THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

Topic Modeling as an Approach to Deduce Tissue-Specific Regions in the Human Genome

RADHA PATEL
SPRING 2023

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Statistics
with honors in Biochemistry and Molecular Biology

Reviewed and approved* by the following:

Shaun Mahony
Associate Professor of Biochemistry and Molecular Biology
Thesis Supervisor

Santhosh Girirajan
Associate Professor of Biochemistry & Molecular Biology
Thesis Honors Adviser

* Electronic approvals are on file.

ABSTRACT

The human genome contains millions of regulatory DNA sequences and it has become increasingly important to identify how the expression and silencing of these sequences can initiate or prevent natural processes in the body. However, with millions of data, researchers need to employ statistical methods to manipulate and uncover tissue-specific regulation of human DNA sequences. Topic models are statistical, unsupervised machine learning algorithms that discover the main themes in large collections of documents. The goal of my thesis is to assess whether topic modeling can be applied to understand regulatory DNA sequences. Here, I manipulated a subset of DNase I hypersensitive sites to obtain a document-term matrix, so that I could employ Latent Dirichlet Allocation (LDA), the simplest topic model. Regulatory regions of DNA acted as the "documents", 5-mers acted as the "words", and each unique occurrence of a 5-mer within a document acted as the "counts." Then I used the most probable DNA sequences within each topic to investigate if each topic contained human phenotype terms and motifs of biological significance. I hypothesized that each topic would contain biologically significant motifs with tissue-specific transcriptional binding sites. However, after employing the GREAT tool and MEME-ChIP software, I ended up seeing little biological significance in my results. Therefore, future studies should take alternative approaches to topic modeling. This includes utilizing motif scanning to replace the word counts to minimize the number of input data into LDA and generate a higher probability of obtaining biologically significant results.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Mahony for his immense helpfulness and patience in guiding me through undergraduate research in his lab. I would also like to thank Dr. Girirajan for allowing me to write an honor thesis under his wing. Translating this research into an honors thesis was an incredible experience.

Additionally, I would like to my family and friends for supporting me as I conducted my research and wrote my thesis; I could not have done this without you all.

Chapter 1

Introduction and Literature Review

An Overview of Gene Regulation and Transcription Factor Binding

Gene regulation — expression and silencing of DNA coding sequences — regulates the basic processes of life in the human body; these include protein synthesis, the presence of a specific trait, disease progression, symptom severity, etc. (Latchman, 2005). Gene regulation processes are essential in maintaining homeostasis and governing gene transcription in organisms (Ladunga, 2010). Therefore, genomic sequences play a significant role in the function of organisms, specifically humans. Starting a few decades ago, sequencing technology has allowed researchers to determine valuable regulatory sequences in the human genome (Velculescu, 1995). In general, gene regulation dictates how organisms develop and respond to their environment (Ladunga, 2010). Regulatory regions of DNA include promoters, enhancers, silencers, and insulators. Promoters initiate transcription, enhancers increase the likelihood of transcription, silencers reduce the likelihood of transcription, and insulators prevent inappropriate interactions between regions of the genome (Riethoven, 2010).

Furthermore, thousands of regulatory regions in an organism recognize many transcription factors: regulatory proteins that turn on or express certain genes in various patterns through binding interactions. However, organisms highly regulate and control transcription factor binding. Transcription factor binding is highly specific and different TFs usually recognize non-overlapping sets of genomic sequences (Aptekmann et al., 2022). One can extract motifs from regulatory regions of DNA to determine correspondence to TF binding preferences. Motifs are patterns of nucleic acid sequences containing biological significance such as binding to a specific TF (Das & Dai, 2007). In eukaryotes, TF binding is

often tissue-specific to regulate certain regions of the body. Many studies have looked at tissue-specific TF binding on model organisms such as mice and cattle (Steuernagel et al., 2019). Furthermore, TF binding has also been extensively studied in humans. For example, researchers employed variations of DNA footprinting on data from the Encyclopedia of DNA Elements (ENCODE) to characterize tissue-specific TF binding sites (Funk et al., 2020). While TF binding has been extensively characterized, researchers still need more methods to understand how TF binding leads to tissue-specific regulatory programs. One can employ computational methods such as topic modeling to determine tissue-specific TF binding events in regulatory regions of DNA.

A Brief Overview of Topic Modeling

Topic models are unsupervised machine learning algorithms that discover the main themes in large collections of documents (Blei, 2012). The simplest topic model approach is called Latent Dirichlet Allocation (LDA) and it traditionally models discrete data such as text (Blei et al., 2003). Topic modeling also uses specific terminology. A word is the basic unit of discrete data, a document is a sequence of words, and a corpus is a collection of documents (Blei et al., 2003). LDA is a parametric approach meaning that its algorithm requires an individual to specify the number of topics. It is an example of a generative statistical model meaning that the model captures joint probabilities (Ng & Jordan, 2002). This is in contrast with discriminative statistical modeling which is a model that captures conditional probabilities (Ng & Jordan, 2002). In addition, LDA is most optimal with longer documents and more words; short bodies of text containing 20 words or less do not perform well with LDA (Yan et al., 2013). Typical output for LDA includes a document-topic probabilities matrix and a word-topic probabilities matrix. Both matrices provide probabilities that indicate the strength of association for each document to the topic and each word to a topic, respectively.

Beyond LDA, some other topic modeling approaches include Non-negative Matrix Factorization (NMF) and Hierarchical Dirichlet Processes (HDP). NMF decompresses high-dimensional data into a lower-dimensional representation. The original matrix (A) produces two lower-ranked matrices: a feature matrix (W) and a coefficient matrix (H) (Zhao et al., 2019). NMF achieves this decomposition through linear algebraic techniques with an iterative approach to ensure that the product of W and H gets as close as possible to A (Lee & Seung, 1999). In contrast to many topic modeling methods, HDP is a nonparametric topic modeling approach. This method possesses the advantage of determining a learned maximum number of topics from the data and therefore, one does not need to specify the number of topics in advance (Teh et al., 2006). While many of these topic modeling techniques originated using text as the basic unit, researchers adapt and implement topic modeling approaches to discover patterns in data beyond text. This includes its application on genetic data, images, and social networks (Blei, 2012).

