown properties can be inferred via posterior inference. Such models are useful for exploratory analysis and visualization, for building density models of data, and for providing features that can be used for later discriminative tasks. A significant limitation of these models, however, is that draws from the prior are often highly redundant due to i.i.d. assumptions on internal parameters. For example, there is no preference in the prior of a mixture model to make components non-overlapping, or in topic model to ensure that co-occurring words only appear in a small number of topics. In this work, we revisit these independence assumptions for probabilistic latent variable models, replacing the underlying i.i.d. prior with a determinantal point process (DPP). The DPP allows us to specify a preference for diversity in our latent vari- ables using a positive definite kernel function. Using a kernel between probability distributions, we are able to define a DPP on probability measures. We show how to perform MAP inference with DPP priors in latent Dirichlet allocation and in mixture models, leading to better intuition for the latent variable representation and quantitatively improved unsupervised feature extraction, without compromising the generative aspects of the model.

6.2 Introduction

The probabilistic generative model is an important tool for statistical learning because it enables rich data to be explained in terms of simpler latent structure. The discovered structure can be useful in its own right, for explanatory purposes and visualization, or it may be useful for improving generalization to unseen data. In the latter case, we might think of the inferred latent structure as providing a feature representation that summarizes complex high-dimensional interaction into a simpler form.

The core assumption behind the use of latent variables as features, however, is that the salient statistical properties discovered by unsupervised learning will be useful for discriminative tasks. This requires that the features span the space of possible data and represent diverse characteristics that may be important for discrimination. Diversity, however, is difficult to express within the generative framework. Most often, one builds a model where the feature representations are independent *a priori*, with the hope that a good fit to the data will require employing a variety of latent variables.

There is reason to think that this does not always happen in practice, and that during unsupervised learning, model capacity is often spent improving the density around the common cases, not allocating new features. For example, in a generative clustering model based on a mixture distribution, multiple mixture components will often be used for a single "intuitive group" in the data, simply because the shape of the component's density is not a close fit to the group's distribution. A generative mixture model will happily use many of its components to closely fit the density of a single group of data, leading to a highly redundant feature representation. Similarly, when applied to a text corpus, a topic model such as latent Dirichlet allocation (31) will place large probability mass on the same stop words across many topics, in order to fine-tune the probability assigned to the common case. In both of these situations, we would like the latent groupings to uniquely correspond to characteristics of the data: that a group of data should be explained by one mixture component, and that common stop words should be one category of words among many. This intuition expresses a need for diversity in the latent parameters of the model that goes beyond what is highly likely under the posterior distribution implied by an independent prior.

In this paper, we propose a modular approach to diversity in generative probabilistic models by replacing the independent prior on latent parameters with a *determinantal* *point process* (DPP). The determinantal point process enables a modeler to specify a notion of similarity on the space of interest, which in this case is a space of possible latent distributions, via a positive definite kernel. The DPP then assigns probabilities to particular configurations of these distributions according to the determinant of the Gram matrix. This construction naturally leads to a generative latent variable model in which diverse sets of latent parameters are preferred over redundant sets.

The determinantal point process is a convenient statistical tool for constructing a tractable point process with repulsive interaction. The DPP is more general than the Poisson process (see, e.g., (77)), which has no interaction, but more tractable than Strauss (86) and Gibbs/Markov (41) processes (at the cost of only being able to capture anticorrelation). Hough *et al.* (70) provides a useful survey of probabilistic properties of the determinantal point process, and for statistical properties, see, e.g., Scardicchio *et al.* (22) and Lavancier *et al.* (82). There has also been recent interest in using the DPP within machine learning for modeling sets of structures (18), and for conditionally producing diverse collections of objects (19). The approach we propose here is different from this previous work in that we are suggesting the use of a determinantal point process within a hierarchical model, and using it to enforce diversity among latent variables, rather than as a mechanism for diversity across directly observed discrete structures.

6.3 Diversity in Generative Latent Variable Models

In this paper we consider generic directed probabilistic latent variable models that produce distributions over a set of N data, denoted $\{x_n\}_{n=1}^N$, which live in a sample space X. Each of these data has a latent discrete label z_n, which takes a value in $\{1, 2, \cdots, J\}$. The latent label indexes into a set of parameters $\{\theta_j\}_{j=1}^J$. The parameters determined by z_n then produce the data according to a distribution $f(x_n \,|\, \theta_{z_n})$. Typically we use independent priors for the θ_j, here denoted by $\pi(\cdot)$, but the distribution over the latent indices z_n may be more structured. Taken together this leads to the generic joint distribution:

$$p(\{x_n, z_n\}_{n=1}^N, \{\theta_j\}_{j=1}^J) = p(\{z_n\}_{n=1}^N) \left[\prod_{n=1}^N f(x_n \,|\, \theta_{z_n}) \right] \prod_{j=1}^J \pi(\theta_j). \tag{6.1}$$

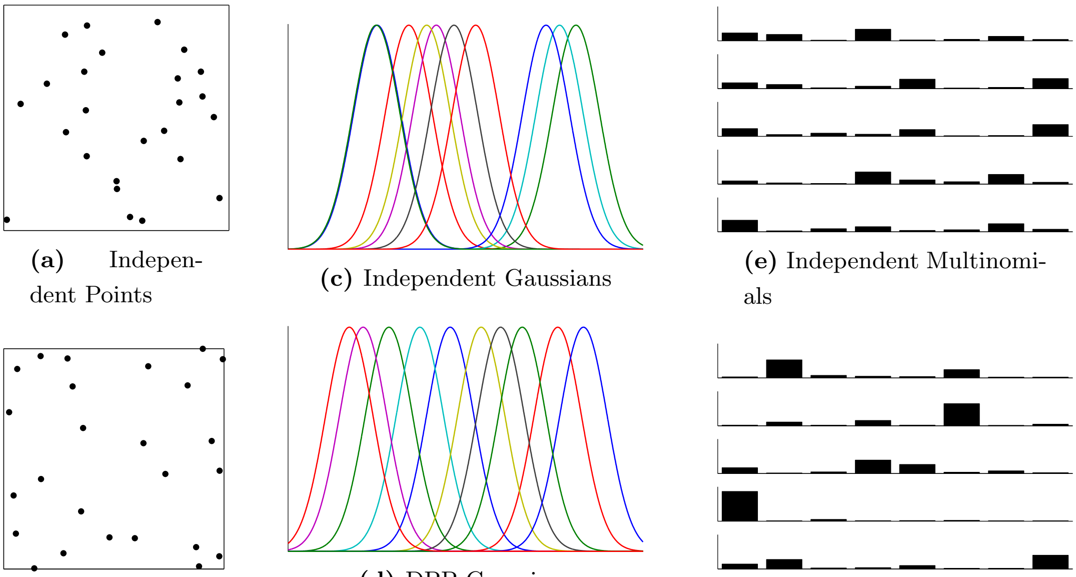
The details of each distribution are problem-specific, but this general framework appears in many contexts. For example, in a typical mixture model, the z_n are drawn independently from a multinomial distribution and the θ_j are the component-specific parameters. In an admixture model such as latent Dirichlet allocation (LDA) (31), the θ_j may be "topics", or distributions over words. In an admixture, the z_n may share structure based on, e.g., being words within a common set of documents.

These models are often thought of as providing a principled approach for feature extraction. At training time, one either finds the maximum of the posterior distribution $p(\{\theta_j\}_{j=1}^J \,|\, \{x_n\}_{n=1}^N)$ or collects samples from it, while integrating out the data-specific latent variables z_n. Then when presented with a test case $x^\star$, one can construct a conditional distribution over the corresponding unknown variable $z^\star$, which is now a "feature" that might usefully summarize many related aspects of $x^\star$. However, this interpretation of the model is suspect; we have not asked the model to make the z_n variables explanatory, except as a byproduct of improving the training likelihood. Different θ_j may assign essentially identical probabilities to the same datum, resulting in ambiguous features.

6.3.1 Measure-Valued Determinantal Point Process

In this work we propose an alternative to the independence assumption of the standard latent variable model. Rather than specifying $p(\{\theta_j\}_{j=1}^J) = \prod_j \pi(\theta_j)$, we will construct a determinantal point process on sets of component-specific distributions $\{f(x \,|\, \theta_j)\}_{j=1}^J$. Via the DPP, it will be possible for us to specify a preference for sets of distributions that have minimal overlap, as determined via a positive-definite kernel function between distributions. In the case of the simple parametric families for $f(\cdot)$ that we consider here, it is appropriate to think of the DPP as providing a "diverse" set of parameters $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^J$, where the notion of diversity is expressed entirely in terms of the resulting probability measure on the sample space $\mathcal{X}$. After MAP inference with this additional structure, the hope is that the θ_j will explain substantially different regions of $\mathcal{X}$ — appropriately modulated by the likelihood — and lead to improved, non-redundant feature extraction at test time.

We will use Θ to denote the space of possible θ. A realization from a *point process* on Θ produces a random finite subset of Θ. To construct a determinantal point process,



(a) Independent Points

(c) Independent Gaussians

(e) Independent Multinomials

(b) DPP Points

(d) DPP Gaussians

(f) DPP Multinomials

Figure 6.1: Illustrations of the determinantal point process prior. (a) 25 independent uniform draw in the unit square; (b) a draw from a DPP with 25 points; (c) ten Gaussian distributions with means uniformly drawn on the unit interval; (d) ten Gaussian distributions with means distributed according to a DPP using the probability product kernel; (e) five random discrete distributions; (f) five random discrete distributions drawn from a DPP on the simplex with the probability product kernel (74).

we first define a positive definite kernel on Θ, which we denote $K : \Theta \times \Theta \to \mathbb{R}$. The probability density associated with a particular finite $\boldsymbol{\theta} \subset \Theta$ is given by

$$p(\boldsymbol{\theta} \subset \Theta) \propto |\mathbf{K}_{\boldsymbol{\theta}}|, \tag{6.2}$$

where $\mathbf{K}_{\boldsymbol{\theta}}$ is the $|\boldsymbol{\theta}| \times |\boldsymbol{\theta}|$ positive definite Gram matrix that results from applying $K(\theta, \theta')$ to the elements of $\boldsymbol{\theta}$. The eigenspectrum of the kernel on Θ must be bounded to $[0, 1]$. The kernels we will focus on in this paper are composed of two parts: 1) a positive definite *correlation function* $R(\theta, \theta')$, where $R(\theta, \theta) = 1$, and 2) the "prior kernel" $\sqrt{\pi(\theta)\pi(\theta')}$, which expresses our marginal preferences for some parameters over others. These are combined to form the kernel of interest:

$$K(\theta, \theta') = R(\theta, \theta') \sqrt{\pi(\theta)\pi(\theta')}, \tag{6.3}$$

which leads to the matrix form $\mathbf{K}_\theta = \mathbf{\Pi}\, \mathbf{R}_{\boldsymbol{\theta}}\, \mathbf{\Pi}$, where $\mathbf{\Pi} = \mathrm{diag}([\sqrt{\pi(\theta_1)}, \sqrt{\pi(\theta_2)}, \cdots])$.

