EXAMINATION AND CHARACTERIZATION OF CANCER RISK VARIANTS

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

James Elliot Hayes

February 2016

EXAMINATION AND CHARACTERIZATION OF CANCER RISK VARIANTS

James Elliot Hayes, Ph.D.

Cornell University, 2016

Genome-wide association studies (GWAS) have been a useful tool in identifying numerous genetic loci that associate with increased risk for numerous cancers. However, as most of the identified risk variants are found in non-coding regions of the genome, the field has been slow in moving beyond identifying the risk variants to functionally determining the mechanism for cancer predisposition. The ultimate goal of genome-wide association studies is to identify and understand the role of these loci disease etiology to ultimately enable more effective screening and therapeutic treatments.

Framed by better understanding GWAS results, my dissertation has 3 main aspects, where I: 1) developed a computational approach to characterize GWAS results using publically-available epigenomic databases, 2) identified novel germline susceptibility loci for myeloproliferative neoplasms, and 3) examined molecular mechanisms by which a prostate cancer single-nucleotide polymorphism may increase risk.

The scientific community has invested great resources into discovering and cataloguing all the functional elements of the genome, through efforts such as ENCODE and Roadmap Epigenomics. First, we developed a computational method, “Understanding Enrichment through Simulation” (UES), which combines GWAS data with these consortia data in order to provide a better understanding of the role of risk-SNPs in cancer predisposition. We validated the approach using a set of lymphoma SNPs and successfully determined that the risk-loci are preferentially found in regulatory elements in lymphoid tissues, suggesting a tissue-specific disruption of regulation may cause lymphomagenesis.

Next we conducted a GWAS of myeloproliferative neoplasms (MPN), a collection of clonal, hematopoietic disorders. We combined multiple MPN GWAS into a larger dataset in order to increase the statistical power needed to identify novel risk-variants. We two risk loci for MPN; rediscovering the *JAK2* locus and identifying a novel association at the *TERT* gene.

Lastly, we studied the mechanism by which microseminoprotein-beta controls prostate cell growth. The risk allele of a prostate cancer risk SNP, rs10993994, had been shown to associate with lower levels of genic transcriptional activity and protein levels in humans. Levels of β-MSP have been shown to lead to decreased prostate cell viability, though our efforts were unable to determine the mechanism explaining this effect.

BIOGRAPHICAL SKETCH

James Elliot Hayes was born in December of 1985 in Livingston, NJ, to his father Robert James Hayes and mother Judy Diane Hayes. He is the middle of three children, having an older sister, Erin Rae, and younger brother, Jonathan Calvin. At the age of 2, he moved to Hawthorne, NJ, where he was raised and where his family still currently resides. After graduating from Hawthorne High School in 2004, James attended The College of New Jersey where he graduated as a Bachelor of Science in Biology and a computer science minor in 2008. As a junior, he began working in the lab of Dr. Sudhir Nayak, where he performed both wet-lab research on *Caenorhabditis elegans* and developed software to detect evolutionary selection on protein coding sequences.

Following graduation, James enrolled in Weill Cornell Graduate School of Medical Sciences in New York, NY, as a Ph.D. student in the Biochemistry, Cellular, and Molecular Biology Allied Program. In 2009, James joined the lab of Dr. Robert J. Klein at Memorial Sloan Kettering Cancer Center where he continued performing both wet- and dry-lab research. In August of 2014 when his thesis lab relocated The Mount Sinai Hospital, James joined the lab of Dr. Christina Leslie at MSKCC and continued his research at both MSKCC and MSSM. His thesis work was focused on characterizing cancer risk variants using both molecular and computational techniques.

His wife, Jaime Lynn Hayes, also a graduate of the 2008 class from The College of New Jersey, is a nurse in the Emergency Department at New York-Presbyterian/Weill Cornell Medical Center. They currently have one son, Weston James, and together, the family eagerly awaits the arrival of their second child in May, 2016.

DEDICATION

For my loving wife and our son, my parents, and my siblings - thank you for your continual love, support, and encouragement.

ACKNOWLEDGEMENTS

First, I would like to express my thanks and gratitude to my mentor, Dr. Robert Klein, for his continual guidance and encouragement during my research. I could not have asked for a more supportive advisor who cared so deeply about my growth, both academically and personally. I will be forever grateful for the time I spent in his lab. Additionally, I am thankful for the other members of the lab with whom I worked closely: Xiaoni Gao, Heriberto Moran, Dr. Semanti Mukherjee, Dr. Jason Willis, Dr. Xing Wang, and Dr. Xing (Dandan) Xu. Thank you for your continual help and conversation, ranging from the academically stimulating to the inane.

I am also thankful for the support and guidance from my Thesis Special Committee members: Dr. Ekta Khurana, Dr. Mark Rubin, Dr. Andrea Ventura, and specifically, Dr. Christina Leslie, who went above and beyond in helping me during our lab's relocation. I gratefully acknowledge the grant support, specifically U01 HG007033, which made a bulk of this research possible. Additionally, I am indebted to Dr. Sudhir Nayak, my undergraduate research advisor at The College of New Jersey, who encouraged me to continue pursuing the field of genetics and genomics.

I'd like to thank God for giving me the wisdom, guidance, and perserverance to complete this degree. I am also eternally grateful for the brothers and sisters in Christ that He inserted into my life at Redeemer Presbyterian Church.

Lastly, I'd like to thank my family. Thank you for your unconditional love and support, both during graduate school and throughout my life. Thank you for the role you played in shaping who I have become today. To my wife, Jaime: thank you for always being there for me. I can't imagine living in New York, let alone my life, without you. And once again, I'm so sorry that I totally underestimated the amount of time this program would take when I brought you to the city. But it sure was fun.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AP	active promoter
ATCC	American Type Culture Colelction
bp	base pairs
β–MSP	microseminoprotein-beta, the protein encoded for by *MSMB*.
CD-CV	common disease-common variant
ChIP-Seq	chromatin immunoprecipitation sequencing
CI	confidence interval
CLL	chronic lymphocytic leukemia
DHS	Deoxyribonuclease I (DNase I) hypersensitivity site
DNA	Deoxyribonucleic acid
ENCODE	Encyclopedia of DNA elements
eQTL	expression quantitative trait loci
ESC	embryonic stem cell
EtOH	ethanol
ET	essential thrombocythemia
FBS	fetal bovine serum
g	grams
GTEx	genotype-tissue expression
GWAS	genome-wide association Study
HD	Hodgkin’s lymphoma
HL	Hodgkin’s lymphoma
HLA	human leukocyte antigen
IBD	inflammatory bowel disease
Kb	kilobases
KSFM	Keratinocyte serum-free medium
L	liter
LCL	lymphoblastoid cell line
LD	linkage disequilibrium
MAF	minor allele frequency
Mb	megabases
mg	microgram
ml	maximum likelihood, or milliliter
mM	millimolar
MPN	myeloproliferative neoplasm
MSKCC	Memorial Sloan Kettering Cancer Center
NHGRI	National Human Genome Research Institute
NHL	non-Hodgkin’s lymphoma
OR	odds ratio
PCA	principle component analysis
PMF	primary myelofibrosis
PrCa	prostate cancer
PV	polycythemia vera
qu	quartile
RFLP	restriction fragment length polymorphism

SA-β-gal	senescence-associated beta-galactosidase
SE	strong enhancer
SNP	single nucleotide polymorphism
TF	transcription factor
TFBS	transcription factor binding site
TSS	transcription start site
UES	Uncovering Enrichment through Simulation method
YAFSNP	“yet another functional single nucleotide polymorphism”

1 Introduction

On April 14, 2003, after a decade of work and an estimated $2.7 billion, the successful completion of the Human Genome Project was announced to great fanfare. The field has since progressed from just determining the sequence of an individual's genome to creating a complete catalogue the genetic variation among populations of people. One outgrowth of having access to such genomic data are phenotypic-based genome-wide association studies (GWAS) which have successfully identified hundreds of genetic loci that contribute to a myriad of traits and diseases. Thus, the logical expansion of these studies is to next determine the underlying mechanism by which these variants lead to the phenotypic outcome. This thesis describes my efforts, both computationally and through bench experiments, to better understand and characterize risk-variants for various cancer types. The thesis work presented herein is divided into 3 distinct chapters and outlined below:

1. **Creating a novel algorithm, "Understanding Enrichment through Simulation (UES)", to analyze GWAS results with publically available databases.** Chapter 2 presents the novel method, UES, which employs a SNP-matching technique to determine statistical enrichment of previously-reported GWAS SNPs in various ENCODE and Roadmap Epigenomic tracks. The method was validated using a set of lymphoma SNPs and further expanded to look at different cancer types.

2. **Identifying germline variants associated with myeloproliferative neoplasm (MPN) risk**. Chapter 3 describes our efforts to combine multiple, smaller MPN datasets with the goal of increasing our statistical power to identify novel associations.

3. **Functionally validating a prostate cancer risk SNP.** Chapter 4 describes the use of multiple molecular biology techniques in order to elucidate the mechanism by which a single SNP increases risk for prostate cancer.

Taken together, this work presents multiple techniques, both computational and molecular, that can be employed to better understand how common variants increase risk in various cancers. In the remainder of this chapter, I present a brief history of the field of human genetics and how this shaped the work and techniques in the present day. Additionally, I introduce consortia data which was integral to providing the context in which the cancer risk-SNPs function. Lastly, I provide introductions to the various cancer types presented in this thesis.

1.1 Genetics of Human Diseases

The history of the of field of genetics began with the work of a friar, Gregor Johann Mendel, and his experiments with pea plants in the monastery's garden, where he observed consistent, reproducible, and predictable ratios of phenotypic traits upon crossbreeding the plants. These observations, which are now called the laws of "Mendelian inheritance," were published in "Experiments in Plant Hybridization" and are the first description of the rules of heredity: 1) traits are determined by discrete "units" or "factors" (now called genes), 2) individuals inherit a unit from each parent, and 3) even though a trait may not be seen in a generation, it can still be passed on to offspring. These rules would become instrumental in allowing the mapping of individual traits to physical locations within a genome (Altshuler et al. 2008).

The predictable, probabilistic outcomes from genetic crosses suggested that there was a physical mode for transmission of traits from parent to offspring, though it was not known whether the transmission of traits came via DNA or proteins. Following a deadly flu outbreak in the early 1900s, Frederick Griffith was studying two strains *Streptococcus pneumoniae* that differed in both appearance and virulence in mice and was able to identify that there was a heat-resistant "transforming principle" which was able to transform the less virulent "R cells" to the more virulent

“S cells” (Griffith 1928). Building on those results, Oswald Avery and colleagues at the Rockefeller University, treated the extract from the heat-killed, virulent “S” strain of *S. pneumoniae* with protease, RNAse, and/or DNAse and, through process of elimination, were able to identify DNA as the transforming agent (Avery et al. 1944). Subsequently, Alfred Hersey and Martha Chase, through experimentation with radio-labeled T2 bacteriophages were able to exclude proteins as heredity material by showing that ^{32}P-labeled DNA was found in infected cells while ^{35}S-labeled proteins were not (Hershey & Chase 1952). DNA was thus established as the physical vector by which genetic information was inherited. Numerous methods were later established to determine the map these phenotypic “units,” or genes, to a precise location of DNA.

1.1.1 Genetic Mapping and Linkage Studies

Linkage studies are the simplest form of performing genetic mapping, which take advantage of the tendency for genetic loci physically close to each other in the genome to be passed along with one another during meiosis. This method was first performed in *Drosophila melanogaster*, the common fruit fly, by setting up mating crosses between parent flies which varied at a Mendelian trait of interest and “markers” or known genetic variants. The resultant progeny were observed for evidence of whether or not the traits were “linked,” or showed correlated segregation, to any of the markers (Sturtevant 1913). It would be years until the technique of positional cloning would be developed, first in *Saccharomyces cerevisiae*, allowing specific genes to be connected with particular traits based on genomic position (Clarke & Carbon 1980). This method proved to be quite powerful in numerous model systems and expanded in use beyond common baker’s yeast, as evidenced when Bender and colleagues used a similar approach back in *D. melanogaster* and developed a genetic map of the bithorax complex (Bender et al. 1983).

It was first proposed that naturally occurring polymorphisms in DNA which disrupted restriction fragment lengths (RFLPs) could be used as genetic markers in humans in order to create a genetic map (Botstein et al. 1980). The first successful example of mapping a to a specific chromosome was demonstrated when the gene causing Huntington disease was the mapped to the short arm of chromosome 4 (Gusella et al. 1983). Even though the study provided a broad locus for Huntington disease, the study showed that linkage analysis was indeed possible to perform in humans for Mendelian disease genes. Since then, linkage studies have identified numerous genetic variants with large effect sizes in other Mendelian diseases, including cancers. One such example of the power of this approach is demonstrated by identifying the *BRCA1* and *BRCA2* familiar risk loci in breast cancer (Hall et al. 1990; Miki et al. 1994; Wooster et al. 1994).

While linkage studies had proven to be quite powerful in model organisms, similar techniques proved to be both difficult and impractical to translate to human genetic research. Firstly, there were not nearly enough genetic markers that to make these types of studies possible, especially when looking at a genomic level. Furthermore, even if those markers were available, human family sizes are too small provide precise results and, more importantly, scientists would be impeded by the ethical issues of controlling the mating of human subjects to design the needed crosses (Altshuler et al. 2008). Nevertheless, while there have been successful linkage studies as previously mentioned, linkage studies are more useful in identifying highly penetrative, rare Mendelian disorders and are not practical to be applied towards studying more common, complex diseases.

The disease risk variants identified by linkage studies were typically highly penetrant and rare within a population (<1% frequency within a population).

However, this framework that only rare-alleles are risk-alleles does not hold together logically for the entire spectrum of human disease. As most Mendelian disorders manifest as strongly damaging phenotypes, natural selection would, over time, cause these deleterious variants to be lost to purifying selection and thus removed from the population (Pritchard & Cox 2002). The exception to this would be in cases where balancing selection keeps the Mendelian risk variant in the population due to some beneficial phenotype, notably as seen in sickle-cell anemia and malaria resistance.

One hypothesis about the common diseases is that they have a different genetic architecture than rare disorders. This hypothesis was supported by the multiple discoveries that risk variants for common diseases, for example Alzheimer's and type II diabetes, had a rather high minor allele frequency (MAF), 13.7% for the *APOE* ε4 risk allele (homozygous odds ratio (OR) = 14.8; 95% confidence interval (CI) = 10.8-20.6) and 16% for the missense mutation in *PPARG* (Pro12Ala) (OR=0.78; 95% CI=0.59-1.05) (Corder et al. 1993; Farrer et al.; Altshuler et al. 2000). These studies eventually led to formulation of the "common disease-common variant" (CD-CV) hypothesis (Reich & Lander 2001), which simply stated, suggests that common diseases are likely affected by genetic variation which is more common in populations though they have much smaller effect than that of highly penetrant, rare variants. Thus, the resultant corollary states that these common variants associated with diseases are not highly penetrant and, since common diseases show heritability, these common variants play a role in susceptibility. This model, particularly in cancer was supported by a meta-analysis of twin- and family-based cancer studies which showed that susceptibility to cancer was individual rare variants had a strong effect whereas multiple common variants showed a more modest effect (Risch 2001). This phenomenon, where allelic frequency is generally inversely proportional to its effect size for a disease, is generally referred to as the allelic spectrum of disease (Figure 1.1).

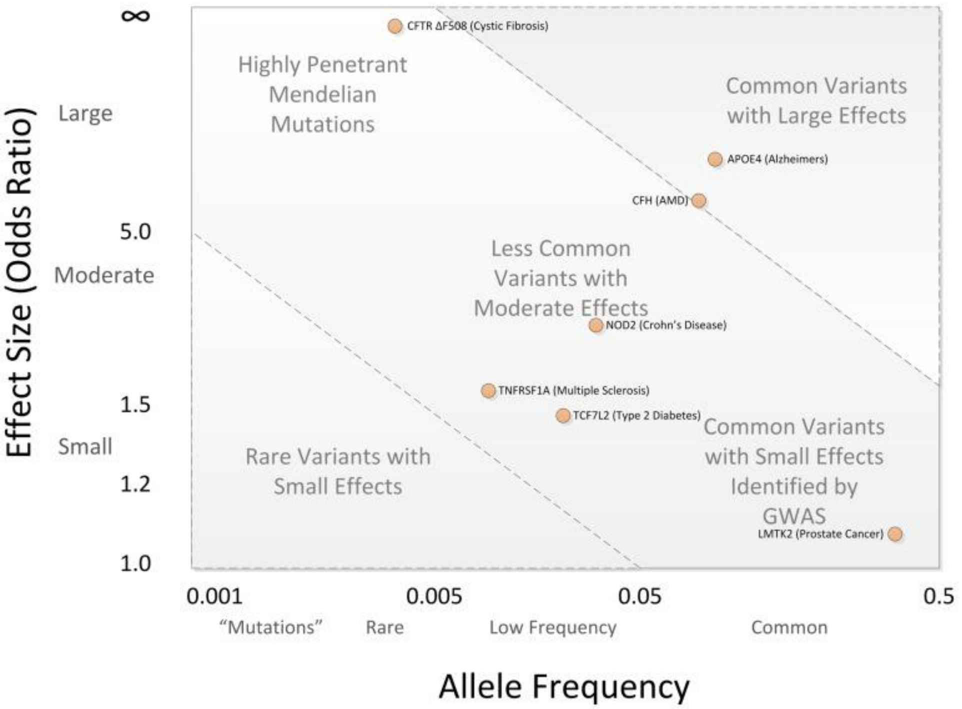


Figure 1.1. Allelic spectrum of disease. Highly penetrant, rare variants are typically have a lower frequency within a population. Conversely, common variants usually show weaker effect (Bush & Moore 2012).

1.1.2 Genome-wide association studies

As linkage studies proved ineffective in discovering common variants with weaker effects, an alternative approach was proposed to test for the association of large numbers of variants across the genome for a particular phenotype or disease of interest (Risch & Merikangas 1996). This new paradigm assumes that the risk-variants had been genotyped and could be tested for association within a population, as opposed to looking at the smaller number of both affected and family members as done in linkage studies.

These studies were enabled by the completion of the Human Genome Project and the International HapMap Project (Collins et al. 2003; Consortium 2003), specifically by the effort of the HapMap Consortium to catalogue the variation of single-nucleotide polymorphisms found within the human genome across multiple populations. These single-nucleotide polymorphisms, or SNPs, are the most common type of variation and consist of the change of an individual nucleotide (A, G, C, or T) to another. The consortium studied 4 geographically-distinct populations: 1) 30 trios (consisting of two parents and a child) from Idaban, Nigeria (abbreviated YRI), 2) 30 trios of Utah residents of northern and western European ancestry obtained from the Centre dEtude du Polymorphisme Humain (CEU), 3) 45 unrelated Han Chinese individuals from Bejing, China (CHB), and 4) 45 unrelated Japanese individuals from Tokyo, Japan (JPN). In 2005, the consortium announce the findings of >1.3 million SNPs that have a minor-allele frequency (MAF) of >1% within a population (The International HapMap Consortium 2005). Since then, the 1000 Genomes Project has continued to expand the effort catalogue human genetic variation through sequencing (The 1000 Genomes Project Consortium et al. 2010; The 1000 Genomes Project Consortium et al. 2012).

It had been observed that SNPs are subject to the phenomenon of linkage disequilibrium (LD), or the non-random association between two alleles of neighboring loci, and cause SNPs to be carried within haplotype blocks (Devlin & Risch 1995; Pritchard & Przeworski 2001). This occurs when two variants are physically located in close proximity to one another and are not typically separated from one another during meiotic recombination. Thus, when these alleles fail to disassociate from one other and tend to be inherited together, these loci are considered "linked." One of the most common measurements of LD is r^2, ranging from 0-1 (no correlation to perfect correlation, respectively). Linkage disequilibrium, in essence, allowed scientists allowed scientists indirectly study a larger swath of genetic variation of proxy, or "tag" SNPs across the genome by performing the association testing on the “lead” SNPs. Combining this knowledge of LD structure with the advancements in SNP genotyping through deoxyribonucleic acid (DNA) microarrays allowed for an appropriate number of SNPs to cover the genome and subsequently usher in the era of genome-wide association studies (GWAS).

A GWAS is a population-based study that typically consists of a two-stages: discovery and validation (Spencer et al. 2009; Klein 2007). For the discovery stage, a large cohort of individuals are chosen and categorized according to whether or not they have the particular phenotype/disease of interest (cases) and those without (controls). All of the individuals from both cases and controls are then genotyped using SNP-arrays, typically using those available from Affymetrix or Illumina, and then, following stringent quality controls, the SNPs are tested for an association with the phenotype as seen by a difference in allelic frequency between cases and controls. Logistic regression is typically used to estimate the log-additive effect of each allele on the odds of disease, or the odds ratio (OR), and its statistical significance is reported by the p-value. However, since numerous tests are performed, a true association must overcome the burden of multiple testing. For example, if we set the significance threshold at $p<0.05$ and test 1 million SNPs in a GWAS, then we would expect to have 50,000 SNPs that met that significance threshold due to nothing more than random chance. Thus, in order to reduce type-I errors, GWAS SNPs usually need to meet the stringent Bonferoni correction level (calculated by 0.05 divided by the number of tests) in order to be deemed a true association. The validation stage of the GWAS is then performed using a smaller subset of the significant SNPs and tested for association in an independent cohort. The first successful GWAS was published on age-related macular degeneration (Klein et al. 2005), and as of November 2015, this approach has been used in 2,305 different studies, ranging from non-disease phenotypes to cancer (Welter et al. 2014).

Most of the identified risk-variants identified from GWAS have a small effect size with ORs typically found in the range of 1.2-1.4, which is in line with the CV-CD hypothesis that these common alleles will have a relatively low effect size (Hindorff et al. 2009). Albeit, these effect sizes are smaller than was hoped for at the beginning of the GWAS era, the use of genome-wide association studies is generally regarded as successful, especially in the field of cancer predisposition (Klein et al. 2010). Though, there are criticisms of the field that remain to be addressed, specifically missing heritability and functional validation. A study 2009 study of height demonstrated that while the heritability of height is near 80%, only 5% of that heritability was explained by the 40 height-associate loci (Manolio et al. 2009). One such explanation of this phenomenon is that the aforementioned Bonferoni-correction method is too stringent and excludes true positive associations. Furthermore, some of this missing heritability could also be may also come from structural variation, synthetic associations, or gene-gene interactions; rare variants unable to be discovered by GWAS may also have a role in this missing heritability, though the tools and platforms used in this technique may make it difficult to detect these risk loci. However, the more severe (and well warranted) critique of the field is the lack of functionally validating the identified variants and determining the mechanism by which they are involved in the etiology of diseases. Analysis by Hindorff et al. (2009) additionally revealed that 88% of the risk-SNPs reported at the time were found in either intergenic or intronic and were not responsible for amino acid changes. Intuitive mechanistic understanding of risk SNPs are further complicated since many are found in areas of the genome devoid of genes, or “gene deserts” (Freedman et al. 2011). A majority of my thesis work focuses on addressing this latter critique of the GWAS field; chapter 2 of describes a computation approach to integrate GWAS results with epigenomic databases and chapter 4 presents molecular-biology driven effort to understand how a particular prostate cancer risk-SNP may be responsible for increased prostate cancer risk.

1.2 Functional Genomic Databases

Once the Human Genome Project was completed and scientist had access sequence, it became abundantly clear that the sequence alone did not fully explain the complexity of human biology, as evidenced by fact that only ~1.5% of the 3 billion nucleotides were responsible for coding the ~20,000 human proteins (Lander et al. 2001). Thus, researchers expanded beyond the genome to study the "epigenome," or how the *in vivo* packaging of DNA. This epigenomic landscape includes a multitude of marks, including histone modifications and positioning, distal chromatin interactions, and DNA-binding proteins including transcription factors. There are two

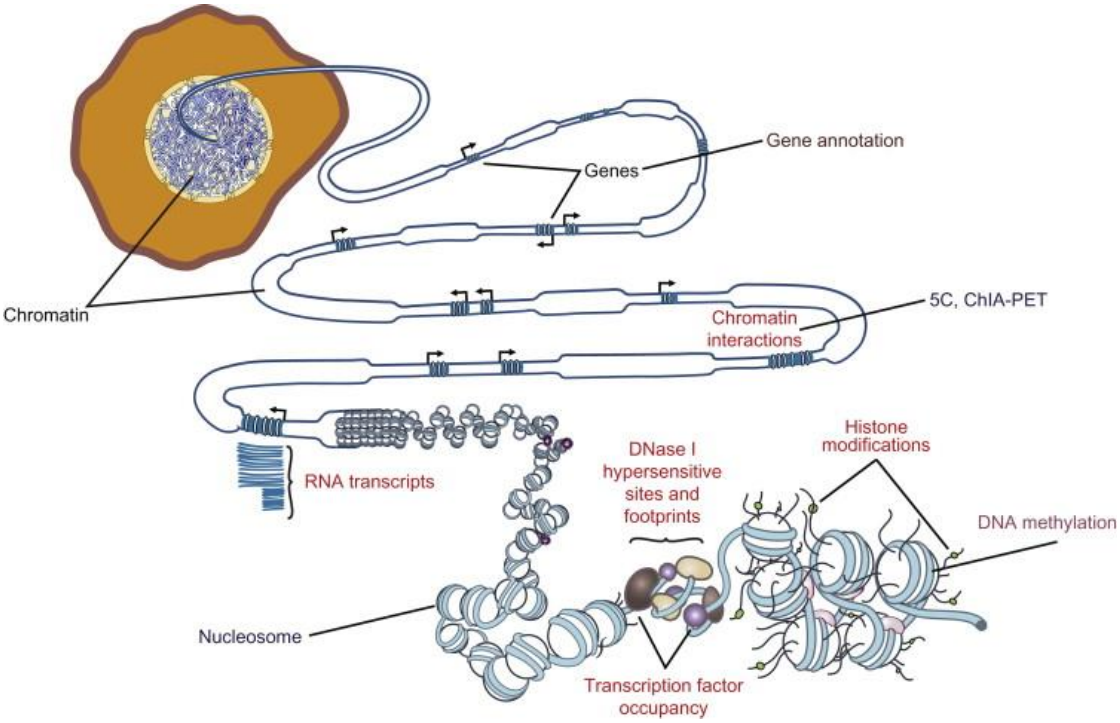


Figure 1.2. Generalized overview of datatypes for the ENCODE Project and Roadmap Epigenomics. This figure provides a visualization of the packaging of DNA, spanning from a broad view at the chromatin level in the upper left to a high-resolution visualized down to the nucleosome. Genomic features and particular assays labeled (Ginsburg et al. 2013).

major consortia that are currently working in parallel to catalog with the goal of ultimately understanding how these differences affect human biology. An overview of the data available from both consortia is visualized in Figure 1.2.

1.2.1 The ENCODE Project

Upon completion of the Human Genome Project in 2003, the National Human Genome Research Institute (NHGRI) launched the ENCODE Project Consortium (ENCODE Project Consortium et al. 2007). This consortium was initially charged identifying all of the functional elements within the human genome, though the pilot study would focus on only 30 megabases (Mb), or just 1%. Half of the loci included in the initial study were of well-characterized regions, such as the *HOXA* and *CFTR* loci, and the other 15 Mb consisted of other regions with varying amounts of gene density and conserved, non-coding regions (Ginsburg et al. 2013). Initially, the ENCODE project focused on using numerous assays to deeply study 3 individual "Tier 1" cell lines: K562 (a erythroleukemia cell line), GM12878 (an Epstein Barr virus-immortalized lymphoblastoid cell line), and H1-hESC (a human embryonic stem cell line). The effort has since expanded to provide data on >200 primary, stem cell, or cancer cell lines.

1.2.2 The NIH Roadmap Epigenomics

Soon afterwards, in 2008, the NIH started a similar effort called the Roadmap Epigenomics Mapping Consortium (referred to throughout as "Roadmap Epigenomics" or simply as "Roadmap"), whose goal was to generate reference epigenomes for normal human cell types, including both adult and fetal tissues. As such, the Roadmap Epigenomics Consortium does not use disease cell lines nor immortalized lines; the cell lines were either obtained as primary cells, generated from tissues, cultured embryonic stem cells (ESC), or differentiated from ESCs. Furthermore, as Roadmap's goal is to create a reference epigenome for these cells, consortium members focused on collecting the same multiple datasets form the same tissues, rather than performing deep assays in relatively few cell lines as done in ENCODE (Bernstein et al. 2010).

1.2.3 Functional Epigenomic Data

Both the ENCODE and Roadmap Consortium performed RNA-seq RNA paired-end tag sequencing (RNA-PET) (Z. Wang et al. 2009). ENCODE specifically has made an effort to fully catalog and annotate the all the genes which encode proteins, though this process has proven difficult to automate computationally (Harrow et al. 2006; Guigó et al. 2006). As such, manual curation constitutes a large amount of the work performed by ENCODE to curate genes. Interestingly, ~50% of transcripts found by ENCODE appear to be either non-coding or pseudogenes, including intronic transcripts and transcripts with opposite polarities (The ENCODE Project Consortium 2012; Park et al. 2012; Djebali et al. 2012). The consortium is also interested in annotating structure of the ribonucleic acid (RNA), such as transcription start and end sites as measured by analyzing the 5’ methyl caps of RNA (CAGE) and determining RNA length by sequencing the amplified 5’/3’ ends of RNA (RACE) (Shiraki et al. 2003; Frohman et al. 1988).

Both consortia employed various sequencing-based assays to extensively catalogue the epigenomic landscape of DNA. These assays included DNase1-seq which provides a comprehensive look at accessible loci and a mark of active regulatory DNA (Boyle et al. 2008; Thurman et al. 2012) and chromatin immunoprecipitation followed by sequencing (ChIP-seq) for various DNA-binding proteins, such as transcription factors (TFs) and RNA Polymerase II, ChIP-seq was also used to catalogue the histone modifications within the cell lines (The ENCODE Project Consortium 2012). ENCODE additionally performed “formaldehyde-assisted isolation of regulatory elements sequencing”, or FAIRE-seq, to identify loci with low occupancy of nucleosomes (Giresi et al. 2007). The true power of these type data can be seen when the TF occupancy data was combined with chromatin accessibility data: a plethora of histone modifications (H3K4me1, H3K4me2, H3K4me3, H3Ac, and H4Ac) were shown to associate with transcription start sites (TSS), both previously known and novel sites (Thurman et al. 2012). Deep-sequencing of these DNase-hypersensitivity data have also allowed researchers to identify TF binding motifs within the data, providing both a validation of the ChIP-seq data and potentially identifying novel TF binding sites (Hesselberth et al. 2009; Neph et al. 2012).