Prior Applications of Topic Modeling in Genomics

Researchers can employ machine learning techniques — such as topic modeling methods — to deduce regulatory complexes between transcription factors and their specific binding regions on the genome. They can also use these techniques to deduce overlapping DNA sequences in the genome. In previous studies, cluster-based techniques such as k-means clustering illustrate that the overlapping of transcriptional signals complicates the interpretation of regulatory complexes (Guo & Gifford, 2017). In other words, cluster-based techniques assume that each observed data point must fit in a singular cluster. Therefore, my research employs topic modeling – algorithms that allow for data points to be generated as a mixture of multiple fundamental topics; unlike hard-cluster approaches, topic models can effectively capture overlapping signals (Blei, 2012). Past studies successfully apply topic modeling to deduce regulatory complexes and overlapping sequences in the genome. For example, Guo & Gifford (2017)

used regulatory module discovery (RMD), originating from Hierarchical Dirichlet Processes, to arrange transcription factors by topic or regulatory module. They hypothesize that this method outperforms hard-clustering approaches such as k-means clustering. Later in the study, they determine their method best captured complex binding regions compared to k-means clustering. Another study applied LDA on 30 genomes from 3 different bacterial families; each genome was considered as a document with 13 base-pair k-mers acting as the words (Borrayo et al., 2020). DNA k-mers are small fragments of the DNA sequence consisting of the four base pairs; these include A for adenine, T for thymine, C for cytosine, and G for guanine. Topic modeling allowed these researchers to compare genomes based on their bacterial composition determined by identifying biological similarities in DNA sequences. In addition, La Rosa et al., (2015) used Probabilistic Topic Modeling obtained through LDA where bacterial DNA sequences acted as the documents and DNA k-mers acted as the words. They hypothesized and concluded that Probabilistic Topic Modeling could classify DNA sequences through this k-mer representation (La Rosa et al., 2015). Similarly, my study applied LDA where human DNA sequences act as the documents and 5-mers act as the words.

Origins of the Dataset Used in this Thesis

The dataset used in this thesis is in .fasta format containing around 3.1 million human DNA sequences after filtering out sequences that were less than 100 base pairs and greater than 1000 base pairs. Meuleman et al. (2020) created the original dataset containing around 3.6 million DNase I hypersensitive sites (DHSs). A DHS is a site on the genome that was characterized to be sensitive to the DNase I enzyme in 438 cell types used in this study. Sensitivity to the DNase I enzyme indicates that the chromatin is more decondensed in a region compared to others; the “open” chromatin is associated with active regulatory processes such as enhancers and promoters (Gross & Garrard, 1988). Therefore, DHSs

represent a large set of regulatory DNA regions on the human genome sourced from a wide range of cell types (Meuleman et al., 2020). DHSs often contain genetic variation associated with diseases and traits. Furthermore, the variation contained in DHSs is often responsible for traits explained by single-nucleotide polymorphism (SNPs) determined through genome-wide association studies (Meuleman et al., 2020). Meuleman et al. (2020) used 733 biosamples representing 438 cell or tissue types to construct a common reference system for regulatory sequences in the human genome. NMF was also applied with the biosamples acting as the documents and the regulatory regions acting as the words with 16 input components. They found that similar DHSs were highly clustered together in the human genome. In fact, DHSs across the genome might share a common regulatory mechanism. My study aims to determine clusters of DHSs on the human genome by applying LDA where DHSs act as the documents and 5 base pair k-mers act as the words.

Purpose of this Thesis

My thesis aims to apply Latent Dirichlet Allocation (LDA) — a topic modeling method — on a subset of a dataset containing around 3.6 million DNase I hypersensitive sites (DHSs). 5-mers, five base pair substrings of each regulatory DNA sequence, will act as the words, and the regulatory DNA sequences will act as the documents in my application. I am interested in identifying specific 5-mers that serve a similar biological purpose on the genome. First, I will obtain the document-topic matrix to identify the most probable documents within a topic. Then, I will generate the word-topic matrix and determine motifs corresponding to each 5-mer. Then, I will determine the biological significance of these motifs with respect to tissue type and TF factor binding to deduce tissue-specific TF factor regulation in humans.

Chapter 2

Methodology

Manipulation of Regulatory DNA Sequence Data

Transforming the Dataset into a Document-Term Matrix

The dataset, DNASequences.fa, applied in this thesis contains 3.1 million human DNA sequences after filtering out sequences that were less than 100 base pairs (too short) and greater than 1000 base pairs (too long). These sequences were extracted and placed into a FASTA format file. In FASTA format, the line before the DNA sequence contains a sequence identifier; in this case, the identifier contains the chromosome number and coordinates for each DNA sequence (Cock et al., 2010). This format is clearly seen in Figure 1 which contains a 10-observation subset of DNASequences.fa. Each DNA sequence is represented using the four base pairs: adenine (A), thymine (T), cytosine (C), and guanine (G). However, due to the difficulty of running large datasets in R, I ended up randomly selecting 10,000 sequences in the data to manipulate in this study. While the file DNASequences.fa is in a useful format for bioinformatics, it is essential to transform these observations into a format more suitable to run the LDA function in R. This function requires the input matrix to be a Document-Term Matrix (DTM). A Document-Term Matrix contains each document, term, and count numerically represented as the DNA sequence, 5-mers, and frequency of occurrence of each 5-mer within a sequence, respectively. However, one must slice each DNA sequence and find its reverse complement to obtain 5-mers.

Figure 1. A Glimpse of 10 Observations of DNASequences.fa Illustrating FASTA Format

Slicing and Counting 5-mers

First, each DNA sequence needed to be sliced into 5-mers. Using a for-loop in R, I moved through each index of each DNA sequence to slice 5-mers. Additionally, since DNA is double-stranded, one must incorporate the reverse complements of each 5-mer into the Document-Term Matrix. Adenine (A) bonds with thymine (T) and cytosine (C) bonds with guanine (G). Therefore, I reversed each DNA sequence, took the complement of the reversed sequence, and moved through each index of the reversed sequence to slice 5-mers. An example of a 5-mer and its reverse complement is illustrated in Figure 2. After finding each 5-mer and reverse complement 5-mer for each DNA sequence, I tallied up the frequency or count of each 5-mer into another column of the matrix.

5-mer: ACATG

Reverse Complement: CATGT

Figure 2. An Example 5-mer and its Reverse Complement

Finalizing the DCM Matrix

After slicing and counting 5-mers in each DNA sequence, I numerically converted each column of the matrix. For the document column, I represented each DNA sequence with a unique number from 1 to 10,000. For the term column, I represented each base pair with the following conversion: A to 1, T to 2, G to 3, and C to 4. Finally, the count column is already numeric and therefore, no extra step was necessary. A 10-observation of the Document-Term Matrix used as input for LDA is represented in Figure 3. The script to obtain the DCM matrix for 10,000 observations took approximately 48 hours to run and required the packages seqinr, dplyr, tidyr, stringr, and data.table in R Version 4.2.1 to sufficiently complete. Additionally, before running the script on all 10,000 DNA Sequences from DNASequences.fa, I ran a test script containing 2 observations to ensure that the final script would run as expected and correctly.

```
> head(DCMMatrix,10)
   Documents Words Count
1          1 11111     6
2          1 11113     2
3          1 11131     3
4          1 11121     1
5          1 11123     1
6          1 11122     1
7          1 11441     1
8          1 11433     1
9          1 11421     1
10         1 11311     2
```