Note that if $R(\theta, \theta') = 0$ when $\theta \neq \theta'$, this construction recovers the Poisson process with intensity measure $\pi(\theta)$. Note also in this case that if the cardinality of $\boldsymbol{\theta}$ is predetermined, then this recovers the traditional independent prior. More interesting, however, are $R(\theta, \theta')$ with off-diagonal structure that induces interaction within the set. Such kernels will always induce repulsion of the points so that diverse subsets of Θ will tend to have higher probability under the prior. See Fig. 6.1 for illustrations of the difference between independent samples and the DPP for several different settings.

6.3.2 Kernels for Probability Distributions

The determinantal point process framework allows us to construct a generative model for repulsion, but as with other kernel-based priors, we must define what "repulsion" means. A variety of positive definite functions on probability measures have been defined, but in this work we will use the *probability product* kernel (74). This kernel is a natural generalization of the inner product for probability distributions. The basic kernel has the form

$$K(\theta, \theta' \,;\, \rho) = \int_{\mathcal{X}} f(x \,|\, \theta)^{\rho} \, f(x \,|\, \theta')^{\rho} \, \mathrm{d}x \tag{6.4}$$

for $\rho > 0$. As we require a correlation kernel, we use the normalized variant given by

$$R(\theta, \theta' \,;\, \rho) = K(\theta, \theta' \,;\, \rho) / \sqrt{K(\theta, \theta \,;\, \rho) K(\theta', \theta' \,;\, \rho)}. \tag{6.5}$$

This kernel has convenient closed forms for several distributions of interest, which makes it an ideal building block for the present model.

6.3.3 Replicated Determinantal Point Process

A property that we often desire from our prior distributions is that they have the ability to become arbitrarily strong. That is, under the interpretation of a Bayesian prior as "inferences from previously-seen data", we would like to be able to imagine an arbitrary amount of such data and construct a highly-informative prior when appropriate. Unfortunately, the standard determinantal point process does not provide a knob to turn to increase its strength arbitrarily.

For example, take a DPP on a Euclidean space and consider a point t, an arbitrary unit vector w and a small scalar ϵ. Construct two pairs of points using a $\delta > 1$: a "near" pair $\{t, t + \epsilon w)\}$, and a "far" pair $\{t, t + \epsilon\delta w\}$. We wish to find some small ϵ such that the "far" configuration is arbitrarily more likely than the "near" configuration under the DPP. That is, we would like the ratio of determinants

$$r(\epsilon) = \frac{p(\{t, t+\epsilon\delta w\})}{p(\{t, t+\epsilon w)\})} = \frac{1 - R(t, t+\epsilon\delta w)^2}{1 - R(t, t+\epsilon w))^2}, \tag{6.6}$$

to be unbounded as ϵ approaches zero. The objective is to have a scaling parameter that can cause the determinantal prior to be arbitrarily strong relative to the likelihood terms. If we perform a Taylor expansion of the numerator and denominator around $\epsilon = 0$, we get

$$r(\epsilon) \approx \frac{1 - (R(t,t) + 2\delta w\epsilon \left[\frac{\mathrm{d}}{\mathrm{d}\tilde{t}} R(t,\tilde{t})\right]_{\tilde{t}=t})}{1 - (R(t,t) + 2w\epsilon \left[\frac{\mathrm{d}}{\mathrm{d}\tilde{t}} R(t,\tilde{t})\right]_{\tilde{t}=t})} = \delta. \tag{6.7}$$

We can see that, when near zero, this ratio captures the difference in distances, but not in a way that can be rescaled to greater effect. This means that there exist finite data sets that we cannot overwhelm with any DPP prior. To address this issue, we augment the determinantal point process with an additional parameter $\lambda > 0$, so that the probability of a finite subset $\boldsymbol{\theta} \subset \Theta$ becomes

$$p(\boldsymbol{\theta} \subset \Theta) \propto |\mathbf{K}_{\boldsymbol{\theta}}|^{\lambda}. \tag{6.8}$$

For integer λ, it can be viewed as a set of λ identical "replicated realizations" from determinantal point processes, leaving our generative view intact. The replicate of $\boldsymbol{\theta}$ is just $\boldsymbol{\theta}_{\lambda} = \{\lambda \text{ copies of } \boldsymbol{\theta}\}$ and the corresponding $\mathbf{K}_{\boldsymbol{\theta}_{\lambda}}$ is a $\lambda|\boldsymbol{\theta}| \times \lambda|\boldsymbol{\theta}|$ block diagonal matrix where each block is a replicate of $\mathbf{K}_{\boldsymbol{\theta}}$. This maps well onto the view of a prior as pseudo-data; our replicated DPP asserts that we have seen λ previous such data sets. As in other pseudo-count priors, we do not require in practice that λ be an integer, and under a penalized log likelihood view of MAP inference, it can be interpreted as a parameter for increasing the effect of the determinantal penalty.

6.3.4 Determinantal Point Process as Regularization.

In addition to acting as a prior over distributions in the generative setting, we can also view the DPP as a new type of "diversity" regularizer on learning. The goal is to solve

$$\boldsymbol{\theta}^{\star} = \operatorname*{argmin}_{\boldsymbol{\theta} \subset \Theta} \mathcal{L}(\boldsymbol{\theta}; \{x_n\}_{n=1}^{N}) - \lambda \ln |\mathbf{K}_{\boldsymbol{\theta}}|, \tag{6.9}$$

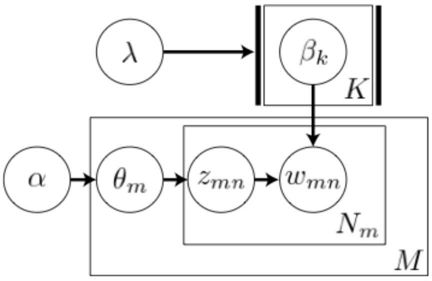


Figure 6.2: Schematic of DPP-LDA. We replace the standard plate notation for i.i.d topics in LDA with a "double-struck plate" to indicate a determinantal point process.

choosing the best *set* of parameters $\boldsymbol{\theta}$ from Θ. Here $\mathcal{L}(\cdot)$ is a loss function that depends on the data and the discrimination function, with parameters $\boldsymbol{\theta}$. From Eqn. (3),

$$\ln |\mathbf{K}_{\boldsymbol{\theta}}| = \ln |\mathbf{R}_{\boldsymbol{\theta}}| + \sum_{\theta_j \in \boldsymbol{\theta}} \ln \pi(\theta_j). \tag{6.10}$$

If $\mathcal{L}(\cdot) = -\ln p(\{x_n\}_{n=1}^{N} | \boldsymbol{\theta})$, then the resulting optimization is simply MAP estimation. In this framework, we can combine the DPP penalty with any other regularizer on θ, for example the sparsity-inducing ℓ_1 regularizer. In the following sections, we give empirical evidence that this diversity improves generalization performance.

6.4 MAP Inference

In what follows, we fix the cardinality $|\boldsymbol{\theta}|$. Viewing the kernel $\mathbf{K}_{\boldsymbol{\theta}}$ as a function of $\boldsymbol{\theta}$, the gradient $\frac{\partial}{\partial \theta} \log |\mathbf{K}_{\boldsymbol{\theta}}| = \text{trace}(\mathbf{K}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\theta}}}{\partial \theta})$. This allows application of general gradient-based optimization algorithms for inference. In particular, we can optimize $\boldsymbol{\theta}$ as a modular component within an off-the-shelf expectation maximization (EM) algorithm. Here we examine two illustrative examples of generative latent variable models into which we can directly plug our DPP-based prior.

Diversified Latent Dirichlet Allocation Latent Dirichlet allocation (LDA) (31) is an immensely popular admixture model for text and, increasingly, for other kinds of data that can be treated as a "bag of words". LDA constructs a set of topics — distributions over the vocabulary — and asserts that each word in the corpus is explained by one of these topics. The topic-word assignments are unobserved, but LDA attempts to find structure by requiring that only a small number of topics be represented in any given document.

In the standard LDA formulation, the topics are K discrete distributions β_k over a vocabulary of size V, where β_{kv} is the probability of topic k generating word v. There are M documents and the mth document has N_m words. Document m has a latent multinomial distribution over topics, denoted θ_m and each word in the document w_{mn} has a topic index z_{mn} drawn from θ_m. While classical LDA uses independent Dirichlet priors for the β_k, here we "diversify" latent Dirichlet allocation by replacing this prior with a DPP. That is, we introduce a correlation kernel

$$R(\beta_k, \beta_{k'}) = \frac{\sum_{v=1}^{V}(\beta_{kv}\, \beta_{k'v})^{\rho}}{\sqrt{\sum_{v=1}^{V} \beta_{kv}^{2\rho}}\sqrt{\sum_{v=1}^{V} \beta_{k'v}^{2\rho}}}, \tag{6.11}$$

which approaches one as β_k becomes more similar to $\beta_{k'}$. In the application below of DPP-LDA, we use $\rho = 0.5$. We use $\pi(\beta_k) = \text{Dirichlet}(\alpha)$, and write the resulting prior as $p(\boldsymbol{\beta}) \propto |\mathbf{K}_{\boldsymbol{\beta}}|$. We call this model "DPP-LDA", and illustrate it with a graphical model in Figure 6.2. We use a "double-struck plate" in the graphical model to represent the DPP, and highlight how it can be used as a drop-in replacement for the i.i.d. assumption.