One of the largest differences between the datasets of the consortia is that ENCODE attempts to catalogue distal interaction that genetic loci has with different other areas of the genome whereas Roadmap does not address this question. It has been shown that DNA can interact with other genomic loci hundreds of Kb away or even located on entirely separate chromosomes (Lieberman-Aiden et al. 2009). The hypothesis is that these distant interactions indicate that one area of the genome act as a regulator of another locus. These interactions are primarily measured through sequencing carbon copy chromatin conformation capture (5C-seq) and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (Dostie et al. 2006; Sanyal et al. 2012; Fullwood et al. 2009).

The real power of all these data when they can be combined to provide greater insight than when analyzed separately. Two different segmentation algorithms have been developed by members of the ENCODE consortium which, using a combination of epigenomic marks, categorize the state of every locus of the genome (Hoffman et al. 2013). The first such method was ChromHMM (Ernst & Kellis 2012) which employs a multivariate Hidden Markov Model to analyze combinations of chromatin marks in order to determine genomic state at 200 bp intervals. Using this technique, a 15-state models were generated and applied to 9 different ENCODE and 127

Roadmap Epigenomics cell lines (Ernst et al. 2011). A few months later, Bill Noble's group at the University of Washington released Segway, another segmentation algorithm (Hoffman et al. 2012). Segway employs a Dynamic Bayesian Network, rather than a Hidden Markov Model, to segment the genome. Using a combination of ChIP-seq and chromatin accessibility data, Segway data tracks with 13 different states were provided for 6 different ENCODE cell types. While both algorithms perform similar analysis and there is high concordance between the outputs, the main difference between the algorithms is the resolution: ChromHMM operates at a 200 bp window whereas Segway segments data at the base pair level. However, this increased resolution also amplifies the output files and increases the time needed to analyze the data.

1.3 Lymphoma and chronic lymphocytic leukemia

Lymphoma is a broad categorization of cancers that develop in the lymphatic cells of the immune system and consists of two broad categories: Hodgkin's lymphoma (HD), as distinguished by the presence of a Reed-Sternberg cell, and non-Hodgkin's lymphoma (NHL). Lymphoma including its various subtypes are the most common type of blood cancer in the United States, being projected to cause 80,900 new cases and 20,940 deaths in 2015, though the new NHL cases (71,850) vastly outnumber the HD cases (9,050) (Siegel et al. 2015). When looking at categories of lymphoma, the median age of onset is 65 years old, however, the median ages are quite disparate when comparing Hodgkin's lymphoma to the non-Hodgkin's subtype, 38 vs 65, respectively. A similar, though smaller, gap is seen between the median ages of death caused by lymphoma when comparing HD and NHL (65 vs 76, respectively) with the median age of death from lymphoma collectively is 75 years old (Howlader et al. 2015).

Patients with lymphoma typically present with lymphadenopathy which can co-occur with additional swelling of the spleen and/or liver as well as night sweats, fever, and weight loss (Pileri et al. 2002). Since the clinical presentation of lymphomas share some commonalities across the various subsets of the disease, a pathological approach is required for a more accurate diagnosis as treatment plans can vary among the different subtypes (Turner et al. 2010). Hodgkin's lymphoma, as previously mentioned, is marked by the presence of Reed-Sternberg cells which appear as bi-nucleated, CD30 positive/CD15 positive/CD45 that are thought to be defective germinal centers (Hartlapp et al. 2009). Advances in treatment have brought the mortality of HD down to the present-day level for 5-year survival to 86% (Howlader et al. 2015). NHL has a slightly lower 5-year survival rate at 70%.

1.3.1 Genetic predisposition to lymphomas

Lymphoma, including the non-Hodgkin's subtype chronic lymphocytic leukemia (CLL), has been shown to be highly heritable. A population-based study of Scandinavian populations revealed that for individuals whom had relatives afflicted with HD, the relative risk increased 3.47-fold (95% CI = 1.77-6.8) in a Swedish population and 2.55-fold (95% CI = 1.01–6.45) in a Danish population (Goldin et al. 2004). A literature review revealed that familial relative risk of CLL was also elevated, ranging from a 1.5-7.5 fold increase (Sellick et al. 2006). Since then, numerous genome-wide association studies have been performed across the subtypes of lymphoma, in order to better understand the mechanism lymphomagenesis (Berndt et al. 2013; Conde et al. 2010; Cozen et al. 2012; Crowther-Swanepoel et al. 2010; Di Bernardo et al. 2008; Enciso-Mora et al. 2010; Kumar et al. 2011; Skibola et al. 2009; Slager et al. 2011; Smedby et al. 2011; Urayama et al. 2012; Vijai et al. 2013; Wade et al. 2011).

Numerous SNPs at the human leukocyte antigen (HLA) region at 6p21 have been identified as associating with lymphoma-risk (Skibola et al. 2009; Conde et al. 2010; Cozen et al. 2012; Smedby et al. 2011; Vijai et al. 2013). On explanation is that these variants may cause dysregulation of the HLA complex and ultimately disrupt B-cell development. Vijai et al. (2013) identified a neighboring, yet still novel risk-SNP, rs707824, at 6p23. This SNPs is located upstream of the gene encoding Jumanji, *JARID,* which has been shown to play a role in the both the differentiation and self-renewal of embryonic stem cells (Shen et al. 2009). Furthermore, this SNP is located downstream of CD83, a known B-cell activation protein (Cao et al. 2005). However, the HLA locus is difficult to study due to the high amount of recombination and little functional evidence confirming the mechanism of increased risk is available for this region.

1.4 Myeloproliferative Neoplasms

Myeloproliferative neoplasms (MPN), officially defined in 2008 by the World Health Organization, are a collection of blood disorders of the myeloid lineage that manifest in the clonal expansion of hematopoietic cells (Ayalew Tefferi et al. 2009). Polycythemia vera (PV), essential thrombocythemia (ET), and primary myelofibrosis (PMF) together make up the set of *BCR-ABL* negative MPNs (Tefferi 2010). Patients afflicted with PV are marked by an increased proliferation of erythroid cells, leading to elevated erythroid cell mass, blood viscosity, hematocrit value, and hemoglobin concentration which, in turn, puts the patient at increased risk hemorrhages, thromboses, and stroke (Dameshek 1951; Ruggeri et al. 2003). Patients with ET are also at risk of thrombosis and excessive bleeding due to an over-proliferation of platelets and megakaryocytes (Tefferi & Murphy 2001). PMF is the most lethal of these three MPNs as evidenced by the high risk of transformation to leukemia (5%-

30%) and is characterized by bone marrow fibrosis and a clinical course of splenomegaly and anemia (Varki et al. 1983; Tefferi 2000; Barosi 1999).

1.4.1 Genetic Predisposition to MPN

Myeloproliferative neoplasms have been observed in familial cluster, suggesting that there is heritable, germline component of the disease's etiology. Studies have shown that between 5-10% of patients afflicted with MPN have a positive family history of the disease; furthermore first-degree relatives are at a higher disease-risk when compared to the normal population: PV (relative risk = 5.7, 95% CI = 3.5-9.1) and ET (relative risk = 7.4, 95% CI = 3.7-14.8). (Landgren et al. 2008). Interestingly, these familial clusters of MPN are also characterized by both genetic and clinical heterogeneity, where individuals within the same cluster are diagnosed with different forms of MPN from one another (Rumi et al. 2007).

Even though MPN is observed to be clustered familial, there have been no linkage studies that have identified the particular germline risk variant. Several genome-wide association studies were conducted in order to identify the heritable component of MPN risk and have successfully identified the *JAK2* locus (found at chromosome 9p) as associating with *JAK2-V617F* positive MPN and explain between 28%-46% of the risk (Kilpivaara et al. 2009; Olcaydu et al. 2009; Jones et al. 2009). Somatic mutations at the *JAK2* locus have been in all three types of *BCR-ABL* negative MPN, found in approximately 96% in PV cases, 55% in ET cases, and 65% in PMF cases (James et al. 2005). This mutation causes *JAK2* to become constitutively active, which in turn, causes a downstream activation of STAT proteins and both the PI3K-AKT and MAP kinase pathways (Ihle & Gilliland 2007; Delhommeau et al. 2007; Jamieson et al. 2006). When a more-mutated molecular haplotype of *JAK2-V617F* was shown to be in *cis* with the risk haplotype, one possible explanatory link was that the risk-variant caused a somatic hypermutability phenotype.

However, a resequencing effort from our group did not find a difference in the mutational load at the V617F locus when comparing 24 MPN cases, 12 homozygous for the G allele at the rs10974944 and 12 homozygous for the C allele, suggesting that neither haplotype gained mutations at a rate quicker than expected (Mukherjee 2011). In chapter 3 of this thesis, I describe a new GWAS that we performed by combining multiple, smaller MPN datasets into a larger set with the goal of increasing power to discover novel associations.

1.5 Prostate Cancer

Prostate cancer (PrCa) is the most common type of cancer among American males. In 2015 alone, there is projected to be 220,800 new cases and 27,540 prostate cancer related deaths, making it the second most deadly cancer in American males behind lung cancer (Siegel et al. 2015). Prostate cancer is typically a slow-growing cancer afflicting older males where the median age of diagnosis in the United states is 66 (age 66 in white males and 63 in black males) and the median age of PrCa-related death is 80 (ages 81 and 77 for whites and blacks, respectively) (Howlader et al. 2015). Joint analysis of the US Surveillance, Epidemiology, and End Results Program and Swedish Cancer Registry have indicated that men who were diagnosed with PrCa were less likely to die from the cancer itself than other causes (Epstein et al. 2012) and further evidenced by the 98.6% five-year survival rate in American men (Howlader et al. 2015). However, highly aggressive PrCa can have a high mortality due to metastasis that preferentially moving beyond the prostate and into bones and lymph nodes and has a 5-year survival rate drops to 28% (Howlader et al. 2015).

1.5.1 Genetic Predisposition to Prostate Cancer

Prostate cancer, though common among men, also has been shown to have a high heritability when compared to other cancer types (Lichtenstein et al. 2000). One of the strongest risk factors for prostate is having a positive family history: if a man’s father and brother(s) had a positive diagnosis, he has a 2.3-fold higher risk of developing the disease (95% CI = 1.76–3.12) (Chen et al. 2008). Additionally, studies performed in Scandinavian monozygotic and dizygotic twins showed that increased risk of PrCa from heritable factors is 42% (95% CI = 29%-50%), noticeably higher than the heritable risk of 27% and 35% found for breast and colorectal cancers, respectively (Lichtenstein et al. 2000).

Numerous linkage studies for PrCa had been attempted though, due to previously described technical difficulties of the assay, the results often failed to replicate and conflicted with similar studies. That is not to say that there were no successes; a linkage study in 2003 identified a non-synonymous variant in the *HOXB13* transcription factor (Lange et al. 2003). This result was verified by fine-mapping of the 17q21-22 locus and by sequencing experiments (Lange et al. 2007; Ewing et al. 2012).

As of November 2015, 39 separate GWAS of PrCa have been reported in the NHGRI-EBI GWAS Catalog (Welter et al. 2014). The first two PrCa GWAS were independently conducted in an Icelandic population and in the United States using cases and controls of European ancestry and simultaneously reported the association of 8q24 with increased risk for PrCa (Gudmundsson et al. 2007; Yeager et al. 2007). Numerous, additional PrCa GWAS in Europeans have been performed and have identified

Returning to 8q24, this region had previously been associated with PrCa risk through admixture mapping (Freedman et al. 2006; Amundadottir et al. 2006). However, while this GWAS did further support the true association with PrCa, the mechanism of increased risk remains unclear. The risk loci at 8q24 lie within a gene desert, a 1 Mb tract of the genome that contains no known genes. The closest gene to the risk loci is the oncogenic transcription factor, *MYC*, and conformation capture assays have identified a long-range chromatin-looping interaction between the risk-locus and the *MYC* gene in prostate cancer cell lines (Pomerantz et al. 2009). These results suggest that the risk-locus at 8q24 may ultimately regulate *MYC* expression through physical interaction and ultimately increase PrCa risk through a dysregulation of *MYC*, however precise details remain unclear.

The SNP, rs10993994 at 10q11, was originally identified as a PrCa risk SNP in two simultaneous GWAS in 2008 (Eeles et al. 2008; Thomas et al. 2008). This SNP is posed as an intriguing candidate for functional validation as it is found in the promoter region, a mere 57 base pairs (bp) upstream of the transcription start site of the gene microseminoprotein-beta, *MSMB*, which encodes for one of the 3 major secretory products of the prostate. My efforts to understand better understand how this SNP affects PrCa etiology is described in Chapter 4 of this thesis.

2 Chapter 2. Enrichment of Cancer Risk SNPs in Functional Elements of DNA

The results of this chapter have been published (Hayes et al. 2015).

Hayes J, Trynka G, Vijai J, Offit K, Raychaudhuri S, Klein RJ. *Tissue-Specific Enrichment of Lymphoma Risk Loci in Regulatory Elements*. PLOS One. 2015. DOI: 10.1371/journal.pone.0139360.

2.1 Introduction

2.1.1 Lymphoma and chronic lymphocytic leukemia

Lymphoma, including the non-Hodgkin's lymphoma subtype chronic lymphocytic leukemia (CLL), was responsible for more than 130,000 new cases of cancer and 44,000 deaths in 2014 (American Cancer Society 2014). Both CLL and Hodgkin's lymphoma are of B-cell origin and have been shown to have a high heritable component (Goldin et al. 2004; Sellick et al. 2006). In an effort to understand this effect and the overall etiology of these diseases, numerous genome-wide association studies (GWAS) have been performed and have identified common genetic variants associated with the risk of developing lymphoma (Berndt et al. 2013; Conde et al. 2010; Cozen et al. 2012; Crowther-Swanepoel et al. 2010; Di Bernardo et al. 2008; Enciso-Mora et al. 2010; Kumar et al. 2011; Skibola et al. 2009; Slager et al. 2011; Smedby et al. 2011; Urayama et al. 2012; Vijai et al. 2013; Wade et al. 2011). Collectively, these studies have identified loci across more than half of the human autosomes that contain low-risk variants that associate with these diseases. Some of the reported risk alleles are found at loci that begin to suggest how this risk is conferred, for example being located in a region containing genes involved in apoptosis (Berndt et al. 2013), a member of the NFκB transcription factor family (Enciso-Mora et al. 2010), and the multi-cancer associated locus 8q24 (Crowther-Swanepoel et al. 2010; Enciso-Mora et al. 2010). However, most of these reported

GWAS hits are non-coding and the direct mechanism by which these allele lead to an increase in risk is yet unresolved.

Previous work from our lab observed that single nucleotide polymorphisms (SNPs) found in evolutionary conserved regions and in regions epigenetically marked for transcriptional regulation are more likely to be under negative selection in humans, suggesting biological function (Levenstien & Klein 2011). Others have shown that risk variants are enriched in particular epigenomic marks of transcriptional regulatory regions (Maurano et al. 2012; Trynka et al. 2013) and that trait-associated SNPs, including GWAS-identified risk SNPs, are often found in expression quantitative trait loci (eQTL) that affect nearby gene expression (Nicolae et al. 2010). Furthermore, recent studies have shown the etiologic nature of transcription factors themselves in some diseases (Shah et al. 2013). Taken together, these data suggest the hypothesis that for many GWAS-identified risk lock, the functional variant may modulate disease risk through alteration of gene regulation rather than coding sequence.

To test the hypothesis that lymphoma risk SNPs, or their LD partners, tend to alter regulatory elements, we interrogated the functional genomics data from ENCODE (Encyclopedia of DNA Elements) (The ENCODE Consortium et al. 2011; The ENCODE Project Consortium 2012) and the Roadmap Epigenomics (Bernstein et al. 2010) consortia. The large amount of data available for the lymphoblastoid cell line GM12878 and other hematologically derived cells allows integrative analysis to give more accurate representation of the segments of the genome that are active regulatory elements (Ernst et al. 2011; Hoffman et al. 2012). To test our hypothesis, we developed a computational pipeline, UES (Uncovering Enrichment through Simulation), that uses a Monte Carlo approach to test whether a set of SNPs is significantly enriched for a particular functional genomic annotation of the genome, taking linkage disequilibrium (LD) patterns into account. We demonstrate a significant enrichment of these lymphoma risk SNPs in regulatory marks specific for lymphoid tissue.

2.2 Materials & Methods

2.2.1 Pipeline Construction & Workflow

We developed a computational pipeline entitled “Uncovering Enrichment through Simulation” (UES) to test if GWAS-identified SNPs are enriched in particular functional annotations through use of Monte Carlo simulations. The pipeline (Figure 2.X) is written predominantly in Perl and accepts 3 parameters: a text file containing the input set of SNPs, the genotyping platform from which to choose the random sets, and the number of random sets to be constructed. SNPs that had been identified at the HLA region – defined as chr6:29570005-33377658 (build 37) – were removed due to the high amount of variability and linkage disequilibrium at that region. LD was calculated using European populations of the 1000 Genomes database, phase 3. We chose the European population specifically since most of the GWAS that were included in our study were of individuals of European descent. The number of LD partners for each SNP was calculated and recorded for various r^2 thresholds: $r^2>0.2$, $r^2>0.4$, $r^2>0.6$, $r^2>0.8$, and $r^2=1$. Furthermore, LD was calculated between both SNPs and indels in the 1000genomes database. We excluded 79364 SNPs or indels that had duplicate start positions, and all the remaining variants were used in the simulations. Each of the initial SNPs is then categorized by its distance from the nearest transcription start site (TSS) and its number of LD partners. Quartiles for both the TSS distance and LD partner count are calculated separately, and the initial SNPs are binned accordingly. The number of each of the initial SNPs contained in each bin (characterized by distance from TSS and LD partner count) is recorded and used for subsequent random SNP set selection. Upon completion of this step, all of the SNPs from the appropriate genotyping platform are loaded (excluding the HLA region) and binned according to the initial SNP criteria. Since it has been shown that disease-associated SNPs have a higher MAF than expected by chance[40], we filter the platform SNPs and keep only those with a MAF >= 5% keeping in concert with the common filter steps when performing GWAS. Random SNP sets are chosen, matching the original bin frequencies, and LD partners are retrieved (r^2>0.8). All the data have been pre-calculated and are retrieved using Tabix (Li 2011). The script executes an instance of BedTool's intersectBed (Quinlan & Hall 2010) in order to determine which SNPs fall directly in a given track . Those resultant SNPs are then collapsed into loci that co-localize with marks based on LD structure. Finally, the empirical p-value for a specific track is calculated by the following formula:

$$p = \frac{r_{loci}}{n}$$

where r_{loci} = the number of instances when the frequency of co-localization of the random SNP sets with the feature >= the number of loci that co-localize with the feature for the initial input set of SNPs, and n = the number of random-SNP sets chosen. Using the Bonferoni method of multiple test correction, the p-value was considered significant if p < 0.05/(number of tracks tested for enrichment). The current pipeline and subsequent versions are available for download from the Klein lab's website http://research.mssm.edu/kleinlab/ues/.

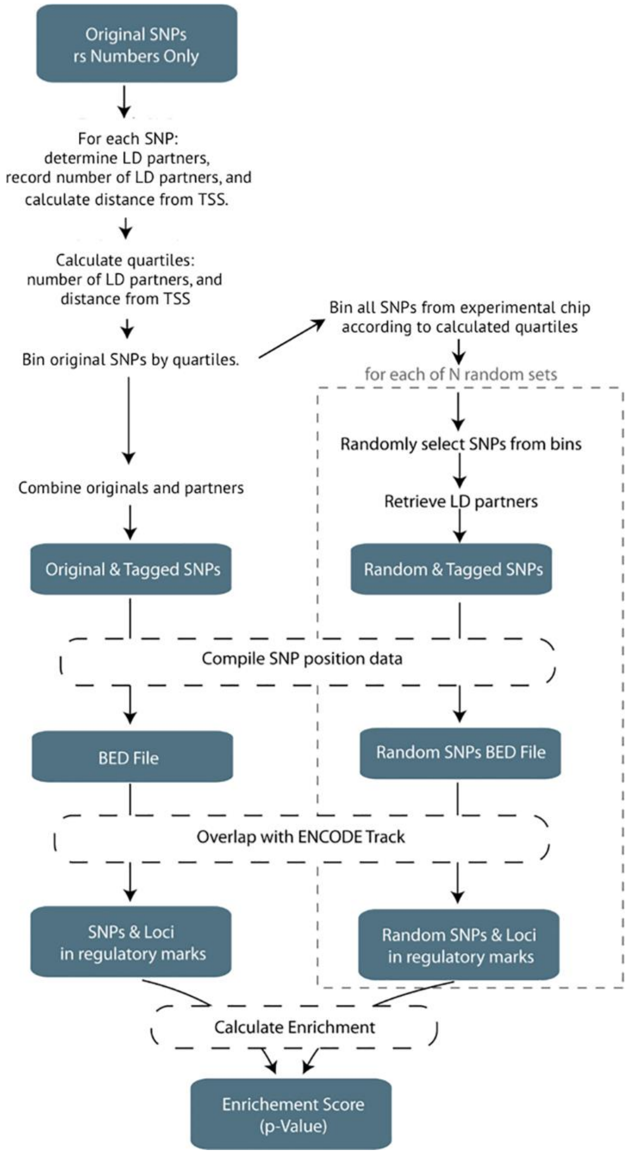

Original SNPs
rs Numbers Only
For each SNP:
determine LD partners,
record number of LD partners, and
calculate distance from TSS.
Calculate quartiles:
number of LD partners, and
distance from TSS
Bin original SNPs by quartiles.
Bin all SNPs from experimental chip
according to calculated quartiles
for each of N random sets
Randomly select SNPs from bins
Retrieve LD partners
Combine originals and partners
Original & Tagged SNPs
Random & Tagged SNPs
Compile SNP position data
BED File
Random SNPs BED File
Overlap with ENCODE Track
SNPs & Loci
in regulatory marks
Random SNPs & Loci
in regulatory marks
Calculate Enrichment
Enrichement Score
(p-Value)

Figure 2.1 UES algorithm visualization. This represents the generalized workflow to determine the SNP enrichment in an ENCODE track.

2.2.2 CLL & Lymphoma Risk SNPs

First, we manually queried the NHGRI GWAS Catalog (Welter et al. 2014) and selected a master list of CLL/lymphoma SNPs that had been reported as having a significant association. To ensure independence, for any SNPs that were correlated ($r^2 > 0.8$), the SNP with the lower, more significant reported p-value was kept. Initially, 56 CLL & lymphoma SNPs were entered into the pipeline, and once the HLA region was excluded, there were 36 SNPs used for the remainder of the analysis (Table 2.1) (Berndt et al. 2013; Conde et al. 2010; Cozen et al. 2012; Crowther-Swanepoel et al. 2010; Di Bernardo et al. 2008; Enciso-Mora et al. 2010; Kumar et al. 2011; Skibola et al. 2009; Slager et al. 2011; Smedby et al. 2011; Urayama et al. 2012; Vijai et al. 2013; Wade et al. 2011). Next the LD partners were found, resulting in 591 SNPs used for analysis of the original lymphoma and CLL data. The enrichment pipeline produced 10,000 sets consisting of 36 matched random SNPs. Once LD partners were included, the sets used for analysis range in size from 331 to 4028 SNPs.

Table 2.1 Lymphoma and CLL SNPs used for enrichment analysis.

SNP	Position	Study	p-value	Type of lymphoma
rs3770745	2:37596088	Berndt S. (2013)	1.68x10^{-08}	Chronic lymphocytic leukemia
rs1432295	2:61066665	Enciso-Mora V. (2010)	2.00x10^{-08}	Hodgkin's lymphoma
rs13401811	2:111616103	Berndt S. (2013)	2.00x10^{-18}	Chronic lymphocytic leukemia
rs17483466	2:111797457	Di Bernardo MC. (2008)	2.00x10^{-10}	Chronic lymphocytic leukemia
rs3769825	2:202111379	Berndt S. (2013)	3.00x10^{-09}	Chronic lymphocytic leukemia
rs13397985	2:231091222	Di Bernardo MC. (2008)	6.00x10^{-10}	Chronic lymphocytic leukemia
rs757978	2:242371100	Slager SL	3.00x10^{-06}	Chronic lymphocytic leukemia
rs898518	4:109016823	Berndt S. (2013)	4.00x10^{-10}	Chronic lymphocytic leukemia
rs27524	5:96101943	Urayama KY. (2012)	7.00x10^{-06}	Hodgkin's lymphoma
rs20541	5:131995963	Urayama KY. (2012)	1.00x10^{-08}	Hodgkin's lymphoma
rs872071	6:411063	Di Bernardo MC. (2008)	6.00x10^{-20}	Chronic lymphocytic leukemia
rs707824	6:14636962	Vijai J. (2013)	6.00x10^{-07}	Lymphoma/Non-Hodgkin's lymphoma
rs2456449	8:128192980	Crowther-Swanepoel D	8.00x10^{-10}	Chronic lymphocytic leukemia
rs2608053	8:129075831	Enciso-Mora V. (2010)	1.00x10^{-07}	Hodgkin's lymphoma
rs2019960	8:129192270	Enciso-Mora V. (2010)	1.00x10^{-13}	Hodgkin's lymphoma
rs1679013	9:22206986	Berndt S. (2013)	1.00x10^{-08}	Chronic lymphocytic leukemia
rs501764	10:8093033	Enciso-Mora V. (2010)	7.00x10^{-08}	Hodgkin's lymphoma
rs4406737	10:90759723	Berndt S. (2013)	1.00x10^{-14}	Chronic lymphocytic leukemia
rs7944004	11:2311151	Berndt S. (2013)	2.00x10^{-10}	Chronic lymphocytic leukemia
rs12289961	11:58060191	Vijai J. (2013)	4.00x10^{-08}	Lymphoma/Non-Hodgkin's lymphoma
rs948562	11:58347764	Vijai J. (2013)	6.00x10^{-07}	Lymphoma/Non-Hodgkin's lymphoma
rs735665	11:123361396	Di Bernardo MC. (2008)	4.00x10^{-12}	Chronic lymphocytic leukemia
rs7097	13:28197435	Kumar V. (2011)	7.00x10^{-06}	Large B-cell Lymphoma
rs751837	14:103484824	Kumar V. (2011)	3.00x10^{-07}	Large B-cell Lymphoma
rs8024033	15:40403656	Berndt S. (2013)	2.71x10^{-10}	Chronic lymphocytic leukemia
rs7169431	15:56340895	Crowther-Swanepoel D	5.00x10^{-07}	Chronic lymphocytic leukemia
rs7176508	15:70018989	Di Bernardo MC. (2008)	5.00x10^{-12}	Chronic lymphocytic leukemia
rs783540	15:83254707	Crowther-Swanepoel D	3.67x10^{-06}	Chronic lymphocytic leukemia
rs391525	16:85944438	Slager SL	3.00x10^{-09}	Chronic lymphocytic leukemia
rs305061	16:85975658	Slager SL	9.00x10^{-08}	Chronic lymphocytic leukemia
rs1036935	18:47843533	Crowther-Swanepoel D	2.28x10^{-06}	Chronic lymphocytic leukemia
rs4368253	18:57622286	Berndt S. (2013)	3.00x10^{-08}	Chronic lymphocytic leukemia
rs4987855	18:60793548	Berndt S. (2013)	3.00x10^{-12}	Chronic lymphocytic leukemia
rs4987852	18:60793920	Berndt S. (2013)	8.00x10^{-11}	Chronic lymphocytic leukemia
rs11083846	19:47207653	Di Bernardo MC. (2008)	4.00x10^{-09}	Chronic lymphocytic leukemia
rs11668878	19:47268372	Crowther-Swanepoel D	8.25x10^{-05}	Chronic lymphocytic leukemia

2.2.3 Location Pruning of CLL & Lymphoma Risk SNPs

In order to ensure that the observed signal was not due to oversampling of a region, we pruned SNPs from the input set so that SNPs were separated by at least a megabase (mb). For those SNPs in close proximity, we retained the SNP that had the lowest reported p-value, resulting in a set of 30 input SNPs.

2.2.4 UES analysis parameters

Since the analysis was run on a collection of SNPs from multiple studies, the parameter that chose the matched-random SNPs from a union set of both Illumina and Affymetrix genotyping chips was used for the random SNP set selection. The pipeline outputted 10,000 sets of feature-matched random SNPs.

2.2.5 Regulatory Track Data

The ENCODE datasets were obtained directly from the ENCODE Consortium's website. The DNase hypersensitivity analysis was performed using the ENCODE Consortium's "unified DNase hypersensitivity" tracks (http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgDnaseUniform/). The ChromHMM track was also downloaded from ENCODE (http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/), after which a Perl script was used to extract the active promoter, strong enhancer, and weak enhancer regions, or combine the active promoter and strong enhancer regions into a combination track. The Segway segmentation was downloaded directly from the Noble lab's website and was modified in the same way as described for the ChromHMM data (http://noble.gs.washington.edu/proj/segway/).

The Roadmap Epigenomics data were downloaded from the consortium's FTP website (ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/roadmapepigenomics/). The segmentation data was processed in the same manner as described for the ENCODE segmentation data.

2.2.6 Extracting cancer risk SNPs

Lists of risk SNPs for 19 different cancer types were obtained by mining NHGRI GWAS (Welter et al. 2014). To ensure independence, LD was calculated using the 1000 Genomes EUR population for variants that were within 500 KB of each other. For variants that were found to be correlated with r^2>0.8, the SNP with the more significant reported p-value was kept for use in the analysis. The UES enrichment pipeline was run separately for each cancer set identical to what was done for the lymphoma analysis. Enrichment scores were calculated for ENCODE DNase hypersensitivity data and all of the Roadmap Epigenomics ChromHMM 15-state model segmentation data.

2.2.7 RegulomeDB Analysis

The 36 lymphoma and CLL SNPs were run through the RegulomeDB website (http://www.regulomedb.org) (Alan P Boyle et al. 2012). The SNPs were inputted into the web-form, 1 SNP per line. There are no additional parameters to adjust.

2.2.8 HaploReg Analysis

The same list of 36 SNPs were analyzed using HaploReg (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) (Ward & Kellis 2012). They were uploaded as a comma-delimited list. The parameters were set as follows: LD threshold, $r^2 > 0.8$, population = EUR, source for epigenomes = ChromHMM 15-state model, conservation = SiPhyh-omega, genes shown relative to GENCODE, condense lists longer than 3, and condense indels longer than 6.

2.2.9 FunciSNP Analysis

FunciSNP was installed and the analysis was performed in accordance with the example vignette (Coetzee et al. 2012). The detailed code to run the FunciSNP analysis on the 326 lymphoma and CLL SNPs is provided in the Appendix.

2.2.10 GWAS-3D Analysis

GWAS-3D (Li et al. 2013) was executed directly from the Wang lab website (http://jjwanglab.org/gwas3d). The “input format” parameter was changed to “Single SNP ID,” the SNPs were subsequently pasted into the web-form containing 1 SNP per line, and the option p-value cutoff option was toggled to positive. The analysis was performed using all the default parameters except for specific cell type, which was changed to specifically perform the analysis using the GM12878 data.