Figure 3. 10-Observation Subset of the Document-Term Matrix Used for LDA

Latent Dirichlet Allocation

Determining Optimal LDA Topics

After obtaining the Document-Term Matrix (DTM), I needed to determine the optimal number (k) of topics to input into LDA. Recall that LDA is a parametric topic model and therefore, one needs to manually input k number of topics. There are a number of metrics that one can employ to determine the number of topics for LDA. For this study, I used four metrics in the ldatuning package; these metrics require the testing of many LDA models to assess the best performance. The results of utilizing these methods can be seen in the plot in Figure 4. I selected 30 topics to fit my final LDA model because it strongly minimizes the Arun2010 and CaoJuan2009 metrics; it also maximizes the Deveaud2014 and Griffiths2004 metrics. The metric Arun2010 is a divergence measure that compares the quality of the contents of the document-term and word-term matrices that serve as output for LDA. When the metric is low for a given number of topics, this indicates when the "right" number of topics is reached (Arun et al., 2010).

Similarly, the metric CaoJuan2009 seeks to use a density-based approach to find the number of topics that minimizes instability (Cao et al., 2009). The Deveaud2014 metric utilizes Latent Concept Modeling (LCM), an unsupervised method to maximize useful information in LDA (Deveaud et al., 2014). Finally, the Griffiths2004 metric utilizes a Bayesian model to select the number of topics that maximizes valuable information in the data (Griffiths & Steyvers, 2004).

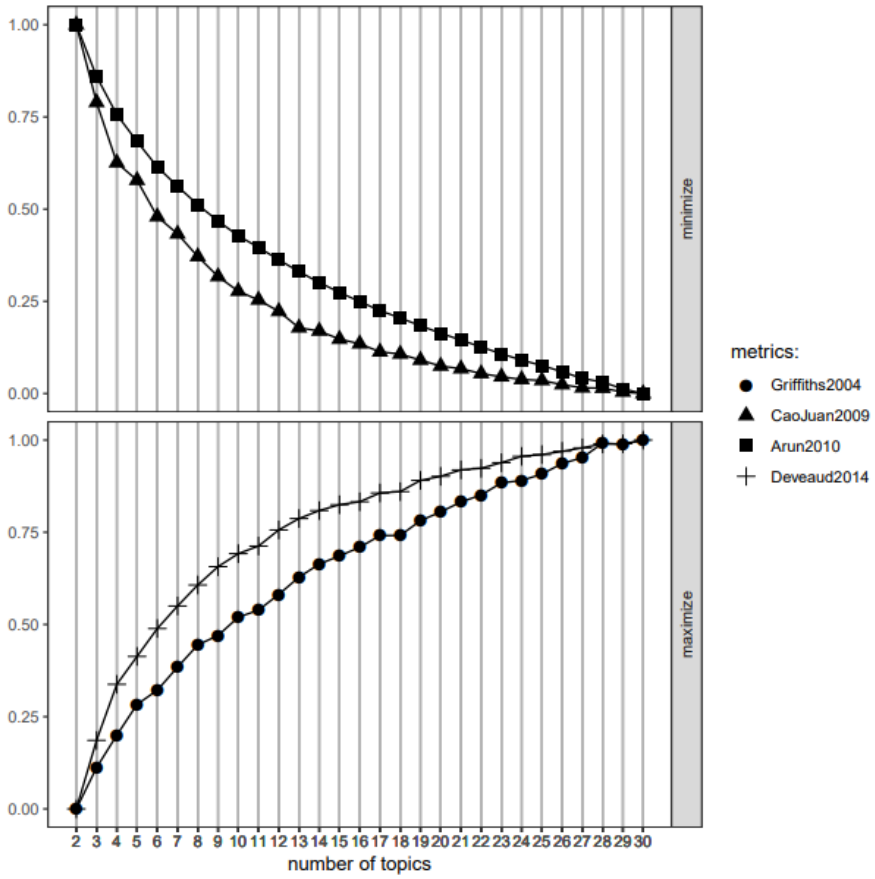


Figure 4. A Plot Displaying the Maximizing and Minimizing Metrics to Select k for LDA

Running LDA and Output

After determining the optimal number of topics (30 topics) for LDA, I successfully ran Latent Dirichlet Allocation on the 10,000 observation Document Term Matrix in R. I obtained the word-topic probability matrix and the document-topic probability matrix. The word-topic matrix in Figure 5 contains each topic, 5-mer, and a beta value. The beta value represents the probability of the 5-mer occurring in the specific topic. Similarly, the document-topic matrix in Figure 6 contains each topic, DNA sequence, and a gamma value. The gamma value represents the probability of the DNA sequence occurring in the specific

document. For my analysis, I am most interested in utilizing the document-topic matrix, particularly the coordinates, to determine tissue-specific motifs and the transcription factors that bind to them.

```
> word_topics <- read.csv('word_topics.csv')
> head(word_topics,10)
   topic  term         beta
1      1 11111 0.0026591132
2      2 11111 0.0062805691
3      3 11111 0.0006197049
4      4 11111 0.0008478076
5      5 11111 0.0078394285
6      6 11111 0.0053194418
7      7 11111 0.0067239150
8      8 11111 0.0035890875
9      9 11111 0.0003534374
10    10 11111 0.0001385162
```

Figure 5. 10-Observation Subset of the Word-Topic Matrix

```
> doc_topics <- read.csv('doc_topics.csv')
> head(doc_topics,10)
   document topic      gamma
1         1     1 0.03319791
2         2     1 0.03317301
3         3     1 0.03288844
4         4     1 0.03365835
5         5     1 0.03346057
6         6     1 0.03393994
7         7     1 0.03337817
8         8     1 0.03392380
9         9     1 0.03307184
10       10     1 0.03423524
```

Figure 6. 10-Observation Subset of the Document-Topic Matrix

Acquiring Probable Genes and Motifs

Obtaining DNA Sequence Coordinates

I filtered each document-topic matrix by assigning each document to its most probable topic. I ended up with 30 data sets each representing one topic. Recall, each document number represents a DNA sequence coordinate in the original data set. Therefore, I merged each DNA sequence coordinate with its corresponding document number. Then, I used Emacs to isolate each coordinate per document in a tab-delimited form to create 30 BED files. An example of the data set for topic 1 is illustrated in Figure 7.