To perform MAP learning of this model, we construct a modified version of the standard variational EM algorithm. As in variational EM for LDA, we define a factored approximation

$$q(\theta_m, z_m | \gamma_m, \phi_m) = q(\theta_m|\gamma_m) \prod_{n=1}^{N} q(z_{mn}|\phi_{mn}). \tag{6.12}$$

In this approximation, each document m has a Dirichlet approximation to its posterior over topics, given by γ_m. ϕ_m is an $N \times K$ matrix in which the nth row, denoted ϕ_{mn}, is a multinomial distribution over topics for word w_{mn}. For the current estimate of β_{kv}, γ_m and ϕ_m are iteratively optimized. See Blei *et al.* (31) for more details. Our extension of variational EM to include the DPP does not require alteration of these steps.

The inclusion of the determinantal point process prior does, however, effect the maximization step. The diversity prior introduces an additional penalty on $\boldsymbol{\beta}$, so that the M-step requires solving

$$\boldsymbol{\beta}^{\star} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{k=1}^{K} \sum_{v=1}^{V} \phi_{mnk} w_{mn}^{(v)} \ln \beta_{kv} + \lambda \ln |\mathbf{K}_{\boldsymbol{\beta}}| \right\}, \tag{6.13}$$

subject to the constraints that each row of $\boldsymbol{\beta}$ sum to 1. For $\lambda = 0$, this optimization procedure yields the standard update for vanilla LDA, $\beta^{\star}_{kv} \propto \sum_{m=1}^{M} \sum_{n=1}^{N_m} \phi_{mnk} w_{mn}^{(v)}$. For $\lambda > 0$ we use gradient descent to find a local optimal $\boldsymbol{\beta}$.

Diversified Gaussian Mixture Model The mixture model is a popular model for generative clustering and density estimation. Given J components, the probability of the data is given by

$$p(x_n \mid \boldsymbol{\theta}) = \sum_{j=1}^{J} \chi_j \, f(x_n \mid \theta_j). \tag{6.14}$$

Typically, the θ_k are taken to be independent in the prior. Here we examine determinantal point process priors for the θ_k in the case where the components are Gaussian.

For Gaussian mixture models, the DPP prior is particularly tractable. As in DPP-LDA, we use the probability product kernel, which in this case also has a convenient closed form. Let $f_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $f_2 = \mathcal{N}(\mu_2, \Sigma_2)$ be two Gaussians, the product kernel is:

$$K(f_1, f_2) = (2\pi)^{(1-2\rho)\frac{D}{2}} \rho^{-\frac{D}{2}} |\hat{\Sigma}|^{\frac{1}{2}} (|\Sigma_1||\Sigma_2|)^{-\frac{\rho}{2}} \exp(-\frac{\rho}{2}(\mu_1^T \Sigma_1^{-1} \mu_1 + \mu_2^T \Sigma_2^{-1} \mu_2 - \hat{\mu}^T \hat{\Sigma} \hat{\mu}))$$

where $\hat{\Sigma} = (\Sigma_1 + \Sigma_2)^{-1}$ and $\hat{\mu} = \Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2$. In the special case of a fixed, isotropic covariance $\sigma^2 I$ and $\rho = 1$, the kernel is

$$K(f(\cdot \mid \mu), f(\cdot \mid \mu')) = \frac{1}{(4\pi\sigma^2)^{D/2}} e^{-||\mu-\mu'||^2/(4\sigma^2)} \tag{6.15}$$

where D is the data dimensionality.

In the standard EM algorithm for Gaussian mixtures, one typically introduces latent binary variables z_{nj}, which indicate that datum n belongs to component j. The E-step computes the responsibility vector $\gamma(z_{nj}) = E[z_{nj}] \propto \chi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)$. This step is identical for DPP-GMM. The update for the component weights is also the same: $\chi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nj})$. The difference between this procedure and the standard EM approach is that the M-step for the DPP-GMM optimizes the objective function (summarizing $\{\mu_j, \Sigma_j\}_{j=1}^{J}$ by $\boldsymbol{\theta}$ for clarity):

$$\boldsymbol{\theta}^{\star} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \left\{ \sum_{n=1}^{N} \sum_{j=1}^{J} \gamma(z_{nj}) \left[\ln \chi_j + \ln \mathcal{N}(x_n | \mu_j, \Sigma_j) \right] + \lambda \ln |\mathbf{K}_{\boldsymbol{\theta}}| \right\}. \tag{6.16}$$

Table 6.1: Top ten words from representative topics learned in LDA and DPP-LDA.

LDA		DPP-LDA					
typical		”stop words”		”Christianity”	”space”	”OS”	”politics”
the	the	the	and	jesus	space	file	ms
to	to	of	in	matthew	nasa	pub	myers
and	and	that	at	prophecy	astronaut	usr	god
in	it	you	from	christians	mm	available	president
of	of	by	some	church	mission	export	but
is	is	one	their	messiah	pilot	font	package
it	in	all	with	psalm	shuttle	lib	options
for	that	but	your	isaiah	military	directory	dee
that	for	do	who	prophet	candidates	format	believe
can	you	my	which	lord	ww	server	groups

Closely related to DPP-GMM is *DPP-K-means.* The kernel acts on the set of centroids as in Eqn. (15), with σ^2 now just a constant scaling term. Let $\boldsymbol{\theta} = \{\mu_j\}$ and z_{nj} be the hard assignment indicator, the maximization step is:

$$\boldsymbol{\theta}^\star = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \left\{ \sum_{n=1}^{N} \sum_{j=1}^{J} z_{nj} ||x_n - \mu_j||^2 + \lambda \ln |\mathbf{K}_{\boldsymbol{\theta}}| \right\}. \tag{6.17}$$

With the product kernel, the similarity between two Gaussians decays exponentially as the distance between their means increases. In practice, we find that when the number of mixture components $|\boldsymbol{\theta}|$ is large, $\mathbf{K}_{\boldsymbol{\theta}}$ is well approximated by a sparse matrix.

6.5 Experiment I: Diversified topic modeling.

We tested LDA and DPP-LDA on the unfiltered 20 Newsgroup corpus, without removing any stop-words. A common frustration with vanilla LDA is that applying LDA to unfiltered data returns topics that are dominated by stop-words. This frustrating phenomenon occurs even as the number of topics is varied from $K = 5$ to $K = 50$. The first two columns of Table 6.1 show the ten most frequent words from two representative topics learned by LDA using $K = 25$. Stop-words occur frequently across all documents and thus are unhelpfully correlated with topic-specific informative keywords. We repeated the experiments after removing a list of 669 most common stop-words. How-

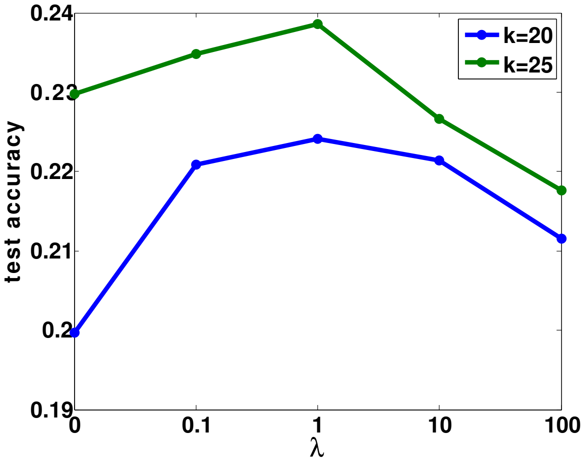


Figure 6.3: Effect of λ on classification error.

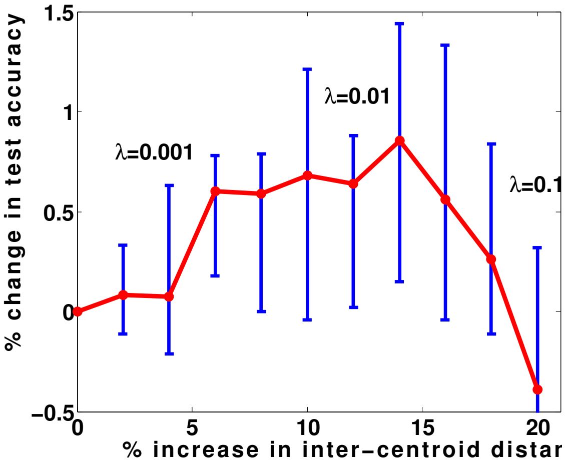


Figure 6.4: Effect of centroid distance on test error.

ever, the topics inferred by regular LDA are still dominated by secondary stop-words that are not informative.

DPP-LDA automatically groups common stop words into a few topics. By finding stop-word-specific topics, the majority of the remaining topics are available for more informative words. Table 6.1 shows a sample of topics learned by DPP-LDA on the unfiltered 20 Newsgroup corpus ($K = 25$, $\lambda = 10^4$). As we vary K or increase λ we observe robust grouping of stop-words into a few topics. High frequency words that are common across many topics significantly increase the similarity between the topics, as measured by the product kernel on the $\boldsymbol{\beta}$ distributions. This similarity incurs a large penalty in DPP and so the objective actively pushes the parameters of LDA away from regions where stop words occupy large probability mass across many topics.

Features learned from DPP-LDA leads to better document classification. It is common to use the γ_m, the document specific posterior distribution over topics, as feature vectors in document classification. We inferred $\{\gamma_{m,train}\}$ on training documents from DPP-LDA variational EM, and then trained a support vector machine (SVM) classifier on $\{\gamma_{m,train}\}$ with the true topic labels from 20 Newsgroups. On test documents, we fixed the parameters α and $\boldsymbol{\beta}$ to the values inferred from the training set, and used variational EM to find MAP estimates of $\{\gamma_{m,test}\}$. The mean test classification accuracy for a range of λ values is plotted in Figure 6.3. The setting $\lambda = 0$ corresponds to vanilla LDA. In each trial, we use the same training set for DPP-LDA on a range of λ values. DPP-LDA with $\lambda = 1$ consistently outperforms LDA in test classification ($p < 0.001$ binomial test). Large values of λ decrease classification performance.