2.2.11 GoShifter Analysis

Using European samples from the 1000 Genomes dataset we first iterated over each of the 36 lymphoma and CLL loci to identify all the variants in tight LD (r2>0.8). Locus boundaries were defined by the most downstream and upstream LD SNP and extended by two times the median size of a tested annotation. The locus was circularized which allowed annotations to randomly shift in 10,000 iterations. For each of the shifting iterations we quantified the number of loci at which a variant overlapped with an annotation. The reported p-value corresponds to the number of iterations where enrichment exceeded the observed value(Trynka et al. 2015).

2.3 Results

2.3.1 Enrichment of CLL & Lymphoma Risk SNPs in GM12878 Regulatory Tracks

We first asked if lymphoma risk SNPs are enriched in regions annotated as putatively regulatory in GM12878 using our novel method, Understanding Enrichment through Simulation (UES). Using the NHGRI GWAS catalog (Welter et al. 2014), we identified 56 risk SNPs for lymphoma, including both the Hodgkin’s lymphoma (HD) and non-Hodgkin’s lymphoma (NHL) types. Once the list was pruned to ensure the SNPs were independent and the HLA region was excluded, the resultant list contained

36 risk SNPs (Table 2.1). The minor allele distribution of the random SNPs were confirmed to be similar to the original input (input SNPs MAF mean = 0.277; random SNP sets mean = 0.259, median = 0.259, min = 0.176, 1st qu. = 0.244, 3rd qu. = 0.274, max = 0.347). I first looked at the Deoxyribonuclease I (DNase I) hypersensitivity sites (DHSs) for GM12878, since genomic regions open to DNase digestion have been shown to be accurate markers of regulatory DNA (Gross & Garrard 1988). We queried the ENCODE "unified DNase" track for GM12878, which identifies regions of open chromatin regardless of the particular factors that bind. The lymphoma risk-SNPs were significantly enriched in GM12878 DNase hypersensitivity sites (p < 0.0001), with 16 distinct regions containing risk SNPs potentially explainable by a variant in a DNase hypersensitive site. The 10,000 control sets of randomly selected SNPs with similar characteristics only showed an average of 4.5 regions potentially explainable by variants overlapping a DNase hypersensitive site (Figure 2.2.A and Table 2.2). The lymphoma risk-SNPs showed equal enrichment (p<0.0001) in the Roadmap Epigenomics DNase data for GM12878 (Table 2.3).

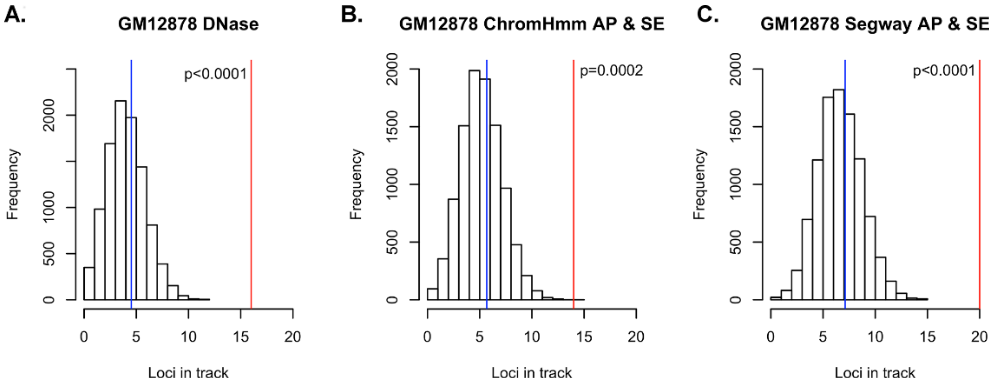


Figure 2.2 Overlap of lymphoma risk SNPs with regulatory regions in GM12878. The histograms represent the distribution of how many random loci overlap a specific annotation. The blue represents the mean of the empirical null distribution while the red line represents the real number of loci from the lymphoma and CLL GWAS that overlap the specific regulatory annotation. A, Overlap of SNPs with DNase hypersensitivity regions in GM12878. B, Overlap of SNPs with active promoters and strong enhancers as annotated by ChromHMM in GM12878. C, Overlap of SNPs with active promoters and strong enhancers as annotated by Segway in GM12878.

Table 2.2 Enrichment of lymphoma SNPs in ENCODE Unified DNase tracks. Enrichment analysis using the UES pipeline were performed for each of the 125 DNase hypersensitivity tracks in the ENCODE database. Cell line information were obtained directed from ENCODE's description of the cells used. The "OrigLoci" column gives the number of loci (once the SNPs are collapsed into loci based on LD partners) for the input lymphoma & CLL SNPs that overlapped with the specific mark. The "Rand>=Orig" column is the number of times a random SNP file had greater than or equal to the number of loci co-localizing with the particular mark. The "Random_Avg" column is the average of the 10,000 random generated SNP sets and the loci that overlap with the mark. The "p-value" is calculated by taking the number of random SNP sets that were greater than or equal to the input SNPs divided by n, in this case, 10,000. The table is sorted according to p-value at the r2>0.8 threshold, as reported in the body of the paper. The "location pruned p-value" is the reported p-value for the rerun of the analysis using the input data where SNPs were removed within one MB of one another.

Cell	**Tissue**	**OrigLoci**	**Rand>= OrigLoci**	**RandAvg**	**p-value (r2>0.8)**	**locationPruned p-value**
GM12878	blood	16	0	4.521	<0.0001	<0.0001
GM19238	blood	16	0	4.066	<0.0001	<0.0001
GM19240	blood	13	0	4.6825	<0.0001	0.0003
Adult_CD4_Th0	blood	15	0	5.0657	<0.0001	0.0001
CD20+	blood	13	0	3.1644	<0.0001	<0.0001
GM06990	blood	13	0	2.7063	<0.0001	<0.0001
GM12864	blood	15	0	3.7906	<0.0001	<0.0001
GM12865	blood	16	0	3.8906	<0.0001	<0.0001
GM19239	blood	11	2	3.5611	0.0002	<0.0001
HRE	epithelium	12	4	4.6192	0.0004	0.0017
GM18507	blood	10	5	3.4146	0.0005	0.0024
Jurkat	blood	11	9	4.246	0.0009	0.0021
GM12892	blood	11	10	4.0804	0.001	0.0007
Monocytes-CD14+_RO01746	monocytes	10	15	3.7611	0.0015	0.0055
Th1	blood	14	16	6.7864	0.0016	0.0003
Th2	blood	8	26	2.7238	0.0026	0.0003
CLL	blood	8	42	2.9196	0.0042	0.0036
GM12891	blood	9	65	3.8591	0.0065	0.0046
CD34+_Mobilized	blood	9	160	4.43	0.016	0.0126
HMVEC-LLy	blood vessel	8	242	3.9195	0.0242	0.0229
AoSMC	blood vessel	8	368	4.1836	0.0368	0.0318
ProgFib	skin	9	379	5.0479	0.0379	0.0088
HPAEC	blood vessel	7	399	3.4612	0.0399	0.0109
PrEC	prostate	8	430	4.3513	0.043	0.0403
LNCaP	prostate	11	438	6.7586	0.0438	0.0679
HTR8svn	blastula	7	457	3.5136	0.0457	0.0122
NHEK	skin	9	503	5.2339	0.0503	0.0374
Urothelia	urothelium	8	573	4.5203	0.0573	0.0156
HMVEC-dNeo	blood vessel	7	641	3.8388	0.0641	0.0179
HUVEC	blood vessel	8	697	4.6863	0.0697	0.0581
HA-h	brain hippocampus	8	734	4.7287	0.0734	0.0683
T-47D	breast	7	768	3.9421	0.0768	0.0718
BJ	skin	7	774	3.9676	0.0774	0.0799
HeLa-S3	cervix	8	796	4.8053	0.0796	0.0663
WI-38	embryonic lung	8	868	4.8793	0.0868	0.0723

(continued) Table 2.2 Enrichment of lymphoma SNPs in ENCODE Unified DNase tracks.

Cell	Tissue	OrigLoci	Rand>= OrigLoci	RandAvg	p-value (r2>0.8)	locationPruned p-value
HCT-116	colon	6	932	3.3267	0.0932	0.0359
HL-60	blood	7	1072	4.2696	0.1072	0.0981
RPTEC	epithelium	7	1075	4.2601	0.1075	0.2474
Caco-2	colon	6	1101	3.4641	0.1101	0.1201
H9ES	embryonic stem cell	7	1102	4.2561	0.1102	0.0379
HMVEC-dBl-Neo	blood vessel	7	1102	4.2971	0.1102	0.1075
HSMMtube	muscle	10	1144	6.9029	0.1144	0.1612
HEEpiC	epithelium	8	1153	5.1812	0.1153	0.1041
HRPEpiC	epithelium	8	1176	5.1665	0.1176	0.1095
A549	epithelium	7	1254	4.4241	0.1254	0.1172
HMF	mammary	7	1356	4.5156	0.1356	0.0528
NH-A	brain	7	1404	4.53	0.1404	0.1418
HMEC	breast	10	1449	7.1997	0.1449	0.2022
Huh-7.5	liver	7	1499	4.5945	0.1499	0.0502
HPIEpC	placenta	7	1596	4.6784	0.1596	0.1573
Fibrobl	skin	11	1615	8.2515	0.1615	0.3414
Ishikawa	uterus	6	1652	3.8188	0.1652	0.0705
HSMM	muscle	9	1722	6.5322	0.1722	0.129
AG09309	skin	7	1774	4.7821	0.1774	0.1628
Urothelia	urothelium	6	1803	3.9366	0.1803	0.0721
HPF	lung	6	1837	3.9532	0.1837	0.0769
K562	blood	7	1900	4.8986	0.19	0.1673
SAEC	epithelium	7	1946	4.9392	0.1946	0.1921
HAEpiC	epithelium	7	1974	4.9174	0.1974	0.0787
MCF-7	breast	7	2004	4.9388	0.2004	0.1908
HCPEpiC	epithelium	7	2053	4.9941	0.2053	0.0838
HFF-Myc	foreskin	7	2091	5.0037	0.2091	0.196
iPS	induced pluripotent stem cell	7	2136	5.0113	0.2136	0.0794
PanIsletD	pancreas	7	2163	5.0524	0.2163	0.0846
H1-hESC	embryonic stem cell	8	2191	5.9703	0.2191	0.1888
Hepatocytes	liver	6	2218	4.1906	0.2218	0.2214
HMVEC-dBl-Ad	blood vessel	6	2255	4.2143	0.2255	0.2317
Gliobla	brain	6	2405	4.2913	0.2405	0.1074
HIPEpiC	epithelium	7	2410	5.2268	0.241	0.2251
SK-N-MC	brain	5	2674	3.5553	0.2674	0.5724
HSMM_emb	muscle	5	2765	3.5835	0.2765	0.1312
HMVEC-dLy-Ad	blood vessel	5	2765	3.5694	0.2765	0.1307
NHDF-neo	skin	6	2776	4.4644	0.2776	0.1292
HRGEC	kidney	5	2945	3.6787	0.2945	0.1515
HFF	foreskin	6	3270	4.7305	0.327	0.3175
HPDE6-E6E7	pancreatic duct	5	3278	3.8144	0.3278	0.3594
HCM	heart	6	3339	4.7802	0.3339	0.1546
HBMEC	blood vessel	6	3377	4.7843	0.3377	0.3525
NHLF	lung	6	3393	4.7947	0.3393	0.3468
HepG2	liver	6	3588	4.9166	0.3588	0.1784
HConF	eye	5	3626	3.9757	0.3626	0.3951
NB4	blood	5	3657	3.9842	0.3657	0.2017
HMVEC-dLy-Neo	blood vessel	5	3686	4.0125	0.3686	0.196
Huh-7	liver	5	3765	4.0479	0.3765	0.1965
RWPE1	prostate	5	3956	4.1185	0.3956	0.2145
WI-38	embryonic lung	5	4136	4.1762	0.4136	0.2295

(continued) Table 2.2 Enrichment of lymphoma SNPs in ENCODE Unified DNase tracks.

Cell	Tissue	OrigLoci	Rand>= OrigLoci	RandAvg	p-value (r2>0.8)	locationPruned p-value
HeLa-S3	cervix	4	4175	3.2753	0.4175	0.2522
PANC-1	pancreas	4	4257	3.2881	0.4257	0.251
Myometr	myometrium	5	4268	4.2808	0.4268	0.2482
NHDF-Ad	skin	6	4270	5.2308	0.427	0.2332
AG10803	skin	5	4285	4.2694	0.4285	0.4555
AoAF	blood vessel	5	4401	4.3201	0.4401	0.2478
Medullo	brain	6	4403	5.2745	0.4403	0.2401
HMVEC-LBl	blood vessel	5	4488	4.3529	0.4488	0.4707
Melano	skin	7	4584	6.3788	0.4584	0.4233
HAc	cerebellar	5	4606	4.4126	0.4606	0.2815
HMVEC-dAd	blood vessel	4	4728	3.4713	0.4728	0.2933
SK-N-SH_RA	brain	3	4888	2.5811	0.4888	0.3244
HPAF	blood vessel	5	5189	4.6668	0.5189	0.3146
AG09319	gingival	4	5274	3.7063	0.5274	0.3358
Osteobl	bone	8	5292	7.7447	0.5292	0.4426
8988T	liver	5	5346	4.7396	0.5346	0.3156
HGF	gingiva	4	5367	3.7331	0.5367	0.3514
CMK	blood	4	5479	3.7863	0.5479	0.3336
WERI-Rb-1	eye	5	5485	4.8081	0.5485	0.3439
SKMC	muscle	5	5514	4.8092	0.5514	0.3517
AG04450	lung	4	5555	3.8283	0.5555	0.3667
FibroP	skin	6	5658	5.9075	0.5658	0.5432
LNCaP	prostate	4	5890	3.9884	0.589	0.6424
MCF-7	breast	4	6019	4.0272	0.6019	0.3992
pHTE	epithelium	6	6192	6.1997	0.6192	0.5946
HVMF	connective	4	6275	4.1624	0.6275	0.4394
HPdLF	epithelium	4	6418	4.2177	0.6418	0.4546
Stellate	liver	4	6508	4.2528	0.6508	0.4373
BE2_C	brain	4	6621	4.3431	0.6621	0.4906
HCF	heart	4	6689	4.3449	0.6689	0.4697
HCFaa	heart	4	6947	4.4782	0.6947	0.5135
PanIslets	pancreas	4	6969	4.5415	0.6969	0.5016
NT2-D1	testis	4	7369	4.7388	0.7369	0.5659
H7-hESC	blood	5	7739	6.0974	0.7739	0.5963
Ishikawa	uterus	3	7756	3.8754	0.7756	0.6264
HNPCEpiC	epithelium	4	7806	4.9796	0.7806	0.8161
AG04449	skin	3	8195	4.1223	0.8195	0.6745
Chorion	fetal membrane	3	8658	4.4445	0.8658	0.7182
HA-sp	spinal cord	2	9560	4.3911	0.956	0.9066

Table 2.3 Enrichment scores calculated for Roadmap Epigenomics DNase hypersensitivity data, ascending.

Cell Line	pValue (r2>0.8)
GM12878 Lymphoblastoid Cells	<0.0001
Primary B cells from peripheral blood	<0.0001
Primary T cells from cord blood	<0.0001
Primary Natural Killer cells from peripheral blood	<0.0001
Fetal Thymus	0.0012
Primary T cells from peripheral blood	0.0028
Primary hematopoietic stem cells G-CSF-mobilized Male	0.0036

(continued) Table 2.3 Enrichment scores calculated for Roadmap Epigenomics DNase hypersensitivity data, ascending.

Cell Line	pValue (r2>0.8)
Monocytes-CD14+ RO01746 Primary Cells	0.0054
Small Intestine	0.0294
Placenta	0.0308
Pancreas	0.0483
Primary hematopoietic stem cells G-CSF-mobilized Female	0.068
Primary monocytes from peripheral blood	0.0696
HeLa-S3 Cervical Carcinoma Cell Line	0.0801
Fetal Muscle Leg	0.0966
H1 BMP4 Derived Trophoblast Cultured Cells	0.1259
Breast variant Human Mammary Epithelial Cells (vHMEC)	0.1378
Foreskin Melanocyte Primary Cells skin01	0.1419
H1 Derived Neuronal Progenitor Cultured Cells	0.1465
Fetal Lung	0.1503
Fetal Kidney	0.1565
NH-A Astrocytes Primary Cells	0.1579
NHEK-Epidermal Keratinocyte Primary Cells	0.1887
Psoas Muscle	0.2024
Fetal Muscle Trunk	0.2362
iPS DF 19.11 Cells	0.2589
H1 Derived Mesenchymal Stem Cells	0.2755
IMR90 fetal lung fibroblasts Cell Line	0.2954
HUVEC Umbilical Vein Endothelial Primary Cells	0.3089
Fetal Stomach	0.3222
Foreskin Melanocyte Primary Cells skin03	0.326
HSMM Skeletal Muscle Myoblasts Cells	0.3406
K562 Leukemia Cells	0.3608
HMEC Mammary Epithelial Primary Cells	0.3744
H1 Cells	0.4654
Foreskin Keratinocyte Primary Cells skin02	0.4683
Fetal Intestine Small	0.4729
A549 EtOH 0.02pct Lung Carcinoma Cell Line	0.4807
Foreskin Fibroblast Primary Cells skin02	0.4883
HepG2 Hepatocellular Carcinoma Cell Line	0.501
HSMM cell derived Skeletal Muscle Myotubes Cells	0.5079
Fetal Intestine Large	0.5268
Gastric	0.5705
Ovary	0.5895
H9 Cells	0.6549
H1 BMP4 Derived Mesendoderm Cultured Cells	0.671
Foreskin Fibroblast Primary Cells skin01	0.6914
NHDF-Ad Adult Dermal Fibroblast Primary Cells	0.6951
NHLF Lung Fibroblast Primary Cells	0.7448
Fetal Heart	0.7673
iPS DF 6.9 Cells	0.8628
Fetal Brain Female	0.9502
Fetal Brain Male	0.9861

As some physical regions of the genome harbor more than one independent risk SNP, we were concerned this could lead to oversampling of a given region and false positives. To test this, we location-pruned the input SNPs, ensuring that none of the SNPs tested were within one MB of one another, reducing the input set from 36 initial SNPs down to 30 SNPs. The results were nearly identical between the location-pruned dataset and the original dataset (Table 2.2) leading to the conclusion that those 6 extra SNPs were not falsely inflating the observed statistical result.

Next, we asked if enrichment could be observed in regulatory elements predicted by genome segmentation of integrated functional genomics data. Rather than querying individual histone modifications, we chose to examine the results from two different segmentation algorithms, ChromHMM (Ernst et al. 2011) and Segway (Hoffman et al. 2012), since both algorithms use a combination of multiple histone modification datasets to determine the segmentation calls. We asked if the lymphoma risk SNPs are enriched in regions identified as active promoters or strong enhancers for GM12878 and observed a significant enrichment of the lymphoma SNPs in regulatory regions as defined by both ChromHMM and Segway with $p=0.0002$ and $p<0.0001$, respectively (Figures 2.2.B and 2.2.C). Upon looking deeper at the ChromHMM data for GM12878, it was observed that the risk SNPs were enriched in each of the four classes of enhancers (2 strong enhancer classes and 2 weak enhancer classes) with $p<=0.0002$ when analyzed separately (Table 2.4). When combined into a separate strong enhancer set and a weak enhancer set, there was a significant enrichment ($p<0.0001$) when compared to random controls for both. Interestingly, the "Active Promoter" state, by itself showed no significant enrichment ($p=0.3845$). Similar results were observed when performing the enrichment analysis for the Segway segmentation track of GM12878: strong enhancers were the most enriched ($p < 0.0001$); weak enhancers and active promoters did not achieve significance at the Bonferoni threshold ($p = 0.0031$ and $p=0.0419$, respectively; Table 2.4).

Table 2.4 Enrichment in GM12878 segmentation data.

ENCODE_Track	OrigLoci	Rand >=OrigLoci	RandAvg	pValue (r2>0.8)
ChromHmm Tracks				
4_Strong_Enhancer.bed	11	0	2.6593	<0.0001
4-7_Enhancers.bed	23	0	9.087	<0.0001
6-7_Weak_Enhancer.bed	17	0	7.7791	<0.0001
7_Weak_Enhancer.bed	15	0	6.051	<0.0001
4-5_Strong_Enchancer.bed	12	2	4.0411	0.0002
AP/SE narrowPeaks	14	2	5.6771	0.0002
2_Weak_Promoter.bed	9	5	3.1866	0.0005
6_Weak_Enhancer.bed	10	20	4.0949	0.002
11_Weak_Txn.bed	18	34	10.2319	0.0034
5_Strong_Enhancer.bed	7	102	2.6873	0.0102
10_Txn_Elongation.bed	8	1662	5.6127	0.1662
9_Txn_Transition.bed	4	1849	2.3141	0.1849
1_Active_Promoter.bed	4	3845	3.1341	0.3845
12_Repressed.bed	6	4424	5.3371	0.4424
3_Poised_Promoter.bed	1	4632	0.6186	0.4632
8_Insulator.bed	3	4818	2.5635	0.4818
13_Heterochrom-lo.bed	17	10000	28.7431	1
14_Repetitive-CNV.bed	0	10000	0.4004	1
14-15_Repetitive-CNV.bed	0	10000	0.5791	1
15_Repetitive-CNV.bed	0	10000	0.2312	1
states12-15.bed	21	10000	29.8544	1
Segway Tracks				
ActivePromoterStrongEnhancer.bed	20	0	7.1181	<0.0001
StrongEnhancer.bed	17	0	5.5632	<0.0001
WeakEnhancer.bed	13	31	6.5206	0.0031
TranscriptionAssociated.bed	18	201	12.2778	0.0201
ActivePromoter.bed	5	419	2.0627	0.0419
PromoterFlanking.bed	6	2803	4.5202	0.2803
Insulator.bed	4	5190	3.6909	0.519
LowActivity.bed	25	8660	26.9396	0.866
HeterochromatinRepetiveCNV.bed	7	9996	14.3001	0.9996
PolycombRepressed.bed	16	9999	25.227	0.9999

The analysis was originally performed with the LD threshold cutoff at $r^2 > 0.8$ since this was the threshold during the design of the GWAS chips. In order to test whether our cutoff was appropriate, the enrichment analysis was repeated using additional r^2 thresholds of $r^2>0.2$, $r^2>0.4$, $r^2>0.6$, and $r^2=1$. The results showed nearly identical enrichment of the same cell lines and tissue specificity for both DNase hypersensitivity marks and segmentation data for every LD threshold, except for r^2=1. Similar levels of enrichment within specific segmentation datatypes of GM12878 are also observed at the other LD threshold levels.

Next we examined whether or not functional SNPs may be those localized at transcription factor binding sites (TFBS). Using the same SNPs, we interrogated the set of ENCODE ChIP-Seq data for GM12878 (January 2011 data freeze). A master dataset consisting of the union of 75 GM12878 transcription factor ChIP-Seq data was created and in which there was a significant enrichment of the lymphoma SNPs in the peaks when compared to the random controls ($p < 0.0001$). Additional sets containing the union of all of the transcription factor peaks with the Gm12878 DNase hypersensitivity and a union of all the LCL DNase hypersensitivity were generated and there was a similar enrichment in both ($p < 0.0001$). In order to identify specific transcription factors that colocalize with lymphoma risk SNPs or their LD partners, the enrichment analysis was performed for each factor in the ChIP-Seq dataset, 4 of which reached the significance threshold once corrected for multiple testing: NFIC, RUNX3, NFκB, and TBLR1 (p< 0.0001; p<0.0001; p=0.0002; p=0.0005).

To test the validity of the findings and verify the approach, we repeated the identical analysis with a set of breast cancer risk SNPs identified from the NHGRI GWAS catalogue as there were a similar number of SNPs in the database at the time (n=31). Since the overall hypothesis states that since the diseases do not share a tissue of origin, there should be no enrichment of the breast cancer SNPs in the GM12878 annotations. There was no observable statistical enrichment (random n=10,000) of the breast cancer risk loci in DNase hypersensitivity, ChromHMM enhancer, Segway enhancer data, or TF union data for GM12878 (Table 2.5). Expanding the scope to see whether or not these tissue-specific observations held true when analyzing a different disease, identical analysis was calculated for prostate cancer which had roughly double the number of input SNPs (n=62). As seen with the breast cancer SNPs, there was no statistical enrichment of the prostate cancer SNPs in any of the datasets for GM12878 after correcting for multiple testing (Table 2.6). These results suggest that the observed signal is specific to the lymphoma and are not generalizable to cancer risk as a whole.

Table 2.5. Breast cancer SNPs analyzed with UES against GM12878 datasets.

Annotation Track	OrigLoci	Rand>=OrigLoci	RandAvg	p-value
DNase Hypersensitivity				
ENCODE	10	631	6.3462	0.0631
Strong Enhancers				
ENCODE-SE	6	3939	5.063	0.3939
Segway-SE	7	5983	7.0854	0.5983
Weak Enhancers				
ENCODE-WE	11	3913	9.899	0.3913
Segway-WE	13	522	8.8717	0.0522
Union Sets				
gm12878tfUnion	10	7855	11.3192	0.7855
lclDnaseUnion	14	3403	12.5424	0.3403
gm12878dnase-gm12878tfUnion.intersect.bed	6	3481	4.8428	0.3481
lclDnaseUnion-gm12878tfUnion.intersect.bed	8	5162	7.6387	0.5162

Table 2.6. Prostate cancer SNPs analyzed with UES against GM12878 datasets.

Annotation Track	OrigLoci	Rand>=OrigLoci	RandAvg	pValue
DNase Hypersensitivity				
ENCODE	14	1268	10.4187	0.1268
Strong Enhancers				
ENCODE-SE	11	2365	8.6987	0.2365
Segway-SE	19	176	12.191	0.0176
Weak Enhancers				
ENCODE-WE	22	945	17.2515	0.0945
Segway-WE	20	1168	15.8539	0.1168

(continued) Table 2.6. Prostate cancer SNPs analyzed with UES against GM12878 datasets.

Annotation Track	OrigLoci	Rand>=OrigLoci	RandAvg	pValue
Union Sets				
gm12878tfUnion	22	3231	19.9967	0.3231
lclDnaseUnion	30	103	21.5959	0.0103
gm12878dnase-gm12878tfUnion.intersect.bed	8	5654	7.984	0.5654
lclDnaseUnion-gm12878tfUnion.intersect.bed	16	1711	12.7851	0.1711

2.3.2 Tissue Specificity of CLL & Lymphoma Risk SNPs

Since it was shown that the lymphoma SNPs were enriched in the LCL, GM12878, the next logical step was to determine if the observed enrichments were seen in the same epigenomic marks of other cell types or whether they were indeed specific to cells of the lymphoid lineage. First, we interrogated the other 124 unified DNase tracks from ENCODE with the same GWAS and random SNP sets used for the GM12878 analysis and observed enrichment of 8 additional cell lines that achieved sub-Bonferoni significance with p<0.0004. Interestingly, all of the lines that showed enrichment at that level were of the lymphoid lineage (Figures 2.3.A-E, J-K, Table 2.2). When the stringency is relaxed and the scope is expanded to any cell lines with p<0.01, 15 out of the 18 cell lines which surpassed that threshold were of the lymphoid lineage. All of the LCLs in the ENCODE database were below this threshold and had a p<=0.0065 (Figure 2.3.F-I). There was only one cell line, HRE (renal epithelial cells), which was not from the lymphoid lineage that almost met the statistical threshold once corrected for multiple testing (p=0.0004). These data show that the previously reported lymphoma risk SNPs are enriched in DNase hypersensitivity regions in a tissue-specific manner (Figure 2.3.L). Similar tissue-specific enrichment results were seen in the Roadmap Epigenomics data (Table 2.3).

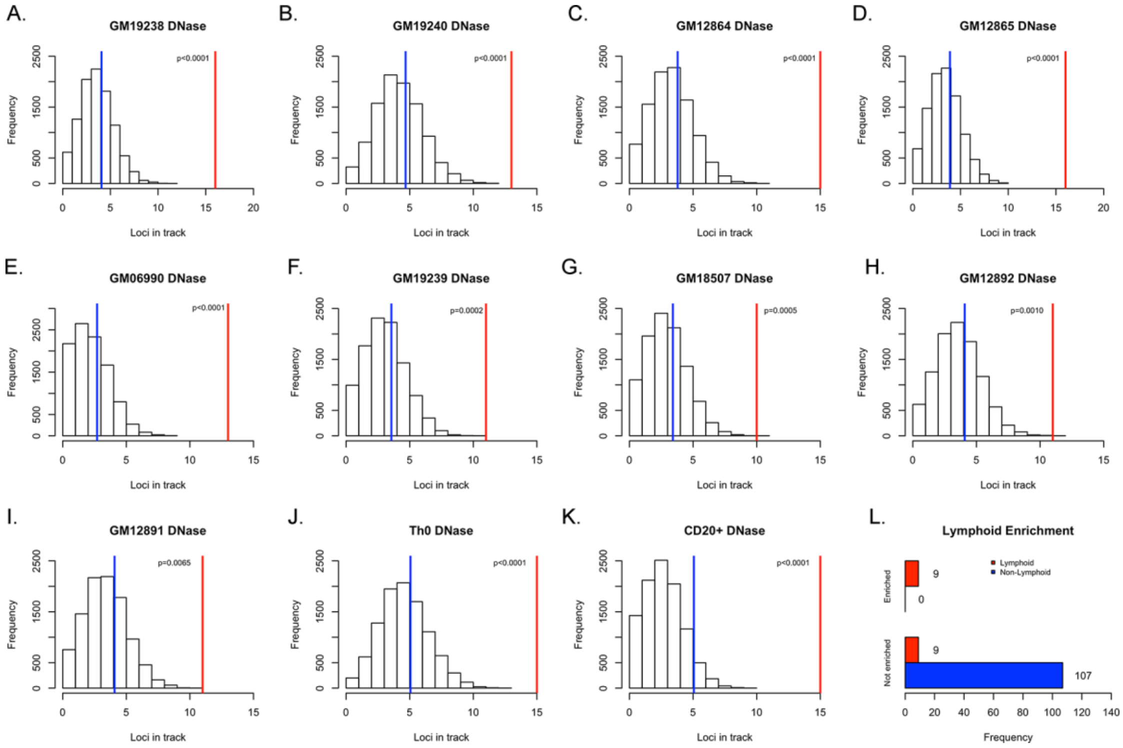


Figure 2.3. Enrichment of lymphoma and CLL risk SNPs in DNase-hypersensitive sites of lymphoblastoid cell lines. (A-I) These histograms represent the distribution of how many random loci overlap a specific annotation. The blue represents the mean of the empirical null distribution while the red line represents the real number of loci from the lymphoma and CLL GWAS that overlap the DNase hypersensitive site in the specified cell line: (A) GM19238 (B) GM19240 (C) GM12864 (D) GM12865 (E) GM06990 (F) GM19239 (G) GM18507 (H) GM12892 (I) GM12891. (J) Th0 (K) CD20+ (L) Summary of distribution of tissue of origin for cell lines in which lymphoma and CLL risk SNPs are either enriched (p<0.0004) in DNase hypersensitive sites or not enriched.