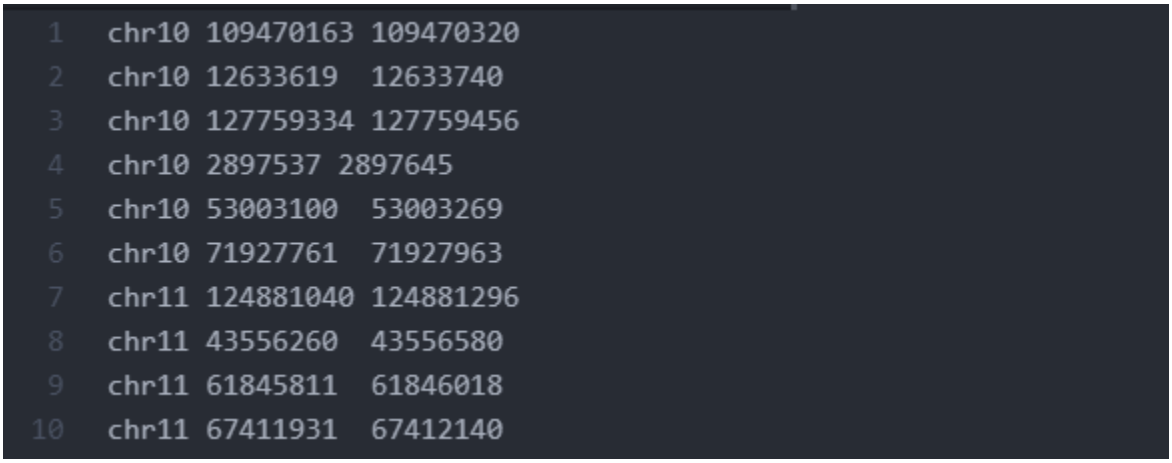


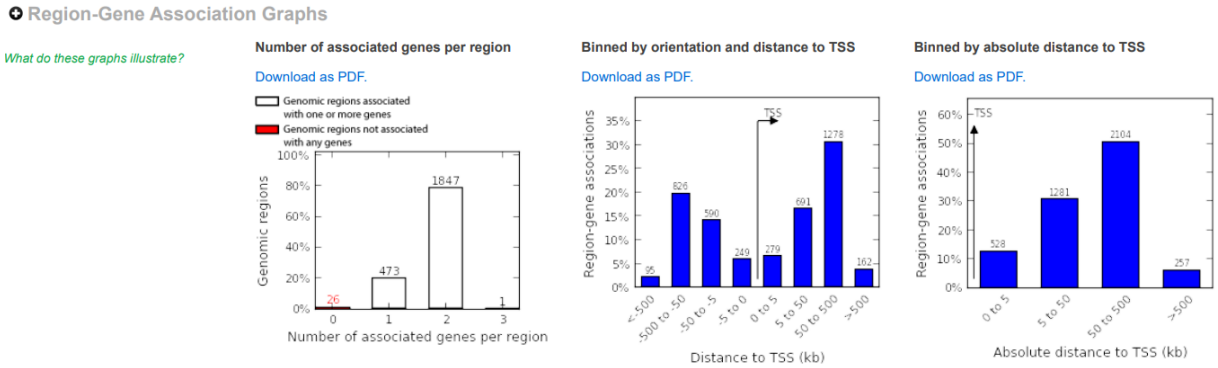
Figure 7. 10-Observation Subset of the Topic 1 BED File Containing the Most Probable Coordinates

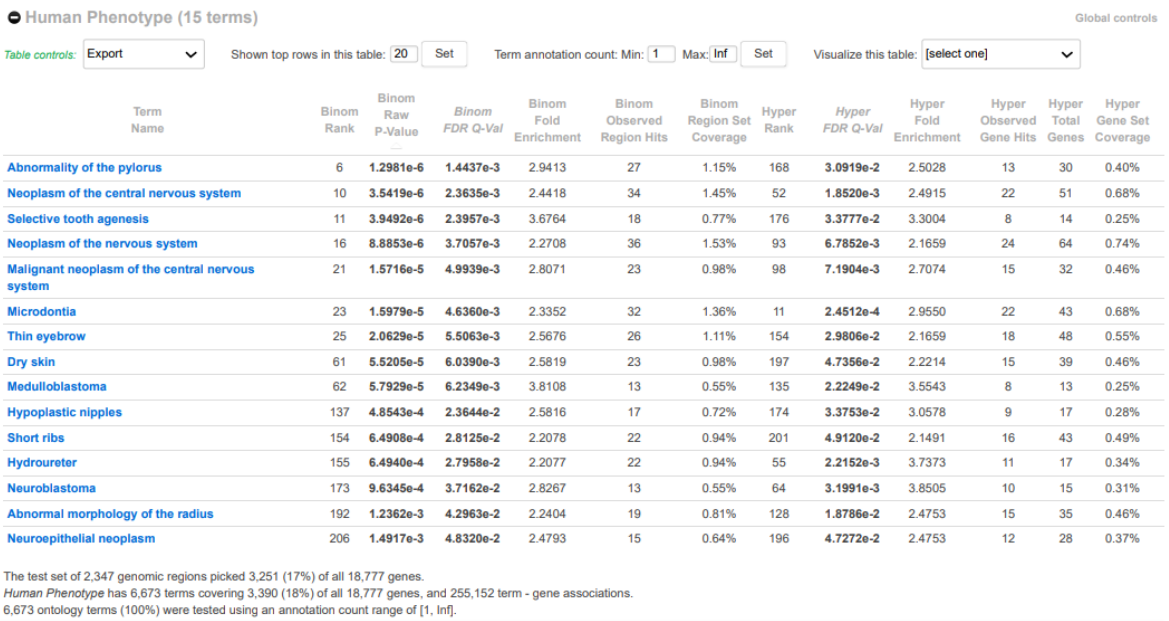
Chapter 3

Results

Running DNA Sequence Coordinates in GREAT Software

After obtaining 30 data sets containing the most probable coordinates per topic, I ran each BED file through software entitled GREAT. McLean et al. (2010) developed GREAT stands for Genomic Regions Enrichment of Annotations Tool; it is employed to determine functional significance within regulatory regions of DNA. Terms are generated for each regulatory DNA sequence by conducting localized measurements of DNA binding events across the genome (McLean et al., 2010). Some examples of these terms include human phenotype terms, mouse phenotype terms, gene oncology cellular component, etc. I was most interested in looking at the human phenotype term to possibly see if topic modeling outputs tissue-specific regulatory DNA sequences. However, after running each of the 30 datasets into GREAT, we only found one topic — namely, topic 23 — to contain many terms of biological significance when looking at human phenotype terms. The output of GREAT for topic 23 is illustrated in Figure 8; there are 15 terms in the Human Phenotype section which shows this BED file holds some biological significance. Looking at Table 1, topic 23 contains the most coordinates compared to the other topics and therefore, the increase of input data for GREAT might explain the output of many terms for that particular dataset. In comparison to topic 23, the lack of data in the other topics might explain why there is a lack of output for human phenotype terms. However, McLean et al (2010) determined that inputting more than 1,000 DNA Coordinates could result in more false positive enriched terms. Therefore, it was imperative that I utilized another tool or software such as MEME-ChIP to determine biologically significant motifs within each DNA coordinate to obtain more accurate results for our data.





Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
Abnormality of the pylorus	6	**1.2981e-6**	**1.4437e-3**	2.9413	27	1.15%	168	**3.0919e-2**	2.5028	13	30	0.40%
Neoplasm of the central nervous system	10	**3.5419e-6**	**2.3635e-3**	2.4418	34	1.45%	52	**1.8520e-3**	2.4915	22	51	0.68%
Selective tooth agenesis	11	**3.9492e-6**	**2.3957e-3**	3.6764	18	0.77%	176	**3.3777e-2**	3.3004	8	14	0.25%
Neoplasm of the nervous system	16	**8.8853e-6**	**3.7057e-3**	2.2708	36	1.53%	93	**6.7852e-3**	2.1659	24	64	0.74%
Malignant neoplasm of the central nervous system	21	**1.5716e-5**	**4.9939e-3**	2.8071	23	0.98%	98	**7.1904e-3**	2.7074	15	32	0.46%
Microdontia	23	**1.5979e-5**	**4.6360e-3**	2.3352	32	1.36%	11	**2.4512e-4**	2.9550	22	43	0.68%
Thin eyebrow	25	**2.0629e-5**	**5.5063e-3**	2.5676	26	1.11%	154	**2.9806e-2**	2.1659	18	48	0.55%
Dry skin	61	**5.5205e-5**	**6.0390e-3**	2.5819	23	0.98%	197	**4.7356e-2**	2.2214	15	39	0.46%
Medulloblastoma	62	**5.7929e-5**	**6.2349e-3**	3.8108	13	0.55%	135	**2.2249e-2**	3.5543	8	13	0.25%
Hypoplastic nipples	137	**4.8543e-4**	**2.3644e-2**	2.5816	17	0.72%	174	**3.3753e-2**	3.0578	9	17	0.28%
Short ribs	154	**6.4908e-4**	**2.8125e-2**	2.2078	22	0.94%	201	**4.9120e-2**	2.1491	16	43	0.49%
Hydroureter	155	**6.4940e-4**	**2.7958e-2**	2.2077	22	0.94%	55	**2.2152e-3**	3.7373	11	17	0.34%
Neuroblastoma	173	**9.6345e-4**	**3.7162e-2**	2.8267	13	0.55%	64	**3.1991e-3**	3.8505	10	15	0.31%
Abnormal morphology of the radius	192	**1.2362e-3**	**4.2963e-2**	2.2404	19	0.81%	128	**1.8786e-2**	2.4753	15	35	0.46%
Neuroepithelial neoplasm	206	**1.4917e-3**	**4.8320e-2**	2.4793	15	0.64%	196	**4.7272e-2**	2.4753	12	28	0.37%

The test set of 2,347 genomic regions picked 3,251 (17%) of all 18,777 genes.
Human Phenotype has 6,673 terms covering 3,390 (18%) of all 18,777 genes, and 255,152 term - gene associations.
6,673 ontology terms (100%) were tested using an annotation count range of [1, Inf].

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
small superior vagus ganglion	6	**2.6020e-7**	**3.9710e-4**	4.0621	20	0.85%	359	**1.1334e-2**	2.8879	11	22	0.34%
decreased basal metabolism	12	**4.5765e-6**	**3.4923e-3**	4.0123	16	0.68%	477	**2.6796e-2**	3.8505	6	9	0.18%
abnormal embryonic hematopoiesis	21	**2.2027e-5**	**9.6047e-3**	2.2011	35	1.49%	255	**3.6567e-3**	2.2885	21	53	0.65%
common ventricle	29	**4.7409e-5**	**1.4970e-2**	4.5260	11	0.47%	404	**1.8063e-2**	3.6755	7	11	0.22%
abnormal vascular endothelial cell development	30	**5.4264e-5**	**1.6563e-2**	2.5245	24	1.02%	182	**6.8077e-4**	2.9810	16	31	0.49%
increased sarcoma incidence	36	**9.7616e-5**	**2.4830e-2**	2.0384	35	1.49%	439	**2.0850e-2**	1.8133	27	86	0.83%
abnormal fibula morphology	43	**1.7877e-4**	**3.8069e-2**	2.0721	31	1.32%	100	**2.0676e-5**	2.9521	23	45	0.71%
abnormal superior vagus ganglion morphology	46	**1.9398e-4**	**3.8615e-2**	2.5469	20	0.85%	517	**2.9953e-2**	2.5413	11	25	0.34%

Figure 8. Output for Topic 23 After Employing the GREAT Tool

Table 1. Number of Most Probable Coordinates Within Each Topic

Topic Number	Number of Coordinates in the Dataset
1	144
2	560
3	126
4	283
5	396
6	59
7	726
8	207
9	117
10	319
11	124
12	156
13	511
14	142
15	302
16	113
17	137
18	233
19	60
20	103
21	66
22	1175
23	2347
24	479
25	155
26	198
27	118
28	244
29	199
30	201

Applying MEME-ChIP on DNA Sequence Coordinates

After employing the GREAT tool on each BED file, I utilized MEME-ChIP to find biologically significant motifs within each topic. Machanick & Bailey (2011) developed MEME-ChIP to conduct motif analysis on large sets of data. MEME-ChIP employs ab initio motif discovery, motif enrichment analysis, motif visualization, binding affinity analysis, and motif identification along with MEME and DREME to discover biologically significant motifs (Machanick & Bailey, 2011). One can align motifs to tissue-specific transcription factors based on TF binding preferences. For example, one can deduce a particular topic as biologically significant for liver-specific regulation if motifs align with TF binding preferences. In a broad sense, I hoped to find tissue-specific groups of transcription factors associated with motifs within each topic. However, I received no biologically significance motifs for 10 out of the 30 topics, namely topics 3, 5, 6, 8, 16, 19, 21, 26, 29, and 30. The other topics contained around 1 to 3 motifs. However, some of the motifs are merely repetitive regions in the genome and are not conducive to my analysis. An example of a repetitive motif is seen in topic 2 in Figure 9. A motif that is less repetitive and seems to align with a transcription factor similar to CTCF, a transcriptional repressor, is seen in topic 11 in Figure 10.