6.6 Experiment II: Diverse clustering.

Mixture models are often a useful way to learn features for classification. The recent work of Coates *et al.* (16), for example, shows that even simple K-means works well as a method of extracting features for image labeling. In that work, K-means gave state of art results on the CIFAR-10 object recognition task. Coates *et al.* achieved these results using a patch-wise procedure in which random patches are sampled from images for training. Each patch is a 6-by-6 square, represented as a point in a 36 dimensional vector space. Patches from the training images are combined and clustered using K-means. Each patch is then represented by a binary K-dimensional feature vector: the k^{th} entry is one if the patch is closer to the centroid k than its average distance to centroids. Roughly half of the feature entries are zero. Patches from the same image are then pooled to construct one feature vector for the whole image. An SVM is trained on these image features to perform classification.

We reason that DPP-K-means may produce more informative features since the cluster centroids will repel each other into more distinct positions in pixel space. We replicated the experiments from Coates *et al.*, using their publicly-available code for identical pre- and post-processing. With this setup, $\lambda = 0$ recovers regular K-means, and reproduces the results from Coates *et al.* (16). We applied DPP-K-means to

Table 6.2: Test classification accuracy on CIFAR-10 dataset.

training set size	K	K-means	DPP K-means	gain (%)	λ
500	30	34.81	**36.21**	1.4	0.01
1000	30	43.32	**44.27**	0.95	0.01
2000	60	52.05	**52.55**	0.50	0.01
5000	150	61.03	**61.23**	0.20	0.001
10000	300	66.36	**66.65**	0.29	0.001

the CIFAR-10 dataset, while varying the size of the training set. For each training set size, we ran regular K-means for a range of values of K and select the K that gives the best test accuracy for K-means. Then we compare the performance with DPP-K-means using the same K. For up to 10000 images in the training set, DPP-K-means leads to better test classification accuracy compared to the simple K-means. The comparisons are performed on matched settings: for a given randomly sampled training set and a centroid initialization, we generate the centroids from both K-means and DPP-K-means. The two sets of centroids were used to extract features and train classifiers, which are then tested on the same test set of images. DPP-K-means consistently outperforms K-means in generalization accuracy ($p < 0.001$ binomial test). For example, for training set of size 1000, with $k = 30$, we ran 100 trials, each with an random training set and initialization, DPP-K-means outperformed K-means in 94 trials. As expected given its role as a regularizer, improvement from DPP-K-means is more significant for smaller training sets. For the full CIFAR-10 with 50000 training images, DPP-K-means does not consistently outperform K-means.

Next we ask if there is a pattern between how far the DPP pushes apart the centroids and classification accuracy on the test set. Focusing on 1000 training images and $k = 30$, for each randomly sampled training set and centroid initialization, we compute the mean inter-centroid distance for K-means and DPP-K-means. We compute the test accuracy for each set of centroids. Fig. 4 bins the relative increase in inter-centroid distance into 10 bins. For each bin, we show the 25^{th}, 50^{th}, and 75^{th} percentile of changes in test accuracy. Test accuracy is maximized when the inter-centroid distances increase by about 14% from K-means centroids, corresponding to $\lambda = 0.01$.

6.7 Discussion.

We have introduced a general approach to including a preference for diversity into generative probabilistic models. We showed how a determinantal point process can be integrated as a modular component into existing learning algorithms, and discussed its general role as a diversity regularizer. We investigated two settings where diversity can be useful: learning topics from documents, and clustering image patches. Plugging a DPP into latent Dirichlet allocation allows LDA to automatically group stop-words into a few categories, enabling more informative topics in other categories. In both document and image classification tasks, there exists an intermediate regime of diversity (as controlled by the hyperparameter λ) that leads to consistent improvement in accuracy when compared to standard i.i.d. models. A computational bottleneck can come from inverting the $M \times M$ kernel matrix $\mathbf{K}$, where M is the number of latent distributions. However in many settings such as LDA, M is much smaller than the data size. We expect that there are many other settings where DPP-based diversity can be usefully introduced into a generative probabilistic model: in the emission parameters of HMM and more general time series, and as a mechanism for transfer learning.

7

The Physarum Solver for Linear Programming Problems

7.1 Overview

Physarum polycephalum (true slime mold) has recently emerged as a fascinating example of biological computation through morphogenesis. Despite being a single cell organism, experiments have observed that through its growth process, the Physarum is able to solve various minimum cost flow problems. This paper analyzes a mathematical model of the Physarum growth dynamics. We show how to encode general linear programming (LP) problems as instances of the Physarum. We prove that under the growth dynamics, the Physarum is guaranteed to converge to the optimal solution of the LP. We further derive an efficient discrete algorithm based on the Physarum model, and experimentally verify its performance on assignment problems.

7.2 Introduction

How do biological systems process information and solve optimization problems? There has been a growing interest to understand these questions within a computation framework. The agenda is two fold: first, can we understand and analyze biological systems in terms of algorithms; and second, can we design new algorithms inspired by biology. Prominent examples of such natural algorithms include flocking algorithms motivated by the collective behavior of birds and fish, and swarming algorithms motived by ant foraging behavior (12, 54, 87).

In this paper, we analyze the morphogenesis of Physarum polycephalum (true slime mold) as a natural algorithm. The Physarum is a single cell organism with multiple nuclei. Recent experiments have shown that it exhibits a surprising ability to solve complex optimization problems (23, 58). In one set of experiments, researchers place food at various places akin to cities on a map (58). As it grows, the Physarum is able to process the input (food) in a de-centralized manner, and converge to the distance minimizing network connecting these food sources1. Mathematicians and biologists have proposed a dynamical systems model that captures essential behaviors of Physarum growth. Simulation and analysis of this model give mechanistic insight of how such a simple organism can solve the Shortest Path Problem (47, 57, 61).

Here we show how a general Linear Programming Problem can be encoded as an instance of a Physarum. We prove that under the Physarum growth model, the appropriate quantity is guaranteed to converge to the optimal solution of the LP. Our result draws on new analysis techniques inspired from the negative cost cycle algorithm. In addition, we derive a discrete algorithm for solving LPs from the dynamical systems. The algorithm is efficient to implement. We apply it to solve instances of the Assignment Problem, and report simulation results.

7.3 Biology of Physarum

Physarum polycephalum is a member of the superclass Myxogastridae commonly referred to as true slime mold (63). In vegetative stage, it is a single cell containing multiple nuclei that divide synchronously. It engulfs bacteria, amoebae and other microbes, and secretes enzymes to digest the engulfed material. Under adverse conditions, it can reversibly transform into a sclerotium, a hardened mass that can survive for long periods. Physarum grows as a complex of interconnected tubes made of a gel-like outer layer enclosing the cytoplasmic fluid. The fluid oscillates back and forth across the tubes with a period of about 100 seconds (46). This flow is shown to be driven by cross-sectional contractions of the tubes through a peristaltic mechanism (46).

7.4 The generalized Physarum Solver

A model of Physarum growth. A Physarum contains a network of veins that it uses to transport nutrients across its body. The veins are modeled as a graph $\mathbf{G}$ (59). The vertices $\mathbf{V}$ are where multiple veins meet, and an edge $e \in \mathbf{E}$ is a segment of the vein. Each edge is a tube described by length c_e and cross-sectional area σ_e. Vertex i is associated with pressure p_i. We assume that $\mathbf{G}$ is directed. So edge ij goes from vertex i to j. The flow on ij is $x_{ij} = \sigma_{ij}\frac{p_i - p_j}{c_{ij}}$. $\mathbf{G}$ may contain multi-edges. In particular, a bidirectional edge can be modeled by two edges: ij and ji, with distinct σ_{ij} and σ_{ji}.

The Physarum growth model describes the time evolution of σ_{ij} and p_i on $\mathbf{G}$. The expansion and contraction of a vein segment is governed by:

$$\frac{d}{dt}\sigma_{ij}(t) = \sigma_{ij}(t)\frac{p_i - p_j}{c_{ij}} - \sigma_{ij}(t). \tag{7.1}$$

The first term on the left captures the positive feedback: the greater the flow on ij, the more the vein segment will expand. If $p_i < p_j$ implying that $x_{ij} < 0$, then σ_{ij} shrinks. The second term reflect the natural decay of the Physarum.

The pressure p_i is obtained from flow conservation: the flow into a vertex (except for source or sink) equals the flow out from it. Let $\mathbf{A}$ be the $|\mathbf{V}|$x$|\mathbf{E}|$ incidence matrix of $\mathbf{G}$. The conservation equation is $\mathbf{Ax(t)} = \mathbf{b}$, where $b_i = 0$ for all i except for $b_s = 1$ and $b_u = -1$. Vertices s and u are the source and sink. The conservation equation can be rewriten as $\mathbf{AW(t)A^T p(t)} = \mathbf{b}$, where $\mathbf{W}(t) = \text{diag}(\sigma_e(t)/c_e)$. If $\mathbf{A}$ has full row rank, then $\mathbf{p(t)}$ can be uniquely solved for.

The Physarum growth dynamic thus has two coupled processes:

1. The vein sizes $\sigma_{ij}(t)$ grow according to Eqn. 1, which depends on $\mathbf{p(t)}$.

2. For a given set of $\sigma_{ij}(t)$, flow conservation uniquely determines $\mathbf{p(t)}$.

Under this growth model, $\{x_{ij}(t)\}$ evolves on $\mathbf{G}$. As $t \to \infty$, the flow $\mathbf{x(t)}$ converges to 1 on edges in the shortest s-u path, and to 0 on all other edges (47).

LP problems. A general linear programming (LP) problem is to $\min \mathbf{c^T x}$ subject to the contraints $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{x} \geq 0$. Define $\mathbf{\Phi} = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}\}$ and $\mathbf{\Phi}^+ = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$. We work in cases where $\mathbf{\Phi}^+$ forms a bounded polytope and A has full row rank, which are common assumptions in LP applications. We assume that the LP has a unique optimal solution. Any small perturbation to $\mathbf{c}$ leads to unique optimum.