Similar to previous analysis, we next looked at the enhancer segmentation data available for other cell lines: 8 additional cell lines with ChromHMM data and 5 additional lines with Segway data. For both the ChromHMM and Segway strong enhancer classifications, GM12878 was the only cell line that showed strong, significant enrichment with p=0.0002 and p<0.0001 for each dataset, respectively. When looking at the weak enhancer classifications, again, GM12878 was the only cell type to demonstrate any significance with p<0.0001 and p=0.0031 for the ChromHMM and Segway data, respectively. Similarly, tissue-specific enhancer enrichment were also observed in the Roadmap Epigenomics data (p<0.0001) while the enrichment calculations for the transcription start site segmentation failed to reach significance (p>0.0497).

2.3.3 Lymphoma & CLL SNPs as eQTLs

Another prediction of the hypothesis that lymphoma risk SNPs alter regulatory regions is that these SNPs will be associated with expression changes in nearby genes. To test this hypothesis, we asked how many of the published lymphoma risk SNPs are expression quantitative trait loci (eQTLs) using a recently published set of eQTLs in blood (Westra et al. 2013). Of the original loci, 21 of them were shown to have at least 1 cis eQTL. Furthermore, most of these eQTLs are tissue specific when queried against the GTEx eQTL database. One example of the power of this overall approach is evidenced by rs7097. This SNP was initially defined as a lymphoma risk SNP(Kumar et al. 2011) but did not intersect with any LCL DNase hypersensitivity sites, nor with promoter or enhancer regions of chromatin segmentation data. However, one of the SNPs it tags, rs694609 (r^2=0.978, D'=0.996), was found in a DNase hypersensitivity site of GM12878, which was categorized as falling in an "Active Promoter" by ChromHMM or a "Transcription Start Site" by Segway. Even as both the original SNP, rs7097, and the tag, rs694609, showed evidence of being *cis*-eQTLs for the same genes (POLR1D, LNX2, and GTF3A), our pipeline would suggest that the tagged SNP is the more likely candidate to be the functionally relevant SNP as it is found in open chromatin and in functional marks (Table 2.7).

Table 2.7 Combination of genomic marks and eQTL status for rs7097 & tag SNP.

Position	RSID	Original SNP	DNase Site	ChromHmm-AP	ChromHmm-SE	Segway-AP	Segway-SE	eQTL (Westra et al.)
13:28194806	rs694609	rs7097	✓	✓		✓		POLR1D,LNX2,GTF3A
13:28197436	rs7097	rs7097						POLR1D,LNX2,GTF3A,NR2E3

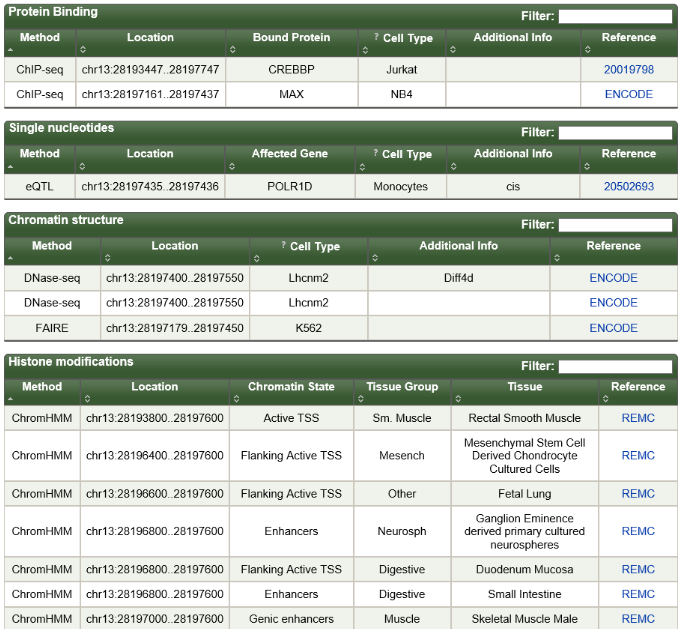
2.3.4 Orthogonal approaches to analyze GWAS SNPs

There has been an interest from the genomics community to develop methods and tools to utilize various publically-available genomic data to analyze SNPs. While these tools are similar in broad terms, each method asks a different question and returns different results. The same list of 36 lymphoma SNPs was run through the pipelines: RegulomeDB, HaploReg, GWAS-3D, FunciSNP, and GoShifter.

2.3.4.1 RegulomeDB

RegulomeDB (Alan P. Boyle et al. 2012) is a web-based, SNP annotation tool. A user can input either a list of SNPs, genomic coordinates, or BED file and RegulomeDB will query its database and return a score value for each SNP. Every SNP in 1000 Genomes was given a RegulomeDB score ranging from 1-7, with lower scores equating to a greater amount of evidence that the SNP may be functional. For example any score that begins with a “1” is considered likely to affect binding in that it was found to an eQTL and to colocalize with a DNase hypersensitivity site or TF motif. Users can then click on any SNP to get a more complete view of the evidence for the score. Our list of 36 lymphoma SNPs was run through RegulomeDB; rs7097 was given a RegulomeDB score of “1f”, indicating that it this SNP is likely to be functional (Figure 2.4).

Though both RegulomeDB and UES can be used to analyze GWAS SNPs, the algorithms are quite different and serve dissimilar functions. However, it fails to consider LD, providing only the information for the specific queried SNPs. While regions can be provided as input, without considering LD, the results will likely contain variants that are physically close though not correlated. Additionally, the provided score only addresses whether or not a SNP colocalizes with genomic annotations, but in the absence of a statistical framework, those results could lead to false positives by just random chance. RegulomeDB’s strength lies in its ability to quickly query its database, annotate SNPs, and provide a score that succinctly encapsulates a breadth of information about the SNP and what types of marks it colocalizes with; it is quite different from UES, whose purpose is to provide a statistical enrichment of SNPs in particular genomic annotation tracks.



Protein Binding

Filter:

Method	Location	Bound Protein	? Cell Type	Additional Info	Reference
ChIP-seq	chr13:28193447..28197747	CREBBP	Jurkat		20019798
ChIP-seq	chr13:28197161..28197437	MAX	NB4		ENCODE

Single nucleotides

Filter:

Method	Location	Affected Gene	? Cell Type	Additional Info	Reference
eQTL	chr13:28197435..28197436	POLR1D	Monocytes	cis	20502693

Chromatin structure

Filter:

Method	Location	? Cell Type	Additional Info	Reference
DNase-seq	chr13:28197400..28197550	Lhcnm2	Diff4d	ENCODE
DNase-seq	chr13:28197400..28197550	Lhcnm2		ENCODE
FAIRE	chr13:28197179..28197450	K562		ENCODE

Histone modifications

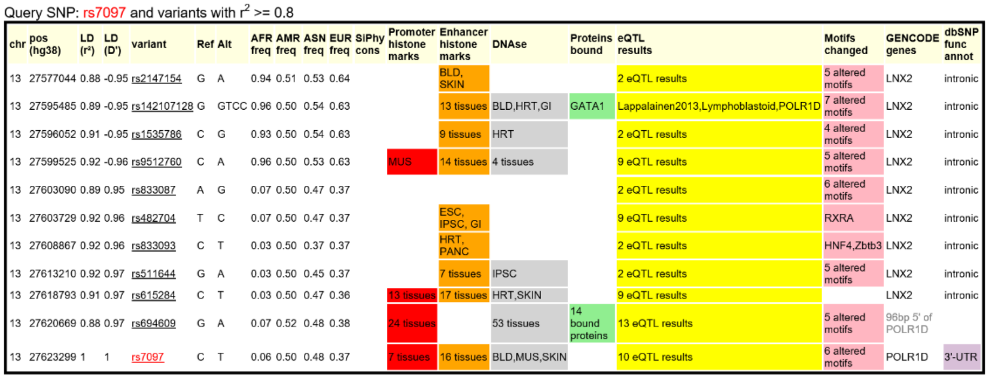
Filter:

Method	Location	Chromatin State	Tissue Group	Tissue	Reference
ChromHMM	chr13:28193800..28197600	Active TSS	Sm. Muscle	Rectal Smooth Muscle	REMC
ChromHMM	chr13:28196400..28197600	Flanking Active TSS	Mesench	Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells	REMC
ChromHMM	chr13:28196600..28197600	Flanking Active TSS	Other	Fetal Lung	REMC
ChromHMM	chr13:28196800..28197600	Enhancers	Neurosph	Ganglion Eminence derived primary cultured neurospheres	REMC
ChromHMM	chr13:28196800..28197600	Flanking Active TSS	Digestive	Duodenum Mucosa	REMC
ChromHMM	chr13:28196800..28197600	Enhancers	Digestive	Small Intestine	REMC
ChromHMM	chr13:28197000..28197600	Genic enhancers	Muscle	Skeletal Muscle Male	REMC

Figure 2.4 Output from RegulomeDB for rs7097.

2.3.4.2 HaploReg

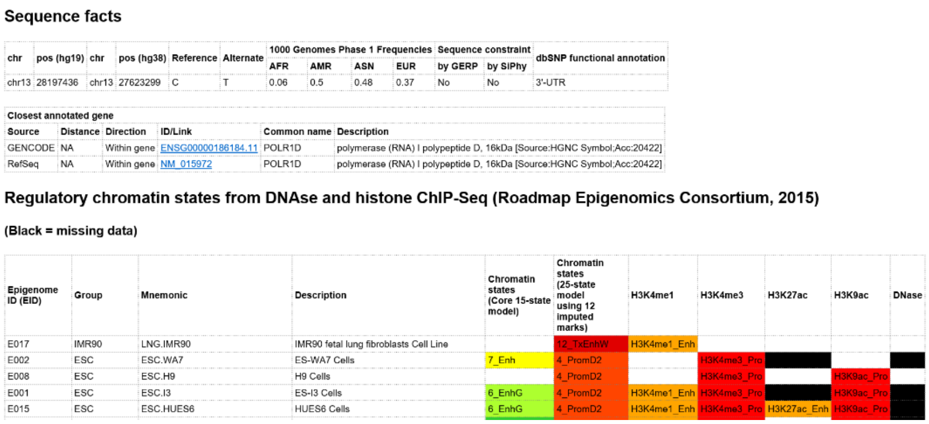
HaploReg (Ward & Kellis 2012) is another web-based, annotation tool that provides a wealth of information about both the queried SNPs and their respective LD partners. It is customizable in how it allows users to choose the r^2 threshold for LD calculations and their choice of HapMap population before querying its database. The results are returned in haplotypes, providing the SNP information, alleles, population frequencies, and the tissues which colocalize with the SNP at promoters, enhancers, and DNase hypersensitivity sites. Additionally, the results indicate whether a SNP has been found to be an eQTL, or is found in a transcription factor motif or any genes. Users can then click on an individual SNPs to be shown more precise information regarding the specific tissues types that colocalize with that particular SNP. The visualization of the haplotype containing rs7097 (Figure 2.5) and the some of the specific results for the SNP (Figure 2.6) are provided.



Query SNP: rs7097 and variants with r^2 >= 0.8

chr	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL results	Motifs changed	GENCODE genes	dbSNP func annot
13	27577044	0.88	-0.95	rs2147154	G	A	0.94	0.51	0.53	0.64			BLD, SKIN			2 eQTL results	5 altered motifs	LNX2	intronic
13	27595485	0.89	-0.95	rs142107128	G	GTCC	0.96	0.50	0.54	0.63			13 tissues	BLD,HRT,GI	GATA1	Lappalainen2013,Lymphoblastoid,POLR1D	7 altered motifs	LNX2	intronic
13	27596052	0.91	-0.95	rs1535786	C	G	0.93	0.50	0.54	0.63			9 tissues	HRT		2 eQTL results	4 altered motifs	LNX2	intronic
13	27599525	0.92	-0.96	rs9512760	C	A	0.96	0.50	0.53	0.63		MUS	14 tissues	4 tissues		9 eQTL results	5 altered motifs	LNX2	intronic
13	27603090	0.89	0.95	rs833087	A	G	0.07	0.50	0.47	0.37						2 eQTL results	6 altered motifs	LNX2	intronic
13	27603729	0.92	0.96	rs482704	T	C	0.07	0.50	0.47	0.37			ESC, IPSC, GI			9 eQTL results	RXRA	LNX2	intronic
13	27608867	0.92	0.96	rs833093	C	T	0.03	0.50	0.37	0.37			HRT, PANC			2 eQTL results	HNF4,Zbtb3	LNX2	intronic
13	27613210	0.92	0.97	rs511644	G	A	0.03	0.50	0.45	0.37			7 tissues	IPSC		2 eQTL results	5 altered motifs	LNX2	intronic
13	27618793	0.91	0.97	rs615284	C	T	0.03	0.50	0.47	0.36		13 tissues	17 tissues	HRT,SKIN		9 eQTL results		LNX2	intronic
13	27620669	0.88	0.97	rs694609	G	A	0.07	0.52	0.48	0.38		24 tissues		53 tissues	14 bound proteins	13 eQTL results	5 altered motifs	96bp 5' of POLR1D	
13	27623299	1	1	rs7097	C	T	0.06	0.50	0.48	0.37		7 tissues	16 tissues	BLD,MUS,SKIN		10 eQTL results	6 altered motifs	POLR1D	3'-UTR

Figure 2.5 Visualization of the haplotype returned for rs7097 from HaploReg v4.



Sequence facts

chr	pos (hg19)	chr	pos (hg38)	Reference	Alternate	1000 Genomes Phase 1 Frequencies				Sequence constraint		dbSNP functional annotation
						AFR	AMR	ASN	EUR	by GERP	by SiPhy	
chr13	28197436	chr13	27623299	C	T	0.06	0.5	0.48	0.37	No	No	3'-UTR

Closest annotated gene					
Source	Distance	Direction	ID/Link	Common name	Description
GENCODE	NA	Within gene	ENSG00000186184.11	POLR1D	polymerase (RNA) I polypeptide D, 16kDa [Source:HGNC Symbol;Acc:20422]
RefSeq	NA	Within gene	NM_015972	POLR1D	polymerase (RNA) I polypeptide D, 16kDa [Source:HGNC Symbol;Acc:20422]

Regulatory chromatin states from DNAse and histone ChIP-Seq (Roadmap Epigenomics Consortium, 2015)

(Black = missing data)

Epigenome ID (EID)	Group	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase
E017	IMR90	LNG.IMR90	IMR90 fetal lung fibroblasts Cell Line		12_TxEnhW	H3K4me1_Enh				
E002	ESC	ESC.WA7	ES-WA7 Cells	7_Enh	4_PromD2		H3K4me3_Pro	■		■
E008	ESC	ESC.H9	H9 Cells		4_PromD2		H3K4me3_Pro		H3K9ac_Pro	
E001	ESC	ESC.I3	ES-I3 Cells	6_EnhG	4_PromD2	H3K4me1_Enh	H3K4me3_Pro	■	H3K9ac_Pro	■
E015	ESC	ESC.HUES6	HUES6 Cells	6_EnhG	4_PromD2	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro	■

Figure 2.6 Specific results for rs7097 returned from HaploReg v4.

While UES is not a SNP annotation program, there is a functional overlap in that HaploReg similarly returns an enrichment score of the queried of SNPs. However there are major differences between how the enrichment is calculated. As described in "2.4.1 Materials & Methods," UES calculates the empirical enrichment score using a Monte Carlo simulation, choosing controlled sets of random SNPs. HaploReg v4 calculates its enrichment score by taking all unique GWAS loci in the HapMap European population and all variants from 1000 Genomes with a frequency greater than 5% in any population. UES, alternatively, only chooses the random SNPs from the specified genotyping platform on which the study was done. Additionally, HaploReg v4 only returns the enrichment score from pre-calculated binomial tests, where the aforementioned SNP sets were used to calculate a background frequency of enhancer overlap in one of four datasets: ChromHMM 15-state model, ChromHmm 25-state model using 12 imputed marks, H3K4me1/H3K4me3 peaks, and H3K27ac/H3K9ac peaks. While this is indeed quicker, UES provides a flexibility for the user to calculate the enrichment of their SNP sets in any genomic annotation track. The returned binomial p-values for GM12878 are p=0.000607 and p=0.005931 for all SNPs and only GWAS SNPs, respectively.

2.3.4.3 FunciSNP

FunciSNP (Coetzee et al. 2012) is a Bioconductor package for R that runs locally and is used to identify putative regulatory SNPs and, unlike the aforementioned methods, is not a SNP annotator. To run FunciSNP, the user provides the SNPs of interest and the biofeature(s),or genomic peak files, of particular interest to the researcher. FunciSNP, similar to both UES and HaploReg, considers LD partners of the original SNPs when performing its functions. For each SNP, the program defines a genomic window, extracts all the SNPs within that region from 1000 Genomes, determines which SNPs are found to be in the biofeatures, and then calculates the R^2 and D' scores using those variants. The YAFSNPs, or "Yet Another Functional SNPs" are identified as they colocalize with numerous biofeature tracks.

We performed a basic FunciSNP run as outlined in the FunciSNP vignette (Coetzee et al. 2012) with the 36 lymphoma and CLL SNPs. As done with the UES analysis, the FunciSNP analysis was performed using the same GM12878 data: ENCODE DNase hypersensitivity, ChromHMM active promoters and strong enhancers, and Segway active promoters and strong enhancers. The identified YAFSNPs are identified in Figure 2.7. The FunciSNP analysis determined that the rs2456449 as the top potential functional SNP based on the number of SNPs that colocalized with biofeatures. The SNP, rs694609 as tagged by rs7097, which we identified based on colocalization with DNase hypersensitivity data, active promoter sites, and eQTL is identified as a YAFSNP, though not the top candidate.

FunciSNP and UES, though they provide complementary results, they fundamentally ask different questions of the data. One of the strengths of FunciSNP is

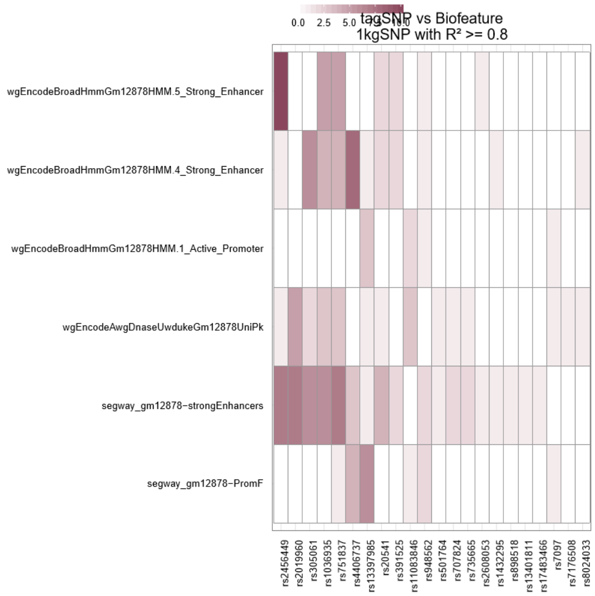


Figure 2.7 Heatmap of the output from FunciSNP.
Darker cells indicate a higher number of YFAPS for the lead SNP in the particular functional element.

to identify individual SNPs as putatively functional elements while UES looks at the whole set of SNPs collectively. FunciSNP outputs YAFSNPs based on the biofeatures provided of particular cell types. The results, unlike UES, do not provide any statistical analysis on whether or not those biofeatures or cell types are the appropriate context in which the YAFSNP may function. Likewise, the goal of GWAS-3D is very similar to that of FunciSNP: to identify putatively functional SNPs based on a combination of genomic signals. As such, the similarities and differences are between GWAS-3D and UES are comparable to that of FunciSNP. GWAS-3D also considers distal interactions when calculating an individual SNP's functional potential.

2.3.4.4 GWAS-3D

GWAS-3D (Li et al. 2013) is a web-based tool which utilizes numerous ENCODE datatypes to calculate the probability that variants affect regulatory pathways. The program uses curated ENCODE data of promoter marks, enhancer marks, insulators marks, chromatin interaction (Hi-C, 5C, and ChIA-PET), and

ChromHMM segmentation peaks. Once GWAS-3D has annotated the input SNPs and LD partners, the variants are subsequently analyzed to gauge whether they affect the binding affinity for various ENCODE transcription binding factor motifs and whether or not they colocalized with evolutionary conserved genetic elements.

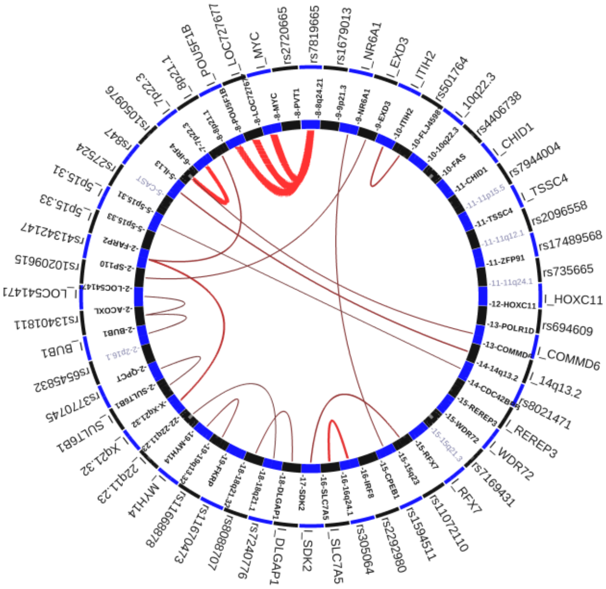
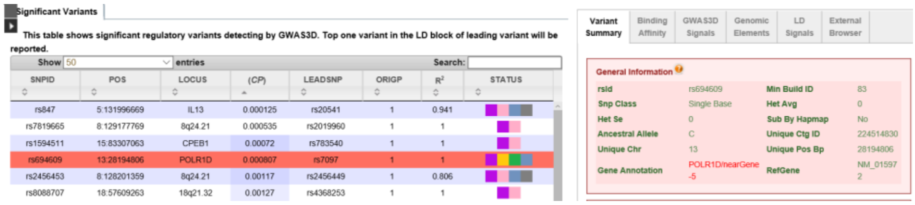


Figure 2.8 Circos-style output from GWAS-3D. The outer axis contains the most significant variants with the highest regulatory potential and the distal interaction partners. The inner axis contains the genes and locations of the variants. Distal interactions are visualized by the lines across the center of the plot, with the width of the line indicating the intensity of the interaction.



Significant Variants

This table shows significant regulatory variants detecting by GWAS3D. Top one variant in the LD block of leading variant will be reported.

Show 50 entries Search:

SNPID	POS	LOCUS	(CP)	LEADSNP	ORIGP	R^2	STATUS
rs847	5:131996669	IL13	0.000125	rs20541	1	0.941	🟪🟥🟦⬛
rs7819665	8:129177769	8q24.21	0.000535	rs2019960	1	1	🟪🟥
rs1594511	15:83307063	CPEB1	0.00072	rs783540	1	1	🟪🟥
rs694609	13:28194806	POLR1D	0.000807	rs7097	1	1	🟪🟨🟩🟦
rs2456453	8:128201359	8q24.21	0.00117	rs2456449	1	0.806	🟪🟥🟦⬛
rs8088707	18:57609263	18q21.32	0.00127	rs4368253	1	1	🟪🟥

General Information

rsId	rs694609	Min Build ID	83
Snp Class	Single Base	Het Avg	0
Het Se	0	Sub By Hapmap	No
Ancestral Allele	C	Unique Ctg ID	224514830
Unique Chr	13	Unique Pos Bp	28194806
Gene Annotation	POLR1D/nearGene-5	RefGene	NM_015972

Figure 2.9 Detailed variant output from GWAS-3D

The 36 lymphoma SNPs from our study were run through the GWAS-3D pipeline specifically querying the GM12878 datasets; results for 29 SNPs were returned. The most significant SNP from the GWAS-3D analysis was rs847 ($p=1.125x10^{-4}$) which was shown to be found in a ChromHMM strong enhancer, found in a conserved region, interact with a distal region, and significantly affect a TFBS (Figures 2.8 and 2.9). The SNP highlighted by our work, rs694609 was the 4th most significant SNP from the GWAS-3D analysis ($p=8.07x10^{-4}$) and was shown to significantly affect a TFBS, map to a promoter of a gene, map to a putative enhancer site, and in a ChromHMM-defined promoter, though it does not have any distal interactions. As seen in our analysis, rs694609 is more likely the functional SNP as opposed to the lead SNP rs7097.

2.3.4.5 GoShifter

It has been previously noted that, under some models in which the functional variants underlying GWAS are not regulatory, enrichment of GWAS-identified SNPs in regulatory regions could occur if proper controls are not used. While our method controls for the major factors that need to be controlled for (LD patterns and distance from transcription start), we nevertheless asked if a similar enrichment could be observed with an alternative approach, GoShifter, that shifts annotations at the associated loci to test the significance of enrichment (Trynka et al. 2015). This approach identified 5 cell lines that showed enrichment for the risk SNPs. Notably, 4 of these 5 lines that showed enrichment at $p < 0.05$– GM19238, Th0, Cd20, and GM06990 – are from the lymphoid lineage (Table 2.8).

Table 2.8. Significant DNase tracks from GoShifter analysis.

ENCODE DNase Track	p-value
wgEncodeAwgDnaseDukeGm19238UniPk	0.0009
wgEncodeAwgDnaseDukeTh0UniPk	0.0104
wgEncodeAwgDnaseUwHreUniPk	0.0228
wgEncodeAwgDnaseUwGm06990UniPk	0.032
wgEncodeAwgDnaseDukeGm19239UniPk	0.0346
wgEncodeAwgDnaseUwdukeGm12878UniPk	0.0354
wgEncodeAwgDnaseUwCd20UniPk	0.0361

Of all the methods GoShifter and UES attempt to answer the same question, albeit through orthogonal approaches. Trynka et al. (2015) demonstrated there is an over-inflation of the results of when failing to utilize proper controls; one of the largest contributors to this effect was due to the failure to consider LD when choosing matching SNPs. Trynka et al. chose to address this effect by developing their alternative approach of shifting the local annotations surrounding the regions of interest. UES was built to specifically address those concerns regarding the proper controls and the general concordance of the results between the two methods validates the work.

2.3.5 “Pan-cancer” enrichment

The NHGRI GWAS catalog (Welter et al. 2014) was mined to generate lists of risk SNPs for 19 different cancer types. The same UES analysis was performed in batch for each caner type separately and enrichment was calculated for ENDODE DNase hypersensitivity data and Roadmap Epigenomics 15-state ChromHMM datasets. Seeing as lymphoma risk showed a strong enrichment in a tissue-specific manner in enhancer sites, we interrogated whether this pattern was true across all these different cancers.

The greatest signal from the enrichment analysis of ENCODE DNase hypersensitivity sites were those signals which had been seen previously in the lymphoma analysis. There was only one other cancer type, breast cancer, that was enriched at a sub-Bonferoni level. This breast cancer enrichment was observed in MCF-7, a breast cancer cell line, showing a similar tissue-specific enrichment in the DNase hypersensitivity data (Figure 2.10). No other significant enrichment of cancer risk-SNPs were observed once corrected for multiple testing. The Roadmap Epigenomics segmentation data was analyzed and patterns started to emerge. The enhancer enrichment (Figure 2.11) again visualized the tissue-specificity of enrichment for the lymphoma SNPs and enrichment across most cell types was observed in the breast cancer risk SNPs containing the newly identified loci from iCOGS, a large-scale study effort by the Collaborative Oncological Gene-environment Study to identify additional common variation in various cancers (Bojesen et al. 2013; French et al. 2013; Gaudet et al. 2013; Couch et al. 2013; Garcia-Closas et al. 2013; Michailidou et al. 2013). When the rick loci that were previously found in MCF-7 peaks were removed and the enrichment analysis was repeated, a similar broad enrichment was observed, suggesting that these weaker-effect iCOGS SNPs are may not be specific to breast cancer. Other patterns of enrichment were quite striking: melanoma and esophageal cancer showed strong enrichment in genic regions (Figure 2.12) across most cell types while, conversely, no significant enrichment was observed in any cancer type when looking at active transcription start sites (Figure 2.13).

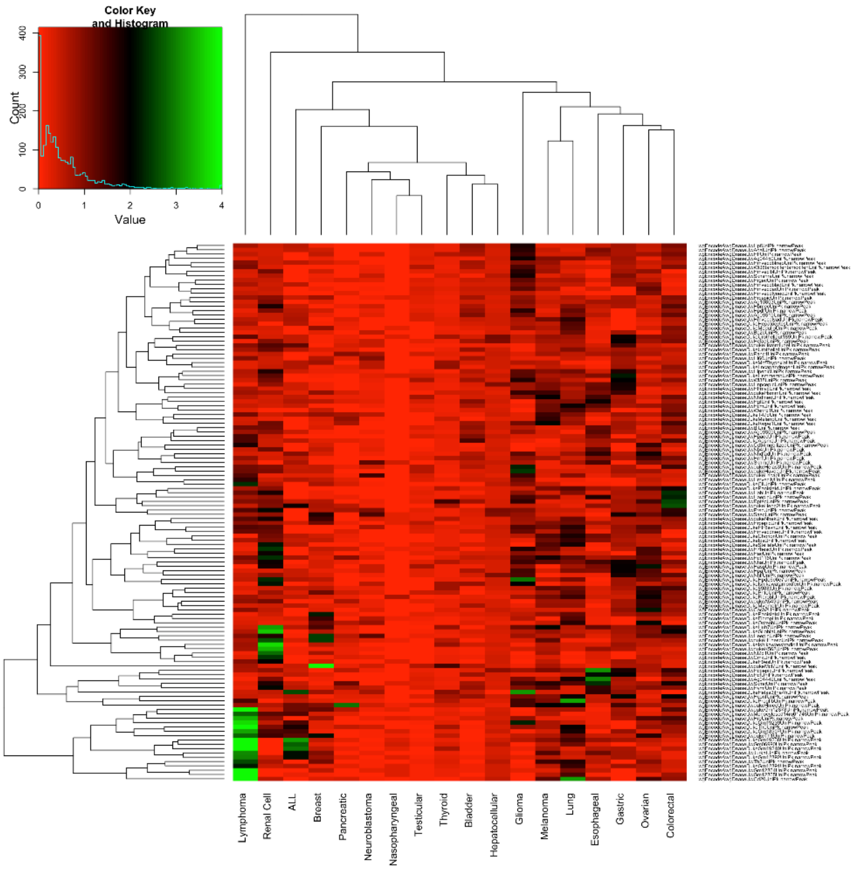


Figure 2.10. Pan-cancer UES enrichment analysis in ENCODE DNase hypersensitivity data. Each column represents a different cancer type. Each row is a separate ENCODE DNase hypersensitivity track. The colors visualize the p-value, with bright green being highly significant and red being insignificant. The tissue specific enrichment of lymphoma SNPs is clearly visualized. The only other statistically significant result is the lone enrichment of breast cancer SNPs in MCF-7, a breast cancer cell line.