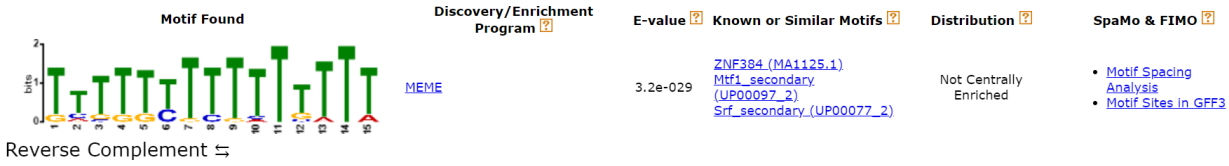


Figure 9. An Example of a Repetitive Motif in Topic 2

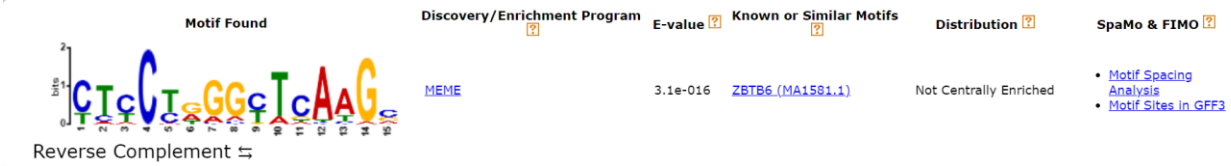


Figure 10. An Example of a Less Repetitive Motif in Topic 11

Chapter 4

Discussion and Concluding Remarks

My research sought to employ LDA on a subset of DNase I hypersensitive sites to obtain a document-term matrix where I can extract the most probable documents in each topic. Then, I used the GREAT tool and MEME-ChIP software to assess the biological significance of the most probable documents (DNA sequences) in each topic. I hypothesized that each topic would possess DNA sequences containing tissue-specific motifs with corresponding transcription factors. However, looking at the results for GREAT and MEME-ChIP, there is not enough evidence to see tissue-specific groupings of biologically significant motifs and transcription factors within the most probable documents per topic generated from LDA. The GREAT tool did not provide human phenotype terms for all 30 topics and MEME-ChIP generated few motifs relevant to my study. However, there are potential explanations for these findings and future directions one can take to obtain results of biological significance. It is possible that each BED file contains a large vocabulary of sequences. Future studies might want to minimize this vocabulary before running LDA to contain the most relevant sequences. Recall that I moved through each index of each DNA sequence to slice 5-mers and took its reverse complement to obtain the DCM matrix for input into LDA. While I included all possible 5-mers when running LDA, some of these "words" include the most frequent repetitive sequences of DNA that are not as meaningful to determine tissue-specific motifs and transcription factors. To minimize the sheer number of words in the DCM matrix, one can take an alternative approach when replicating this study. Instead of taking all length 5-mers, one can represent DNA regulatory regions by examining catalogues of motifs. In LDA, the documents are still regulatory regions of DNA, but the word count contains how many types of motifs are hits. Motif scanning means finding all known or similar motifs in a DNA sequence. Essentially, one can employ motif scanning prior to running LDA to ensure there is a smaller vocabulary so that there is a higher

probability of obtaining biologically significant results. This would also help to prevent large overlaps between topics which could complicate classifying a topic to tissue-specific regulation.

Additionally, one might want to experiment more with utilizing a different number of topics. Despite using four metrics through the ldatuning package in R to determine the optimal number of input topics, it seemed that as k increased, the optimization of LDA increased. Therefore, I selected 30, the maximum k that I tested, to run LDA. Future studies could employ a larger k value beyond 30 to run LDA with the hopes of increasing model stability. One can also use other parameter selection methods to determine the optimal number of topics for LDA. For example, Gan & Qi (2021) created a comprehensive judgment index based on perplexity, isolation, stability, and coincidence to determine the optimal k for LDA. Zhao et al. (2015) also looked at the rate of perplexity change as a function of the number of topics to select the value of k. One can also use a different topic modeling approach beyond LDA such as Hierarchical Dirichlet Allocation (HDP) which is a non-parametric model. Hence, one does not need to input an exact number of topics into HDP model. This would eliminate the extra step of parameter tuning and allow the model to tune itself to reduce instability.

Appendix A

test_script_cluster.R

```
library(seqinr)
library(dplyr)
library(tidyr)
library(stringr)
library(data.table)

DNAseq <- read.fasta('DNASequences.fa', as.string = TRUE)
DNA <- unlist(head(DNAseq,2))
namesDNA <- names(DNA)
DNAmatrix <- data.frame(namesDNA, DNA)
DNAmatrix <- setnames(DNAmatrix, c("Documents", "Words"))
splitwords <- c()
docnames <- c()

for(i in 1:length(DNA)){
  string <- toString(DNAmatrix[i,2])
  dname <- as.character(DNAmatrix[i,1])
  characters <- nchar(string)
  while(characters >=5){
    kmer <- substr(string,1,5) # getting the first 5 characters
    splitwords <- append(splitwords, kmer)
    docnames <- append(docnames, dname)
```

```
    string <- paste(unlist(strsplit(string,""))[-1], collapse = "") # removing the first index
    characters <- nchar(string)
  }
}
DNAmatrix <- data.frame(docnames, splitwords)
DNAmatrix <- setnames(DNAmatrix, c("Documents", "Words"))

backwardssplitwords <- c()
docnames <- c()

for(i in 1:nrow(DNAmatrix)){
  kmer <- as.character(DNAmatrix[i,2])
  dname <- DNAmatrix[i,1]
  string_split <- strsplit(kmer, "")[[1]]
  reversed_string <- paste(rev(string_split), collapse="")
  backwardssplitwords <- append(backwardssplitwords, reversed_string)
  docnames <- append(docnames, dname)
}

backward <- data.frame(docnames, backwardssplitwords)
backward <- setnames(backward, c("Documents", "Words"))

DNAmatrix <- rbind(DNAmatrix, backward)
DNAmatrix <- DNAmatrix %>% group_by(Documents,Words) %>% tally()
```

```
DNAmatrix <- setnames(DNAmatrix, c("Documents", "Words", "Count"))
DNAmatrix$Words <- as.character(DNAmatrix$Words)

integer_kmer <- c()

for (i in 1:nrow(DNAmatrix)){
  kmer <- DNAmatrix[i,2]
  kmer <- gsub("a", "1", kmer)
  kmer <- gsub("t", "2", kmer)
  kmer <- gsub("g", "3", kmer)
  kmer <- gsub("c", "4", kmer)
  integer_kmer <- append(integer_kmer, kmer)
}
DNAmatrix$Words <- as.numeric(integer_kmer)
doc_coordinates <- distinct(data.frame(as.character(DNAmatrix$Documents)))
write.csv(x = doc_coordinates, file = "coordinates.csv", row.names = FALSE)

Unique_Document<- unique(as.character(DNAmatrix$Documents))
doc_coordinates <- data.table(unique(DNAmatrix$Documents))

Unique_Document<
data.frame(Documents=as.numeric(factor(DNAmatrix$Documents,levels=Unique_Document)))
DNAmatrix$Documents <- Unique_Document$Documents
write.csv(x = DNAmatrix, file = "DNAMatrix.csv", row.names = FALSE)
```