Keeping the graph notations, we label columns of $\mathbf{A}$ by $e \in \mathbf{E}$. Each column $\mathbf{A_e}$ is analogous to an edge and is associated with a "conductance "$\sigma_e \geq 0$, which is an auxiliary variable. Every constraint equation in $\mathbf{Ax} = \mathbf{b}$ corresponds to a vertex and is associated with a pressure p_i. We define the generalized *Physarum Solver* by the dynamics

$$\frac{d}{dt}\sigma_e = \sigma_e(\psi_e - 1) \tag{7.2}$$

where $\psi_e = (\mathbf{A_e^T p})/\mathbf{c_e}$ is the "pressure gradient"and $\mathbf{p}$ satisfies Kirchhoff's Law

$$\mathbf{AWA^T p} = \mathbf{b} \text{ with } \mathbf{W} = \operatorname{diag}(\sigma_e/c_e). \tag{7.3}$$

As before, the "flow "is given by $x_e(t) = \sigma_e(t)\psi_e(t)$. Note that $\mathbf{Ax} = \mathbf{AWA^T p} = \mathbf{b}$, though $x_e(t)$ could be negative if $\psi_e < 0$. So $\mathbf{x(t)} \in \mathbf{\Phi}$.

Duality. The pressure, p_i, is the dual variable associated with each constraint. If the dynamical system reaches a stationary point, then for all e, $0 = \frac{d}{dt}\sigma_e = \sigma_e(\frac{\mathbf{A_e^T P}}{c_e} - 1)$. If $x_e > 0$, then $\sigma_e > 0$ and stationarity implies $\mathbf{A_e^T P} = c_e$. This is precisely the complementary slackness condition. The stationary point of the Physarum Solver is the optimal solution to the LP.

The main result of the paper is the following theorem.

Theorem 1. *Suppose the LP problem has a unique optimal solution. Then from any positive inital configuration $\sigma(\mathbf{0}) > \mathbf{0}$, the flow of the Physarum Solver $\mathbf{x(t)}$ converges exponentially fast to the optimal solution.*

7.5 Proof of convergence

7.5.1 LP with positive costs.

We first assume that $c_e > 0$ $\forall e$. In the next section we will show how to extend the proof to general $\mathbf{c}$. Here we give the outline of the proof; the technical details are below.

We multiply Eqn. 2 by e^t and integrate to obtain

$$\sigma(t) = (1 - e^{-t})\tilde{\mathbf{x}}(t) + e^{-t}\sigma(0) \tag{7.4}$$

with

$$\tilde{\mathbf{x}}(t) \doteq \frac{1}{1-e^{-t}} \int_0^t \mathbf{x}(\mathbf{s}) e^{-(t-s)} ds. \tag{7.5}$$

Since $\tilde{\mathbf{x}}(t)$ is a weighted average of $\mathbf{x}(s)$ for $s \leq t$ and each $\mathbf{x}(s)$ satisfies $\mathbf{A}\mathbf{x}(s) = \mathbf{b}$, we also have $\mathbf{A}\tilde{\mathbf{x}}(t) = \mathbf{b}$.

Definition 1. *A circulation is a $|\mathbf{E}|$ dimensional vector γ such that $\mathbf{A}\gamma = \mathbf{0}$.*

There must exists e such that $\gamma_e < 0$, for otherwise the feasible region $\mathbf{\Phi}^+$ would be unbounded. Let $\gamma_- \doteq \{e : \gamma_e < 0\}$ denote the negative edges of γ. Similarly define $\gamma_+ \doteq \{e : \gamma_e > 0\}$. We say a circulation has negative cost if $\mathbf{c}^{\mathbf{T}}\gamma < \mathbf{0}$.

Lemma 1. *In a bounded LP problem, $\mathbf{x}$ is an optimal solution if and only if there is no negative cost circulation γ such that $\gamma_- \subseteq supp(\mathbf{x})$.*

Our strategy is to prove that under Eqn. 2, the flow on γ_- becomes exponentially small for any negative circulation γ. We start with the following Lemma.

Lemma 2. *Let γ be a negative cost circulation. Then for all $\epsilon > 0$, $\exists$ e with $\gamma_e < 0$ and t such that $\sigma_e(t) < \epsilon$.*

Proof. Let $Z_\gamma \doteq e^{-\sum_{e\in\mathbf{E}} \gamma_e c_e log\sigma_e}$. Since $\sum_{e\in\mathbf{E}} \gamma_e c_e \psi_e = \mathbf{p}^{\mathbf{T}}\mathbf{A}\gamma = \mathbf{0}$,

$$\frac{d}{dt} Z_\gamma = Z_\gamma \sum_{e\in\mathbf{E}} \gamma_e c_e (1-\psi_e) = Z_\gamma \mathbf{c}^{\mathbf{T}}\gamma < \mathbf{0}. \tag{7.6}$$

Therefore Z_γ decays exponentially. On the other hand,

$$Z_\gamma(t) = \frac{\Pi_{e\in\gamma_-} \sigma_e(t)^{|\gamma_e|c_e}}{\Pi_{e\in\gamma_+} \sigma_e(t)^{\gamma_e c_e}}, \tag{7.7}$$

and by Lemma 4 (Appendix), all σ_e are bounded above for large t and hence the denominator is bounded above as t increases. As the ratio decays exponentially, the numerator must decay exponentially and at least one of the terms in its product must be exponentially small. □ □

The next lemma shows that we can clean up a solution $\mathbf{x}$ of $\mathbf{A}\mathbf{x} = \mathbf{b}$ by essentially setting x_e to 0 if x_e is sufficiently small.

Lemma 3. *Given $\mathbf{x} \in \mathbf{\Phi}$, and $\mathbf{S} \subseteq \mathbf{E}$. There are constants M_1 and K such that if $\forall e \in \mathbf{S}$, $|x_e| < \epsilon^{|\mathbf{S}|+\mathbf{1}} < M_1/2m$ for some $\epsilon < min\{1, \frac{M_1}{2}, \frac{1}{2K}\}$, then there exists $\mathbf{y} \in \mathbf{\Phi}$ satisfying:*

- $||\mathbf{x} - \mathbf{y}||_\infty < K\epsilon$,
- *supp*($\mathbf{y}$) $\subseteq$ *supp*$(\mathbf{x})\backslash\mathbf{S}$,
- $x_e y_e \geq 0\ \forall e$.

Combining the previous two lemmas, we prove that $\mathbf{x}(t)$ converges to the unique optimal solution under the Physarum dynamics.

Proof of Theorem 1 for LP with positive costs.
Consider the time-averaged solution $\tilde{\mathbf{x}}(t)$ defined above. Let $\mathbf{S} = \{e : \tilde{x}_e(t) < 0\}$. From Eqn. 3 and the fact that $\sigma(t)$ is bounded for large t, it follows that

$$\tilde{\mathbf{x}}(t) = \sigma(t) + O(e^{-k_1 t}) \tag{7.8}$$

for some constant k_1. Since $\sigma(t) > 0$, we have $|\tilde{x}_e(t)| \sim O(e^{-k_1 t})$, $\forall e \in \mathbf{S}$. For sufficiently large t so that the inequality requirements of Lemma 3 are satifistied, we obtain a $\mathbf{y}(t)$ such that

$$\mathbf{y}(t) = \tilde{\mathbf{x}}(t) + O(e^{-k_2 t}) \tag{7.9}$$

for constant k_2. Moreover $\mathbf{y}(t) \geq 0$ is feasible.

Let $\hat{\mathbf{y}}$ denote the optimal feasible solution of the LP. Let $e \in \text{supp}(\mathbf{y})/\text{supp}(\hat{\mathbf{y}})$. Then by Lemma 6 (Appendix), there is a basic feasible solution τ such that

$$y_e \leq k_3 \min\{y_{e'} : e' \in \text{supp}(\tau)\} \tag{7.10}$$

for constant k_3. Consider $\gamma = \hat{\mathbf{y}} - \tau$. This is a circulation since $\mathbf{A}\gamma = \mathbf{0}$, and it has negative cost. By Lemma 2, $\exists\, e' \in \text{supp}(\tau)$ such that $\sigma_{e'}(t)$ decays exponentially. Eqns. 8 and 9 implies that $y_{e'}(t) \sim O(e^{-k_4 t})$ for some constant k_4. Combined with Eqn. 10, this implies that $y_e \sim O(e^{-k_4 t})$ for all $e \in \text{supp}(\mathbf{y})/\text{supp}(\hat{\mathbf{y}})$. Applying Lemma 2 again to remove all such e, we find $\tilde{\mathbf{y}} \in \mathbf{\Phi}^+$ such that

$$\tilde{\mathbf{y}} = \mathbf{y}(\mathbf{t}) + O(e^{-k_5 t}) = \tilde{\mathbf{x}}(t) + O(e^{-k_2 t}) + O(e^{-k_5 t}). \tag{7.11}$$

But $\tilde{\mathbf{y}}$ has support contained in supp($\hat{\mathbf{y}}$), implies $\tilde{\mathbf{y}} = \hat{\mathbf{y}}$ since the optimal solution is a unique basic feasible solution. This concludes that $\tilde{\mathbf{x}}(t) = \hat{\mathbf{y}}(t) + O(e^{-k_6 t})$.

As $\mathbf{x}(\mathbf{t})$ is continuous, the weighted average $\tilde{\mathbf{x}}(t)$ converges to $\hat{\mathbf{y}}$ exponentially implies $\mathbf{x}(t)$ also converges to $\hat{\mathbf{y}}$ exponentially. □

7.5.2 LP with general costs.

First we show how to deal with when $c_e = 0$ for some e. The assumption that $\mathbf{\Phi}^+$ is a bounded polytope implies $\exists\ M$ such that $x_e < M$ for all $e \in \mathbf{E}$ and $\mathbf{x} \in \mathbf{\Phi}^+$. Given $\{\mathbf{A}, \mathbf{b}, \mathbf{c}\}$, consider the LP $\{\mathbf{A}, \mathbf{b}, \hat{\mathbf{c}}\}$, where $\hat{c}_e = c_e$ if $c_e > 0$ and $\hat{c}_e = \epsilon$ if $c_e = 0$. This problem can be solved by the Physarum Solver since $\hat{\mathbf{c}} > \mathbf{0}$. Let $\mathbf{x_1}$ be the optimal solution of $\{\mathbf{A}, \mathbf{b}, \mathbf{c}\}$ and let $\mathbf{x_2}$ be its second lowest cost solution. Define the gap $\delta = \mathbf{c^T x_2} - \mathbf{c^T x_1}$ and set $\epsilon < \frac{\delta}{|E|M}$. It's easy to see that $\mathbf{x_1}$ is also the optimal solution of $\{\mathbf{A}, \mathbf{b}, \hat{\mathbf{c}}\}$. Hence it suffices to apply the Physarum Solver to solve $\{\mathbf{A}, \mathbf{b}, \hat{\mathbf{c}}\}$.