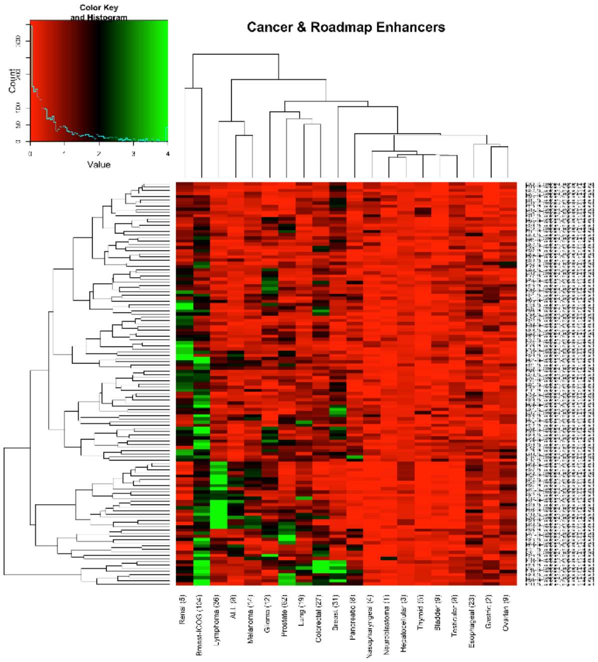


Figure 2.11. UES Enrichment in Roadmap Epigenomics enhancers. Green is a more significant p-value. Red is less significant.

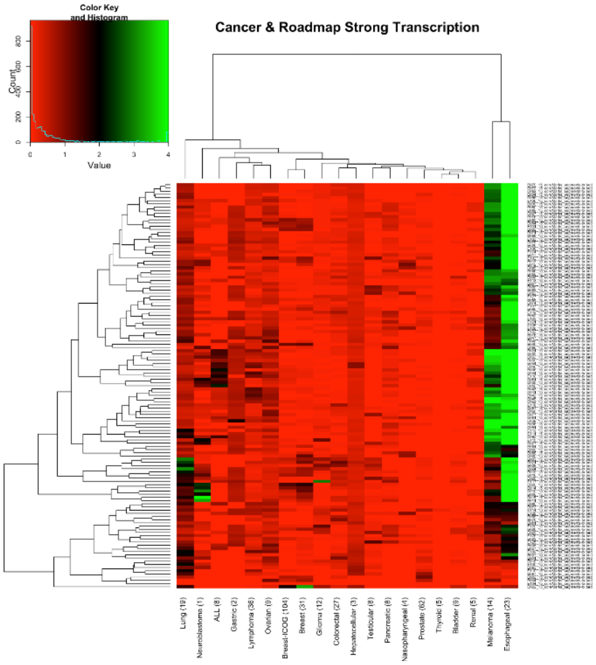


Figure 2.12. UES Enrichment in Roadmap Epigenomics genic regions. Green is a more significant p-value. Red is less significant.

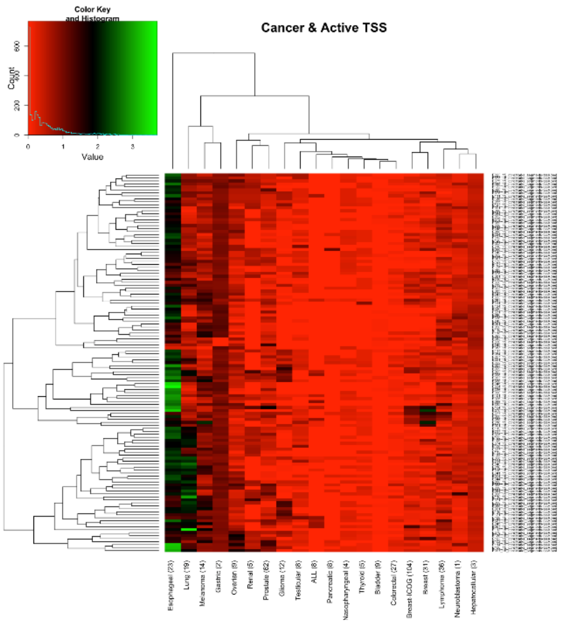


Figure 2.13. UES Enrichment in Roadmap Epigenomics Active TSS sites. Green is a more significant p-value. Red is less significant

2.4 Discussion

The UES algorithm is a novel, well-controlled method to determine the enrichment of GWAS SNPs in any genomic or epigenomic dataset. The pipeline was tested and validated using a set of lymphoma and CLL GWAS and has shown the set to be significant enrichment in the regulatory elements in lymphoblastoid cell lines. This suggests that there is a tissue-specific manner through which these genetic loci may confer increased risk for lymphomagenesis. Looking specifically at the analyses in GM12878, we observed this enrichment in DNase hypersensitivity loci as well as numerous enhancer sites; there was no observable enrichment when looking at risk loci for other cancer types. Our research also identified candidate functional SNPs that co-localize with these genomic marks and have also been shown to be eQTLs in published blood datasets.

These analyses were made possible due to the significant amounts of functional genomic data available for the cell line GM12878. Taking a deeper look at those results, there are a similar number of loci for which a candidate functional SNP can be found in DNase regions, ChrommHMM-Strong Enhancers, and Segway-Strong Enhancers (n=16, 12, and 17, respectively). While none of these 3 datasets were complete subsets of each other, there is significant overlap. However, as DNase hypersensitivity data can be obtained from a single assay as opposed to a combination of multiple assays for the segmentation data, in the case where data on relevant cell types do not yet exist, DNase data may be sufficient to identify putatively functional SNPs before investing the time and resources to generate all the assays needed for segmentation analysis.

These enrichment studies can provide valuable insight into the potential etiology of the disease of interest. For example, looking at the enrichment of lymphoma SNPs in ChIP-Seq data, we see an enrichment of risk SNPs in *RUNX3* binding sites ($p<0.0001$). *RUNX3* is a gene which is highly expressed in LCLs (Spender et al. 2002) and has been shown, paradoxically, to act both in promoting and suppressing tumor growth (Ito et al. 2015). We also observed enrichment of risk SNPs in binding sites for NfκB and TNF ($p<0.0001$); variation within these two pathways have also been shown to associate with non-Hodgkin's lymphoma risk (S. S. Wang et al. 2009). Lymphoma risk SNPs are also enriched at binding sites of TBLR1 ($p=0.0005$); disruptions at the TBLR1 locus in diffuse large B-cell lymphoma have been seen through a deletion of the locus (Pasqualucci et al. 2011) and the identification of a novel fusion between it and *TP63*, a paralogue of *TP53* (Scott et al. 2012).

It should be no surprise that we observed an enrichment of lymphoma risk-SNPs in the DNase hypersensitivity sites for multiple B-cells lines, as a majority of the reported SNPs are from studies on CLL. However, we also see an enrichment in lymphoid but non-B cell types such as adult CD4+ Th0 cells ($p < 0.0001$). If we slightly relax the stringent definition of significance and look to those lines that just failed to meet statistical significance, such as $p < 0.001$, the vast majority of those cells are also T-cells. These data makes sense since the B-cell and T-cells lines share a common lineage. This hypothesis is further supported as a hematopoietic progenitor cell line, CD34+ Mobilized cells, was approaching significance in the DNase hypersensitivity enrichment analysis (p=0.016). This type of analysis could also reveal novel insight into how different cell-types work in concert to lead towards the progression of a disease. For example, in our own results I saw an enrichment of lymphoma risk-SNPs in the DNase hypersensitivity track of human renal epithelial cells (DNase: p=0.0004). A literature search revealed that there is a known, rare disease – primary renal lymphoma – with 31 reported cases through 1991 (Harris & Lager 1991). Little is known beyond that is a non-Hodgkin's lymphoma affecting large B-cells; our results suggest that this rare condition may share commonalities between itself and more familiar lymphomas.

There are many different tools developed by the community which, using the same datasets and input, answer vastly different questions and provide complementary results. A summary of the different analyses are provide in Table 2.9. In brief, some of the tools, like ReglomeDB and HaploReg work best as efficient genomic annotators, though since their speed is attained by retrieving pre-calculated data from databases, there is not the flexibility there to ask the same question of enrichment in particular tracks of interest as can be done with UES. FunciSNP, on the other hand, provides the ability to look for the colocalization of SNPs though it's main purpose is to identify putatively functional SNPs based on these overlaps and doesn't provide a statistical measure of enrichment. GWAS-3D is unique among the aforementioned approaches as it considers distal interactions when determining whether a SNP is likely functional. Of all these methods, GoShifter is the most similar to UES in its aim and the questions it can answer though through an orthogonal approach and as such, it may be useful to use both methods when calculating enrichment in a genomic track file as a means of conformation.

Table 2.9. Comparison of different genomic analysis tools for GWAS SNPs.

	UES	GoShifter	RegulomeDB	HaploReg	FunciSNP	GWAS-3D
Run locally	✓	✓			✓	
Web-based			✓	✓		✓
Annotates individual SNPs			✓	✓	✓	✓
Results for SNP list as a whole	✓	✓				
Accounts for LD	✓	✓		✓	✓	✓
Adjustable R^2 threshold	✓			✓	✓	✓
Choice of population				✓	✓	✓
Provides output for any user supplied track	✓	✓			✓	
Motif alteration			✓	✓		✓
Considers distal interactions						✓
Calculates enrichment score	✓	✓		✓		
Enrichment calculation for any tack	✓	✓				
Identifies relevant tissue type	✓	✓				
Identifies relevant genomic mark	✓	✓				
Enrichment through SNP matching	✓			v3 - ✓ v4 - ×		
Enrichment by local shifting		✓				
Speed	✓	✓✓	✓✓✓	✓✓✓	✓✓	✓✓

Interestingly, while the lymphoma risk-SNPs were highly enriched in a tissue-specific manner in regulatory elements, analysis of the additional cancer types showed that this observation is not true across the cancer landscape. Breast cancer, for example is enriched solely in MCF-7, a breast cancer cell line, though they are not enriched in a tissue specific manner in other regulatory elements. Conversely, while the lymphoma risk-SNPs showed no enrichment in genic regions whereas there was significant enrichment in melanoma and esophageal cancer, however, as most of the esophageal cancer GWAS were performed in Chinese populations and UES that could confound the current results. Taken together, pan-cancer enrichment analysis suggests that not all inherited risk of cancer work identically and thus broad generalizations of how predisposition confers risk should be avoided. Furthermore, these results warrant more complete studies to better understand the etiology of these disease.

Overall, the data presented support the hypothesis that regulatory variants that influence transcription in cells of the lymphoid lineage contribute to inherited risk of lymphoma and chronic lymphocytic leukemia. However, based on enrichment of other cancer types, this hypothesis does not appear to hold true across different diseases. These results validate our computational approach that, moving forward, could provide novel insight into disease etiology when applied to other diseases.

3 Genome-wide association study of myeloproliferative neoplasms identifies *TERT* as a risk locus

The work presented in this chapter was conducted in conjunction with the lab of Dr. Ross Levine at Memorial Sloan Kettering Cancer Center. The GWAS data used for the subsequent analysis was both produced in and graciously provided by his lab.

3.1 Introduction

Myeloproliferative neoplasms (MPN), including polycythemia vera (PV), essential thrombocythemia (ET), and primary myelofibrosis (PMF), are chronic myeloid disorders characterized by clonal expansion of hematopoietic stem cells, and overproduction of mature blood elements (Campbell & Green 2006). The genetic basis for PV, ET, and PMF remained unknown until multiple groups identified a somatic activating mutation in the JAK2 tyrosine kinase (JAK2V617F) in ~90% of PV and in 50-60% of ET/PMF (Levine et al. 2006). Genetic instability may be induced also by mutations of genes involved in epigenetic regulation and chromatin remodeling such as *TET2*, *ASXL1*, and *EZH2* (A Tefferi et al. 2009; Carbuccia et al. 2009; Ernst et al. 2010). These latter somatic mutations might cause selective pressure for the acquisition of additional genomic lesions responsible for disease progression, such as *TP53* lesions (Harutyunyan et al. 2011). A close relationship was observed between aberrations of chromosome 9p (UPD and/or gain) and progression from PV to post-PV MF.

Prior genome wide association studies have identified a germline variant in the *JAK2* gene that predisposes to the development of *JAK2*-mutant MPN that are preferentially associated with specific MPN phenotypes (Kilpivaara et al. 2009; Jones et al. 2009; Olcaydu et al. 2009). Under a dominant genetic model, the risk genotype at *JAK2* rs10974944 contributes significantly to the excess familial risk of MPN (OR=3.1, population attributable risk=46.0%). These effects are most evident in

JAK2-positive MPN (OR=4.0, population attributable risk=55.3%), suggesting that germline variation at *JAK2* is a major determinant for the predisposition to develop *JAK2*-positive MPN. Our group has also observed that somatic *JAK2* mutations were most commonly acquired in *cis* with the *JAK2* predisposition haplotype, suggesting a direct interaction between haplotype-specific genetic variation in the *JAK2* locus and secondary acquisition of somatic mutations on the same strand (Kilpivaara et al. 2009).

However, since the study sample size of our previous MPN GWAS was small, we were unable to have complete coverage of the genome. Therefore, it was likely there are additional germline loci important in MPN predisposition and pathogenesis.

Here we report a larger genome wide association study to identify MPN risk variants. We tested 217 cases from our studies and an additional 361 cases obtained from publicly available cohorts along with 7,787 controls at over nine million SNPs imputed using data from the 1000 Genomes Project. Besides the previously known MPN risk locus at *JAK2* gene, we observed a statistically significant association signal at *TERT* (rs7717443; p-value=8.42x10^{-10} and OR=0.716, 95% CI=0.634-0.808).

3.2 Materials & Methods

3.2.1 SNP Array Analysis of MPN Samples

MPN patients were recruited in Boston and New York City under IRB approved protocols in which all patients provided informed consent. DNA was extracted from granulocytes and buccal swabs as previously described (Kilpivaara et al. 2009). 217 granulocyte DNA samples, which included 113 PV patient samples and 68 ET patient samples, were chosen for SNP array analysis based on clonality studies and JAK2V617F mutational burden in order to limit analysis to samples with

>80% MPN cells (Levine et al. 2006). DNA samples were genotyped using the Illumina Omni1-Quad genotyping array following the manufacturer's instructions.

To increase the power of this study, additional SNP array data was taken from several data sources. Additional data on 408 MPN cases was obtained from ArrayExpress (E-MTAB-608) from a genomic profiling study of MPN using the Affymetrix 6.0 SNP array3. Control data was obtained from both dbGaP (phs000187, phs000209v7, phs000167v1, phs000091v1) and from a study of schizophrenia in the Ashkenazi Jewish population (Vijai et al. 2011). For the Illumina data (phs000187 and the Ashkenazi schizophrenia study), data was merged after genotyping calling by the standard Illumina software independently in each study. For the Affymetrix data, genotypes were jointly called using the Birdseed algorithm using the Affymetrix Power Tools software from the raw CEL files.

The Affymetrix and Illumina datasets were then processed separately for quality control. SNPs were filtered on the following bases: call rate <98%, minor allele frequency <0.02 and Hardy–Weinberg exact test P< $1.0x10^{-7}$ in controls. Samples were filtered based on genotype quality control filtration (sample call rate <97%, gender mismatch). In total, 9,034,812 SNPs markers were identified for analysis and used in the merged case and control dataset.

3.2.2 Principal Component Analysis of MPN Patients/Controls

For principal component analysis we used all of the genome-wide data for our samples in order to correct for any chip and/or batch effects and, thus, allow for us to join the multiple datasets for joint analysis. Before analysis, we performed quality control filtering of both samples and SNP separately for cases and controls and then merged the dataset using the common set of SNPs present in the two cohorts. To do so, we first filtered out the ambiguous SNPs (A/T or G/C alleles) to ensure we unambiguously know strand when we merge the two datasets.

To investigate potential population stratification biases that could be introduced by the shared controls we performed principal component analysis using EIGENSTRAT (Price et al. 2006). To reduce the linkage disequilibrium between markers, we first used PLINK (Purcell et al. 2007) to filter markers such that all remaining markers are in low LD (r2 < 0.1, calculated in sliding windows 50 SNPs wide, shifted and recalculated every five SNPs). We applied the EIGENSTRAT program with default parameters and no outlier removal to infer axes of variation in the combined dataset. The case and controls that cluster together on the eigenvector plot (with the first two axes of variation) were used for the association analysis.

3.2.3 Imputation and association tests

We performed imputation analysis to merge the two datasets and determine the full extent of the whole genome, and to test for any untyped variants than those available on the original GWAS platform. After pre-phasing of the original data using SHAPEIT (Delaneau et al. 2012), genome-wide imputation was performed using IMPUTE2 (Howie et al. 2009) using 1000 genomes reference panel.

3.2.4 Association testing

To test for association of each imputed SNP with MPN risk, we used SNPTEST v2.5 beta 4 (Marchini & Howie 2010). Specifically, we used frequentist statistics to test for association under an additive model using the maximum likelihood (ml) fitting method in the program. We adjusted for the top 5 principal components of ancestry as well as a binary variable representing which chip a sample was genotyped on (Affymetrix or Illumina). Any SNP for which the information ("info" column in SNPTEST output) is < 0.4, a p-value was not computed, the minor allele frequency is <= 0.01, or the p-value for Hardy-Weinberg equilibrium in controls was <= 0.001 was removed. As the controls in both cohorts were genotyped at separate sites than the cases, we also removed any SNP whose frequency differed between the Illumina controls and the Affymetrix controls after adjusting for the top 5 principal components with $p<1x10^{-09}$ (Mukherjee et al. 2011). Finally, we computed the association separately in the Affymetrix and Illumina cohorts for each SNP with $p<1x10^{-05}$.

3.2.5 Pleiotropy analyses

To investigate if SNPs associated with blood phenotypes, inflammatory bowel disease, or cancer, are also MPN risk SNPs, we used several different data sources. We manually extracted SNPs that had been reported to associate with blood-related phenotypes in the NHGRI GWAS Catalog (Welter et al. 2014), resulting in 113 SNPs used in the analysis. Those SNPs were then extracted for analysis from our genotyping data and the following quality control filters were applied to keep SNPs in the analysis: genotyping rate > 0.05, individual missingness < 0.1, minor allele frequency > 0.05, and Hardy-Weinberg exact test of p<1.0x10-7. A list of 463 cancer-associated risk SNPs was generated from the same source (Welter et al. 2014) as the blood SNPs. Identical quality control was performed on the cancer SNPs. Additionally, 8 SNPs were shared between the blood and cancer lists. For each list, we asked if those SNPs were associated with MPN in our full, imputed data set.

3.3 Results

3.3.1 A larger genome-wide association study for MPN risk SNPs

To identify additional genetic variants associated with the risk of developing MPN, we conducted a genome-wide association study using denser genotyping platforms than we had used previously. Specifically, we genotyped 217 individuals with MPN and generated 2739 controls from studies of melanoma and schizophrenia; both sets were genotyped on the Illumina Omni1 Quad chip. Additionally, we obtained genome-wide SNP chip data from the Affymetrix SNP6 platform on 408 MPN cases and 5025 controls. We performed individual and SNP level quality control on the individual cohorts, jointly determined the principal components of genetic variation as a marker of genetic ancestry, and then imputed both datasets using data from the 1000 Genomes Project. This resulted in 9,034,812 SNPs genotyped in 578 cases and 7771 controls. We tested each SNP for association with the risk of developing MPN, adjusting for the top 5 principal components and genotyping platform. After removing SNPs that did not meet our stringent quality control criteria, we tested 8,010,302 SNPs for association. Of these, 1162 were significant after correcting for multiple testing ($p<6.2x10^{-09}$). However, upon further analysis of the date, we noted that there were a multitude of imputed SNPs widely disparate ORs when comparing the original genotyping platform. We hypothesized that these differences may be confounding our results and inflating the number of true associations. To correct for this, we removed SNP for which the estimated odds ratio on one platform was outside the 95% confidence interval estimated on the other platform. This resulted in 75 significant SNPs. Of these, 74 were at the previously identified *JAK2* locus, where a common risk haplotype greatly increases the risk of developing MPN. The most significant SNP at this locus is chr9:5074466:D (p = $5.06x10^{-60}$, OR = 3.17606, 95% CI = 2.81-3.58). The only significant SNP not at *JAK2* is rs7717443, at the TERT locus (p = $8.42x10^{-10}$; OR = 0.72, 95% CI = 0.63 - 0.81).

Table 3.1. Significant MPN GWAS hits at a Bonferoni Level

SNP	CHR	POS	OR (95% CI)	P
chr9:5074466:D	9	5074466	3.17606 (2.81434 - 3.58426)	5.06E-60
rs12348771	9	5083634	3.07135 (2.72164 - 3.46598)	2.17E-58
rs12343065	9	5083533	3.08042 (2.72972 - 3.47617)	8.20E-58
rs11788834	9	5092466	3.10975 (2.75568 - 3.50931)	1.98E-57
chr9:5090966:D	9	5090966	3.07617 (2.72601 - 3.47132)	3.38E-57
rs11788790	9	5092263	3.10058 (2.7476 - 3.4989)	4.62E-57
rs1034072	9	5088903	3.06834 (2.71913 - 3.46239)	1.65E-56
rs10283564	9	5075628	3.03685 (2.69125 - 3.42684)	2.31E-56
rs12349785	9	5076613	2.98901 (2.64863 - 3.37315)	4.23E-56

(continued) Table 3.1. Significant MPN GWAS hits at a Bonferoni Level

SNP	CHR	POS	OR (95% CI)	P
chr9:5090970:D	9	5090970	3.00607 (2.66397 - 3.39211)	1.23E-55
rs1159782	9	5078117	2.9886 (2.64843 - 3.37246)	2.91E-55
chr9:5076945:I	9	5076945	2.97928 (2.64025 - 3.36184)	3.63E-55
rs62543863	9	5085417	3.02019 (2.67637 - 3.40818)	4.64E-55
rs10283563	9	5075603	3.05575 (2.70802 - 3.44814)	1.95E-54
rs7038763	9	5076399	3.02286 (2.67892 - 3.41096)	8.28E-54
rs10974952	9	5079828	2.8473 (2.52268 - 3.21369)	2.59E-52
chr9:5137970:D	9	5137970	2.87064 (2.54204 - 3.24173)	9.13E-52
rs2146041	9	5262911	2.93143 (2.5981 - 3.30752)	1.02E-49
rs10975033	9	5262567	2.88817 (2.55979 - 3.25867)	4.20E-49
rs1887428	9	4984530	0.397599 (0.352353 - 0.448655)	6.74E-48
rs12350079	9	5259620	2.78915 (2.47202 - 3.14697)	9.45E-48
rs72703608	9	5258127	2.74781 (2.43513 - 3.10064)	2.90E-47
rs12351715	9	5262349	2.76196 (2.448 - 3.11617)	4.13E-47
rs11506293	9	5257430	2.77481 (2.4593 - 3.1308)	4.76E-47
rs11506292	9	5257048	2.76925 (2.45439 - 3.1245)	5.38E-47
rs28872016	9	5253558	2.75402 (2.44087 - 3.10735)	8.37E-47
rs12349113	9	5254224	2.72304 (2.41361 - 3.07214)	1.07E-46
chr9:5252803:D	9	5252803	2.72041 (2.41079 - 3.06979)	1.11E-46
rs12349508	9	5184222	2.89186 (2.56164 - 3.26466)	1.28E-46
rs10118267	9	5243736	2.75717 (2.44352 - 3.11107)	1.33E-46
rs7035456	9	5261440	2.74323 (2.43154 - 3.09489)	2.06E-46
rs7025005	9	5261794	2.74136 (2.42988 - 3.09276)	2.57E-46
rs60768043	9	5224676	2.75649 (2.44341 - 3.10968)	2.64E-46
rs13440043	9	5268264	2.74183 (2.4303 - 3.0933)	3.07E-46
rs11506668	9	5252789	2.70209 (2.39464 - 3.04902)	7.71E-46
rs10283473	9	5244708	2.69721 (2.39017 - 3.04369)	2.00E-45
rs7862042	9	5268139	2.72087 (2.4118 - 3.06955)	2.33E-45
rs1575285	9	5267440	2.71457 (2.40621 - 3.06246)	6.76E-45
chr9:5250918:D	9	5250918	2.65694 (2.3544 - 2.99836)	7.06E-45
rs1853221	9	5249364	2.65558 (2.35309 - 2.99696)	9.59E-45
rs36051895	9	4981866	2.42145 (2.14584 - 2.73245)	9.74E-45
rs11790680	9	5248768	2.65417 (2.35188 - 2.99532)	1.10E-44
rs10975028	9	5249020	2.64973 (2.34799 - 2.99025)	1.36E-44
rs10975027	9	5248827	2.64997 (2.3482 - 2.99052)	1.36E-44
rs2208685	9	5251758	2.64116 (2.34071 - 2.98016)	2.03E-44
chr9:5076938:I	9	5076938	2.6432 (2.3375 - 2.98888)	6.59E-44
rs11999802	9	5189773	2.8514 (2.52562 - 3.21922)	9.32E-44
rs2381215	9	5262607	2.59387 (2.29879 - 2.92682)	1.17E-43

(continued) Table 3.1. Significant MPN GWAS hits at a Bonferoni Level

SNP	CHR	POS	OR (95% CI)	P
rs10975024	9	5246403	2.64395 (2.34293 - 2.98365)	3.82E-43
chr9:4980929:I	9	4980929	2.32254 (2.05843 - 2.62053)	2.07E-42
chr9:4980756:D	9	4980756	2.26775 (2.01031 - 2.55816)	5.54E-41
rs10758669	9	4981602	0.461145 (0.408939 - 0.520016)	3.37E-39
rs7865719	9	5082333	1.99205 (1.75641 - 2.25929)	1.95E-32
rs1327497	9	4967539	0.538753 (0.470002 - 0.617561)	1.33E-28
rs10815141	9	4962247	0.566688 (0.494494 - 0.649421)	9.80E-25
rs113657238	9	5180065	2.43109 (2.09397 - 2.82249)	3.67E-23
rs62554837	9	5266200	2.17184 (1.90159 - 2.4805)	8.29E-22
rs9987451	9	5113452	1.63889 (1.45268 - 1.84896)	2.39E-21
chr9:5269166:I	9	5269166	2.15366 (1.8782 - 2.46951)	5.09E-21
rs72701691	9	5229419	2.40486 (2.06564 - 2.79979)	6.88E-21
rs1322223	9	5264425	2.0827 (1.82419 - 2.37785)	1.40E-20
rs1327500	9	4961260	0.583081 (0.513902 - 0.661573)	4.77E-16
rs72701653	9	5156285	2.38029 (1.97127 - 2.87418)	1.38E-15
chr9:5166295:D	9	5166295	2.36974 (1.96176 - 2.86256)	1.75E-15
rs2381216	9	5270603	2.1579 (1.81748 - 2.56208)	5.09E-14
rs59966455	9	5271028	1.88144 (1.61996 - 2.18512)	2.23E-10
rs72701669	9	5186616	0.285054 (0.160362 - 0.5067)	3.55E-10
rs72701644	9	5142495	0.282555 (0.158464 - 0.503819)	3.86E-10
rs72701646	9	5148564	0.283061 (0.158827 - 0.504473)	4.05E-10
rs143944808	9	5180144	0.346777 (0.216108 - 0.556454)	4.12E-10
rs56385018	9	5178033	0.28778 (0.162375 - 0.510037)	4.54E-10
rs72701648	9	5149890	0.345836 (0.214791 - 0.55683)	5.46E-10
rs7717443	5	1283486	0.71576 (0.633926 - 0.808158)	8.42E-10
rs62541542	9	5040876	0.527881 (0.395486 - 0.704596)	1.44E-09
chr9:5272475:D	9	5272475	1.87456 (1.60589 - 2.18817)	2.61E-09

3.3.2 Pleiotropy with inflammatory bowel disease

We noted that rs10758669, an SNP at JAK2 previously associated with risk of inflammatory bowel disease (IBD) (Polgar et al. 2012), was also an MPN risk SNP (p=3.4x10^{-39}, OR=0.46, 95% CI=0.41-0.52). We therefore wondered if there was a larger overlap between risk alleles for IBD and MPN. To address this, we collected a set of 164 IBD risk SNPs as enumerated in a recent paper (Jostins et al. 2012). For each SNP, we asked if it was associated with MPN risk. Of 161 IBD risk SNPs considered, only rs10758669 (3.373x10-39, OR=0.461, 95% CI=0.409-0.520) is associated with MPN risk after correcting for multiple testing. We next asked if a polygenic model built using the reported associations with IBD could predict MPN status in our data. For models build using ulcerative colitis alone or Crohn's disease alone, no association with MPN was observed (p=0.54 and 0.73, respectively). For a model built using SNPs associated with IBD in general, a significant association with MPN was observed (p=0.03). However, this association disappears when rs10758669 is removed (p=0.88), suggesting that the association was mediated solely through the JAK2 locus.

3.3.3 Pleiotropy with hematological traits

We next chose to investigate whether any additional SNPs that had been reported to associate with various blood phenotypes were also associated with MPN risk. Our working hypotheses was that by reducing the initial input set of SNPs, we would reduce the number of tests to correct for which would allow for the identification of true associations that may have failed to meet the most stringent Bonferoni threshold at a genome-wide level. Of the 113 blood-phenotype reported SNPs that passed QC, only one SNP, rs2736100, a SNP that had previously been identified as association with red blood-cell counts (Kamatani et al. 2010), showed a significant association with MPN once adjusted for multiple testing ($p=7.65x10^{-13}$, adjusted p=0.01153 OR=0.628, 95% CI=0.555-0.710). Notably, rs2736100 is found in the second intron of *TERT*, a gene which encodes a telomere reverse transcriptase and has been implicated with increasing risk for numerous cancer types and telomere length (Rafnar et al. 2009; Haiman, Chen, Vachon, et al. 2011; Huang et al. 2013; Bojesen et al. 2013).

3.3.4 Pleiotropy with cancer

Since numerous SNPs at TERT have been associated with the risk of developing cancer, we therefore asked if other known cancer risk SNPs may predispose to MPN. Of 389 identified cancer risk SNPs that met our quality control criteria, only rs10974944 at *JAK2* (p=4.501x10-58, OR=3.297, 95% CI=2.921-3.721) and rs2736100 at *TERT* (p=7.650x10-13, OR=0.628, 95% CI=0.555-0.710) were significant after accounting for 389 tests. Notably, these two SNPs showed differences in the odds ratios between the Affymetrix and Illumina samples that were outside of the respective 95% confidence intervals, which is why they were not reported in the initial GWAS described above. However these associations are validated by the fact that the 76 significant SNPs reported in the initial GWAS identified the same two loci. Taken together, these results confirm the role of *JAK2* in MPN predisposition and suggest an possible additional mechanism for increasing MPN risk, namely through modulating TERT activity and dysregulating telomere length.