Appendix B

topic_modeling_DNA.R

```
library(seqinr)
library(dplyr)
library(tidyr)
library(stringr)
library(data.table)

DNAseq <- read.fasta('DNASequences.fa', as.string = TRUE)
DNA <- unlist(sample(DNAseq, 10000))
namesDNA <- names(DNA)
DNAmatrix <- data.frame(namesDNA, DNA)
DNAmatrix <- setnames(DNAmatrix, c("Documents", "Words"))

splitwords <- c()
docnames <- c()

for(i in 1:length(DNA)){
  string <- toString(DNAmatrix[i,2])
  dname <- as.character(DNAmatrix[i,1])
  characters <- nchar(string)
  while(characters >=5){
    kmer <- substr(string,1,5) # getting the first 5 characters
    splitwords <- append(splitwords, kmer)
```

```
    docnames <- append(docnames, dname)
    string <- paste(unlist(strsplit(string,""))[-1], collapse = "") # removing the first index
    characters <- nchar(string)
  }
}
DNAmatrix <- data.frame(docnames, splitwords)
DNAmatrix <- setnames(DNAmatrix, c("Documents", "Words"))

backwardssplitwords <- c()
docnames <- c()

for(i in 1:nrow(DNAmatrix)){
  kmer <- as.character(DNAmatrix[i,2])
  dname <- as.character(DNAmatrix[i,1])
  string_split <- strsplit(kmer, "")[[1]]
  reversed_string <- paste(rev(string_split), collapse="")
  backwardssplitwords <- append(backwardssplitwords, reversed_string)
  docnames <- append(docnames, dname)
}
backward <- data.frame(docnames, backwardssplitwords)
backward <- setnames(backward, c("Documents", "Words"))
DNAmatrix <- rbind(DNAmatrix, backward)
DNAmatrix <- DNAmatrix %>% group_by(Documents,Words) %>% tally()
DNAmatrix <- setnames(DNAmatrix, c("Documents", "Words", "Count"))
```

```
DNAmatrix$Words <- as.character(DNAmatrix$Words)
integer_kmer <- c()

# note each base to integer conversion
for (i in 1:nrow(DNAmatrix)){
  kmer <- DNAmatrix[i,2]
  kmer <- gsub("a", "1", kmer)
  kmer <- gsub("t", "2", kmer)
  kmer <- gsub("g", "3", kmer)
  kmer <- gsub("c", "4", kmer)
  integer_kmer <- append(integer_kmer, kmer)
}
DNAmatrix$Words <- as.numeric(integer_kmer)
doc_coordinates <- distinct(data.frame(as.character(DNAmatrix$Documents)))
write.csv(x = doc_coordinates, file = "coordinates_3.csv", row.names = FALSE)

Unique_Document<- unique(as.character(DNAmatrix$Documents))
Unique_Document<-
data.frame(Documents=as.numeric(factor(DNAmatrix$Documents,levels=Unique_Document)))
DNAmatrix$Documents <- Unique_Document$Documents
write.csv(x = DNAmatrix, file = "DNAMatrix_3.csv", row.names = FALSE)
```

Appendix C

best_k.R

```
library(seqinr)
library(dplyr)
library(tidyr)
library(stringr)
library(data.table)
library(tm)
library(tidytext)
library(ldatuning)

# 10,000 observation matrix import
DNAmatrix <- read.csv('DNAMatrix_3.csv')
DTM <- DNAmatrix %>% cast_dtm(document = Documents, term = Words, value = Count)
#TDM <- DNAmatrix %>% cast_tdm(document = Documents, term = Words, value = Count)
# Figuring out k value based on metrics
result <- FindTopicsNumber(
  DTM,
  topics = seq(from = 2, to = 30, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  mc.cores = 2L,
  verbose = TRUE
)
FindTopicsNumber_plot(result)
```

Appendix D

document_term_prob.R

```
library(seqinr)
library(dplyr)
library(tidyr)
library(stringr)
library(data.table)
library(topicmodels)
library(tm)
library(tidytext)

# Refer to best_k.R file before this
# 10,000 observation matrix import
DNAmatrix <- read.csv('DNAMatrix_3.csv') #comment this to use test script
DTM <- DNAmatrix %>% cast_dtm(document = Documents, term = Words, value = Count)
LDAModel <- LDA(DTM, k = 30, control = list(seed = 1234))

doc_topics <- tidy(LDAModel, matrix = "gamma")
write.csv(x = doc_topics, file = "doc_topics.csv", row.names = FALSE)
doc_topics <- as.data.table(doc_topics)
doc_topics <- doc_topics[doc_topics[, .I[which.max(gamma)], by=document]$V1]

indiv_doc_topics <- list()
```

```
 for (i in 1:30){
indiv_doc_topics <- append(indiv_doc_topics, list(assign(paste0("doc_topics", as.character(i)),
filter(doc_topics, topic==i))))
 }

 doc_coordinates <- read.csv('coordinates_3.csv')
 doc_coordinates <- doc_coordinates %>% mutate(document = row_number())
 doc.list <- list(doc_topics1, doc_topics2, doc_topics3, doc_topics4, doc_topics5, doc_topics6,
doc_topics7, doc_topics8, doc_topics9, doc_topics10, doc_topics11, doc_topics12, doc_topics13,
doc_topics14, doc_topics15, doc_topics16, doc_topics17, doc_topics18, doc_topics19, doc_topics20,
doc_topics21, doc_topics22, doc_topics23, doc_topics24, doc_topics25, doc_topics26, doc_topics27,
doc_topics28, doc_topics29, doc_topics30)

 for (i in 1:length(indiv_doc_topics)) {
  doc.list[[i]]<-merge(doc.list[[i]], doc_coordinates, all.x = TRUE)
 }
 for (i in 1:length(indiv_doc_topics)) {
 write.csv(x = doc.list[[i]], file = paste0("doc_topics", as.character(i)), row.names = FALSE)
 }
```

REFERENCES

Aptekmann, A. A., Bulavka, D., Nadra, A. D., & Sánchez, I. E. (2022). Transcription factor specificity limits the number of DNA-binding motifs. *PLOS ONE*, *17*(1). https://doi.org/10.1371/journal.pone.0263307

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. *Advances in Knowledge Discovery and Data Mining*, 391–402. https://doi.org/10.1007/978-3-642-13657-3_43

Blei, D. M., Jordan, M. I., & Ng, A. Y. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

Borrayo, E., May-Canche, I., Paredes, O., Morales, J. A., Romo-Vázquez, R., & Vélez-Pérez, H. (2020). Whole-genome K-mer topic modeling associates bacterial families. *Genes*, *11*(2), 197. https://doi.org/10.3390/genes11020197

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA Model Selection. *Neurocomputing*, *72*(7-9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