Now consider if $c_e < 0$ for some $e \in \mathbf{U} \subseteq \mathbf{E}$. For each such e add a new variable $\hat{e}$ and the constraint equation $x_e + x_{\hat{e}} = M$. Modify the costs by setting $c'_e = 0$ and $c'_{\hat{e}} = |c_e|$ for $e \in \mathbf{U}$. The new LP problem $\{\mathbf{A}', \mathbf{b}', \mathbf{c}'\}$ can by solved by the Physarum Solver. There is a one-to-one mapping between the feasible sets of $\{\mathbf{A}, \mathbf{b}, \mathbf{c}\}$ and $\{\mathbf{A}', \mathbf{b}', \mathbf{c}'\}$ that maps the optimal solutions to each other.

7.5.3 Relations to other LP solvers

The Physarum dynamics does not maintain primal feasibility. Even though $\mathbf{Ax(t)} = \mathbf{b}$ always holds, the gradient $\frac{A_e^T p}{c_e}$ could be negative, resulting in $x_e(t) < 0$ for some t. Though our proof shows that $\mathbf{x}(t)$ stays exponentially close to the feasible polytope $\mathbf{\Phi}^+$. Similarly, the potentials may violate dual feasibility: $\mathbf{A_e^T p}(t) > c_e$ for some e.

In the Physarum solver, $\mathbf{x}(t)$ does not live on the boundary of $\mathbf{\Phi}^+$ and in this sense is more related to interior point methods than to the Simplex Algorithm. Unlike interior point methods, the Physarum is not guaranteed to monotonically reduce the cost $\mathbf{c^T x}(t)$ or increase the dual objective $\mathbf{p^T b}$, and is not restricted to be inside the polytope (8).

The Physarum is conceptually most similar to the negative cost cycle algorithm. In this algorithm, a negative cost cycle is identified and a flow is pushed on this cycle until one of the directed edges has flow 0 (8). As our proof shows, this is also the Physarum's strategy. Think of the graph case for intuition. If there is a negative cost oriented cycle γ, such that all edges ij of γ are unsaturated: $p_i - p_j < c_{ij}$. Then σ_{ij} increase for all $ij \in \gamma$, and an additional negative cost flow is pushed $\{x_{ij}(t + \triangle) - x_{ij}(t) : ij \in \gamma\}$.

7.6 The discrete Physarum Solver

7.6.1 Finite difference approximation

The continuous dynamics of the generalized Physarum Solver makes it difficult to be implemented and analyzed as an algorithm. We use the finite difference method to discretize Eqn. 1

$$\sigma_e(t+\triangle) \approx \sigma_e(t) + \left(\frac{\mathbf{A_e^T p}}{c_e} - 1\right)\sigma_e(t)\triangle. \tag{7.12}$$

By setting $\triangle = 1$, we obtain a simple discrete algorithm.

Algorithm 4 Discrete Physarum Solver

```
INPUT: LP problem instance $(\mathbf{A}, \mathbf{b}, \mathbf{c})$
Initialize $\{\sigma_e(0) > 0, e \in \mathbf{E}\}$
while $\sigma(t)$ not converged do
    Compute $\mathbf{p}$(t) by solving $\mathbf{AWA^T p} = \mathbf{b}$, where $\mathbf{W} = \mathrm{diag}(\frac{\sigma_e(t)}{c_e})$.
    if $\mathbf{A^T p}(t) > 0$ then
        $\sigma_e(t+1) = \sigma_e(t)\frac{\mathbf{A_e^T p}(t)}{c_e} \doteq x(t)$
    else
        $\sigma_e(t) = \epsilon$
    end if
    $t = t + 1$
end while
```

In the algorithm, $\epsilon \ll 1$ is a positive constant. We say σ has converged if $\sigma(t+1) - \sigma(t)$ is small. We check for whether the algorithm has converged to the optimal solution by testing for approximate complementary slackness: $|\frac{A_e^T p}{c_e} - 1| < \delta, \forall e$ for some small δ. The most computationally intensive step is to solve Kirchhoff's Equation for $\mathbf{p}(t)$. Recent progress shows this can be done in near-linear time in $|\mathbf{E}|$. Because of the large finite difference approximation we have taken, the candidate solution $\mathbf{x}(t)$ is not guaranteed to converge to the optimal. We show via simulation that the Discrete Physarum Solver does find the optimum for the special class of Assignment problems.

7.6.2 The Assignment Problem

We test the discrete Physarum algorithm on a standard class of LP problems: the Assignment Problem (8). In the Assignment Problem, there are N workers and N

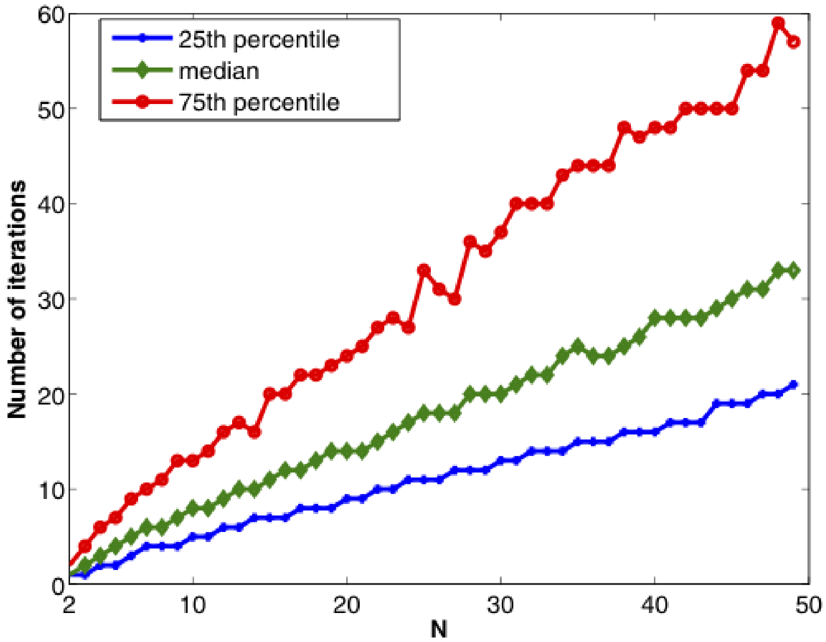


Figure 7.1: Number of iterations to find the optimal solution versus the size of the Assignment Problem. N is the number of tasks. For each N, plotted are the 25^{th}, median, and 75^{th} percentile of the number of iterations.

tasks. Every worker and task pair has an associated cost. The goal is to assign each worker to a unique task such that the total cost incurred is minimal.

This can be represented by a complete bipartite graph **G** with worker nodes and task nodes. Edges are directed from worker nodes to task nodes. Edge ij connecting worker i and task j is associated a cost c_{ij}. **A** is the incidence matrix, and $b_i = 1$ if i is a worker vertex; $b_j = -1$ for task vertex. The update is $\sigma_{ij}(t+1) = \sigma_{ij}(t)\frac{p_i - p_j}{c_{ij}} = x_{ij}(t)$ if $p_i > p_j$; else $\sigma_{ij}(t+1) = \epsilon$. To check for optimality, we derive a hard assignment from $\mathbf{x}(t)$. For each i, identify $j_i = \arg\max_j\{x_{ij}(t)\}$. Set $x_{ij_i} = 1$ and to 0 on all other edges. Then we test for complementary slackness.

For each N from 2 to 50, we randomly generate 1000 weighted complete bipartite graphs. The costs are iid samples from the uniform distribution. We applied the Discrete Physarum Algorithm to each of these assignment problems. After every iteration of the algorithm, we check for optimality. If the optimal solution is found, the number of iterations so far is recorded; otherwise the algorithm continues.

In all the assignment problem instances tested, the Discrete Physarum Solver always converged to the optimal solution. For each N, we plot the 25^{th}, median, and 75^{th}

percentile of the number of iterations the Discrete Physarum Solver took before reaching the optimal solution. The median number of iterations grows roughly linear in the size of the problem.

7.7 Proofs of Technical Lemmas

Proof of Lemma 1.

Suppose there exists a negative cost circulation γ. Let $\epsilon = max_e\{\frac{x_e}{|\gamma_e|}\}$. Then $\mathbf{y} = \mathbf{x}+\epsilon\gamma$ satisfies $\mathbf{y} \geq 0$, $\mathbf{Ay} = \mathbf{b}$, and $\mathbf{y}$ has lower cost than $\mathbf{x}$. This implies $\mathbf{x}$ is not an optimal solution. Conversely, suppose $\mathbf{x}$ is not optimal. Let $\mathbf{y}$ be the optimal solution and $\gamma = \mathbf{y} - \mathbf{x}$. Then γ is a negative cost circulation. □

Lemma 4. *For all e, $\sigma_e(t)$ is bounded for large t.*

Proof. Rearranging Eqn. 3, $\tilde{\mathbf{x}}(t) > -e^{-t}\sigma(0)$. Let $\mathbf{S}(t) = \{e : \tilde{x}_e(t) < 0\}$. Then for t sufficiently large, the condition for Lemma 3 is satisfied and there exists $\mathbf{y}(t)$ such that $\mathbf{y}(t) \geq 0$, $\mathbf{Ay}(\mathbf{t}) = \mathbf{b}$, and $||\tilde{\mathbf{x}}(t) - \mathbf{y}(t)||_\infty < Ce^{-t}$ for some constant C. The vector $\mathbf{y}(t)$ lies in the feasible region of the LP, which we assume to be a bounded polytopy, implies $||\mathbf{y}(t)||_\infty$ is bounded. This implies that $||\tilde{\mathbf{x}}(t)||$ is bounded and hence $\sigma(t)$ is bounded. □ □

Definition 2. *Given $\mathbf{A}$ and $\mathbf{x}$, a* ***positive re-orientation*** *of $\mathbf{A}$ with respect to $\mathbf{x}$ is $\mathbf{A}'$ and $\mathbf{x}'$ such that*

- $x'_e = |x_e|$ *for all e*
- $\mathbf{A'_e} = \mathbf{A_e}$ *for e where $x_e \geq 0$*
- $\mathbf{A'_e} = -\mathbf{A_e}$ *for e where $x_e < 0$*

Definition 3. *A* ***positive cycle*** *is a vector γ indexed by $e \in \mathbf{E}$ such that $\mathbf{A}\gamma = \mathbf{0}$, $\gamma_e \geq 0, \forall e$, and supp($\gamma$) is minimal. Since γ is ambigous up to a constant, we set $\min_{e \in supp(\gamma)}\{\gamma_e\} = 1$. The set of positive cycles of A is denoted by pcyc(A).*

Let bfs(A) denote the set of basic feasible solutions of the LP $\mathbf{Ax} = \mathbf{b}$, $\mathbf{x} \geq 0$.