3.4 Discussion and future studies

We have presented further analysis on the predisposition myeloproliferative neoplasm enabled by the combination of multiple datasets. By merging these multiple MPN GWAS, we were able to increase our statistical power and detect previously unreported variants as associating with increased risk of MPN. Our study identified two loci associated with MPN risk: the known *JAK2*-risk locus and the previously unknown *TERT* locus.

The *JAK2* locus at 9p24 as shown a strong association in both previous GWAS and is further confirmed by our expanded analysis. However the precise mechanism by which this locus increase risk remains unknown. One hypothesis stated that the risk haplotype, JAK2-V617F, predisposes one to MPN by creating a “hyper-mutable” phenotype of *JAK2* that ultimately leads to a dysregulation of STAT proteins and the PI3K-AKT and MAP-kinase pathways. However, previous work from our lab showed no difference in the amount of somatic mutations at JAK2 when comparing individuals homozygous for the risk haplotype to induvial homozygous for the protective haplotype (Mukherjee 2011). Additional, unpublished analysis from our lab also was unable to show a significant difference in the rate of single nucleotide variants present in the MPN cases relative to the ancestral sequence of the individuals of European ancestry in the 1000 Genomes Project Thus, the mechanism by which the *JAK2*-V617F haplotype predisposes to MPN remains elusive.

Further research is required to determine the precise mechanism by which the *JAK2* locus influences MPN risk. Since the risk V617F haplotype does not appear to act through hyper-mutability at the somatic level, nor is there a significant amount of variants in MPN patients when compared to the ancestral alleles, an alternative hypothesis states that there is some functional variant on the *JAK2*-V617F locus that cause an allele-specific expression or regulation of *JAK2*. Under such a hypothesis, the V617F somatic mutation may arise on all haplotypes at equal rates, but the risk haplotype may confer selective advantage to the V617F-positive clone. This could be examined by performing eQTL analysis, and correlating the genotype with gene expression levels from patients. A difference in expression would suggest that there is potentially a change in the regulation of *JAK2* between the haplotypes. It should be noted that our previous work did not provide of evidence that the risk haplotype was associated with *JAK2* expression levels and, as such, would argue against this hypothesis (Mukherjee 2011). However, there is the possibility that an effect that is either small in magnitude or that alters allelic ratios of expression without altering total expression levels may not have been detected previously. Additionally, this question could be addressed by examining patients who are heterozygous for the risk variants to determine if one of *JAK2* that is preferentially expressed in the individuals who are heterozygotes.

Conducting the pleiotropy analysis allowed us focus specifically on smaller sets of previously reported relevant SNPs and loci in order to reduce our multiple testing stringency threshold. As *JAK2* had previously been identified as a risk locus for developing IBD (Jostins et al. 2012), we interrogated the other IBD-associated SNPs with the goal that they may provide additional insight into the etiology of MPN. However, the only significant IBD SNP, rs10758669, was at the *JAK2* locus, thus reaffirming the previously observed association but failing to provide additional, novel insights about MPN predisposition.

The cancer pleiotropy analysis further confirmed the association of increased MPN risk at the *JAK2* locus with the identification of rs10974944. Additionally, the analysis of both cancer and blood-trait associated SNPs shared a single commonality: rs2736100. This SNP piqued our interest as it is located at 5p15 and is located in the second intron of *TERT*, a gene which has been implicated in numerous cancer types (Rafnar et al. 2009; Haiman, Chen, Vachon, et al. 2011; Huang et al. 2013; Bojesen et al. 2013). This result suggests that there may be mechanistic commonalities between MPN and the other cancer types. While this SNP has previously been shown to be associated with erythrocyte counts in a Japanese population (Kamatani et al. 2010), this is the first replication of this SNP in individuals of European descent. As rs2736100 is found in the intron of *TERT*, this suggest that the SNP may change the regulation and function of TERT. Variants for this SNP have been shown to associate with the mean telomere length in the Han Chinese and replicated in Europeans, with the C-allele (the protective allele) associating with longer telomere length (Liu et al. 2014). Further analysis would be needed to see if this association hold true in a cohort of MPN patients, linking the genetic variant to both telomere length and disease.

3.5 Conclusion

This chapter described the work done to follow up on previous studies of MPN in our lab and to better understand the etiology of MPN. Combining our MPN samples with an additional publically-available cohort, we were able to successfully conduct a GWAS that replicated the importance of the *JAK2* locus in predisposition to MPN and additionally identified *TERT* as an additional risk locus of MPN. Subsequently, using previously reported risk SNPs to inflammatory bowel disease, various blood phenotypes, and numerous cancers, we were able to identify additional SNPs at the *JAK2* and *TERT* loci associating with MPN risk and, thus replicate our findings from the full GWAS. However, as the precise mechanism by which these loci confer risk remains unknown, further research is needed in order to answer the remaining questions.

4 Investigating the function of microseminoprotein-beta in prostate cancer

4.1 Introduction

In 2008, two groups published separate genome-wide association studies both identified the SNP rs10993994 as associating with prostate cancer risk (Eeles et al. 2008; Thomas et al. 2008). Since then, this association has been replicated in numerous studies in additional populations beyond Europeans (Chang et al. 2009; Lou et al. 2009; Takata et al. 2010; Haiman, Chen, Blot, et al. 2011; Lange et al. 2012; Xu et al. 2012). The concordance of the evidence across multiple studies from disparate populations suggests that this SNP or a highly-correlated variant is indeed a function SNP.

This risk SNP, rs10993994, is a C/T polymorphism and is found at 10q11. Interestingly, this SNP is found just 57 base pairs (bp) upstream of the TSS for the gene *MSMB*, a gene that encodes one of the major secreted proteins of the prostate. This protein, microseminoprotein-beta (β-MSP), is a small protein consisting of only 94 amino acids and, as such, was originally identified as "prostatic secretory protein of 94 amino acids" or PSP94 (Seidah et al. 1984; Sheth et al. 1984; Akiyama et al. 1985). This 16 kiloDalton protein has 10 evolutionarily conserved cysteine residues between human and mouse (Xuan et al. 1999).

The physiological role of β-MSP is not fully known, though as it is one of the 3 major secreted proteins of the prostate, there is a body of work examining its role in prostate cancer and its ability to be used as a prognostic marker of prostate cancer. In one study, patients that had higher levels of β-MSP had a higher chance of survival after having a radical prostatectomy (Bjartell et al. 2007). Additionally, staining for β-MSP has shown that prostate cancer tissue has lower levels of β-MSP staining when compared to benign tissue (Whitaker et al. 2010). Though the precise role of β-MSP is not yet know, it has been shown in numerous studies to inhibit cell growth. A dose-dependent decrease in growth was observed when adding exogenous β-MSP to PC3 cells in culture (Garde et al. 1999). This inhibitory effect has also been observed in prostate cancer xenograft models, leading to smaller tumor sizes in mice (Garde et al. 1999; Shukeir et al. 2003) which led to the isolation and development of a synthetic 15-mer which has been shown to have the same effect (Shukeir 2004). Lastly, at the genetic level, knocking down *MSMB* in the LHS-AR prostate cell line allowed the cells to achieve anchorage-independent growth (Pomerantz et al. 2010). Additionally, that group also showed that the risk-SNP, rs10993994, acts as an eQTL for *MSMB* (Pomerantz et al. 2010).

Our group has previously shown that rs10993994 is correlated with prostatic secretions of β-MSP in both the blood and semen of healthy, young men, with the protective C-allele associating with higher levels of β-MSP and the risk T-allele association with lower levels (Xu et al. 2010). Following up on that work, our group attempted to elucidate the mechanism by which rs10993994 regulates *MSMB*. The rs10993994-C allele is predicted to be at a CREB-binding site; we hypothesized that the T-allele would disrupt the binding affinity. However, a siRNA-mediated knockdown of CREB1 did not alter the transcription activity between the two alleles. Additionally, ChIP experiments did not show an allele-specific preference sequence they pulled down (Xu 2014). The regulatory mechanism remains yet unknown.

Simultaneously, our group also began to investigate the mechanism by which cell growth is reduced. We also observed the dose-dependent decrease of prostate cancer cells when treated with exogenous β-MSP though the mechanism was still undetermined. However, in contrast to a previously reported study (Garde et al. 1999), we did not see the decrease in viability as a function of increased apoptosis when measured by caspase 3/7 activity (Figure 4.1). Thus, the mechanism of β-MSP's tumor-suppressive properties also remains unknown.

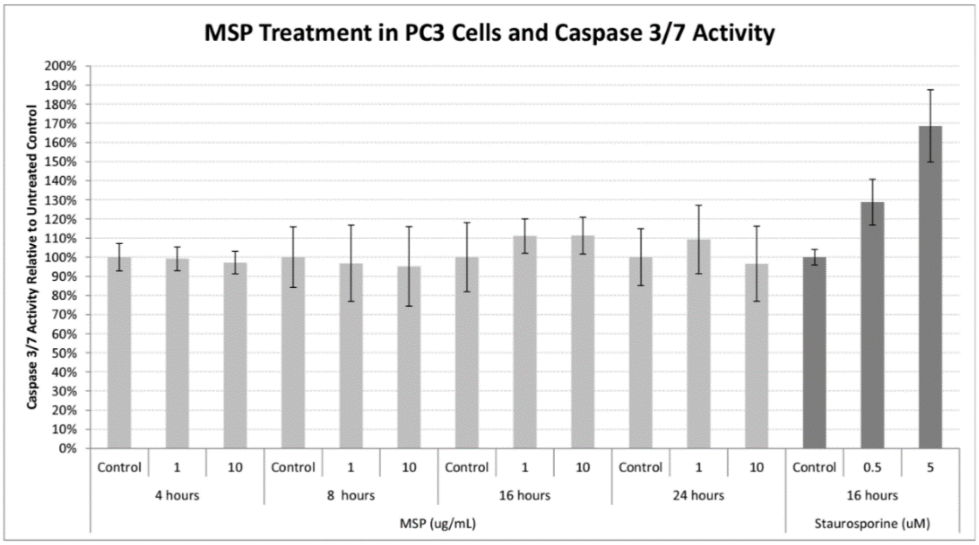


Figure 4.1. Apoptosis is not induced in PC3 cells by addition of β-MSP. Caspase 3/7 activity, a readout of apoptosis, remains unchanged between two treatments of exogenous β-MSP when compared to controls. Staurosporine, known to induce apoptosis, is included as a positive control.

We hypothesize that β-MSP has tumor-suppressor in the prostate. As such, the risk-SNP, rs10993994, disrupts regulation of *MSMB* which, in turn, leads to lower levels of β-MSP and eventually contributes to prostatic tumorigenesis. The goal of this chapter is describe the work performed to determine the mechanism by which β-MSP decreases viability in cells and thus contributes to prostate cancer.

4.2 Materials & Methods

4.2.1 Cell Lines

The cell lines used for experimentation were mostly obtained from the American Type Culture Collection (ATCC) with the exceptions being noted. The cells were grown at standard cell-culture conditions (37°C, 5% CO_2) in the appropriate media and supplemented with 10% FBS, 1% penicillin/streptomycin, and 0.2% fungizone. Particular growth conditions are given in Table 4.1.

Table 4.1. Cell lines used in experimental work.

Cell Line (Citation)	Description	Cell culture notes
AGS (Barranco et al. 1983)	Gastric cancer cells	F12K media
BPH-1 (Hayward et al. 1995)	Immortalized prostate epithelial cells	Keratinocyte serum-free medium (KSFM); 50 µg/ml bovine pituitary extract; 5 ng/ml EGF.
DU145 (Stone et al. 1978)	Prostate cancer cells	EMEM Media
LHSR (Berger 2004)	Immortalized prostate epithelial cells	PREGM bullet kit; cells provided by William Hahn
LNCaP (Horoszewicz et al. 1983)	Androgen-dependent epithelial cells	RPMI 1640 media
Myc-CaP (Jiao et al. 2007)	Murine, prostate cancer cell line	DMEM media, high-glucose (4.5 g/L; cells provided by Charles Sawyer's lab
RWPE-1 (Bello et al. 1997)	Immortalized prostate epithelia cells	KSFM; 50 µg/ml bovine pituitary extract; 5 ng/ml EGF.
PC3 (Kaighn et al. 1979)	Prostate cancer cells	F12K media
VCaP (Korenchuk et al.)	Immortalized prostate epithelial cells	DMEM media; cells provided by Charles Sawyer's lab

4.2.2 Overexpression of *MSMB*-plasmids

Cells were plated to achieve 90% confluency the night before the transfection. The following day, the *MSMB*-plasmids were transfected into the cells using Lipofectamine 2000 (ThermoFisher Scientific, 11668500). For over-expression, functional assays performed in 24-well plates, 0.8 ng of total plasmid was diluted in 50 µl of Opti-MEM Reduced Serum Media (ThermoFisher Scientific, 31985070) per well which was combined with 2 µl of Lipofectamine 2000 diluted in 50 µl of Opti-MEM Reduced Serum Media according to the provided protocol. For transfections performed in 6-well plates, 4 µg of plasmid DNA and 10 µl were diluted in 250 µl of media, respectively. The transfection-media was removed and fresh media was re-added after 4 hours. Depending on the subsequent assay, the cells were allowed to grow for 24, 48, or 72 hours before they were tested.

4.2.3 Cell viability

The viability of cells was measured by use of the cell health and viability indicator alamarBlue (ThermoFisher Scientific, DAL1100). The substrate of alamarBlue, resazurin, is reduced by living cells to resorufin, a fluorescent molecule. For experiments done in 96-well plates, 10 µl of alamarBlue was added to the 100 µl of media already in the wells of the plate. For experiments performed in 6-well plates, 200 µl of alamarBlue was added to the 2 ml of media in the plate wells. Plates were incubated for 1 hour at 37C and read using a fluorescence spectrophotometer.

4.2.4 Cell-cycle analysis

PC-3 Cells were grown in a T-75 flask until they were 90% confluent, split 1:2, and re-plated into 6-well plates. Once they had reached 60-70% confluence, they were transfected with the appropriate *MSMB*-overexpression plasmids or control treatment using the standard Lipofectamine 2000 procedure. Cells were collected at 24, 48, and 72 hours. For collection, first they were washed with 3 mls of PBS and then treated with 1 ml of Trypsin (ThermoFisher Scientific, 25300-054) allowing for removal from. Once the digestion was complete, 3 mls of Trypsin was added to deactivate the trypsin. The cells were colleted in 50 ml Falcon tubes and spun down for 5 minutes at 450 g at room temperature. The supernatant was removed and the cells were washed with PBS and spun down again. The cells were spun down again, and once the supernatant was removed, they were resuspended in 500 µl of 70% EtOH.

The cells were transferred to round-bottomed 96-well plates for analysis on a Guava PCA-96 flow cytometer. The plate was spun down for 5 minutes at 450 g at room temperature after which the supernatant was removed. Cells were washed using 200 µl of 1X PBS. After washing, the cells were stained with 200 µl of the Guava Cell Cycle Staining reagent, a propidium iodide solution, and analyzed on the Guava PCA-96 system.

4.2.5 Staining for senescence

Cells used in the senescence assays were plated in 6-well plates containing microscope slide covers in each well and transfected in an identical manner as previously described. Cells were collected at five different timepoints: 24h, 36h, 48h, 60h, and 72h. Once sufficient time had passed since transfection, the slide covers containing the cells were collected, washed with 2 mls of PBS 3 times for 2 minutes each on an orbital shaker. The cells fixed using 1.5 ml of a fixing solution (6190.8 µl PBS, 356.4 µl formaldehyde, and 52.8 µl glutaraldehyde) and shaking for 5 minutes on an orbital shaker. The fixing solution was removed and the cells were immediately washed with 3 mls of PBS on the orbital shaker. The cells were stained with a staining solution (300 µl 1 mg/ml Xgal in DMF, 1.2 ml citric acid/sodium phosphate solution, 300 µl 100 mM potassium ferrocyanide, 300 µl 100 mM potassium ferricyanide, 180 µl 5M NaCl, and 12 µl 1M $MgCl_2$) for 16 hours at 37C with no CO_2 circulation. Slide covers were mounted and stored at 4°C until analysis.

4.2.6 KI-67 immunohistochemistry staining

PC3 cells were grown on microscope chamber slides and transfected for one of the following conditions: no treatment control, Lipofectamine 2000 only control, empty PCDNA3 vector and MSMB-PCDNA. Transfections were performed according to the standard protocol as previously given and cells were collected at 24 and 48 hour timepoints. The cells were fixed and passed along to the Immunohistochemistry Core facility at MSKCC for staining with a KI-67 antibody.

4.2.7 Phospho-kinase analysis

The Proteome Profiler Human Phospho-Kinase (R&Dsystems: ARY003B) was used to determine signaling changes for numerous pathways simultaneously. PC3 cells were transfected with transfected with either the empty PCDNA3 plasmid or the MSMB-PCDNA3 construct as previously described. Transfections were done in duplicate. The standard protocol for the proteome profiler assay provided with the kit was performed as directed. The intensities of the individual dots were extracted and analyzed using ImageJ v1.45 (Schneider et al. 2012) to calculate the ratio between the two transfection treatments.

4.2.8 Exogenous Microseminoprotein-beta

Microseminoprotein-beta (β-MSP), the product of *MSMB*, was generously provided by Hans Lilja at MSKCC. It was dissolved in water to a concentration of 5 mg/ml, aliquoted, and frozen until use. The cells were treated at a final concentration of 1 mg/ml.

4.2.9 Western blotting

Total cell protein extracts were prepared using a nonindet-P40 detergent lysis buffer along with a protease inhibitors. Once the protein concentration was measured with a standard BCA assay, it was run on an SDS-PAGE gel and transferred to a PVDF membrane using a BioRad semi-dry transfer module. Blotting was performed using an anti-beta-catenin antibody from OriGene. The Pierce ECL reagent from ThermoFisher Scientific was used to visualize the blots.

4.3 Results

Previous work from our group has confirmed that there is a cell-type specific growth-inhibitory effect on prostate cancer cells: there is a dose-dependent decrease in viability of PC3 and LNCaP with higher levels of β–MSP, while no effect was observed in DU145 or RWPE-1. This result was also observed when *MSMB* was transfected and overexpressed in PC3 cells. However, there was no supporting evidence that this decreased viability was due to increased apoptosis as measured by Caspase 3/7 activity (Xu 2014).

4.3.1 Cell Cycle Profiling

One hypothesis was that β–MSP was causing prostate cancer cells to arrest in a particular stage of the cell cycle and causing decreased viability. The cell cycle profiles of PC3 were analyzed after transfection with a *MSMB*-overexpression plasmid. There was no observable difference between the cell cycle profiles of the *MSMB*-overexpressing cells when compared to controls (Figure 4.2).

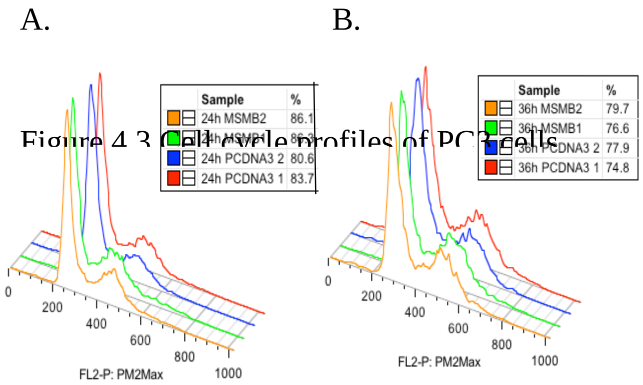


Figure 4.3 Cell cycle profiles of PC3 cells

Figure 4.3 Cell cycle profiles of PC3 cells. Cells were transfected with either an *MSMB*-overexpression or control vector. Profiles shown are for 24 hours (A) and 36 hours (B). There is no observable difference among the G1, S, or G2/M stages.

4.3.2 Proliferation analysis

We next chose to examine whether or not β–MSP caused a change in the proliferation of prostate cancer cells as measured by Ki-67 staining. Low levels of Ki-67 are observed in the nucleus of cells at the G1 phase and increase the levels of the protein increase with progression to the S and G2/M phases. As done for the cell cycle assay, PC3 cells were transfected with either control or *MSMB*-overexpression vectors for 24 and 48 hours. There was no observable difference between the Ki-67 staining at either time point at all treatments (Figure 4.3).

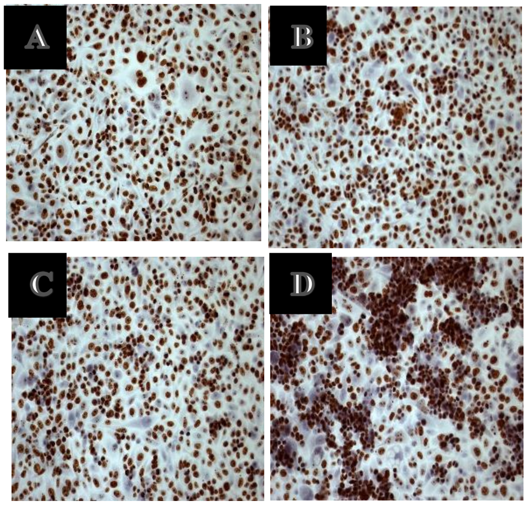


Figure 4.5 Ki-67 staining of PC3 cells at 24 hours. Cells positive for PC3 are stained dark brown. There is no difference between A) no treatment control, B) Lipofectamine only control, C) MSMB overexpression, or D) empty-vector control.

4.3.3 Senescence analysis

Since the decrease in cell viability was not due to cell cycle arrest, decrease in proliferation, or an increase in apoptosis, we proposed that β-MSP inhibited growth by causing the cells senesce, or an arrest of growth, rather than death. PC3 cells were subject to the same controls and overexpression transfections as previously reported. Cells were then stained for the presence of senescence-associated beta-galactosidase (SA-β-gal), a phenomenon by which senescing cells express beta-galactosidase activity (Dimri et al. 1995; Lee et al. 2006). As with previous functional assays, there was no observable difference among the SA-β-gal treatments between conditions (Figure 4.4).

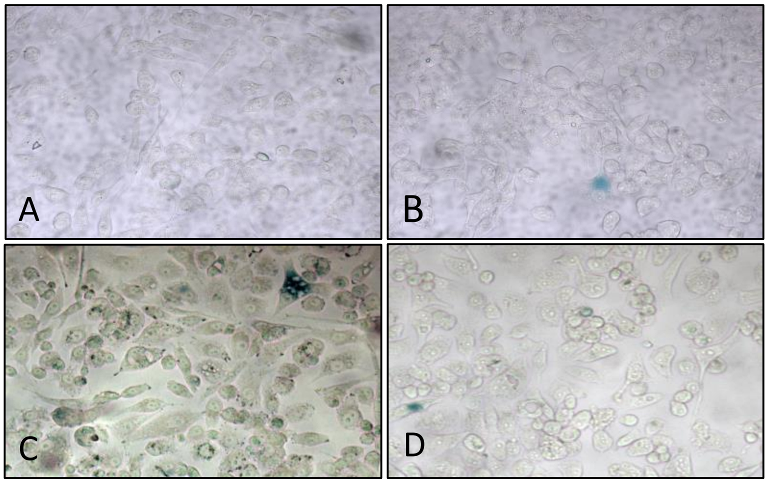


Figure 4.7. SA-β-gal staining of PC3 cells. Cells positive for SA-β-gal are stained dark blue. There is no difference between A) no treatment control, B) Lipofectamine only control, C) MSMB overexpression, or D) empty-vector control.

4.3.4 Cell Viability

We next performed a series of experiments using both *MSMB*-overexpression plasmids and exogenous β-MSP obtained from human semen samples. We transfected both BPH-1 and PC3 cells as previously described and determined their viability after 24 hours with alamarBlue. There was no difference among the treatments for the non-cancerous prostate line BPH-1 (MSMB-O/E vs empty vector, p=0.25). While there was a significant difference between the “no treatment control” and “MSMB overexpression” (p=0.02), there was no significant difference between the overexpression vector and empty vector (p=0.45). Similar trends were also observed in PC3 when media from control and treated cells were transferred to untreated cells, indicating that we were observing a true effect (MSMB-O/E vs empty vector, p=0.58; Figure 4.5).

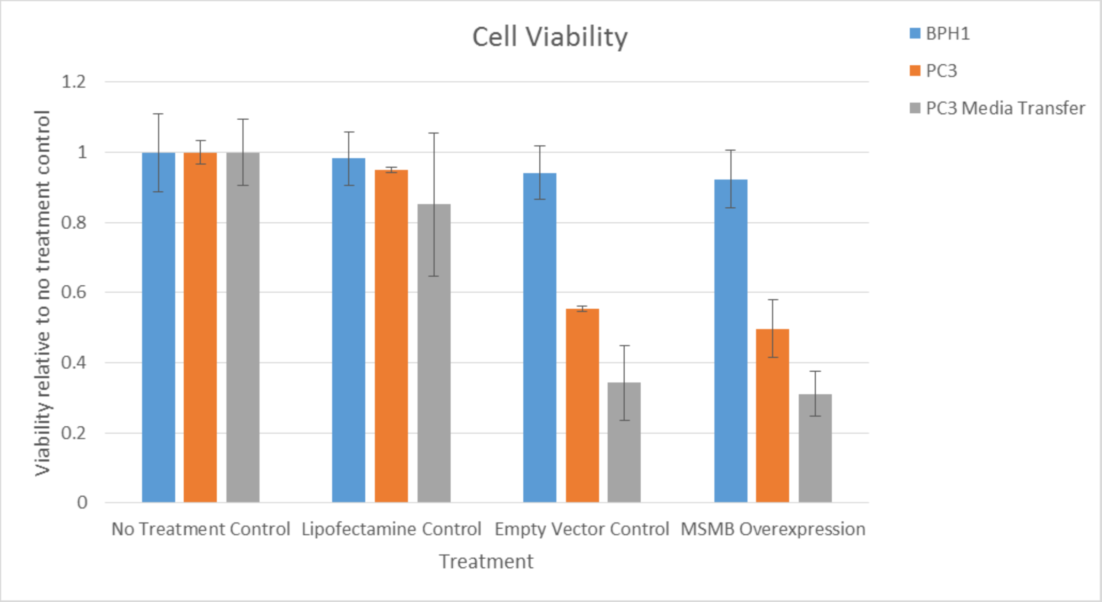


Figure 4.10. Cell viability assays. Cells were transfected with control vectors or the MSMB overexpression vector and their viability was assayed 24-hours post transfection with alamarBlue. General patterns held true across BPH-1 (blue), PC3 (orange), and PC3 where media from transfected cells was transferred (gray).

One critique of overexpression vectors is that you cannot control the amount of protein that is overexpressed; it may not be in the normal range of physiological conditions. To test for this, we used exogenous β–MSP from the Lilja lab at MSKCC, dissolved in water, and directly applied to PC3 cells at both physiological concentration (1 mg/ml) and 1/10th of physiological. The viability of the cells was assayed after 24 hours. While there was a decrease in viability from the non-treated cells to 1/10th physiological, to physiological conditions (1, 0.995, 0.939, respectively), there was no significant difference between the physiological condition and the water control (0.979, p=0.41, Figure 4.6).

4.3.5 Kinase signaling analysis

Since the mechanism was not immediately clear on how the previously described growth inhibition was mediated, we utilized a human phospho-kinase

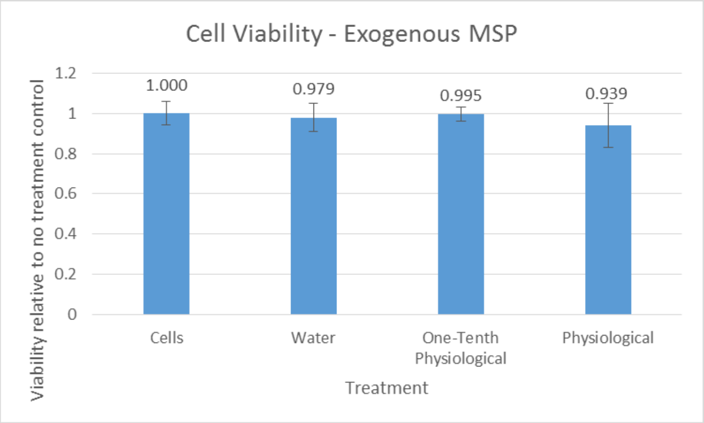


Figure 4.11. PC3 viability with exogenous β–MSP. Purified human β–MSP was added to PC3 cell cultures and cell viability was measured after 24 hours. While there appeared to be a dose dependent reduction in viability, it was not significantly significant.

antibody array from R&D systems. This system contained antibodies blotted on membranes for 43 different human kinases, allowing us to quickly assay to determine the relative levels of kinase phosphorylation for all targets simultaneously. PC3 cells were transfeted with the *MSMB*-overexpression or empty vector control constructs in duplicate. After a 24-hour transfection, the antibody array indicated that there were 9 different kinases that showed at least a 2-fold change between overexpression and control (Figure 4.7, Table 4.2).

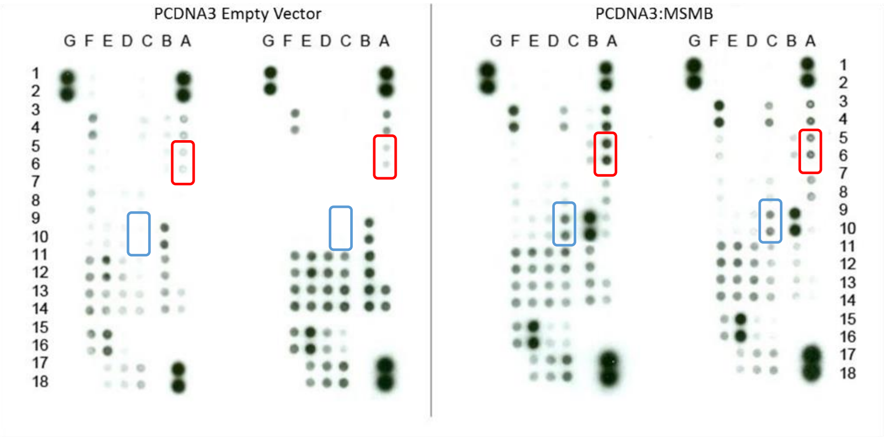


Figure 4.12. Human phospho-kinase antibody array for PC3 cells. The kinases with the largest changes are outlined: ERK1/2 (red, 6.46), and beta-catenin (blue, undetected to detectable with MSMB overexpression).

Table 4.2. Largest changes in phosphorylated kinases on antibody array MSMB-overexpression vs. control.