Das, M. K., & Dai, H.-K. (2007). A survey of DNA motif finding algorithms. *BMC Bioinformatics*, *8*(S7). https://doi.org/10.1186/1471-2105-8-s7-s21

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Funk, C. C., Casella, A. M., Jung, S., Richards, M. A., Rodriguez, A., Shannon, P., Donovan-Maiye, R., Heavner, B., Chard, K., Xiao, Y., Glusman, G., Ertekin-Taner, N., Golde, T. E., Toga, A., Hood, L., Van Horn, J. D., Kesselman, C., Foster, I., Madduri, R., Price N. D., & Ament, S. A. (2020). Atlas of transcription factor binding sites from encode DNase hypersensitivity data across 27 tissue types. *Cell Reports*, *32*(7), 108029. https://doi.org/10.1016/j.celrep.2020.108029

Gan, J., & Qi, Y. (2021). Selection of the optimal number of topics for LDA Topic Model—taking patent policy analysis as an example. *Entropy*, *23*(10), 1301. https://doi.org/10.3390/e23101301

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*, 5228–5235. https://doi.org/10.1073/pnas.0307752101

Gross, D. S., & Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. *Annual Review of Biochemistry*, *57*(1), 159–197. https://doi.org/10.1146/annurev.bi.57.070188.001111

Guo, Y., & Gifford, D. K. (2017). Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC Genomics*, 18(1). https://doi.org/10.1186/s12864-016-3434-3

Ladunga, I. (2010). *Computational biology of transcription factor binding*. Humana Press.

La Rosa, M., Fiannaca, A., Rizzo, R., & Urso, A. (2015). Probabilistic topic modeling for the analysis and classification of genomic sequences. *BMC Bioinformatics*, *16*(S6). https://doi.org/10.1186/1471-2105-16-s6-s2

Latchman, D. S. (2005). *Gene regulation: A eukaryotic perspective*. Taylor & Francis.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Machanick, P., & Bailey, T. L. (2011). MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*, *27*(12), 1696–1697. https://doi.org/10.1093/bioinformatics/btr189

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). Great improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. https://doi.org/10.1038/nbt.1630

Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., Reynolds, A., Haugen, E., Nelson, J., Johnson, A., Frerker, M., Buckley, M Sandstrom, R., Vierstra, J., Kaul, R., & Stamatoyannopoulos, J. (2020). Index biological spectrum of human dnase I hypersensitive sites. *Nature*, 584(7820), 244–251 https://doi.org/10.1038/s41586-020-2559-3

Ng, A. Y., & Jordan, M. I. (2002). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, *14*.

Riethoven, J.-J. M. (2010). Regulatory regions in DNA: Promoters, enhancers, silencers, and insulators. *Methods in Molecular Biology*, 33–42. https://doi.org/10.1007/978-1-60761-854-6_3

Steuernagel, L., Meckbach, C., Heinrich, F., Zeidler, S., Schmitt, A. O., & Gültas, M. (2019). Computational identification of tissue-specific transcription factor cooperation in ten cattle tissues. *PLOS ONE*, *14*(5). https://doi.org/10.1371/journal.pone.0216475

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302.

Velculescu, V. E., Zhang, L., Volgelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, *270*(5235), 484–487. https://doi.org/10.1126/science.270.5235.484

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *Proceedings of the 22nd International Conference on World Wide Web*. https://doi.org/10.1145/2488388.2488514

Zhao, J., Feng, Q. P., Wu, P., Warner, J. L., Denny, J. C., & Wei, W.-Q. (2019). Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of lipoprotein(a) (LPA). *PLOS ONE*, *14*(2). https://doi.org/10.1371/journal.pone.0212112

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16*(S13). https://doi.org/10.1186/1471-2105-16-s13-s8

ACADEMIC VITA

Radha Patel (she/her/hers)

Education

Pennsylvania State University, **Schreyer Honors College**, Class of 2023 *Dean's List*

- *Bachelor of Science*, Major in Statistics, Minor in Biology
- **Senior Honors Thesis**: Topic Modeling as an Approach to Deduce Tissue-Specific Regions in the Human Genome
- **Relevant Coursework**: Data Analysis in R, Introductory Python, ANOVA, Applied Regression Analysis, Survey Sampling, Honors Advanced Technical Writing

Research

The Mahony Lab @ Penn State, *Undergraduate Researcher* | University Park, PA Aug 2020 - present

- Pioneer bioinformatic research, specifically performing topic modeling on a 10,000 subset of regulatory DNA sequences
- Communicate results to scientific communities through graphical displays, writeups, and presentations.

Work Experience

Varsity Tutors, *Tutor* | Virtual May 2021 - present

- Implement specialized AP Biology and Chemistry curriculums for high school students
- Tutor high school students from ages 16 to 18 by teaching strategies to improve their SAT scores

Penn State Health Promotion and Wellness Internship, *Intern* | University Park, PA Aug 2021 - Dec 2021

- Created wellness workshops and coordinated personalized nutrition appointments, one-on-one wellness services, alcohol counseling, etc.

Volunteer Experience

Volunteer at UPMC Shadyside, *Volunteer* | Pittsburgh, PA May 2022 - Aug 2022

- Provided emotional support and nail polish/hand massage services to pre- and post-op patients
- Assisted nurses with filing patient's identifying information and medical history

Arts in Health Initiative, *Feature Post Coordinator* | University Park, PA Feb 2020 - Sept 2021

- Designed Instagram posts that feature student and professional artists to educate individuals on art's healing power and vitality in medical settings

Days for Girls, *Volunteer* | University Park, PA Oct 2019 - May 2020

- Alleviated barriers to quality menstrual care by sewing and globally distributing over 60 reusable, menstrual bags

Leadership

Schreyer for Women, *Career Development Chair ('21-'22), Treasurer ('22-'23)* | University Park, PA Sept 2019 - present

- Hosts creative and restaurant fundraisers to support and manage SfW's finances throughout the year
- Connect female alumni and honors students through large-scale networking events that highlight many professions

South Asian Student Association, *Treasurer ('21-'22), Senior Adviser ('22-'23)* | University Park, PA Oct 2019 - May 2022

- Coordinate events such as Garba, Diwali, and Holi to promote Indian culture and revitalize cultural traditions at Penn State

Awards

Student Leadership Scholarship Dec 2022

- Received $500.00 for outstanding leadership in student organizations at Penn State

Travel Grant to Masai Mara, Kenya Nov 2022

- Granted $1,500.00 to conduct ecological fieldwork in the Masai Mara

Emerson Electric Engineering College Scholarship May 2019

- Awarded $10,000.00 for excellence in scholarship, service, and leadership

Skills

- Advanced in R for Data Analysis and all Google Applications
- Proficient in Microsoft Applications
- Basic Skills in Python, Visual Basic, HTML, and JavaScript
- Certified in Mental Health First Aid