Lemma 5. *Given $\mathbf{A}$, $\mathbf{x} \in \mathbf{\Phi}$ and $e \in \mathbf{E}$ such that $|x_e| < \epsilon < M_1/2$. Then $\exists$ $\mathbf{y} \in \mathbf{\Phi}$ such that:*

- $||\mathbf{x} - \mathbf{y}||_\infty < 3\epsilon\frac{M_2}{M_1}$,

- *supp*$(\mathbf{y})$ $\subseteq$ *supp*$(\mathbf{x})\backslash e$,
- $x_e y_e \geq 0$ *for all* e,

Here

$$M_1 = \min_{A'} \min_{\tau \in bfs(A') \cup pcyc(A')} \{\tau_e : e \in \mathbf{E} \text{ and } \tau_e > 0\} \tag{7.13}$$

$$M_2 = \max_{A'} \max_{\tau \in bfs(A') \cup pcyc(A')} \{\tau_e : e \in \mathbf{E}\} \tag{7.14}$$

where min and max are over all re-orientation A' *of* A.

Proof. After re-orientation, we can assume wlog that $\mathbf{x} \geq 0$. Then $\mathbf{x}$ lies in the polyhedron defined by $\mathbf{A'x} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$. We can express $\mathbf{x}$ as a convex linear combination of basic feasible solutions and positive cycles (8): $\mathbf{x} = \sum_{\tau \in \mathrm{bfs}(A') \cup \mathrm{pcyc}(A')} c_\tau \tau$, such that $c_\tau > 0$ and $\sum_{\tau \in \mathrm{bfs}(A')} c_\tau = 1$. We can split the sum into two groups of τ depending on if $\tau_e = 0$ or $\tau_e > 0$:

$$\mathbf{x} = \sum_{\tau : \tau_e = 0} c_\tau \tau + \sum_{\tau : \tau_e > 0} c_\tau \tau. \tag{7.15}$$

Let $\alpha = \sum_{\tau : \tau_e > 0} c_\tau$. Then

$$\alpha M_1 \leq \sum_{\tau : \tau_e > 0} c_\tau \tau_e = x_e < \epsilon < \frac{M_1}{2} \tag{7.16}$$

implying $\alpha < \frac{\epsilon}{M_1} < \frac{1}{2}$. Therefore there is at least one bfs τ such that $c_\tau > 0$ and $\tau_e = 0$. Define

$$\mathbf{y} = \frac{1}{1-\alpha} \sum_{\tau \in \mathrm{bfs}(A') : \tau_e = 0} c_\tau \tau + \sum_{\tau \in \mathrm{pcyc}(A') : \tau_e = 0} c_\tau \tau. \tag{7.17}$$

Then y is the convex combination of at least one bfs and positive cycles, and is also a feasible solution. In particular $x_e y_e \geq 0$ for all e. It's also clear that supp$(\mathbf{y}) \subseteq$ supp$(\mathbf{x})$. Finally,

$$||\mathbf{x} - \mathbf{y}||_\infty \leq \frac{\alpha}{1-\alpha} || \sum_{\tau \in bfs(A') : \tau_e = 0} c_\tau \tau ||_\infty + || \sum_{\tau \in \mathrm{pcyc} : \tau_e > 0} c_\tau \tau ||_\infty. \tag{7.18}$$

Since $\alpha < \frac{1}{2}$, $\frac{\alpha}{1-\alpha} < 2\alpha$. Moreover,

$$|| \sum_{\tau \in bfs(A') : \tau_e = 0} c_\tau \tau ||_\infty < M_2 \sum_{\tau \in bfs(A') : \tau_e = 0} c_\tau < M_2 \tag{7.19}$$

, and $|| \sum_{\tau : \tau_e > 0} c_\tau \tau ||_\infty < M_2 \alpha$. We have

$$||x - y||_\infty \leq 3\alpha M_2 < 3\epsilon \frac{M_2}{M_1}. \tag{7.20}$$

□ □

We use induction to remove multiple small edges.

Proof of Lemma 3.
Let M_1, M_2 be the same as in the previous lemma, and set $K = 3\frac{M_2}{M_1}$. Let $S = \{e_1, e_2, ..., e_{|S|}\}$. It's clear that properties 2 and 3 are satisfied during induction, so we focus on 1. In the base case, we apply the previous lemma to e_1 to construct $\mathbf{y_1}$, such that $||\mathbf{x} - \mathbf{y_1}||_\infty < \epsilon^{|S|+1}K$. The induction claim that we will prove is: if $\exists$ $\mathbf{u}$ that removes $\{e_1, e_2, ...e_{l-1}\}$ from $\mathbf{x}$, and $||\mathbf{x} - \mathbf{u}||_\infty < K\epsilon^n$, $n \leq |S| + 1$, then we can construct $\mathbf{y}$ such that $\mathbf{y}$ removes $\{e_1, ..., e_l\}$ from $\mathbf{x}$, and $||\mathbf{x} - \mathbf{y}||_\infty < K\epsilon^{n-1}$. We can then iterate $|S|$ times to construct $\mathbf{y}$ such that $||\mathbf{x} - \mathbf{y}||_\infty < K\epsilon$.

To prove the induction claim, observe that $||\mathbf{x} - \mathbf{u}||_\infty < K\epsilon^n < \frac{\epsilon^{n-1}}{2}$. Since $n \leq |S| + 1$, $x_{e_l} < \epsilon^{|S|+1} < \epsilon^n$ and

$$u_{e_l} < x_{e_l} + \frac{\epsilon^{n-1}}{2} < \epsilon^{n-1} < \frac{M_1}{2}. \tag{7.21}$$

Apply Lemma to remove e_l from $\mathbf{u}$ and construct $\mathbf{y}$ such that $||\mathbf{y} - \mathbf{u}||_\infty < K\epsilon^{n-1}$. Then

$$||\mathbf{x} - \mathbf{y}||_\infty < ||\mathbf{x} - \mathbf{u}||_\infty + ||\mathbf{y} - \mathbf{u}||_\infty < K\epsilon^n + K\epsilon^{n-1} < K\epsilon^{n-1}. \tag{7.22}$$

□

Let $\beta_{\min} = \min_{\tau \in \text{bfs}}\{\tau_e : e \in \text{supp}(\tau)\}$ and $\beta_{\max} = \max_{\tau \in \text{bfs}}\{\tau_e : e \in \text{supp}(\tau)\}$.

Lemma 6. *Let $\mathbf{x} \in \mathbf{\Phi}^+$ and $e \in \mathbf{E}$ be in the support of $\mathbf{x}$. Then there is a basic feasible solution $\hat{\tau} \in \mathbf{\Phi}^+$ such that*

$$\min \mathbf{x}(supp(\hat{\tau})) \geq \frac{\beta_{min}}{\beta_{max}} \frac{x_e}{|supp(\mathbf{x})|} \tag{7.23}$$

Proof. Since $\mathbf{\Phi}^+$ is a bounded polytope, $\mathbf{x}$ is a convex linear combination of basic feasible solutions. In particular, there exists $\tau \in$ bfs such that $\text{supp}(\tau) \subseteq \text{supp}(\mathbf{x})$. Let $c_\tau = min\{\frac{x_e}{\tau_e} : e \in \text{supp}(\tau)\}$. Then $\mathbf{x} - c_\tau \tau \in \mathbf{\Phi}^+$ and has support strictly contained in the support of $\mathbf{x}$. Iterating this shows that we can write $\mathbf{x} = \sum c_\tau \tau$, where the number of positive c_τ is at most $|\text{supp}(\mathbf{x})|$. Let $\hat{\tau} = \text{argmax}_\tau\{c_\tau \tau_e\}$, then

$$c_{\hat{\tau}} \beta_{max} |\text{supp}(\mathbf{x})| \geq c_{\hat{\tau}} \hat{\tau}_e |\text{supp}(\mathbf{x})| \geq x_e. \tag{7.24}$$

Furthermore

$$\min \mathbf{x}(\text{supp}(\hat{\tau})) \geq c_{\hat{\tau}} \min \hat{\tau}(\text{supp}(\hat{\tau})) \geq c_{\hat{\tau}} b_{min} \tag{7.25}$$

and the desired inequality follows. □

References

[1] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. K. Liu. A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 25*, 2012. 64, 65, 66, 73

[2] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and T. Telgarsky. Tensor decompositions for learning latent variable models, 2012. arXiv:1210.7559. 64, 70, 72, 73, 74

[3] J. Leek at el. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3:1724–35, 2007. 8

[4] J. Zou et al B. Zhao. Epstein-barr virus exploits intrinsic b-lymphocyte transcription programs to achieve immortal cell growth. *PNAS*, 108, 2011. 3

[5] D. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7:781–91, 2006. 6

[6] B. Balle, A. Quattoni, and X. Carreras. Local loss optimization in operator models: A new insight into spectral learning. In *Twenty-Ninth International Conference on Machine Learning*, 2012. 64

[7] Leonard E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73(3):360–363, 1967. 62

[8] D. Bertsimas and J. Tsitsiklis. Introduction to linear optimization. *Athena Scientific*, 1997. 101, 102, 105

[9] David M. Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003. 62, 65

[10] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, 1992. 73

[11] A. T. Chaganty and P. Liang. Spectral experts for estimating mixtures of linear regressions. In *Thirtieth International Conference on Machine Learning*, 2013. 64

[12] B. Chazelle. Natural algorithms. *Proceedings of the 20th Symposium of Discrete Algorithms*, 2009. 95