Target	Phosphorylation site	Ratio (MSMB:Control)
B-Catenin	Presence or absence	Detectable in Treatment
ERK1/2	T202/Y204, T185/Y187	6.46
p27	T157	3.36
Akt	S473	2.767
p38-alpha	T180/Y182	2.53
PLC-gamma-1	Y78	2.32
Chk-2	T68	2.25
Paxillin	Y118	2.22
p70 S6 Kinase	T421/S424	2.08

We next followed up by overexpressing *MSMB* in PC3, LNCaP, and RWPE-1 cells and performed western blots for beta-catenin. However, for the antibodies we used, we were able to detect the presence of beta-catenin in every cell line for every treatment (Figure 4.8) Additional work will be required to determine to validate this observation and to reconcile our western blot with the kinase array.

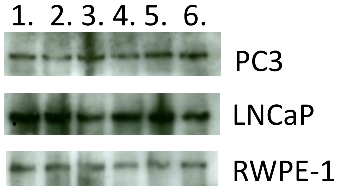


Figure 4.13. Western blot for beta-catenin in various cell lines. The treatments are as follows: 1) no treatment, 2) PCDNA3 empty, 3) PCDNA3-MSMB, 4) PCDNA3-MSMB Δ2-20, 5) PCDNA3.1 empty, and 6)PCDNA3.1-MSMB. Beta-catenin was detected for each treatment in each cell line.

4.4 Discussion

It has been well established that β–MSP reduces cell viability when added exogenously or overexpressed, however, the mechanism by which this happens has yet to be understood. Since previous work from our lab did not observe an increase in apoptosis (Xu 2014) although it had previously been reported based on microscopy alone (Garde et al. 1999), we chose to follow up and better understand this growth-arrest process.

We first examined whether or not cell viability was decreased due to cells becoming arrested in a particular stage of the cell cycle. Cells were transfected to overexpress *MSMB* or control vectors, fixed, and stained with propidium iodide for analysis by flow cytometry. However, there was no noticeable changes in the profiles of the controls compared to overexpression. Additionally, this analysis provided insight into whether or not apoptosis is responsible for the observed effect. If the overexpression construct caused a greater number of dead cells, there would be a larger fraction of cells in the sub-G1 phase causing the cell-cycle profiles to shift to the right. However, as this shift was not present in any of the cell cycle profiles, this indicates that there was not a marked increase in cell death.

We have observed the same repeatable decrease in cell viability when cells are β–MSP is added exogenously or overexpressed. However, we must consider the assay being used for our observations: alamarBlue. Since alamarBlue's reduction of resazurin to resorufin is a function of the number of live cells, this could also mean that our observed decrease in viability is not actually a decrease in viability but rather that cells proliferate slower in higher concentrations of not β–MSP grow slower. This slower growth would lead to fewer cells by the end of the assay growth period and thus provide a lower fluorescent readout. We analyzed this by staining the cells for Ki-67, a marker of cellular proliferation. However, once again there was no discernable difference between the β–MSP overexpressing cells and the controls (Figure 4.3). Similarly, we did not see an increase in senescencing cells, akin to the slowing of cell growth, when base treated with β–MSP. Additionally, the other viability experiments, through both overexpression and the addition of exogenous β–MSP once again confirm that the observed effect is quite subtle. Based on these results we can conclude that β–MSP does not decrease cell viability by causing cell-cycle arrest or senescence.

The phosopho-kinase antibody array was an efficient way to assay a multitude of potential signaling pathways that could be affected by an increased level of β–MSP. However, the results do not make the mechanism abundantly clear. The most striking increase when comparing the two conditions was the observation of beta-catenin's presence in the *MSMB*-overexpression when it was previously unexpressed in the controls. This is in conflict with the expected result as higher levels of β–MSP lead to a decrease in cellular proliferation whereas increased levels of beta-catenin has been shown to accompany increased cellular proliferation (Sellin et al. 2001). Additionally, we observed the presence of beta-catenin in multiple cell lines, both treated and untreated, by western blot. Further analysis is required to determine the difference between the antibodies used for both sets of experiments.

The signals with the greatest fold-change on the phosopho-array ran counter to the expected direction. For example, we observed a nearly 6.5 fold increase in phosho-ERK in the overexpression condition. Activation of ERK has been shown to increase as normal cells progress to cancer (Grubb et al. 2003). Again, we would have expected the inhibitory nature of overexpressing *MSMB* to lead to a decrease in ERK activation. Also, activated AKT was increased almost 3-fold in the overexpression compared to control even though activated AKT has been shown to protect LNCaP cells from TRAIL-induced apoptosis (Nesterov et al. 2001). AKT that is phosphorylated at S473, the phosphorylation mark tested for on the antibody array, has previously been shown to phosphorylate androgen-receptor at S210 and lead to AR transcriptional activity (Ha et al. 2011; Facompre et al. 2010).

Thankfully, not all of the activated kinases are in the wrong direction. For example, we observed a 3.3-fold increase in p27 in the presence of increased *MSMB.* It has been reported that lower levels of p27 predict poor disease-free survival in prostate cancer patients (Yang et al. 1998). The results from the antibody array are not clear cut. Much more precise research will be required to fully understand the signaling pathways that are affected by β–MSP.

The true function of β–MSP and the mechanism by which it acts has remained elusive even after 3 decades of research. These most recent experiments once again confirm that there is a true, though incredibly subtle effect that β–MSP has on cells. Nevertheless, this does make sense in the concept of the genome-wide association studies. The reported ORs for rs10993994 (1.11-1.27) show that the risk is only slightly increased when compared to the protective allele. While this effect is most likely true, it does not appear that it is sufficient enough to lead to prostate cancer on its own, rather it will work in a systems context. It is in this context that the work must be done if we want to understand the mechanism of this protein.

5 Conclusions

While genome-wide association studies have been successful in identifying regions and variants that predispose to multiple cancers, the field has much more work to do in functionally describe these loci. The overall goal of my thesis was to both generate and employ various experimental methods to characterize the reported variants.

We first created a computational approach to characterize previously reported GWAS using publically-available epigenomic datasets, such as those from ENCODE and Roadmap Epigenomics. Our method, entitled “Uncovering Enrichment through Simulation (UES),” employs a SNP matching technique to calculate the empirical enrichment of the input SNPs in various user-supplied genomic annotation tracks. During the construction of the pipeline, we took great care to properly control the random SNP selection process and ensure that the random sets were architecturally similar to the input set with regards to the presence on a genotyping platform, the number of LD-partners, and its distance to a TSS.

We validated the pipeline using a set of lymphoma SNPs curated from the NHGRI-EBI GWAS catalog. Analysis of these SNPs using UES revealed a tissue-specific enrichment of these risk SNPs in DNase hypersensitivity sites and enhancer loci. These SNPs were not enriched in similar tracks from additional cell lines, suggesting that this observed result is a true observation of the nature of lymphoma risk. Complementary analysis of both breast and prostate cancer risk SNPs, did not shown enrichment in the LCL tracks which had previously shown enrichment for the lymphoma SNPs, confirming validating the observation and excluding the possibility that the enrichment was observed strictly as a function of the tracks analyzed.

Our analyses lay the groundwork for additional follow-up. With lymphoma, our study suggests the dysregulation of regulatory elements is partially responsible for lymphomagenesis, Additional wet-lab work, such as molecular cloning of high-priority variants, will be required in order to more completely understand the context in which the SNPs function. Additionally, when we expanded our analysis to the enrichment of other cancer types, interestingly, the observed pattern of tissue-specific, regulatory enrichment did not hold true among the cancers. For example, esophageal cancer was found to be significantly enriched in both genic regions and active transcription start sites across multiple cell lines, thought, unlike lymphoma, it showed no enrichment in DNase hypersensitivity sites nor enhancer regions. This supports the notion that the generalized use of "cancer" is much too broad of a classifier; different cancers diseases develop and manifest by unique methods. Each cancer type, and even subtype, would require a more detailed analysis to better understand the context of their representative risk SNPs.

Next, we combined multiple GWAS for myeloproliferative neoplasm in order to increase statistical power and thus allow for the identification of novel associations. Our analysis revealed two risk loci at *JAK2* (75 sub-Bonferoni SNPs) and *TERT* (rs7717443; p-value=8.42x10^{-10} and OR=0.716, 95% CI=0.634-0.808). Subsequent pleiotropy analysis to IBD, various blood phenotypes, and numerous cancers, we were able to identify additional SNPs at the *JAK2* and *TERT* loci associating with MPN risk and, thus replicate our findings from the full GWAS. While there have been multiple hypothesis suggesting how the JAK2V617F haplotype risk, such as through hyper-mutability, though the precise mechanism by which these loci confer risk still unknown. The *TERT* risk locus suggests a possible mechanism of increasing risk through a dysregulation of telomere-length, though further analysis is required to confirm the associations of SNP alleles at *TERT* with telomere length.

Lastly, we spent great effort to understand the molecular mode of action by which *MSMB* is involved in PrCa etiology. Since the risk allele of rs10993994 associated with lower levels of both transcript and protein levels and β-MSP has been shown to reduce viability of prostate cancer cells, we had hypothesized that β-MSP has putative tumor-suppressive properties and controls proliferation of prostate cells *in vivo.* Previous work from members of our lab showed that the decreased viability of prostate cancer cells was not due to apoptosis. Thus, we explored different means by which the cell growth may be decreased including: senescence, cell-cycle profiling, and staining for marks of proliferation. However, none of these analyses provided any significant difference between treated and untreated, leaving the mode of action an open mystery. We identified a few differentially expressed kinase pathways when comparing *MSMB*-overexpression transfected cell lines against controls. This analysis too did not provide a definitive answer as some signaling cascades were modulated in the logically appropriate direction, such as p27, and others seemingly in the opposite of the expected direction, i.e. ERK and AKT.

To concluded, as the speed and ease of performing GWAS continue to increase, efforts to functionally validate the reported variants must increase even more so. Validation efforts will provide valuable, concrete understandings of the genetic etiology of diseases, opening the door to better screening and treatment techniques, and ultimately reducing human mortality.

BIBLIOGRAPHY

Akiyama, K., Yoshioka, Y., Schmid, K., Offner, G.D., Troxler, R.F., Tsuda, R. & Hara, M., 1985. The amino acid sequence of human beta-microseminoprotein. *Biochimica et biophysica acta*, 829(2), pp.288–94.

Altshuler, D., Daly, M.J. & Lander, E.S., 2008. Genetic mapping in human disease. *Science (New York, N.Y.)*, 322(5903), pp.881–888.

Altshuler, D., Hirschhorn, J.N., Klannemark, M., Lindgren, C.M., Vohl, M.C., Nemesh, J., Lane, C.R., Schaffner, S.F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T.J., Daly, M., Groop, L. & Lander, E.S., 2000. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature genetics*, 26(1), pp.76–80.

American Cancer Society, 2014. Cancer Facts & Figures. *Cancer Facts and Figures*.

Amundadottir, L.T., Sulem, P., Gudmundsson, J., Helgason, A., Baker, A., Agnarsson, B.A., Sigurdsson, A., Benediktsdottir, K.R., Cazier, J.-B., Sainz, J., Jakobsdottir, M., Kostic, J., Magnusdottir, D.N., Ghosh, S., Agnarsson, K., Birgisdottir, B., Le Roux, L., Olafsdottir, A., Blondal, T., Andresdottir, M., Gretarsdottir, O.S., Bergthorsson, J.T., Gudbjartsson, D., Gylfason, A., Thorleifsson, G., Manolescu, A., Kristjansson, K., Geirsson, G., Isaksson, H., Douglas, J., Johansson, J.-E., Bälter, K., Wiklund, F., Montie, J.E., Yu, X., Suarez, B.K., Ober, C., Cooney, K.A., Gronberg, H., Catalona, W.J., Einarsson, G. V, Barkardottir, R.B., Gulcher, J.R., Kong, A., Thorsteinsdottir, U. & Stefansson, K., 2006. A common variant associated with prostate cancer in European and African populations. *Nature genetics*, 38(6), pp.652–8.

Avery, O.T., Macleod, C.M. & McCarty, M., 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of experimental medicine*, 79(2), pp.137–58.

Barosi, G., 1999. Myelofibrosis with myeloid metaplasia: diagnostic definition and prognostic classification for clinical studies and treatment guidelines. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 17(9), pp.2954–70.

Barranco, S.C., Townsend, C.M., Casartelli, C., Macik, B.G., Burger, N.L., Boerwinkle, W.R. & Gourley, W.K., 1983. Establishment and characterization of an in vitro model system for human adenocarcinoma of the stomach. *Cancer research*, 43(4), pp.1703–9.

Bello, D., Webber, M.M., Kleinman, H.K., Wartinger, D.D. & Rhim, J.S., 1997. Androgen responsive adult human prostatic epithelial cell lines immortalized by human papillomavirus 18. *Carcinogenesis*, 18(6), pp.1215–23.

Bender, W., Akam, M., Karch, F., Beachy, P.A., Peifer, M., Spierer, P., Lewis, E.B. & Hogness, D.S., 1983. Molecular Genetics of the Bithorax Complex in Drosophila melanogaster. *Science (New York, N.Y.)*, 221(4605), pp.23–9.

Berger, R., 2004. Androgen-Induced Differentiation and Tumorigenicity of Human Prostate Epithelial Cells. *Cancer Research*, 64(24), pp.8867–8875.

Di Bernardo, M.C., Crowther-Swanepoel, D., Broderick, P., Webb, E., Sellick, G., Wild, R., Sullivan, K., Vijayakrishnan, J., Wang, Y., Pittman, A.M., Sunter, N.J., Hall, A.G., Dyer, M.J.S., Matutes, E., Dearden, C., Mainou-Fowler, T., Jackson, G.H., Summerfield, G., Harris, R.J., Pettitt, A.R., Hillmen, P., Allsup, D.J., Bailey, J.R., Pratt, G., Pepper, C., Fegan, C., Allan, J.M., Catovsky, D. & Houlston, R.S., 2008. A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nature genetics*, 40(10), pp.1204–1210.

Berndt, S.I., Skibola, C.F., Joseph, V., Camp, N.J., Nieters, A., Slager, S.L., et al., 2013. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nature genetics*, 45(8), pp.868–76.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S. & Thomson, J.A., 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, 28(10), pp.1045–8.

Bjartell, A.S., Al-Ahmadie, H., Serio, A.M., Eastham, J.A., Eggener, S.E., Fine, S.W., Udby, L., Gerald, W.L., Vickers, A.J., Lilja, H., Reuter, V.E. & Scardino, P.T., 2007. Association of Cysteine-Rich Secretory Protein 3 and -Microseminoprotein with Outcome after Radical Prostatectomy. *Clinical Cancer Research*, 13(14), pp.4130–4138.

Bojesen, S.E., Pooley, K.A., Johnatty, S.E., Beesley, J., Michailidou, K., Dunning, A.M., et al., 2013. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature genetics*, 45(4), pp.371–84, 384e1–2.

Botstein, D., White, R.L., Skolnick, M. & Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms.

American journal of human genetics, 32(3), pp.314–31.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. & Crawford, G.E., 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2), pp.311–22.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., Cherry, J.M. & Snyder, M., 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, 22(9), pp.1790–1797.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., Cherry, J.M. & Snyder, M., 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, 22(9), pp.1790–7.

Bush, W.S. & Moore, J.H., 2012. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12), p.e1002822.

Campbell, P.J. & Green, A.R., 2006. The Myeloproliferative Disorders. *New England Journal of Medicine*, 355(23), pp.2452–2466.

Cao, W., Lee, S.H. & Lu, J., 2005. CD83 is preformed inside monocytes, macrophages and dendritic cells, but it is only stably expressed on activated dendritic cells. *The Biochemical journal*, 385(Pt 1), pp.85–93.

Carbuccia, N., Murati, A., Trouplin, V., Brecqueville, M., Adélaïde, J., Rey, J., Vainchenker, W., Bernard, O.A., Chaffanet, M., Vey, N., Birnbaum, D. & Mozziconacci, M.J., 2009. Mutations of ASXL1 gene in myeloproliferative neoplasms. *Leukemia*, 23(11), pp.2183–2186.

Chang, B.-L., Cramer, S.D., Wiklund, F., Isaacs, S.D., Stevens, V.L., Sun, J., Smith, S., Pruett, K., Romero, L.M., Wiley, K.E., Kim, S.-T., Zhu, Y., Zhang, Z., Hsu, F.-C., Turner, A.R., Adolfsson, J., Liu, W., Kim, J.W., Duggan, D., Carpten, J., Zheng, S.L., Rodriguez, C., Isaacs, W.B., Grönberg, H. & Xu, J., 2009. Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. *Human molecular genetics*, 18(7), pp.1368–75.

Chen, Y.-C., Page, J.H., Chen, R. & Giovannucci, E., 2008. Family history of prostate and breast cancer and the risk of prostate cancer in the PSA era. *The Prostate*, 68(14), pp.1582–91.

Clarke, L. & Carbon, J., 1980. Isolation of the centromere-linked CDC10 gene by complementation in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 77(4), pp.2173–2177.

Coetzee, S.G., Rhie, S.K., Berman, B.P., Coetzee, G.A. & Noushmehr, H., 2012. FunciSNP: An R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Research*, 40(18).

Collins, F.S., Morgan, M. & Patrinos, A., 2003. The Human Genome Project: lessons from large-scale biology. *Science (New York, N.Y.)*, 300(5617), pp.286–90.

Conde, L., Halperin, E., Akers, N.K., Brown, K.M., Smedby, K.E., Skibola, C.F., et al., 2010. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nature genetics*, 42(8), pp.661–664.

Consortium, T.I.H., 2003. The International HapMap Project. , 426(6968), pp.789–796.

Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L. & Pericak-Vance, M.A., 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer’s disease in late onset families. *Science (New York, N.Y.)*, 261(5123), pp.921–3.

Couch, F.J., Wang, X., McGuffog, L., Lee, A., Olswold, C., Antoniou, A.C., et al., 2013. Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk K. W. Hunter, ed. *PLoS Genetics*, 9(3), p.e1003212.

Cozen, W., Li, D., Best, T., Van Den Berg, D.J., Gourraud, P.A., Cortessis, V.K., Skol, A.D., Mack, T.M., Glaser, S.L., Weiss, L.M., Nathwani, B.N., Bhatia, S., Schumacher, F.R., Edlund, C.K., Hwang, A.E., Slager, S.L., Fredericksen, Z.S., Strong, L.C., Habermann, T.M., Link, B.K., Cerhan, J.R., Robison, L.L., Conti, D. V. & Onel, K., 2012. A genome-wide meta-analysis of nodular sclerosing Hodgkin lymphoma identifies risk loci at 6p21.32. *Blood*, 119(2), pp.469–475.

Crowther-Swanepoel, D., Broderick, P., Di Bernardo, M.C., Dobbins, S.E., Torres, M., Mansouri, M., Ruiz-Ponte, C., Enjuanes, A., Rosenquist, R., Carracedo, A., Jurlander, J., Campo, E., Juliusson, G., Montserrat, E., Smedby, K.E., Dyer, M.J.S., Matutes, E., Dearden, C., Sunter, N.J., Hall, A.G., Mainou-Fowler, T., Jackson, G.H., Summerfield, G., Harris, R.J., Pettitt, A.R., Allsup, D.J., Bailey, J.R., Pratt, G., Pepper, C., Fegan, C., Parker, A., Oscier, D., Allan, J.M., Catovsky, D. & Houlston, R.S., 2010. Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nature* *genetics*, 42(2), pp.132–136.

Dameshek, W., 1951. Some speculations on the myeloproliferative syndromes. *Blood*, 6(4), pp.372–5.

Delaneau, O., Marchini, J. & Zagury, J.-F., 2012. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2), pp.179–81.

Delhommeau, F., Dupont, S., Tonetti, C., Massé, A., Godin, I., Le Couedic, J.-P., Debili, N., Saulnier, P., Casadevall, N., Vainchenker, W. & Giraudier, S., 2007. Evidence that the JAK2 G1849T (V617F) mutation occurs in a lymphomyeloid progenitor in polycythemia vera and idiopathic myelofibrosis. *Blood*, 109(1), pp.71–7.

Devlin, B. & Risch, N., 1995. A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics*, 29(2), pp.311–322.

Dimri, G.P., Lee, X., Basile, G., Acosta, M., Scott, G., Roskelley, C., Medrano, E.E., Linskens, M., Rubelj, I. & Pereira-Smith, O., 1995. A biomarker that identifies senescent human cells in culture and in aging skin in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 92(20), pp.9363–7.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Gingeras, T.R., et al., 2012. Landscape of transcription in human cells. *Nature*, 489(7414), pp.101–8.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., Green, R.D. & Dekker, J., 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10), pp.1299–309.

Eeles, R.A., Kote-Jarai, Z., Giles, G.G., Olama, A.A. Al, Guy, M., Jugurnauth, S.K., Mulholland, S., Leongamornlert, D.A., Edwards, S.M., Morrison, J., Field, H.I., Southey, M.C., Severi, G., Donovan, J.L., Hamdy, F.C., Dearnaley, D.P., Muir, K.R., Smith, C., Bagnato, M., Ardern-Jones, A.T., Hall, A.L., O'Brien, L.T., Gehr-Swain, B.N., Wilkinson, R.A., Cox, A., Lewis, S., Brown, P.M., Jhavar, S.G., Tymrakiewicz, M., Lophatananon, A., Bryant, S.L., Horwich, A., Huddart, R.A., Khoo, V.S., Parker, C.C., Woodhouse, C.J., Thompson, A., Christmas, T., Ogden, C., Fisher, C., Jamieson, C., Cooper, C.S., English, D.R., Hopper, J.L., Neal, D.E. & Easton, D.F., 2008. Multiple newly identified loci associated with prostate cancer susceptibility. *Nature genetics*, 40(3), pp.316–21.

Enciso-Mora, V., Broderick, P., Ma, Y., Jarrett, R.F., Hjalgrim, H., Hemminki, K., van den Berg, A., Olver, B., Lloyd, A., Dobbins, S.E., Lightfoot, T., van Leeuwen, F.E., Försti, A., Diepstra, A., Broeks, A., Vijayakrishnan, J., Shield, L., Lake, A., Montgomery, D., Roman, E., Engert, A., von Strandmann, E.P., Reiners, K.S., Nolte, I.M., Smedby, K.E., Adami, H.-O., Russell, N.S., Glimelius, B., Hamilton-Dutoit, S., de Bruin, M., Ryder, L.P., Molin, D., Sorensen, K.M., Chang, E.T., Taylor, M., Cooke, R., Hofstra, R., Westers, H., van Wezel, T., van Eijk, R., Ashworth, A., Rostgaard, K., Melbye, M., Swerdlow, A.J. & Houlston, R.S., 2010. A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). *Nature genetics*, 42(12), pp.1126–1130.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., de Jong, P.J., et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799–816.

Epstein, M.M., Edgren, G., Rider, J.R., Mucci, L.A. & Adami, H.-O., 2012. Temporal trends in cause of death among Swedish and US men with prostate cancer. *Journal of the National Cancer Institute*, 104(17), pp.1335–42.

Ernst, J. & Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, 9(3), pp.215–6.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. & Bernstein, B.E., 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345), pp.43–49.

Ernst, T., Chase, A.J., Score, J., Hidalgo-Curtis, C.E., Bryant, C., Jones, A. V, Waghorn, K., Zoi, K., Ross, F.M., Reiter, A., Hochhaus, A., Drexler, H.G., Duncombe, A., Cervantes, F., Oscier, D., Boultwood, J., Grand, F.H. & Cross, N.C.P., 2010. Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nature genetics*, 42(8), pp.722–6.

Ewing, C.M., Ray, A.M., Lange, E.M., Zuhlke, K.A., Robbins, C.M., Tembe, W.D., Wiley, K.E., Isaacs, S.D., Johng, D., Wang, Y., Bizon, C., Yan, G., Gielzak, M., Partin, A.W., Shanmugam, V., Izatt, T., Sinari, S., Craig, D.W., Zheng, S.L., Walsh, P.C., Montie, J.E., Xu, J., Carpten, J.D., Isaacs, W.B. & Cooney, K.A., 2012. Germline mutations in HOXB13 and prostate-cancer risk. *The New England journal of medicine*, 366(2), pp.141–9.

Facompre, N.D., El-Bayoumy, K., Sun, Y.-W., Pinto, J.T. & Sinha, R., 2010. 1,4-phenylenebis(methylene)selenocyanate, but not selenomethionine, inhibits androgen receptor and Akt signaling in human prostate cancer cells. *Cancer prevention research (Philadelphia, Pa.)*, 3(8), pp.975–84.

Farrer, L.A., Cupples, L.A., Haines, J.L., Hyman, B., Kukull, W.A., Mayeux, R., Myers, R.H., Pericak-Vance, M.A., Risch, N. & van Duijn, C.M., Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA*, 278(16), pp.1349–56.

Freedman, M.L., Haiman, C.A., Patterson, N., McDonald, G.J., Tandon, A., Waliszewska, A., Penney, K., Steen, R.G., Ardlie, K., John, E.M., Oakley-Girvan, I., Whittemore, A.S., Cooney, K.A., Ingles, S.A., Altshuler, D., Henderson, B.E. & Reich, D., 2006. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proceedings of the National Academy of Sciences of the United States of America*, 103(38), pp.14068–73.

Freedman, M.L., Monteiro, A.N.A., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., James, M., Liu, P., Tichelaar, J.W., Vikis, H.G., You, M. & Mills, I.G., 2011. Principles for the post-GWAS functional characterization of cancer risk loci. *Nature genetics*, 43(6), pp.513–8.

French, J.D., Ghoussaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Dunning, A.M., et al., 2013. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *American journal of human genetics*, 92(4), pp.489–503.

Frohman, M.A., Dush, M.K. & Martin, G.R., 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences of the United States of America*, 85(23), pp.8998–9002.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G.Y., Huang, P.Y.H., Welboren, W.-J., Han, Y., Ooi, H.S., Ariyaratne, P.N., Vega, V.B., Luo, Y., Tan, P.Y., Choy, P.Y., Wansa, K.D.S.A., Zhao, B., Lim, K.S., Leow, S.C., Yow, J.S., Joseph, R., Li, H., Desai, K. V, Thomsen, J.S., Lee, Y.K., Karuturi, R.K.M., Herve, T., Bourque, G., Stunnenberg, H.G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E.T., Wei, C.-L., Cheung, E. & Ruan, Y., 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269), pp.58–64.

Garcia-Closas, M., Couch, F.J., Lindstrom, S., Michailidou, K., Schmidt, M.K., Kraft, P., et al., 2013. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nature genetics*, 45(4), pp.392–8, 398e1–2.

Garde, S. V, Basrur, V.S., Li, L., Finkelman, M.A., Krishan, A., Wellham, L., Ben-Josef, E., Haddad, M., Taylor, J.D., Porter, A.T. & Tang, D.G., 1999. Prostate secretory protein (PSP94) suppresses the growth of androgen-independent prostate cancer cell line (PC3) and xenografts by inducing apoptosis. *The Prostate*, 38(2), pp.118–25.

Gaudet, M.M., Kuchenbaecker, K.B., Vijai, J., Klein, R.J., Kirchhoff, T., Offit, K., et al., 2013. Identification of a BRCA2-specific modifier locus at 6p24 related to breast cancer risk. *PLoS genetics*, 9(3), p.e1003173.

Ginsburg, G.S., Willard, H.F., John, S., Wang, H. & Stamatoyannopoulos, J.A., 2013. *Genomic and Personalized Medicine*, Elsevier.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D., 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17(6), pp.877–885.

Goldin, L.R., Pfeiffer, R.M., Gridley, G., Gail, M.H., Li, X., Mellemkjaer, L., Olsen, J.H., Hemminki, K. & Linet, M.S., 2004. Familial Aggregation of Hodgkin Lymphoma and Related Tumors. *Cancer*, 100(9), pp.1902–1908.

Griffith, F., 1928. The Significance of Pneumococcal Types. *The Journal of hygiene*, 27(2), pp.113–59.

Gross, D.S. & Garrard, W.T., 1988. Nuclease hypersensitive sites in chromatin. *Annual review of biochemistry*, 57, pp.159–197.

Grubb, R.L., Calvert, V.S., Wulkuhle, J.D., Paweletz, C.P., Linehan, W.M., Phillips, J.L., Chuaqui, R., Valasco, A., Gillespie, J., Emmert-Buck, M., Liotta, L.A. & Petricoin, E.F., 2003. Signal pathway profiling of prostate cancer using reverse phase protein arrays. *Proteomics*, 3(11), pp.2142–6.

Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L.T., Gudbjartsson, D., Stefansson, K., et al., 2007. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature genetics*, 39(5), pp.631–7.

Guigó, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T.R., Harrow, J., Hubbard, T., Lewis, S.E. & Reese, M.G., 2006.

EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome biology*, 7 Suppl 1, pp.S2.1–31.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R. & Sakaguchi, A.Y., 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940), pp.234–8.

Ha, S., Ruoff, R., Kahoud, N., Franke, T.F. & Logan, S.K., 2011. Androgen receptor levels are upregulated by Akt in prostate cancer. *Endocrine-related cancer*, 18(2), pp.245–55.

Haiman, C.A., Chen, G.K., Blot, W.J., Strom, S.S., Berndt, S.I., Henderson, B.E., et al., 2011. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nature genetics*, 43(6), pp.570–3.

Haiman, C.A., Chen, G.K., Vachon, C.M., Canzian, F., Dunning, A., Couch, F.J., et al., 2011. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nature genetics*, 43(12), pp.1210–4.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B. & King, M.C., 1990. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science (New York, N.Y.)*, 250(4988), pp.1684–9.

Harris, G.J. & Lager, D.J., 1991. *Primary renal lymphoma.*,

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J.G.R., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S.E. & Guigo, R., 2006. GENCODE: producing a reference annotation for ENCODE. *Genome biology*, 7 Suppl 1, pp.S4.1–9.

Hartlapp, I., Pallasch, C., Weibert, G., Kemkers, A., Hummel, M. & Re, D., 2009. Depsipeptide induces cell death in Hodgkin lymphoma-derived cell lines. *Leukemia research*, 33(7), pp.929–36.

Harutyunyan, A., Klampfl, T., Cazzola, M. & Kralovics, R., 2011. p53 Lesions in Leukemic Transformation. *New England Journal of Medicine*, 364(5), pp.488–490.

Hayes, J.E., Trynka, G., Vijai, J., Offit, K., Raychaudhuri, S. & Klein, R.J., 2015. Tissue-Specific Enrichment of Lymphoma Risk Loci in Regulatory Elements L. Prokunina-Olsson, ed. *PLOS ONE*, 10(9), p.e0139360.

Hayward, S.W., Dahiya, R., Cunha, G.R., Bartek, J., Deshpande, N. & Narayan, P., 1995. Establishment and characterization of an immortalized but non-transformed human prostate epithelial cell line: BPH-1. *In vitro cellular & developmental biology. Animal*, 31(1), pp.14–24.

Hershey, A.D. & Chase, M., 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of general physiology*, 36(1), pp.39–56.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., Fields, S. & Stamatoyannopoulos, J.A., 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods*, 6(4), pp.283–9.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. & Manolio, T.A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), pp.9362–7.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A. & Noble, W.S., 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5), pp.473–476.