[13] S. B. Cohen, K. Stratos, M. Collins, D. P. Foster, and L. Ungar. Spectral learning of latent variable PCFGs. In *Proceedings of Association of Computational Linguistics*, 2012. 64

[14] S. B. Cohen, K. Stratos, M. Collins, D. P. Foster, and L. Ungar. Experiments with spectral learning of latent-variable PCFGs. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics*, 2013. 64

[15] R. Dowell. Transcription factor binding variation in the evolution of gene regulation. *Trends in Genetic*, 26:468–475, 2010. 32

[16] A. Coates et al. An analysis of single-layer networks in unsupervised feature learning. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011. 92

[17] A. Krejci et al. Notch activation stimulates transient and selective binding of su(h)/csl to target enhancers. *Genes and Development*, 21:1322–1327, 2007. 32

[18] A. Kulesza et al. Structured determinantal point processes. *Advances in Neural Information Processing Systems*, 23, 2011. 82

[19] A. Kulesza et al. Learning determinantal point processes. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 2011. 82

[20] A. Price et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–9, 2006. 8

[21] A. Rickinson et al. Epstein-barr virus. *Fields Virology*, pages 2655–2700, 2007. 47

[22] A. Scardicchio et al. Statistical properties of determinantal point processes in high-dimensional euclidean spaces. *Physical Review E*, 79, 2009. 82

[23] A. Tero et al. Rules for biological inspired adaptive network design. *Science*, 327: 439–442, 2010. 96

[24] B. Johansson et al. Epstein-barr virus (ebv)-associated antibody patterns in malignant lymphoma and leukemia. i. hodgkins disease. *International Journal of Cancer*, 6:450–462, 1970. 47

[25] C. Alfieri et al. Early events in epstein-barr virus infection of human b lymphocytes. *Virology*, (181):595–608, 1991. 47

[26] C. Del Bianco et al. Notch and maml-1 complexation do not detectably alter the dna binding specificity of the transcription factor csl. *PLoS ONE*, 5, 2010. 32

[27] C. Kaiser et al. The proto-oncogene c-myc is a direct target gene of epstein-barr virus nuclear antigen 2. *Journal of Virology*, 73:4481–4484, 1999. 47

[28] C. Lippert et al. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8:833–5, 2011. 8

[29] C. Lippert et al. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports*, 2013. 8

[30] D. Beck et al. Signal analysis for genome wide maps of histone modifications measured by chip-seq. *Bioinformatics*, 28(8):1062–9, 2012. 72

[31] D. Blei et al. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 81, 83, 87, 88

[32] E. Gomariz-Zilber et al. Isolation procedures for blood lymphocytes produce artifacts. *Cell Biophysics*, 16:55–69. 7

[33] E. Houseman et al. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13, 2012. 7

[34] E. Robertson et al. The amino-terminal domains of epstein-barr virus nuclear proteins 3a, 3b, 3c interact with rbpj. *Journal of Virology*, 70:3068–3074, 1996. 31

[35] F. Wang et al. Epstein-barr virus nuclear protein 2 transactivates a cis-acting cd23 dna element. *Journal Virology*, 65:4101–4106, 1991. 47

[36] H. Kang et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42:348–54, 2010. 8

[37] J. Aster et al. Notch signalling in t-cell lymphoblastic leukaemia/lymphoma and other haematological malignancies. *Journal of Pathology*, 223:262–273, 2011. 31

[38] J. Ernst et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011. 71

[39] J. Lin et al. Transcription factor binding and modified histones in human bidirectional promoters. *Genome Research*, 17:818–827, 2007. 36

[40] J. Listgarten et al. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics*, 29:1526–33, 2013. 6, 8

[41] J. Moller et al. Statistical inference and simulation for spatial point processes. *Monographs on Statistics and Applied Probability.*, 2004. 82

[42] J. Yu et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38:203–8, 2006. 8

[43] J. Zhu et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*, 152(3):642–54, 2013. 71

[44] J. Zou et al. Ewasher: Epigenome-wide association studies without the need for cell-type composition. *Under submission at Nature Methods*, 2013. 2

[45] J. Zou et al. Contrastive learning using spectral methods. *Neural Information Processing Systems*, 2013. 3

[46] K. Alim et al. Random network peristalsis in physarum polycephalum organizes fluid flows across an individual. *Proceedings of the National Academy of Sciences*, 110(33), 2013. 96

[47] K. Ito et al. Convergence properties for the physarum solver. *arXiv*, 2011. 96, 97

[48] M. Epstein et al. Virus particles in cultured lymphoblasts from burkitts lymphoma. *Lancet*, 1:702–703, 1964. 47

[49] M. Ho et al. Epstein-barr virus infections and dna hybridization studies in post-transplantation lymphoma and lymphoproliferative lesions: The role of primary infection. *Journal of Infectious Disease*, 152:876–886, 1985. 47

[50] M. Lechner et al. Cancer epigenome. *Advances in Genetics*, 70:247–76, 2010. 6

[51] P. Hollenhorst et al. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ets gene family. *Genes and Development*, 21:1882–1894, 2007. 34

[52] P. Hollenhorst et al. Dna specificity determinants associate with distinct transcription factor functions. *PLoS Genetics*, 5, 2009. 34

[53] P. Nikitin et al. An atm/chk2-mediated dna damage-responsive signaling pathway suppresses epstein-barr virus transformation of primary human b cells. 8:510–522, 2010. 47

[54] S. Camazine et al. Self-organization in biological systems. *Princeton University Press*, 2003. 95

[55] S. Lee et al. Epstein-barr virus nuclear protein 3c domains necessary for lymphoblastoid cell growth: Interaction with rbp-jkappa regulates tcl1. *Journal of Virology*, 83:12368–12377, 2009. 31

[56] T Henkel et al. Mediation of epstein barr virus ebna2 transactivation by recombination signal-binding protein j kappa. *Science*, (265):92–95, 1994. 31

[57] T. Miyaji et al. Physarum can solve the shortest path problem on riemannian surface mathematically rigorously. *International Journal of Pure and Applied Mathematics*, 47:353–369, 2008. 96

[58] T. Nakagaki et al. Maze-solving by an amoeboid organism. *Nature*, 407:470, 2000. 96

[59] T. Nakagaki et al. A mathematical model for adaptive transport network in path finding by true slime mold. *Journal of Theoretical Biology*, 244:553–564, 2007. 97

[60] T. Palomero et al. Cutll1, a novel human t-cell lymphoma cell line with t(7;9) rearrangement, aberrant notch1 activation and high sensitivity to gamma-secretase inhibitors. *Leukemia*, 20:1279–1287, 2006. 32

[61] V. Bonifaci et al. Physarum can compute the shortest paths. *Proceedings of the 23th Symposium of Discrete Algorithms*, 2012. 96

[62] V. Rakyan et al. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12:529–41, 2011. 6

[63] W. Gawlitta et al. Studies on microplasmodia of physarum polycephalum. *Cell and Tissue Research*, 209(1), 1980. 96

[64] W. Henle et al. Herpes-type virus and chromosome marker in normal leukocytes after growth with irradiated burkitt cells. *Science*, 157:1064–1065, 1967. 47

[65] W. Qiao et al. Pert: a method for expression deconvolution of human blood samples from varied microenvironmental and development conditions. *PLoS Computational Biology*, 8, 2012. 7

[66] Y. Anno et al. Genome-wide evidence for an essential role of the human staf/ znf143 transcription factor in bidirectional transcription. *Nucleic Acids Research*, 39:3116–3117, 2010. 36

[67] Y. Liu et al. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31: 142–7, 2013. 6, 7, 11, 21

[68] Grossman. The epstein-barr virus nuclear antigen 2 transactivator is directed to response elements by the j kappa recombination signal binding protein. *PNAS*, 91: 7568–7572, 1994. 47

[69] J. Zou et al H. Wang. Genome-wide analysis reveals conserved and divergent features of notch1/rbpj binding in human and murine t-lymphoblastic leukemia cells. *PNAS*, 108, 2011. 3

[70] J. Hough. Determinantal processes and independence. *Probability Surveys*, 3: 206–229, 2006. 82

[71] J. Hsieh. Masking of the cbf1/rbpj kappa transcriptional repression domain by epstein-barr virus ebna2. *Science*, 268:560–563, 1995. 31

[72] D. Hsu, S. M. Kakade, and P. Liang. Identifiability and unmixing of latent parse trees. In *Advances in Neural Information Processing Systems 25*, 2012. 64

[73] D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012. 65

[74] T. Jebara. Probability product kernels. *Journal of Machine Learning Research*, 5: 819–844, 2004. 84, 85

[75] A. Johannson and J. Zou. A slime mold solver for linear programming problems. *Lecture Notes in Computer Science*, 7318, 2012. 4

[76] P. Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13:484–92, 2012. 6

[77] J. Kingman. Poisson processes. *Oxford University Press*, 1993. 82

[78] A. Kontorovich, B. Nadler, and R. Weiss. On learning parametric-output HMMs. In *Thirtieth International Conference on Machine Learning*, 2013. 70

[79] R. Kopan and M. Ilagan. The canonical notch signaling pathway: unfolding the activation mechanism. *Cell*, 137:216–233, 2009. 31

[80] M. Kulis and M. Esteller. Dna methylation and cancer. *Advances in Genetics*, 70: 27–56, 2010. 6

[81] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-$(R_1, R_2, ..., R_n)$ approximation and applications of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4):1324–1342, 2000. 72

[82] F. Lavancier. Statistical aspects of determinantal point processes. 2012. 82

[83] A. Portela and M. Esteller. Epigenetic modifications and human disease. *Nature Biotechnology*, 28:1057–68, 2010. 6

[84] D. Schmidt. Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328:1036–1040, 2010. 32

[85] S. Siddiqi, B. Boots, and G. Gordon. Reduced rank hidden markov models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010. 65

[86] D. Strauss. A model for clustering. *Biometrika*, 62:467–475, 1975. 82

[87] D. Sumpter. Collective animal behavior. *Princeton University Press*, 2010. 95

[88] J. Zou and R. Adams. Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems 25*, 2012. 62

[89] J. Zou and R. Adams. Priors for diversity in generative latent variable models. *Neural Information Processing Systems*, 2012. 3