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., Hardison, R.C., Dunham, I., Kellis, M. & Noble, W.S., 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, 41(2), pp.827–41.

Horoszewicz, J.S., Leong, S.S., Kawinski, E., Karr, J.P., Rosenthal, H., Chu, T.M., Mirand, E.A. & Murphy, G.P., 1983. LNCaP model of human prostatic carcinoma. *Cancer research*, 43(4), pp.1809–18.

Howie, B.N., Donnelly, P. & Marchini, J., 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6), p.e1000529.

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Miller, D., Altekruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Chen, H., Feuer, E. & Cronin, K. (eds)., 2015. *SEER Cancer Statistics Review, 1975-2012*, Bethesda, MD.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G. V, Chin, L. & Garraway, L.A., 2013.

Highly recurrent TERT promoter mutations in human melanoma. *Science (New York, N.Y.)*, 339(6122), pp.957–9.

Ihle, J.N. & Gilliland, D.G., 2007. Jak2: normal function and role in hematopoietic disorders. *Current opinion in genetics & development*, 17(1), pp.8–14.

Ito, Y., Bae, S.-C. & Chuang, L.S.H., 2015. The RUNX family: developmental regulators in cancer. *Nature reviews. Cancer*, 15(2), pp.81–95.

James, C., Ugo, V., Le Couédic, J.-P., Staerk, J., Delhommeau, F., Lacout, C., Garçon, L., Raslova, H., Berger, R., Bennaceur-Griscelli, A., Villeval, J.L., Constantinescu, S.N., Casadevall, N. & Vainchenker, W., 2005. A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature*, 434(7037), pp.1144–8.

Jamieson, C.H.M., Gotlib, J., Durocher, J.A., Chao, M.P., Mariappan, M.R., Lay, M., Jones, C., Zehnder, J.L., Lilleberg, S.L. & Weissman, I.L., 2006. The JAK2 V617F mutation occurs in hematopoietic stem cells in polycythemia vera and predisposes toward erythroid differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 103(16), pp.6224–9.

Jiao, J., Wang, S., Qiao, R., Vivanco, I., Watson, P.A., Sawyers, C.L. & Wu, H., 2007. Murine Cell Lines Derived from Pten Null Prostate Cancer Show the Critical Role of PTEN in Hormone Refractory Prostate Cancer Development. *Cancer Research*, 67(13), pp.6083–6091.

Jones, A. V, Chase, A., Silver, R.T., Oscier, D., Zoi, K., Wang, Y.L., Cario, H., Pahl, H.L., Collins, A., Reiter, A., Grand, F. & Cross, N.C.P., 2009. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nature genetics*, 41(4), pp.446–9.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Cho, J.H., et al., 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422), pp.119–24.

Kaighn, M.E., Narayan, K.S., Ohnuki, Y., Lechner, J.F. & Jones, L.W., 1979. Establishment and characterization of a human prostatic carcinoma cell line (PC-3). *Investigative urology*, 17(1), pp.16–23.

Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y. & Kamatani, N., 2010. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature genetics*, 42(3), pp.210–5.

Kilpivaara, O., Mukherjee, S., Schram, A.M., Wadleigh, M., Mullally, A., Ebert, B.L., Bass, A., Marubayashi, S., Heguy, A., Garcia-Manero, G., Kantarjian, H., Offit, K., Stone, R.M., Gilliland, D.G., Klein, R.J. & Levine, R.L., 2009. A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. *Nature genetics*, 41(4), pp.455–9.

Klein, R.J., 2007. Power analysis for genome-wide association studies. *BMC genetics*, 8, p.58.

Klein, R.J., Xu, X., Mukherjee, S., Willis, J. & Hayes, J., 2010. Successes of genome-wide association studies. *Cell*, 142(3), pp.350–1; author reply 353–5.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C. & Hoh, J., 2005. Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720), pp.385–9.

Korenchuk, S., Lehr, J.E., MClean, L., Lee, Y.G., Whitney, S., Vessella, R., Lin, D.L. & Pienta, K.J., VCaP, a cell-based model system of human prostate cancer. *In vivo (Athens, Greece)*, 15(2), pp.163–8.

Kumar, V., Matsuo, K., Takahashi, A., Hosono, N., Tsunoda, T., Kamatani, N., Kong, S.-Y., Nakagawa, H., Cui, R., Tanikawa, C., Seto, M., Morishima, Y., Kubo, M., Nakamura, Y. & Matsuda, K., 2011. Common variants on 14q32 and 13q12 are associated with DLBCL susceptibility. *Journal of human genetics*, 56(6), pp.436–439.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Szustakowki, J., et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Landgren, O., Goldin, L.R., Kristinsson, S.Y., Helgadottir, E.A., Samuelsson, J. & Björkholm, M., 2008. Increased risks of polycythemia vera, essential thrombocythemia, and myelofibrosis among 24,577 first-degree relatives of 11,039 patients with myeloproliferative neoplasms in Sweden. *Blood*, 112(6), pp.2199–204.

Lange, E.M., Gillanders, E.M., Davis, C.C., Brown, W.M., Campbell, J.K., Jones, M., Gildea, D., Riedesel, E., Albertus, J., Freas-Lutz, D., Markey, C., Giri, V., Dimmer, J.B., Montie, J.E., Trent, J.M. & Cooney, K.A., 2003. Genome-wide scan for prostate cancer susceptibility genes using families from the University of Michigan prostate cancer genetics project finds evidence for linkage on chromosome 17 near BRCA1. *The Prostate*, 57(4), pp.326–34.

Lange, E.M., Robbins, C.M., Gillanders, E.M., Zheng, S.L., Xu, J., Wang, Y., White, K.A., Chang, B.-L., Ho, L.A., Trent, J.M., Carpten, J.D., Isaacs, W.B. & Cooney, K.A., 2007. Fine-mapping the putative chromosome 17q21-22 prostate cancer susceptibility gene to a 10 cM region based on linkage analysis. *Human genetics*, 121(1), pp.49–55.

Lange, E.M., Salinas, C.A., Zuhlke, K.A., Ray, A.M., Wang, Y., Lu, Y., Ho, L.A., Luo, J. & Cooney, K.A., 2012. Early onset prostate cancer has a significant genetic component. *The Prostate*, 72(2), pp.147–56.

Lee, B.Y., Han, J.A., Im, J.S., Morrone, A., Johung, K., Goodwin, E.C., Kleijer, W.J., DiMaio, D. & Hwang, E.S., 2006. Senescence-associated beta-galactosidase is lysosomal beta-galactosidase. *Aging cell*, 5(2), pp.187–95.

Levenstien, M.A. & Klein, R.J., 2011. Predicting functionally important SNP classes based on negative selection. *BMC bioinformatics*, 12, p.26.

Levine, R.L., Belisle, C., Wadleigh, M., Zahrieh, D., Lee, S., Chagnon, P., Gilliland, D.G. & Busque, L., 2006. X-inactivation-based clonality analysis and quantitative JAK2V617F assessment reveal a strong association between clonality and JAK2V617F in PV but not ET/MMM, and identifies a subset of JAK2V617F-negative ET and MMM patients with clonal hematopoiesis. *Blood*, 107(10), pp.4139–41.

Li, H., 2011. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, 27(5), pp.718–719.

Li, M.J., Wang, L.Y., Xia, Z., Sham, P.C. & Wang, J., 2013. GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic acids research*, 41(Web Server issue).

Lichtenstein, P., Holm, N. V, Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A. & Hemminki, K., 2000. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine*, 343(2), pp.78–85.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S. & Dekker, J., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950), pp.289–93.

Liu, Y., Cao, L., Li, Z., Zhou, D., Liu, W., Shen, Q., Wu, Y., Zhang, D., Hu, X., Wang, T., Ye, J., Weng, X., Zhang, H., Zhang, D., Zhang, Z., Liu, F., He, L. & Shi, Y., 2014. A genome-wide association study identifies a locus on TERT for mean telomere length in Han Chinese. *PloS one*, 9(1), p.e85043.

Lou, H., Yeager, M., Li, H., Bosquet, J.G., Hayes, R.B., Orr, N., Yu, K., Hutchinson, A., Jacobs, K.B., Kraft, P., Wacholder, S., Chatterjee, N., Feigelson, H.S., Thun, M.J., Diver, W.R., Albanes, D., Virtamo, J., Weinstein, S., Ma, J., Gaziano, J.M., Stampfer, M., Schumacher, F.R., Giovannucci, E., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole, G.L., Crawford, E.D., Anderson, S.K., Tucker, M., Hoover, R.N., Fraumeni, J.F., Thomas, G., Hunter, D.J., Dean, M. & Chanock, S.J., 2009. Fine mapping and functional analysis of a common variant in MSMB on chromosome 10q11.2 associated with prostate cancer susceptibility. *Proceedings of the National Academy of Sciences of the United States of America*, 106(19), pp.7933–8.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A. & Visscher, P.M., 2009. Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp.747–53.

Marchini, J. & Howie, B., 2010. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics*, 11(7), pp.499–511.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R. & Stamatoyannopoulos, J.A., 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099), pp.1190–1195.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Easton, D.F., et al., 2013. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics*, 45(4), pp.353–61, 361e1–2.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M. & Ding, W., 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science (New York, N.Y.)*, 266(5182), pp.66–71.

Mukherjee, S., 2011. *Integrative genome-wide analysis to study the germline genetics of myeloproliferative neoplasms*. Memorial Sloan Kettering Cancer Center.

Mukherjee, S., Simon, J., Bayuga, S., Ludwig, E., Yoo, S., Orlow, I., Viale, A., Offit, K., Kurtz, R.C., Olson, S.H. & Klein, R.J., 2011. Including Additional Controls from Public Databases Improves the Power of a Genome-Wide Association Study. *Human Heredity*, 72(1), pp.21–34.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., Maurano, M.T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R.S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M.J., Akey, J.M., Bender, M.A., Groudine, M., Kaul, R. & Stamatoyannopoulos, J.A., 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414), pp.83–90.

Nesterov, A., Lu, X., Johnson, M., Miller, G.J., Ivashchenko, Y. & Kraft, A.S., 2001. Elevated AKT activity protects the prostate cancer cell line LNCaP from TRAIL-induced apoptosis. *The Journal of biological chemistry*, 276(14), pp.10767–74.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E. & Cox, N.J., 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4), p.e1000888.

Olcaydu, D., Harutyunyan, A., Jäger, R., Berg, T., Gisslinger, B., Pabinger, I., Gisslinger, H. & Kralovics, R., 2009. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. *Nature genetics*, 41(4), pp.450–4.

Park, E., Williams, B., Wold, B.J. & Mortazavi, A., 2012. RNA editing in the human ENCODE RNA-seq data. *Genome research*, 22(9), pp.1626–33.

Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., Wells, V.A., Grunn, A., Messina, M., Elliot, O., Chan, J., Bhagat, G., Chadburn, A., Gaidano, G., Mullighan, C.G., Rabadan, R. & Dalla-Favera, R., 2011. Analysis of the coding genome of diffuse large B-cell lymphoma. *Nature genetics*, 43(9), pp.830–837.

Pileri, S.A., Ascani, S., Leoncini, L., Sabattini, E., Zinzani, P.L., Piccaluga, P.P., Pileri, A., Giunti, M., Falini, B., Bolis, G.B. & Stein, H., 2002. Hodgkin’s lymphoma: the pathologist's viewpoint. *Journal of clinical pathology*, 55(3), pp.162–76.

Polgar, N., Csongei, V., Szabo, M., Zambo, V., Melegh, B.I., Sumegi, K., Nagy, G., Tulassay, Z. & Melegh, B., 2012. Investigation of JAK2, STAT3 and CCR6 polymorphisms and their gene-gene interactions in inflammatory bowel disease. *International journal of immunogenetics*, 39(3), pp.247–52.

Pomerantz, M.M., Beckwith, C.A., Regan, M.M., Wyman, S.K., Petrovics, G., Chen, Y., Hawksworth, D.J., Schumacher, F.R., Mucci, L., Penney, K.L., Stampfer, M.J., Chan, J.A., Ardlie, K.G., Fritz, B.R., Parkin, R.K., Lin, D.W., Dyke, M., Herman, P., Lee, S., Oh, W.K., Kantoff, P.W., Tewari, M., McLeod, D.G., Srivastava, S. & Freedman, M.L., 2009. Evaluation of the 8q24 prostate cancer risk locus and MYC expression. *Cancer research*, 69(13), pp.5568–74.

Pomerantz, M.M., Shrestha, Y., Flavin, R.J., Regan, M.M., Penney, K.L., Mucci, L.A., Stampfer, M.J., Hunter, D.J., Chanock, S.J., Schafer, E.J., Chan, J.A., Tabernero, J., Baselga, J., Richardson, A.L., Loda, M., Oh, W.K., Kantoff, P.W., Hahn, W.C. & Freedman, M.L., 2010. Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. *PLoS genetics*, 6(11), p.e1001204.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8), pp.904–9.

Pritchard, J.K. & Cox, N.J., 2002. The allelic architecture of human disease genes: common disease-common variant...or not? *Human molecular genetics*, 11(20), pp.2417–23.

Pritchard, J.K. & Przeworski, M., 2001. Linkage disequilibrium in humans: models and data. *American journal of human genetics*, 69(1), pp.1–14.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. & Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3), pp.559–75.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841–842.

Rafnar, T., Sulem, P., Stacey, S.N., Geller, F., Gudmundsson, J., Stefansson, K., et al.,

2009. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nature genetics*, 41(2), pp.221–7.

Reich, D.E. & Lander, E.S., 2001. On the allelic spectrum of human disease. *Trends in genetics : TIG*, 17(9), pp.502–10.

Risch, N., 2001. The Genetic Epidemiology of Cancer: Interpreting Family and Twin Studies and Their Implications for Molecular Genetic Approaches. *Cancer Epidemiol. Biomarkers Prev.*, 10(7), pp.733–741.

Risch, N. & Merikangas, K., 1996. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)*, 273(5281), pp.1516–7.

Ruggeri, M., Tosetto, A., Frezzato, M. & Rodeghiero, F., 2003. The rate of progression to polycythemia vera or essential thrombocythemia in patients with erythrocytosis or thrombocytosis. *Annals of internal medicine*, 139(6), pp.470–5.

Rumi, E., Passamonti, F., Della Porta, M.G., Elena, C., Arcaini, L., Vanelli, L., Del Curto, C., Pietra, D., Boveri, E., Pascutto, C., Cazzola, M. & Lazzarino, M., 2007. Familial chronic myeloproliferative disorders: clinical phenotype and evidence of disease anticipation. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 25(35), pp.5630–5.

Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J., 2012. The long-range interaction landscape of gene promoters. *Nature*, 489(7414), pp.109–13.

Schneider, C.A., Rasband, W.S. & Eliceiri, K.W., 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature methods*, 9(7), pp.671–5.

Scott, D.W., Mungall, K.L., Ben-Neriah, S., Rogic, S., Morin, R.D., Slack, G.W., Tan, K.L., Chan, F.C., Lim, R.S., Connors, J.M., Marra, M.A., Mungall, A.J., Steidl, C. & Gascoyne, R.D., 2012. TBL1XR1/TP63: A novel recurrent gene fusion in B-cell non-Hodgkin lymphoma. *Blood*, 119(21), pp.4949–4952.

Seidah, N.G., Arbatti, N.J., Rochemont, J., Sheth, A.R. & Chrétien, M., 1984. Complete amino acid sequence of human seminal plasma beta-inhibin. Prediction of post Gln-Arg cleavage as a maturation site. *FEBS letters*, 175(2), pp.349–55.

Sellick, G.S., Catovsky, D. & Houlston, R.S., 2006. Familial Chronic Lymphocytic Leukemia. *Seminars in Oncology*, 33(2), pp.195–201.

Sellin, J.H., Umar, S., Xiao, J. & Morris, A.P., 2001. Increased beta-catenin expression and nuclear translocation accompany cellular hyperproliferation in vivo. *Cancer research*, 61(7), pp.2899–906.

Shah, S., Schrader, K. a, Waanders, E., Timms, A.E., Vijai, J., Offit, K., et al., 2013. A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. *Nature genetics*, 45(10), pp.1226–31.

Shen, X., Kim, W., Fujiwara, Y., Simon, M.D., Liu, Y., Mysliwiec, M.R., Yuan, G.-C., Lee, Y. & Orkin, S.H., 2009. Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. *Cell*, 139(7), pp.1303–14.

Sheth, A.R., Vanage, G.R., Hurkadli, K.S. & Sheth, N.A., 1984. Role of the prostate in the regulation of pituitary secretion of follicle stimulating hormone. *Medical hypotheses*, 15(2), pp.141–8.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. & Hayashizaki, Y., 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26), pp.15776–81.

Shukeir, N., 2004. A Synthetic 15-mer Peptide (PCK3145) Derived from Prostate Secretory Protein Can Reduce Tumor Growth, Experimental Skeletal Metastases, and Malignancy-Associated Hypercalcemia. *Cancer Research*, 64(15), pp.5370–5377.

Shukeir, N., Arakelian, A., Kadhim, S., Garde, S. & Rabbani, S.A., 2003. Prostate secretory protein PSP-94 decreases tumor growth and hypercalcemia of malignancy in a syngenic in vivo model of prostate cancer. *Cancer research*, 63(9), pp.2072–8.

Siegel, R.L., Miller, K.D. & Jemal, A., 2015. Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1), pp.5–29.

Skibola, C.F., Bracci, P.M., Halperin, E., Conde, L., Craig, D.W., Agana, L., Iyadurai, K., Becker, N., Brooks-Wilson, A., Curry, J.D., Spinelli, J.J., Holly, E.A., Riby, J., Zhang, L., Nieters, A., Smith, M.T. & Brown, K.M., 2009. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nature genetics*, 41(8), pp.873–875.

Slager, S.L., Rabe, K.G., Achenbach, S.J., Vachon, C.M., Goldin, L.R., Strom, S.S., Lanasa, M.C., Spector, L.G., Rassenti, L.Z., Leis, J.F., Camp, N.J., Glenn, M., Kay, N.E., Cunningham, J.M., Hanson, C.A., Marti, G.E., Weinberg, J.B., Morrison, V.A., Link, B.K., Call, T.G., Caporaso, N.E. & Cerhan, J.R., 2011. Genome-wide association study identifies a novel susceptibility locus at 6p21.3

among familial CLL. *Blood*, 117(6), pp.1911–1916.

Smedby, K.E., Foo, J.N., Skibola, C.F., Darabi, H., Conde, L., Liu, J., et al., 2011. GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. *PLoS Genetics*, 7(4).

Spencer, C.C.A., Su, Z., Donnelly, P. & Marchini, J., 2009. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip J. D. Storey, ed. *PLoS Genetics*, 5(5), p.e1000477.

Spender, L.C., Cornish, G.H., Sullivan, A. & Farrell, P.J., 2002. Expression of transcription factor AML-2 (RUNX3, CBF(alpha)-3) is induced by Epstein-Barr virus EBNA-2 and correlates with the B-cell activation phenotype. *Journal of virology*, 76(10), pp.4919–4927.

Stone, K.R., Mickey, D.D., Wunderli, H., Mickey, G.H. & Paulson, D.F., 1978. Isolation of a human prostate carcinoma cell line (DU 145). *International Journal of Cancer*, 21(3), pp.274–281.

Sturtevant, A., 1913. *The Journal of Experimental Zoology, Volume 14*, Wiley-Liss.

Takata, R., Akamatsu, S., Kubo, M., Takahashi, A., Hosono, N., Kawaguchi, T., Tsunoda, T., Inazawa, J., Kamatani, N., Ogawa, O., Fujioka, T., Nakamura, Y. & Nakagawa, H., 2010. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nature Genetics*, 42(9), pp.751–754.

Tefferi, A., 2000. Myelofibrosis with myeloid metaplasia. *The New England journal of medicine*, 342(17), pp.1255–65.

Tefferi, A., 2010. Novel mutations and their functional and clinical relevance in myeloproliferative neoplasms: JAK2, MPL, TET2, ASXL1, CBL, IDH and IKZF1. *Leukemia*, 24(6), pp.1128–38.

Tefferi, A., Lim, K.-H., Abdel-Wahab, O., Lasho, T.L., Patel, J., Patnaik, M.M., Hanson, C.A., Pardanani, A., Gilliland, D.G. & Levine, R.L., 2009. Detection of mutant TET2 in myeloid malignancies other than myeloproliferative neoplasms: CMML, MDS, MDS/MPN and AML. *Leukemia*, 23(7), pp.1343–1345.

Tefferi, A. & Murphy, S., 2001. Current opinion in essential thrombocythemia: pathogenesis, diagnosis, and management. *Blood reviews*, 15(3), pp.121–31.

Tefferi, A., Thiele, J. & Vardiman, J.W., 2009. The 2008 World Health Organization classification system for myeloproliferative neoplasms: order out of chaos. *Cancer*, 115(17), pp.3842–7.

The 1000 Genomes Project Consortium, Abecasis, G., R, A., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. & McVean, G.A., 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp.1061–73.

The 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. & McVean, G.A., 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), pp.56–65.

The ENCODE Consortium, Myers, R.M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Good, P.J., et al., 2011. A user's guide to the Encyclopedia of DNA elements (ENCODE). *PLoS Biology*, 9(4).

The ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

The International HapMap Consortium, 2005. A haplotype map of the human genome. *Nature*, 437(7063), pp.1299–320.

Thomas, G., Jacobs, K.B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., Crenshaw, A., Cancel-Tassin, G., Staats, B.J., Wang, Z., Gonzalez-Bosquet, J., Fang, J., Deng, X., Berndt, S.I., Calle, E.E., Feigelson, H.S., Thun, M.J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F.R., Giovannucci, E., Willett, W.C., Cussenot, O., Valeri, A., Andriole, G.L., Crawford, E.D., Tucker, M., Gerhard, D.S., Fraumeni, J.F., Hoover, R., Hayes, R.B., Hunter, D.J. & Chanock, S.J., 2008. Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics*, 40(3), pp.310–315.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Stamatoyannopoulos, J.A., et al., 2012. The accessible chromatin landscape of the human genome. *Nature*, 489(7414), pp.75–82.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S. & Raychaudhuri, S., 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, 45(2), pp.124–30.

Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B.E., Klein, R.J., Han, B. & Raychaudhuri, S., 2015. Disentangling the Effects of

Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics*, 97(1), pp.139–152.

Turner, J.J., Morton, L.M., Linet, M.S., Clarke, C.A., Kadin, M.E., Vajdic, C.M., Monnereau, A., Maynadié, M., Chiu, B.C.-H., Marcos-Gragera, R., Costantini, A.S., Cerhan, J.R. & Weisenburger, D.D., 2010. InterLymph hierarchical classification of lymphoid neoplasms for epidemiologic research based on the WHO classification (2008): update and future directions. *Blood*, 116(20), pp.e90–8.

Urayama, K.Y., Jarrett, R.F., Hjalgrim, H., Diepstra, A., Kamatani, Y., McKay, J.D., et al., 2012. Genome-wide association study of classical hodgkin lymphoma and epstein-barr virus status-defined subgroups. *Journal of the National Cancer Institute*, 104(3), pp.240–253.

Varki, A., Lottenberg, R., Griffith, R. & Reinhard, E., 1983. The syndrome of idiopathic myelofibrosis. A clinicopathologic review with emphasis on the prognostic variables predicting survival. *Medicine*, 62(6), pp.353–71.

Vijai, J., Kirchhoff, T., Gallagher, D., Hamel, N., Guha, S., Darvasi, A., Lencz, T., Foulkes, W.D., Offit, K. & Klein, R.J., 2011. Genetic architecture of prostate cancer in the Ashkenazi Jewish population. *British journal of cancer*, 105(6), pp.864–9.

Vijai, J., Kirchhoff, T., Schrader, K.A., Brown, J., Dutra-Clarke, A.V., Manschreck, C., Hansen, N., Rau-Murthy, R., Sarrel, K., Przybylo, J., Shah, S., Cheguri, S., Stadler, Z., Zhang, L., Paltiel, O., Ben-Yehuda, D., Viale, A., Portlock, C., Straus, D., Lipkin, S.M., Lacher, M., Robson, M., Klein, R.J., Zelenetz, A. & Offit, K., 2013. Susceptibility Loci Associated with Specific and Shared Subtypes of Lymphoid Malignancies. *PLoS Genetics*, 9(1).

Wade, R., di Bernardo, M.C., Richards, S., Rossi, D., Crowther-Swanepoel, D., Gaidano, G., Oscier, D.G., Catovsky, D. & Houlston, R.S., 2011. Association between single nucleotide polymorphism-genotype and outcome of patients with chronic lymphocytic leukemia in a randomized chemotherapy trial. *Haematologica*, 96(10), pp.1496–1503.

Wang, S.S., Purdue, M.P., Cerhan, J.R., Zheng, T., Menashe, I., Armstrong, B.K., Lan, Q., Hartge, P., Kricker, A., Zhang, Y., Morton, L.M., Vajdic, C.M., Holford, T.R., Severson, R.K., Grulich, A., Leaderer, B.P., Davis, S., Cozen, W., Yeager, M., Chanock, S.J., Chatterjee, N. & Rothman, N., 2009. Common gene variants in the Tumor Necrosis Factor (TNF) and TNF receptor superfamilies and NF-kB transcription factors and non-hodgkin lymphoma risk. *PLoS ONE*, 4(4).

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), pp.57–63.

Ward, L.D. & Kellis, M., 2012. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40(D1).

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H., 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1).

Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Franke, L., et al., 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10), pp.1238–1243.

Whitaker, H.C., Kote-Jarai, Z., Ross-Adams, H., Warren, A.Y., Burge, J., George, A., Bancroft, E., Jhavar, S., Leongamornlert, D., Tymrakiewicz, M., Saunders, E., Page, E., Mitra, A., Mitchell, G., Lindeman, G.J., Evans, D.G., Blanco, I., Mercer, C., Rubinstein, W.S., Clowes, V., Douglas, F., Hodgson, S., Walker, L., Donaldson, A., Izatt, L., Dorkins, H., Male, A., Tucker, K., Stapleton, A., Lam, J., Kirk, J., Lilja, H., Easton, D., Cooper, C., Eeles, R. & Neal, D.E., 2010. The rs10993994 risk allele for prostate cancer results in clinically relevant changes in microseminoprotein-beta expression in tissue and urine. *PloS one*, 5(10), p.e13363.

Wooster, R., Neuhausen, S.L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T. & Averill, D., 1994. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science (New York, N.Y.)*, 265(5181), pp.2088–90.

Xu, J., Mo, Z., Ye, D., Wang, M., Liu, F., Sun, Y., et al., 2012. Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31.2 and 19q13.4. *Nature genetics*, 44(11), pp.1231–5.

Xu, X., 2014. *Functional characterization of common prostate cancer risk variants*. Cornell University.

Xu, X., Valtonen-André, C., Sävblom, C., Halldén, C., Lilja, H. & Klein, R.J., 2010. Polymorphisms at the Microseminoprotein-beta locus associated with physiologic variation in beta-microseminoprotein and prostate-specific antigen levels. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 19(8), pp.2035–42.

Xuan, J.W., Kwong, J., Chan, F.L., Ricci, M., Imasato, Y., Sakai, H., Fong, G.H., Panchal, C. & Chin, J.L., 1999. cDNA, Genomic Cloning, and Gene Expression Analysis of Mouse PSP94 (Prostate Secretory Protein of 94 Amino Acids). *DNA and Cell Biology*, 18(1), pp.11–26.

Yang, R., Naitoh, J., Murphy, M., Wang, H., Phillipson, J., Dekernion, J., Loda, M. & Reiter, R., 1998. Low p27 expression predicts poor disease-free survival in patients with prostate cancer. *The Journal of Urology*, 159(3), pp.941–945.

Yeager, M., Orr, N., Hayes, R.B., Jacobs, K.B., Kraft, P., Wacholder, S., Minichiello, M.J., Fearnhead, P., Yu, K., Chatterjee, N., Wang, Z., Welch, R., Staats, B.J., Calle, E.E., Feigelson, H.S., Thun, M.J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F.R., Giovannucci, E., Willett, W.C., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole, G.L., Gelmann, E.P., Tucker, M., Gerhard, D.S., Fraumeni, J.F., Hoover, R., Hunter, D.J., Chanock, S.J. & Thomas, G., 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics*, 39(5), pp.645–9.

6 Appendix

6.1 FunciSNP Analysis Code

The following R code was executed to perform the FunciSNP analysis. This analysis follows the vignette provided in the FunciSNP documentation (Coetzee et al. 2012).

```
#Load package.
> library(FunciSNP)

#Import SNPS
> lymphoma.snp <-
  file.path("/Users/hayes/Documents/projects/funciAnalysis/lymphoma.s
  np")
> lsnp <- read.delim(file=lymphoma.snp,sep="\t",header=FALSE)
> lymphoma.bio <-
  file.path("/Users/hayes/Documents/projects/funciAnalysis/gm12878Fil
  es/")

# Load Biofeatures
> segwayprom.filename <- list.files(lymphoma.bio, pattern='.bed$')[1]
> segwayStrongEnhancer.filename <- list.files(lymphoma.bio,
  pattern='.bed$')[2]
> encodeDnase.filename <- list.files(lymphoma.bio,
  pattern='.bed$')[3]
> encodeProm.filename <- list.files(lymphoma.bio, pattern='.bed$')[4]
> encodeStrongEnhancer4.filename <- list.files(lymphoma.bio,
  pattern='.bed$')[5]
> encodeStrongEnhancer5.filename <- list.files(lymphoma.bio,
  pattern='.bed$')[6]
> SegwayProm <- read.delim(file=paste(lymphoma.bio,
  segwayprom.filename, sep="/"), sep="\t", header=FALSE)
> SegwayStrongEnhancer <- read.delim(file=paste(lymphoma.bio,
  segwayStrongEnhancer.filename, sep="/"), sep="\t", header=FALSE)
> EncodeDNase <- read.delim(file=paste(lymphoma.bio,
  encodeDnase.filename, sep="/"), sep="\t", header=FALSE)
> EncodeProm <- read.delim(file=paste(lymphoma.bio,
  encodeProm.filename, sep="/"), sep="\t", header=FALSE)
> EncodeStrongEnhancer4 <- read.delim(file=paste(lymphoma.bio,
  encodeStrongEnhancer4.filename, sep="/"), sep="\t", header=FALSE)
> EncodeStrongEnhancer5 <- read.delim(file=paste(lymphoma.bio,
  encodeStrongEnhancer5.filename, sep="/"), sep="\t", header=FALSE)

# Generate lymphoma SNP object.
> lymphoma <- getFSNPs(snp.regions.file=lymphoma.snp,
  bio.features.loc=lymphoma.bio, built.in.biofeatures=FALSE)
> lymphoma.anno <- FunciSNPAnnotateSummary(lymphoma)
> ly.anno <- lymphoma.anno
> rownames(ly.anno) <- c(1:length(rownames(ly.anno)))
```